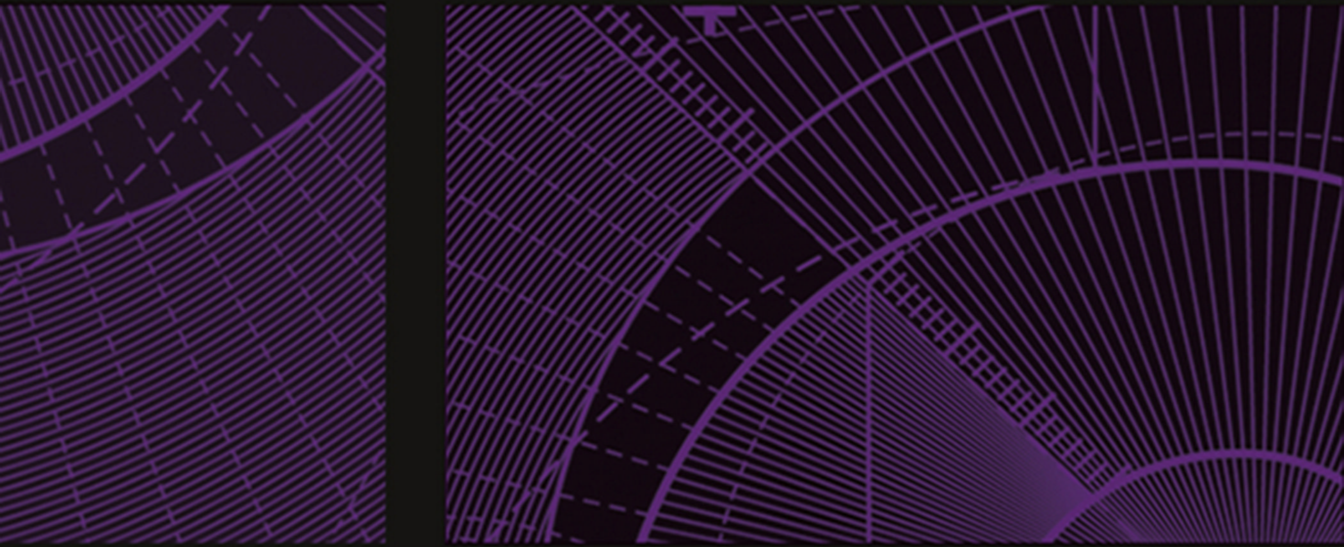
HANDBOOK OF MEASUREMENT

IN SCIENCE AND ENGINEERING

Volume 3



EDITED BY MYER KUTZ

WILEY

HANDBOOK OF MEASUREMENT IN SCIENCE AND ENGINEERING

HANDBOOK OF MEASUREMENT IN SCIENCE AND ENGINEERING

Volume 3

Edited by

MYER KUTZ

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Handbook of Measurement in Science and Engineering / Myer Kutz, editor.
volumes cm
Includes bibliographical references and index.
ISBN 978-0-470-40477-5 (volume 1) – ISBN 978-1-118-38464-0 (volume 2) – ISBN 978-1-118-38463-3 (set) – ISBN 978-1-118-64724-0 (volume 3) 1. Structural analysis (Engineering) 2. Dynamic testing. 3. Fault location (Engineering) 4. Strains and stresses–Measurement. I. Kutz, Myer.
TA645.H367 2012
620′.0044–dc23
2012011739

Set in 10.5/13.5pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Mark Berger, teacher

CONTENTS

VOLUME 3

LIST OF CONTRIBUTORS

M. Al Akerman, Emco-Williams Inc, Knoxville, TN, USA

Stephen W. Allison, Emco-Williams Inc, Knoxville, TN, USA

Beniamino Barbieri, ISS, Champaign, IL, USA

David L. Beshears, Emco-Williams Inc, Knoxville, TN, USA

Sandya Beeram, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Cong Bi, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

John D. Bullough, Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY, USA

Grant Bunker, Department of Physics, Illinois Institute of Technology, Chicago, IL, USA

Michael R. Cates, Emco-Williams Inc, Knoxville, TN, USA

Dominic M. Desiderio, The Charles B. Stout Neuroscience Mass Spectrometry Laboratory, Department of Neurology, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, USA

Marco Faifer, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

Alessandro Ferrero, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

Robert H. Giles, Biomedical Terahertz Technology Center and Submillimeter-Wave Technology Laboratory, Department of Physics and Applied Physics, University of Massachusetts Lowell, Lowell, MA, USA

Thomas M. Goyette, Submillimeter-Wave Technology Laboratory, Department of Physics and Applied Physics, University of Massachusetts Lowell, Lowell, MA, USA

David S. Hage, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Charles D. Hoyle, Jr., Department of Physics and Astronomy, Humboldt State University, Arcata, CA, USA

Cecil S. Joseph, Biomedical Terahertz Technology Center, Department of Physics and Applied Physics, University of Massachusetts Lowell, Lowell, MA, USA

Ellis Kaufmann, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Byung I. Kim, Department of Physics, Boise State University, Boise, ID, USA

Judah Levine, Time and Frequency Division and JILA, NIST and the University of Colorado, Boulder, CO, USA

Zhao Li, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Nicola Locci, Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Ryan Matsuda, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Kenneth R. Metz, Chemistry Department, Merkert Chemistry Center, Boston College, Chestnut Hill, MA, USA

Carlo Muscas, Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Kwok-Wai Ng, Department of Physics and Astronomy, University of Kentucky, Lexington, KY, USA

James E. Parks, Department of Physics, University of Tennessee, Knoxville, TN, USA

Maria Podariu, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Yevgen Povrozin, ISS, Champaign, IL, USA

Ray Radebaugh, Applied Chemicals and Materials Division, National Institute of Standards and Technology, Boulder, CO, USA

Elliott Rodriguez, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

Matthew B. Scudiere, Emco-Williams Inc, Knoxville, TN, USA

W. Rudolf Seitz, Department of Chemistry, University of New Hampshire, Durham, NH, USA

Gargi Sharma, Biomedical Terahertz Technology Center, Department of Physics and Applied Physics, University of Massachusetts Lowell, Lowell, MA, USA

Paul J. Simmonds, Departments of Physics & Materials Science and Engineering, Boise State University, Boise, ID, USA

Clair J. Sullivan, Department of Nuclear, Plasma, and Radiological Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Brad Swarbrick, Quality by Design Consultancy, Sydney, New South Wales, Australia

Frank Westad, CAMO Software AS, Oslo, Norway

Xianquan Zhan, Key Laboratory of Cancer Proteomics of Chinese Ministry of Health, Xiangya Hospital, Central South University, Changsha, P. R. China

Xiwei Zheng, Department of Chemistry, University of Nebraska, Lincoln, NE, USA

PREFACE

The idea for the *Handbook of Measurement in Science and Engineering* came from a Wiley book first published over 30 years ago. It was *Fundamentals of Temperature, Pressure and Flow Measurements*, written by a sole author, Robert P. Benedict, who also wrote Wiley books on gas dynamics and pipe flow. Bob was a pleasant, unassuming, and smart man. I was the Wiley editor for professional-level books in mechanical engineering when Bob was writing such books, so I knew him as a colleague. I recall meeting him in the Wiley offices at a time when he seemed to be having some medical problems, which he was reluctant to talk about. Recently, I discovered a book published in 1972 by a London firm, Pickering & Inglis, which specializes in religion. This book was *Journey Away from God*, an intriguing title. The author's name was Robert P. Benedict. I do not know whether the two Benedicts are in fact the same person, although Amazon seems to think so. (See the Robert P. Benedict page.) In any case, I do not recall Bob's mentioning the book when we had an occasion to talk.

The moral of this story, if there is one, is that the men and women who contributed the chapters in this handbook are real people, who have real-world concerns, in addition to the expertise required to write about technology. They have families, jobs, careers, and all manner of cares about the minutia of daily life to deal with. And that they have been able to find the time and energy to write these chapters is remarkable. I salute them. I have spent a lot of time in my life writing and editing books. I wrote my first Wiley book somewhat earlier than Bob Benedict wrote his. When Wiley published *Temperature Control* in 1967, I was in my mid-twenties and was a practicing engineer, working on temperature control of the Apollo inertial guidance system at the MIT Instrumentation Lab, where I had done my bachelor's thesis. One of the coauthors of my book was to have been a Tufts Mechanical Engineering Professor by the name of John Sununu (yes, that John Sununu), but he and the other coauthor dropped out of the project before the contract was signed. So I wrote the short book myself.

Bob Benedict's measurement book, the third edition of which is still in print, surfaced several years ago, during a discussion I was having with one of my Wiley editors at the time, Bob Argentieri, about possible projects we could collaborate on. It turned out that no one had attempted to update Benedict's book. I have not been a practicing engineer for some time, so I was not in a position to do an update as a single author—or even with a collaborator or two. Most of my career life has been in scientific and technical publishing, however, and for over a decade I have conceived of, and edited, numerous handbooks for several publishers. (I also write fiction, but that is another story.) So, it was natural for me to think about using Benedict's book as the kernel of a much larger and broader reference work dealing with engineering measurements. The idea, formed during that discussion, that I might edit a contributed handbook on engineering measurements took hold, and with the affable and expert guidance of my other Wiley editor at the time, George Telecki, the volume you are holding in your hands, or reading on an electronic device, came into being.

Like many such large reference works, this handbook went through several iterations before the final table of contents was set, although the general plan for arrangement of chapters has been the same throughout the project. The initial print version of the handbook was divided into two volumes. The chapters were arranged essentially by engineering discipline. The first volume contains 30 chapters related to five engineering disciplines, which are divided into three parts:

Part I, "Civil and Environmental Engineering," which contains seven chapters, all but one of them dealing with measurement and testing techniques for structural health monitoring, GIS and computer mapping, highway bridges, environmental engineering, hydrology, and mobile source emissions (the exception being the chapter on traffic congestion management, which describes the deployment of certain measurements)

Part II, "Mechanical and Biomedical Engineering," which contains 16 chapters, all of them dealing with techniques for measuring dimensions, surfaces, mass properties, force, resistive strain, vibration, acoustics, temperature, pressure, velocity, flow, heat flux, heat transfer for nonboiling two-phase flow, solar energy, wind energy, human movement, and physiological flow

Part III, "Industrial Engineering," which contains seven chapters dealing with statistical quality control, evaluating and selecting technology-based projects, manufacturing systems evaluation, measuring performance of chemical process equipment, industrial energy efficiency, industrial waste auditing, and organizational performance measurement

The second volume contains 23 chapters divided into three parts:

Part IV, "Materials Properties and Testing," which contains 15 chapters dealing with measurement of viscosity, tribology, corrosion, surface properties, and thermal conductivity of engineering materials; properties of metals, alloys, polymers, and particulate composite materials; nondestructive inspection; and testing of metallic materials, ceramics, plastics, and plastics processing

Part V, "Instrumentation," which contains five chapters covering electronic equipment used for measurements

Part VI, "Measurement Standards," which contains three chapters covering units and standards, measurement uncertainty, and error analysis

This new, third volume of the handbook expands the range of the handbook to cover measurements in physics, electrical engineering, and chemistry. This volume contains 20 chapters divided into two major parts:

Part VII, "Physics and Electrical Engineering," which contains 11 chapters, covering laser-based measurement systems, scanning probe microscopy, scanning tunneling microscopy, photometry, detection and measurement of ionizing radiation, time measurement systems, laboratory-based gravity measurement, cryogenic measurements, temperature-dependent fluorescence measurements, measurement of electrical quantities, and electrical power measurement

Part VIII, "Chemistry" which contains nine chapters, covering chemometrics/ chemical metrology, liquid chromatography, mass spectrometry measurements, basic principles of fluorescence spectroscopy, X-ray absorption spectroscopy, NMR spectroscopy, near-infrared spectroscopy, nanomaterials properties, and chemical sensing

Thanks to Brett Kurzman, my new editor for this volume, and Kari Capone, Allison, McGinniss, and Alex Castro for shepherding the manuscript toward production and to the stalwarts Kristen Parrish and Shirley Thomas for bringing this handbook volume home. Thanks, also, to my wife Arlene, who helps me with everything else.

MYER KUTZ

Delmar, NY
October 2015

PART VII

PHYSICS AND ELECTRICAL ENGINEERING

54

LASER MEASUREMENT TECHNIQUES

Cecil S. Joseph[1], Gargi Sharma[1], Thomas M. Goyette[2], and Robert H. Giles[1,2]

[1] *Biomedical Terahertz Technology Center, Department of Physics and Applied Physics, University of Massachusetts Lowell, Lowell, MA, USA*

[2] *Submillimeter-Wave Technology Laboratory, Department of Physics and Applied Physics, University of Massachusetts Lowell, Lowell, MA, USA*

54.1 INTRODUCTION

54.1.1 History and Development of the MASER

Lasers have enabled us to investigate the structure of atoms and molecules as well as accurately measure fundamental constants and observe natural phenomenology over a broad range of frequencies. But the applications of laser measurement technologies have advanced rapidly, for the instrumentation is little more than half a century in the making, and many laser-based measurement systems are now commercially available.

Albert Einstein first suggested the concept of stimulated emitted radiation in 1916. Indicating that a photon could cause the energy transition of an atom from an upper level to a lower level, Einstein proposed that the atom would emit a photon with the same energy as the photon initiating the energy transition. It was not until 1928 that Ladenburg observed stimulated emission.

As a faculty member at Columbia University, Charles H. Townes applied the concept to stimulating molecules in a resonant cavity to generate microwave radiation. His insight enabled postdoctoral fellow Herbert Zeiger and graduate student James P. Gordon to build a working maser by 1953. In 1954, Prokhorov and Basov of Moscow's Lebedev Physical Institute published the complete details for establishing

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

stimulated microwave emission. Prokhorov, Basov, and Townes shared the 1964 Nobel Prize for their research.

54.1.1.1 From Maser to Laser: Extending the Operable Region The science of stimulated emissions at wavelengths shorter than the microwave regime is very different. Though many physicists were now in pursuit of the technology, it was Charles Townes and Bell Lab's researcher Arthur L. Schawlow who first published the requirements for generating visible radiation in 1959. Detailing the parameters such as cavity structures, spontaneous emission ratios, and transition energy levels for generating visible radiation, Townes and Schawlow filed a patent for their development of "optical masers." However, Gordon Gould's graduate work at Columbia University predated Townes and Schawlow's patent and publication so after extended litigation Gould was awarded patents on the optical pumping techniques and his cavity designs using Brewster windows and a number of applications of what he referred to as "lasers."

While most researchers were building gas lasers at this point, Theodore H. Maiman was investigating the energy levels of ruby crystals. By the mid-1960s he demonstrated the first solid-state laser using a rod of synthetic ruby, thereby expanding the possibilities of light sources and the type of devices available. Through continued innovations by the research community, the performance characteristics of lasers were optimized, and these devices became the primary source for a rapidly growing market of measurement technologies.

54.1.2 Basic Laser Physics

Though there exist a large variety of laser devices that cover the electromagnetic spectrum from the microwave to the ultraviolet (higher frequency, i.e., X-ray lasers are also under development), the core elements of any laser device are:

1. A laser medium
2. A pump process
3. Optical feedback elements

Different approaches to these basic elements and varied combinations thereof lead to the output frequency range of the device. Put simply, the pump process injects energy into the laser media and via optical feedback techniques, and part of this energy is recovered from the media in the form of coherent, stimulated photons, which make up a laser beam. Here we discuss the basic requirements of these elements and how they combine to form a laser.

54.1.2.1 Stimulated Emission and Atomic Rate Equations A laser medium is made up of a collection of atoms/molecules in a gas, liquid, or solid phase. When the medium absorbs energy (e.g., by heating), some of the atoms transition to higher energy levels in the quantum mechanical energy structure. A fraction of these atoms spontaneously lose energy by emitting a photon and transition back to a lower energy state. In general, the spontaneous decay rate of any state is proportional to the number of atoms in that state. So if $N_i(t)$ is the instantaneous population of an energy level, E_i, the spontaneous decay rate is given by $dN_i/dt_{\text{Spontaneous}} = -\gamma_i N_i$, where γ_i is the spontaneous decay rate of the energy level E_i with $\gamma_i = 1/\tau_i$ and τ_i being the lifetime of the state. (Note that both radiative and nonradiative transitions are allowed and implicitly included. Emitted photons are radiative transitions, while dissipation of energy via lattice phonons is nonradiative transitions.)

It is also possible to stimulate absorption from a lower energy level to a higher energy level by providing a photon that corresponds to the energy difference between the levels (i.e., $\Delta E = h\upsilon$, where h is the Planck constant and υ is the photon frequency). This stimulates both absorption and emission at the applied signal frequency. The primary difference between stimulated transitions and spontaneous transitions is that stimulated transitions are caused by an applied signal and as such the photons emitted are coherent with that signal. Spontaneous transitions, on the other hand, are radiated out by atoms driven independently of each other and as such are incoherent.

Consider a two-energy-level system, where level E_i has a population N_i and E_j has a population N_j. Moreover, let $E_j > E_i$, corresponding to an energy difference, ΔE_{ji}. The spontaneous decay rate of level E_j to E_i is given by γ_{ji}. If we now apply a signal to this system that corresponds to the energy difference, that is, $\Delta E_{ji} = h\upsilon_{ji}$, then we stimulate absorption from E_i to E_j, and we stimulate emission from E_j to E_i. If $n(t)$ is the photon density of the incident signal, then the change in population of the higher energy state is given by

$$\begin{aligned}\frac{dN_j(t)}{dt} &= \text{stimulated absorption from } E_i - \text{stimulated emission from } E_j \\ &\quad - \text{spontaneous emission from } E_j,\end{aligned}$$

that is,

$$\frac{dN_j(t)}{dt} = Kn(t)N_i(t) - Kn(t)N_j(t) - \gamma_{ji}N_j(t)$$

where K is a constant that measures the strength of the stimulated response.

Thus the rate at which atoms make *stimulated* transitions is proportional to the population difference of the two energy levels and the applied signal intensity. Each absorbed photon attenuates the applied signal, and each emitted photon amplifies it.

Thus the change in photon density of the applied signal due to stimulated transitions can be represented as $dn/dt = -K\Delta N_{ij}n(t)$, where ΔN_{ij} is the population difference between the lower and upper energy state. When there are more atoms in the lower energy state, $\Delta N_{ij} > 0$, and the applied signal is attenuated. If, however, $\Delta N_{ij} < 0$, the applied signal is amplified; this condition that requires a larger number of atoms in the higher energy state for signal amplification is one of the basic requirements of a working laser and is referred to as *population inversion*.

Before discussing the means to achieving population inversion in laser media, it is important to consider the difference between spontaneously emitted photons and stimulated photons. Stimulated transitions are caused by an applied signal and as such are coherent with the applied signal, unlike spontaneous emission, which is incoherent.

54.1.2.2 Population Inversion and Laser Amplification Population inversion is an essential condition for laser amplification, that is, there must be a larger number of atoms in the upper excited state than the lower in order for the applied signal to be amplified. According to the Boltzmann principle, the relative populations of any two energy levels are given by $N_2/N_1 = e^{-\Delta E/kT}$ where $\Delta E = E_2 - E_1$ and $E_2 > E_1$. Therefore in order to create population inversion, we need to "pump" atoms into the upper excited state since the population difference is always attenuating at equilibrium for a two-level system. Typically other upper energy levels are used to "feed" the upper lasing level, thereby creating inversion.

As an example, consider the lasing levels and population inversion in a ruby laser. Ruby is essentially sapphire (Al_2O_3) doped with chromium. The Cr^{3+} ions replace a fraction of the Al^{3+} ions in the sapphire lattice. These Cr^{3+} ions in the lattice have energy levels in the red at approximately 694 nm. A representative energy-level diagram is shown in Figure 54.1.

In a typical ruby laser, atoms are optically pumped using a xenon arc lamp from the ground state to higher energy levels. Atoms in these levels relax rapidly down to highly metastable "R" levels via nonradiative transitions. With sufficient pumping, it is

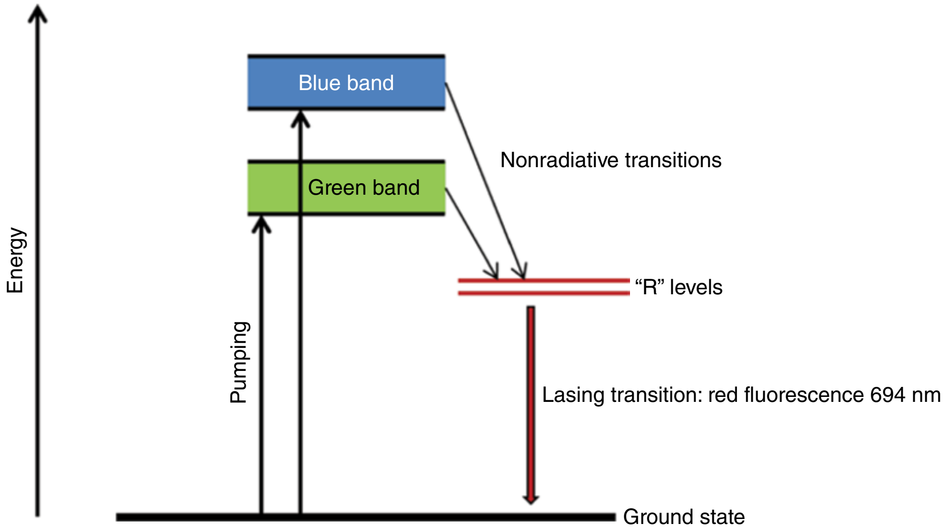


FIGURE 54.1 Energy-level transitions in ruby.

possible to create population inversion between the metastable (long lifetime) R levels and the ground state. This transition can then be amplified.

The reason inversion occurs without violating the Boltzmann principle is that through sufficient pumping more than half of the atoms transfer from the ground state to the higher energy states. These states have very short lifetimes and relax with close to 100% quantum efficiency down to the R levels. Since the R levels have a relatively long lifetime (≈4.3 ms), they are being fed with atoms at a significantly faster rate than atoms decaying (spontaneously) from them, leading to population inversion [1].

54.1.2.3 Laser Cavity or Oscillator In order to amplify stimulated emission as photons pass through laser media, one must generate population inversion using a pump process. With the condition of inversion met, any signal at the lasing (inversion) frequency that passes through the gain media is slightly amplified. If the media are then enclosed in an oscillator cavity that forces this signal to repeatedly pass back and forth through the gain media, then the signal strength is amplified in each pass as long as inversion exists.

For example, suppose a laser medium pumped to maintain population inversion is enclosed in an optical cavity of length L as shown in Figure 54.2. Initially there is always some spontaneous emission at the laser frequency. However, if one of these spontaneously emitted photons is aligned with the optical axis of the cavity, then while it travels through the medium it causes stimulated emission along the same axis. When this radiation reflects from either end of the cavity, it retraces its path through the gain media and is further amplified by stimulated emission. This process continues and the stimulated transition at the laser frequency in a direction aligned with the optical axis of the cavity is continuously amplified until the stimulated emissions negate population inversion.

In general the amplification process will continue until the pump process can maintain population inversion. That is the point at which the net gain in any round trip through the medium is balanced by the net loss; this is the steady-state condition. In order to extract a portion of the stimulated signal, the mirrors at either end of the cavity are generally partially transmitting. For the system shown in Figure 54.2, r_1 and r_2 are

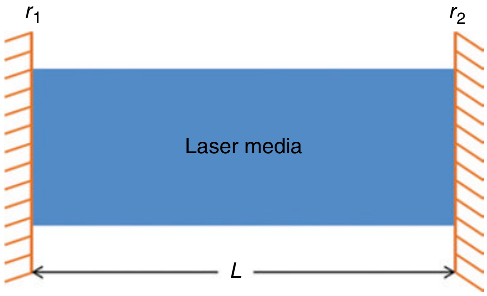


FIGURE 54.2 A laser medium enclosed in an optical cavity of length L.

the Fresnel reflection coefficients of the mirror surfaces. Then, the previously discussed steady state requires that the amplitude gain in one round trip through the length L cavity be equal to 1, that is,

$$r_1 r_2 e^{2\alpha L} = 1,$$

where α is the gain of the laser medium.

Note that r_1, r_2, and L are characteristics of the optical cavity; thus the steady-state amplitude condition yields the laser gain coefficient, α, that must be maintained in that specific cavity:

$$\alpha = \frac{1}{2L}\ln\left(\frac{1}{r_1 r_2}\right) = \frac{1}{4L}\ln\left(\frac{1}{R_1 R_2}\right)$$

where R_1 and R_2 are the power reflectivity of the two mirrors.

The laser gain coefficient depends on the population inversion in the laser media. The greater the inversion, the larger the gain. Assuming that the lasing transition has a Lorentzian line shape, then the *threshold inversion density* (ΔN_t) required for lasing action in a cavity is given by

$$\Delta N_\mathrm{t} = \frac{\pi \Delta\omega}{\lambda^2 \gamma L}\ln\left(\frac{1}{R_1 R_2}\right)$$

where $\Delta\omega$ is the transition linewidth, λ is the transition wavelength, and γ is the transition decay rate [1].

Steady state also introduces a round-trip phase condition for the oscillator. The steady-state frequencies must correspond to standing waves in the oscillator. For the simple cavity shown in Figure 54.2, this implies that the length L must be an integral number of half wavelengths for sustained oscillations, that is,

$$L = \frac{m\lambda}{2};\ m = \text{integer}$$

which yields discrete axial mode frequencies given by

$$\omega_m = m 2\pi\left(\frac{c}{2L}\right)$$

For most lasers, several axial mode frequencies exist within the transition linewidth and m is a large integer.

Finally, we also need to consider the wave nature of light and its effects on the transverse spatial properties of the laser beam. The laser beam oscillating in the cavity has a spatial extent along the transverse axis and, as it oscillates, will spread due to diffraction. After the beam size exceeds the cavity reflectors, the lost signal can be accounted for as lowered reflectivity or increased round-trip propagation loss. However, the pattern still needs to be self-consistent over one round trip, that is, the beam parameters must also match for a sustainable laser mode. Therefore, as the reflections on the ends introduce changes to the transverse mode profile of the beam, sustained oscillations depend upon the geometry of the cavity. The laser resonator system discussed so far is essentially two flat mirrors at the ends of the cavity; however, several other cavity designs are possible, such as curved mirrors, which focus the radiation in the transverse direction (see Suggested Readings at the end of this section).

In general, for any given laser cavity, one can find a set of discrete transverse eigenmodes that can propagate self-consistently in that cavity. Thus, for any working laser, one requires a pumping process that is capable of sustaining population inversion above a cavity defined threshold in the lasing media. The laser cavity or resonator will also define the axial frequencies that are capable of oscillations, and the cavity's geometry will then dictate which transverse modes are capable of maintaining sustained oscillations in the system.

54.1.3 Laser Beam Characteristics

When extracting the output beam from a laser cavity for an application, it is the frequency and amplitude characteristics of the output radiation that are of practical relevance. Moreover, it is also important to understand the propagation of laser beams through space and optical elements. We start by describing the frequency and amplitude characteristics of the output radiation and then discuss the solution of the wave equation and explore the fundamental mode solution. Finally we cover a basic introduction to the ABCD law for beam propagation.

54.1.3.1 Laser Frequency and Amplitude Characteristics Ideally, the output of a laser is essentially single frequency, amplitude stabilized, and highly directional. In practice however, the cavity and laser setup introduce slight variations from the ideal output. As previously described, the laser cavity defines the allowed axial modes of oscillation that ultimately set the laser frequency. In general, real lasers can operate at several axial modes, still within the atomic linewidth of the lasing transition. Several steps can be taken to improve frequency stability in real laser systems, and the ultimate limit on frequency stability is set by the spontaneous emission in the gain media. Amplitude stability is generally achieved by the gain balancing mechanism (round-trip gain = 1). In practical systems, slight amplitude instability is introduced by pumping mechanisms and thermal cavity effects. Also, laser beams are temporally and spatially coherent.

54.1.3.2 Fundamental Mode Several laser modes can be supported by the lasing cavity, provided the field component, u, of the laser beam satisfies the scalar wave equation [2]

$$\nabla^2 u + k^2 u = 0, \quad \text{where } k = 2\pi/\lambda \text{ is the propagation constant.}$$

Assuming that the function varies slowly along the propagation axis, the second derivative along this axis $\left(\frac{\partial^2}{\partial z^2}\psi\right)$ can be neglected, and we get

$$\frac{\partial^2}{\partial x^2}\psi + \frac{\partial^2}{\partial y^2}\psi - 2ik\frac{\partial}{\partial z}\psi = 0$$

One solution to the above equation is a Gaussian beam profile:

$$\psi = \exp\left\{-i\left(P + \frac{k}{2q}r^2\right)\right\}$$

where $r^2 = x^2 + y^2$ and P and q are functions of z.

The parameter q is called the complex beam parameter and it can be related to real beam parameters R and w using the following relation:

$$\frac{1}{q} = \frac{1}{R} - i\frac{\lambda}{\pi w^2}$$

Both R and w are functions of position on the propagation axis z. $R(z)$ represents the radius of curvature of the wavefront at z, and $w(z)$ represents the "beam radius." Figure 54.3 shows the intensity distribution as a function of radial distance r from the propagation axis. Note that intensity is proportional to the square of the field amplitude.

The intensity of the laser is typically concentrated near the propagation axis. As shown in Figure 54.3, $w(z)$ measures the decrease in field strength with distance from

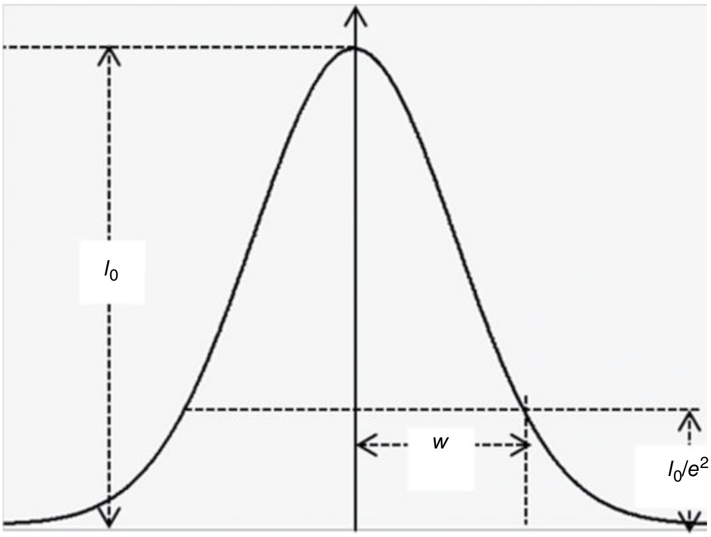


FIGURE 54.3 Intensity distribution of a Gaussian beam.

the propagation axis. In general, the "beam radius" corresponds to $w(z)$, which is defined as the radius at which the beam intensity drops to $1/e^2$ of its peak value. The beam diameter is $2w$.

The Gaussian beam contracts to a minimum value along the z-axis; the waist at this point is referred to as the "beam waist," w_0. Noting that at the minimum waist the wavefront is a plane wave, which indicates that $R = \infty$ at this point, thus the complex beam parameter q at this point (denoted q_0) becomes

$$q_0 = i\frac{\pi w_0^2}{\lambda}$$

The size and curvature of a wavefront as it propagates along the z-axis are described by [2]

$$w^2(z) = w_0^2\left[1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2\right]$$

and

$$R(z) = z\left[1 + \left(\frac{\pi w_0^2}{\lambda z}\right)^2\right]$$

From Figure 54.4 and the equations above we can see that the far-field diffraction angle, θ, is given by

$$\theta = \frac{\lambda}{\pi w_0}$$

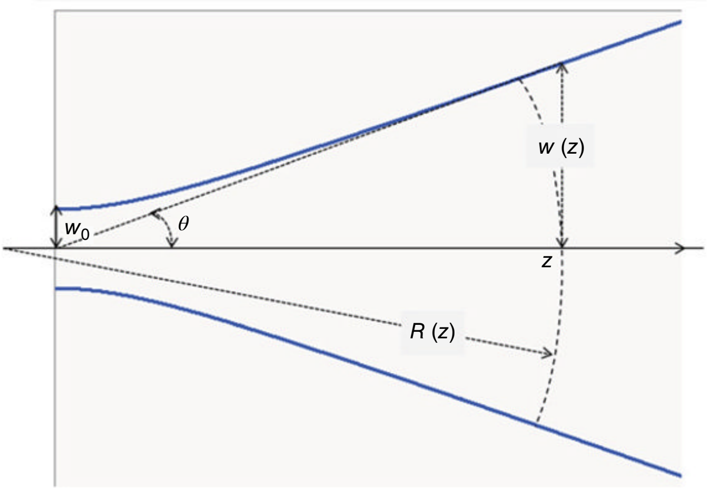


FIGURE 54.4 The contour of a Gaussian beam along the z-axis.

The solution to the wave equation can be expressed as [2]

$$u(r,z)=\frac{w_0}{w}\exp\left\{-i(kz-\Phi)-r^2\left(\frac{1}{w^2}+\frac{ik}{2R}\right)\right\}$$

where

$$\Phi=\tan^{-1}\left(\frac{\lambda z}{\pi w_0^2}\right)$$

This solution is not the only solution to the scalar wave equation; however, it is the desired output mode for most laser applications and is referred to as the "*fundamental mode*." More complicated solutions to the wave equation exist and are generally referred to as "higher-order" modes. Solving the equation in Cartesian coordinates for the higher-order modes generally yields a combination of Gaussian and Hermite functions [2]. Solving equation in cylindrical coordinates (for systems with cylindrical symmetry) yields generalized Laguerre polynomials for solutions [2].

54.1.3.3 Beam Propagation The path of a ray can be characterized by its radial distance from the optical axis (r) and by its slope with respect to the optical axis (θ), as described in Figure 54.5a. The path of this ray through an optical element is then dependent on the optical properties of the element and the input beam parameters. The parameters of the output beam can then be determined by using the ABCD matrix of the optical element:

$$\begin{bmatrix} r' \\ \theta' \end{bmatrix}=\begin{bmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{bmatrix}\begin{bmatrix} r \\ \theta \end{bmatrix}$$

where r and θ are parameters describing the input beam and r' and θ' describe the output beam. The ABCD matrix describes the optical properties of the element and is also known as the ray transfer matrix.

As the beam travels through the optical elements, the ray transfer matrices can be stacked to determine the output beam parameters. In Figure 54.5b, an input beam

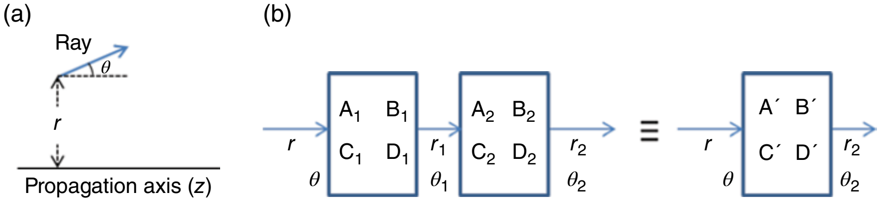


FIGURE 54.5 (a) Shows the beam parameters (r and θ) for a paraxial ray and (b) demonstrates stacking of ABCD matrices.

(parameters r and θ) travels through two optical elements sequentially. The stacked ABCD matrix for this system is given by

$$\begin{bmatrix} \mathrm{A}' & \mathrm{B}' \\ \mathrm{C}' & \mathrm{D}' \end{bmatrix} = \begin{bmatrix} \mathrm{A}_2 & \mathrm{B}_2 \\ \mathrm{C}_2 & \mathrm{D}_2 \end{bmatrix} \begin{bmatrix} \mathrm{A}_1 & \mathrm{B}_1 \\ \mathrm{C}_1 & \mathrm{D}_1 \end{bmatrix}$$

The same process can be generalized for N optical elements; if the input beam passes through N elements in sequence, going from element 1 to element N, and if the corresponding ray transfer matrices for each element are labeled $[\mathrm{T}_1]$, $[\mathrm{T}_2]$, …, $[\mathrm{T}_N]$, then the ray transfer matrix for the entire system $[\mathrm{T}_{\mathrm{tot}}]$ can be evaluated as

$$[\mathrm{T}_{\mathrm{tot}}] = [\mathrm{T}_N][\mathrm{T}_{N-1}]\ldots[\mathrm{T}_2][\mathrm{T}_1]$$

Table 54.1 gives the ABCD matrices of several common optical elements.

While the discussion has focused on ray optics, ray transfer matrices can also be used to map the effect of optical elements on laser beams. For a Gaussian laser beam, the complex q parameter is defined as

$$\frac{1}{q} = \frac{1}{R} - i\frac{\lambda}{\pi w^2}$$

where λ is the laser wavelength and $R(z)$ and $w(z)$ are the radius of curvature and beam radius at a point z on the optical axis.

If q_1 is the complex beam parameter for a Gaussian beam entering an optical system characterized by ray transfer matrix $\begin{bmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{bmatrix}$, then the output beam parameter q_2 is given by

$$q_2 = \frac{\mathrm{A}q_1 + \mathrm{B}}{\mathrm{C}q_1 + \mathrm{D}}$$

This is called the ABCD law for Gaussian beams.

TABLE 54.1 ABCD Matrices of Common Optical Elements[a]

Propagation through free space by a distance d	$\begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}$
Propagation through a medium of refractive index n by a distance d	$\begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix}$
Reflection at a mirror of the radius of curvature R for normal incidence	$\begin{pmatrix} 1 & 0 \\ -2/R & 1 \end{pmatrix}$
Thin lens of focal length "f"	$\begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}$
Dielectric interface	$\begin{pmatrix} 1 & 0 \\ 0 & n_1/n_2 \end{pmatrix}$

[a] Ref. 2.

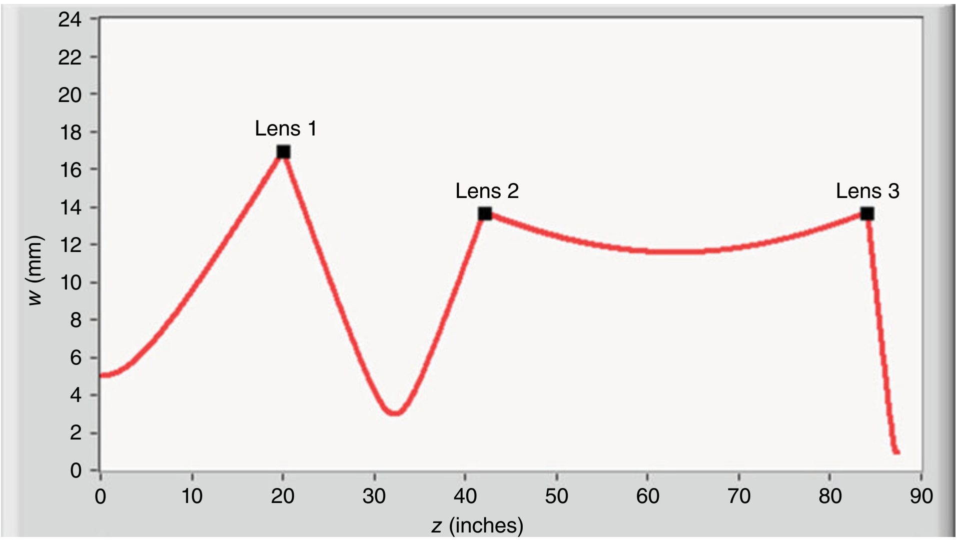


FIGURE 54.6 Tracking the beam diameter of a 500 μm Gaussian laser beam with propagation along the optical axis through three lenses using the ABCD law.

Using the ABCD law the size of a Gaussian beam propagating through an optical system can be mapped as a function of distance along the propagation axis. Figure 54.6 depicts the beam radius as function of propagation distance for a beam of wavelength 500 μm. The beam travels along the axis and interacts with three lenses of varying focal lengths placed at specific locations along the optical path. ABCD matrices for both the propagation distances and lenses determine the beam's profile along the propagation axis.

As shown in Figure 54.6, the beam diameter varies considerably with position. This variation can be controlled using focusing elements placed along the optical axis. For example, in Figure 54.6, lens 2 is used to collimate the Gaussian beam as it propagates through free space. Another aspect of laser optics that the ABCD law yields is the required optic size. As Figure 54.6 indicates, the beam size varies considerably with location on the optical axis; this defines the size of optical elements (mirrors, lenses) that are required to effectively redirect or reshape the beam. For a Gaussian beam of beam radius $w(z)$, a good rule of thumb for determining required optic size is that the diameter of the optic should be larger than $4w$ at that point on the axis. Smaller optics induce distortions on the beam profile.

54.1.4 Example: CO_2 Laser Pumped Far-Infrared Gas Laser Systems

As described in Section 54.1.1, lasers require a cavity/oscillator, a gain medium, and a pumping process. This pump process supplies the energy required to create inversion. This can be achieved using a flashlamp broadband source (as with the previously discussed ruby laser), an electrical voltage, or another laser as well. A specific example of this technique is far-infrared (FIR) gas lasers that can be "pumped" using CO_2 gas

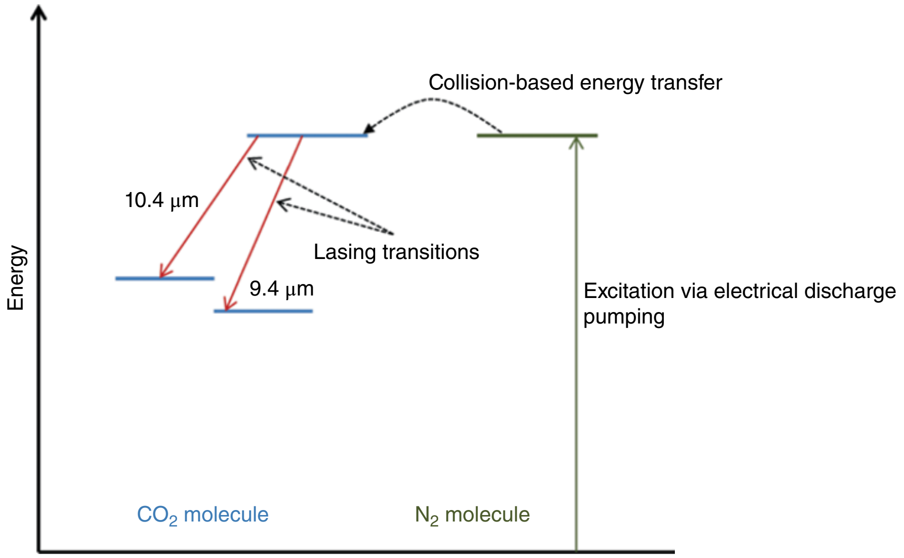

Collision-based energy transfer
10.4 µm
Lasing transitions
9.4 µm
Energy
Excitation via electrical discharge pumping
CO_2 molecule
N_2 molecule

FIGURE 54.7 Energy-level diagram for CO_2 gas laser.

lasers. By tuning the CO_2 laser's frequency for an appropriate selection of gain media (gas), single-frequency laser systems can be adjusted to cover the terahertz (FIR) portion of the electromagnetic spectrum. This example describes the setup of such a system, along with the relevant measurement techniques.

A CO_2 laser is a gas laser that lases at approximately 10 µm. The lasing medium is a mix of CO_2, nitrogen, and helium. The lasing transitions are in the CO_2 molecule. Figure 54.7 shows the energy-level diagram for the relevant lasing levels of the CO_2 molecules. Typical CO_2 lasers are pumped via a process called gas-discharge pumping, wherein a current is arced through the gas-discharge tube containing the gas mixture. Collisions between electrons in the discharge and nitrogen molecules excite the nitrogen molecules to a vibrational excitation mode that closely corresponds to the desired vibrational mode in the CO_2 molecule as shown in Figure 54.7. Collisions between the nitrogen and CO_2 molecules excite the CO_2 molecules to this excited metastable state (asymmetric stretch vibrational mode), and population inversion is obtained. The lower lasing level of a CO_2 laser is actually two-level bands: one corresponding to a transition of 10.4 µm (symmetric stretch) and the other corresponding to 9.4 µm (bending mode).

After the stimulated emission to the lower CO_2 levels, they are "deexcited" by collisions with helium molecules, which act as a buffer gas. Thus, the gas mix is of critical importance for achieving high laser output power with standard CO_2 lasers. The nitrogen is needed to effectively excite the upper lasing levels of the CO_2 molecule, and the helium is needed as a buffer gas to deexcite the CO_2 molecule from the lower lasing levels.

The output frequency of the CO_2 laser can be tuned within a certain frequency bandwidth. As pointed out in Section 54.1.2.3 any laser cavity has several axial modes; the

TABLE 54.2 Far-Infrared Laser Transitions Pumper by CO_2 Lasers[a]

Gas	Frequency (GHz)	CO_2 Pump Line
HCOOH formic	247.08	9.62P28
HCOOH formic	381.34	9.17R40
HCOOH formic	403.72	9.17R40
HCOOH formic	561.72	9.23R28
HCOOH formic	584.39	9.23R28
HCOOH formic	673.99	10.53P14
HCOOH formic	692.95	9.27R20
HCOOH formic	740.23	9.60P26
HCOOH formic	832.99	9.20R34
HCOOH formic	991.78	9.37R04
CH_2F_2 difluoromethane	586.17	9.23R28
CH_2F_2 difluoromethane	1110.32	9.26R22
CH_2F_2 difluoromethane	1267.08	9.35R06
CH_2F_2 difluoromethane	1272.12	9.21R32
CH_2F_2 difluoromethane	1397.12	9.20R34
CH_2F_2 difluoromethane	1546.08	9.26R22
CH_2F_2 difluoromethane	1626.6	9.21R32
CH_2F_2 difluoromethane	1891.27	9.47P10
CH_2F_2 difluoromethane	2237.3	9.57P22
CH_2F_2 difluoromethane	2447.97	9.26R22

[a] Adapted from various sources.

spacing between these axial modes depends upon the cavity length. Typical CO_2 lasers that are used for research also utilize a blazed grating as one of the reflectors in the cavity. Appropriate orientation of the grating end and cavity length allows for research-grade CO_2 lasers to lase specific transitions within a certain frequency bandwidth [3]. Figure 54.1 depicts the vibrational energy levels of the CO_2 molecule, and for each vibrational level there are several rotational levels. The J quantum number labels the rotational state of the molecule. When transitions take place between energy levels, certain quantum mechanical selection rules apply. These selection rules limit the change in the rotational quantum state (ΔJ) to span -1 to $+1$, though for many vibrational transitions $\Delta J=0$ is forbidden. To accurately describe which transition is lasing in a CO_2 laser, we need to define which vibrational transition occurred (10.4 or 9.4 μm), the change in J, and the final value of J. Knowing all of these quantities determines which exact rovibrational transition occurred (and hence the wavelength of the emitted photon). For CO_2 laser transitions the convention is to indicate which wavelength (9 or 10), followed by ΔJ (P for -1 and Q for $+1$), followed by the final J value, for example, 10P(20), is a 10.4 μm transition from a $J=21$ to $J=20$ rovibrational state in the CO_2 molecule.

The CO_2 laser can be used as a pump laser for FIR gas lasers. As shown in Table 54.2, the output frequency of CO_2 lasers can be tuned to efficiently pump lasing transitions in a gas laser. Selecting an appropriate laser gas media and CO_2 laser transition allows one to lase at a range of frequencies in the FIR. Table 54.2 lists some of these gases, the lasing frequency (FIR), and the CO_2 pump frequency. Notice that pumping different transitions in the same gas molecule leads to different FIR laser frequencies. Also note that sometimes the same pump laser frequency can excite multiple lasing transitions in the same gas. In such situations, the exact frequency of the FIR laser is determined by the relative gain of the lasing transitions and construction of the cavity. Reference textbooks [4] list measured transitions and the appropriate CO_2 pump frequencies.

54.1.5 Heterodyned Detection

Laser measurement systems focus on studying the interaction of laser radiation with different materials. For several applications it is the laser beam that has interacted with the sample (e.g., transmitted through or reflected from) that is of interest for the measurement. Detecting the laser beam requires a detector that is sensitive to light at the laser frequency. Depending on the frequency range of the laser, several detection options with varying sensitivities are available. The basic principle of the detector is that it produces a measurable output for a given amount of laser input.

One way to broadly classify detection techniques is coherent versus incoherent detection. For incoherent detection the output signal is dependent on the intensity of laser beam. Coherent detection schemes allow for the determination of phase information in the laser measurement as well as field strength. This section discusses heterodyned detection, which is a technique commonly used in telecommunications, astronomy, and FIR measurement systems for coherent detection.

The basic principle of heterodyned detection requires a detector that has a nonlinear output with input electric field amplitude. An example is a diode mixer, the output of which is proportional to the square of the electric field amplitude.

Consider a detector where the output current is proportional to the square of the input amplitude:

$$I \propto E^2$$

We have two input signals, a transmit signal given by

$$E_{\mathrm{t}} = E_{\mathrm{t}} \cos\left(\omega_{\mathrm{t}} t + \phi_{\mathrm{t}}\right)$$

where E_{t} is the amplitude of the transmit beam, ω_{t} is the frequency, and ϕ_{t} is the phase and a Local Oscillator (LO) signal given by

$$E_{\mathrm{LO}} = E_{\mathrm{LO}} \cos\left(\omega_{\mathrm{LO}} t + \phi_{\mathrm{LO}}\right)$$

where E_{LO} is the amplitude of the LO, ω_{LO} is the frequency, and ϕ_{LO} is the phase.

If these two signals are simultaneously incident on the detector, the output is then proportional to the square if the input amplitude is the superposition of these two signals:

$$I \propto \left[E_{\mathrm{t}} \cos(\omega_{\mathrm{t}} t + \phi_{\mathrm{t}}) + E_{\mathrm{LO}} \cos(\omega_{\mathrm{LO}} t + \phi_{\mathrm{LO}})\right]^2$$

$$\rightarrow I \propto E_{\mathrm{t}}^2 \cos^2(\omega_{\mathrm{t}} t + \phi_{\mathrm{t}}) + E_{\mathrm{LO}}^2 \cos^2(\omega_{\mathrm{LO}} t + \phi_{\mathrm{LO}}) + 2E_{\mathrm{t}} E_{\mathrm{LO}} \cos(\omega_{\mathrm{t}} t + \phi_{\mathrm{t}}) \cos(\omega_{\mathrm{LO}} t + \phi_{\mathrm{LO}})$$

Using trigonometric identities and rearranging the terms yield

$$\begin{aligned} I & \\ & \propto \underbrace{\frac{1}{2}\left(E_{\mathrm{t}}^2 + E_{\mathrm{LO}}^2\right)}_{\text{DC term}} \\ & + \underbrace{\left[\frac{E_{\mathrm{t}}^2}{2} \cos 2(\omega_{\mathrm{t}} t + \phi_{\mathrm{t}}) + \frac{E_{\mathrm{LO}}^2}{2} \cos 2(\omega_{\mathrm{LO}} t + \phi_{\mathrm{LO}}) + E_{\mathrm{t}} E_{\mathrm{LO}} \cos\left((\omega_{\mathrm{t}} + \omega_{\mathrm{LO}}) t + (\phi_{\mathrm{t}} + \phi_{\mathrm{LO}})\right)\right]}_{\text{High frequency terms}} \\ & + \underbrace{E_{\mathrm{t}} E_{\mathrm{LO}} \cos\left((\omega_{\mathrm{t}} - \omega_{\mathrm{LO}}) t + (\phi_{\mathrm{t}} - \phi_{\mathrm{LO}})\right)}_{\text{Intermediate frequency term}} \end{aligned}$$

If at this point the signal is passed through a low-pass filter and the DC term is subtracted, the output of the detector is at the difference frequency also called the Intermediate Frequency (IF):

$$I \propto E_{\mathrm{t}} E_{\mathrm{LO}} \cos(\omega_{\mathrm{IF}} t + \Delta\phi)$$

Thus looking for signal at the IF yields information corresponding to the amplitude and phase of the laser beam. Figure 54.8 shows the schematic for coherent single-frequency detection of the radiation reflected back from some target. The transmit laser and the LO are both FIR laser beams such as those generated by the FIR laser systems described in the previous sections. The reference diode and receiver diode are both nonlinear mixers such as Schottky diodes. The reason for a heterodyned approach is that FIR frequencies are too high for electronics to respond; thus a heterodyned system is used to down-convert the frequency of the laser signal.

In Figure 54.8, the purpose of the measurement is to measure the response (reflection) from the target at the transmit frequency (ω_{t}). The LO frequency is offset from the transmit frequency by an amount called the IF. A series of beam splitters (an optical device used for partially reflecting and transmitting beams, discussed later) is used to combine the transmit and LO signals at the two detectors. Following the beam path in Figure 54.8 shows that the transmit signal output from the laser source is overlapped with the LO at Beam Splitter 4 (BS4) and sent to the reference detector, while the transmit laser signal reflected from the target is combined with the same LO at BS5 and sent to the receiver diode. The reference diode is used to measure the amplitude

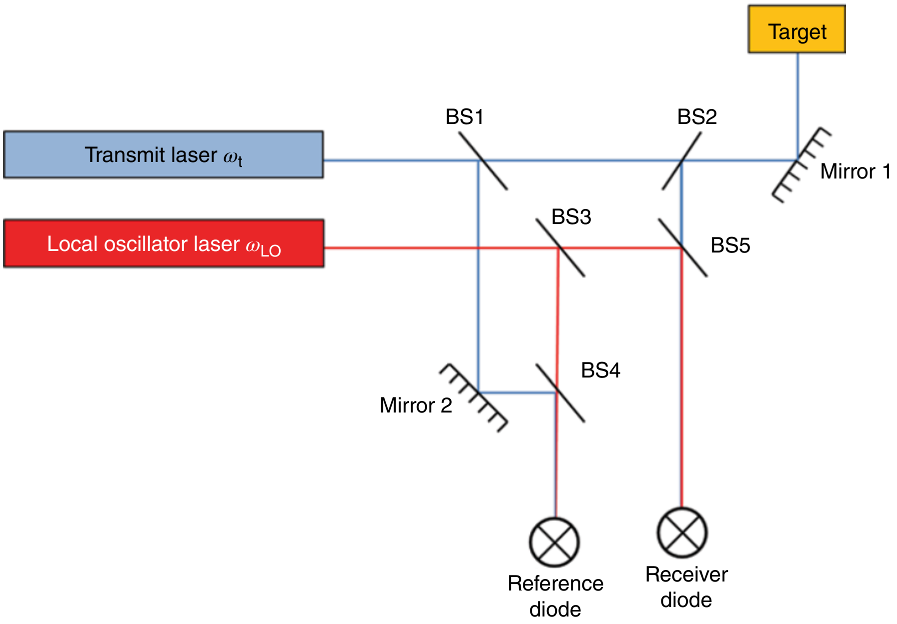


FIGURE 54.8 Schematic of heterodyned detection using a transmit laser and local oscillator.

and phase of the signal that has not been modified by the target, while the receiver diode measures the signal that has been modified by the sample. Comparing these two signals allows one to compute the change in amplitude and phase of the transmitted beam caused by the target.

A Schottky diode can be used as the required nonlinear mixer for heterodyned detection. Further details can be found in *Microwave Engineering* by Pozar [5]. The small signal current–voltage (IV) relationship of a Schottky diode is given as

$$I(V) = I_s\left(e^{\alpha V} - 1\right)$$

where $\alpha = q/nkT$, where q is electron charge, k is Boltzmann's constant, T is temperature and n is the ideality factor for the diode. I_s is the diode saturation current [5].

Now if the voltage across the diode is a DC bias voltage (V_{DC}) and a small AC voltage (v),

$$V = V_{\text{DC}} + v$$

Expanding the diode current as a Taylor series about V_{DC} and keeping the first three terms yield

$$I(V) = I_{\text{DC}} + v\left(\alpha\left(I_{\text{DC}} + I_s\right)\right) + \frac{1}{2}v^2\left(\alpha^2\left(I_{\text{DC}} + I_s\right)\right) + \cdots$$

This is termed the *small signal approximation* for diode current. Note that the output is proportional to the square of the input AC voltage as is required for a nonlinear mixer in a heterodyned receiver.

In later sections we discuss the Noise Equivalent Power (NEP) for incoherent detection schemes. For incoherent detection setups, which are sensitive to the intensity of the laser signal, the NEP is proportional to the square root of the dwell time, and detector NEPs are quoted as W/Hz$^{1/2}$. For coherent detection, such as the heterodyne system described earlier, the NEP is proportional to the dwell time and measured in units of W/Hz. Thus coherent detection offers better noise Figures for the same dwell time per point when compared with incoherent detection. Mixers can also be used to generate radiation; if instead of the transmit and LO as inputs we provided a transmit frequency and IF input, then the output of the diode mixer will include frequencies of transmit ± IF. If the IF is varied over a range that the mixer can respond, the output frequency of the mixer is the lasing frequency modulated over the varying bandwidth. This process is called “frequency chirping.”

54.1.6 Transformation of Multimode Laser Beams from THz Quantum Cascade Lasers

54.1.6.1 Quantum Cascade Lasers Quantum Cascade Lasers (QCLs) are semiconductor lasers based on intersubband transitions within the conduction band in quantum wells. This differs from semiconductor diode lasers where the lasing transition is interband (between conduction and valence bands in the semiconductor). The intersubband energy levels responsible for lasing transitions in QCLs are based on layer thicknesses in semiconductor materials and can be tailored, using appropriate growth and fabrication techniques, over a wide frequency range.

The schematic energy diagram of a QCL is shown in Figure 54.9. Barriers and wells (within the conduction band) are created by combining semiconductor materials with different bandgaps. Applying a bias voltage across the structure allows electrons to tunnel through the structure. The laser structure essentially consists of several periods; within each period there is an active region and an injector. The lasing transition occurs in the active region (level 3 to level 2 in Fig. 54.9), and then after relaxation processes (from level 2 to level 1 in Fig. 54.9), the electron tunnels across the barrier and is “injected” into the upper lasing levels of the next active region, where the process repeats. As the number of periods can be quite large, a single electron is capable of generating several photons as it cascades down the energy levels.

Faist et al. demonstrated the first QCL in 1994 with an emission frequency of 4.2 μm based on a superlattice structure [7]. Since then QCLs have been designed and fabricated to span the near- to mid-infrared spectrum quite effectively. FIR or terahertz QCLs are currently under development, and while some have been demonstrated, there are significant challenges to lowering the operating frequency.

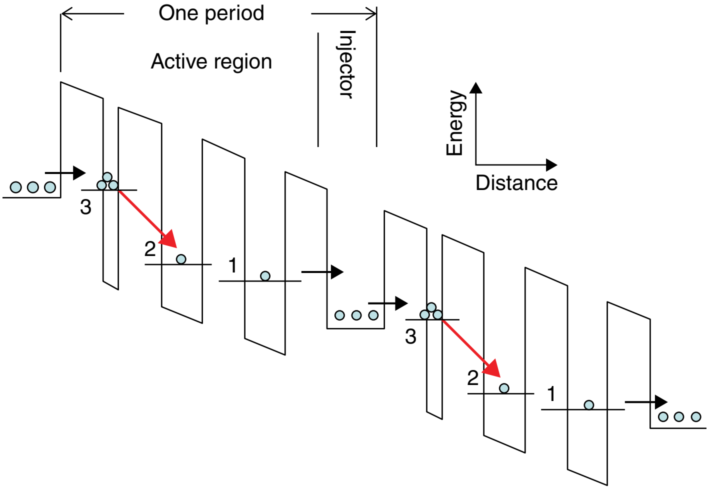


FIGURE 54.9 Conduction band energy diagram of two periods of a QCL. Source: Danylov [6]. Reproduced with permission from Andriy Danylov.

54.1.6.2 Transforming Multimode Laser Beams into Gaussian Beams As discussed in Section 54.1.2.3, laser oscillators are capable of supporting several discrete eigenmodes depending on the cavity's geometry. For several applications of laser beams (specifically imaging applications as discussed in later sections), the Gaussian output mode (TEM_{00}) is desired. Terahertz QCLs suffer from multimode output beams that diverge rapidly. A dielectric tube can be attached to the output face of the waveguide in order to improve the beam profile of the THz QCL. The method is equally relevant for other lasers with multimode, divergent output beams.

The basic idea is to attach a hollow dielectric tube to the laser output face (Fig. 54.10). Since the beam propagates in the hollow core, losses are minimal. The lossy dielectric material assists in cleaning up the mode profile as higher-order modes are more severely attenuated in the waveguide. The modes that propagate in these circular cylindrical structures are of three basic types: transverse circular magnetic (TM_{0m}), transverse circular electric (TE_{0m}), and nontransverse hybrid (EH_{nm}). If the dielectric waveguide has a radius "a," which is significantly larger than the free space wavelength of the laser beam (λ), then in making this approximation the EH_{1m} modes are transverse and linearly polarized. The far-field pattern of the EH_{11} mode essentially resembles a Gaussian profile. The attenuation constant for a given mode is proportional to λ^2/a^3; thus larger-diameter waveguides experience lower propagation loss.

For a THz QCL operating at 2.960 THz, University of Massachusetts (UMass) Lowell researchers selected a hollow pyrex tube of inner diameter 1.8 mm with a tube length of 43 mm [8]. Thus λ/a is negligible ($\lambda = 101.4\,\mu m$) and the EH_{1m} mode is transverse and linearly polarized. They found that the TE_{01} mode is significantly less lossy than the TM_{01} mode; however it exhibited slightly higher loss than the TE_{01} mode. However, as both the TE and TM modes are not linearly polarized like the QCL, the

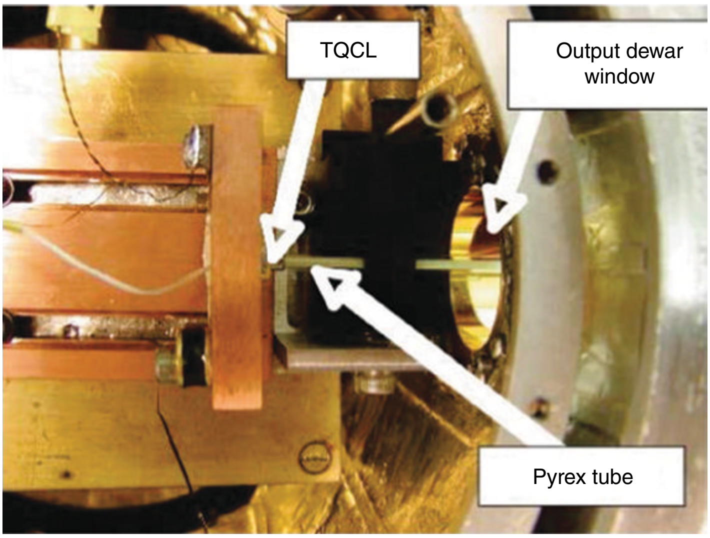


FIGURE 54.10 Photograph of terahertz QCL with dielectric tube attached at the output. Source: Danylov [8]. Reproduced with permission from the Optical Society.

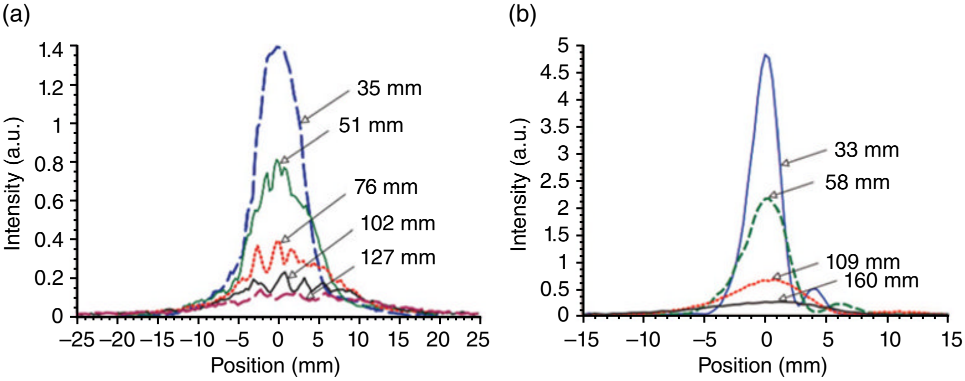


FIGURE 54.11 (a) THz QCL beam profile as a function of distance from laser output end. (b) Beam profile of the same laser as a function of distance after the dielectric tube was inserted. Notice the higher-order modes are essentially replaced by a Gaussian mode [8].

EH_{11} linearly polarized mode is dominant within the waveguide. This mode couples efficiently (98%) to the TEM00 (Gaussian) mode in free space, provided that the beam waist to tube radius ratio is 0.6435.

As can be seen in Figure 54.11, Danylov et al. measured the output beam profile of a THz QCL as a function of beam propagation distance both without and with a hollow dielectric waveguide (Fig. 54.11a and b, respectively). They were able to reduce the beam divergence and propagate a Gaussian mode in free space. The primary disadvantage to this technique is the loss in beam power due to the attenuation losses within the waveguide and coupling losses into the waveguide.

54.1.7 Suggested Reading

A very good standard reference textbook for understanding laser fundamentals is *Lasers* by Anthony E. Siegman, published by University Science Books. A standard textbook for optics is *Principles of Optics* by Born and Wolf, published by Cambridge University Press. A more basic optics textbook is *Optics* by Eugene Hecht, published by Pearson Education Limited. Another good reference textbook for laser optics is *Quantum Electronics* by Amnon Yariv, published by John Wiley & Sons.

54.2 LASER MEASUREMENTS: LASER-BASED INVERSE SYNTHETIC APERTURE RADAR SYSTEMS

Round-the-clock detection and location of man-made structures in all weather conditions have long been of interest. Active illumination *Ra*dio *D*etection *a*nd *R*anging (RADAR) systems have provided this capability through the coherent signal processing of propagating pulses transmitted and received using directional antennas. Since radio waves propagate at the speed of light c, the time delay (t_{delay}) between the transmitted and received pulses provides a measure of the round-trip distance to the object detected where the object's range r from the radar can be expressed as $r = c\ t_{\text{delay}}/2$. Using large aperture antennas to direct the frequency-chirped narrow beam pluses illuminating the scene, range (distance) and angular information may be acquired to form two-dimensional (2D) or three-dimensional (3D) imagery. The resolution of the imagery is inversely proportional to the bandwidth of the frequency chirp.

To produce azimuth resolution in the imagery, in 1951 Carl Wiley proposed coherent pulse-to-pulse processing of the backscattered signal captured from the scene while moving the radar. Since this technique artificially increased the antenna aperture, the process of forming these images was referred to as Synthetic Aperture Radar (SAR). Given the same imagery could be formed by pulse-to-pulse processing of backscattered signals from a rotating object with stationary radar, the popularity of collecting high-resolution radar imagery in controlled environments through Inverse Synthetic Aperture Radar (ISAR) also grew, and a large number of turntable radar facilities were established.

By the late 1970s, radar scientist Dr. Jerry Waldman [9], working at MIT Lincoln Laboratories in Bedford, Massachusetts, recognized the potential to acquire radar ISAR imagery in a laboratory setting by using very-high-frequency radar beams and scale models of objects of interest. Drawing from Maxwell's equations the use of millimeter-wave (wavelengths near, but longer than 1 mm) and submillimeter-wave radar beams (wavelengths shorter than 1 mm), the proportional size of the target and millimeter wave beam would reproduce a full-size ISAR turntable measurement.

Waldman et al. constructed an optically pumped submillimeter-wave laser and "compact" radar range and demonstrated the concept of calibrated signature similitude.

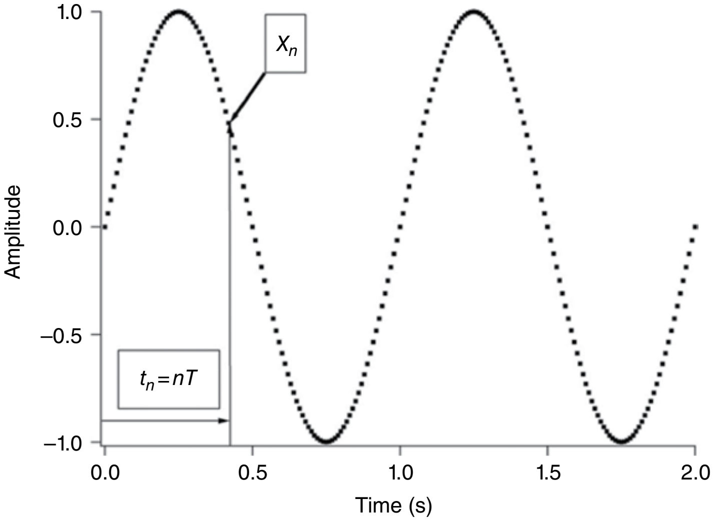


FIGURE 54.12 Example of discretely sampled data as a function of time.

The first imagery was published in the fall of 1979. Today, the UMass Lowell Submillimeter-Wave Technology Laboratory continues to produce high-quality and high-accuracy radar signatures of objects of interest over every radar band and has received numerous grants from DARPA, NASA, DoD, DHS, and the NSF.

54.2.1 ISAR Theory

54.2.1.1 Digital Fourier Transform The primary mathematical tool that is used to analyze radar data into imagery is the Fourier Transform (FT). Consider the discrete sampling of data in Figure 54.12. The data in this Figure can be represented by the array $\{x_n\}$ where $n = 0$ to $N - 1$. In general the discrete values of this array are composed of complex numbers and therefore

$$x_n = A_n e^{i\phi_n}$$

By using the Discrete Fourier Transform (DFT), the digitized time-varying signal $\{x_n\}$ can be represented as a summation of N uniformly spaced sinusoidal phasors such that

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{i\omega_k t_n}$$

where

$$t_n = nT$$
$$\omega_k \equiv \frac{2\pi k}{NT}$$

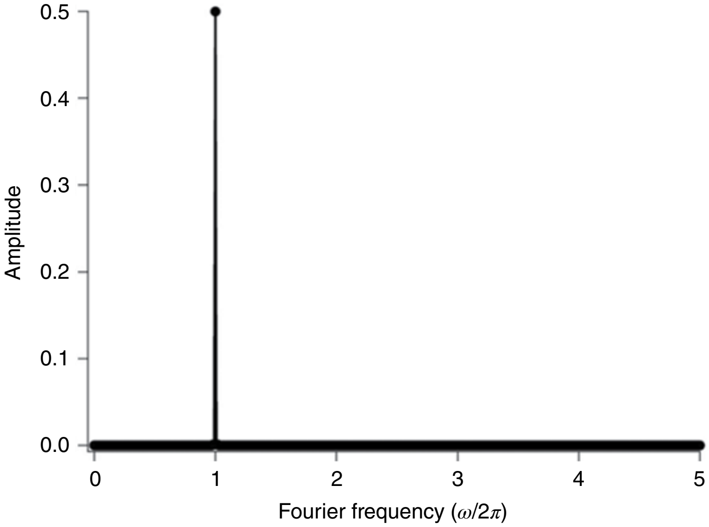


FIGURE 54.13 Discrete Fourier transform of a 1 Hz sine wave as a function of Fourier frequency.

In general the coefficients X_k are complex numbers as well such that

$$X_k = A_k e^{i\varphi_k}$$

The values of these Fourier coefficients are given by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i\omega_k t_n}$$

Figure 54.13 shows a sample calculation for the data given in Figure 54.12. In this case the data represents a 1 Hz sine wave, and when a DFT is applied to the data array, only the Fourier coefficient at $f = \omega/2\pi = 1$ Hz is nonzero. In this example the result of the DFT is to represent the spectral content of the original time-domain signal in the frequency-domain space.

54.2.2 DFT in Radar Imaging

The Fourier analysis described in the previous section can be optimized for use in radar analysis and image formation. Figure 54.14 shows the general geometry for a single isolated object illuminated by a radar beam. In general, objects are more complex. However, any complex object can be viewed as a linear sum of individual scattering centers, each of which can be considered individually. Therefore, the geometry of Figure 54.14 is still the fundamental approach to the formation of images.

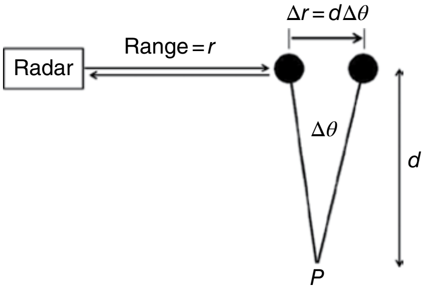


FIGURE 54.14 Range shift of a target rotated through an angle $\Delta\theta$.

In this Figure the object that is in the radar beam has a *total phase* due to its distance and the wavelength of the radar of

$$\varphi(r) = \frac{2\pi}{\lambda}[2r] = \frac{4\pi r}{\lambda}$$

Note the fact that there is an extra factor of 2 due to the round-trip path of the radar. If the target is located a distance *d* from the center of rotation and the target is rotated through an angle Δθ, this equation becomes

$$\Delta\varphi(r) = \frac{4\pi(r_2 - r_1)}{\lambda} = \frac{4\pi}{\lambda} d \cdot \Delta\boldsymbol{\theta}$$

In terms of a discrete sampling of points, the phase is given by

$$\Delta\varphi_n = \frac{4\pi(r_2 - r_1)}{\lambda} = \frac{4\pi}{\lambda} d \cdot \boldsymbol{\theta}_{\boldsymbol{n}}$$

where

$$\theta_n = n \cdot \Delta\boldsymbol{\theta}_0$$

Therefore the discretely sampled signal in the time domain can be represented by

$$x_n = A_n e^{i\left[\frac{4\pi}{\lambda} d \cdot \boldsymbol{\theta}_{\boldsymbol{n}}\right]}$$

The DFT coefficients can then be written as

$$X_k = A \sum_{n=0}^{N-1} e^{i\left\{\left[\frac{4\pi}{\lambda} d \cdot \boldsymbol{\theta}_{\boldsymbol{n}}\right] - \omega_k t_n\right\}}$$

It can be shown that the value of X_k rapidly approaches 0 except for the case where

$$\left[\frac{4\pi}{\lambda} d \cdot \boldsymbol{\theta}_{\boldsymbol{n}}\right] - \omega_k t_n = 0$$

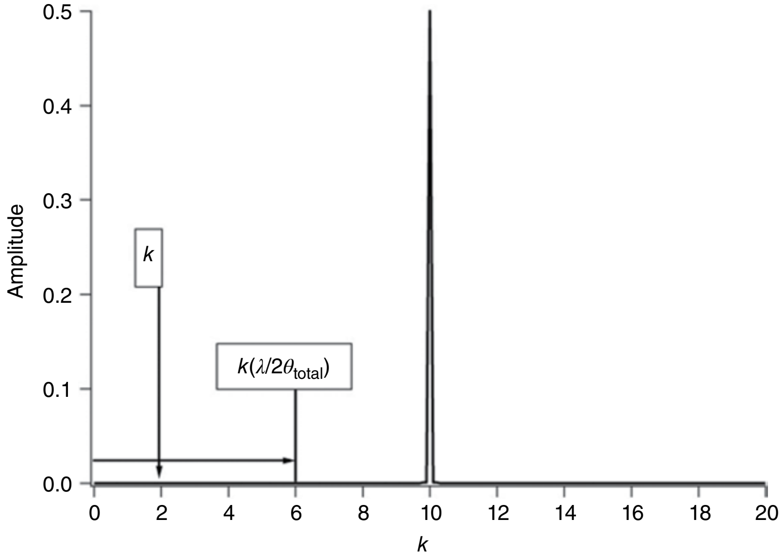


FIGURE 54.15 Fourier transform of an arbitrary signal from a target rotated through the radar beam. Each individual value of k represents another increment of the cross-range resolution.

Solving for d gives the result

$$d = k\left[\frac{\lambda}{2\theta_{\text{total}}}\right]$$

The physical interpretation of the value of k that is the integer vector from the DFT is that of equally spaced distances. An example is given in Figure 54.15. In this Figure the results of the DFT are displayed where the amplitude of the Fourier components are plotted versus the integer index k. The time-domain signal that has been transformed represents an object slowly turned through the radar beam in a manner to that shown in Figure 54.14. Complex amplitude and phase information is collected at uniform increments of θ. The resultant DFT yields a nonzero Fourier amplitude only when the constraint in the previous equation is satisfied. The quantity $[\lambda/2\theta_{\text{total}}]$ is known as the cross-range resolution. The example of Figure 54.15 shows that the target is located 10 resolution bins from the center of rotation.

The same analysis can be used for measuring the range seen in Figure 54.14 directly by calculating the phase change as a function of changing frequency. Following the same formalism given earlier it can be shown that for a series of complex measurements taken at uniformly spaced frequencies, the down-range position will be given by $r = k[c/2B]$. In this equation c is the speed of light and B is the total bandwidth that the radar frequency is changed by. The formalism given above can be extended to 2D and 3D measurements in order to form radar imagery.

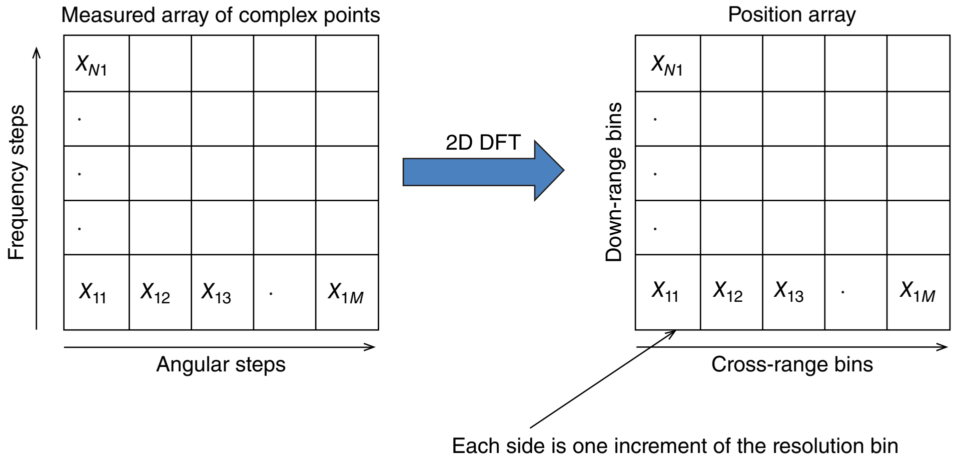


FIGURE 54.16 An example of the analysis of measured complex voltages as a function of frequency and angle transformed into positions by use of a Fourier transform.

Figure 54.16 shows a representation of a 2D data collection and DFT varying the angle and the frequency in a controlled manner. Once the DFT is performed, that index of the resulting array has the meaning of the down-range and cross-range resolutions that are given earlier. The frequency and angular information are thereby converted directly into a 2D array where the position within that array has the meaning of k times the resolution size.

54.2.3 Signal Processing Considerations: Sampling Theory

Theoretically in the course of using the FT, one is unfolding continuous periodic signals with infinite extent. Since practical measurements involve discrete sampling over limited bandwidths and displacements, ambiguities and spurious responses can form as a result of the sampling process. These spectral discontinuities at the end of the measurement interval cause spurious responses in the image. Multiplicative weighting functions (referred to as windowing, W) can be applied to smoothly taper the signal to zero at the ends of the measurement interval and reduce the effect of these discontinuities. The windowing can be applied directly to discrete sequenced signals within FFT in the form of

$$X_k = \sum_n x_n W_n e^{\frac{-i2\pi kn}{N}}$$

for n and $k=0$ to $N-1$.

By executing the FFT on the measured sequence of N backscattered signals, x_n, as a product with the windowing function, W_n, the computed signal values (i.e., resolution cells), X_k, will still represent the scattered signal at uniformly spaced locations along the range direction, but with reduced discontinuities.

54.2.4 Measurement Calibration

The electric field of an electromagnetic wave traveling in the z-direction is given by the equation

$$\vec{E} = E_{0x}\cos\left(\omega t - kz + \alpha_x\right)\hat{x} + E_{0y}\cos\left(\omega t - kz + \alpha_y\right)\hat{y}$$

where ω represents the angular frequency, t is time, k is the wavenumber, and α is an absolute phase constant. This equation can be rewritten in matrix form such that

$$\vec{E} = \begin{bmatrix} E_x \\ E_y \end{bmatrix}$$

where the components in the x and y directions are given by E_x and E_y. Radar systems normally represent the x-axis of the electric field as an electromagnetic wave whose electric field is in the horizontal plane. Similarly the y-axis is represented in the vertical plane. Thus we relabel the x- and y-axis as H and V to represent the polarization of the radar beam, H for horizontal and V for vertical. The received electric field for a radar system can be written in matrix form as

$$\begin{bmatrix} E_{\mathrm{H}}^{\mathrm{r}} \\ E_{\mathrm{V}}^{\mathrm{r}} \end{bmatrix} = \frac{e^{-ikR}}{2\sqrt{\pi}R} \begin{bmatrix} \mathrm{S}_{\mathrm{HH}} & \mathrm{S}_{\mathrm{HV}} \\ \mathrm{S}_{\mathrm{VH}} & \mathrm{S}_{\mathrm{VV}} \end{bmatrix} \begin{bmatrix} E_{\mathrm{H}}^{\mathrm{t}} \\ E_{\mathrm{V}}^{\mathrm{t}} \end{bmatrix}$$

where the transmitted signal E^{t} is multiplied by a scattering matrix S.

Since radar systems are composed of transmitter and receiver optics and electronics, it is necessary to calibrate the response of the system both in polarization and in frequency. The method for this calibration is described in detail by Chen et al. [10] and is briefly given here and in DeMartinis et al. [11]. Consider the measurement matrix S^{m} given in the equation below. In this equation the measured values $\mathrm{S}_{\mathrm{HH}}^{\mathrm{m}}$, $\mathrm{S}_{\mathrm{HV}}^{\mathrm{m}}$, $\mathrm{S}_{\mathrm{VH}}^{\mathrm{m}}$, and $\mathrm{S}_{\mathrm{VV}}^{\mathrm{m}}$ are complex numbers and represent the measured amplitude and phase of the back-reflected signal of a target illuminated by a radar beam and received through the radar's optics and electronics. The target itself has reflected the radiation according to the values given in the S matrix such that the ideal radar return is represented by S_{HH}, S_{HV}, S_{VH}, and S_{VV}. The values of S undergo distortions that are represented by the R, T, and I matrices. These matrices are the receiver, transmitter, and isolation distortion matrices, respectively:

$$\begin{bmatrix} \mathrm{S}_{\mathrm{HH}}^{\mathrm{m}} & \mathrm{S}_{\mathrm{HV}}^{\mathrm{m}} \\ \mathrm{S}_{\mathrm{VH}}^{\mathrm{m}} & \mathrm{S}_{\mathrm{VV}}^{\mathrm{m}} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_{\mathrm{HH}} & \mathrm{I}_{\mathrm{HV}} \\ \mathrm{I}_{\mathrm{VH}} & \mathrm{I}_{\mathrm{VV}} \end{bmatrix} + \begin{bmatrix} \mathrm{R}_{\mathrm{HH}} & \mathrm{R}_{\mathrm{HV}} \\ \mathrm{R}_{\mathrm{VH}} & \mathrm{R}_{\mathrm{VV}} \end{bmatrix} \begin{bmatrix} \mathrm{S}_{\mathrm{HH}} & \mathrm{S}_{\mathrm{HV}} \\ \mathrm{S}_{\mathrm{VH}} & \mathrm{S}_{\mathrm{VV}} \end{bmatrix} \begin{bmatrix} \mathrm{T}_{\mathrm{HH}} & \mathrm{T}_{\mathrm{HV}} \\ \mathrm{T}_{\mathrm{VH}} & \mathrm{T}_{\mathrm{VV}} \end{bmatrix}$$

The equation given above can be solved for the S matrix by inverting the transmit and receive distortion matrices:

$$[\mathrm{S}] = [\mathrm{R}]^{-1}\left(\left[\mathrm{S}^{\mathrm{m}} - [\mathrm{I}]\right]\right)[\mathrm{T}]^{-1}$$

Chen et al. [10] have shown that the R, T, and I matrices can be calculated by scanning a series of known objects with the radar system and solving a series of equations. The known objects are a flat plate, a dihedron with the seam in the horizontal plane, and a dihedron with the seam oriented at an angle q to the horizontal plane. The scattering matrices of the ideal objects are given by

$$S_{\text{plate}} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$S_{\text{dihedron, }0^\circ} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$S_{\text{dihedron, }\theta^\circ} = \begin{bmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix}$$

This formalism is used to accurately calculate the undistorted scattering matrix, S, from the measured scattering matrix, S^m.

54.2.5 Example Terahertz Compact Radar Range

Researchers at the UMass Lowell have acquired 2D and 3D ISAR imagery using two CO_2 lasers to optically pumped FIR lasers (Goyette et al. [12, 13]). Configured with a difference frequency of 1.9 GHz between the two 1.56 THz lasers, one laser served as the transmitter, while the second was used as an LO for the heterodyne receiver. To establish a tunable frequency, the laser transmitter was mixed with a microwave sweeper (10–18 GHz) using a Schottky diode. Producing two sideband frequencies that can be swept, the receiver's electronics were designed to only detect the lower sideband. The swept frequency system had sufficient bandwidth to achieve a range resolution of 0.625″ (Fig. 54.17).

Using single-frequency terahertz laser data over a 5 × 5 solid angle measured in azimuth (side to side) and elevation (up and down), UMass Lowell scientists achieved a resolution on the order of a millimeter for 2D cross-range-only imagery. Figure 54.18 shows the complex amplitude and phase data taken across the angular extent in azimuth and elevation. This data has been analyzed in a manner similar to what has been described in the previous sections whereby the data, spaced incrementally in the two angular directions, is Fourier transformed in order to calculate the image. Figure 54.19 shows the result of this DFT. The image of the 1/16th scale model truck is easily recognized once the DFT is plotted in a manner similar to that shown in Figure 54.16. In this case the two axes are variations in the viewing angle producing an image in azimuth cross-range and elevation cross-range.

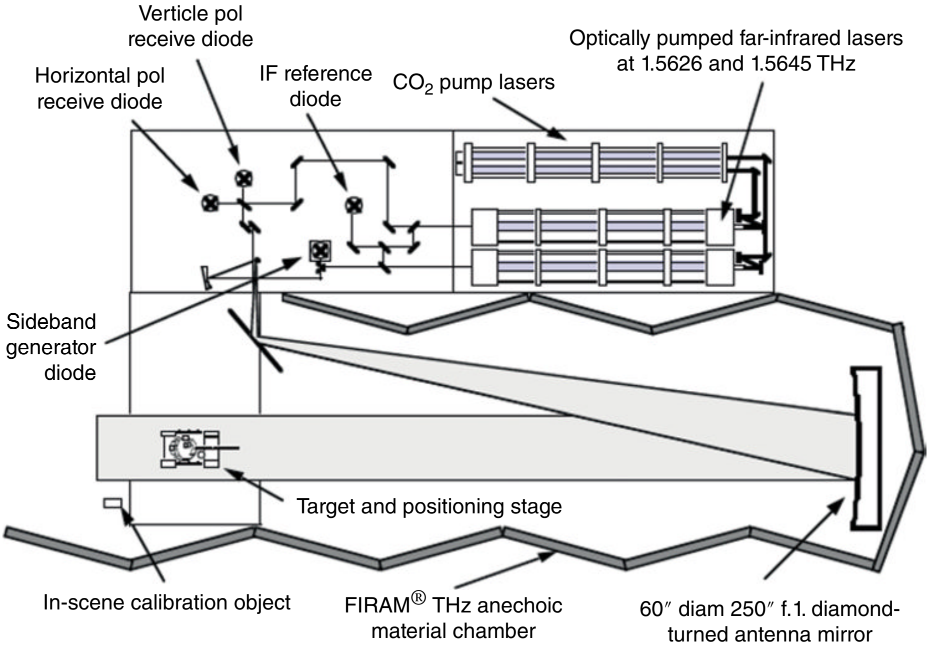


FIGURE 54.17 A 1.56 THz laser-based compact radar range. Source: Goyette et al. [12]. Reproduced with permission from SPIE.

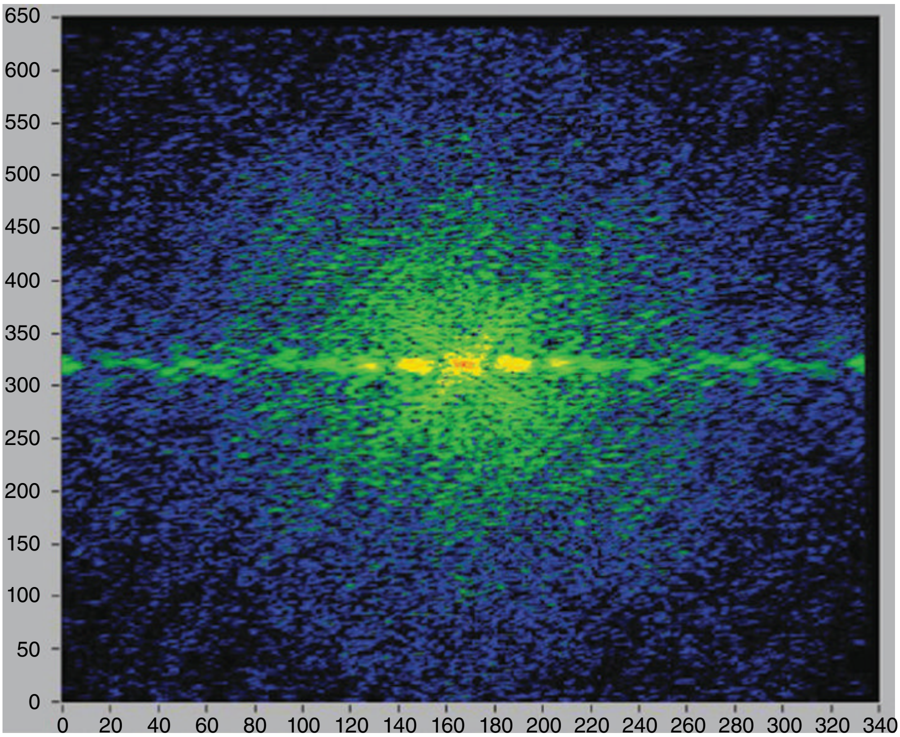


FIGURE 54.18 Amplitude plot of the complex radar return from a truck measured through a 5-degree × 5-degree solid viewing angle at 1.56 THz.

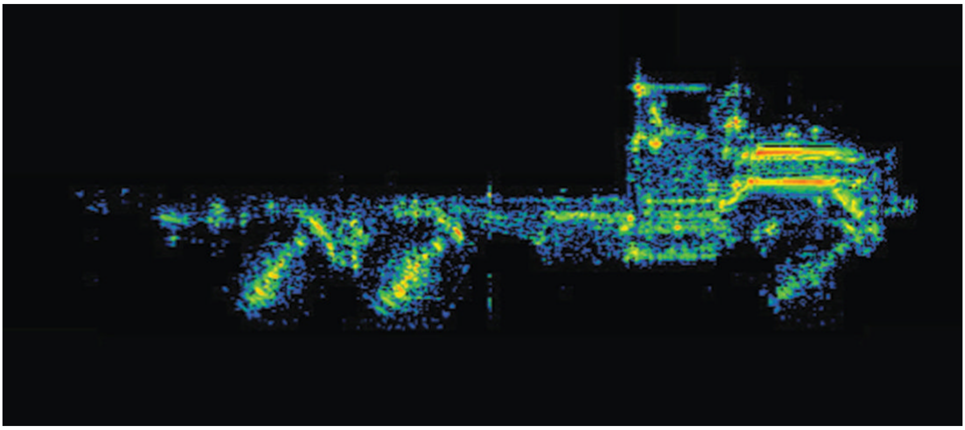


FIGURE 54.19 DFT of the complex radar return from a truck measured through a 5-degree × 5-degree solid viewing angle, forming a side view image of the target at 1.56 THz.

54.2.6 Suggested Reading

Radar Cross Section by Knott, Shaeffer, and Tuley published by Artech House, Inc. provides a good introduction to radar measurement systems, while *High Resolution Radar Cross-Section Imaging* by Dean L. Mensa also published by Artech House Radar Library also provides an excellent reference for advanced users.

54.3 LASER IMAGING TECHNIQUES

The time required to evolve scientific discoveries into practical products has always been the challenge for any technical community. While teams of researchers with great insight and persistence establish the foundation of breakthrough science, significant engineering is also necessary for maturing technologies with these discoveries to finally create products for new applications.

The time cycle from discovery to product appears to be diminishing as global communication rapidly moves ideas from research to industry internationally. But development of source technologies, like the tunable CO_2 optically pumped FIR lasers, stretched across half a century as many sought to turn this laboratory instrument into a precision product for terahertz (THz) frequency spectroscopic imaging applications. Now with evidence of applications in the fields of forensics, security, medicine, and communications, a number of new THz source/detection technologies offer the promise of commercialization.

From the gamma to microwave frequency regime, lasers in general have always provided a significant platform for materials science. Whether using broadband or narrowband tunable sources, lasers are capable of facilitating precise amplitude, phase-stable, polarization-sensitive measurements for characterizing the properties of a never-ending list of novel materials.

A laser is essentially a monochromatic coherent light source. The interaction of any material with a laser source is determined by the refractive index of that material, which is a function of frequency. Laser imaging typically examines light, which is reflected from, or transmitted through, a sample illuminated by a laser source. Correlating the collected signal with sample geometry yields the image. In general, when laser light is incident on a material, several processes occur: reflection at the interface (specular and diffuse), transmission through the interface, absorption and scattering within the sample, and in some cases excitation and emission of light (fluorescence). All these processes are determined by the interaction of light with materials and can be used to generate images of the samples that highlight features that are of interest.

In previous sections ISAR imaging was introduced, wherein the entire sample is illuminated by a laser beam and the image is reconstructed by changing the orientation of the sample with respect to the beam. In this section two more conventional approaches to laser-based imaging are presented: point scanning, wherein a laser beam is focused onto a moving sample, and camera imaging, wherein the beam illuminates the entire sample and the response is collected by a camera. When designing laser imaging systems, several factors are considered; these include imaging resolution, system dynamic range, and data acquisition rates. Specific examples of these parameters are discussed.

54.3.1 Imaging System Measurement Parameters

54.3.1.1 Detectors and NEP Apart from a laser source, laser imaging systems require a detector to measure the signal remitted by the sample. In Section 54.1.5, heterodyned detection was discussed. This was an example of a coherent detection scheme; other detection schemes used in imaging applications employ incoherent detection wherein the detector output is proportional to the intensity of the laser beam incident on it.

For example, a liquid helium-cooled bolometer can be used as an incoherent detector for terahertz frequency applications; it is sensitive to the intensity of the terahertz beam. The bolometer operates on the following principle: the silicon semiconductor is placed in a cold bath, which is in contact with the liquid helium ($T \approx 4.2\,\mathrm{K}$). The crystal is also connected to an absorber that heats up when terahertz radiation is incident. The amount of heat generated in the absorber is proportional to the intensity of the terahertz beam. As the absorber is in contact with the silicon, the silicon also absorbs heat; this creates electron–hole pairs in the silicon and reduces the resistivity. Thus monitoring the resistivity of the silicon measures the intensity of the terahertz beam. Modulating the incident intensity at a specific frequency allows for tracking changes in beam intensity. This is accomplished by optically chopping the terahertz beam.

The highest modulation/chop frequency is determined by the response time of the detector. The response time of the detector is a measure of how fast a detector can respond to incident signal. It is generally specified by quoting the time constant of the detector (τ). The time constant is the amount of time taken after the detector is exposed to the signal for the detector's output to rise to $(1-1/e)$ of its final output value. The modulation frequency can be related to the detector time constant as $f_{\mathrm{M}} = 1/2\pi\tau'$ where f_{M} is the frequency corresponding to half the peak detector output [14]. For accurate measurements, modulation frequencies are lower than f_{M}.

For a silicon bolometer, the response time is also dependent on the thermal contact between the silicon and the cold bath; higher thermal conductivity implies that the detector cools down faster. Generally the trade-off is between response time and sensitivity; collecting photons for longer time periods increases the response, while cooling faster decreases the response and provides a faster response time. At terahertz frequencies, the frequency response curve of a bolometer is fairly flat; thus the detector can be used over a range of frequencies.

Also a critical characteristic for an imaging system is the Signal-to-Noise Ratio (SNR). This depends on two factors: the maximum power received at the detector and the NEP of the detector. The NEP of the detector is the signal power required for the SNR to be 1 over a 1 Hz bandwidth [14]. Thus, the smaller the NEP, the better the detector. As discussed previously, the NEP of a coherent detection scheme is measured in W/Hz, while the NEP of an incoherent detector is measured in W/Hz$^{1/2}$. Thus using the same detector in coherent as opposed to incoherent detection schemes yields higher SNRs. Another system characteristic often confused with SNR is dynamic range. The dynamic range of the system is the difference between the noise floor of the system and the saturation point of the detector.

54.3.1.2 Beam Waist Measurements There are several techniques to measure the beam waist at a point. Discussed here are two common techniques: the knife-edge method and profiling the beam expansion.

For the knife-edge method, a straight edge is moved across the beam perpendicular to the propagation axis, and the transmitted signal is collected by a detector as shown in Figure 54.20. If the beam intensity profile is Gaussian, then the intensity incident on the detector ($I(x)$) as the knife-edge is scanned across the beam is given by

$$I(x) = \frac{I_0}{2}\left[1 + \operatorname{erf}\left(2\sqrt{\ln(2)}\left(\frac{x - x_{\mathrm{c}}}{d}\right)\right)\right]$$

where I_0 is the total intensity, x_{c} is the position of the intensity peak, and d is the Full Width at Half Maximum (FWHM) of the Gaussian beam [14]. The Gaussian beam waist (w) can be calculated from the FWHM using the following relation:

$$w = 0.8493218 * \mathrm{FWHM}$$

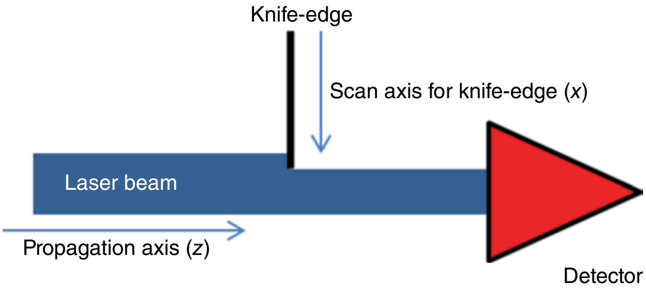


FIGURE 54.20 Knife-edge scan technique for laser beam propagating along the *z*-axis; the transmitted fraction of the beam is collected as a function of the knife-edge position perpendicular to the beam (*x*).

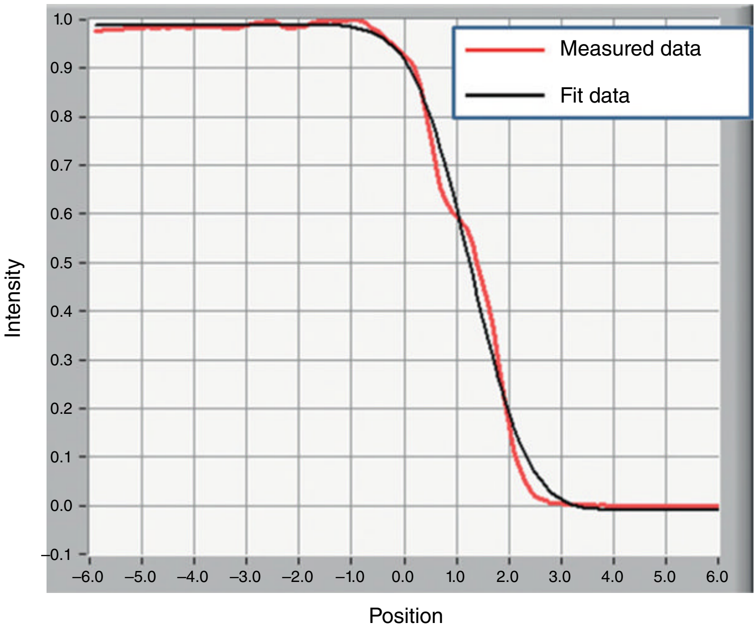


FIGURE 54.21 Knife-edge scan and best fit curve of a 1.4 mm FWHM 2.5 THz beam. The black line is the best fit curve.

Figure 54.21 shows the knife-edge scan of a 2.5 THz laser beam and the best fit curve. Using this one can extract the FWHM and beam waist at that point along the propagation axis. Other commonly used expressions for beam size specify the 10/90 and 20/80 points of the knife-edge scan. If x_{10} and x_{90} are the positions of 10% intensity measured and 90% intensity measured by a knife-edge measurement [15, 16], then, for a Gaussian beam profile, the waist can be calculated as

$$w \approx 0.7803\left(\left|x_{10} - x_{90}\right|\right)$$

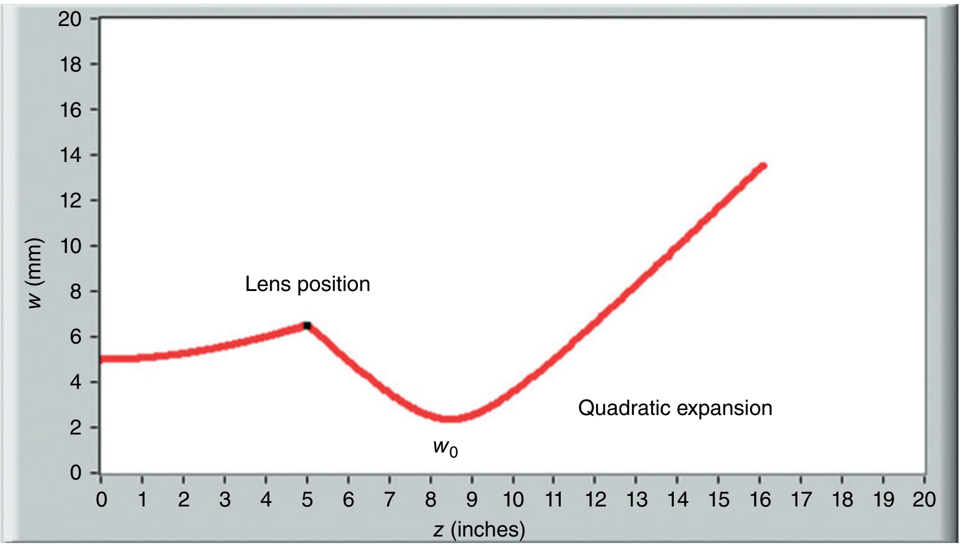


FIGURE 54.22 The quadratic expansion of a 513 μm laser beam after it is brought to focus (w_0) by a lens.

Another technique used to determine the beam waist and position of a Gaussian beam is to profile the beam expansion with an aperture. As the beam expands along the propagation axis, the waist as a function of axial distance (z) is given by

$$w^2(z) = w_0^2\left[1+\left(\frac{\lambda z}{\pi w_0^2}\right)^2\right]$$

where w_0 is the beam waist at the focus (i.e., the radius of curvature at that point is infinity).

As shown in Figure 54.22, after the beam is brought to a focus, it expands quadratically in free space. Thus, using an aperture to measure the beam and fitting a Gaussian beam profile at various points along the z-axis allow one to map the quadratic expansion. Fitting the measured beam profile allows one to determine w_0 and its axial location, which is the focal plane.

54.3.1.3 Data Acquisition Data acquisition for imaging systems has two primary aspects: adequate sampling of the imaging target (scanning resolution) and the acquisition speed.

Scanning resolution essentially translates to sampling of the target with the laser beam. According to the Nyquist sampling limit, the signal must be sampled at a rate twice that of its frequency. What this implies for spatial imaging is that to generate a well-resolved image, at least two data points must be collected within the imaging beam's FWHM. This means if the beam FWHM is 1 mm, the scanning resolution should be at most 0.5 mm in order to generate a well-resolved image. Collecting more information (higher scanning resolution) does not yield significantly more information.

Another aspect of imaging is the acquisition speed. This is determined by two primary interrelated factors: scanning speed and the detector's acquisition speed. The scanning speed determines how long it takes to move the sample across the beam or alternatively move the beam across the sample. While this is sometimes mechanically limited, as in the case of optomechanical scanning systems that direct the beam across the target area, it is ultimately limited by how fast the detector is capable of responding and the requisite signal averaging required.

When considering the speed of the detector's response and its relation to determining system acquisition speed, it is useful to consider a specific example. Let us consider the liquid helium-cooled silicon bolometer. As discussed previously, the detector is modulated at a frequency that is related to its response time, and any higher speed modulation will not yield the requisite response. Another factor that needs to be considered is the "dwell time" per data point. The detector receives the desired signal and the background noise; since background is random, it can be eliminated by time-averaging the received signal, for example, with a lock-in amplifier averaging the signal at the detector modulation frequency. Increasing the averaging time will lower the background noise. As a rule of thumb, the dwell time for a data point collected using a lock-in amplifier should be at least 3 times the time constant of the lock-in amplifier. Thus, given a certain scanning range, scanning resolution, and dwell time, the image acquisition time can be computed.

54.3.1.4 Optical Elements Laser measurement systems consist of several optical elements that are designed to alter characteristics of the laser beam. Lenses and curved mirrors are used to reshape the laser beam. Dichroic mirrors and gratings are used to separate different frequency components. In this section we discuss two commonly used optical elements: beam splitters and wire-grid polarizers.

A commonly used beam splitter is a thin film. This works based of the Fresnel reflection and transmission coefficients for a thin film. The fraction of light that is reflected, transmitted, and absorbed by the thin film depends on the frequency-dependent refractive index of the material, the film thickness, wavelength, polarization, and the angle of incidence of the incident beam. If the incoming beam is polarized, the physical orientation of the polarizer is determined by the requisite *s*- or *p*-Fresnel coefficient.

Consider a thin Mylar film. Mylar has a complex refractive index of $1.73 + i0.030$ at 513 μm. Figure 54.23a and b shows the *s*- and *p*-Fresnel reflectance for this Mylar film (76.2 μm thick) as a function of incidence angle. For *s*-polarized light, the reflectance is 42% and the transmittance is 53%, while for *p*-polarized light the reflectance is 7.6% and transmittance is 87%. Thus, for the beam splitter to approximate a 50–50 splitter, it is critical that the film be oriented such that the beam polarization is perpendicular to the plane of incidence (*s*-polarized). It is important to note that the reflectance and

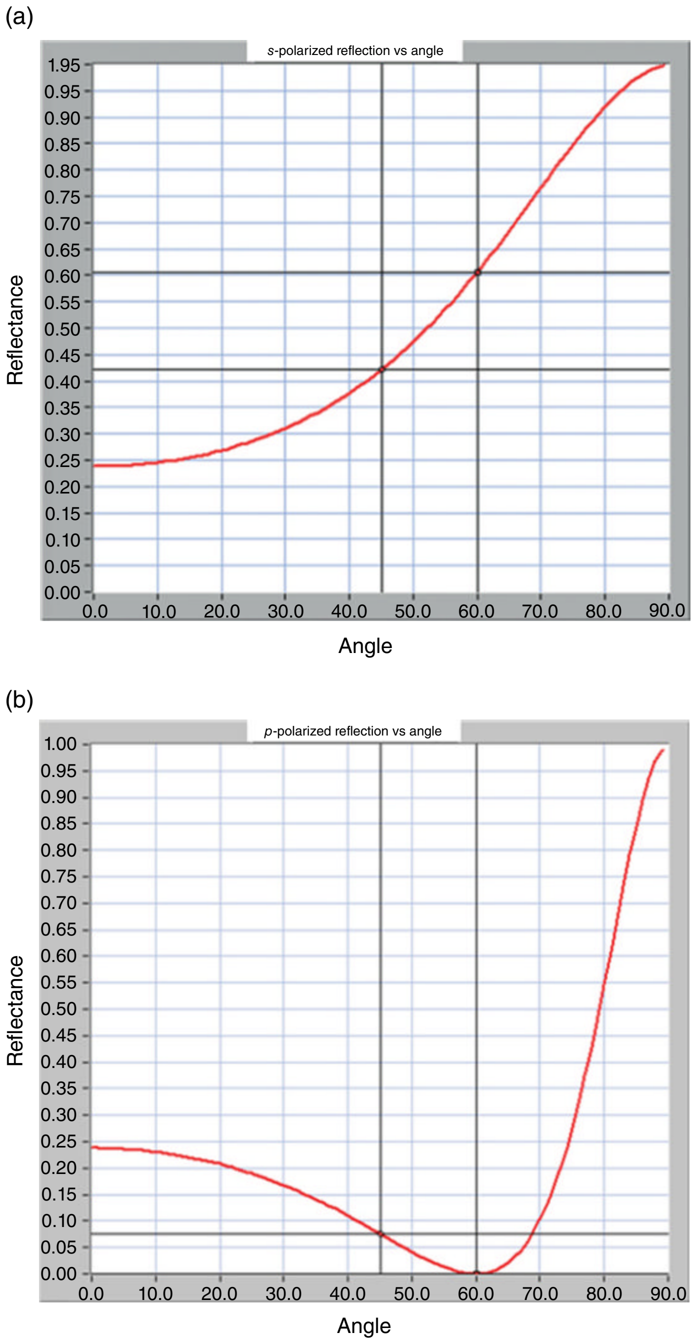


FIGURE 54.23 (a) The *s*-polarized Fresnel reflectance as function of incident angle and (b) the *p*-polarized Fresnel reflectance.

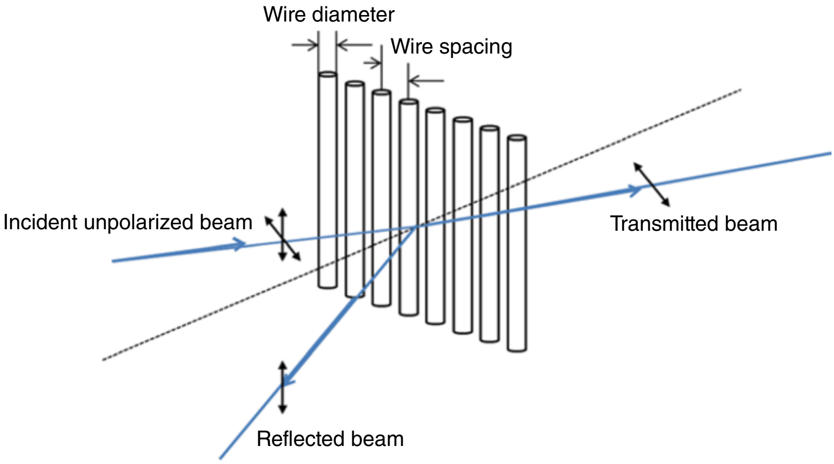


FIGURE 54.24 Reflection and transmission from a wire-grid polarizer.

transmittance do not add up to one, with the remaining signal being absorbed in the film. The signal loss due to absorption is determined by the complex part of the refractive index and film thickness. Beam splitters are typically used when normal-incidence reflectance measurements are desired. Even in the case of a lossless 50–50 beam splitter, a 100% reflective sample, and no signal loss in system optics, at most 25% of the generated laser intensity can reach the detector.

A polarizer is an optical device that selectively transmits or reflects radiation based on the polarization state of the incident light. A metallic wire-grid polarizer, such as the one shown in Figure 54.24, separates different components of linearly polarized light. The wire grid is essentially transparent to radiation that is polarized perpendicular to the orientation of the wires and strongly reflects polarization that is parallel to the wire orientation. This happens because the electric field polarized parallel to the wire grid establishes a current in the wires, which then radiates a reflected beam; however, if the electric field is perpendicular to the wires, no current is set up and the beam is transmitted. The rejection ratio is determined by the wire material, the wire diameter, the wire spacing, and the wavelength of light used. A laser beam is always polarized. The polarization can be linear or circular and quarter-wave plates can be used to convert linearly polarized light to circular polarized and vice versa.

54.3.2 Terahertz Polarized Reflection Imaging of Nonmelanoma Skin Cancers

One of the primary uses for laser imaging is in medical diagnostics. For example, optical frequency systems are used in Optical Coherence Tomography (OCT) and confocal microscopy, among others, for a variety of biomedical imaging applications [17–23]. Terahertz systems are also being developed as possible diagnostic imaging modalities [24–28].

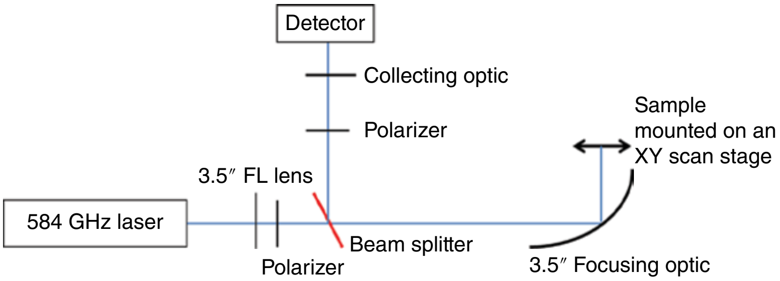


FIGURE 54.25 Schematic of a point scan terahertz reflectance system [29].

In this section we consider Continuous-Wave (CW) terahertz reflectance imaging of skin cancer. Terahertz imaging offers a way to image intrinsic contrast between healthy and diseased tissue and is nonionizing. The presented technique is an example of point scanning and uses the polarized nature of the terahertz laser beam to generate images of intrinsic contrast between normal skin and nonmelanoma skin cancer. This research was performed by Joseph et al. at UMass Lowell [29].

This investigation used a CO_2 optically pumped FIR gas laser as the source. The 584 GHz (513 μm) vertically polarized transition in HCOOH was pumped by the 9R28 transition of the CO_2 laser. A hollow glass waveguide was used to achieve a Gaussian mode, and the measured output power was 10.23 mW. An IR Labs liquid helium-cooled silicon bolometer is used as the imaging detector. The NEP was 1.13E–13 W/Hz$^{1/2}$ and the responsivity was 2.75E+05 V/W. The bolometer had a response time of 5 ms and the gain was 200. A crystalline quartz garnet powdered window on the bolometer cutoff wavelengths below 100 μm.

Figure 54.25 depicts the optical configuration for the point scan imaging system. The imaging system can be summarized as follows: the laser source emits a vertically polarized beam at 513 μm. This beam is collimated by a TPX lens and then hits a fast-focusing Off-Axis Parabolic (OAP) mirror that focuses the beam down to a spot of waist size 570 μm. A sample is placed at the focal spot such that the laser beam is incident normal to the sample surface and the sample is then raster scanned across the focal spot. The acquired signal is correlated to the sample position to generate a 2D image of the sample. The reflected beam retraces the path of the incident beam. A Mylar beam splitter deflects a fraction of the reflected signal into the detection arm. Using a wire-grid polarizer in the detection arm allows one to image different polarization states of the reflected signal and generates images that are co- and cross-polarized relative to the incident beam.

The beam waist at the focal plane determines the imaging resolution. For the 584 GHz optical configuration described in this section, the beam waist was measured to be 0.57 mm at the focal plane. Standard reflection imaging systems can be set up to either move the sample across the focal spot (as was done in this case) or to scan the beam across a stationary target. During the imaging procedure several factors need to be taken into account, including scanning resolution and speed. For the experiment described, the laser beam was optically chopped, and the chopping frequency served as the reference frequency for a lock-in amplifier. The data collected by the bolometer was then sent to a lock-in amplifier that had a time constant of 30 ms. The dwell time per point in the image was around 150 ms. For the laser-based reflection imaging experiment described here, the system SNR was measured to be 65 dB.

54.3.2.1 Sample Processing and Mounting The samples imaged were fresh thick excess skin cancer specimens obtained after Mohs micrographic surgeries performed at Massachusetts General Hospital. Prior to imaging, 5 μm thick horizontal sections were cut from the sample for histopathology. These sections were stained with Hematoxylin and Eosin (H&E) and were used to evaluate the results. For imaging, the specimens were covered with a 1 mm thick z-cut quartz window. The z-cut quartz window was selected because it is relatively low loss and its refractive index 2.117 closely matches the refractive index of the human skin (≈2.2 at 600 GHz). Index matching allows for efficient transfer of beam power across the quartz–sample interface. To prevent dehydration during the imaging experiment, the samples were placed on a gauze soaked in pH balanced (pH 7.4) saline solution. A total of nine samples, six Basal Cell Carcinomas (BCC) and three Squamous Cell Carcinomas (SCC), were measured during this experiment.

54.3.2.2 Image Processing and Analysis Copolarized and cross-polarized images were acquired by selecting the appropriate orientation with the analyzing polarizer in the reflectance arm of the system. In order to calibrate the percent reflectance of the images, they were calibrated against the full-scale return of a flat front-surface gold mirror. Figure 54.26 shows examples of the co- and cross-polarized images (in logarithmic scale) along with the H&E-stained histopathology of the sample.

In Figure 54.26c, the tumor is outlined with the black dotted line. Easily observed, the cross-polarized image (Fig. 54.26b) correlates low terahertz cross-polarized reflectance with the tumor area, while the copolarized image (Fig. 54.26a) does not correlate well. This same observation was seen in all nine samples. This investigation demonstrates that cross-polarized terahertz reflectance imaging offers intrinsic contrast between normal and cancerous tissue. However the origin of the contrast is unclear. A possible explanation is scattering within the tissue volume. While contrast in the copolarized image is possibly obscured by Fresnel reflections from interfaces, the cross-polarized reflectance requires repolarization, which in turn could be caused by multiple scattering events within the tissue volume. At terahertz wavelengths, the primary mechanism is absorption as water is highly absorbing at these frequencies. Scattering requires structures on the order of the wavelength, which is fairly long for terahertz radiation when compared with most cellular structures; thus scattering is generally neglected when compared to absorption.

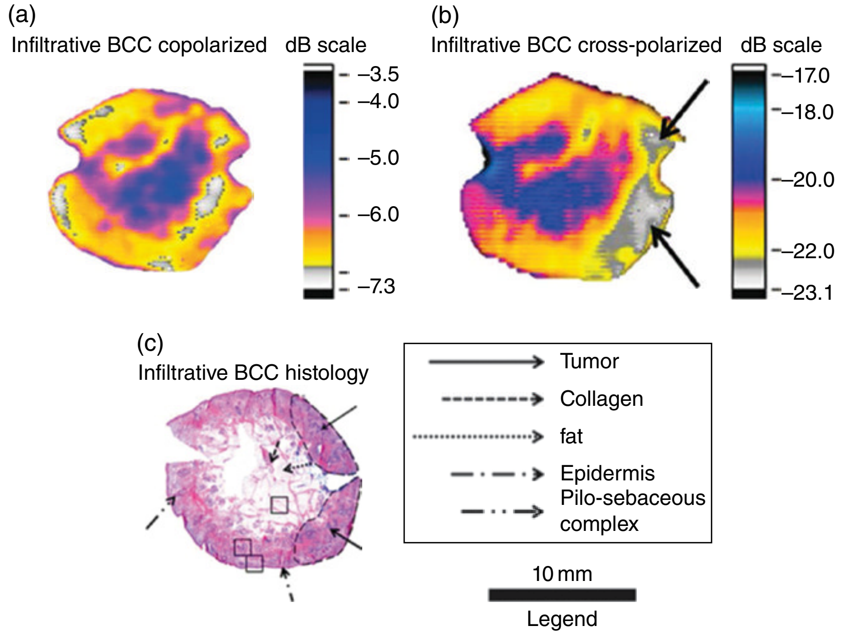


FIGURE 54.26 (a) The copolarized terahertz reflectance, (b) the cross-polarized terahertz reflectance, and (c) the H&E-stained histopathology of the corresponding 5 μm thick section of a sample of infiltrative Basal Cell Carcinoma (BCC).

The image contrast indicates that while scattering of terahertz radiation is low, it may not be negligible. The measured cross-polarized reflectance is very low (<1%) as opposed to copolarized reflectance. However, it indicates that the terahertz radiation possibly undergoes scattering events within the tissue volume. Moreover, in CW-THz imaging, copolarized reflectance suffers from Fresnel artifacts at the air–window interface and at the surface of the tissue, thus obscuring contrast. Using the fact that repolarization of the backscattered beam requires several scattering events in the tissue implies that imaging cross-polarized terahertz imaging penetrates deeper into tissue and the signal collected is representative of the tissue volume.

The postulate that scattering possibly contributes to terahertz contrast is consistent with work done on terahertz dark-field imaging of dehydrated formalin-fixed tumor tissue that also showed intrinsic contrast [30]. Furthermore, work on terahertz polarization-sensitive reflection imaging of colon cancers shows contrast between normal and cancerous colon tissue as well; however in the case of colon cancer the remittance from the cancerous tissue is higher than that of normal colon [31]. The possible cause of this is that cancerous colon contains structures that are of the order of the imaging wavelength (≈500 μm), while normal colon is very homogeneous. Further work is required to study the refractive index variation within the tissue volume at terahertz frequencies in order to determine whether scattering is an additional contrast mechanism as opposed to just water content differences.

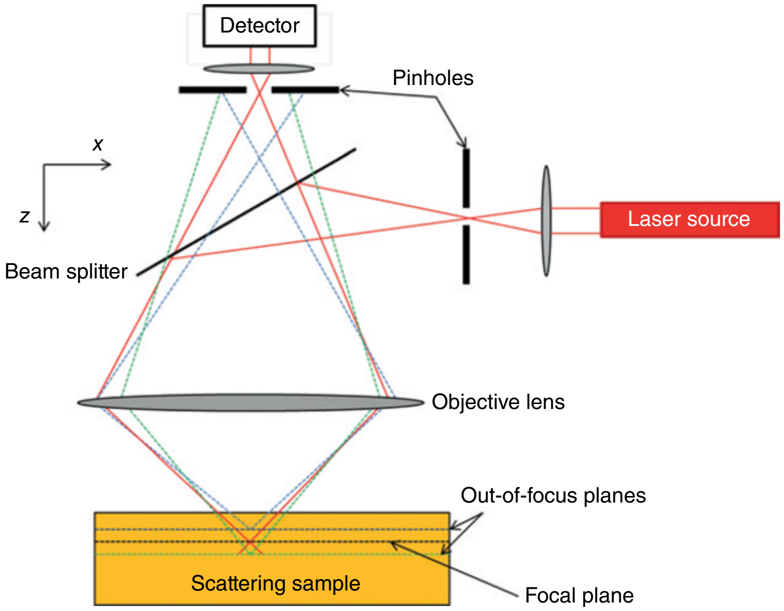


FIGURE 54.27 Schematic diagram of the confocal principle.

54.3.3 Confocal Imaging

Confocal microscopes are based on the confocal principle invented by Marvin Minsky [32]; they are able to use a pinhole to reject light that is scattered from outside the focal plane and produce high-resolution images. Confocal images are generally formed by optomechanically scanning the laser spot across the target area. The confocal principle is used to eliminate scattering from unwanted regions of the sample and thereby improve image contrast. Figure 54.27 presents the schematic of the confocal principle.

As shown in Figure 54.27, when light propagates through scattering media, signal is remitted back to the detector from the entire sample volume. This remission of light from planes that are not in focus leads to blurring and a loss of image quality and contrast. The confocal principle uses pinholes that are placed in conjugate planes to reject the unwanted reflections. As seen in Figure 54.30, the solid line traces the signal from the desired focal plane, while the dotted lines trace the scattered remission from out-of-focus planes. The out-of-focus light is rejected by the detector pinhole, the position of which is in the conjugate plane. Confocal laser scanning microscopes offer very high resolution of the order of the emission wavelength of the laser. The lateral and axial resolutions are determined by the numerical aperture of the objective lens and the emission wavelength of the laser source [21]. The depth discrimination, which is the ability to reject out-of-focus planes, is ultimately a function of the pinhole diameter; smaller pinholes better confine the focal plane. For a confocal setup, it is possible to

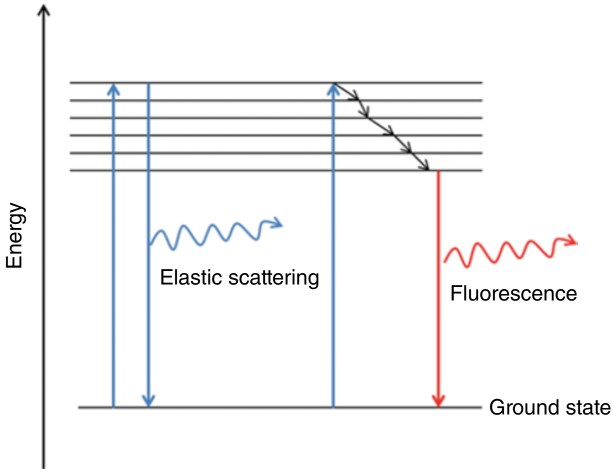


FIGURE 54.28 Absorption and possible scattering mechanisms.

image a scattering sample at varying depths by scanning the sample along the optical axis (z-axis in Fig. 54.27). This generates a sequence of imaging planes across the depth of the sample and can be used to generate a 3D map of the target. This technique is sometimes referred to as "optical sectioning."

Confocal microscopes can be used for fluorescence and reflectance imaging of biomedical samples [33–35]. Confocal Raman microscopes use the confocal principle to generate images of the Raman scattering cross section across a target volume [36, 37]. For optical wavelengths, the interaction with biological tissue is dominated by the scattering coefficient. Figure 54.28 shows an energy-level diagram showing possible scattering mechanisms.

The absorption of photons in Figure 54.28 takes place at the excitation (or laser emission wavelength). If the photon is elastically scattered, the remitted signal is at the same wavelength as the excitation wavelength and is measured in the reflectance channel. For inelastic scattering one of the mechanisms possible is fluorescence. In this case, as shown in Figure 54.28, after the photon is absorbed, the molecule relaxes via nonradiative process to a lower energy level and then relaxes back to ground state. In this case, as the emitted photon has lower energy than the absorbed photon, it has a longer wavelength than the emission laser. This is measured in the fluorescence channel. Other inelastic scattering processes include phosphorescence and Raman scattering.

Contrast in confocal imaging can be either intrinsic or extrinsic. Intrinsic contrast measures signal remitted from the sample constituents, and for several biomedical applications at optical frequencies, this contrast is limited. Extrinsic contrast is generated by adding a dye or contrast agent that binds to specific aspects of the sample. The fluorescence or reflectance of these dyes is then used to generate high contrast images [33, 34].

54.3.4 Optical Coherence Tomography

54.3.4.1 Introduction Technologies such as OCT have allowed researchers to analyze the 3D structures of various materials. OCT works in the optical to near-infrared range and therefore provides high axial and lateral resolution. Depending upon how the signal is measured, OCT can broadly be divided in two types: time-domain OCT and spectral-domain OCT. OCT has been an established technique for medical imaging. It is commonly used for eye and retinal imaging [38]. Some recent research shows that it can be a useful tool to monitor the progression of glaucoma [39]. It has also been used for the imaging of coronary arteries [40]. As this technique uses optical or near-infrared range of frequencies, it is limited by scattering in biological tissue samples.

Broadband diode lasers are used as a light source in a typical OCT system. For example, a diode laser centered at 800 nm with a bandwidth of 50 nm is used for retinal imaging [38].

54.3.4.2 Time-Domain OCT A typical OCT system uses broadband diode lasers that have low coherence length. This property has been utilized to investigate the axial properties of the sample. The schematic diagram of time-domain coherence tomography using broadband sources is shown in Figure 54.29a.

The optical configuration is simply a Michelson interferometer. The broadband input beam is split between two arms using a beam splitter. One arm acts as a reference with signal reflected from a mirror, while the other acts as a sample arm in which the signal is reflected from the sample under investigation. The combined

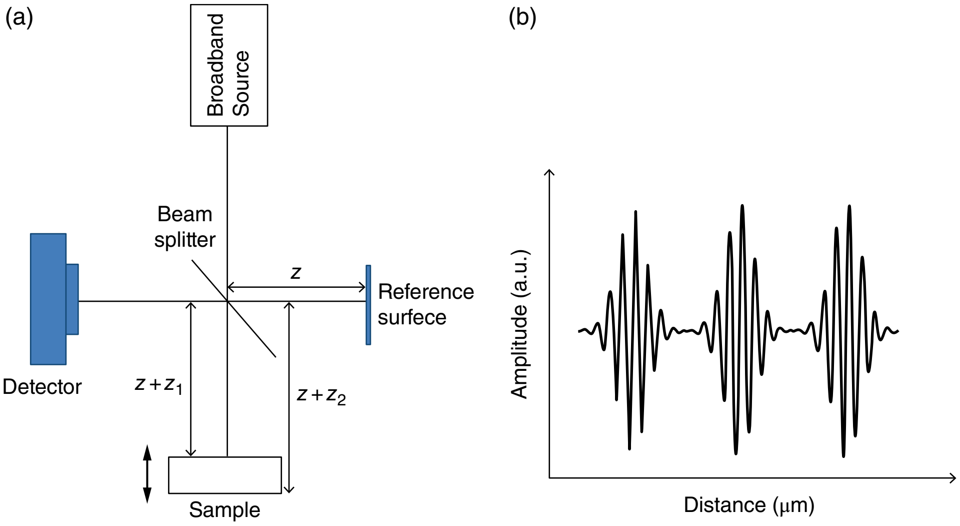


FIGURE 54.29 (a) Time-domain OCT system. (b) Interference pattern of the signal reflected from three different layers of the sample with respect to signal reflected from reference mirror surface.

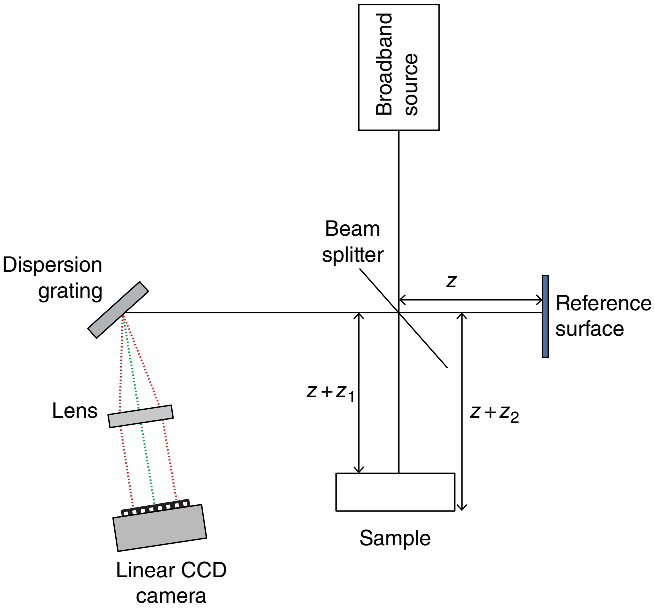


FIGURE 54.30 Schematic of spectral-domain interferometry.

reflected signal from the reference mirror and sample is measured at the detector. The signals reflected from different layers of the sample are delayed in time with respect to each other and also with respect to the reference mirror. In order to obtain the axial profile, the reference mirror is scanned axially. A sinusoidal signal is obtained at the detector when the optical path length of the reference mirror matches with a certain layer of the sample. The envelope of the interference pattern gives the axial profile of the sample. For example, if the sample shown in Figure 54.29a has three layers, it yields the axial profile shown in Figure 54.29b. The 3D profile of the sample is obtained by either raster scanning the probe beam or the sample itself and at the same time scanning the reference mirror to profile the sample axially. The speed of the scanning stage poses a limitation on how fast one can acquire the sample information.

54.3.4.3 Spectral-Domain OCT The spectral-domain low coherence interferometry technique is based on a Michelson interferometer, where a broadband, low-coherence source is used to illuminate a reference surface and a sample as shown in Figure 54.30. Consider a sample composed of two reflecting layers: one at an Optical Path Difference (OPD) of Z_1 and another at an OPD of Z_2 from the reference surface. The reflected light from the two illuminated objects is combined onto a dispersion grating that angularly separates the different wavelength components, which form an interference pattern on a linear CCD camera using a lens.

At the image plane, the phase difference between the signal reflected from the reference surface and the first layer of the sample is given by

$$\phi_1(k) = \phi_0 + \frac{4\pi}{\lambda} z_1$$

where λ is the wavelength, ϕ_0 is the phase change introduced by reflection at the first layer, and z_1 is the OPD between the reference surface and the first layer.

The phase difference can be rewritten in terms of the wavenumber, $k=2\pi/\lambda$, such that

$$\phi_1(k) = \phi_0 + 2kz_1$$

which is a linear relationship between the phase and the wavenumber. As the frequency along the k-axis is given by the rate of change of the phase with respect to the wavenumber, this leads to

$$f_{k1} = \frac{1}{2\pi} . \frac{\partial \phi_1(k)}{\partial k} = \frac{z_1}{\pi}$$

Similarly for the second reflecting layer in the sample, the frequency of the phase is given by

$$f_{k2} = \frac{1}{2\pi} . \frac{\partial \phi_2(k)}{\partial k} = \frac{z_2}{\pi}$$

One should note that analysis in the k-space is preferred to λ-space since f_k is independent of k, whereas the equivalent frequency in λ- space would vary with λ. Thus, sampling the interferogram intensity data uniformly along the k-axis would cause a broadening in the frequency spectrum, which reduces the axial depth resolution of the system.

The intensity distribution along the k-axis on the linear CCD camera as a result of the interference between the signal reflected from the reference surface and the signal reflected from the two layers of the sample can be expressed as

$$I(k) = I_0(k) + 2\sqrt{I_r(k) I_1(k)} \cos(\phi_1(k)) + 2\sqrt{I_r(k) I_2(k)} \cos(\phi_2(k))$$

where $I_0(k)$ is a DC term; $I_r(k)$, $I_1(k)$, and $I_2(k)$ are the intensity of the signals coming from the reference surface, the first layer, and the second layer, respectively; and $\phi_1(k)$ and $\phi_2(k)$ are the phase differences between the signal from the reference surface and the first layer and the reference surface and the second layer, respectively. Because of the presence of the cosine term and the fact that the phase difference is dependent on k, a modulation in the spectrum intensity along the k-axis is introduced.

The modulation is the result of two signals with frequencies given by $f_{k1} = z_1/\pi$ and $f_{k2} = z_2/\pi$, which are directly proportional to z_1 and z_2, respectively. One should note that if there were N_i layers in the sample, the signal from each layer would interfere with the reference signal and produce a modulation in the spectrum intensity along the k-axis whose frequencies would be proportional to the OPD between the reference surface and all N_i corresponding reflecting layers. The FT of the intensity pattern at the camera, after proper calibration from pixel to k-space, directly leads to the frequencies of the modulation and thus the position of the reflecting layers within the sample with respect to the reference surface. Such frequency spectrum obtained after the FT of the spectrum intensity is commonly called axial scan (A-Scan), or depth scan. In this technique, the maximum measurable OPD is limited by the depth range of the system, the details of which are discussed further.

In spectral-domain low coherence interferometry, the available spectral bandwidth of the light source is spread over the limited number of CCD pixels. If the total bandwidth of the interfering signal acquired with N_p number of pixels of the CCD camera is Δk, then the interval along the distance axis or the distance per pixel (Δd) after the FT is given by $\Delta d = \frac{1}{2}\frac{2\pi}{\Delta k}$. The 1/2 factor in this expression accounts for the doubling of the OPD after reflection. Thus the maximum OPD that can be measured as a function of N_p is given by

$$d_\mathrm{max} = \frac{1}{2}\frac{N_\mathrm{p}}{2}\Delta d = \frac{1}{2}\frac{N_\mathrm{p}}{2}\frac{2\pi}{\Delta k} = \frac{1}{2}\frac{N_\mathrm{p}}{2}\frac{\lambda_0^2}{\Delta\lambda}$$

A factor of 2 appears in the denominator of the above equation, since the signal after an FT is symmetric around zero OPD. Thus, the signals on the opposite sides of zero give the same information. For a Gaussian profiled spectrum, the depth range can be written as

$$d_\mathrm{max} = \frac{1}{2}\frac{2\ln 2}{\pi}\frac{N_\mathrm{p}}{2}\frac{\lambda_0^2}{\Delta\lambda}$$

A typical example of the frequency-domain interferometric signal at different depths for a given depth range is shown in Figure 54.31.

54.3.5 Femtosecond Laser Imaging

The branch of ultrashort laser imaging and spectroscopy is used in a wide range of areas, both scientific and industrial. The key point in ultrashort laser pulses is their time- and frequency-domain properties. In the time domain, the output consists of high-intensity pulses on the order of femtoseconds. In the frequency domain, the pulse train produced by a mode-locked laser consists of a broad spectrum of equidistant

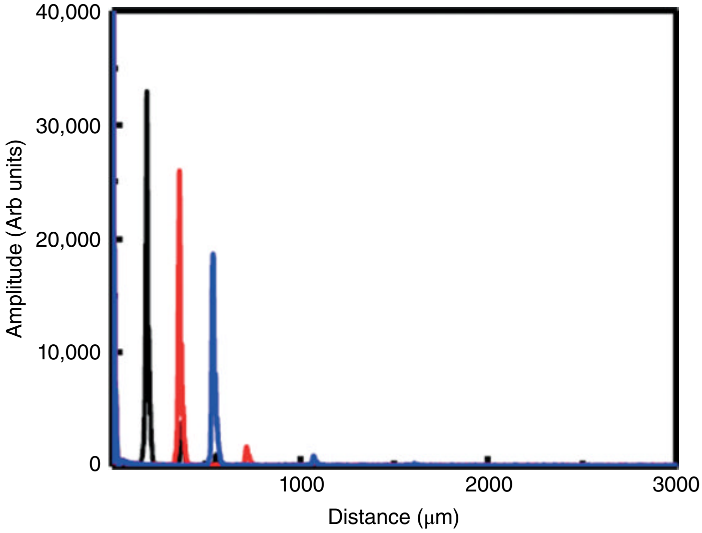


FIGURE 54.31 Change in the FFT signal with change in the optical path difference between sample and reference surface.

modes with a defined phase relationship. Ultrashort pulse lasers make it possible to probe samples on a femtosecond time scale, allowing a number of ultrafast chemical, biological, and physical processes [41–43].

Today, the titanium–sapphire (Ti:Al_2O_3) laser is the most widely used commercially available tunable laser. Like a ruby laser, to make Ti:sapphire, Ti_2O_3 is doped into a crystal of Al_2O_3, and Ti^{3+} ions then occupy some of the Al^{3+}-ion sites in the lattice. The typical concentrations range between 0.1 and 0.5% by weight. The laser is based on a four-level energy scheme. CW Ti:sapphire lasers are pumped by the green output of an argon laser. In pulsed operation, frequency-doubled Nd:YAG or Nd:YLF lasers as well as flashlamps are used. The tuning curve of this laser in a CW regime spans the wavelength range of over 400 nm between 670 and 1050 nm. This laser possesses a favorable combination of properties that are up to now the best among all known broadband laser materials. First, the active medium is solid state, which means long operational time and laser compactness. Second, sapphire has high thermal conductivity, exceptional chemical inertness, and mechanical resistance. Third, it has a very broad generated spectrum. The combination of all these properties made the Ti:sapphire crystal the most popular laser medium in the industry.

After selecting an appropriate lasing medium the next step is to generate ultrashort pulses. The following two techniques are used for generating short pulses:

1. Q-switching
2. Mode-locking

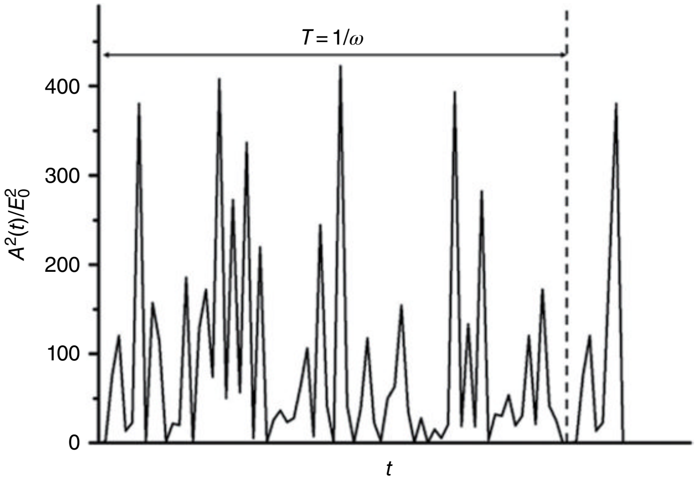


FIGURE 54.32 Squared amplitude of random oscillating modes in a cavity as a function of time.

54.3.5.1 Laser Q-Switching As been discussed, under CW operation, the population inversion reaches to its threshold value when oscillation starts. If a shutter is introduced in the cavity; when the shutter is closed, the population inversion can exceed a value far more compared to value when the shutter is open. When the shutter is closed, there will be gain in the cavity that greatly exceeds losses, and when the shutter is opened suddenly, the stored energy will be released in the form of short and intense light pulses. As this operation involves switching the Q-factor of the cavity from a low to high value, this technique is called Q-switching. This technique allows for the generation of laser pulses of the duration of photon decay times (few tens of nanoseconds) and power of the order of megawatts.

Several techniques have been applied to switch the cavity Q [44]. Broadly the devices can be grouped into active and passive Q-switches. Active Q-switching involves some kind of external operation to actively control the switching mechanism, for example, electro-optical Q-switching, where an external voltage is applied to switch the cavity Q-value. In passive Q-switching the switching takes place automatically using nonlinearity of some medium. An example of a passive Q-switch is the saturable absorber, which has low value of saturation intensity. When the laser power is low, the material does not allow the light to transmit through it; this means the switch is closed. When the laser power exceeds the saturation threshold of the medium, it becomes transparent, which in turn opens the switch.

54.3.5.2 Laser Mode-Locking In a laser cavity, many longitudinal modes can oscillate at a frequency given by $\omega_m = m2\pi\frac{c}{2L}$. These modes exhibit no phase relationship and oscillate independent of each other as shown in Figure 54.32. It is possible to make these random modes oscillate with a definite phase relation within the cavity; the process used is known as mode-locking and such lasers are referred to

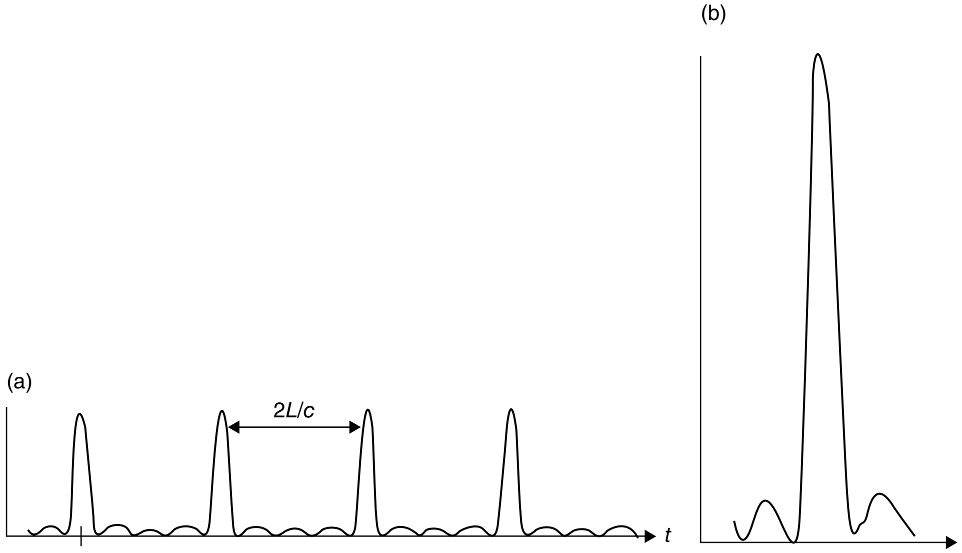


FIGURE 54.33 (a) Phase-locked modes propagating in the cavity. (b) Coherent sum of these pulses results in short but much higher-amplitude pulse.

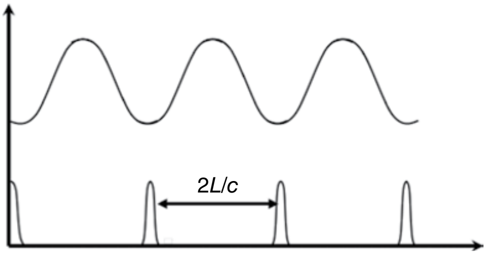


FIGURE 54.34 Cavity round-trip losses for amplitude-modulated mode-locking.

as mode-locked lasers. These locked modes can be considered as a Fourier series expansion of a periodic function in time, given by $T = 1/\omega = 2L/m2\pi c$, in which case they constitute a periodic pulse train, as shown in Figure 54.33a. The coherent sum of all the locked phases would interfere constructively, and a pulse with high peak power is obtained as shown in Figure 54.33b.

Like Q-switching the mode-locking mechanism can also be classified as active and passive. In this section, only active mode-locking (amplitude modulation) is discussed, and other active and passive mode-locking techniques are discussed in detail in reference [44]. In the case of amplitude modulation mode-locking, the modulator is placed at one end of cavity.

Figure 54.34 shows cavity round-trip losses, which are modulated by a time interval $T = 2L/c$. If the frequency of the modulation (ω_m) is equal to the frequency difference between two consecutive modes ($\Delta\omega$), the light pulses will pass through the modulator at the time of minimum loss. This is the steady-state condition: as if a light pulse passes

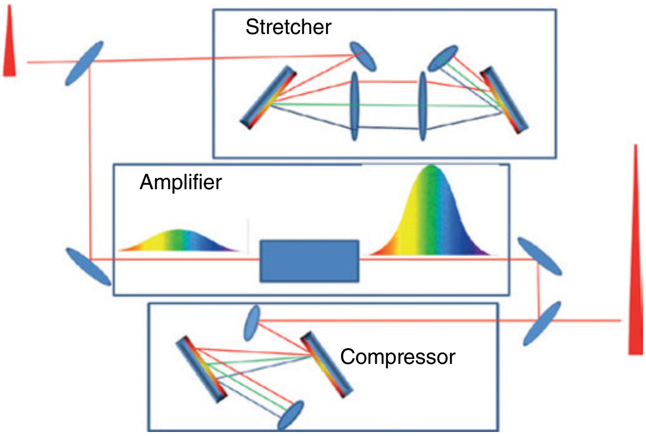


FIGURE 54.35 Chirped pulse amplification of a femtosecond laser pulse.

through the modulator at a time of minimum loss, which is the round-trip time of the cavity as well (2*L*/*c*), it will return to the modulator at its minimum loss as well. The time-varying losses of the modulator will damp the pulses reaching before and after the minimum loss time, thus mode-locking the laser by forcing a definite phase relationship between the propagating modes. In this case after each pass the pulse duration gets shorter, as the leading and trailing edge of the pulse attenuates in each pass. The shortening of the pulse is eventually limited by bandwidth of the gain medium. As the pulse width becomes shorter, the spectrum becomes large and at one point would fill the bandwidth of the laser medium. In this situation, the wings of the spectrum would no longer be amplified due to losses of the modulator, and this would limit the pulse duration of the laser.

54.3.5.3 Chirped Pulse Amplification Using mode-locking it is possible to generate laser pulses less than 10 fs, producing a peak power on the order of megawatts. Further amplification of the laser pulse is limited by nonlinear processes due to the high power and intensity of the laser pulse in the gain medium. To amplify the pulse further a technique adopted from radar technology is used, where a few tens of femtosecond short pulse is first stretched in time to several picoseconds, then amplified, and then finally compressed back to the ultrashort femtosecond pulse. The stretching reduces the intensity of the laser beam by 3–4 orders of magnitude and thus makes further amplification of the pulse possible. A typical Chirped Pulse Amplification (CPA)-based Ti:sapphire laser amplifier system is shown in Figure 54.35, where a grating pair is used as a stretcher and another pair as compressor.

54.3.5.4 Two-Photon Fluorescence Microscopy One application of high-power lasers for imaging is the two-photon fluorescence microscopy. In typical single-photon fluorescence microscopy, a molecule in its ground state will absorb a photon and is

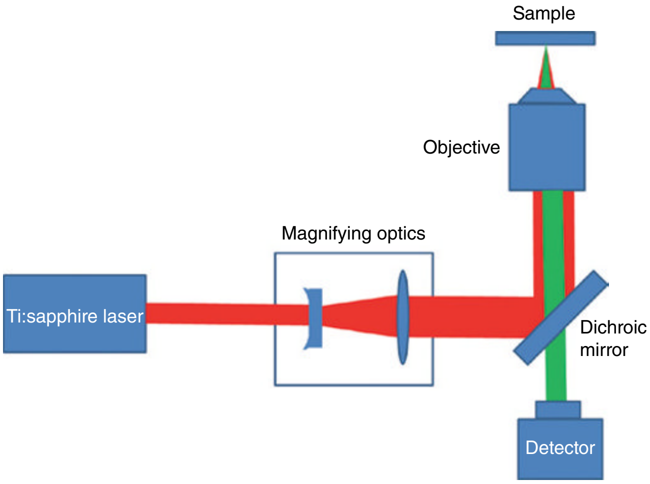


FIGURE 54.36 Schematic of two-photon fluorescence microscopy imaging setup.

excited to a higher energy state if the energy of the photon is equal to or higher than the energy difference between the two states. After relaxing to a lower vibrational state, the molecule will return to its electronic ground state, emitting a lower-energy photon (fluorescence) compared to the absorbed one (Fig. 54.28). The same process can take place if the sum of the energies of two photons is enough to reach the first excited state of the molecule. In this case, the probability of simultaneous absorption of two photons depends on the square of the intensity of the exciting beam.

The experimental design for a two-photon microscopy system is shown in Figure 54.36. In the setup, the femtosecond laser pulse from Ti:sapphire is magnified using negative and positive lens. The expanded beam is reflected using a dichroic mirror, which reflects infrared and transmits visible light. The reflected beam is focused using an objective lens to the diffraction limit at the sample position. At the focus, due to the high intensity of the beam, two-photon absorption will take place. Visible light, for example, green, will be generated due to fluorescence. The dichotic mirror will transmit the visible light to the detector.

Two-photon microscopy offers several advantages over other conventional techniques like confocal microscopy. The two-photon wavelength is twice as high as the wavelength of single photon; this wide difference between the excitation and emission wavelength makes sure that the excitation light and other scattering can be filtered out from the fluorescence signal. Secondly, two-photon microscopy is well suited for optically thick samples, as infrared radiation used in the two-photon excitation has greater penetration and less absorption/scattering in biological samples compared to green, blue, or ultraviolet light. The use of pinhole in confocal microscopy limits the number of photons reaching the detector, as the scattering of fluorescence deviates the path of the reflected photons. In two-photon microscopy the signal reaching at the detector is much higher as the deviated light will still reach the detector. Several biomedical applications of two-photon fluorescence microscopy are currently under investigation [45].

54.3.6 Laser Raman Spectroscopy

Given the myriad of spectroscopic measurement systems developed, Raman spectroscopy exemplifies one of the methodologies engineered for performing chemical analysis. Pioneering the technique, C.V. Raman recognized that an energy shift occurred for photons scattered off molecules. He realized that the inelastic process of scattering light caused vibrational and/or rotational transitions of the molecules that shifted the scattered photons to lower or higher energy levels.

If the molecule transitions to a higher energy level when illuminated, then the photon scatters with a lower energy level exhibiting a longer wavelength referred to as a Stokes shift. Likewise, if the molecular transition is to a lower level, the scattered photon carries away the excess energy and thus has a higher energy and shorter wavelength than the incident light referred to as an anti-Stokes shift. The Raman shift is expressed in units of wavenumber as

$$\Delta w\left(\mathrm{cm}^{-1}\right)=\left(\frac{1}{\lambda_{\mathrm{i}}}-\frac{1}{\lambda_{\mathrm{s}}}\right)$$

where λ_{i} incident wavelength and λ_{s} is the scattered wavelength. Raman received the 1930 Nobel Prize in Physics for his discovery.

Since the vibrational and/or rotational transitions of a molecular structure are uniquely driven by the chemical bonds, Raman spectroscopy can be used to identify chemical species of molecules. However the Raman scattering is weak. A variety of techniques have been developed to improve sensitivity and resolution. Taking advantage of the phenomena, the ultraviolet wavelength pulses of an excimer laser can be passed through a gas cavity to generate wavelengths as short as 35.5 nm, the seventh harmonic of KrF, enabling spectrometry in the extreme ultraviolet. Raman interactions in molecular gases such as hydrogen produce smaller frequency shifts when a tightly focused excimer laser beam is passed through the hydrogen gas cell. Though the Raman shift can produce longer or shorter wavelengths, generally longer wavelength light is generated more efficiently.

Given the commercial availability of Raman frequency shifters, Raman spectroscopy can be readily used to observe vibrational, rotational, and other low-frequency modes in molecules. With both pulsed and CW lasers generating sufficient power densities for driving the harmonic generation technique, the inelastic scattering of monochromatic laser light in the ultraviolet, visible, and near-infrared regime yields information about the vibrational modes of molecules and therefore the molecular structure of materials. Raman spectroscopy is particularly useful for mineral identification because chemical species exhibit unique spectral characteristics and can identify molecular structure by chemical bonds [46]. Techniques such as Resonance Raman Scattering (RRS), Surface-Enhanced Raman Scattering (SERS), and coherent anti-Stokes Raman spectroscopy have been developed to observe low Raman Scattering cross sections [23].

54.3.7 Suggested Reading

Several reference books cover different imaging and spectroscopic systems introduced in this section. *Terahertz Techniques* by Bründermann, Hübers and Kimmitt, published by Springer, provides an excellent overview of terahertz frequency measurement systems and covers various aspects of imaging and detector technology. *Handbook of Photonics for Biomedical Science*, edited by V. Tuchin and published by CRC Press, covers a wide range of imaging systems used in biomedical applications including OCT, terahertz imaging, confocal microscopy, and Raman spectroscopy. *Handbook of Biological Confocal Microscopy* by J. B. Pawley, published by Springer, also provides an excellent textbook for confocal microscopy.

REFERENCES

1. Siegman, A. E., "*Lasers*," University Science Books, Mill Valley, CA, 1986.
2. Kogelnik, H. and Li, T., "Laser beams and resonators," *Applied Optics*, 5(10), 1550–1567, 1966.
3. Guenther, R. D., "*Modern Optics*," John Wiley & Sons, Inc., New York, 1990.
4. Weber, M. J., "*CRC Handbook of Laser Science and Technology, Volume II Gas Lasers*," CRC Press, Inc., Boca Raton, FL, (1982).
5. Pozar, D. M., "*Microwave Engineering*," John Wiley & Sons, Inc., Hoboken, NJ, 2005.
6. Danylov, A., "Frequency stabilization, tuning, and spatial mode control of terahertz quantum cascade lasers for coherent transceiver applications," Ph.D. dissertation, Department of Physics and Applied Physics, University of Massachusetts Lowell, USA, 2010.
7. Faist, J., Capasso, F., Sivco, D. L., Sirtori, C., Hutchinson, A. L., and Cho A. Y., "Quantum cascade laser," *Science*, 264(5158), 553–556, 1994.
8. Danylov, A. A., Waldman, J., Goyette, T. M., Gatesman, A. J., Giles, R. H., Linden, K. J., Neal, W. R., Nixon, W. E., Wanke, M. C., and Reno, J. L., "Transformation of the multimode terahertz quantum cascade laser beam into a Gaussian, using a hollow dielectric waveguide," *Applied Optics*, 46(22), 5051–5055, 2007.
9. Waldman, J., Fetterman, H. R., Goodhue, W. D., Bryant, T. G., and Temme, D. H., "Submillimeter modeling of millimeter radar systems," *Proceedings of SPIE*, 259 Millimeter Optics, 152–157, 1980.
10. Chen, T. J., Chu, T. H., and Chen, C., "A new calibration algorithm of wideband polarimetric measurement system," *IEEE Transactions of Antennas and Propagation*, 39(8), 1188–1192, August 1991.

11. DeMartinis, G. B., Coulombe, M. J., Horgan, T. M., Giles, R. H., and Nixon, W. E., "A 240 GHz Polarimetric Compact Range for Scale Model RCS Measurements," Antenna Measurements Techniques Association (AMTA), Atlanta, GA, pp. 3–8, October 2010.
12. Goyette, T. M., Dickinson, J. C., Waldman, J., and Nixon, W. E., "1.56-THz compact radar range for W-band imagery of scale-model tactical targets," *Proceedings of SPIE*, 4053, 615–622, Algorithms for Synthetic Aperture Radar Imagery VII; Edmund G. Zelnio; Ed. August 2000.
13. Goyette, T. M., Dickinson, J. C., Waldman, J., and Nixon, W. E., "Three dimensional fully polarimetric W-band ISAR imagery of scale-model tactical targets using a 1.56 THz compact range," *Proceedings of SPIE*, 5095, 66–74, Algorithms for Synthetic Aperture Radar Imagery X; Edmund G. Zelnio, Frederick D. Garber; Ed. September 2003.
14. Brundermann, E., Hubers H.-W., and Kimmit, M. F., "*Terahertz Techniques*," Springer Series in Optical Sciences, Springer, Berlin/London, 2012.
15. Suzaki, Y. and Tachibana, A., "Measurement of the μm sized radius of Gaussian laser beam using the scanning knife-edge," *Applied Optics*, 14(12), 2809–2810, 1975.
16. Khosrofian, J. M. and Garetz, B. A., "Measurement of a Gaussian laser beam diameter through the direct inversion of knife-edge data," *Applied Optics*, 22(21), 3406–3410, 1983.
17. Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A., and Fujimoto, J. G., "Optical coherence tomography," *Science*, 254, 1178–1181, 1991.
18. Schmitt, J. M., "Optical coherence tomography (OCT): a review," *IEEE Journal of Selected Topics in Quantum Electronics*, 5(4), 1205–1215, July/August 1999.
19. Wojtkowski, M., Leitgeb, R., Kowalczyk, A., Bajraszewski, T., and Fercher, A. F., "In vivo human retinal imaging by Fourier domain optical coherence tomography," *Journal of Biomedical Optics*, 7(3), 457–463, 2002.
20. Paddock, S. W., "Principles and practices of laser scanning confocal microscopy," *Molecular Biotechnology*, 16(2), 127–149, 2000.
21. Pawley, J. B., "*Handbook of Biological Confocal Microscopy*," 3rd ed., Springer Science + Business Media, New York, 2006.
22. Rajadhyaksha, M., Grossman, M., Esterowitz, D., Webb, R. H., and Anderson, R. R., "In vivo confocal scanning laser microscopy of human skin: melanin provides strong contrast," *Journal of Investigative Dermatology*, 104(6), 946–952, 1995.
23. Tuchin, V. V. (Editor), "*Handbook of Photonics for Biomedical Science*," CRC Press, Taylor & Francis Group, Boca Raton, FL, 2010.
24. Pickwell, E. and Wallace, V. P., "Biomedical applications of terahertz technology," *Journal of Physics D: Applied Physics*, 39(17), R301, 2006.
25. Kim, S. M., Baughman, W., Wilbert, D. S., Butler, L., Bolus, M., Balci, S., and Kung, P., "High sensitivity and high selectivity terahertz biomedical imaging," *Chinese Optics Letters*, 9(11), 110009, 2011.
26. Oh, S. J., Choi, J., Maeng, I., Park, J. Y., Lee, K., Huh, Y. M., Suh, J., Haam, S., and Son, J. H., "Molecular imaging with terahertz waves," *Optics Express*, 19(5), 4009–4016, 2011.

27. Arbab, M. H., Dickey, T. C., Winebrenner, D. P., Chen, A., Klein, M. B., and Mourad, P. D., “Terahertz reflectometry of burn wounds in a rat model,” *Biomedical Optics Express*, 2(8), 2339–2347, 2011.
28. Son, J. H. (Editor), “*Terahertz Biomedical Science and Technology*,” CRC Press, Taylor & Francis Group, Boca Raton, FL, 2014.
29. Joseph, C. S., Patel, R., Neel, V. A., Giles, R. H., and Yaroslavsky, A. N., “Imaging of ex vivo nonmelanoma skin cancers in the optical and terahertz spectral regions. Optical and terahertz skin cancers imaging,” *Journal of Biophotonics*, 7, 295–303, 2014, published online 2012.
30. Löffler, T., Bauer, T., Siebert, K., Roskos, H., Fitzgerald, A., and Czasch, S., “Terahertz dark-field imaging of biomedical tissue,” *Optics Express*, 9, 616–621, 2001.
31. Doradla, P., Alavi, K., Joseph, C., and Giles, R., “Detection of colon cancer by continuous-wave terahertz polarization imaging technique,” *Journal of Biomedical Optics*, 18(9), 090504–090504, 2013.
32. Minsky, M., “Memoir on inventing the confocal scanning microscope,” *Scanning*, 10(4), 128–138, 1988.
33. Wirth, D., Snuderl, M., Sheth, S., Kwon, C. S., Frosch, M. P., Curry, W., and Yaroslavsky, A. N., “Identifying brain neoplasms using dye-enhanced multimodal confocal imaging,” *Journal of Biomedical Optics*, 17(2), 0260121–0260127, 2012.
34. Snuderl, M., Wirth, D., Sheth, S. A., Bourne, S. K., Kwon, C. S., Ancukiewicz, M., Curry, W. T., Frosch, M. P., and Yaroslavsky, A. N., “Dye-enhanced multimodal confocal imaging as a novel approach to intraoperative diagnosis of brain tumors,” *Brain Pathology*, 23(1), 73–81, 2013.
35. Al-Arashi, M. Y., Salomatina, E., and Yaroslavsky, A. N., “Multimodal confocal microscopy for diagnosing nonmelanoma skin cancers,” *Lasers in Surgery and Medicine*, 39(9), 696–705, 2007.
36. Caspers, P. J., Lucassen, G. W., Carter, E. A., Bruining, H. A., and Puppels, G. J., “In vivo confocal Raman microspectroscopy of the skin: noninvasive determination of molecular concentration profiles,” *Journal of Investigative Dermatology*, 116(3), 434–442, 2001.
37. Caspers, P. J., Lucassen, G. W., and Puppels, G. J., “Combined in vivo confocal Raman spectroscopy and confocal microscopy of human skin,” *Biophysical Journal*, 85(1), 572–580, 2003.
38. Swanson, E., Izatt, J., Lin, C., Fujimoto, J., Schuman, J., Hee, M., Huang, D., and Puliafito, C., “In vivo retinal imaging by optical coherence tomography,” *Optics Letters*, 18, 1864–1866, 1993.
39. Schuman, J. S., Hee, M. R., Puliafito, C. A., Wong, C., Pedut-Kloizman, T., Lin, C. P., Hertzmark, E., Izatt, J. A., Swanson, E. A., and Fujimoto, J. G., “Quantification of nerve fiber layer thickness in normal and glaucomatous eyes using optical coherence tomography: a pilot study,” *Archives of Ophthalmology*, 113, 586–596, 1995.
40. Jang, I.-K., Tearney, G. J., MacNeill, B., Takano, M., Moselewski, F., Iftima, N., Shishkov, M., Houser, S., Aretz, H. T., Halpern, E. F., and Bouma, B. E., “In vivo characterization of coronary atherosclerotic plaque by use of optical coherence tomography,” *Vascular Medicine*, 111, 1551–1555, 2005.

41. Stolow, A., Bragg, A. E., and Neumark, D. M., "Femtosecond time-resolved photoelectron spectroscopy," *Chemical Reviews*, 104, 1719–1758, 2004.

42. Kukura, P., McCamant, D. W., and Mathies, R. A., "Femtosecond stimulated Raman spectroscopy," *Annual Review of Physical Chemistry*, 58, 461–488, 2007.

43. Holzwart, A. R., "Applications of ultrafast laser spectroscopy for the study of biological systems," *Quarterly Reviews of Biophysics*, 22, 239–326, 1989.

44. Svalto, O., "*Principles of Lasers*," 5th ed., Springer Science + Business Media, LLC, New York, 2010.

45. So, P. T., Dong, C. Y., Masters, B. R., and Berland, K. M., "Two-photon excitation fluorescence microscopy," *Annual Review of Biomedical Engineering*, 2(1), 399–429, 2000.

46. Chen, H. and Stimets, R. W., "Fluorescence of trivalent neodymium in various materials excited by a 785 nm laser," *American Mineralogist*, 99(2–3), 332–342, 2014.

55

MAGNETIC FORCE IMAGES USING CAPACITIVE COUPLING EFFECT

Byung I. Kim
Department of Physics, Boise State University, Boise, ID, USA

55.1 INTRODUCTION

There are several conventional approaches for mapping magnetic field distributions. Optical techniques based on the Kerr effect have moderate spatial resolution of about 0.5 μm. Bitter pattern technique causes degradation of the sample surface. Although electron beam imaging techniques like Lorentz microscopy, scanning electron microscope with polarization analysis (SEMPA), and differential phase contrast (DPC) STEM are known to have higher spatial resolution, sample preparation and operation are difficult. But magnetic force microscopy (MFM) needs no special sample preparation and provides different and somewhat complementary information to e-beam imaging techniques as well as high spatial resolution (10–100 nm).

Owing to these merits, MFM has been applied to the study of recording media such as longitudinal media and magnetic multilayer film [1]. As the recording density is increased, it is important to understand the submicron magnetic structure behavior. In this respect, MFM could be a promising tool. Furthermore, MFM has been successfully applied to the study of fundamental science: For example, the macroscopic quantum tunneling (MQT) [2], the biological magnetism [3], the domain-structure change with the film thickness [4], the magnetic field [5], and so forth. There also have been several

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

recent attempts to study the vortex structure of the high-temperature superconductors with MFM [6, 7].

MFM has matured and developed into a routine method for imaging magnetic surface structures [8]. Although MFM is one of the most important imaging tools [9] of nanoscale magnetic structures such as interspin interactions [10], vortex ratchets and cores [11, 12], carrier-controlled ferromagnetism [13], superconducting vortices [14], the separation of magnetic and topographic signals in MFM has been a long-standing issue since its development 20 years ago [8]. This issue still remains largely unsolved and thus has limited the current capability of the MFM as a quantitative magnetic imaging tool [9, 15].

Since magnetic forces between the tip and the sample are very weak (~10^{-12} N), an ac-mode operation with a small vibration amplitude between the tip and the sample is used to detect these weak magnetic interactions [16]. When the oscillation amplitude is used as a feedback signal, the image obtained represents a constant force gradient contour of the magnetic sample surface. Most of the systems currently implemented in MFM are based on mechanically driven vibrations utilizing a piezo device called a bimorph (see Fig. 55.1a).

The bimorph-driven system is limited in its ability to separate the MFM signal from the topographic signal on surfaces where topographic features match or exceed the tip height where imaging occurs [15, 17, 18]. Mixing of the topographic and magnetic

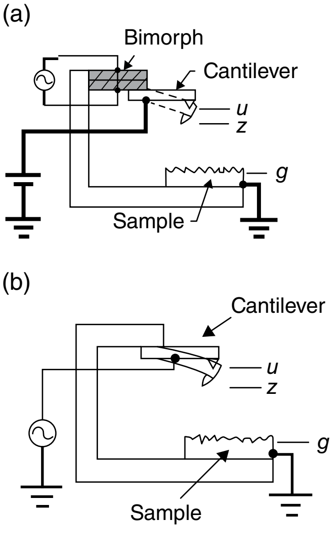


FIGURE 55.1 (a) Schematic of the bimorph-driven system. (b) Schematic of the electrostatic force modulation system. Source: Reprinted with permission from Ref. 25. Copyright (2012), American Institute Physics.

signals can make it difficult to extract the intrinsic magnetic structure from MFM images, thereby limiting our understanding of magnetic surfaces.

Existing separation methods have focused on monitoring magnetic signals as the tip follows the sample topography. Schönenberger et al. developed an MFM technique that monitors the dc magnetic force (magnetic image) under the constant amplitude (topography image) between the sample and the tip [19, 20]. The technique gives *reasonably good* separation of topographic and magnetic information. Later, Giles et al. made an important contribution to the separation of the two signals through the development of a two-pass technique, the so-called "tapping/lift mode" MFM [21]. In this technique, the first scan is for the topographic signal, then the second scan is repeated to record magnetic information (either as variations in amplitude, frequency, or phase of the cantilever oscillation) in the same line scan at a constant elevated height above the surface. The technique is good at eliminating topographic features in the magnetic image, but the tip contact to the sample during the tapping scan causes some problems.

First, the tip stray field can frequently and significantly disturb the magnetization distribution in a sample, especially in a soft magnetic material, during tapping [22]. Second, the sharp magnetic tip may be easily worn in the presence of topographic variations on the sample surface [15]. Third, the imaging mode places higher demands on instrument stability, or the continual alternating line scan causes serious reduction of the correlation between the two scans due to drift [15]. To minimize these problems, the topographic features need to be separated from MFM images *without touching the sample surface*, thus minimizing the influence of the tip stray field on a sample magnetization.

As we shall see in this chapter, however, magnetic images taken with the conventional noncontact amplitude modulation (AM) MFM frequently picked up topographic features even at larger distances (100–200 nm) [15, 23]. Here, the mixing mechanism of the two signals in the conventional AM MFM is identified using nonlinear dynamics between the tip and the surface. An MFM method has recently been designed and developed using electrostatic force modulation to separate the magnetic domain structure from the topographic structure on magnetic samples with rough surfaces [24, 25]. In this method, a capacitive coupling is introduced between the tip and the sample using electrostatic force modulation (see Fig. 55.1b).

In this chapter, an electrostatic force modulation technique is introduced and the detailed mechanism of the stability improvement is described through a direct comparison of the traditional bimorph method and the electrostatic force modulation method. This will be done by comparing the amplitude–frequency curves for the two systems. The effects on imaging stability of tip–sample separation and perturbations (e.g., collisions between the cantilever tip and a tall hillock structure) are also investigated. Through this investigation, it will be revealed that the superior performance of the electrostatic force modulation system results from the long-range electrostatic capacitance effect and the direct force modulation effect.

55.2 EXPERIMENT

55.2.1 Principle

The technique can be understood by considering the equation of motion of the cantilever probe with a single-point mass m undergoing one-dimensional (1D) forced harmonic oscillation along the vertical z axis:

$$m\frac{\partial^2 z}{\partial t^2}+\gamma\frac{\partial z}{\partial t}+k\left(z-u\right)=F\left(z\right) \tag{55.1}$$

The probe–surface interaction force can be incorporated as

$$F\left(z\right)=F_0+F'\left(z_0\right)\zeta \tag{55.2}$$

and the positions of the bimorph and the lever, u and z, respectively, can be given by

$$u=u_0+a\exp\left(i\omega t\right) \tag{55.3}$$

$$z=z_0+\zeta \tag{55.4}$$

Then,

$$m\frac{\partial^2 \zeta}{\partial t^2}+\gamma\frac{\partial \zeta}{\partial t}+k\left[\zeta-a\exp\left(i\omega t\right)\right]=F'\zeta \tag{55.5}$$

The amplitude of vibration of lever is given by

$$A_\mathrm{b}\left(\omega,F'\right)=\frac{ak\exp\left(i\theta\right)}{k'-\omega^2 m+i\omega\gamma} \tag{55.6}$$

where $k'=k-F'$. When the driving frequency is near the resonance frequency, $\omega\sim\omega_0$, the local force gradient F' will shift the resonance frequency by an amount $\Delta\omega_0\sim\omega_0\left(F'/2k\right)$. This resonance frequency shift causes the change of the amplitude in the cantilever vibration. The constant force gradient contour of a selected area can be obtained by taking the servo electronics input as data.

55.2.2 Instrumentation

A heavily doped Si cantilever coated with a thin magnetic Co layer is used (nanosensors) for the detection of the magnetic force. The cantilever has a force constant of 2.1 N/m and a resonant frequency of about 106 kHz. The lock-in phase was set to make the driving signal be "in phase" with the response signal of the cantilever at the tip–sample distance greater than 1000 nm from the sample surface. The in-phase ω component amplitude, X_ω ($=R_\omega \cdot \cos \phi_\omega$), was measured as a function of tip–sample distance to understand its behavior. The in-phase amplitude is selected as a feedback signal because of its higher sensitivity with the tip–sample distance over the amplitude R_ω [25]. The tip is magnetized by placing the cantilever in a magnetic field of 0.2 T, aligned perpendicularly to the lever, for 3 min [18]. The electrical contact for the tip and sample was made with a silver paste. The scan rate is 1 Hz, and the time constant is set at 100 μs in a lock-in amplifier (EG&G Princeton Applied Research, Model 5302). A CoCr film deposited on the glass substrate by the dc magnetron sputtering method up to a thickness of 300 nm was used as a sample for this study. All the data shown here are obtained with a commercial AutoProbe LS in air [26].

MFM cantilevers attached on the bimorph were driven by the voltage-controlled oscillator (VCO). For MFM imaging using the bimorph-driven method in Figure 55.1a, a dc voltage +10 V was applied between the tip and the sample to provide the servo force F_C for feedback to keep the tip from crashing into the surface during scanning [27]. For the electrostatic force modulation method in Figure 55.1b, a sinusoidal signal V_{ac} sin(ωt) was applied using a function generator (Hewlett Packard, HP 33120A) to modulate the gap between the tip and the sample. In this experiment, we applied $V_{ac} = 14$ V between the tip and the sample surface at an operating frequency of 53 kHz, half the resonance frequency of the cantilever. For both magnetic imaging methods, magnetic and topographic images were obtained under a constant amplitude feedback condition by using the difference between the output of the amplitude and set amplitude as the error signal. The typical scan rate was 1 Hz. The cantilevers used here are the microfabricated triangular Si_3N_4 cantilevers with pyramidal tips.

The AutoProbe LS system employed an optical beam deflection–detection method, which is one of the most commonly used methods to detect interaction between the magnetic tip and the magnetic sample. The optical-detection system is composed of a laser diode, a bicell photodiode as a position sensitive detector, and a mirror for aligning laser beam on the backside of cantilever. Under the cantilever, there is the sample stage mounted on the tube piezo scanner. The tip can be positioned on the selected area of the sample surface by an XY micrometer stage combined with CCD camera.

55.2.3 Approach

The tip is approached to the sample after loading the sample and focusing the laser beam on the backside of the cantilever. At the moment the cantilever amplitude drastically decreases, the tip is retracted from the contact point about 100 nm by the stepping motor. This trapping of the tip can also be observed easily in the CCD image. It implies that there is a strong adhesive force directly on the magnetic sample. From the amplitude–frequency curve obtained at that point, we choose an optimum frequency at which the amplitude is 81.6% of the resonance peak [28]. In this experiment, the operating frequency is 53 kHz. The absolute vibration amplitude of the cantilever can be determined using the Michelson–Morley interferometry at this operating frequency before the experiment. After the scan size, the offset, the time constant, the gain, and the slope setting are determined, the images are obtained.

The amplitude changes are measured with the lock-in amplifier as the error signal of the feedback. If the output of the bicell photodiode is the error signal of feedback directly, MFM can be operated as an atomic force microscope (AFM). This is useful when a topographic image of a magnetic sample is needed. The feedback circuits are composed of the control circuit and the high-voltage amplifier. The feedback-control parameters, like the time constant and the gain, can be manually adjusted for the optimum feedback condition.

It takes 5–10 min to get an MFM image for 256 × 256 data acquisition. In the data file, the scan frequency, the scanning area, the offset, the bias voltage, the data gain, and the force gradient set value are also stored for each image. Image process and analysis can be done after data acquisition. When data are acquired in the constant force gradient mode with the feedback on, the image represents a mapping of the force gradient experienced by the magnetic tip as a function of position. Maximum scanning area is 50 × 50 μm with the system used. The finest structure observed in our MFM images is in the region of 50–100 nm. All data presented here are raw data.

55.3 RESULTS AND DISCUSSION

We must approach to improve the resolution of MFM and study any relation between topography and magnetic structure, if we want magnetic tip to the surface more closely (~200 Å). In this case, it is important to separate the topographic and magnetic features of the sample. One way to solve this problem is to apply a sinusoidal voltage to the tip to induce the modulated electrostatic force. The information of modulated electrostatic signal extracted by the lock-in will be used to control the height of the scan [24].

If we assume that the cantilever is parallel to the sample and that the tip is a point dipole with fixed moment, $\vec{m} = (m_x, m_y, m_z)$, in the stray magnetic field of the sample surface, $\vec{B} = (B_x, B_y, B_z)$, then we can write down force derivative as

$$F' = m_x \frac{\partial^2 B_x}{\partial z^2} + m_y \frac{\partial^2 B_y}{\partial z^2} + m_z \frac{\partial^2 B_z}{\partial z^2}. \tag{55.7}$$

Since the force derivative change is related with the second derivative of the sample stray field, the contrast will be obvious at the abrupt change of the vertical component and the horizontal component of the magnetization like the domain boundary. Therefore, the lines of the image seem to be the natural magnetic domain walls with the thickness of about 50–100 nm. For more complete understanding, this value must be compared with the result based on the magnetic anisotropy measurement [29].

55.3.1 Separation of Topographic Features from Magnetic Force Images Using Capacitive Coupling Effect

55.3.1.1 Topographic Features in a Magnetic Force Image A magnetic tip is mechanically vibrated at the resonance frequency of the cantilever with the free oscillation amplitude of 96 nm by the acoustic excitation method such as the bimorph-driven system (inset of Fig. 55.2a). A dc bias voltage $V_{dc} = +10\text{V}$ is applied between the tip and the surface to create a long-range attractive electrostatic interaction, which is essential to make the feedback polarity stay constant, regardless of the attractive or repulsive magnetic forces for stable feedback [20, 23]. In the in-phase amplitude–distance curve, the amplitude decreases monotonically as the tip moves toward the surface for the distance between 64 and 1000 nm ("noncontact region") in Figure 55.2a. For magnetic imaging, the average tip–sample distance of the oscillating tip should be positioned in the noncontact region where the magnetic signal is dominant over the topographic signal. Since the sign of slope is positive (i.e., $\partial X_\omega(d)/\partial d > 0$) in the noncontact region, it is necessary to set the feedback polarity to positive for stable magnetic imaging. The linear tapping region also has the same sign of slope [30, 31], indicating that two stable states exist for a given set-amplitude ($X_{\omega,\text{SP}}$): one in the tapping region and the other in the noncontact region.

The key approach of this chapter is to solve a 20-year-old problem by using the fundamental understanding of nonlinear stochastic physics in the tip–sample interactions of AM AFM, recently discovered and published by Garcia and San Paulo [32].

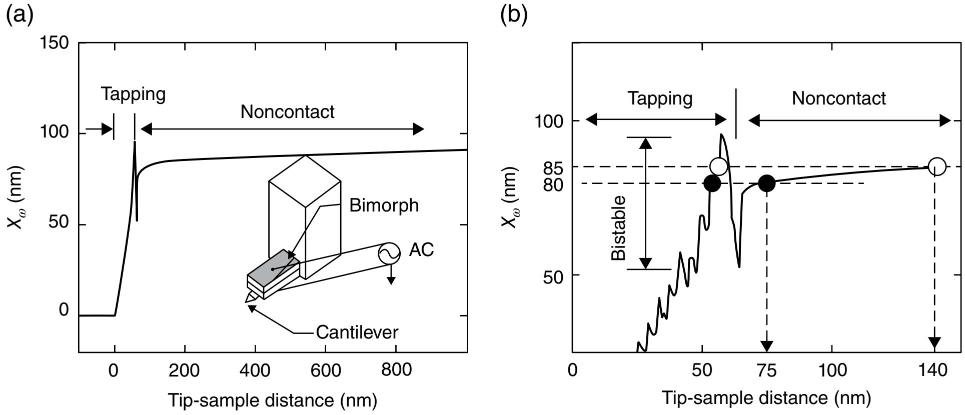


FIGURE 55.2 (a) A typical in-phase amplitude–distance curve at the operating frequency of resonance 106 kHz and the free oscillation amplitude of 96 nm on a CoCr film. (inset) A schematic sketch of the acoustic excitation method. (b) An enlarged amplitude–distance curve in the distance range between 0 and 150 nm. Horizontal dashed lines represent the feedback set amplitudes of 80 and 85 nm that correspond to the average noncontact distances of 75 and 140 nm from the surface, respectively, as marked with dashed arrows. The two stable states are marked with two circles for each set amplitude. Source: Reprinted with permission from Ref. 24. Copyright (2009), American Institute Physics.

Garcia and San Paulo associated abrupt changes in height of topographic features with the continual switching of the oscillating tip between the two stable states during the tip scanning over the surface in the AM AFM. Similarly, in this chapter, the topographic features in the magnetic image are attributed to the switching between the bistable states. This effect is analogous to the sudden transition from the noncontact to contact states in contact mode AFM [22]. In principle, both effects result from the intrinsic nonlinear mechanical bistability of the sensor–sample assembly. In the contact AFM, the so-called "snap-to-contact" issue has been resolved by removing the bistability using a voltage-activated force feedback [33, 34]. Similarly, in this chapter, this issue has been addressed by removing the stable state in the tapping region. The bistability originates from the dc bias voltage to induce *monotonic decrease* of the long-range amplitude as the distance becomes smaller in the noncontact region between 64 and 1000 nm (Fig. 55.2a).

In order to verify this concept, the magnetic imaging was performed repeatedly at two different set-amplitudes ($X_{\omega,\mathrm{SP}}$) of 80 and 85 nm at the same tip location. The average operating spacing can be determined graphically by projecting to the x-axis from the intersecting points where the open-loop curve and a feedback set point line meet, as shown in Figure 55.2b. The origin is chosen as the point where the linear-extrapolated line in the tapping region crosses with the x-axis, and thus the x-coordinate represents the average tip–sample distance [31]. The average noncontact distances are determined to be 75 and 140 nm for the set amplitudes of 80 and 85 nm, respectively. Figure 55.3a and b shows stripe-like structures that correspond to the magnetic domain features because of a higher contrast variation at a larger tip–sample distance (see line scans in Fig. 55.3a and b), a well-known signature for magnetic features [35]. Topographic hillocks and grains appear as spots with diameters of 1000–2000 nm and spots with diameters of 100–300 nm, respectively, in Figure 55.3a (see more details in Fig. 55.5b). Most of the topographic features drastically disappear except a few hillocks marked with arrows in Figure 55.3b. The remarkable change of pickup ratio results from the noncontact lift-height change by 65 nm (= 140 − 75 nm) in Figure 55.2b. In the constant amplitude mode, the average tip–sample distance continues to vary in order to maintain the set-amplitude constant. Because the set-amplitude 80 nm is comparable to the average tip–sample distance 75 nm, feedback perturbations allow for switching between the bistable states for picking up topographic features. The switching between two stable states would be almost equally probable during the data acquisition, explaining the pickup ratio of nearly one as shown in Figure 55.3a. At the set amplitude of 85 nm, the average noncontact distance of 140 nm is somewhat bigger than, but still smaller than, the heights of bigger hillocks of 200–300 nm. The oscillating tip spends most of its time collecting magnetic features *in the noncontact region* except in the regions of the bigger hillocks, explaining the small pickup ratio in Figure 55.3b. The bistable states exist for almost all of the average tip–sample spacing from 64 nm up to >1000 nm in Figure 55.2a. This indicates that topographic features always have a chance to appear in a magnetic image, except for smooth and homogeneous sample surfaces.

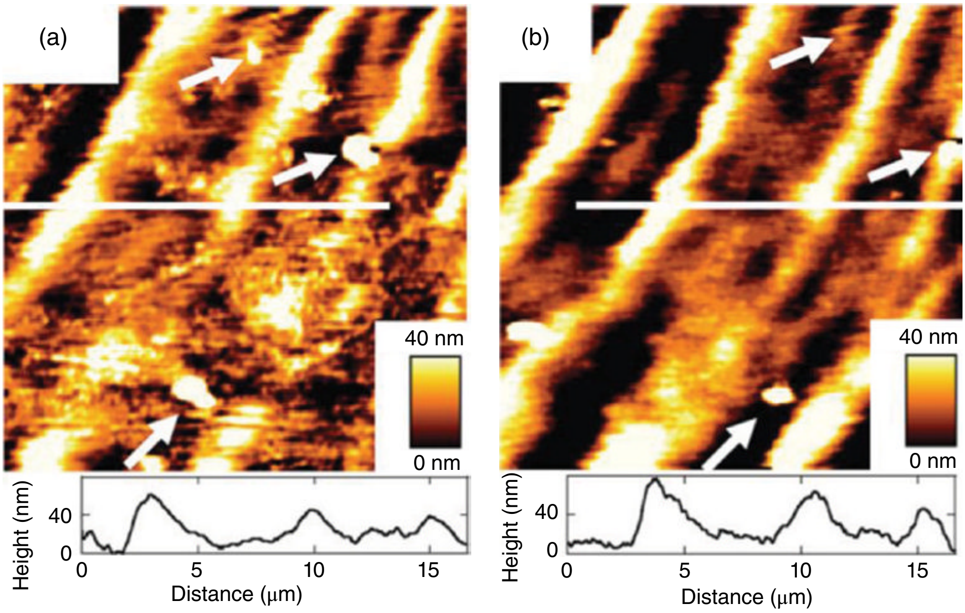


FIGURE 55.3 (a) A stripe-like magnetic domain image with hillocks and magnetic grains on the CoCr film (scan area: 20×20 μm). (b) The same stripe-like magnetic domain image with sporadic hillocks with 4 μm shift to the left from the position of (a) (scan area: 20×20 μm). (insets) Line scans along the white lines in each image for comparison of the contrast variation between two images. Common topographic hillocks in both images are as marked with arrows for comparison. Source: Reprinted with permission from Ref. 24. Copyright (2009), American Institute Physics.

55.3.1.2 Separation of Topography and Magnetic Structures As a method to make the *amplitude increase* in the noncontact region as the tip approaches the surface, an electrostatic force modulation method is introduced to use the capacitive coupling effect for magnetic imaging (inset of Fig. 55.4a). The same rough surface is used for a direct comparison between both methods in separating topographic features from an MFM image. A sinusoidal signal $V_{ac} \cdot \sin(\omega t)$ with $V_{ac} = 10\,\mathrm{V}$ is applied between the same tip and the same surface using a function generator (Hewlett Packard, HP 33120A). The operating frequency (f_{op}) was set to 53 kHz where in-phase 2ω component amplitude, $X_{2\omega}$ ($=R_{2\omega} \cdot \cos \phi_{2\omega}$), has the resonance peak for the cantilever with the resonance frequency of 106 kHz [36]. In Figure 55.4a, the in-phase amplitude increases in the noncontact region as the tip approaches the surface. The enlarged curve (Fig. 55.4b) shows that the horizontal set-amplitude line ($X_{2\omega}=X_{2\omega,SP}$) meets twice with the in-phase amplitude–distance curve, but the signs of the slope $\partial X_{2\omega}/\partial z$ (i.e., the feedback polarity) at the two intersecting points are different from each other. The result shows that only one stable state (in either tapping or noncontact region) is available for a given feedback polarity.

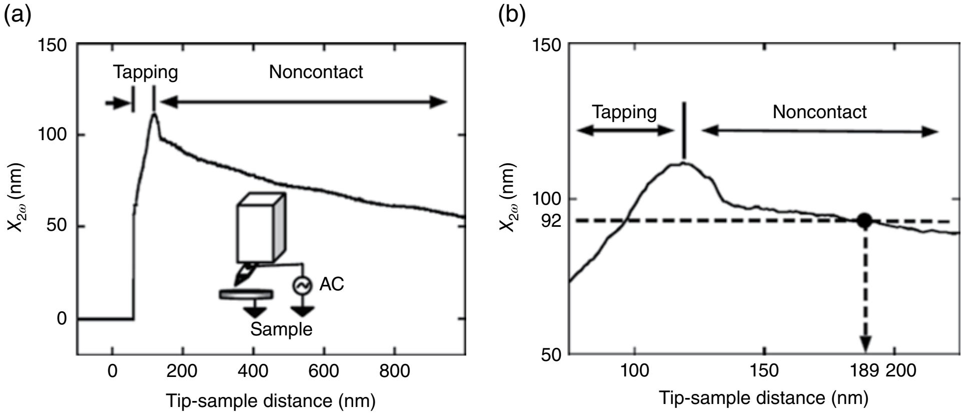


FIGURE 55.4 (a) A typical in-phase amplitude–distance curve at the operating frequency of resonance 53 kHz. (inset) A schematic sketch of the electrostatic force modulation method. (b) An enlarged amplitude–distance curve in the distance range between 75 and 225 nm. A dashed line representing the feedback set amplitude of 92 nm corresponds to the average noncontact distance of 189 nm from the surface. A stable state is marked with a solid circle in the noncontact region. Source: Reprinted with permission from Ref. 24. Copyright (2009), American Institute Physics.

Magnetic imaging of the same CoCr magnetic film was performed under the feedback condition of a constant set amplitude of 92 nm in the noncontact region to investigate the effect of the capacitive coupling on the separation. The imaging condition is similar to the conditions of Figure 55.3b because the noncontact distance of 189 nm is much bigger than the set amplitude of 92 nm, but still smaller than the heights of the hillocks. Figure 55.5a shows a magnetic image of well-connected labyrinthine structures with the periodicity of 4.5–5 μm. The image does not show any evidence of the topographic features. More magnetic images are collected repeatedly in different scan areas from 100 × 100 nm to 40 × 40 μm for the different set values between 82 and 96 nm (not shown). However, neither image shows the topographic features, indicating that there is no mechanical contact between the tip and the surface during MFM data acquisition. The reproducibility of MFM data without the tip crashing toward the surface suggests the enhanced stability of MFM method using the capacitive coupling over the conventional AM MFM. In order to observe the topographic features, the oscillating tip is brought into the tapping region by reversing feedback polarity with the same set amplitude of 92 nm (see Fig. 55.4b). Figure 55.5b shows several hillocks with the diameters of 1000–3000 nm and heights of 200–300 nm and aggregations of small grains with the sizes of 100–300 nm and the heights of 10–40 nm, consistent with those observed in Figure 55.3a and b. The consistency indicates that the charging effect due to the electrostatic modulation is not usually important for general conductive magnetic samples [19].

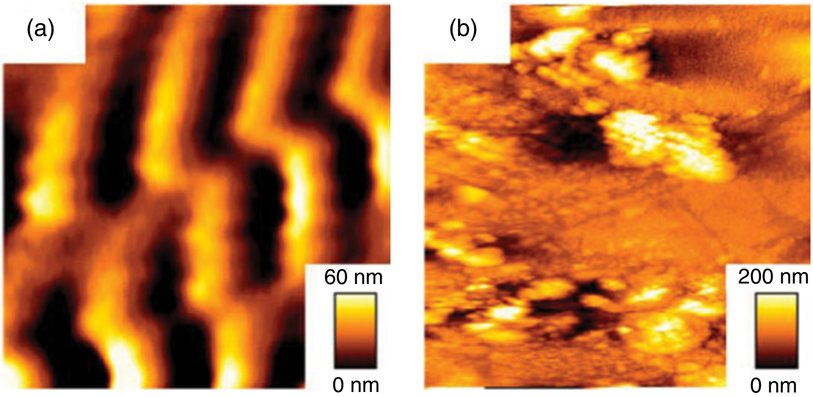


FIGURE 55.5 (a) The stripe-like magnetic domain image in the noncontact regime with the set amplitude of 92 nm (scan area: 20 × 20 μm). (b) Topographic image of CoCr film taken with the electrostatic tapping mode with the set amplitude of 92 nm (scan area: 20 × 20 μm). (insets) Line scans along the white lines in each image for comparison of the contrast variation between two images. Common topographic hillocks in both images are as marked with arrows for comparison. Source: Reprinted with permission from Ref. 24. Copyright (2009), American Institute Physics.

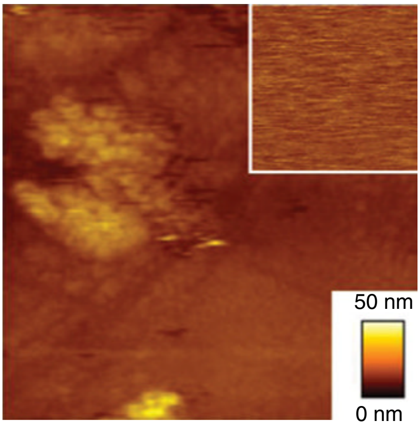


FIGURE 55.6 A topographic image taken by a nonmagnetic tip in the tapping regime with the set amplitude of 92 nm (scan area: 20 × 20 μm). (inset) An image taken by the same nonmagnetic tip in the noncontact regime with the set amplitude of 92 nm (scan area: 20 × 20 μm). Source: Reprinted with permission from Ref. 24. Copyright (2009), American Institute Physics.

An image taken in a tapping regime by a nonmagnetic commercial conductive Si cantilever [36] is shown in Figure 55.6 where the topographic features are similar to those in Figure 55.5b. When an image was taken in a noncontact regime (inset of Fig. 55.6), it was completely featureless without showing any evidence of magnetic features within the current noise level of the system. The result indicates that topographic interactions (e.g., electrostatic force) are too weak to provide observable topographic features. It also suggests that the magnetic interactions should be dominant over the topographic interactions at a noncontact distance around approximately 200 nm where the magnetic image of Figure 55.5a was taken.

55.3.2 Effects of Long-Range Tip–Sample Interaction on Magnetic Force Imaging: A Comparative Study Between Bimorph-Driven System and Electrostatic Force Modulation

In this section, the detailed mechanism of the stability improvement is described through a direct comparison of the traditional bimorph method and the electrostatic force modulation method. Figure 55.7a and b compares the amplitude–frequency curves for both the electrostatic force modulation MFM system and the bimorph-driven MFM system. In the bimorph-driven system (Fig. 55.7a), unwanted peaks around the cantilever resonance result in unstable feedback conditions during MFM imaging. This poor frequency response of the cantilever is due to the additional frequency components integrated near the primary resonance peak, as shown in Figure 55.7a. The existence of several resonance peaks imposes drastic restrictions on the operating frequency range at which useful measurements can be made. In the electrostatic force modulation system (Fig. 55.7b), the cantilever frequency response has two well-defined resonance peaks (ω and 2ω components only). This indicates that the cantilever is the only component being driven by electrostatic force modulation.

Figure 55.8a and b shows maze-like magnetic domain structures with periodicity 3–5 μm. These results are consistent with our previous study [24]. However, topographic features frequently overlap in magnetic images obtained by MFM using bimorph-driven modulation, as shown in Figure 55.8a. The overlapped topographic features appear to be associated with previously observed large hillocks with diameter 5–10 μm and height 100–300 nm [24]. Figure 55.8b shows an MFM image of the striped magnetic domains on the CoCr magnetic film using electrostatic force modulation. We reproduced the same image repeatedly at the same location, indicating the improved stability of the electrostatic force modulation technique.

Figure 55.9a and b illustrates the differences in imaging stability obtained by bimorph-driven and electrostatic MFM systems. Figure 55.9a shows an MFM image of stripe domain structures with periodicity 4.5–5 μm on CoCr magnetic film, consistent with previously reported MFM images [16]. As the bimorph-driven tip scans over the magnetic surface, it appears to crash into a large hillock that is taller than the tip height. The crash of the tip and a hillock structure (as marked by an arrow in the middle of y-scan) creates a feedback disturbance, overcoming the barrier for the imaging mode transition from magnetic mode to topographic mode [30]. After the crash, the tip remains in tapping mode as it continues to scan in the y-direction [30]. The topographic image mode stays until the tip completes scanning, different from previous observations of reversible switches between noncontact and topographic modes [24, 32]. The persistence of topographic imaging mode in the bimorph-driven system indicates that

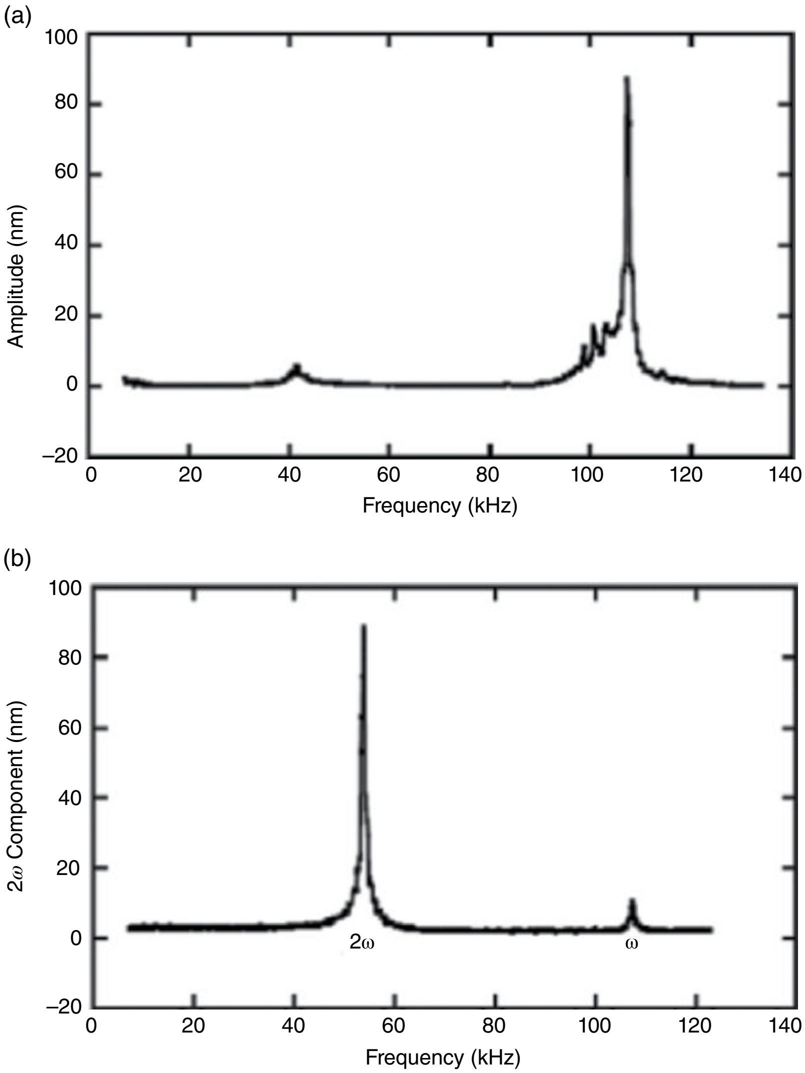


FIGURE 55.7 Amplitude–frequency response for the cantilever. (a) The bimorph-driven system. (b) The electrostatic force modulation system. Source: Reprinted with permission from Ref. 25. Copyright (2012), American Institute Physics.

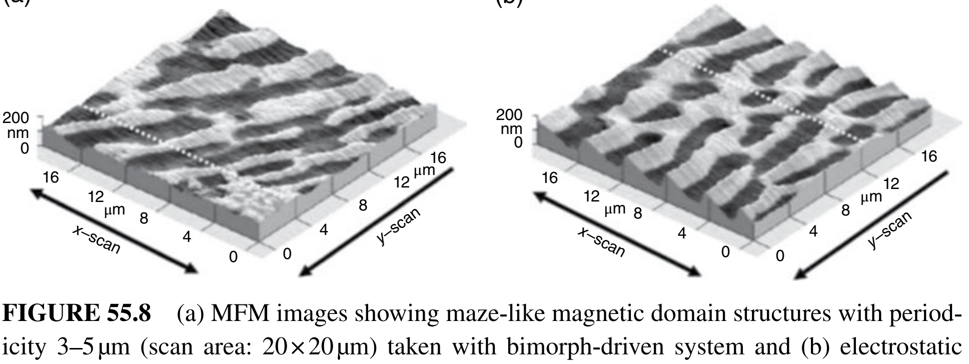


FIGURE 55.8 (a) MFM images showing maze-like magnetic domain structures with periodicity 3–5 μm (scan area: 20 × 20 μm) taken with bimorph-driven system and (b) electrostatic force modulation system. Source: Reprinted with permission from Ref. 25. Copyright (2012), American Institute Physics.

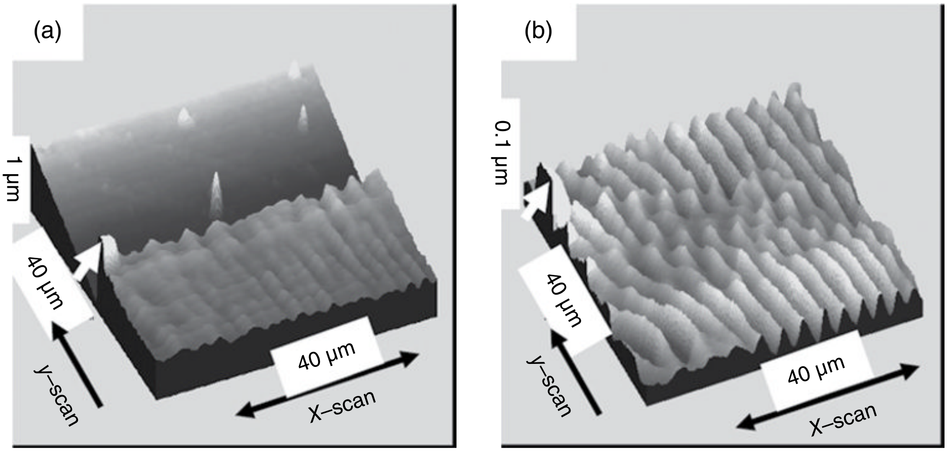


FIGURE 55.9 (a) Transition from magnetic domain imaging mode to topographic imaging mode during tip scanning of CoCr film with set-amplitude 85 nm using bimorph-driven modulation (scan area: 40 × 40 μm). The bimorph-driven system is more susceptible to large topographic features such as the feature indicated by an arrow, which causes the imaging mode to change from magnetic to topographic. (b) Magnetic domain image of CoCr film in the noncontact regime with set-amplitude 95 nm operating at the resonance frequency 53 kHz using the electrostatic force modulation system (scan area: 40 × 40 μm). The system is robust against frequent tip collisions with topographic hillocks such as the feature indicated by an arrow. Source: Reprinted with permission from Ref. 25. Copyright (2012), American Institute Physics.

there is a hysteresis in switching between tapping mode and the noncontact magnetic mode. This irreversible process indicates that there might be a change in the energy barrier between the noncontact and tapping regions in the amplitude–distance curve. The disappearance of noncontact mode in Figure 55.9a can possibly be explained by a change in the tip structure: a tip radius decrease during the crash causes the noncontact region of the amplitude–distance curve to move upward due to a capacitance decrease. This upward move eliminates the intersection point between the horizontal set amplitude and the amplitude–distance curve in the noncontact region of Figure 55.10b [24].

Figure 55.9b shows a magnetic image from the same sample surface, this time using electrostatic force modulation. As the electrostatically driven tip scans over the magnetic surface, it again encounters the large hillock that is taller than the tip height. After the encounter, the tip stays in noncontact mode as it scans in the *y*-direction. The persistence of noncontact mode in the electrostatically driven system can possibly be understood by the opposite polarity of slope of the amplitude–distance curve in the tapping and noncontact regions, respectively, as shown in Figure 55.10a [24]. In addition, the completeness of the magnetic image beyond the encounter indicates that the enhanced barrier height in amplitude (i.e., the difference between peak height and the horizontal set amplitude) of the electrostatic force modulation system (over the

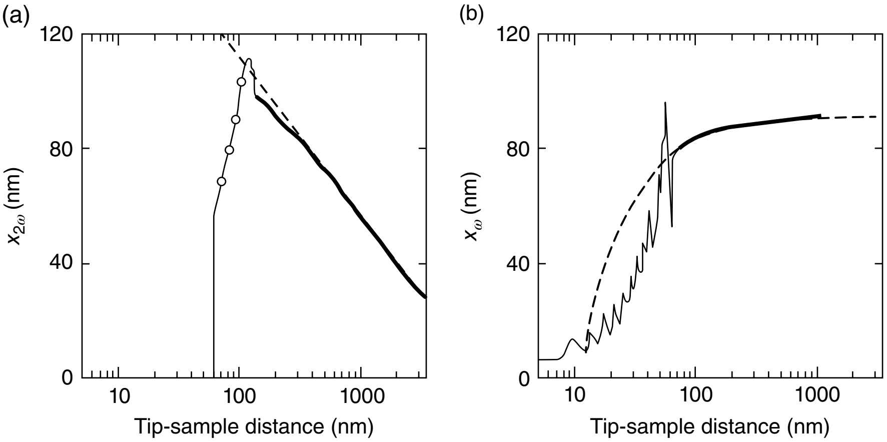


FIGURE 55.10 (a) Electrostatic force modulation system: Solid curve fits the amplitude–distance data. Note the barrier between the noncontact and tapping regions and the opposite feedback polarity (sign of slope) in each region. Dashed line fits the data to the logarithmic function $b \cdot \log(D/z)$ over the entire noncontact region from 200 nm to 4 μm with the fitting parameters $b = 1.9\,\mathrm{nN/V^2}$, $D = 1$ μm. (b) Amplitude–distance data for the bimorph system. Note the identical feedback polarity in the noncontact and tapping regions. Source: Reprinted with permission from Ref. 25. Copyright (2012), American Institute Physics.

bimorph system) prevents the cantilever from snapping upon contact with the surface. Since there are no appreciable height changes in the magnetic image, the tip appears to be more secure under the electrostatically modulated system.

In the bimorph-driven system, feedback bistability between the noncontact and tapping regions in the amplitude–distance curve frequently causes overlap of the topographic signal onto the magnetic signal during scanning. Figure 55.10a shows the amplitude change as the tip approaches the sample surface in the electrostatic force modulation system. The amplitude increases in the noncontact region but decreases in the tapping region. This behavior is very different from that obtained with the bimorph-driven system, as shown in Figure 55.10b. In the bimorph-driven system, the feedback polarity is the same in both the noncontact and tapping regimes.

In the usual constant force gradient mode with bimorph-driven modulation, a dc voltage is applied between the tip and the sample to prevent the tip from crashing into the surface during scanning. Instead, we employed the electrostatic force modulation technique for self-actuation of the cantilever to satisfy the requirements for MFM imaging. For the direct force modulation in a novel MFM, we utilized capacitive force coupling between the cantilever and the sample to induce the modulation (Fig. 55.1b). To avoid mixing of the topographic signal with the MFM image, we must increase the operating distance of the noncontact region to avoid the tip crashing near the tapping region. The opposite feedback polarity between noncontact mode and tapping mode is necessary to exclude possible mixing of topographic signal due to the tapping mode operation during noncontact MFM imaging. Figure 55.1b shows the electrostatic force modulation as a servo force to keep the polarity of feedback constant for both repulsive and attractive magnetic forces. The electrostatic force keeps the total force gradient always positive regardless of changes in the polarity of the magnetic force. The decrease in amplitude results from the sudden involvement of the repulsive tapping force. The peak in Figure 55.10a represents a transition from an electrostatic noncontact interaction to tapping interaction between the tip and the sample. The linear part in the tapping regime is used to find the amplitude of the tip vibration in the noncontact regime because the z-piezo is precisely calibrated [32, 37, 38].

However, in the bimorph-driven system, the linear tapping region also has the same sign of slope as the noncontact region [24]. This indicates that there exist two *stable* states for a given set-amplitude ($X_{\omega,\mathrm{SP}}$): one in the tapping region and the other in the noncontact region. The coexistence of two stable states is analogous to that in the AFM [32, 39]. Garcia and San Paulo associated continual switching of the oscillating tip between the two stable states with abrupt changes in height of topographic features. Similarly, we attribute the appearance of topographic features in a magnetic image to the switching between the bistable states.

If this switching mechanism is the major channel of picking up topographic features in a magnetic image, the pickup ratio should depend upon the average distance of the noncontact state from the sample surface. In order to check this idea, we performed the magnetic imaging as a function of feedback set point amplitude $X_{\omega,\mathrm{SP}}$ at the same tip location repeatedly. Using electrostatic force modulation, tip crashing toward the surface upon encountering the peak barrier in Figure 55.10a rarely happens during repeated magnetic imaging. Hong et al. reported the same type of stable imaging condition in tapping mode AFM using the electrostatic force modulation [40].

The equation of motion of the cantilever probe is employed to reveal the stability of the mechanism: the probe with a single-point mass m undergoes 1D forced harmonic oscillation along the vertical z axis in Figure 55.1a and b for bimorph-driven modulation and electrostatic force modulation, respectively. This oscillation is described by the following equation:

$$\frac{\partial^2 z}{\partial t^2}+\frac{\omega_0}{Q}\cdot\frac{\partial z}{\partial t}+\omega_0^2\cdot\left(z-u\right)=\frac{1}{m}\left(F_{\mathrm{m}}+F_{\mathrm{C}}+F_{\mathrm{VdW}}\right) \tag{55.8}$$

where z is the coordinate perpendicular to the sample surface, t is time, u is the undeflected cantilever position, k is the force constant, ω_0 is the resonance angular frequency, Q is the quality factor, and ω_0/Q corresponds to the damping constant per unit mass in the presence of magnetic force F_{m}, capacitance force F_{C}, and van der Waals force F_{VdW}. Equation 55.8 can be solved analytically when the average tip–sample distance u is much bigger than the set-amplitude $X_{\omega,\mathrm{SP}}$, for both bimorph-driven and electrostatic force modulation systems. The position u in the equation of motion (55.8) is constant while ac voltage is applied.

For electrostatic force modulation, the square law dependence of the capacitive force F_C on the driving signal $V_{ac}\cos(\omega t)$ induces a mechanical vibration of the cantilever:

$$F_C = \frac{1}{4}\frac{\partial C}{\partial z} V_{ac}^2 \left(1 + \cos 2\omega t\right) \tag{55.9}$$

where C is the capacitance between the tip and the sample. The in-phase amplitude of the 2ω component, $X_{2\omega}$, is predicted as follows:

$$X_{2\omega} = \frac{1}{2m} \cdot \frac{-F_C\left(z, V_{ac}\right) \cdot 2\left(\omega_0'^2 - 4\omega^2\right)}{\sqrt{\left(\omega_0'^2 - 4\omega^2\right)^2 + \left(2\omega\omega_0/Q\right)^2}} \sin\left(\phi_{2\omega} + \phi_0\right) \tag{55.10}$$

where ϕ_0 is 108° that maximizes the $X_{2\omega}$ near $u = 100\,\text{nm}$.

$$\tan\phi_{2\omega} = \frac{2\omega\omega_0}{Q \cdot \left(\omega_0'^2 - 4\omega^2\right)} \tag{55.11}$$

where $\omega_0'^2 = \omega_0^2\left(1 - \frac{1}{k}\left(F_m' + \frac{1}{2} \cdot F_C'\left(z, V_{ac}\right)\right)\right)$.

$$F_C\left(z, V_{ac}\right) = \frac{1}{2}\frac{\partial C}{\partial z} V_{ac}^2 \tag{55.12}$$

$$F_C'\left(z, V_{ac}\right) = \frac{1}{2}\frac{\partial^2 C}{\partial z^2} V_{ac}^2 \tag{55.13}$$

The Lorentzian Equation 55.10 accounts for the behavior of $X_{2\omega}$ in Figure 55.7b. The earlier predicted solutions, Equations 55.10 and 55.13, are compared with the experimental data to find the dependence of the interaction with distance in Figure 55.10a and b. The amplitude increases in the noncontact region using electrostatic force modulation since the change of the electrostatic capacitance force $-F_C$ (z, V_{ac}) in Equation 55.10 increases as the distance z becomes smaller. Experimental data in Figure 55.10a shows that the electrostatic interaction extends to 3 μm (equivalent to the limit of our z-piezo). The solid line of the prediction of Equation 55.10 matches well with the experimental data. In the prediction, the electrostatic capacitance force $-F_C$ (z, V_{ac}) is described by assuming a conical tip. In this case, the capacitance gradient in Equation 55.12 is known to vary with a single logarithmic function $\partial C/\partial z = b \cdot \log(D/z)$, where the constants b and D depend on the angle and length of the cone [41, 42]. They are determined to be 1.9 nN/V^2 and 1000 nm, respectively, from the predictive amplitude curve (solid line). The amplitude data curve (Fig. 55.9a) in the distance range 200–4000 nm fits with the predictive curve that assumes the logarithmic function (i.e., conical tip shape). The fitting result suggests that the amplitude depends logarithmically on the tip–sample distance z. Due to the logarithmic long-range electrostatic interaction, good feedback stability was achieved during the image acquisition of magnetic domains in the electrostatic force-driven MFM.

In case of the bimorph-driven system, the equation of motion (55.8) has $F_\mathrm{C} = \dfrac{1}{2}\dfrac{\partial C}{\partial z}V_\mathrm{dc}^2$. The bimorph-driven modulation can be described as follows:

$$u = u_0 + a \cdot e^{i\omega t} \tag{55.14}$$

where u_0 is the midpoint of the undeflected cantilever position, a is the driving amplitude, and ω is the angular frequency. For bimorph-driven modulation, the in-phase component is predicted as follows:

$$X_\omega = \frac{a \cdot \omega_0^2}{Q} \cdot \frac{a \cdot \omega_0^2 \cdot \left(\omega_0'^2 - \omega^2\right)}{\left(\omega_0'^2 - \omega^2\right)^2 + \left(\omega\omega_0/Q\right)^2} \tag{55.15}$$

where $\omega_0'^2 = \omega_0^2\left(1 - \dfrac{1}{k}\left(F_\mathrm{m}' + F_\mathrm{C}'\left(z, V_\mathrm{dc}\right)\right)\right)$ and

$$F_\mathrm{C}'\left(z, V_\mathrm{dc}\right) = \frac{1}{2}\frac{\partial^2 C}{\partial z^2}V_\mathrm{dc}^2 \tag{55.16}$$

Due to the long-range nature of the electrostatic capacitance force $-F_\mathrm{C}$ (z, V_ac), the amplitude varies significantly for a much longer range (up to 3000 nm) in the noncontact region when electrostatic force modulation is used.

The concavity of the approaching curve is down since the amplitude in the noncontact region is determined by the change of capacitance with the tip–sample distance, $\partial C/\partial z$, in Equation 55.10; we fitted the curve in Figure 55.10a, assuming the conical tip model as a good approximation in the noncontact regime [35, 41]. To support this argument, the $X_{2\omega}$ versus distance curve was fitted with a single logarithmic function representing capacitance coupling between a conical tip and a flat sample surface [40]. Figure 55.10a shows an excellent agreement of the logarithmic function with the curve up to the distance 4000 nm that the current z-piezo tube can maximally extend from the surface, indicating that the conical tip model is a good approximation in the noncontact region. The contrast mechanism during the image acquisition of magnetic domain in the electrostatic force modulation results from the long-range electrostatic servo interaction that depends on the tip–sample distance z logarithmically. Due to the long-range nature of capacitance coupling in the electrostatic interaction, the noncontact region is much greater for electrostatic force modulation than for bimorph-driven modulation in Figure 55.10b. Therefore, the electrostatic force F_C(z, V_ac) with logarithmic z dependence is the origin of the long-range servo interaction rather than the electrostatic force gradient

$$F_\mathrm{C}'\left(z, V_\mathrm{dc}\right) \propto \frac{1}{z},$$

commonly used as the servo interaction in bimorph-driven modulation [43]. Figure 55.10b also shows that the 1/z function represents nicely over the noncontact region in the in-phase X_ω–distance curve of the bimorph-driven system. Due to the long-range electrostatic interaction, good feedback stability during the imaging of magnetic domain can be achieved in the electrostatic force-driven MFM up to the operation distance of a few hundred nanometers.

As the tip approaches the surface, $\partial C/\partial z$ in the numerator $F_C(z, V_{ac})$ contributes more to the $X_{2\omega}$ amplitude than the denominator due to the change of $F_C'(z, V_{ac})$ with z. In the bimorph-driven system, on the other hand, the in-phase vibration amplitude X_ω is controlled by the change of $F_C'(z, V_{dc})$ due to the absence of the factor $\partial C/\partial z$ in the numerator. When the tip experiences the change of magnetic force $\Delta F_m'$ during a lateral move across a domain wall, both the electrostatic force $F_C(z, V_{ac})$ and the electrostatic force gradient $F_C'(z, V_{ac})$ compensate for a given change in magnetic force gradient by moving the tip vertically relative to the sample surface to maintain the $X_{2\omega}$ constant under feedback. In a bimorph-driven MFM, only an electric force gradient $F_C'(z, V_{dc})$ plays this role. The bigger amplitude in $X_{2\omega}$ at a closer distance in the noncontact region indicates that the in-phase amplitude of the 2ω component is controlled more by $F_C(z, V_{ac})$ than by $F_C'(z, V_{ac})$. The change of magnetic force gradient ($\Delta F_m'$) is compensated by the factor $F_C(z, V_{ac})$ to maintain the in-phase amplitude constant under feedback through the movement of the piezo tube. In the MFM using bimorph-driven modulation, the change of magnetic force gradient ($\Delta F_m'$) is simply compensated by the change of the servo force gradient ($\Delta F_C'$) [16, 35, 44]. The shorter range indicates that due to the collision, the excited tip can easily overcome the crossover barrier.

Based on the understanding of the interactions of the tip and the sample surface of Figure 55.10a and b, the separation mechanism is described through the comparison between the bimorph-driven system and electrostatic force modulation system in Figure 55.11a and b. In Figure 55.11a, topographic features (shaded profile) and the magnetic features (solid profile line extracted from the dotted line in Fig. 55.8a) appear together across the vertical dashed line due to the involvement of the short-range interaction during data acquisition. This is because the tip is likely to crash with tall topographic features during the oscillatory motion of the cantilever, as illustrated on the right side of Figure 55.11a. Such a coexistence of topographic and magnetic structures on an MFM image is known to depend on the distance between the tip and the surface and oscillatory amplitude [24]. In the electrostatic force modulation system, the long-range nature of the interaction keeps the tip from being crashed into the topographic features, as shown in Figure 55.11b. Due to such long-range of interactions, the MFM image has only magnetic structures without having any topographic features as found in Figure 55.8b. This is because the oscillatory amplitude is smaller than the tip–sample distance as depicted in Figure 55.11b.

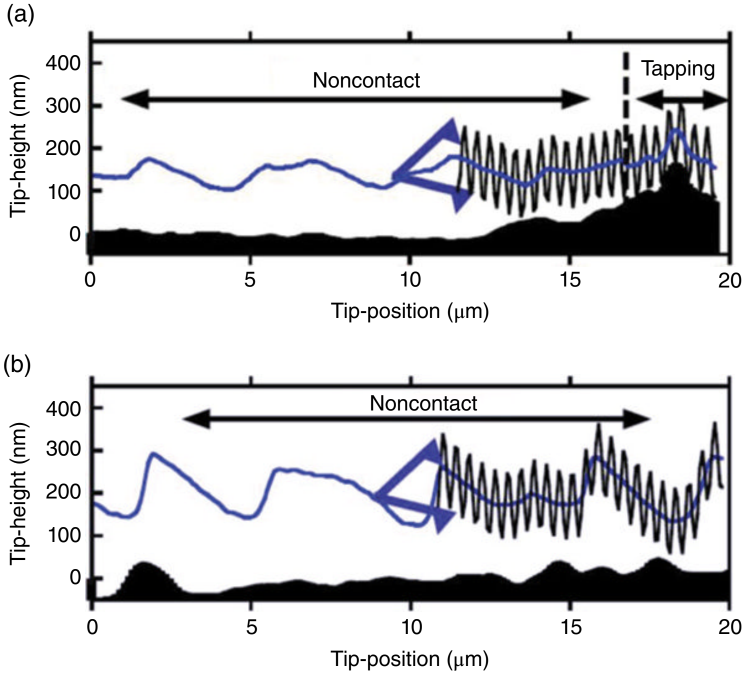


FIGURE 55.11 A comparison between the bimorph-driven system (a) and electrostatic force modulation system (b) during the magnetic force imaging acquisition. The shaded profiles represent the topographic features, whereas the solid lines represent the sectional profiles in the magnetic force images. The oscillatory lines depict the motion of the cantilever with magnetic probes. Source: Reprinted with permission from Ref. 25. Copyright (2012), American Institute Physics.

55.4 CONCLUSION

For improvement of the MFM resolution and study of the relation between topography and magnetic structure, it is important to separate the topographic and magnetic features of the sample. One way to do this is to use modulated electrostatic bias field between the tip and the sample and then the information of modulated electrostatic signal as feedback input.

Both nonmagnetic images again support the separation of the topographic features from the magnetic images through the removal of one stable state using the capacitive coupling. The novel approach presented in this chapter should have a dramatic impact on not only on the MFM field, but also on other scanning probe microscopy fields such as electrostatic force microscopy and Kelvin probe microscopy. Furthermore, the technique is expected to allow for less invasive observation of superconducting vortex structures and soft magnetic structures by avoiding the sudden transition from the noncontact state to the tapping state in the conventional bimorph-driven system, thus leading to a better understanding of the relationship between magnetic structures and topographic pinning sites such as grain boundaries.

We have developed MFM using an electrostatic force modulation that shows excellent separation of the magnetic signal from the topographic signal. Compared to the bimorph-driven system, the observed magnetic images do not show any topographic features, clearly indicating the separation of topographic and magnetic signals in the noncontact region. We attribute this separation to the opposite feedback polarity in the noncontact region to the one in tapping mode for topographic imaging, thus preventing the magnetic signal from mixing with the topographic signal under feedback condition for the constant amplitude. The origin of the feedback polarity difference is discussed with the electrostatic capacitive coupling and is compared with the bimorph-driven system. This enhanced stability may come from differences between the two systems in terms of the servo forces and modulation method. We attribute the higher stability of MFM using electrostatic force modulation (instead of bimorph-driven modulation) to the long-range electrostatic interaction between the tip and the sample surface.

We also report the successful application of this system for imaging the magnetic structure and topography of CoCr thin film surfaces through separation of the magnetic and topographic structures, with enhanced stability. The system has been tested on the magnetic domain structures of CoCr magnetic film, with a periodicity of 4.5–5 μm. The system has good stability due to the long-range electrostatic capacitance force and also the well-defined single resonance peak in the frequency response. This system will be a promising tool for studying magnetic and topographic structures on the magnetic sample surfaces. In addition to the magnetic imaging, the system can be easily expanded to image the electronic potential structures using ω component of the output signal due to the tip–sample interaction. The simultaneous measurement of magnetic structure and the charge–potential distribution would be especially useful for studying the relationship between magnetic structures and electronic structures in semiconducting magnetic materials for future spintronics applications.

REFERENCES

1. M. S. Valera, A. N. Farley, S. R. Hoon, L. Zhou, S. McVitie, and J. N. Chapman, *Appl. Phys. Lett.* 67, 2566 (1995).
2. G. Bochi, H. J. Hug, D. I. Paul, B. Stiefel, A. Moser, I. Parashikov, H.-J. Güntherodt, and R. C. O'Handley, *Phys. Rev. Lett.* 75, 1839 (1995).
3. M. Lederman, S. Schulz, and M. Ozaki, *Phys. Rev. Lett.* 73, 1986 (1994).
4. R. Proksch, T. E. Schaffer, B. M. Moskowitz, E. D. Dahlberg, D. A. Bazylinski, and R. B. Frankel, *Appl. Phys. Lett.* 66, 2582 (1995).
5. M. Löhndorf, A. Wadas, R. Wiesendanger, and H. W. van Kesteren, *J. Vac. Sci. Technol. B* 14, 1214 (1996).

6. A. Moser, H. J. Hug, I. Parashikov, B. Stiefel, O. Fritz, H. Thomas, A. Baratoff, and H.-J. Güntherodt, *Phys. Rev. Lett.* 74, 1850 (1995).
7. C. W. Yuan, Z. Zheng, A. L. de Lozanne, M. Tortonese, D. A. Rudman, and J. N. Eckstein, *J. Vac. Sci. Technol. B* 14, 1210 (1996).
8. Y. Martin and H. K. Wickramasinghe, *Appl. Phys. Lett.* 50, 1455 (1987).
9. M. R. Freeman and B. C. Choi, *Science* 294, 1484 (2001).
10. U. Kaiser, A. Schwarz, and R. Wiesendanger, *Nature* 446, 522 (2007).
11. C. C. de Souza Silva, A. V. Silhanek, J. Van de Vondel, W. Gillijns, V. Metlushko, B. Ilic, and V. V. Moshchalkov, *Phys. Rev. Lett.* 98, 117005 (2007).
12. T. Shinjo, T. Okuno, R. Hassdorf, K. Shigeto, and T. Ono, *Science* 289, 930 (2000).
13. J. Philip, A. Punnoose, B. I. Kim, K. M. Reddy, S. Layne, J. O. Holmes, B. Satpati, P. R. Leclair, T. S. Santos, and J. S. Moodera, *Nat. Mater.* 5, 298–304 (2006).
14. E. W. J. Straver, J. E. Hoffman, O. M. Auslaender, D. Rugar, and K. A. Moler, *Appl. Phys. Lett.* 93, 172514 (2008).
15. S. Porthun, L. Abelmann, and C. Lodder, *J. Magn. Magn. Mater.* 182, 238 (1998).
16. H.-J. Güntherodt and R. Wiesendanger, eds., *Scanning Tunneling Microscopy I-II* (Springer-Verlag, Berlin, 1992).
17. (a) X. Zhu, P. Grütter, V. Metlushko, and B. Ilic, *Phys. Rev. B* 66, 024423 (2002); (b) X. B. Zhu and P. Grutter, *IEEE Trans. Magn.* 39, 3420 (2003).
18. B. I. Kim, J. W. Hong, J. I. Kye, and Z. G. Khim, *J. Korean. Phys. Soc.* 31, s79 (1997).
19. C. Schönenberger and S. F. Alvarado, *Z. Phys. B* 80, 373 (1990).
20. C. Schönenberger, S. F. Alvarado, S. E. Lambert, and I. L. Sanders, *J. Appl. Phys.* 67, 7278 (1990).
21. R. Giles, J. P. Cleveland, S. Manne, P. K. Hansma, B. Drake, P. Malvald, C. Boles, J. Gurley, and V. Elings, *Appl. Phys. Lett.* 63, 617 (1993).
22. E. Meyer, H. Heinzelmann, P. Grütter, Th. Jung, Th. Weisskopf, H.-R. Hidber, R. Lapka, H. Rudin, and H.-J. Güntherodt, *J. Microsc.* 269, 152 (1988).
23. P. Grütter, H. J. Mamin, and D. Rugar, *Scanning Tunnelling Microscopy II* (Springer, Berlin, 1992), pp. 151–207.
24. B. I. Kim, *Rev. Sci. Instrum.* 80, 023702 (2009).
25. B. I. Kim, *J. Appl. Phys.* 111, 104313 (2012).
26. Bruker AFM Probes, Camarillo, CA, http://www.brukerafmprobes.com (accessed December 19, 2015).
27. H. J. Mamin, D. Rugar, J. E. Stern, B. D. Terris, and S. E. Lambert, *Appl. Phys. Lett.* 53, 1563 (1988).
28. D. Sarid, *Scanning Force Microscopy* (Oxford University, New York, 1991), p. 31.
29. R. M. White and T. H. Geballe, *Long Range Order in Solids* (Academic, New York, 1979), Chapter VIII.
30. Q. Zhong, D. Inniss, K Kjoller, and V. B. Ellings, *Surf. Sci. Lett.* 290, L688 (1993).
31. F. Perez-Murano, G. Abadal, N. Barniol, X. Aymerich, J. Servat, P. Gorostiza, and F. Santz, *J. Appl. Phys.* 78, 6797 (1995).

32. R. Garcia and A. San Paulo, *Phys. Rev. B* 61, R13381 (2000).
33. S. A. Joyce and J. E. Houston, *Rev. Sci. Instrum.* 62, 710 (1991).
34. J. R. Bonander and B. I. Kim, *Appl. Phys. Lett.* 92, 103124 (2008).
35. Y. Martin, D. Rugar, and H. K. Wickramasinghe, *Appl. Phys. Lett.* 52, 244 (1988).
36. B. I. Kim, U. H. Pi, Z. G. Khim, and S. Yoon, *Appl. Phys. A* 66, s95 (1998).
37. B. Anczykowski, D. Kruger, and H. Fuchs, *Phys. Rev. B* 53, 15485 (1996).
38. A. Kuhle, A. H. Sorensen, J. B. Zandbergen, and J. Bohr, *Appl. Phys. A* 66, S329 (1998).
39. B. N. J. Persson, *Sliding Friction: Physical Principles and Applications*, 2nd ed. (Springer, Heidelberg, 2000), pp. 54–77.
40. J. W. Hong, Z. G. Khim, A. S. Hou, and S. I. Park, *Appl. Phys. Lett.* 69, 2831 (1996).
41. H. Yokoyama, T. Inoue, and J. Itoh, *Appl. Phys. Lett.* 65, 3143 (1994).
42. S. Belaidi, P. Girard, and G. Leveque, *J. Appl. Phys.* 81, 1023 (1997).
43. A. S. Hou, Ultrafast electric force microscope for probing integrated circuits, PhD dissertation, Stanford University, Palo Alto, CA (1995).
44. A. Wadas, P. Grütter, and H.-J. Güntherodt, *J. Appl. Phys.* 67, 3462 (1990).

56

SCANNING TUNNELING MICROSCOPY

Kwok-Wai Ng

Department of Physics and Astronomy, University of Kentucky, Lexington, KY, USA

56.1 INTRODUCTION

Scanning tunneling microscope (STM) was first invented by Binnig and Rohrer in the early 1980s to study surface reconstruction on silicon surface [1]. This was the first time that atoms were imaged in real space. Besides its unsurpassed resolution, STM also has the convenience that imaging can be performed at ambient atmosphere or even chemical solution. Comparing to scanning electron microscope, the cost of STM can be significantly lower since there is no expensive electron optics in STM, and high vacuum equipment is not required in many STM applications. We should note that STM is a surface probe and its operation highly depends on the cleanliness and purity of the surface. A heavily oxidized surface will make imaging impossible or produce data that is hard to interpret. For this reason, ultrahigh vacuum environment is quite common for STM—especially as a high-resolution surface study tool. This will unavoidably increase the cost of the setup.

Quantum mechanically it is possible for an electron to coexist in two conductors separated by a small gap. This phenomenon is called quantum tunneling. When a voltage bias is applied between the two conductors, a tunneling current can be established along the direction of the bias. Obviously if we look at the electron as a wave, the separation between the conductors has to be comparable to the lattice parameter for tunneling to occur. From this we can expect the magnitude of the tunneling current depends sensitively on the separation between the conductors. In STM, this tunneling current is measured and used to deduce the distance between a conducting

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

sharp tip and the conducting sample surface. If we taste the tip on the surface and measure and record the distance at the grid points, a topographical image can then be formed. One can see that the hardware of an STM is extremely simple. It basically comprises of a battery and some sensitive electronics to measure small currents. Most STM nowadays can measure a current as small as a few pA. We can find many cases in the Internet that even a hobbyist can build an STM at home with a very low budget. This performance per price ratio makes STM a very powerful research tool.

STM has one major shortcoming in requiring the sample under study to be conducting or at least semiconducting. To overcome this shortcoming, since the inception of STM in the 1980s, many sensing techniques other than tunneling current have been introduced. The atomic force microscope (AFM) is a good example [2]. In an AFM the deflection of a cantilever is measured either with a laser beam [3] or shift in resonance frequency [4] as the cantilever scans along the surface. Since no current is involved, even the surface of an insulator (I) can be imaged. AFM is now becoming very common and can be found in many laboratories. AFM is most conveniently used to image larger features, and it is more difficult to achieve ultrahigh resolution like the STM. Many other probing techniques like the sensing of magnetic force (magnetic force microscope (MFM)), magnetic field (scanning Hall probe), and electromagnetic radiation (scanning near-field optic microscope (SNOM or NSOM)) have been used to image different surface properties accordingly. Besides the different sensing methods, the control and instrumentation of these microscopes are quite similar. This type of microscopy technique is in general known as scanning probe microscopy. In this review we will focus on the STM.

56.2 THEORY OF OPERATION

In quantum mechanics a particle is described by a wave function $\psi(x)$ satisfying the time-independent Schrödinger equation:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x)+V(x)\psi(x)=E\psi(x)$$

where $V(x)$ is the potential and E is the energy of the particle. The probability of finding the particle in the range $[a, b]$ is given by

$$P(a \le x \le b)=\int_a^b \psi^* \psi \, dx$$

where ψ^* is the complex conjugate of ψ. This probability allows the particle to be found even in the classical forbidden region and leads to the tunneling phenomenon. Consider

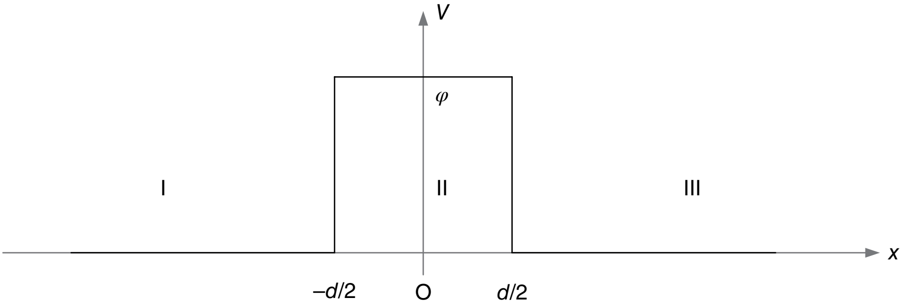


FIGURE 56.1 A square potential area of width *d* and height *φ*. In the discussion here the energy of the electron is less than *φ*, so the barrier is classically forbidden.

a simple square potential in Figure 56.1; if a particle entering from the left with $E < V_0$, classically, it will be bounced back by the potential barrier.

Quantum mechanically, solving Schrödinger equation for this potential will give the wave function of the particle as

$$\psi_{\mathrm{I}} = e^{ikx} + Re^{-ikx}$$
$$\psi_{\mathrm{II}} = Ae^{\kappa x} + Be^{-\kappa x}$$
$$\psi_{\mathrm{III}} = Te^{ikx}$$
$$\text{where } k = \frac{\sqrt{2mE}}{\hbar} \quad \text{and} \quad \kappa = \frac{\sqrt{2m(\phi - E)}}{\hbar}$$

The transmission coefficient is important here, and it can be gotten by the continuity conditions in ψ and $d\psi/dx$ between the wave functions at $x=-d/2$ and $x=d/2$:

$$|T|^2 = \frac{(2k\kappa)^2}{(k^2-\kappa^2)^2 \sinh^2 \kappa d + (2k\kappa)^2 \cosh^2 \kappa d}$$

The fact that ψ_{III} is not zero means there is a certain probability for the particle to tunnel through the barrier, giving rise to a probability current

$$j = \frac{i\hbar}{2m}\left(\psi \frac{\partial \psi^*}{\partial x} - \psi^* \frac{\partial \psi}{\partial x}\right) = \frac{\hbar k}{m}|T|^2$$

In low tunneling rate limit $\kappa d \gg 1$, $\sinh \kappa d \sim \cosh \kappa d \sim e^{\kappa d}$ and $|T|^2 \sim (2k\kappa/(k^2+\kappa^2))^2 e^{-2\kappa d}$. While k and κ are mostly constant, $I \propto e^{-2\kappa d}$, the current is decaying exponentially with the barrier thickness d. A slight increase in d will produce a sharp drop in the current. For a reasonable tunneling rate, we require $2\kappa d \sim 1$. Roughly speaking, this means the barrier thickness d should be the order of the wavelength of the particle given by $\lambda=2\pi/k$. The wavelength of a 1 eV electron is about 1.23 nm. If the barrier height ϕ is not a constant across the barrier, we can conveniently replace it with the average height because of the exponential dependence:

$$\kappa=\frac{\sqrt{2m\left(\bar{\phi}-E\right)}}{\hbar}$$

The preceding discussion can be applied to the real situation when two metals are separated by a vacuum gap (the barrier) of distance z. The barrier height will now become the average work potential of the metals at two sides. All conducting electrons of different energies at both sides will involve in the tunneling process.

For simplicity we will ignore thermal excitation and assume the Fermi–Dirac distribution function is a step function. Only the electrons in the energy range $0 \le E \le |\mathrm{eV}|$ at the left side in Figure 56.2 can tunnel because only they can find unoccupied states at the opposite side for tunneling to occur. If $\rho_{\mathrm{L}}(E)$ and $\rho_{\mathrm{R}}(E)$ are the electron density of states of the materials at the left- and right-hand sides in the figure, the tunnel current will be given as

$$I\sim\int_0^{\mathrm{eV}} e\left(\frac{2k\kappa}{k^2+\kappa^2}\right)^2 e^{-2\kappa z}\cdot\rho_{\mathrm{L}}\left(E\right)\rho_{\mathrm{R}}\left(E\right)dE$$

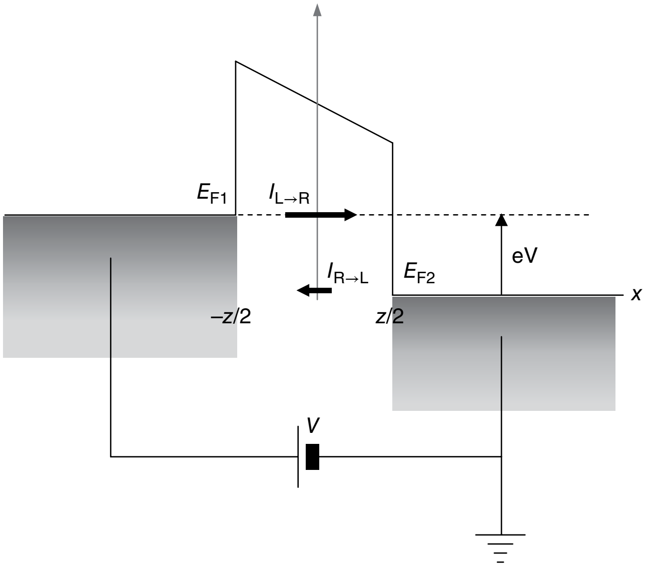


FIGURE 56.2 The Fermi level at one side of the barrier is raised by the applied potential V. This will tip the balance and create a net tunnel current through the barrier.

Note that the current is a convolution of the occupied and unoccupied density of states at both sides. Useful result can be derived if we assume the density of state of one material is more constant than the other, say, $\rho_R(E)=\rho_R(0)=$constant. The earlier equation can now be reduced to

$$I \sim e\rho_R(0)\int_0^{eV}\left(\frac{2k\kappa}{k^2+\kappa^2}\right)^2 e^{-2\kappa z}\cdot\rho_L(E)dE$$

Differentiating I with respect to d,

$$\frac{d}{dz}\ln I = \frac{1}{I}\frac{dI}{dz} = -2\kappa = -\frac{\sqrt{8m(\bar{\phi}-E)}}{\hbar}$$

From this equation, we can see how an STM can be used to determine the average work function of the two metals by a dI/dz measurement, where z is the sample–tip separation.

Besides the work function, STM can also be used to measure the density of state of a material. We should point out that the transmission coefficient is not as well understood and well predicted as described previously. This is especially true in the case of STM, when the wave function is not really a plane wave [5] and it is highly sensitive to the surface condition. For this reason, we will replace the transmission coefficient with an arbitrary function $M(E, z)$:

$$I \sim e\rho_R(0)\int_0^{eV}|M(E,z)|^2\cdot\rho_L(E)dE$$

and

$$\frac{dI}{dV} \sim e^2\rho_R(0)|M(eV,z)|^2\cdot\rho_L(eV)$$

If the variation of $M(E, z)$ is not as strong as $\rho_L(E)$ and it can be approximated as constant in the energy range of measures, then the tunneling conductance dI/dV is proportional to the density of state $\rho_L(eV)$. Tunneling spectroscopy is a powerful technique to measure the electron density of states in a conductor, and it is widely used to measure the energy gap [6] and phonon spectrum [7] in a superconductor. STM can also be used to measure the local density of state, and this type of measurement is known as scanning tunneling microscopy and spectroscopy (STMS). In the STM community, it is a common practice to rid the unknown proportional constant by dividing dI/dV with I/V and quote it as "normalized density of states," which actually equal to

$$\frac{1}{I/V}\frac{dI}{dV} = V\cdot\frac{e^2\rho_R(0)|M(eV,z)|^2\cdot\rho_L(eV)}{e\rho_R(0)\int_0^{eV}|M(E,z)|^2\cdot\rho_L(E)dE} = eV\cdot\frac{\rho_L(eV)}{\int_0^{eV}\rho_L(E)dE}$$

56.3 MEASUREMENT OF THE TUNNEL CURRENT

The previous probabilistic current can be measured as an electrical current because electron carries charge. This tunnel current depends on the bias voltage and the sample–tip distance. For topographical imaging, the bias voltage should be set as high as possible to increase the tunnel current. However, this bias voltage should be less than the work functions of the sample and tip as field emission will replace quantum tunneling at higher voltages. The work function of most metals is in the order of a few electron volts. For this reason, the bias voltage used for most imaging work is between 0.1 and 1 V or slightly higher. With such a bias voltage, the tunnel current should be in the range of nA (10^{-9}A) to pA (10^{-12}A). A greater current indicates direct touch between the tip and sample. Measuring a smaller current is pushing the limit of the electronics, and noise will eventually become intolerable. It is theoretically correct to use as small a tunnel current as possible. Since the tunnel current depends exponentially on the sample–tip distance, pulling back the tip by merely an angstrom will require a lot of reduction in tunnel current, so we should not compromise too much on the noise level in achieving a small tunnel current.

To measure such a small current, most STM put the first-stage amplifier as close to the junction as possible, in many cases just next to the tip or sample. This first-stage amplifier is actually a current to voltage converter, converting the tunneling current to a few volts at the low-impedance output. This output can then be connected to the main electronics over long cables. Since the circuit is close to the junctions and very likely mounted on the STM body, it has to be simple with not too many components. Many STMs just use an operational amplifier to accomplish this, schematically shown in Figure 56.3. Since the current is less than 1 nA with a bias voltage of 1 V, so the junction resistance is at least 1 GΩ. For this large load resistance, the bias voltage can be easily provided by a constant voltage source. In contrast, the junction resistance of a planar tunnel junction is often less than 10 Ω. Small-resistance junction should be current biased with a constant current source, so STM electronics is not very suitable for spectroscopic measurement of planar junctions. It is easier to ground and guard the tip than the sample, so the tip in Figure 56.3 is virtually grounded, and the bias is applied directly to the sample. If the tunnel current is I, the output of this current to voltage converter will be $-IR$ where R is the circuit feedback resistance of the circuit. In many cases, a logarithmic amplifier is used to remove the exponential dependence of the tunnel current. In constant height mode (Section 56.5 of this chapter), this data will be recorded as the sample to tip distance, and its magnitude will be represented by a gray scale in the topographic image. In constant current mode, the output from the current to voltage converter will be compared to a preset value, and the difference or error will then be proportional, integration, and differentiation (PID) processed, magnified to high voltage and fed back negatively to the z-electrode of the scanner to maintain a constant tunnel current. The voltage to the

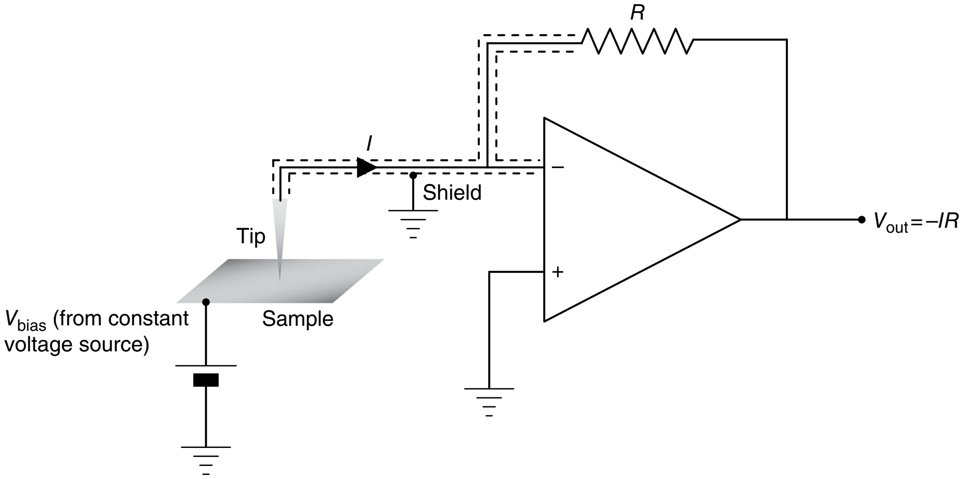


FIGURE 56.3 A current to voltage converter can be used to measure the tunnel current. The sensitivity is determined by the feedback resistance *R*. The operational amplifier has to be extremely low in input bias current.

z-electrode will be attenuated and recorded as the tip position that represents the topographic height.

In many experiments it is necessary to measure *dI*/*dV* for information in density of states. While one can do numerical differentiation on *I*/*V* data, it is preferable to measure *dI*/*dV* directly with a lock-in amplifier for a better signal-to-noise ratio. To do this a small sinusoidal modulation $\delta V \sin \omega t$ is added to the DC bias applied to the sample; the output of the converter will then equal to (δI sin ωt) *R*. Commercial STM electronics often provide external modulation input for this purpose. The modulation signal and the current to voltage converter output should be connected to the reference and voltage input of the lock-in amplifier, respectively. The output of the lock-in amplifier $V_{\text{lock-in}}$ is DC and equals to the product of the amplitude of the sin ωt component in the input voltage and cosine of the phase difference between the input and reference voltage,

$$V_{\text{lock-in}} = -\delta IR\cos\phi = -\left(\frac{dI}{dV}\right)\delta V\cos\phi$$

so if δV is kept constant, $V_{\text{lock-in}}$ is proportional to *dI*/*dV*. The phase of the lock-in amplifier can be adjusted for maximum signal when $\cos\phi = 1$. We now can measure *I* from the converter output and *dI*/*dV* from the lock-in amplifier output simultaneously. Note that for planar tunnel junction, when the junction resistance is small, current source has to be used, and δI will be the control variable, so only *dV*/*dI* can be measured in this case (Fig. 56.4).

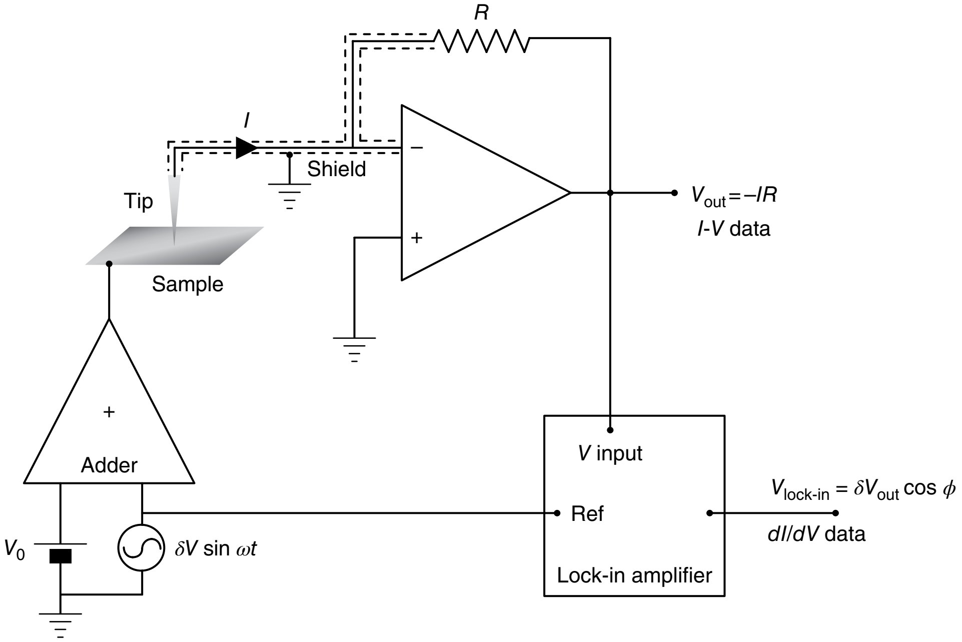


FIGURE 56.4 Lock-in amplifier can be used to measure *dI*/*dV*. A constant voltage modulation δV is added to the bias, and the lock-in amplifier is used to measure the resultant current modulation δI, which is proportional to *dI*/*dV*.

56.4 THE SCANNER

For atomic imaging, it is necessary to control the tip position to angstrom (10^{-10} m) resolution. The only mechanical method to achieve this is the use of piezoelectric materials to produce the motion. Piezoelectricity is the production of accumulated charges and electric field when the material is under mechanical stress. We are using the reverse of this effect to produce motion by applying an electric field to the piezoelectric material. The electric field is conveniently generated by applying a voltage difference across two electrodes deposited on the surface of the component. The piezoelectric material is lead zirconate titanate (PZT), and there are many types of them like PZT 4, PZT 5A, PZT 5H, PZT 8, etc. [8]. All these have slightly different physical properties, and each STM may choose a type that is most suitable for the experimental conditions it is designed for. The piezoelectric component is poled by applying a high voltage across the electrodes when it is manufactured. This will align the polarization of the domain and induce a permanent dipole in the ceramic. The material can be depolarized when it is heated above the Curie temperature, or too high a voltage is applied in reverse to the poling voltage. Piezoelectricity will be lost after the material is depolarized, and the component has to be replaced or repolarized when this happens.

Applying a voltage with the same polarity as the poling voltage will further align the polarization, and the ceramic will elongate between the electrodes (Fig. 56.5b). In reverse, it will contract if the applied voltage is in reverse to the poling voltage (Fig. 56.5c).

The polarity of a component can be determined by reducing the dipole. The dipole will become smaller when the ceramic is compressed between the electrodes or heat to a slightly higher temperature. A voltage in the same polarity as the poling voltage will be induced at the two electrodes, and the polarity of this voltage can be measured with either a voltmeter or oscilloscope (Fig. 56.6).

The most common shape of piezoelectric material used in STM scanner is probably a cylindrical tube, so we will use this as an example in the discussion here. The two electrodes are on the inner and outer walls of the tube. Let us assume the dipole is pointing radially outward, so when a higher voltage is applied to the outer electrode (opposite polarity to the poling voltage), the element will contract radially and elongate along the length of the tube. Reversing the polarity of the applied voltage will make the tube contract. It is this motion that can be used to scan and move the tip. The capacitance between the two electrodes can be measured as $C = 2K_{33}^{T}\varepsilon_0\pi\, L \big/ \ln\left(\dfrac{\mathrm{OD}}{\mathrm{ID}}\right)$,

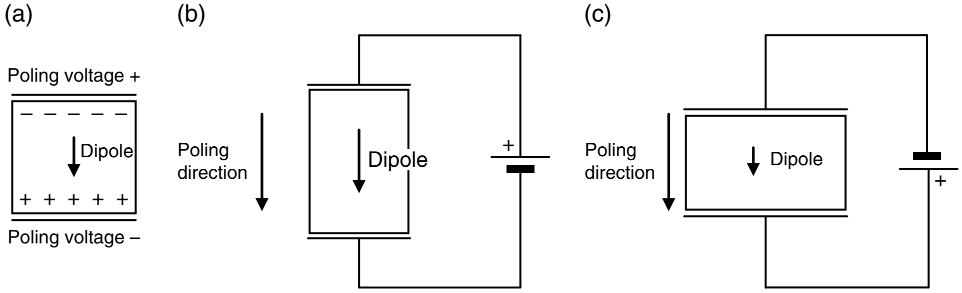


FIGURE 56.5 (a) A piezoelectric material has polarity indicated by the poling direction. In most cases it will (a) elongate when an electric field is applied in the same direction as the poling direction and (b) contract when the electric field is opposite to the poling direction.

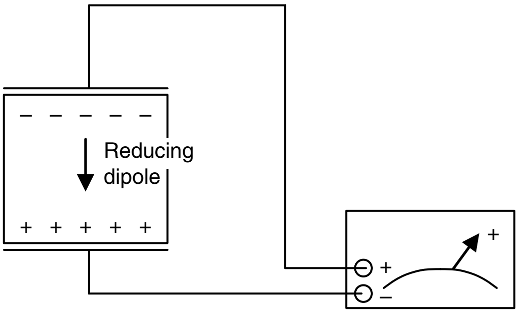


FIGURE 56.6 Determination of the polarity of a piezoelectric component.

TABLE 56.1 d_{31}, d_{33}, and K_{33}^T Values for Different Piezoelectric Materials

	PZT 4	PZT 4D	PZT 5A	PZT 5B	PZT 5J	PZT 5H	PZT 5R	PZT 7A	PZT 7D	PZT 8
d_{31}	−122	−135	−171	−185	−220	−274	−195	−60	−100	−97
d_{33}	285	315	374	405	500	593	450	153	225	225
K_{33}^T	1300	1450	1700	2000	2600	3400	1950	425	1200	1000

d_{31} and d_{33} are in units of 10^{-12} m/V, and K_{33}^T is in 10^{-15}.

where K_{33}^T is the free dielectric constant (given in Table 56.1 for different PZT materials), L is the length of the tube, and OD and ID are the outer and inner diameter, respectively. The capacitance has to be measured at a frequency much lower than the resonance frequency of the tube. In some cases, with some detective work, the type of the ceramic used in an STM can be determined by measuring the capacitance and the K_{33}^T value. Though the capacitance will not tell the condition of the piezoelectric tube, it can be used to confirm the proper connection of the wiring to the electrodes. The extension of the tube depends on the potential difference across the two electrodes as $\Delta L = \frac{2d_{31}\,L}{\mathrm{OD}-\mathrm{ID}}\Delta V$, where d_{31} is the piezoelectric constant with the value of different materials given in Table 56.1. With the scanner tube we can easily produce motion with angstrom resolution. For example, for a quarter-inch-long PZT 8 tube of 1/8-inch outer diameter and 0.020-inch wall thickness, 1 V can extend or contract the tube by 25 Å in length.

We can now construct the scanner of the STM with piezoelectric tube described in the preceding text. There are two common types of scanner, the tripod and the single-tube scanner. The tripod uses three tubes, one for each axis of motion, connected orthogonally at the tip holder as shown in Figure 56.7a. The (x, y, z) coordinates of the tip will be represented by the voltages applied to the three piezoelectric tubes. Another more compact design is to use a single tube to produce all the motion along the three axes. The outer electrode is sliced into four equal pieces along the length of the tube (Fig. 56.7b). By adjusting the voltage applied to a quadrant electrode, we can just expand or contract that quarter of the tube. This will bend the tube and move the tip in the x–y plane. To move the tip along the tube (i.e., z-) direction, we just need to set the voltage to the inner electrode, which is common to all four quadrants. The inner electrode of some scanners has to be permanently grounded to better shield the tip and the wiring running through it. In this case we can electronically add the z-voltage to the four outside voltages as common. Since the applied voltage cannot exceed the depolarization voltage, this will reduce the dynamic range of the scanner and cause inconvenient complication during operation. Alternatively we can also use a longer tube, with one-half for the z-motion and the other half for the x- and y-motion. This will unavoidably lower the resonant frequency of the scanner.

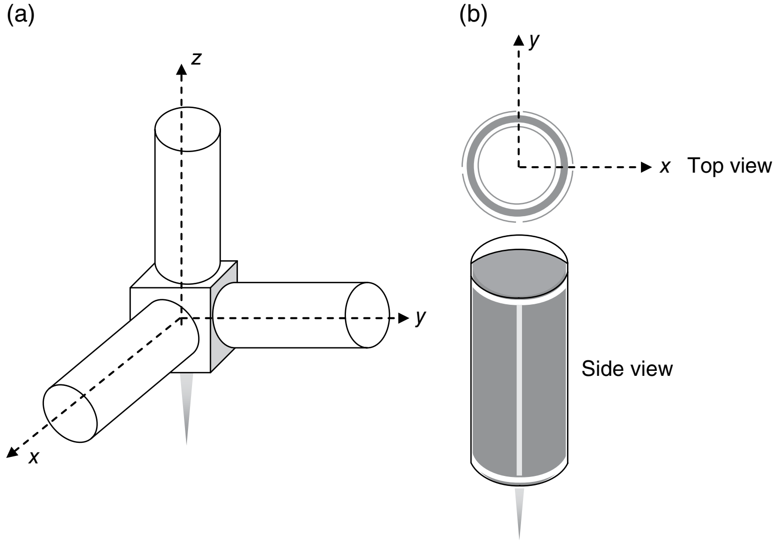


FIGURE 56.7 (a) Tripod design of the tip scanner. (b) A single-tube scanner.

56.5 OPERATING MODE

There are two operating modes of STM. In constant height mode, the tip is scanned with a constant z-voltage, and the tip is maintained at a constant height (Fig. 56.8a), disregarding the topographical variation as it scans along the surface. The tunnel current will be used to represent the topographical height and plotted as a pixel in the 2D image according to a color or gray scale. This requires an atomically flat surface, so the constant height mode is mostly for atomic resolution imaging. For larger scan area, there will be atomic steps and other higher features that the tip may crash into. The STM has to be operated in constant current mode in this case. In constant current mode, the current is compared to a preset value, and the difference is considered as an error between the actual position of the tip and the preset position. Either the positive or negative of this error is added to the z-voltage and negatively fed back to the piezoelectric tube to keep the tip position as close to the preset value as possible. The PID parameters of the feedback loop have to be adjusted to compromise between accuracy, response time, and stability for the best result. The total z-voltage (the offset together with the feedback) will now be used as topographic height to construct the image. The tunneling current should now be roughly constant because of the feedback, but it is advisable to monitor the current map also to ensure the feedback is operating properly.

Since tunnel current is directly measured without any feedback delay in constant height mode, it is easier to obtain crispy high-resolution pictures in constant height mode. However, constant current mode should always be used to scan an unknown area first. The same area can then be imaged with constant current mode after the surface flatness is confirmed.

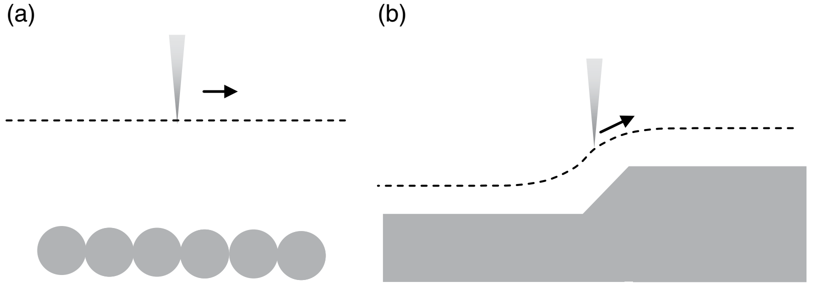


FIGURE 56.8 (a) Constant height mode. (b) Constant current mode.

56.6 COARSE APPROACH MECHANISM

A piezoelectric tube can produce very high-resolution motion, but the range it can scan is very limited. If 1 V can extend the tube by 25 Å, then it can extend its length at most by 5000 Å with 200 V. Higher voltage can break down the piezoelectricity and cause permanent damage to the scanner. Until now, STM with long-range capability is still not readily available. The reason is resolution and scan range are two contradictory properties of an STM. To extend the scanning range of an STM from a single-tube scanner, additional mechanism has to be introduced. These extra constructions will unavoidably make the STM more bulky and massive and weaken the rigidity of the structure. This will lower the resonance frequency of the STM and make it more vulnerable to external vibration and eventually reduce the resolution. The short motion of a scanner also poses a problem in positing the sample within its scanning range without crashing the tip to the sample surface. The coarse approach mechanism is one of the most critical components in the construction of an STM, as it will ultimately affect the performance of an STM.

A coarse approach mechanism is some kind of a mechanical device that can move the scanner and the tip over a long distance of several millimeters but with a poor resolution. Though the mechanism has a poor resolution in its motion, it should still be able to move in step significantly smaller than the maximum range of the scanner. With the coarse approach mechanism, the sample can be placed millimeters away from the tip. The scanner will slowly push the tip toward the sample, and tunnel current is monitored at the same time. If there is no tunnel current after the scanner is fully extended, the scanner will retreat to its natural length. Since the sample is still out of the scanning range, it is safe for the coarse approach mechanism to move one step forward. This cycle will be repeated until a tunnel current is established.

It is not difficult to produce fine motion that is good enough for the coarse approach purpose. Earlier STMs often used simple machines like lever, differential springs and screws, gears, or a combination to reduce the motion (Fig. 56.9). The mechanism is most likely driven by a computer-controlled step motor. These simple machines can produce reliable and robust motions, but they are in general massive, bulky, and not

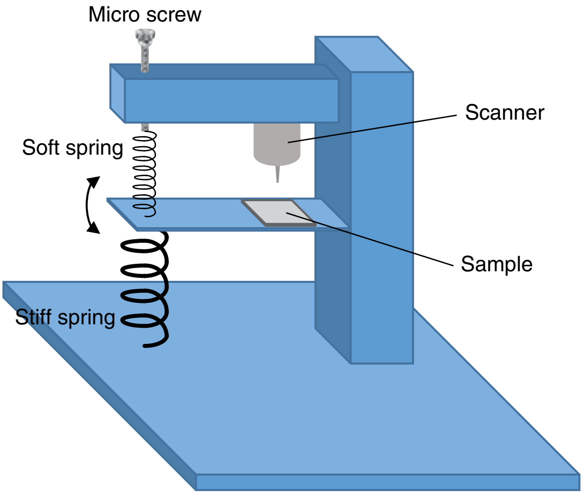


FIGURE 56.9 An example of coarse approach by mechanical components, using the fine motion of a micro screw, and further reduction of motion by differential spring and lever.

very rigid. STMs nowadays mostly employ some kinds of piezoelectric motor for coarse approach. While piezoelectric motors are readily available, most STMs have special designs to integrate the coarse approach motor with the STM structure for the most needed compactness and stability. In here we will discuss some major development in designs over the years as examples to demonstrate the idea.

The first STM developed by Binnig and Rohrer used a device called “louse,” which was a kind of piezoelectric motor already (Fig. 56.10a). The louse is a triangular piezoelectric plate that can expand or contract by applying a voltage across the two sides of the plate. Each vertex is attached to a smooth circular metal foot (MF) coated with a thin I. This structure is placed on three smooth ground plates (GP) of large size. As shown in Figure 56.10, the MFs are grounded to zero volt. When a voltage V_F is applied to a GP, it will be held and locked to the fixed GP by the electric force. Let us label the three vertices as A, B, and C. If we lock down B and C and expand the piezoelectric plate, it can only expand in the vertex A direction. We can then lock down A and unlock B and C before relaxing the piezoelectric plate to its natural size. Repeating this sequence will move the louse with A as spearhead. By the similar method we can choose to move the louse in the direction with any one of the vertices as spearhead. The louse can move to any point on the ground plane, providing a two-dimensional movement. Binnig and Rohrer utilize the louse to transport the sample and use one of the dimensions for coarse approach, as shown in Figure 56.10b. The remaining dimension can be used to move the tip over a long distance over the sample. All piezoelectric coarse approach motors use similar inching mechanism like the louse. The louse uses electric force to cramp one end of the actuator while it is extending or contracting. Another approach is to make use of the static friction, and this type of device is often called inertial motor.

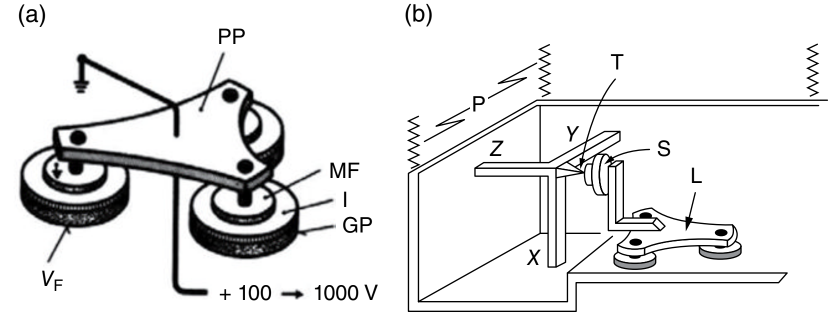


FIGURE 56.10 (a) The louse used by Binnig and Rohrer. GP, ground plate; I, insulator; MF, metal foot; PP, piezoelectric plate; V_F, voltage applied to the ground plate [1]. (b) Louse can be used to coarse approach the sample toward the tip [9]. X, Y, and Z are piezoelectric scanner in tripod configuration, L, louse; S, sample; T, tip. P is springs for vibrational isolation.

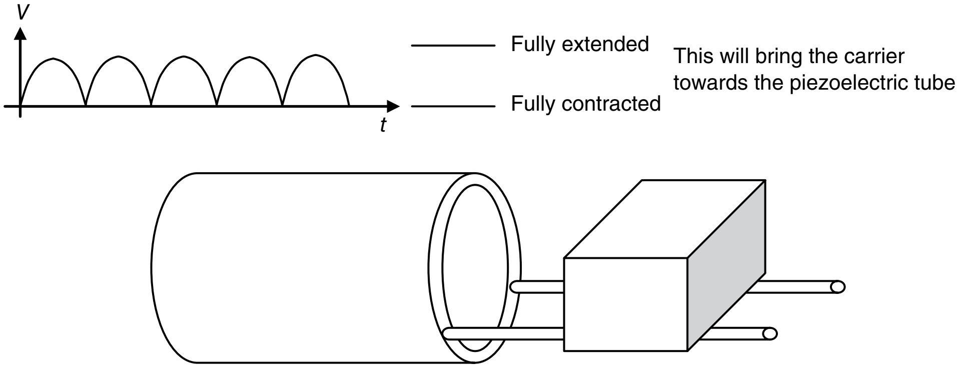


FIGURE 56.11 An example of inertial approach mechanism in which the piezoelectric tube expands and contracts at different rates causing the carrier to slip toward or away from the tube. Top: voltage pattern applied to the piezoelectric tube.

The inertial approach was first introduced by Pohl in 1987 [10]. We will use Figure 56.11 to demonstrate the idea. In this particular example, two rails or the supporter are attached to the end of a piezoelectric tube. The piezoelectric tube has only inner and outer electrodes for extension and contraction. A carrier (mostly for the sample) is then placed on the rails. We need to pay attention when the tube is fully extended and begin to contract, and vice versa, because acceleration is maximum at these extreme points. Suppose we want to move the carrier toward the piezoelectric tube. The acceleration has to be maximized when the tube is fully contracted. Slipping will occur if the acceleration exceeds the limit of the static friction, and the carrier will stay roughly at the same place without following the rail. Now when the rail is fully extended, it has to return and contract with a small acceleration not exceeding the friction limit, and the rail will pull the carrier toward the piezoelectric tube by repeating this cycle. The upper insert of Figure 56.11 shows the voltage applied to the piezoelectric tube. Note that the acceleration is proportional to *dV/dt*. Besides the value of the acceleration at the turnaround points, the actual waveform is not that important in most cases, unless in some special situations like vertical or low-temperature applications when the performance becomes very critical. The scanner tube is often installed concentrically inside the piezoelectric tube holding the rails. If both tubes have the same length and are made of the same material, they will compensate each other in thermal expansion and make the STM less vulnerable to thermal fluctuation and variation. It has been demonstrated that springs could be added to hold the carrier in place, and the inertial motor could be installed in the vertical direction [11].

K. Besocke [12] later extended the inertial motor idea to construct a two-dimensional transporter like the louse but without the cramping force. This device is called a beetle for its appearance, and it is still used by many STMs today because of its versatility. A beetle has three legs built of scanner piezoelectric tube (each four outer electrodes and one inner electrode) supporting a platform at the top. Each tube is attached to a ball bearing at the other end for point contact with the floor (Fig. 56.12a). The *x*- and *y*-electrodes of these tubes have to be aligned in parallel. By applying *x*- and *y*-voltages to a piezoelectric tube and if the acceleration exceeds the friction limit, the ball will slip, and the tube will swing to the designated point on the floor. If all three legs swing with the same displacement, the whole station will transport in the same amount.

The scanner is often installed at the center of the platform, vertically with the tip pointing downward to the sample at the floor. By applying different combinations of voltages to the three piezoelectric tubes, it can produce more complicated motion. For example, if the three legs swing tangentially along the circle joining them, the beetle will rotate around that circle. This provides a method for coarse approach by forcing the beetle to rotate up and down a circular ramp, as shown in Figure 56.12b.

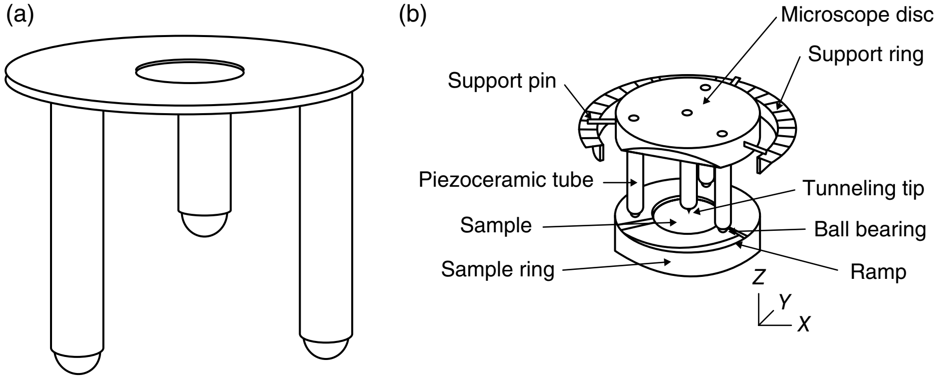


FIGURE 56.12 The beetle. (a) The three legs are piezoelectric tubes like the one used in single-tube scanner. Each leg can swing and slip in any predetermined direction to cause the beetle to rotate or translate to any position. (b) Rotating the beetle on a circular ramp will force it to move up and down, which is useful for coarse approach [13].

While the beetle structure has the advantage of thermal compensation and the possibility of three-dimensional motion, its stability is not outstanding, and it also has to move together with many high-voltage wires attached to the piezoelectric tube. There is a newer design [14] commonly used by homebuilt STM in research laboratories, and it is becoming more popular lately. In this design the motion is produced by shear mode piezoelectric plates. The polarization direction of a shear mode electric plate is within the plane of the plate. A voltage applied across the two sides of the plate will cause them to "slide" on each other and produce the shear motion. Several plates are stacked together to enhance the magnitude of the motion. The STM body has a V-shaped groove along which the carrier will move. Two stacks of piezoelectric plates are glued to each side of the groove. The carrier is a sapphire prism resting on the top of the piezoelectric stacks. The assembly is completed by clamping the prism with another two piezoelectric stacks attached to the top plate, as shown in Figure 56.13a. Figure 56.14 shows the voltage sequence applied to these piezoelectric stacks. The basic idea is to apply a voltage step to a piezoelectric one by one with a time delay, and this will cause the stack to slip on the prism surface. The voltages will then return to zero slowly, and this will cause the piezoelectric plates to relax to the neutral position all together with the prism. Though this motor can only produce one dimension of motion, it can naturally move in a vertical direction, and the rigidity makes it one of the most stable designs. Unlike the beetle, the piezoelectric plates are not part of the

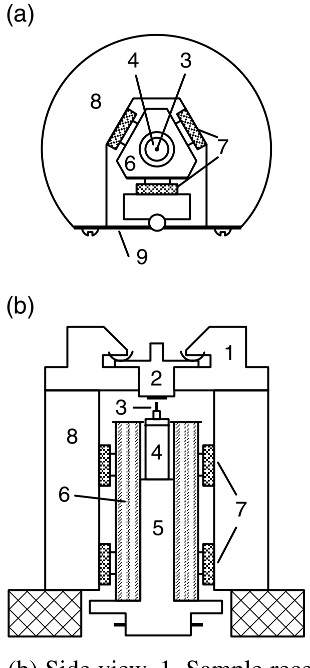


FIGURE 56.13 (a) Top view. (b) Side view. 1, Sample receptacle; 2, sample holder; 3, tip; 4, single-tube scanner; 5, scanner holder; 6, sapphire prism; 7, shear piezo stacks; 8, macor body; 9, spring plate [14].

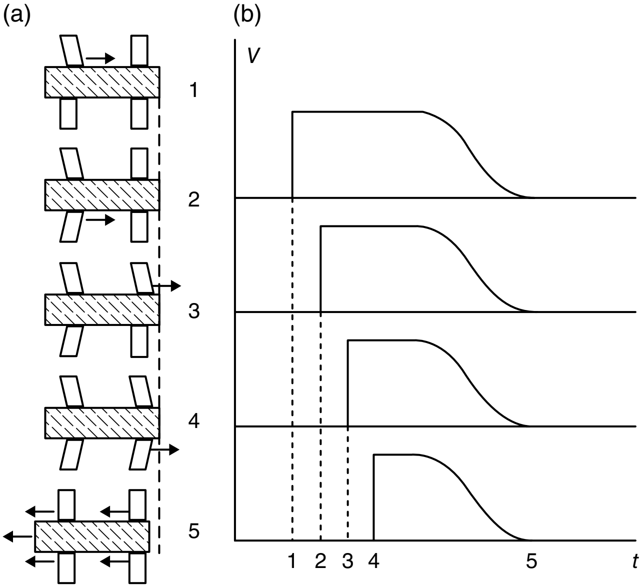


FIGURE 56.14 (a) Positions of the piezoelectric stacks at different times (1–5) when the voltage pattern in (b) is applied to the appropriate stack. Note that for simplicity only four stacks are shown here to demonstrate the idea [14].

carrier, and hence the high-voltage wirings are stationarily attached to the STM body. This reduces the carrier load significantly, and the motion produced by this design should be more reliable.

56.7 SUMMARY

In this chapter we have reviewed the principle of operation of an STM. An STM can be operated in an ambient atmosphere or ultrahigh vacuum, at room temperature or ultralow temperatures. It can provide high-resolution topographical image of conducting surfaces. STM is not just a high-power microscope. Measurement of dI/dV will give information on the electron density of states, and the surface work function can be measured by dI/dz. STM has also been used to manipulate individual atom on metallic surfaces. Other types of scanning probe microscope like the AFM and the SNOM or NSOM are developed based on the idea of STM.

One major advantage of STM is in the simplicity of instrumentation. The highest voltage used in an STM is at most 100–200V. It does not involve expensive electron optics. The component that gives rise to the high resolution is a small tip that is commercially available at a low cost or can be prepared by simple methods in a lab (not reviewed here). The most critical part in the electronics is probably in the measurement of the extremely small tunnel current in nA to pA range. A preamplifier is often put next to the tip in converting the tunnel current to a voltage signal, and the requirements on measurement will be minimal after this point. A good STM should be small in size but rigidly built. For this reason the coarse approach mechanism will determine the performance of an STM, and it has to be carefully designed.

REFERENCES

1. G. Binnig and H. Rohrer, “Scanning tunneling microscopy”, *Helvetica Physica Acta* 55, 726 (1982).
2. G. Binnig, C. F. Quate, and Ch. Gerber, “Atomic force microscope”, *Physical Review Letters* 56, 930 (1986).
3. G. Meyer and N. M. Amer, “Novel optical approach to atomic force microscopy”, *Applied Physics Letters* 53, 1045 (1988).
4. F. J. Giessibl, “Advances in atomic force microscopy”, *Reviews of Modern Physics* 75, 949 (2003).
5. J. Tersoff and N. D. Lang, “Theory of scanning tunneling microscopy”, *Methods of Experimental Physics* (Ed. J. A. Stroscio and W. J. Kaiser), Academic Press, San Diego, pp. 1–29 (1993).
6. M. Tinkham, *Introduction to Superconductivity*, 2nd edition, McGraw-Hill, New York, pp. 71–78 (1996).
7. E. L. Wolf and G. B. Arnold, “Proximity electron tunneling spectroscopy”, *Physics Reports* 91, 33 (1982).
8. H. Jaffe and D. A. Berlincourt, “Piezoelectric transducer materials”, *Proceedings of the IEEE* 53, 1372 (1965).
9. G. Binnig and H. Rohrer, “Scanning tunneling microscopy”, *IBM Journal of Research and Development* 30, 355 (1986).
10. D. W. Pohl, “Sawtooth nanometer slider: a versatile low voltage piezoelectric translation device”, *Surface Science* 181, 174–175 (1987).
11. Ch. Renner, Ph. Niedermann, A. D. Kent, and O. Fischer, “A vertical piezoelectric inertial slider”, *Review of Scientific Instruments* 61, 965 (1990).
12. K. Besocke, “An easily operable scanning tunneling microscope”, *Surface Science* 181, 145–153 (1987).
13. J. Frohn, J. F. Wolf, K. Besocke, and M. Teske, “Coarse tip distance adjustment and positioner for a scanning tunneling microscope”, *Review of Scientific Instruments* 60, 1200 (1989).
14. S. H. Pan, E. W. Hudson, and J. C. Davis, “3He refrigerator based very low temperature scanning tunneling microscope”, *Review of Scientific Instruments* 70, 1459 (1999).

57

MEASUREMENT OF LIGHT AND COLOR

John D. Bullough
Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY, USA

57.1 INTRODUCTION

This chapter provides the reader with some of the basic terminology and concepts used in the measurement of light (photometry) and color (colorimetry) pertaining to lighting systems. Also described are some of the types of instrumentation used to make photometric and colorimetric measurements.

57.2 LIGHTING TERMINOLOGY

57.2.1 Fundamental Light and Color Terms

57.2.1.1 Light Light is defined as radiant energy in the electromagnetic spectrum that is capable of producing a visual sensation in humans through stimulation of the retina [1]. Electromagnetic radiation includes gamma rays, X-rays, ultraviolet (UV) energy, infrared (IR) energy, and radio frequencies as illustrated in Figure 57.1. Light comprises of only a small portion of the entire electromagnetic spectrum.

The wavelength band containing light is from approximately 380 to 780 nm (1 nm = 10^{-9} m). Light itself (i.e., rays of light) cannot be perceived directly but must be directed from a luminous surface or reflected from an object toward the eye. When the light reaches the retina (the photosensitive layer in the back of the eye), the photoreceptors in the retina transmit signals to the brain, which are interpreted by the visual centers of the brain as visual information about the luminous environment. Importantly,

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

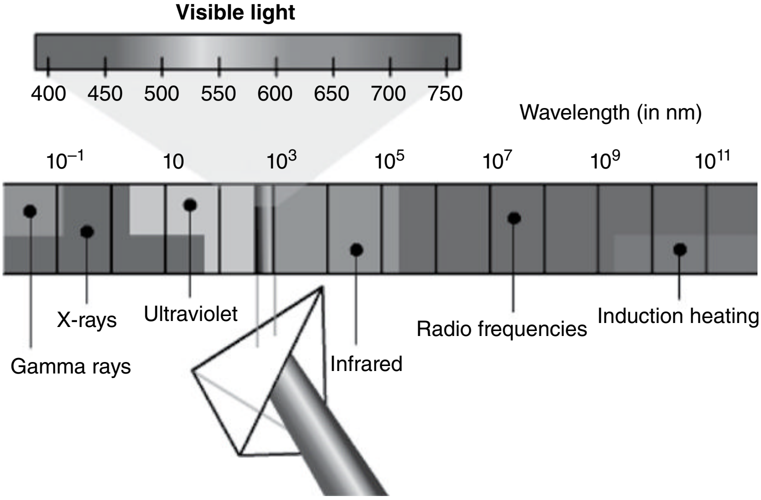


FIGURE 57.1 Electromagnetic spectrum showing the location of visible light. Source: Reproduced with permission of the Lighting Research Center.

this definition of light is made with reference to human visual responses and not the responses of any other organism to radiant energy in the wavelength band defined as light. Some animal species, for example, may be able to respond to UV energy and in some sense could be considered to be light for that species, but there are no formal definitions of light that are not related to human visual responses. Light is the only of the fundamental physical quantities (the others are length, mass, time, electric current, temperature, amount of substance) that depends upon human experience [2].

Collectively, light, in addition to electromagnetic radiation in the bands adjacent to light, UV, and IR radiation, is often referred to as optical radiation because these forms of radiant power can enter and interact with the optical tissues of the human eye (primarily the cornea and lens).

57.2.1.2 Spectral Power Distribution The radiant output of a light source, such as the sun or an electric lamp, can be expressed graphically as a spectral power distribution (SPD) as illustrated in the following. The SPD curve shows the relative distribution of radiant power in the different visible spectral bands (see Fig. 57.2).

Very approximately, the wavelength bands of visible light correspond to different perceived colors as follows:

- 380–430 nm: violet
- 430–490 nm: blue
- 490–560 nm: green
- 560–600 nm: yellow
- 600–620 nm: orange
- 630–780 nm: red

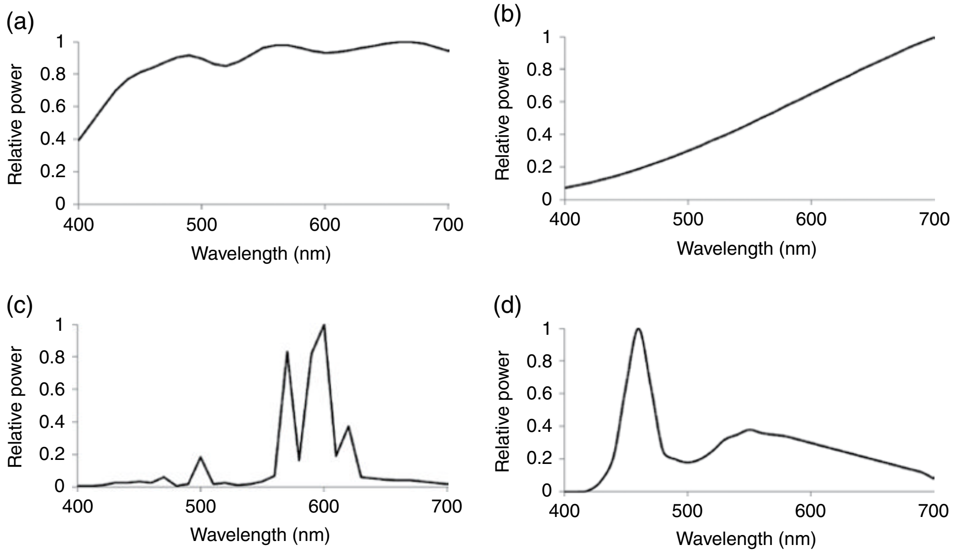


FIGURE 57.2 SPD curves for various light sources. (a) Sunlight, (b) incandescent, (c) high-pressure sodium, and (d) light-emitting diode (LED).

SPD curves like those in Figure 57.2 provide some clues about how different colored objects might appear under a given light source. For example, sunlight (Fig. 57.2a) produces a broad SPD with radiant power across the entire visible light bandwidth, and incandescent light bulbs (Fig. 57.2b) produce more radiant power in the longer visible wavelengths, corresponding to light that is perceived as yellow, orange, and red. This accounts for the warm or yellowish appearance of incandescent bulbs compared to other sources. A high-pressure sodium (HPS) lamp (Fig. 57.2c), commonly used for street and outdoor lighting, produces most of its energy in the portion of the spectrum perceived as yellow and orange (570–620 nm) resulting in the very yellowish appearance of this light source.

57.2.1.3 Correlated Color Temperature As the tungsten filament of an incandescent bulb is heated, it produces light, first a reddish light when the filament is beginning to warm up, then yellowish light until finally it produces its familiar warm white light. Tungsten behaves almost exactly like an ideal blackbody, a material that radiates energy (including light) as a function of its temperature. The higher the temperature, the greater the proportion of short visible wavelengths (bluish light) relative to long visible wavelengths (reddish light) is. The temperature of an ideal blackbody is used as a common metric of the relative coolness or warmness of the appearance of the light. Since no real-world light source is an ideal blackbody, the term correlated color temperature (CCT) is used to define the temperature of an ideal blackbody (in kelvins: K) that most closely matches the color of a light source.

Somewhat confusingly [3], low CCTs that might be considered relatively cool (in terms of temperature) actually produce more yellowish light commonly called warm light, and high CCTs corresponding to a higher blackbody temperature produce more bluish light commonly called cool light:

- CCTs below 3200 K: warm white light
- CCTs between 3200 and 4000 K: neutral white light
- CCTs above 4000 K: cool white light

57.2.1.4 Color Rendering Index As might be deduced from the SPD curves in Figure 57.2, different light sources will perform differently regarding the way they make objects of different colors look. Under sunlight (Fig. 57.2a), which produces energy across the entire visible spectrum, objects of any colors might be expected to look natural, but a blue car seen under HPS illumination (Fig. 57.2c), which has very little spectral power in the wavelength band corresponding to blue light (~430–490 nm), might not even be seen as blue but rather a distorted color like black or brown, whereas a yellow car under HPS might be expected to look yellow because of the large amount of spectral output in the wavelength range between approximately 560 and 600 nm.

The color rendering index (CRI) is a measure of the color shift of a range of colors that occurs for a given light source compared to the color shift under a comparison, reference source. By convention, the reference source is an ideal blackbody (almost identical to an incandescent bulb's tungsten filament) when the lamp's CCT is 5000 K or lower, and the reference source is daylight when the lamp's CCT is higher than 5000 K. Since incandescent lamps have traditionally been the most common warm white (low CCT) light source experienced in many locations (especially in residences), and daylight was the most common cool white (high CCT) source experienced in the early to middle part of the twentieth century, the CRI is often used as a measure of the naturalness of the appearance of colored objects under a given source. It is expressed as a numerical quantity with a maximum value of 100, representing color appearance identical to that under the reference source. Lower CRI values represent larger color shifts and often less natural appearance (at least, compared to incandescent or daylight illumination). Table 57.1 lists CRI values for several common light sources.

Several alternatives and supplements to CRI are presently under discussion in the lighting industry [4], but at present, CRI is the primary metric used to assess the quality of color rendering of a light source.

TABLE 57.1 CRI Values for Several Common Light Sources

Light Source	CRI
Incandescent lamp	99
4000 K fluorescent lamp	89
High-pressure sodium	22
5000 K light-emitting diode	78

57.2.2 Terms Describing the Amount and Distribution of Light

57.2.2.1 Luminous Intensity Luminous intensity is a measure of the amount of light emitted by a light source in a particular direction, within a particular angular cone from the source. It is measured in units of candelas (cd) and is sometimes referred to as candlepower. As stated in its definition, luminous intensity is direction specific; the luminous intensity of a light source in one direction can, and often will, differ from that in another direction. Luminous intensity is an inherent characteristic of a light source and is not dependent upon the distance from the source. As long as the direction from the light source remains the same, the luminous intensity from the source in that direction will also remain the same. Historically, the candela was defined in terms of the luminous intensity of a standard candle flame, giving the term its name.

57.2.2.2 Luminous Flux Luminous flux is a measure of the total amount of light produced by a light source in all directions around the source. It is measured in lumens (lm). It is the quantity seen on lamp packages to express the total light output of the lamp. Geometrically, the lumen is the amount of luminous flux produced by a point source with a uniform luminous intensity of 1 cd, within an angular cone having a solid angle of 1 steradian. A steradian (Fig. 57.3) is the solid angle subtended by a cone that, when projected onto a sphere, has an area equal to the square of the sphere's radius. Therefore, 1 cd represents 1 lm/steradian (in a given direction).

To determine luminous flux (lumen) ratings of light sources, the conditions under which the light source is measured must be carefully controlled, including the ambient temperature, the orientation of the lamp, the input voltage and current, and vibration, because most light sources are sensitive to these conditions [1].

57.2.2.3 Luminous Intensity Distribution Luminous flux is a useful quantity to help understand how much light a given source produces when the specific direction of the light is not important, such as for general room lighting, but it is less useful for directional lights such as flashlights, spotlights, vehicle headlights, or display lighting, where the lighting system must produce a narrower distribution of light. Two

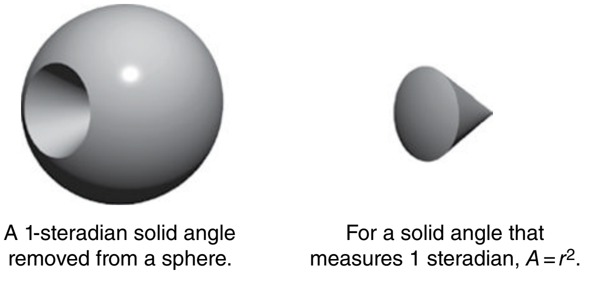

A 1-steradian solid angle removed from a sphere.
For a solid angle that measures 1 steradian, $A = r^2$.

FIGURE 57.3 Graphical representation of a steradian. Source: Reproduced with permission of the Lighting Research Center.

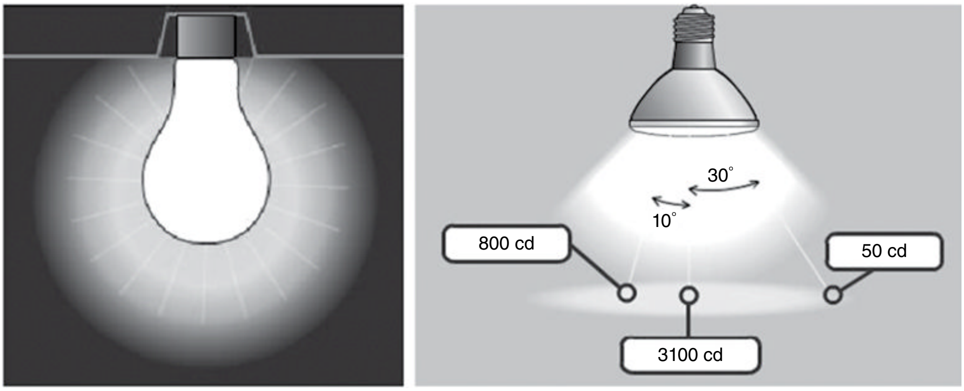


FIGURE 57.4 The general service incandescent lamp shown in the left panel may produce a uniform luminous intensity of 200 cd in most directions. The spot lamp shown in the right panel can produce an equivalent luminous flux but would have much higher intensity in one particular direction and much lower intensity in other directions. Source: Reproduced with permission of the Lighting Research Center.

incandescent lamps, a general service light bulb and a spot lamp (Fig. 57.4), with the same wattage, will have similar lumen outputs, but the spot lamp will be more useful as a headlight or flashlight because it has a high intensity in a particular direction and relatively low intensity elsewhere, whereas the general service bulb will have a modest intensity in nearly every direction from the bulb.

The luminous intensity distribution from a light source can be represented on a polar coordinate graph (Fig. 57.5) that shows the luminous intensity from the light source as a function of the angular direction from the front of the source, represented by a polar angle of 0°.

The luminous intensity distribution can be represented graphically as illustrated in Figure 57.5 or in tabular form for different angles from the light source.

When a light source or system would be expected to have different distributions in different directions, two or more luminous intensity distribution curves can be shown. For many fluorescent lighting systems where the light source is a linear tube, the distribution along the length of the lamp would differ substantially from that across the lamp. In this case, two luminous intensity distributions (Fig. 57.6) would be used to illustrate the distribution of light from such a lighting system.

Sometimes the luminous intensity distribution is graphed using rectangular rather than polar coordinates; this is common for many LED sources.

57.2.3 Terms Describing Lighting Technologies and Performance

57.2.3.1 Luminaire A luminaire, commonly called a light fixture, is a complete lighting unit consisting of a lamp or lamps, a ballast or driver (if needed to operate the lamp), and parts designed to position and protect the lamps, to connect the lamps or

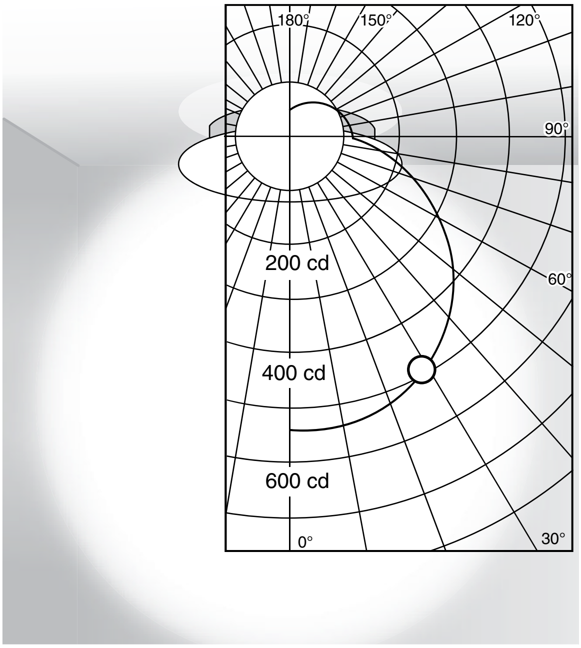


FIGURE 57.5 A luminous intensity distribution of a ceiling luminaire plotted on a polar coordinate graph; the luminous intensity is approximately 550 cd at 0° (directly below the source) and is 500 cd at 30°. Source: Reproduced with permission of the Lighting Research Center.

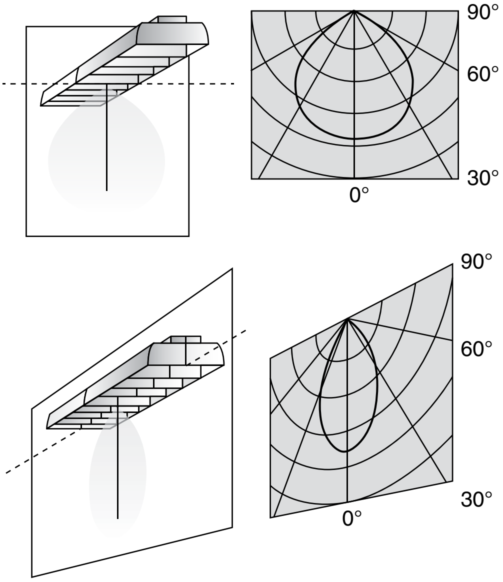


FIGURE 57.6 Top: luminous intensity distribution for a fluorescent luminaire, in the plane across the lamp. Bottom: luminous intensity distribution for the same luminaire, in the plane along the length of the lamp. Source: Reproduced with permission of the Lighting Research Center.

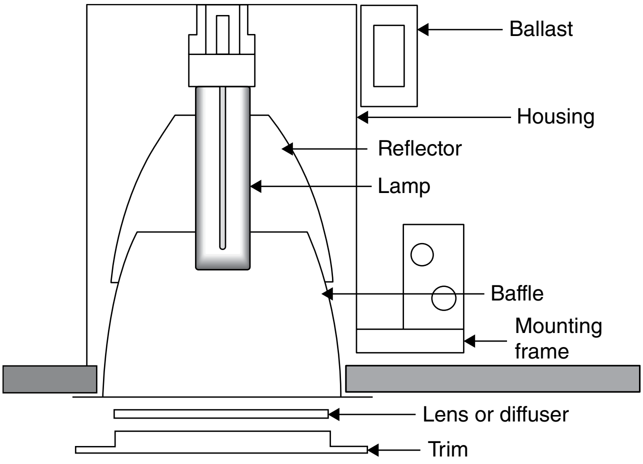


FIGURE 57.7 Cutaway diagram of a compact fluorescent downlight luminaire, showing the individual parts. Source: Reproduced with permission of the Lighting Research Center.

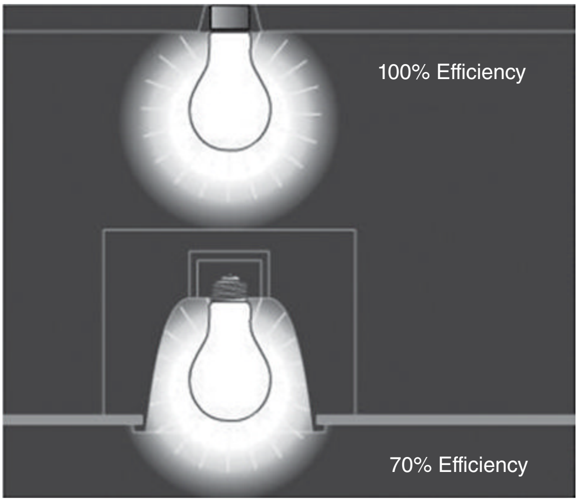


FIGURE 57.8 Top: a bare lamp without a luminaire emits all of its lumens. Bottom: only a percentage of the lumens emitted by the lamp will exit a luminaire; this percentage is the luminaire's efficiency. Source: Reproduced with permission of the Lighting Research Center.

ballasts to the power supply, and to direct the light. Light-directing components may be reflectors, diffusers, baffles, or lenses. An example of a luminaire is the downlight illustrated in Figure 57.7.

57.2.3.2 Luminaire Efficiency The efficiency of a luminaire is defined as the ratio (in percent) of the luminous flux emitted by a luminaire to that emitted by the lamp or lamps within the luminaire. It is a percentage of the lamp lumens that are ultimately emitted by the entire luminaire (see Fig. 57.8).

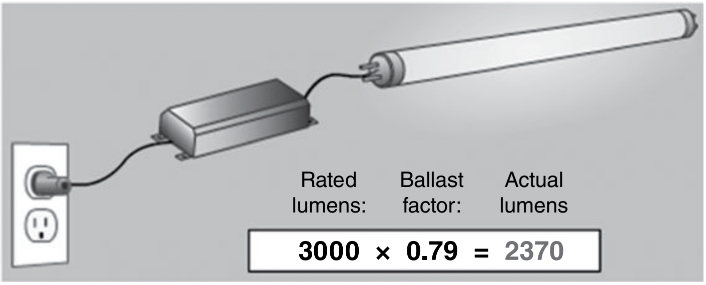


FIGURE 57.9 A fluorescent lamp operated on a reference circuit might produce 3000 lm but when operated on one specific ballast might produce 2370 lm, a ballast factor (BF) of 0.79. Source: Reproduced with permission of the Lighting Research Center.

57.2.3.3 Ballast Factor For a luminaire that contains a ballast to operate the lamp (such as those using fluorescent or high-intensity discharge (HID) lamps), the ballast uses some of the power necessary to operate the luminaire. Different ballasts will use different amounts of power, depending upon the type of functionality they provide (e.g., dimming, cold temperature operation, etc.). When determining the luminous flux produced by a lamp, a specific reference ballast circuit for each type of lamp is specified [1] to ensure repeatable and consistent results, but the ballast in a particular luminaire is likely to differ from this specific type of reference circuit. The ratio between the luminous flux produced by a lamp using a particular ballast and that produced when using the reference ballast circuit is defined as the ballast factor (commonly abbreviated BF; see Fig. 57.9).

The BF value allows a lighting specifier to predict the lamp lumens in a given luminaire, relative to the rated lumen value for the lamp(s) provided by the lamp manufacturer for the reference ballast circuit. Most of the time, but not always, BF values are less than 1.0.

57.2.3.4 Rated Life For most conventional light sources, the rated life is defined as the amount of time a large group of lamps would be operated (usually, in hours), before half of the lamps in the group would be expected to have failed or burned out (Fig. 57.10). Lamp life is assessed under specific conditions (e.g., temperature, lamp orientation, voltage) and using specific operating cycles [1]. For example, the life of an incandescent lamp is determined by operating the lamp continuously until failure. For fluorescent lamps, where failure of the electrodes in the lamp is a common failure mechanism, lamps are operated on a constant cycle of 3 h on, followed by 20 min off, continuously until failure. For HID lamps, the lamps are operated for 11 h followed by 1 h off until failure.

Because burnout is not a common mechanism for the failure of light-emitting diode (LED) sources, the lighting industry has worked to develop more practical and meaningful measures of useful life for these sources. A common approach [5] is to specify

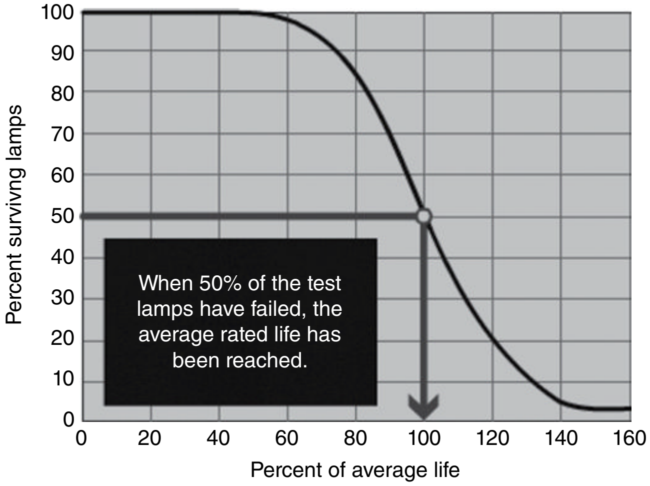


FIGURE 57.10 Rated lamp life is usually defined by the operating hours at which 50% of the lamps in a sample have failed. Source: Reproduced with permission of the Lighting Research Center.

the operating time (in hours) to reach a reduction in light output (such as to 50 or 70%) from the initial light output.

57.2.3.5 Luminous Efficacy (Electrical) Luminous efficacy is defined as the quotient of the total luminous flux (in lm) produced by a lamp or luminaire, by the total electrical power input (W) of the lamp or luminaire. For luminaires, the luminous flux should be modified by the BF, and the power should include power used by both lamp(s) and ballast (if any). Luminous efficacy is expressed in units of lumens per watt (lm/W).

Sometimes the stated luminous efficacy of LED sources is the optical luminous efficacy, given as the lumens per watt of radiant power produced by the source, not the watts used to provide power to the source. Despite having identical units, optical luminous efficacy and electrical luminous efficacy values are different and cannot be compared in a meaningful way.

57.2.4 Common Quantities Used in Lighting Specification

57.2.4.1 Illuminance

The definition of illuminance is the density of luminous flux incident on a surface. It is most commonly expressed in units of lux (lx) or footcandles (fc). 1 lux is equivalent to 1 lumen per square meter (lm/m^2), and 1 fc is equivalent to 1 lumen per square foot (lm/ft^2). If a point source of light (Fig. 57.11) with a uniform luminous intensity of 1 cd were surrounded by a sphere with a radius of 1 m, the illuminance on the interior surface of the sphere would be 1 lx. If the same light source were surrounded by a sphere

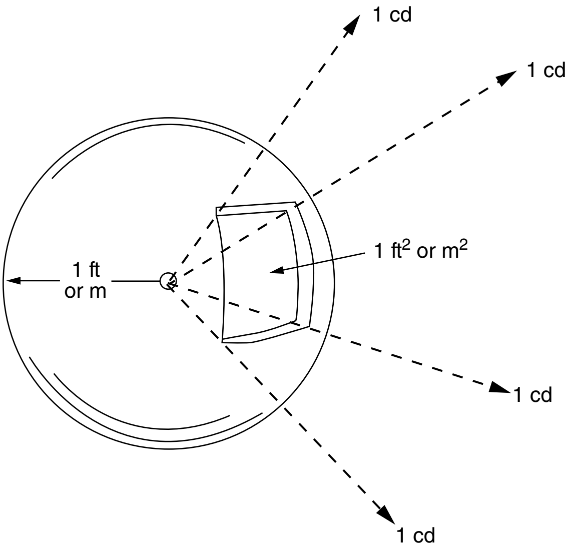


FIGURE 57.11 A point source in the center of the sphere with a uniform luminous intensity of 1 cd would produce an illuminance of 1 lx on if the sphere's radius were 1 m, and 1 fc if the sphere's radius were 1 ft.

with a radius of 1 ft, the illuminance on the interior surface of the sphere would be 1 ft. Because there are 10.76 ft^2 in 1 m^2, 1 fc is equal to 10.76 lx. Commonly, a rounded factor of 10 is used to relate fc to lx (i.e., 30 fc is ≅300 lx).

Unlike luminous intensity and luminous flux, which are properties of the light source, illuminance is dependent upon the geometry between the light source and the surface being illuminated. For example, the 1-cd source shown in Figure 57.11 produces an illuminance of 1 lx on a surface that is 1 m away but produces only 0.25 lx on a surface that is 2 m away. The relationship between the luminous intensity of a light source and the illuminance it produces at a given distance follows the inverse-square law, where the illuminance is inversely proportional to the square of the distance. The inverse-square law is commonly written as follows:

$$E = \frac{I}{d^2} \tag{57.1}$$

where E is the illuminance in lx (or fc),

I is the luminous intensity from the source in the specific direction in cd,

and d is the distance between the light source and the surface being illuminated, in meters (or feet).

When d in Equation 57.1 is in meters, E is in lx. When d is in feet, E is in fc.

Because illuminance is a representation of the amount of light falling on a surface such as a countertop, a desk, a chalkboard, or a piece of machinery, it is the most

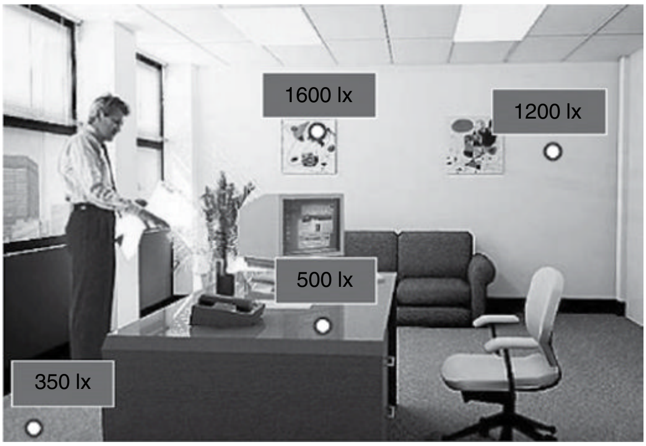


FIGURE 57.12 Distribution of selected illuminances on various surfaces within a space. Source: Reproduced with permission of the Lighting Research Center.

common quantity used in the specification of recommended light levels for many different applications in buildings and outdoors [1]. In offices, for example, the illuminance on desks is the common light level specification; in parking lots, the illuminance on the pavement surface is the required light level. However, illuminances (Fig. 57.12) can be horizontal (as on desks or pavement), vertical (as on paintings hung on walls or a person's face in front of a bathroom mirror), or in any other plane (such as an inclined instrument panel on a factory machine).

57.2.4.2 Luminance Luminance is defined as the amount of light directed (or reflected) from a surface, in a particular direction and within a particular solid angular cone. Luminance is most commonly expressed in units of candelas per square meter (cd/m^2), sometimes called nits. More practically, luminance represents a quantity that is somewhat analogous to the brightness of a surface. The luminance of a surface is specific to the direction from which the surface is viewed (Fig. 57.13) but does not change as the distance from the surface is changed. For example, if the luminance of a vertical wall at the end of a corridor is 100 cd/m^2 when viewed from the opposite end of the corridor, this luminance would not change as an observer moved along the corridor toward the wall because the direction of view from the wall to the observer would not change.

57.2.4.3 Reflectance Consider a black desk located underneath a luminaire that produces an illuminance of 300 lx on the desktop and a white piece of paper sitting on the desk. Both the black desk and the white paper would have the same illuminance (300 lx) incident upon them, but the paper would appear substantially brighter than the desk. This is because the reflectance of the paper is much higher than that of the desk. The reflectance is defined as the ratio of the luminous flux incident on a surface to the

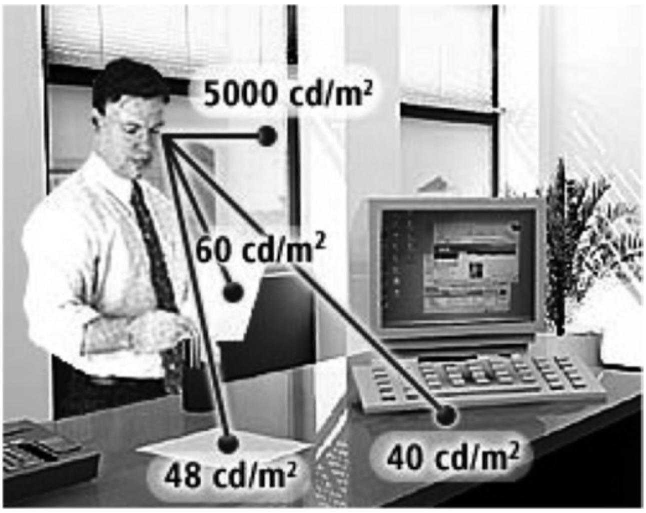


FIGURE 57.13 Surface luminances are expressed with respect to a particular direction, in the previous illustration, toward the observer's eyes. Source: Reproduced with permission of the Lighting Research Center.

amount of flux that is reflected, from the surface. Reflectances are expressed as unitless quantities from 0 (a perfectly black surface that reflects no light) to 1 (a perfectly white source that reflects all of the light that reaches it). Sometimes reflectances are given in terms of percentages from 0 to 100%. Some typical room and surface reflectances are as follows:

- White painted ceiling: 0.8
- Light finished/painted walls: 0.5
- Floors: 0.2
- White office paper: 0.8
- Asphalt pavement: 0.1

The luminance of a matte (diffuse) surface can be estimated if its reflectance and the illuminance falling on the surface are known, using Equation 57.2:

$$L = \frac{E\rho}{\pi} \tag{57.2}$$

where L is the luminance (in cd/m^2),
E is the illuminance on the surface (in lx),
ρ is the reflectance of the surface,
and $\pi \approx 3.14$.

For matte surfaces, Equation 57.2 predicts the luminance in all viewing directions. For glossy or semiglossy surfaces, the relationship among luminance, illuminance, and reflectance is very complex and dependent upon the specific geometry among the light source, the surface, and the observer. Lighting calculations often make the simplifying assumption that all surfaces within an illuminated space are matte, so that their luminances can be predicted by Equation 57.2. However, it should be recognized that for highly shiny or "specular" surfaces like polished metal or glass, luminances cannot be readily estimated by this simple calculation.

57.3 BASIC PRINCIPLES OF PHOTOMETRY AND COLORIMETRY

57.3.1 Photometry

Photometry is a simple, mathematically precise system of measuring and specifying light agreed to by an international community involved with its commerce and specification. It is the basis for the illuminance and luminance quantities described in earlier sections of this chapter. In this section, the relationship between a light source's SPD, its luminous flux, and the luminous efficiency functions used to define light is described.

57.3.1.1 Luminous Efficiency Function The luminous efficiency function is used to relate the relative effectiveness of radiant power along the visible spectrum (between about 380 and 780 nm) at creating a visual sensation, since equal amounts of radiant power at different wavelengths will not necessarily produce equal visual sensations. Several luminous efficiency functions have been developed [6], but the most common is the photopic luminous efficiency function [2], often denoted $V(\lambda)$, where V stands for visibility and λ represents the wavelength. The peak value of the photopic luminous efficiency function is at 555 nm (Fig. 57.14), where the luminous efficiency is defined as 1.0. The photopic luminous efficiency function represents the combined spectral sensitivity of the cone photoreceptors in the central portion of the human retina, and the peak spectral sensitivity of the combination of cones in this portion of the retina is 555 nm. There is also a scotopic luminous efficiency function, denoted $V'(\lambda)$, representing the spectral sensitivity of the human retina's rod photoreceptors, which have a peak spectral sensitivity at 507 nm rather than 555 nm (Fig. 57.14).

All luminous efficiency functions are expressed in unitless quantities between 0 and 1, with the wavelength having the greatest visual sensation defined to have a value of 1 and all other wavelengths having lower values. In practice, only the photopic luminous efficiency function is used to characterize light in almost every situation [1]. This is because light levels must be very low in order to be in a state where only rod photoreceptors contribute to human vision [6]. The presence of nearly any electric light source will place an observer above the scotopic luminance range. Most interior

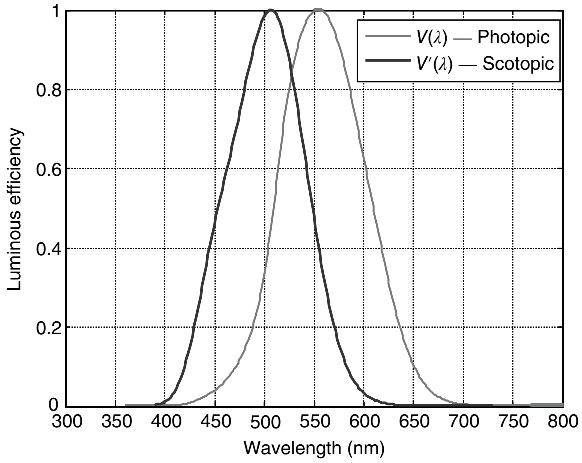


FIGURE 57.14 Photopic (right, gray) and scotopic (left, black) luminous efficiency functions. Source: Reproduced with permission of the Lighting Research Center.

lighting applications are of sufficient intensity to place the observer in the photopic luminance range, where cone photoreceptors contribute to vision. There is also a range of lighting applications, primarily outdoor nighttime applications, where a mixture of rods and cones contribute to vision (denoted the mesopic luminance range), but there is no single luminous efficiency function that can characterize mesopic luminous efficiency [7], and this special set of conditions is not discussed further in this chapter.

57.3.1.2 Calculating Luminous Flux In order to calculate the luminous flux from the SPD of a light source, several steps must be undertaken. The SPD should be expressed in terms of the radiant power (in W) produced by the light source at each wavelength in the visible spectrum, and the power at each wavelength is multiplied by the value of the luminous efficiency function at each wavelength (Fig. 57.15), integrated across the wavelength limits of the visual spectrum (e.g., 380–780 nm).

Mathematically, the calculation can be expressed as the following integral equation:

$$\Phi = k \int_{380\ \mathrm{nm}}^{830\ \mathrm{nm}} P(\lambda)\cdot V(\lambda)\cdot d\lambda \tag{57.3}$$

where Φ is the luminous flux (in lm),
k is a constant equal to 683 lm/W,
P(*λ*) is the light source radiant power at wavelength *λ*,
and *V*(*λ*) is the photopic luminous efficiency function.

An important step in this process is the inclusion of the constant *k* in the aforementioned equation, which allows the conversion of specific amounts of radiant power

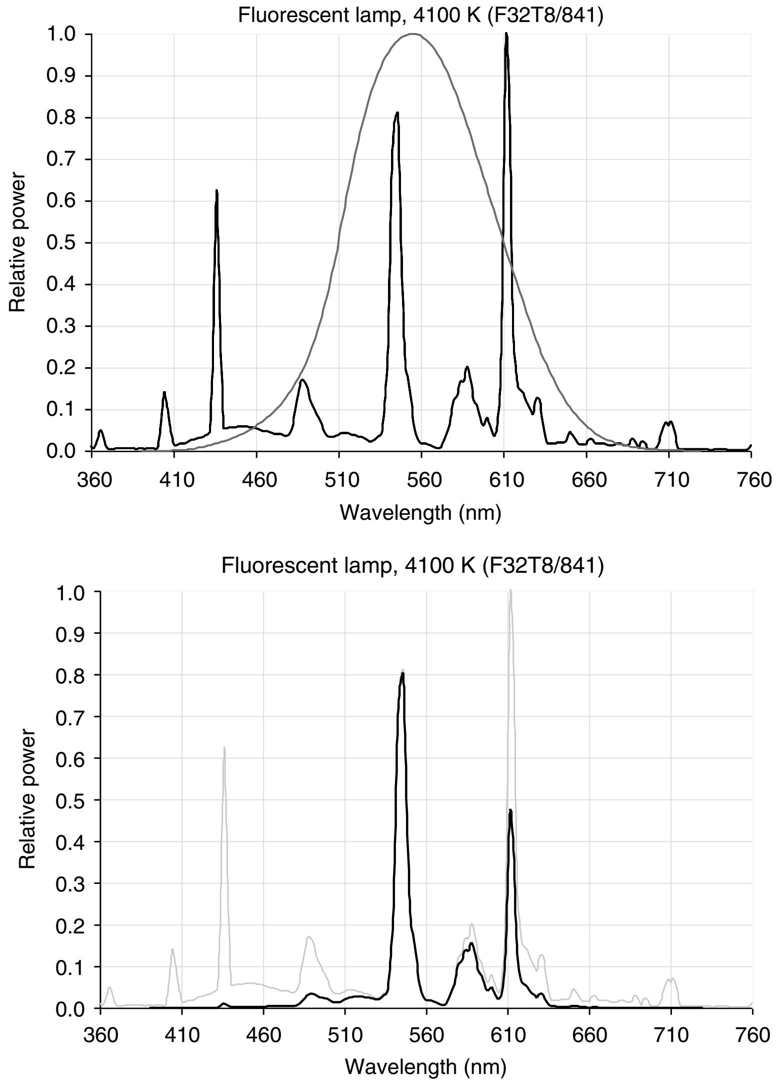


FIGURE 57.15 The light source radiant power and the luminous efficiency at each wavelength are shown at top. The products of the radiant power and luminous efficiency values at each wavelength give the quantities shown at bottom. Source: Reproduced with permission of the Lighting Research Center.

(in W) to specific amounts of luminous flux (in lm). By international convention, a light source that produces exactly 1 W of radiant power at 555 nm, the peak of the photopic luminous efficacy function, is defined to produce 683 lm, for a maximum photopic luminous efficacy of 683 lm/W. At different wavelengths, 1 W of radiant power would produce fewer than 683 lm. As an example, Figure 57.16 shows, for LED light sources having different peak wavelengths, how much relative radiant power is required

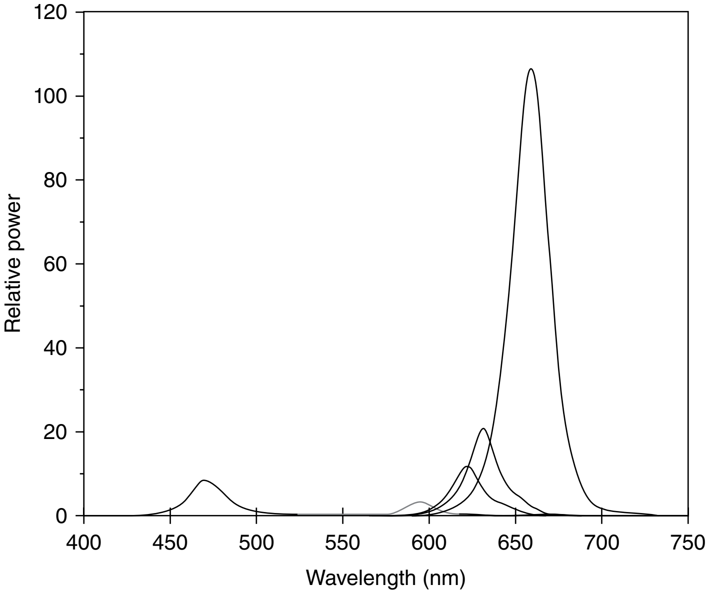


FIGURE 57.16 Relative radiant power needed for LED light sources with different peak wavelengths, in order to produce equivalent amounts of luminous flux.

to produce equivalent amounts of luminous flux. The further the spectral emission is from 555 nm, the more radiant power is required to achieve the same luminous flux.

Importantly, the value of luminous efficacy at 555 nm of 683 lm/W applies to both the photopic and scotopic luminous efficacy functions. Since the scotopic luminous efficiency function has a peak value not at 555 nm but at 507 nm, the maximum scotopic luminous efficacy is not 683 lm/W. Since the value of scotopic luminous efficiency at 555 nm is 0.402, the scotopic luminous efficacy at 507 nm, where the scotopic luminous efficiency is 1, is actually 683 lm/W ÷ 0.402, or 1700 lm/W. It is important to recognize that this difference in maximum luminous efficacies between the photopic and scotopic functions is merely a mathematical artifact and has no basis in physiology. Because the maximum scotopic luminous efficacy is 1700 lm/W, the value of k in the aforementioned equations should be 1700 lm/W, and not 683 lm/W, if a scotopic luminous flux quantity is to be calculated.

57.3.1.3 Photometric Measurement Spectral Considerations Photometric quantities such as illuminance and luminance are all related to luminous flux by geometrical relationships, such as the luminous flux density (illuminance), the luminous flux per solid angle (luminous intensity), and the density of luminous flux per solid angle (luminance). Most commonly used photometric instruments for measuring illuminance and luminance do not contain individual spectrally tuned elements at each wavelength but rather employ a material with a broad spectral response, such as selenium or silicon [1]. Silicon, for example, responds to radiant power in the UV and IR bands (Fig. 57.17)

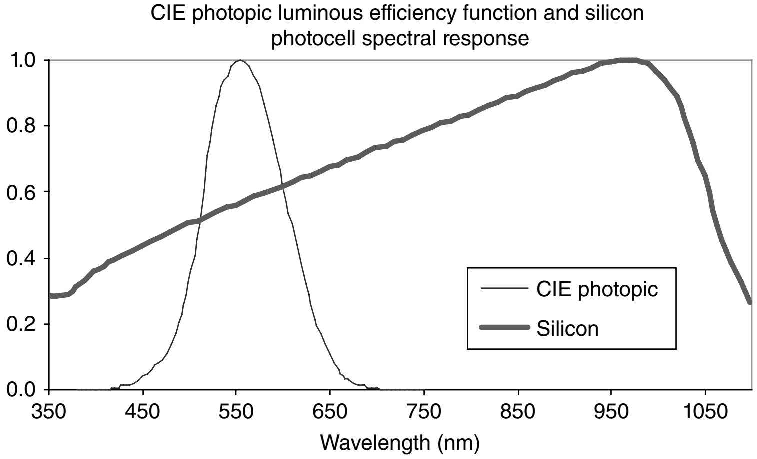


FIGURE 57.17 Spectral response of silicon (heavy curve) shown alongside the photopic luminous efficiency function commonly used to characterize light (lighter curve). Source: Reproduced with permission of the Lighting Research Center.

so a simple silicon cell alone would not be a suitable detector for a photometric instrument, since it would register a response to UV or IR radiation that would not be detected by the human eye as light.

Instead, such materials are fitted with filters that approximate the spectral response of the luminous efficiency function. No filter or combination of filters can provide a perfect match, but for broadband light sources producing “white” light, the mismatches largely cancel each other out with a reasonable estimate of the luminous quantity being measured. However, the user of such photometric instruments should be aware that they can yield quite large errors (Fig. 57.18) when measuring narrowband (colored) sources near the extremes of the visible spectral range, such as blue LED sources.

57.3.1.4 Luminous Intensity Measurement Geometrical Considerations When measuring the luminous intensity distribution from a light source such as an LED, one could in principle measure the illuminance from the source at various angular directions and use the inverse-square law to calculate the luminous intensity using Equation 57.1. However, it is necessary to remember that the inverse-square law is strictly applicable only to pure point sources with infinitesimally small sizes. Since LEDs and all other light sources have a finite size, several constraints should be met in order to ensure reasonable accuracy. One is the so-called five-times rule [1], which states that the measurement distance should be no less than five times the maximum dimension of the light source. It can be applied to diverging sources of light (without imaging optics such as lenses) and for diffuse emitters ensures that the error in estimating luminous intensity from the illuminance should be less than 0.5%.

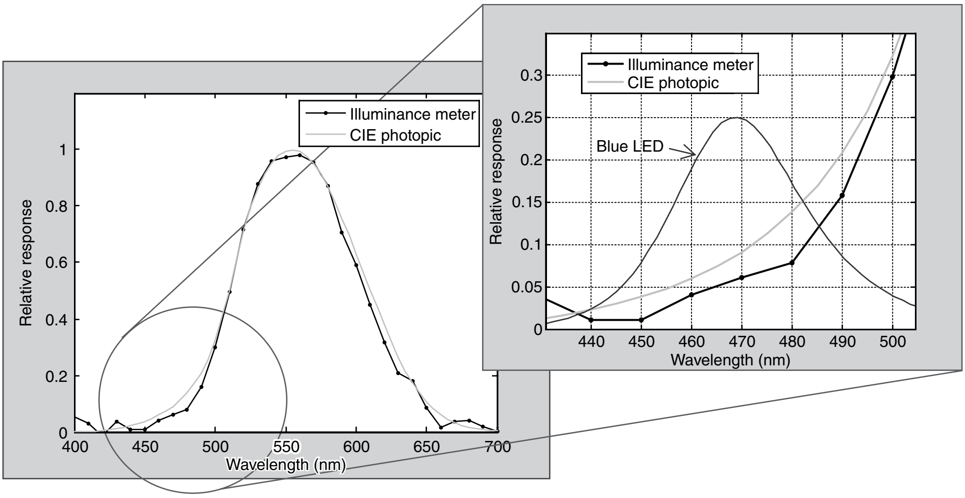


FIGURE 57.18 Left panel: a filter can provide a reasonably close match to $V(\lambda)$ for broadband light sources. Right panel: when used to measure a narrowband source such as the blue LED, an instrument using a filter can produce substantial errors because of mismatches between the spectral response and $V(\lambda)$ in a narrow wavelength range. Source: Reproduced with permission of the Lighting Research Center.

Many luminaires are outfitted with optics, such as the epoxy capsule of a 5-mm LED source, which can help collimate the light into a narrow beam. When optics are incorporated into a light source or luminaire, the test measurement distance must usually exceed the five-times rule by a substantial amount in order to ensure that the measured values will accurately represent the actual luminous intensity from the source being measured [1].

57.3.1.5 Luminous Flux Measurements Provided illuminance measurements at various angles from a light source are made at distances sufficient to employ the inverse-square law in order to accurately estimate the luminous intensity, these values can in turn be used to estimate the total luminous flux produced by the source. For example, consider a luminaire such as a downlight that produces a radially symmetric distribution downward (Fig. 57.19). An angle of 0° is defined as the angle directly below the luminaire. One could divide the angles from 0° to 90° into four zones with angular widths of 22.5°. If luminous intensity measurements are made at the midpoints of each of these zones (i.e., at 11.25°, 33.75°, 56.25°, and 78.75°), constants (denoted Z_n) relating the luminous intensity to the luminous flux within these zones around the luminaire can be calculated. These constants are specific to the width and number of zones, but for the simple example of the four zones in Figure 57.19, the luminous intensity within a zone is multiplied by the zonal constant for that zone to estimate the luminous flux within that zone. These four products would be summed to obtain the total luminous flux produced by the downlight luminaire.

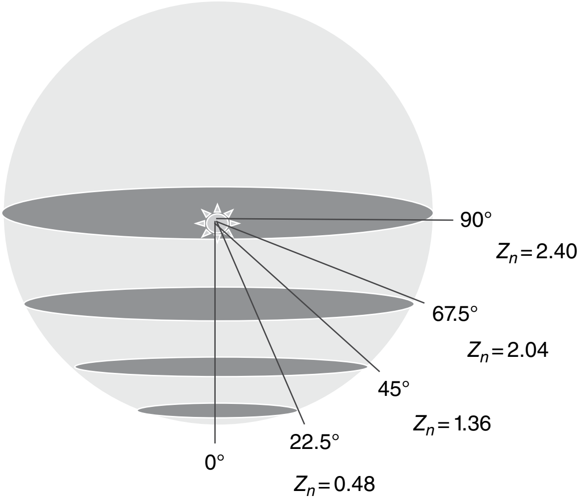


FIGURE 57.19 Illustration of the use of zonal constants (Z_n) to calculate the luminous flux produced by a light source or luminaire. Source: Reproduced with permission of the Lighting Research Center.

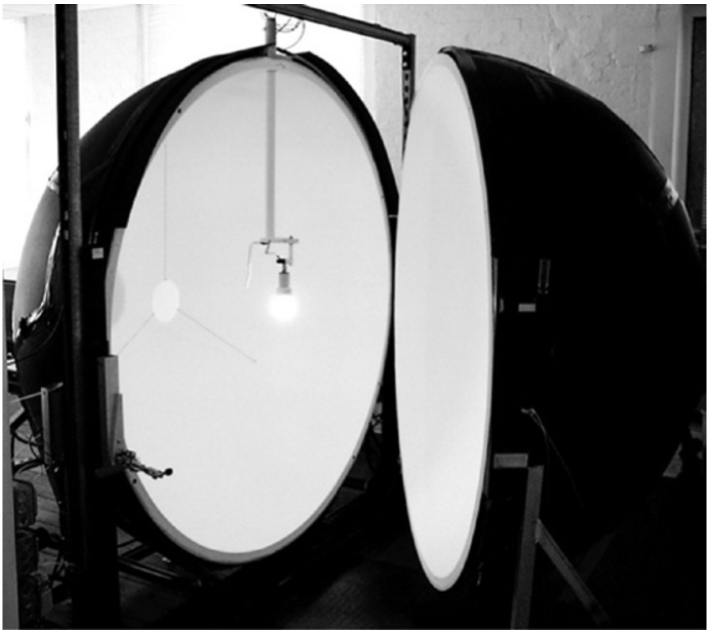


FIGURE 57.20 Photograph of an integrating sphere. Source: Reproduced with permission of the Lighting Research Center.

Another way to measure the luminous flux from a light source is through the use of an integrating sphere (Fig. 57.20). This is a large sphere painted matte white on the interior surface with a special high-reflectance paint. Because of multiple reflections from the white painted surface when a light source is placed inside the sphere, the illuminance on the sphere's interior surface can be estimated by the following equation:

$$E = \frac{\rho\Phi}{4\pi r^2 (1-\rho)} \tag{57.4}$$

where ρ is the reflectance of the white paint,
r is the radius of the sphere (in m),
Φ is the luminous flux from the source (in lm),
and E is the illuminance (in lx) on the interior surface of the sphere.

By rearranging terms in Equation 57.4, it is possible to solve for Φ (in lm) for a measured value of E (in lx). Of course, this method is subject to some limitations. The reflectance of paint used in spheres is not perfectly spectrally flat, so interreflections within the sphere create some spectral distortion that can affect the resulting measurement of illuminance. In addition, spheres and baffles used in the sphere will create more interreflections than would be caused within a completely empty sphere, which also influences the measured value. In conjunction with standard lamps of known calibration, however, the use of an integrating sphere can provide reasonably good accuracy for luminous flux measurements [1].

57.3.2 Colorimetry

The human retina contains three types of cone photoreceptors, which respond to light in different parts of the visible spectrum. Short-wavelength (S) cones are responsive to short wavelengths, medium-wavelength (M) cones to intermediate wavelengths, and long-wavelength (L) cones to long wavelengths (Fig. 57.21).

The three types of signals generated by these cone types offer the potential for color vision because the visual system can compare signals from different cone types and use this information to determine the likely wavelength band of a light source. For example, a blue LED producing most of its light near 460 nm will elicit a relatively strong S cone signal but weak M and L cone signals. A red LED near 630 nm would generate a strong L cone signal but a weaker M cone signal (and little S cone signal either). These combinations of cone inputs are denoted color channels. There are two color channels in the human visual system: a blue-yellow channel that compares signals from S cones to those from M and L cones, and a red-green channel that compares signals from M cones to those from S and L cones. If both channels are relatively balanced by the stimulation of all cone types, the resulting color appearance is white.

Interestingly, sources with different SPDs can produce the same relative blue-yellow and red-green channel signals and will appear to have the same color appearance. The illustration in Figure 57.22 shows two light sources that would both appear identical to the human eye [8].

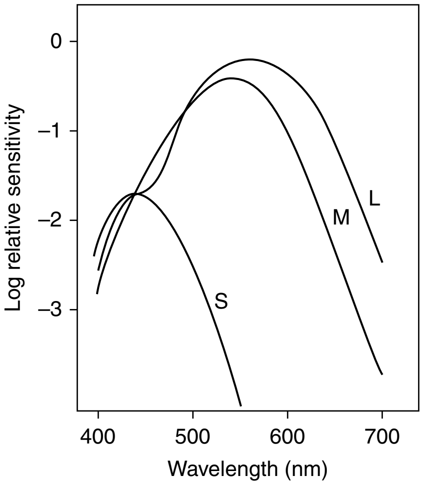


FIGURE 57.21 Spectral sensitivity of the three cone photoreceptor types in the human retina. Source: Reproduced with permission of the Lighting Research Center.

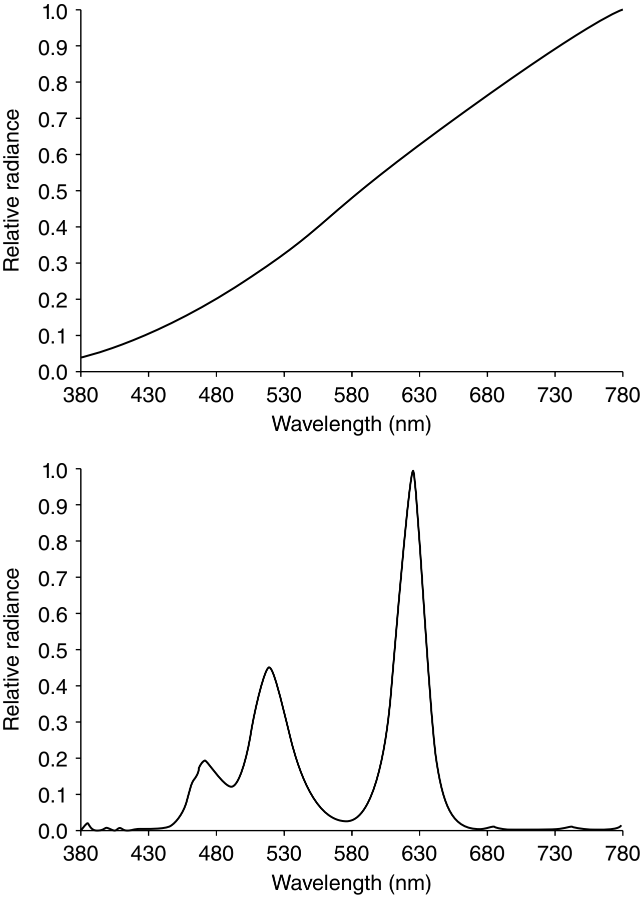


FIGURE 57.22 Spectral power distributions that will have the same color appearance to a human observer. The top panel represents the spectrum for an incandescent lamp; the bottom panel represents the spectrum for a light emitting diode source. Source: Reproduced with permission of the Lighting Research Center.

These types of SPDs are called metamers. Although the light from light sources would appear to be the same, it is important to understand that colored objects illuminated by them can and often would look very different. A very deep red object illuminated by the SPD in the left panel of Figure 57.22 would probably look red, whereas under the SPD in the right panel of Figure 57.22 it would probably appear much darker or even black. In general, light sources with SPDs having energy distributed throughout the visible spectrum are more likely to produce better color rendering (see Section 57.2.1.4 of this chapter for information on color rendering).

57.3.2.1 CIE Colorimetric System The lighting industry uses a system developed by the International Illumination Commission (CIE, for its French name, Commission Internationale de l'Éclairage) to communicate information about the color of light sources and of surfaces illuminated by different light sources. This system is not based on color appearance or the opponent channels described previously but rather on the concept of color matching: if two light sources or illuminated surfaces match each other, they can be said to be identical.

57.3.2.2 Color Matching Functions By using three primary light sources, a red, a green, and a blue source, it is theoretically possible to mix them to create an exact match to any of the individual visible wavelengths from 380 to 780 nm. Experiments with three such primary light sources, having wavelengths of 435.8 nm (blue [b]), 546.1 nm (green [g]), and 700.0 nm (red [r]), were conducted to demonstrate this [6]. Strictly, it is not possible to match all wavelengths with combinations of all three of these primaries. Wavelengths between 435.8 and 546.1 nm cannot be matched in this fashion. However, it is possible to add one of the primaries (i.e., the 700.0 nm red primary) to the individual wavelengths between 435.8 and 546.1 nm to create a match to a combination of the other two primary sources. Using arithmetical logic, adding a quantity to a term on one side of an equation can be canceled by subtracting the same quantity from the other side of the equation. The proportions of each primary needed to match all of the visible wavelengths (using negative proportions for the red primary between 435.8 and 546.1 nm) can be shown graphically; the resulting curves in Figure 57.23 are called color matching functions.

Color matching functions like the ones shown in Figure 57.23 can be generated for any three primary sources consisting of single wavelengths; in some wavelength regions, negative values would have to be used to produce the necessary color matches. To avoid the unpleasantness of color matching functions having negative values, it is possible to create color matching functions that have only positive values by using imaginary primaries, analogous to colors more saturated than any that can physically be generated. Figure 57.24 shows the three color matching functions that were standardized by the CIE in 1931 and used today to characterize color matches.

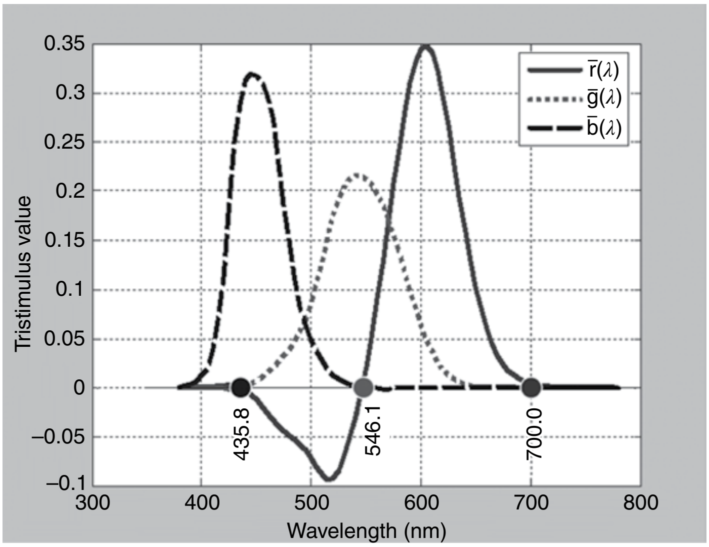


FIGURE 57.23 Color matching functions, showing the relative amounts (denoted tristimulus values) of three primary colors (b, 435.8 nm; g, 546.1 nm; and r, 700.0 nm) needed to match each wavelength in the visible spectrum.

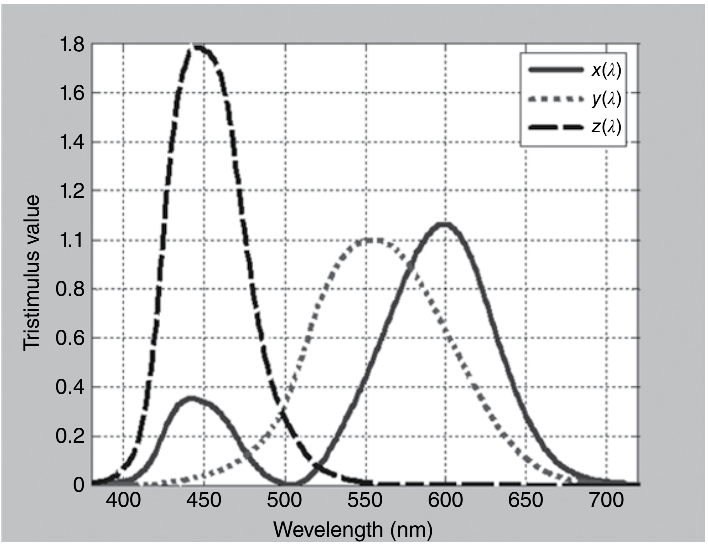


FIGURE 57.24 Color matching functions (x, y, z) based on imaginary primary color stimuli.

Each of the three functions, denoted by symbols $\overline{x}$, $\overline{y}$, and $\overline{z}$, can be used to calculate tristimulus values for each function, denoted by uppercase X, Y, and Z from an SPD in a similar manner used to calculate the luminous flux of an SPD using the photopic luminous efficiency function (Eq. 57.5a–c). In fact, the $\overline{y}$ color matching function was adjusted to match the photopic luminous efficiency function exactly:

$$\begin{aligned}
X &= \int P(\lambda)\bar{x}(\lambda)d\lambda \\
Y &= \int P(\lambda)\bar{y}(\lambda)d\lambda \\
Z &= \int P(\lambda)\bar{z}(\lambda)d\lambda \\
x &= \frac{X}{X+Y+Z} \\
y &= \frac{Y}{X+Y+Z}
\end{aligned} \tag{57.5a–e}$$

In Equation 57.5a–c, *X*, *Y*, and *Z* are the tristimulus values corresponding to the $\bar{x}$, $\bar{y}$, and $\bar{z}$ color matching functions, respectively; *P*(*λ*) is the SPD of the light source. The lowercase symbols *x* and *y* are the chromaticity coordinates for the SPD.

57.3.2.3 Chromaticity Coordinates From the tristimulus values *X*, *Y*, and *Z*, the previous equations can be used to calculate the chromaticity coordinates of the SPD in question. These coordinates, denoted by lowercase *x*, *y*, and *z*, indicate the relative proportion of each tristimulus value to the sum of all three values (Eq. 4.5d and e). By definition, therefore, the sum of the *x*, *y*, and *z* chromaticity coordinates is always 1. Because of this, once *x* and *y* are known, the value of *z* is fixed at $1-x-y$, and it provides no additional information. For this reason the *z* chromaticity coordinate is almost never used.

The chromaticity coordinates for a particular SPD can be plotted on a set of rectangular coordinates with *x* and *y* axes. Since it is possible to calculate the chromaticity coordinates of the individual wavelengths from 380 to 780 nm, many chromaticity diagrams show these values, and they are often referred to as the spectrum locus (Fig. 57.25). The spectrum locus forms an inverted "U" shape in the chromaticity diagram. Since it is possible to mix very short wavelengths (i.e., near 380 nm) with very long wavelengths (i.e., near 780 nm), a straight line can join these extreme wavelengths where such combinations will appear as various shades of purple, and the line is called the purple boundary.

The chromaticity coordinates of any SPD that can be produced or imagined will always have chromaticity coordinates within the area bounded by the spectrum locus and the purple boundary. In addition, mixing the light from any two sources of light will result in a mixture that has chromaticity coordinates along the line connecting the two components' chromaticity coordinates. Observing the illustration in Figure 57.25, it can be seen that it is possible to find some mixture of two wavelengths, such as 490 and 570 nm, that will produce a match to a particular mixture of two other wavelengths, such as 500 and 600 nm, because the line segments connecting each pair of wavelengths would intersect.

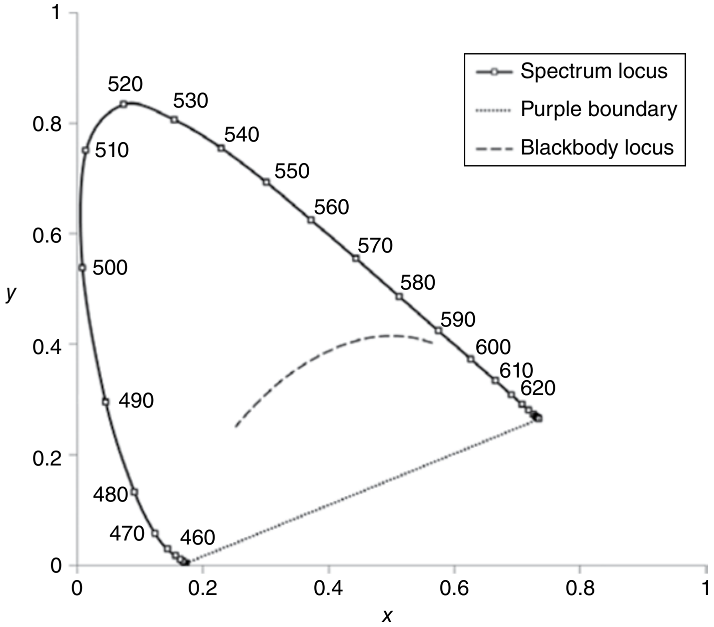


FIGURE 57.25 Chromaticity diagram showing (x, y) coordinates for several wavelengths along the spectrum locus. Also shown are the purple boundary and the blackbody locus (for a range of color temperatures between 1,667 and 25,000 K).

57.3.2.4 Color Metrics Based on Chromaticity Coordinates The chromaticity diagram and its system of chromaticity coordinates can provide many insights regarding the color properties of different light sources, and two useful color metrics can be determined from the chromaticity coordinates. One is the dominant wavelength. Selecting the chromaticity of a particular white light source as a reference (commonly, either the chromaticity of an incandescent source or, as shown in Fig. 57.26, the chromaticity of a spectrum having equal radiant power across all visible wavelengths, called an equal-energy SPD), it is possible to extend a line from the chromaticity coordinates of the reference to those of the light source in question until it reaches either the spectrum locus or the purple boundary. If the line reaches the purple boundary, then the light source in question has no dominant wavelength, but if it intersects the spectrum locus, the wavelength corresponding to the intersection point is the dominant wavelength.

Another metric is called the complementary wavelength, and it is determined by extending a line from the chromaticity coordinates of the light source in question through those of the reference source, until it intersects the spectrum locus or the purple boundary. If the line intersects the purple boundary, the source in question has no complementary wavelength, and if it intersects the spectrum locus, the complementary wavelength is determined as for the dominant wavelength. Some light sources will have both a dominant wavelength and a complementary wavelength, while others will have only one or the other.

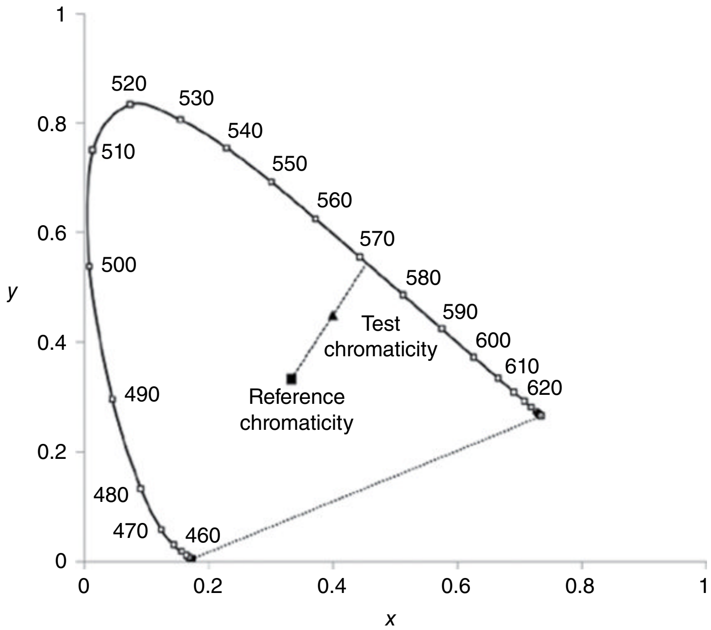


FIGURE 57.26 Illustration of the graphical determination of dominant wavelength. For the example shown here, the line segment intersecting the reference chromaticity (equal-energy SPD) and the test chromaticity intersects the spectrum locus at a dominant wavelength slightly longer than 570 nm.

Related to the dominant wavelength is another metric called the excitation purity. Using the same line as that used to determine dominant wavelength, it is determined by taking the length of the line between the reference source's chromaticity coordinates and those of the source in question and dividing that length by the length of the line between the reference source's chromaticity coordinates and the spectrum locus or purple boundary. It is generally expressed as a percentage. In Figure 57.26, the excitation purity of the test chromaticity source (using the shown reference source) is about 50% because the test chromaticity lies about halfway between the reference chromaticity and the spectrum locus.

57.3.2.5 Chromaticity Discrimination It has been stated previously that two light sources that have identical chromaticity coordinates will produce matches to the human eye. It should seem reasonable to suppose that if two light sources have chromaticity coordinates that do not match exactly, but are very close to one another, they would be judged to match most of the time. Since hardly any two light sources will ever match exactly in terms of their chromaticity, not even two lamps of the exact same type and make, because of factors such as variations in manufacturing, it is reasonable to want to know how different chromaticity coordinates can be before they will be reliably judged as different.

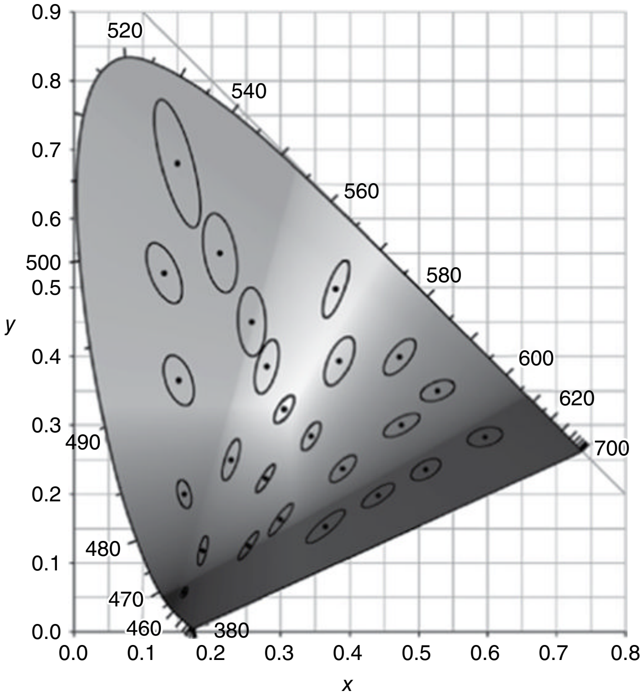


FIGURE 57.27 MacAdam [9] ellipses for various chromaticities, increased in size by a factor of 10.

So-called MacAdam ellipses [9] have been determined experimentally for different regions within the chromaticity diagram. These ellipses indicate, for a given reference source having certain chromaticity coordinates, the boundary indicating the chromaticity coordinates of light sources that most people would judge to match the appearance of the reference source. They are shown, 10 times larger than their actual size to make them easier to see, in Figure 57.27.

MacAdam ellipses are used in the specification of chromaticity tolerances for commercial lamps such as fluorescent lamps [8]. Lamps having various nominal CCTs must fall within four-step MacAdam ellipses having radii four times larger than the original MacAdam ellipses (and just under half the size of the 10-step ellipses shown in Fig. 57.27).

57.3.2.6 Uniform Chromaticity Diagram If the CIE 1931 chromaticity diagram were perceptually uniform, equal vector distances in the diagram would correspond to equal perceived color differences. The varied shapes and sizes of the MacAdam ellipses demonstrate that this is not the case. If the chromaticity diagram were perceptually uniform, the sizes of the ellipses would all be the same, and moreover, the ellipses would instead be perfect circles. Several attempts to develop more uniform chromaticity diagrams have been made, with one of the more successful being the CIE 1976 uniform

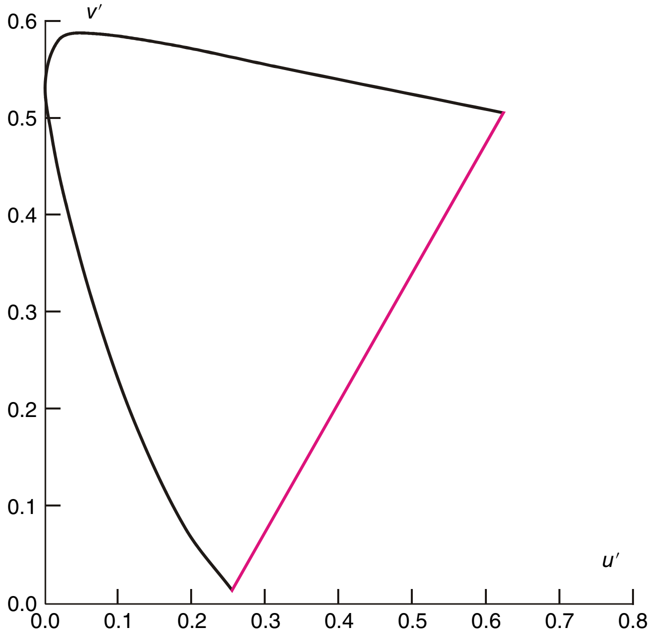


FIGURE 57.28 CIE 1976 uniform chromaticity diagram.

chromaticity diagram, using coordinates denoted u' and v' in place of x and y. The 1976 uniform chromaticity diagram was developed through a linear transformation of the x and y coordinates in the 1931 diagram:

$$u' = \frac{4x}{(-2x + 12y + 3)} \tag{57.6a}$$

$$v' = \frac{9y}{(-2x + 12y + 3)} \tag{57.6b}$$

While the 1976 CIE chromaticity diagram is more perceptually uniform than the 1931 diagram, it is not perfect in that MacAdam ellipses are not transformed into perfectly equally sized circles, but it is a marked improvement over the 1931 diagram. Nonetheless, the 1931 diagram is much more commonly used in the lighting industry to characterize the chromaticity of different light sources, and the 1976 diagram is mainly used in color research and in specific applications where color discrimination might be extremely important to predict, such as some printing processes (Fig. 57.28).

57.3.2.7 Munsell Color System Another system is used in many architectural applications [1] to indicate the color of a surface, known as the Munsell color system. Unlike the CIE chromaticity diagram, which provides a quantitative framework for characterizing light source and surface color, the Munsell system is a color order system based on three attributes of color known as the hue, value, and chroma.

The hue is the quality identified with the color name, such as blue, green, yellow, red, or violet. Certain color combinations are also allowed, such as blue–green or yellow–red. The value indicates the lightness of the color with a value of 0 corresponding to perfect black, and a value of 10 corresponding to perfect white. The chroma is an indication of how saturated a color appears. Different hue and value combinations can have different maximum chroma values. As an example, a color with a very high value will appear like a pastel and cannot be highly saturated.

Steps along the hue, chroma, and value dimensions are judged to be approximately equal. Sets of Munsell chips containing finely gradated color chip samples along the hue, value, and chroma dimensions can be used by architects and interior designers to specify paint and finish colors in built spaces. Munsell chips are also used in some research studies as stimuli of known, repeatable colors.

57.4 INSTRUMENTATION

57.4.1 Illuminance Meters

One of the most commonly used and least expensive types of photometric instruments is the illuminance meter (Fig. 57.29). Many illuminance meters are handheld devices consisting of a sensor element, usually a silicon or selenium detector with a photopic filter having spectral characteristics similar to those illustrated in Figure 57.18. The readout displays the measured illuminance value (lx or fc). When measuring illuminance, the sensor element should be placed as close as possible and parallel to the surface on which the illuminance is being measured. The sensor element's spatial sensitivity corresponds to a cosine distribution; incident light from directly above the sensor is weighted fully, while light incident from an angle is proportional to the cosine of that angle.

Care should be taken to avoid producing shadows on the sensor element when making measurements and to avoid the possibility of reflected light from the user's clothing that can increase the illuminance at the sensor element's location. A quality illuminance meter can be purchased for between US$100 and US$300.

57.4.2 Luminance Meters

A number of portable luminance meters are commercially available (i.e., Fig. 57.30). Unlike illuminance meters, luminance meters require imaging optics to capture light emitted or reflected from a relatively narrow entrance angle (1° and 0.33° entrance angles are representative values). Luminance meters have view finders to allow the user to position the aperture over the part of the visual scene of interest. Spectral correction is usually achieved by using a filter over a silicon sensor element.

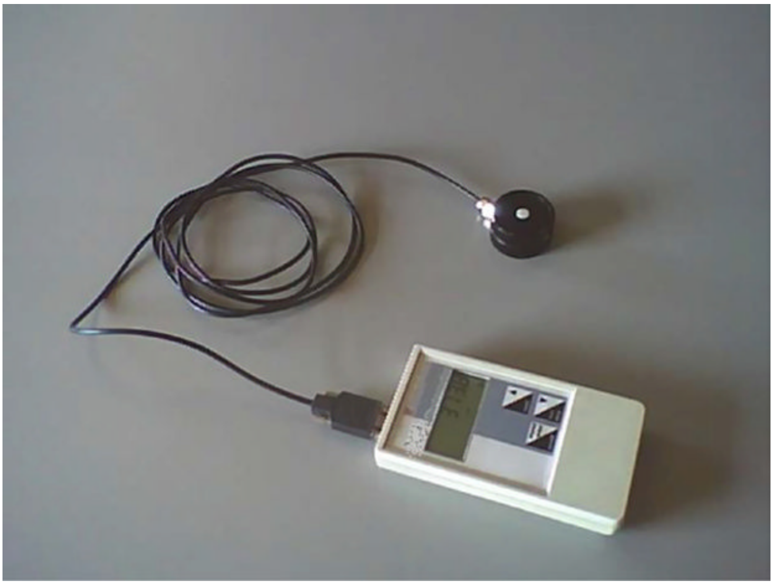


FIGURE 57.29 A portable illuminance meter with a detachable sensor element.

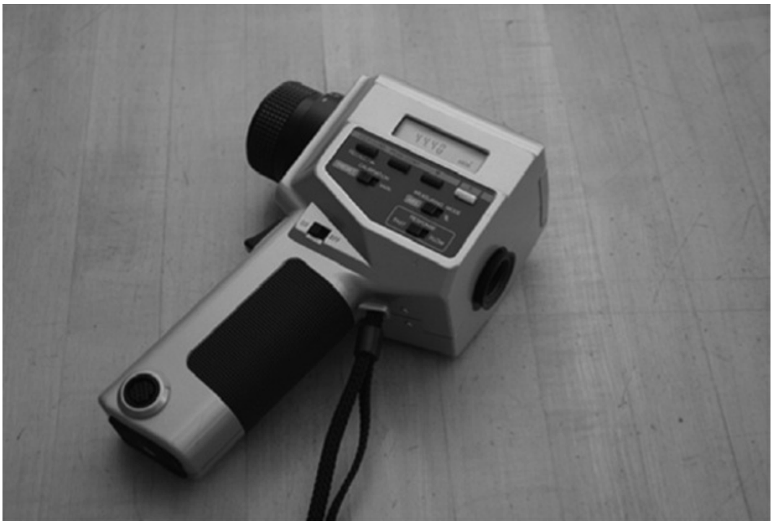


FIGURE 57.30 Portable luminance meter. Source: Reproduced with permission of the Lighting Research Center.

When using a luminance meter, it is important that the part of the visual field that is being measured is seen in clear focus through the view finder to minimize the effects of stray light from influencing the measurement. Because of the additional optics required by luminance meters, they generally have substantially higher costs than illuminance meters, on the order of US$1000 to US$3000.

57.4.3 Spectroradiometers

Most illuminance and luminance meters provide only photometric measurements without regard for the colorimetric properties of the sources or surfaces being measured because of the use of broadband detectors and filters that integrate responses according to the appropriate luminous efficiency function (usually, the photopic function). For this reason they cannot provide information about the color properties of the lighting conditions they are used to measure. Spectroradiometers are used to measure the spectral power at various wavelengths across the visible spectrum. They incorporate collection optics to receive the radiant power from the object being measured, usually in a manner similar to luminance meters. A monochromator with an element such as a diffraction grating or a prism disperses the light of varying wavelengths onto an array of detectors. The circuitry from the detectors processes their signals and stores the results for each wavelength band so that the SPD curve can be generated, displayed, and stored. Most spectroradiometers will also have additional processing software to calculate quantities such as the CCT, CRI, and chromaticity coordinates. Because of their complexity relative to illuminance and luminance meters, spectroradiometers are generally quite expensive, ranging from US$5,000 to US$50,000.

REFERENCES

1. Rea MS, ed. 2000. *IESNA Lighting Handbook: Reference and Application*, 9th ed. New York: Illuminating Engineering Society.
2. Commission Internationale de l'Éclairage. 1978. *Light as a True Visual Quantity*, No. 41. Paris: Commission Internationale de l'Éclairage.
3. Bullough JD. 2005. Research matters: What's cooler than cool? Warm! *Lighting Design and Application* 35(2): 12–14.
4. Rea MS, Freyssinier JP. 2010. *ASSIST Recommends: Recommendations for Specifying Color Properties of Light Sources for Retail Merchandising*. Troy, NY: Rensselaer Polytechnic Institute. Accessed on February 28, 2014 at http://www.lrc.rpi.edu/programs/solidstate/assist/recommends/lightcolor.asp.
5. Bullough JD, Gu Y, Narendran N, Taylor J. 2005. *ASSIST Recommends: LED Life for General Lighting: Definition of Life*. Troy, NY: Rensselaer Polytechnic Institute. Accessed on February 28, 2014 at http://www.lrc.rpi.edu/programs/solidstate/assist/recommends/ledlife.asp.
6. Wyszecki G, Stiles WS. 1982. *Color Science*, 2nd ed. New York: Wiley-Interscience.
7. Commission Internationale de l'Éclairage. 2010. *Recommended System for Mesopic Photometry Based on Visual Performance*, No. 191. Vienna: Commission Internationale de l'Éclairage.
8. Rea MS, Deng L, Wolsey R. 2004. *Lighting Answers: Light Sources and Color*. Troy, NY: Rensselaer Polytechnic Institute. Accessed on February 28, 2014 at http://www.lrc.rpi.edu/nlpip/publicationDetails.asp?id=901.
9. MacAdam DL. 1942. Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America* 32(5): 247–273.

58

THE DETECTION AND MEASUREMENT OF IONIZING RADIATION

Clair J. Sullivan

Department of Nuclear, Plasma, and Radiological Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

58.1 INTRODUCTION

Detection of ionizing radiation can be traced by to 1895 when W.C. Rontgen made the first medical radiograph using X-rays. Since that time, many high-quality textbooks have been written on the subject [1–7]. We will not attempt to recreate those works here but rather to convey a working familiarity with the concepts behind how these detectors work and are used in everyday measurements. While there are many types of radiation such as radio-frequency emissions, infrared, etc., this chapter seeks to address specifically the detection of ionizing radiation.

The detection and measurement of ionizing radiation can be described in one sentence: *it is all about converting the incident radiation to charge (or light and then charge) and measuring that charge, whose magnitude is ideally proportional to the type, energy, and/or quantity of incident quanta.* While this statement on its surface is simple, there is much subtlety conveyed. It is important to understand what type of radiation is incident, how that type of radiation interacts with matter and ideally creates charge, and then how that charge is read out of the system and analyzed. Therefore, to truly understand radiation detection, it is necessary to understand some basic physics, statistics, materials science, and electronics design. In this chapter, we hope to provide a working knowledge of the basics necessary to understand each of these concepts.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

58.2 COMMON INTERACTIONS OF IONIZING RADIATION

Ionizing radiation comes in two major categories, charged particulate radiation and uncharged radiation, with multiple different types of each. Examples of charge particulate radiation include alpha particles (which is just the nucleus of a helium atom containing two protons and two neutrons), beta particles (which is either an electron or a positron), protons, fission fragments, etc. Uncharged ionizing radiation includes photons such as gamma rays and X-rays as well as neutrons.

Most ionizing radiation is created through the radioactive decay of an unstable parent isotope. This decay occurs within some statistical time distribution and results in a daughter nucleus and the ionizing radiation particle. How quickly a given isotope decay determines is so-called radioactivity, which is mathematically described as

$$\frac{dN}{dt} = -\lambda t$$

where N is the number of nuclei, t is time, and λ is called the decay constant, which is related to the isotope's half life, $t_{1/2}$, through the relation

$$\lambda = \frac{\ln 2}{t_{1/2}}.$$

There are several units that are assigned to radioactivity, the most common being the Curie (Ci) and the becquerel (Bq). 1 Bq is equivalent to 1 decay per second or 2.703×10^{-11} Ci.

In addition to radioactive decay, ionizing radiation can be produced through other processes such as spontaneous fission, excitation resulting from an energizing source impacting a target, etc. These processes are beyond the scope of this chapter but are discussed in greater detail in many other texts [3, 7].

58.2.1 Radiation Interactions

In order to detect ionizing radiation, we must first get it to interact with a detecting medium so charge can be produced. How radiation interacts with matter is a function of the type of radiation, its energy, and what it is interacting with. The probability of an incident particle interacting with the medium can be statistically described and has been measured for many different types of materials. The measurement process is reasonably simple: a calibrated source of radiation is directed through a material of thickness x, as shown in Figure 58.1. The number of particles that make it through the material can be described by the equation

$$I = I_0 e^{-\mu x}$$

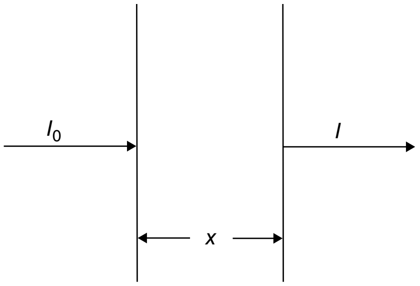


FIGURE 58.1 Example of source attenuation of initial intensity, I_0, to measured intensity, I, through an absorber of thickness x.

where μ is called the linear attenuation coefficient. It is related to the mean free path (mfp), the average distance before an interaction takes place in an absorber, as

$$\text{mfp} = \frac{1}{\mu}.$$

It is important to note that μ, and hence the mfp, takes into account all types of interactions and is also a function of the type of radiation, energy, and what it is interacting with. For example, gamma rays can interact in one of three different ways depending on their energy: the photoelectric effect, Compton scatter, or pair production. So μ is a linear combination of the interaction probabilities of each individual type of interaction.

Another important parameter for radiation interaction is how much energy, E, is deposited as a function of path length, x. This is called the linear stopping power, which is differentially defined as

$$S = -\frac{dE}{dx}.$$

Particles with higher stopping power, such as alpha particles, deposit more energy per interaction, meaning that they lose energy quickly within an absorber. Ideally, it is this energy loss that creates the charge that will be the basis for detection.

58.3 THE MEASUREMENT OF CHARGE

There are three predominant ways that charge can be measured from a detector: the measurement of current across a resistor, voltage across a capacitor, or the mean square voltage. Generally speaking, the charge created in a radiation detector is small, necessitating a certain emphasis on charge amplification and electronics. Prior to a discussion of these various modes or the actual measurement of current or voltage, it is necessary to briefly discuss the statistics associated with radiation detection.

58.3.1 Counting Statistics

There are three predominant probability distributions that apply to radiation detection. The first and most basic of these is the binomial distribution, which is given by

$$P(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

where $P(x)$ is the normalized probability of counting x successes, p is the probability of success for a single event, and n is the number of trials. It is clear then that, since x is a discrete random variable, $P(x)$, called the probability mass function (PMF), is the probability of obtaining exactly x discrete successes. This should not be confused with a cumulative probability distribution (CDF) but can lead to the CDF by considering what happens when x becomes infinitely small. Since in real measurement applications it is not possible to have infinite precision, the PMF is sufficient for our purposes.

Using this equation, it is possible to determine the mean of the distribution, μ, which can be found to be

$$\mu = \sum_{x=0}^{n} xP(x) = pn.$$

This fact will be useful in further simplifications described in the following. Additionally, the predicted variance of the distribution can be calculated to be

$$\sigma^2 = \sum_{x=0}^{n} (x-\mu)^2 P(x) = \mu(1-p).$$

To calculate $P(x)$, μ, and σ^2, we must understand how a success, p, is defined. There are many trivial examples for probability students that involve coin flip (where success could be defined as the number of times the experimenter gets heads) or rolling a six-sided die (where success might be considered as rolling a three). In the case of coin tosses, the probability of success is obviously one half whereas for rolling the dice it is one sixth. For radiation detection, the probability of success is considered to be the probability of an individual, specific nucleus decaying. This is given by

$$p = 1 - e^{-\lambda t}$$

where λ is the decay constant for the isotope, which is generally known and is very small.

An important simplification is possible when p is small and n is large through the use of the so-called Poisson limit theorem. The product of the two approach a limiting value, or $np \to \mu$, which can be used to rewrite the binomial distribution as

$$P(x) \to \frac{\mu^x}{x!} e^{-\mu},$$

which is known as the Poissonian distribution. Since we know that p is small for radioactive decay, we can default to using the Poissonian distribution for radiation counting statistics. Like for the binomial distribution, the mean is given by $\mu = pn$. However, the predicted variance is slightly simpler:

$$\sigma^2 = \mu.$$

Further simplification is possible if we consider what happens when μ is large and take advantage of the fact that the predicted variance is equal to the mean. (Most practitioners consider "large" to be around 25–30 or when $n(1-p) \geq 5$.) In this case, we can rewrite the Poissonian distribution as the well-known Gaussian distribution given by

$$P(x) = \frac{1}{\sqrt{2\pi\mu}} \exp\left(-\frac{(x-\mu)^2}{2\mu}\right)$$

or

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Examples of the binomial, Poissonian, and Gaussian distributions are shown in Figure 58.2.

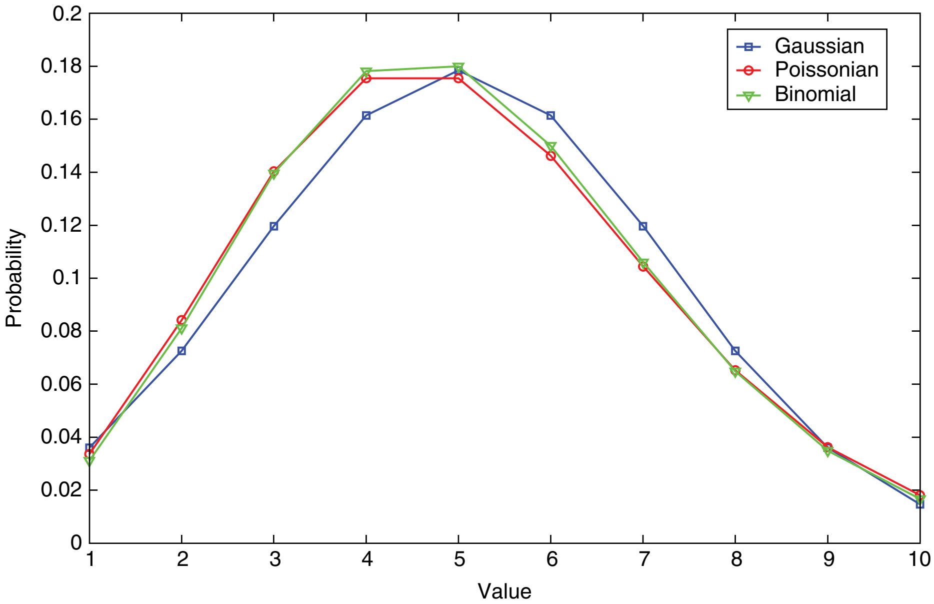


FIGURE 58.2 Examples of the binomial, Poissonian, and Gaussian distributions around a mean value of 5. Note the peak value differs based on distribution.

58.3.2 The Two Measurement Modalities

The deposition of energy from incident radiation in a detector results in the creation of charge, which must be moved through the application of an external field and then collected. This can be done as a function of time, observing individual radiation interaction events or "counts" as they are created in the detector, as a time-integrated average whereby averages of that moving charge are measured in some time window. The former case, referred to as "pulse mode operation," measures the movement of charge within the detector on an event-by-event basis. The detector is connected to some readout electronics as shown in Figure 58.3. Assuming the time it takes for the charge to travel is significantly smaller than the RC time constant of this circuit, a signal is created representing the change in signal voltage as a function of time, $V(t)$, whose maximum is given by

$$V_{\text{max}} = \frac{Q}{C},$$

where Q is the amount of charge created per event in the detector. This can be calculated as

$$Q = \frac{E}{W} e_0$$

where W is the mean ionization energy for the detecting material, E is the energy deposited in the interaction, and e_0 is the elementary charge of an electron (1.6×10^{-19} C). Thus it is clear that a measurement of V_{max} is directly proportional to the energy deposited in the detector (which is hopefully proportional to the energy of the source). Because of this fact, pulse mode operation is the most common mode of operation for radiation detection.

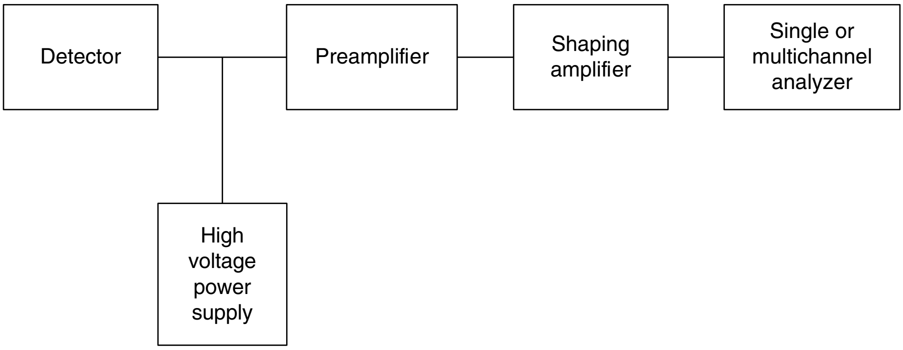


FIGURE 58.3 Standard setup for radiation measurements. This setup can be used either to count radiation events through the single channel analyzer or to collect spectra through the multichannel analyzer.

Another common method of measurement involves measuring the integrated current that is the charge moving under the application of the external field. This type of measurement is most common when an average event rate, *r*, must be known, such as is common in radiation dosimetry and health physics. In this case, the current is measured and averaged over some integration time. Then, *r* can be calculated by

$$r = \frac{\bar{I}}{Q} = \bar{I}\,\frac{W}{Ee_0}$$

where $\bar{I}$ is the average current generated through the movement of charge of many incident quanta averaged over a period of time. While this measurement is not done on an event-by-event basis, the electronics necessary to measure a current signal tend to be simpler and are perfectly suitable when rate-based information (such as dose rate) is all that the user requires.

58.4 MAJOR TYPES OF DETECTORS

As previously stated, the goal of any radiation detector is to convert the incident radiation to charge through the interaction within some detecting medium. The three most common types of detecting media are gases, scintillating materials (solid, liquid, or gas), and semiconductors. (It should be noted that scintillators do not directly convert the incident radiation to charge, but they convert it to light that is later converted through charge via optical readout techniques.) Each of these detector types shall be described in detail in the following.

58.4.1 Gas Detectors

The ionization of a gas atom or molecule is the basis for the simplest type of radiation detector. The general concept behind the gas detector is that the radiation enters a volume of gas and interacts through excitation and ionization. Excitation does not directly result in the formation of charge (although charge can be created indirectly through the deexcitation process, which can sometimes release secondary radiation such as an X-ray that is later collected), so we will focus on the ionization process. In ionization, the incident ionizing radiation interacts with an atomic electron, giving it enough energy to escape the atom and become a free particle. The result of ionization is the creation of both a free electron and an ion of the gas molecule—an electron-ion pair. The minimum energy required to liberate an electron is called the ionization energy and is a function of the element as well as which shell the electron originates (outer electrons have lower ionization potentials than those in the inner shells). The ionization potentials for some key elements are provided in Table 58.1 and a photo of several commercial gas detectors is provided in Figure 58.4.

TABLE 58.1 Energies Required to Ionize the Outer Electron Shell of Select Elements[a]

Element	First Ionization Energy (eV)
Ge	7.899
Si	8.151
Xe	12.13
Ar	15.759
Ne	21.564
He	24.587

[a] From Ref. 6.

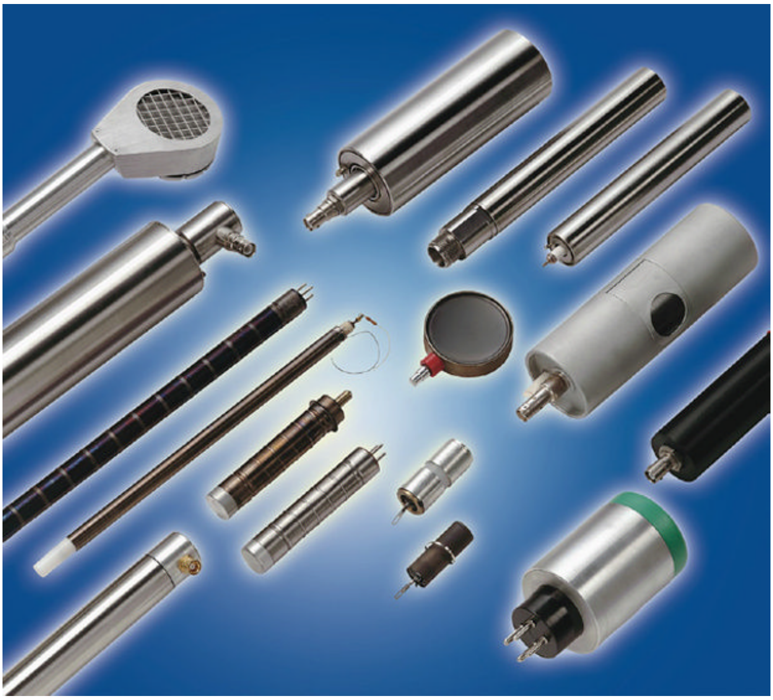


FIGURE 58.4 Examples of a variety of gas detectors of different sizes and configurations. Source: Reproduced with permission of Saint-Gobain Crystals.

In general, the radiation will likely undergo several ionization and excitation interactions along its path before it is completely absorbed. The result is that several electron-ion pairs will be created. The average amount of total charge created is determined by

$$Q = \frac{E}{W} e_0$$

TABLE 58.2 Average Values of the Energy Required to Create an Ion Pair in Some Select Gases[a]

Gas	Average *W* Value (eV/Ion Pair)
Xe	21.5
Ar	26.2
CH_4	28.2
O_2	31.5
Air	34.5
N_2	35.6
Ne	36.2
He	42.0

[a] From Ref. 3.

where E and W, the average energy required to create an electron-ion pair, are in units of energy (usually the electron volt, eV, for radiation detection). Some values of W for common fill gases are provided in Table 58.2.

It is also important to understand the statistics of the creation of this charge. In general, we assume Poissonian statistics to govern the variation in the number of charge pairs created, N. If this was truly a Poissonian process, then we would expect the variance in the number of charge pairs created, σ_N^2, to just be N. However, by simply assuming Poissonian statistics, we have inherently assumed that the processes that create individual charge pairs are independent. In reality, this is not quite true, so a correction factor called the Fano factor has been introduced to account for this variability:

$$\sigma_N^2 = FN$$

where F is the so-called Fano factor, which is usually less than 1 for most gases. The variance in the number of charge pairs created relates to the energy resolution of the detector, so keeping this value small is desirable.

Once the charge Q is created within the gas, all that remains is to collect it. Like charge-based radiation detectors (i.e., not scintillators), the charge pairs are separated through the creation of an electric field. This is usually done by creating two electrodes separated by both a spatial and potential difference. Detectors can be designed to use electrodes in rectilinear, cylindrical, or spherical geometries, as shown in Figure 58.5. The magnitude of the electric field, $|\vec{E}|$, is a function of this geometry and the separation of the two electrodes, d. For rectilinear coordinates, the relationship is simply

$$|\vec{E}| = \frac{V}{d}$$

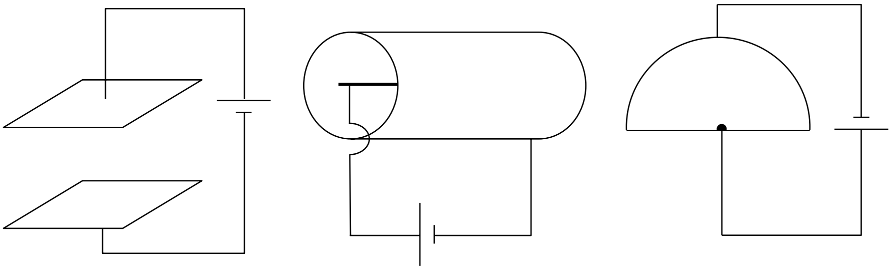


FIGURE 58.5 Examples of planar (left), cylindrical (center), and hemispherical (right) detector geometries with the anode as the high potential electrode and the cathode as the low potential electrode.

TABLE 58.3 Sample Values of Electron Mobilities for a Variety of Different Gases[a]

	E/p (V m/N)	
Gas	0.6	1.5
He	8.5	15
Ar	3.9	6.2
N_2	7	14
Air	11	17
CH_4	96	117
Ne	12	26

[a] From Ref. 3.

where V is the potential difference between the two electrodes. The electrode at the higher potential is called the anode whereas that of the lower potential is referred to as the cathode. In cylindrical coordinates, it can be shown that

$$\left|\overrightarrow{E(r)}\right| = \frac{V}{r\ln(b/a)}$$

where a is the radius of the inner electrode and b is the radius of the outer electrode.

It is also relevant to consider how fast the electron and ion move in the presence of the electric field. The velocity of each charged particle can be determined by

$$\overline{v_{\mathrm{e,i}}} = \frac{\mu_{\mathrm{e,i}}\vec{E}}{p}$$

where $\mu_{\mathrm{e,i}}$ is the mobility of the electron and ion, respectively (of units V m/N or m^2 atm/V s), and p is the gas pressure. Sample values of mobility are provided in Table 58.3. The mobility of the electron is around three orders of magnitude larger

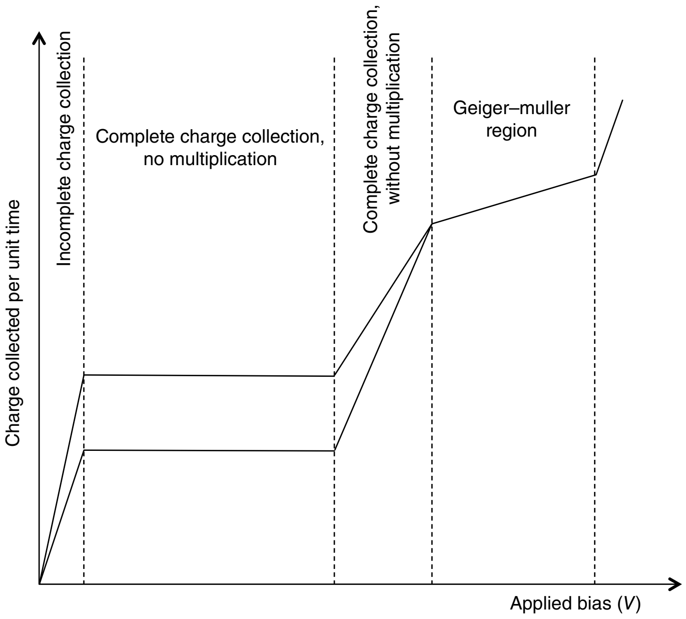


FIGURE 58.6 The four regions of gas detector operation shown for two different energy depositions. Complete charge collection without multiplication corresponds to ionization chamber operation. The next region where multiplication is added is called the proportional region. The final region, where there is no difference in collected charge with energy deposition, is the Geiger–Muller region.

than that of the ions due to the mass difference between the two particles. However, it puts significant constraints on the operation of a gas detector, if full signal generation requires the complete collection of both the electrons and the ions. Collection of the ions will take time, and this means that the count rates that can be handled will be limited. Additionally, the ions easily recombine during this time, resulting in incomplete charge collection. In the following sections, we shall present a method for overcoming this limitation.

Gas detectors can be operated either in current mode or in pulse mode. When in steady state, average values are to be measured, such as average dose rate, and current mode is typically employed due to its simplicity. However, gas detectors can also operate in pulse mode and used to perform spectroscopy. These techniques shall be discussed in detail in the following.

Beyond just variations in the overall geometry of the detector, we can describe gas detectors as being of one of three different types depending on their internal electric fields: ionization chambers, proportional counters, and Geiger–Muller (GM) tubes, as will be described in the following sections. Since the electric field is established by the applied bias, *V*, it is possible to measure how the number of collected ion pairs changes with *V*, which is shown in Figure 58.6. When examining this figure, there are four key

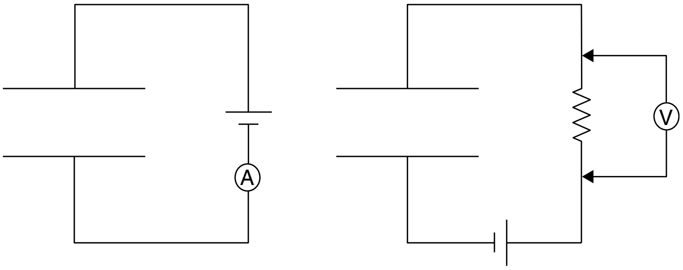


FIGURE 58.7 Circuit diagram for radiation detector operating in current (left) versus pulse (right) modes.

regimes to consider. In the first regime, the number of charge carriers collected increases with applied bias. This corresponds to the fact that the bias is increasing so a larger and larger volume of the detector becomes active. (The inactive regions are where the charge pairs are created too far from their electrodes, resulting in recombination before the charges can be collected.) The second region, which lies soundly on a plateau, is where the electric field is sufficient to collect all charges created within the detector. This is where ionization chambers operate. The third region occurs when the electric field becomes large enough to allow charge multiplication. Called the proportional region, this corresponds to the place where the output signal amplitude is proportional through a constant of multiplication to the input energy. Finally the fourth regime is another plateau where all energy-depositing events in the detector create pulses of equal amplitude. This is the GM region.

It is also important to note that most gas detectors are capable of either current or pulse mode operation, as shown in Figure 58.7, making them a very flexible choice for a variety of applications.

58.4.2 Ionization Chambers

Ionization chambers (or “ion chambers” for short) represent the simplest type of gas detector. In this configuration, the charge Q is created, moved to the electrodes by the internal electric field, and directly counted. One key parameter for an ion chamber is the applied bias at which all charges will be collected, V_s, which is called the saturation voltage. This voltage represents the minimum bias necessary to collect all charge pairs before they recombine and is evident in Figure 58.6 as where the saturation plateau begins. Increasing the applied bias further will not change the amount of charge collected and, therefore, the current amplitude assuming constant, monoenergetic irradiation because all of the charge that is generated is fully collected. Assuming the applied bias is greater than V_s, the output signal is directly proportional to the charge created through ionization interactions.

When operated in current mode, ion chambers are useful for measuring the radiation exposure rate, which has SI units of C/kg air per second. This can be related to the traditional unit of exposure, the roentgen (R) where 1 R is 2.58×10^4 C/kg air. One roentgen is defined to be the exposure due to ionizing photons in 1 cm^3 of dry air at standard temperature and pressure. In the presence of a source of radiation, a detector operated at a bias of at least V_s will have a steady-state saturation current of i_s. If the ion chamber can be thought of as being air equivalent, then the exposure rate in SI units is

$$\dot{X} = \frac{i_s}{M}$$

where M is the air mass at standard temperature and pressure. (Note that it may be necessary to adjust the value of M if the detector is at a pressure or temperature other than standard.) Therefore, the measurement of the saturation current directly provides the exposure rate.

Pulse mode operation of an ion chamber, while more complicated, can provide a wealth of additional information about a radiation source, such as type of incident radiation and/or its energy. Prior to discussing practical pulse mode operation, it is necessary to consider how a signal is formed within the detectors as a function of the movement of charge. When a point charge is created at a distance d from an electrode (it doesn't matter if this charge is the electron or the ion), it induces a charge on the surface of the electrode through the method of images or mirror charges. While the actual derivation of the method of images is beyond the scope of this book (the reader is encouraged to consult [8] or [9] for more information), the result is useful. The method of images equates the solution of the potential and electric fields of a point source above a grounded conducting plane to that of two equal but opposite charges separated by the same distance, as shown in Figure 58.8. The surface charge induced on a grounded conducting surface under these conditions is

$$\sigma(x,y) = \frac{-qd}{2\pi\left(x^2 + y^2 + d^2\right)^{3/2}}$$

where q is the charge amplitude and x and y are the rectilinear coordinates of the grounded plane. Then the total induced charge on the surface can be found integrating this expression as

$$Q = \int \sigma dxdy = -q.$$

So the total induced charge is equal to the amount of charge. Note though that this integral is done over the area of an electrode. When the charge is far away, it still induces q on the surface of the electrode, but the area this is spread over is much larger since the electric flux density is smaller. As the charge gets closer to the surface, the flux density increases until the limit where the charge is at the surface of the conductor when all induced charge is concentrated in a single point directly beneath q.

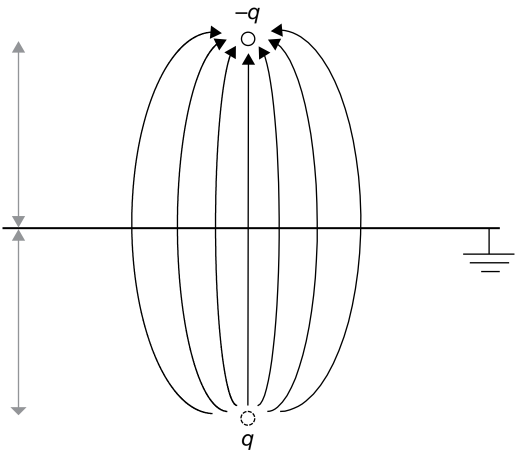


FIGURE 58.8 Illustration of the approach to use the imaginary mirror charge, *q*, to determine the induced charge on the surface of a grounded conductor by the real charge, −*q*.

Let us now apply this concept to the movement of electrons and ions within a gas detector. Upon their generation by ionization, each charge pair will induce a surface charge on the electrodes of the detector. They immediately begin to move because of the application of the electric field with electrons moving toward the anode and the ions moving toward the cathode. As the move closer to their respective anodes, the amount of charge they induce increases steadily until it reaches the charge value of the individual carrier ($-e_0$ for electrons and $+e_0$ for the ions) and the charge is collected on the electrode surface.

However, it is important to realize that the amount of time it takes for these charges to be collected varies significantly between the electrons and the ions due to the large mobility difference of the two. In reality, the electron is collected much sooner that the ion, so the induced signal initially increases quickly as the electrons and ions are both drifting. But once the electrons are collected, the slope of the time rate of change of induced charge decreases to reflect the fact that the only signal being induced is from the movement of the ions. This is represented schematically in Figure 58.9.

Because the ions move so slowly, this creates limitations on the rate of incident radiation that can be detected if one has to wait for all of the ions to be collected. It is therefore desirable to create a detector capable of creating signal only through the movement of electrons. One way of doing this is by choosing a charge collection time that is small relative to the collection time of the ions. In this case, the movement of the electrons from generation to collection induces a voltage signal, V_e, equal to

$$V_e = \frac{Q}{C}\frac{x}{d}$$

where *x* is the distance the electron must travel from generation to the anode and *d* is the distance between the cathode and the anode (the movement of ions is assumed to

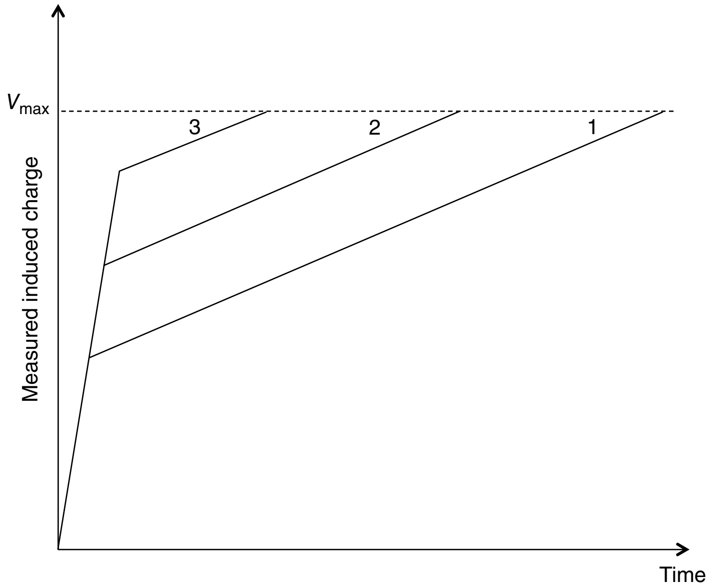


FIGURE 58.9 Pulses created through the measurement of induced charge where the electron–ion pair are created (1) close to the anode, (2) at a point midway between the cathode and anode, and (3) at a point close to the cathode. The fast rise component corresponds to the movement of both the electrons and holes. Once the electrons are collected, which happens much sooner than the ions, then the induced charge is only a function of the slow-moving ions. The measured induced charge reaches its maximal value of Q/C when both the electrons and ions are fully collected.

be negligible in this short time span). As is evident from the equation, this approach suffers from the limitation that the amplitude of the resulting signal is a function of where in the detector the ionization occurred.

The Frisch grid was invented to overcome this problem, as shown in Figure 58.10. In a “gridded” detector, a wire mesh is placed between the anode and cathode and kept at a potential that is intermediate to that of the cathode–anode potential difference. As before, the signal is measured on the anode as the induced charge from the movement of the electrons. The grid acts as a type of Faraday cage for the anode. When the electrons are generated and begin drifting in the drift region of the detector, no charge is induced on the anode. But eventually they pass through the grid and continue to drift toward the anode. Once this occurs, they induce charge on the anode until their collection, as shown in the figure. This does several things. First, the movement of the ions is not measured, so their low mobilities are not a problem. Second, the only portion of the electron movement that induces any signal is between the grid and the anode, a distance that is usually very small. The electrons still induce their full signal but over a very small distance. Therefore, the Frisch grid allows for what is called “single polarity charge sensing” with a signal amplitude that is no longer a function of the interaction location in the detector.

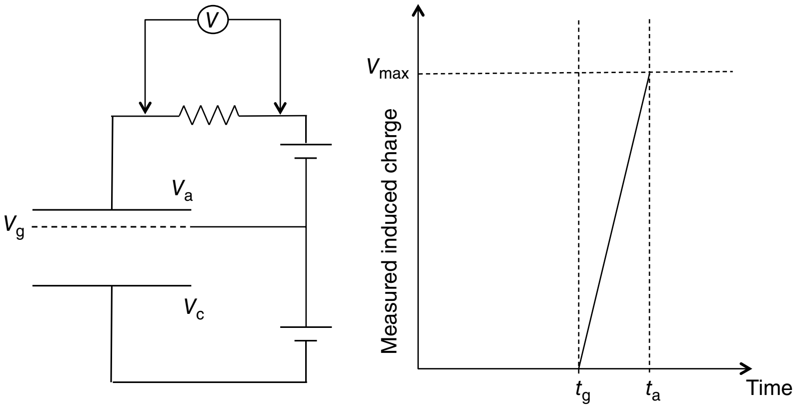


FIGURE 58.10 Schematic of a gas detector with a Frisch grid (left). The anode is held at potential V_a, the cathode at V_c, and the grid at the intermediate potential V_g. The corresponding induced charge profile shows that no charge is induced on the anode as the electron moves from creation to the grid, where it arrives at t_g (right). From there, it induces charge on the anode until it reaches its maximal value at time t_a.

58.4.3 Proportional Counters

While ionization chambers are very simple, their pulse amplitudes are very small since the only charge that can be measured is what is directly created through ionization. However, when the magnitude of the electric field in the gas detector gets large enough, it is possible that the secondary electrons created through ionization are accelerated by the field to such a point where they too can ionize gas molecules, resulting in more charge pairs being created. This process, called the Townsend avalanche, results in the multiplication in the number of charge pairs created, which results in a significantly larger signal and better overall signal-to-noise statistics. It should be noted that the ions are typically not accelerated sufficiently since the acceleration due to an electric field, a, is given by

$$\vec{a} = \frac{q\vec{E}}{m}$$

where m is the mass of the accelerated particle. So it is clear that most ions are too massive to be accelerated enough by the electric field to create further ionizations.

The minimum necessary electric field to achieve charge multiplication is a function of the gas with typical values around 5×10^4 V/cm atm [3]. While it is theoretically possible to achieve this magnitude of electric field in a parallel plate configuration, it is practically difficult since the necessary applied bias is so large. However, these fields can be achieved in a cylindrical geometry, as described in the

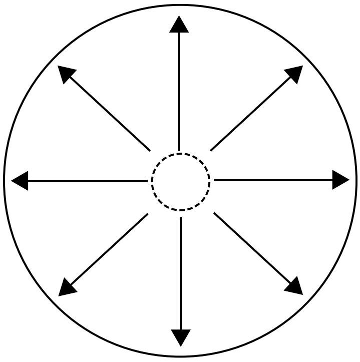


FIGURE 58.11 Illustration of the electric field inside a cylindrical detector. Note that the critical radius, r_c, where the electric field is large enough to support charge multiplication is represented by the dashed line.

previous section. In fact, if we consider how the drift velocity changes with distance, we observe

$$\vec{v} = \frac{\mu\vec{E}}{p} = \frac{\mu V}{pr\ln(b/a)}.$$

So it is evident that the electron actually increases in velocity the closer it gets to the anode, further increasing the probability of avalanche.

An example of the electric field in a cylindrical detector is shown in Figure 58.11. As is evident in the figure, the threshold for charge multiplication occurs at $r = r_c$. So an ion pair would be generated outside this volume through normal ionization and then the electron would migrate into the multiplication region as a result of the internal electric field to then be multiplied by several thousands. The net effect is a significant increase in the size of the overall signal in the detector given by

$$Q = ne_0M$$

where n electrons are initially created and M is the gas multiplication factor. Values of 10^5–10^6 are not uncommon for M.

It is important to realize that, despite the gas multiplication, the amount of charge that is measured is proportional to the amount of charge originally created, which is proportional to the energy deposited in the detector. This is shown in the third region of Figure 58.6. However, unlike for ionization chambers where the variance in pulse amplitude is limited to the Poissonian statistics of charge creation, in proportional counters the variance in the amount of multiplication in any given avalanche must be included. In this case, the overall statistical limit of the proportional counter can be calculated as

$$\left(\frac{\sigma_Q}{Q}\right)^2 = \frac{F}{n} + \frac{1}{n}\left(\frac{\sigma_M}{M}\right)^2$$

where n is the original number of charge carriers created through ionization and F is the Fano factor for the gas. It can be shown that this is inversely proportional to energy of the incident particle.

Due to the size of the output signal and overall ease of use, proportional counters are widely used for a variety of applications including the measurements of beta particles, low energy gamma and X-rays, and neutrons. Their use in mixed radiation fields is also common.

58.4.4 GM Detectors

GM detectors or counters have designs that are very similar to proportional counters. They are also cylindrical in geometry, although they operate at larger electric fields as indicated in the fourth region of Figure 58.6. During the acceleration in the electric field in either a proportional or a GM counter, the electron can ionize the gas molecules or it can just excite them. The excited gas will eventually deexcite through the emission of an ultraviolet photon. This photon can then trigger further ionization in the gas and more avalanches. In a proportional counter, a different gas called a quench gas is usually added in small concentrations to the mixture to absorb the UV photons. However, in a GM counter, no quench gas is added thus allowing these UV photons to create their own avalanches. Most avalanches will create at least a few photons. The result is that the UV photons create avalanches throughout the entire detector volume in what is termed the Geiger discharge.

This would be a runaway, continuous discharge if it were not for the resulting gas ions. In a GM detector, there are a significant number of ions created. Once their number reaches a critical value, the electric field they create as they move toward the cathode is enough to distort and reduce the detector's overall electric field. The reduction in electric field caused by the ions is enough to keep the electrons from being accelerated sufficiently to create an avalanche, thus terminating the discharge process. Once the ions have been collected, the detector returns to its normal operating electric field and can begin the process again.

Because the Geiger discharge continues until the ion concentration is sufficient to terminate it, the pulses created in a GM counter are all the same size since all that is required is that a specific number of ions be created. Therefore, the pulses from a GM detector cannot be used to determine the energy of the incident radiation. They are always operated as counters and are usually used to survey for the presence of radiation and to measure exposure rates.

58.4.5 Scintillators

A scintillator is any material that absorbs energy from ionizing radiation and reemits the energy in the form of light, usually in the optical or ultraviolet wavelengths. Scintillation light can be created through either fluorescence or phosphorescence, with

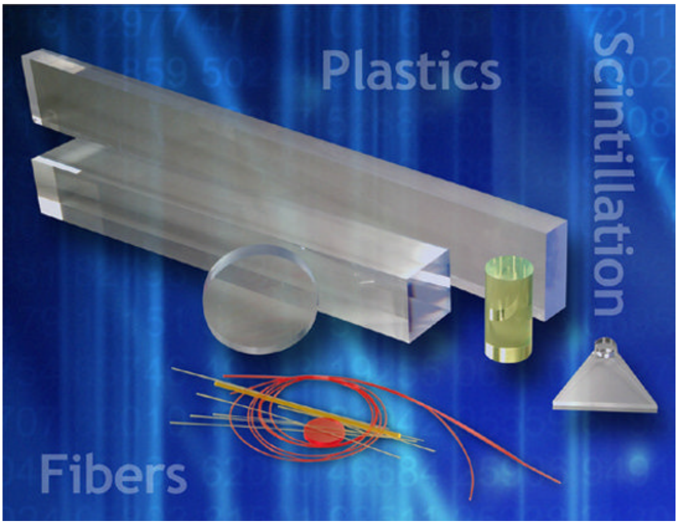


FIGURE 58.12 Example of plastic scintillators made in a variety of shapes. Source: Reproduced with permission of Saint-Gobain Crystals.

the former being the preferred method since the emission of light occurs much faster. The scintillation photons are then directly counted by a light sensor such as a photomultiplier tube (PMT) or a photodiode to create a spectrum, much in the same way that charge pairs are counted in a gas detector.

Scintillators are categorized as either organic or inorganic. Organic scintillators can be in any form—solid, liquid, or gas. They emit light through the deexcitation of an excited molecule. The most common types of organic scintillators are anthracene ($C_{14}H_{10}$), stilbene ($C_{14}H_{12}$), and polyvinyltoluene (PVT). Many plastics work well as organic scintillators. Because these types of materials can be made in many different form factors and easily produced, it is reasonably simple to mold them into desired shapes, including very large slabs and very small fibers, as shown in Figure 58.12.

Inorganic scintillators function differently than organic scintillators in that the scintillation light is emitted as the result of the deexcitation of the crystal lattice of the material. This implies that inorganic scintillators are all solids and care must be given to how they are grown. As a result, larger inorganic detectors are significantly more expensive than the comparable volume of organic scintillators. The properties of several inorganic scintillators are shown in Table 58.4 [10]. Thallium-doped sodium iodide, NaI(Tl), has long been the standard by which all inorganic scintillators are measured due to its exceptional light yield or how many scintillation photons are emitted per unit of energy deposited. However, recent research into new materials such as the lanthanum halides ($LaBr_3$ and $LaCl_3$, for example) [11–14] and elpasolite scintillators (most notably Cs_2LiYCl_6, $Cs_2LiLaCl_6$, and $Cs_2LiLaBr_6$, called CLYC, CLLC, and CLLB, respectively) [15–17] has resulted in a significant improvement in the energy resolutions attributed to scintillators.

TABLE 58.4 Select Common Inorganic Scintillators and Their Properties[a]

Scintillator	Light Yield (Photons/keV)	Light Output (%) of NaI with Bialkali PMT	Wavelength of Maximum Emission (nm)	Thickness to Stop 50% of 662 keV Photons (cm)	Density (g/cm^3)
NaI(Tl)	38	100	415	2.5	3.67
$LaCl_3$(Ce)	49	70–90	350	2.3	3.85
$LaBr_3$(Ce)	63	165	380	1.8	5.08
CsI(Na)	41	85	420	2.0	4.51
CsI(Tl)	54	45	550	2.0	4.51
BGO	8–10	20	480	1.0	7.13
$CdWO_4$	12–15	30–50	475	1.0	7.9

[a] Data provided courtesy of Saint-Gobain Crystals.

When deciding between scintillators, it is important to consider what type of radiation is to be detected. For gamma rays, it is important that a material with high atomic number, *Z*, be chosen to maximize the probability of the photoelectric effect within the detector volume (which is proportional to $Z^{4.5}$). Organic scintillators have low atomic number and therefore are not usually used for gamma-ray spectroscopy. However, because of their size, they can make very large detectors that would be suitable as counters, as is often the case in portal monitoring applications. Inorganic scintillators, on the other hand, can be made with high atomic numbers, so their probability of photoelectric effect is high, but they cannot be made very large.

While the detection of neutrons is discussed in more detail in later sections, it is useful to realize that hydrogenous materials make good neutron detectors. Thus organic scintillators are frequently chosen for neutron detection.

58.4.6 Readout of Scintillation Light

Once the scintillation photons have been produced in the scintillator, they must be counted. This is usually achieved by use of a PMT. (Photodiodes can also be used, but their use is not as common as PMTs and beyond the scope of this chapter. The interested reader is encouraged to consult [3] for additional information on their operation.) The job of the PMT is to convert the scintillation light to electrons. The conversion itself happens at the surface of the PMT, which is called the photocathode. From there, the free electrons enter the tube itself and are accelerated by an electric field to a small metal dynode. If the electron is sufficiently accelerated, it has enough energy to liberate a few more electrons at this dynode. These electrons are then guided by electric field to the second dynode where they liberate a few more electrons and so on through the dynode structure until all of the electrons are finally collected at the anode. A schematic showing the internal structure of a PMT is shown in Figure 58.13.

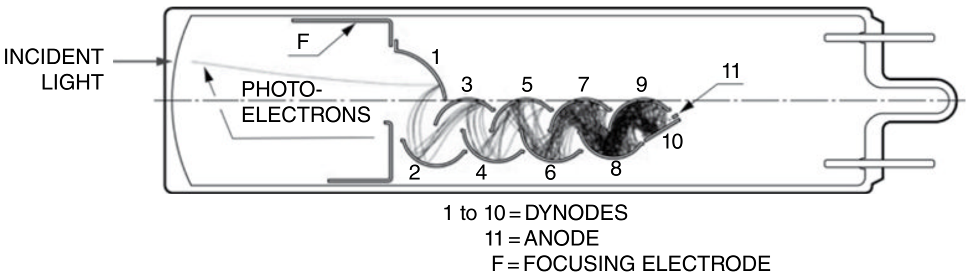


FIGURE 58.13 Schematic of a photomultiplier tube including photocathode (far left), dynodes, and anode. Source: Reproduced with permission of Hamamatsu Corporation.

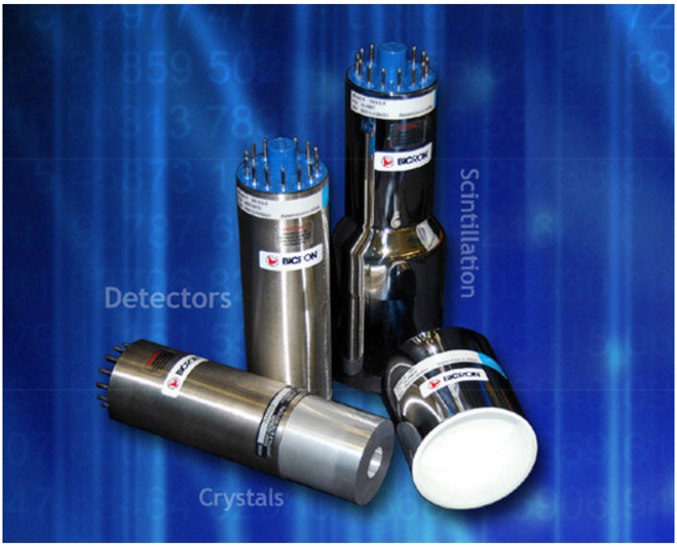


FIGURE 58.14 A variety of different inorganic scintillators coupled to photomultiplier tubes. Source: Reproduced with permission of Saint-Gobain Crystals.

If an electron, on average, liberates M electrons from the surface of each dynode, then the overall gain of the PMT, G, can be approximately calculated as

$$G = M^N$$

where N is the number of dynodes. It is easy to see then that for a typical PMT with 10 stages and M on the order of 3–5, the gain can easily reach 10^6. An example of a PMT mated with a scintillator is shown in Figure 58.14.

The selection of which PMT to use is a function of many things. First, it is ideal that the size and shape of the photocathode surface closely match the surface of the scintillator to minimize light loss. Second, the sensitivity of the photocathode is a function of wavelength, as shown in Figure 58.15. It is therefore important that the wavelength band of absorption of the PMT overlap well with the band of emission of the

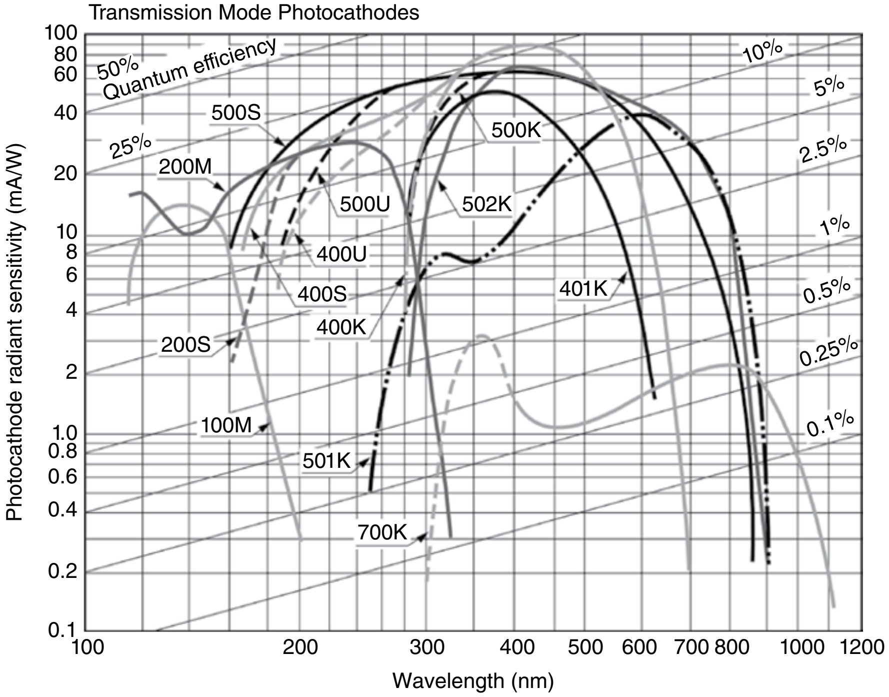


FIGURE 58.15 Sensitivity to a variety of commercial PMTs to different wavelengths of scintillation light. Source: Reproduced with permission of Hamamatsu Corporation.

scintillator. (Note that this is why scintillators like CsI(Tl) have a poor light output relative to NaI(Tl)—because the wavelength of emission is not well matched to the common bialkali PMTs used for NaI.) Lastly, the gain of the PMT is a strong function of the internal electric field uniformity. A small change in the internal electric field results in a significant change in the acceleration of the electrons and therefore seriously impacts *M*. The end result is that the output of the PMT has a great deal of noise. So consideration should be made on how much stability is available in the high-voltage bias supply of the PMT.

58.4.7 Semiconductors

The final type of radiation detector is the semiconductor, the physics of which is well described in many classic texts on the subject [18–20]. Semiconductors work similarly to gas detectors, where signal is created by the excitation of charge carriers to different energy bands of the crystal lattice. In many ways, this is not unlike an ionization chamber; however for semiconductors the charge carriers are electron–hole pairs. When zero energy is deposited in the semiconducting material, the electrons are bound to the atoms at energy levels determined by which shell they are bound to. This can be

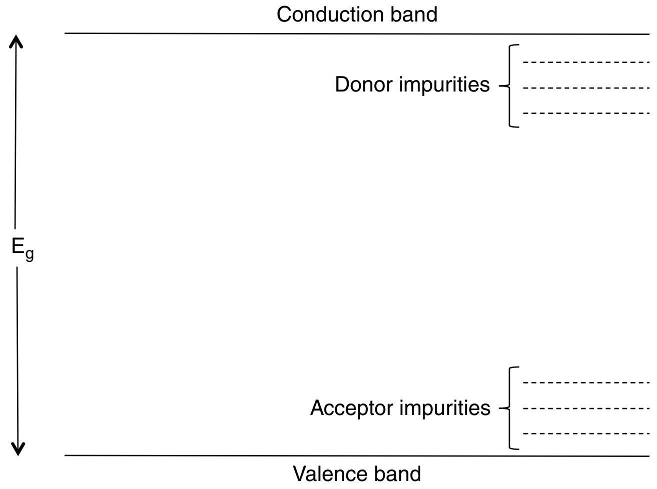


FIGURE 58.16 The band structure of a semiconductor with bandgap energy of E_g.

energetically represented by what is called the valence band. However, when energy is deposited of at least a certain minimum value, an electron from the outer shell can be liberated from the atom and is free to migrate through the material. Energetically speaking, the electron is considered to have entered the conduction band. The energy difference between the valence band and the conduction band is called the bandgap energy, E_g, and represents the required energy to create an electron–hole pair. The resulting vacancy in the valence band is a hole, which is the other charge carrier. This band structure is represented schematically in Figure 58.16.

It should be noted that the model represented in Figure 58.16 is very simplistic. In reality, semiconductor materials can be tailored by deliberately adding dopant impurities to alter the material's overall properties. These impurities can add intermediate energy levels within the bandgap, thus altering the minimum energy required to generate an electron–hole pair. The concentration of different impurity types, whether deliberately added or existing as a result of the crystal growth process, determines whether the detector is considered to be intrinsic (very low dopant concentration), n-type (high concentration of impurities with loosely bound electrons), or p-type (high concentration of impurities with holes available for electrons to occupy at much lower levels). In practice, it is ideal to combine different impurity types in different locations to form a p-n junction or diode or a p-i-n junction detector. For further details on the process of doping semiconductors and the physics resulting from the addition of impurities, the reader is encouraged to consult [19].

There are many ways to generate charge carriers in semiconductors. One common, problematic way is through thermal generation. For gas detectors, the typical ionization energies are on the order of tens of eV. However, for semiconductors, it is more common to have E_g on the order of 1–3 eV. This implies that it is possible for thermal energy to create charge pairs. The probability of thermal generation at temperature T of a charge pair is proportional to

$$p(T) \propto T^{3/2} \exp\left(-\frac{E_\mathrm{g}}{2kT}\right).$$

Thus it is clear that for detectors with smaller bandgap energies such as Ge require cooling in order to minimize thermal generation of charge carriers.

Ideally, the charge carrier generation will occur due to the deposition of energy from incident radiation. Contacts are applied to the surface of the semiconductor to create the anode and cathode through the application of a voltage bias. This results in the collection of the charges in a way that is similar to ionization chambers. Once the electron–hole pair are created, they are collected in the normal way by the application of an electric field between an anode and cathode. Similar to gas detectors, the drift velocity of each charge carrier can be calculated as

$$\overline{v_{\mathrm{e,h}}} = \mu_{\mathrm{e,h}} \vec{E}$$

where $\mu_{\mathrm{e,h}}$ is the mobility of the electrons and holes, respectively. Unlike gas detectors where the mobility of the ions is orders of magnitude less than that of the electrons, for semiconductors the mobilities tend to be close to the same order of magnitude. Several common semiconductor radiation detectors and their properties are presented in Table 58.5.

For the detection and measurement of ionizing radiation, semiconductor devices have many benefits. One key benefit is their superior energy resolution. Assuming trapping of the charge carriers is small and the readout electronics do not contribute much noise to the overall system, the noise is just a function of charge carrier statistics. Assuming Poissonian statistics and including the Fano factor, F, the variance in the number of charge carriers, N, created during the deposition of E energy is

$$\sigma_N^2 = FN = F\frac{E}{W}$$

TABLE 58.5 Select Physical Properties of Some Common Semiconductor[a]

Parameter	Si	Ge	GaAs	CdZnTe	CdTe	HgI_2	TlBr
Density (g/cm^3)	2.33	5.33	5.32	5.78	5.85	6.4	7.56
Average atomic number	14	32	31.5	49.1	50	62	58
Bandgap (eV)	1.12	0.67	1.43	1.572	1.44	2.15	2.68
Electron $\mu\tau$ product (cm^2/V)	>1	>1	8×10^{-5}	4×10^{-3}	3×10^{-3}	3×10^{-4}	2×10^{-6}
Hole $\mu\tau$ product (cm^2/V)	~1	>1	4×10^{-6}	1.2×10^{-4}	2×10^{-4}	4×10^{-5}	2×10^{-6}
Resistivity (Ω cm)	$<10^4$	50	10^7	3×10^{10}	10^9	10^{13}	10^{12}

[a] From Ref. 21.

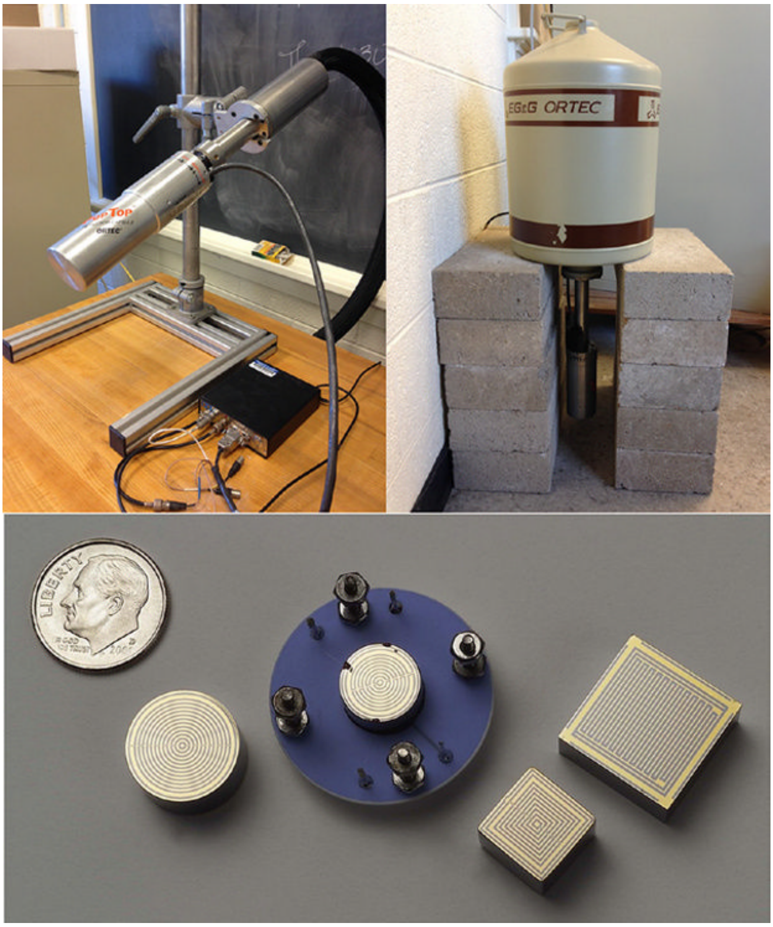


FIGURE 58.17 Examples of different types of high-purity germanium (HPGe, top row) and CdZnTe (bottom row) semiconductor devices.

where W is the average energy to generate an electron–hole pair in the medium and E is the energy deposited in the interaction. For most common semiconductors, F is on the order of 0.1 and W is around 1 eV/charge carrier [22]. Therefore, σ_N^2 is much smaller for semiconductor radiation detectors than any other detector.

Semiconductors, because of their excellent energy resolution and high density, are widely used in a number of different types of applications. They are an obvious choice for spectroscopy, both of gamma rays and charged particles. However, they also can be used as imagers, primarily in the medical industry. Most recent research has focused on identifying new materials appropriate for detectors as well as sophisticated readout techniques that mimic the Frisch grid for gas detectors [21, 23–27]. Several commercial and research-grade semiconductors are shown in Figure 58.17. A comparison of the energy spectra possible with semiconductors is shown in Figure 58.18.

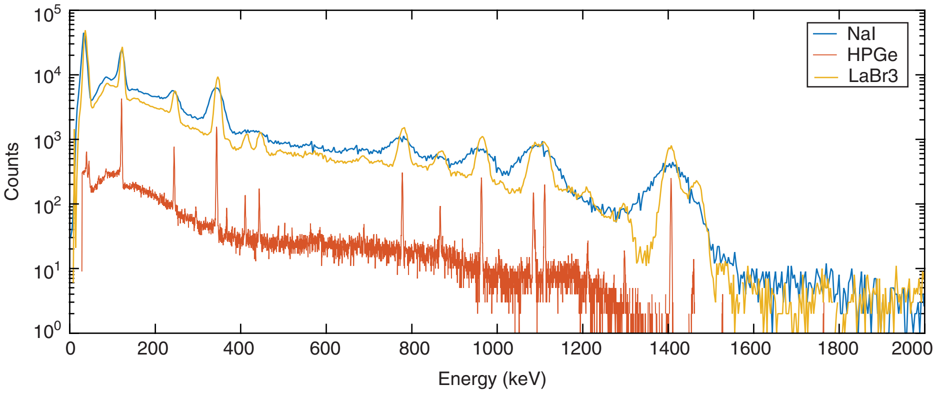


FIGURE 58.18 Comparison of the energy spectrum of Eu-152 obtained by a HPGe (top), LaBr_3 (middle), and NaI (bottom) spectrumeters, illustrating the difference between high, medium, and low resolution.

58.5 NEUTRON DETECTION

Unlike other forms of radiation, neutrons do not directly ionize atoms, so any method for detecting them is based on indirect processes where the neutron creates either a charged particle or a photon that is subsequently detected. Neutrons can interact in one of six different ways that fall into the broader categories of scattering or absorption, as shown in Figure 58.19. An individual scatter cannot create a secondary particle that will generate ionization; however in most materials, a neutron will scatter many times before being absorbed.

When considering any neutron interaction, it is necessary to understand the energy of the neutron since the probability of any of these interactions (called the cross section) is strongly a function of the neutron energy. Unlike other types of radiation, the energy of a neutron source is specified based on a continuous distribution rather than a discrete energy, and all interaction probabilities are statistically derived from this. Neutrons can be thought of as an ideal gas at a given temperature, which is statistically described by the Maxwell–Boltzmann distribution shown in Figure 58.20:

$$p(E)=\frac{2\pi}{(\pi kT)^{3/2}}e^{-\frac{E}{kT}}E^{1/2}$$

where T is the temperature of the gas. Based on this distribution, the average energy of the neutron source can be calculated as

$$\bar{E}=\frac{3}{2}kT=5.227\times10^{-15}v^2$$

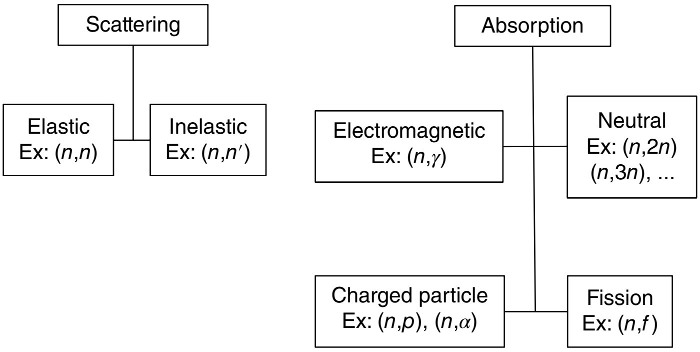


FIGURE 58.19 The six different neutron interaction mechanisms grouped into the categories of scattering versus absorption.

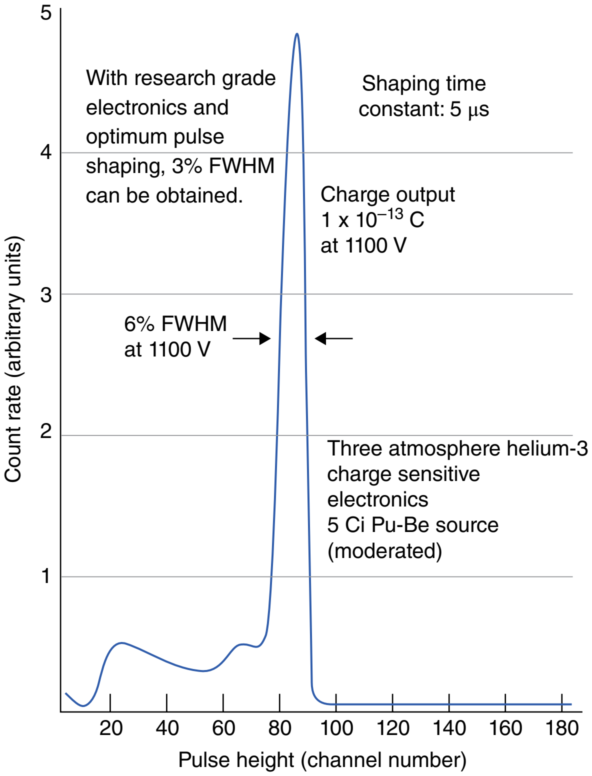


FIGURE 58.20 Sample pulse height spectrum with varying gamma-ray noise from a ^{3}He proportional counter, showing the peak at full *Q* energy deposition and the wall effect associated with partial charge collection when the proton or triton deposit a portion of their energy in the wall of the tube. Source: Reproduced with permission of GE Reuter-Stokes.

where v is the velocity of the neutron. Because of this temperature or velocity dependence on the energy distribution, neutrons are given terms like "cold," "thermal," "fast," "slow," etc. to describe their energies. For the sake of this text, we shall refer to thermal neutrons as those with $\bar{E}$ less than around 0.5 eV whereas a fast neutron will have $\bar{E}$ greater than 1 eV. (The intermediate region is considered to be epithermal but will not be further discussed in this chapter since the most important differences in neutron detection occur between the thermal and fast regimes.)

When an absorption interaction occurs, two daughter products are created traveling in directions opposite each other, and a certain amount of energy, Q, is liberated in the reaction. This energy is shared among the daughter products with the exact value imparted to the daughter determined by the conservation of energy and momentum. (In other words, lighter daughter products will receive more energy and heavier ones.) Note that there is no relationship between Q and the energy of the incident neutron. So a measurement of Q does not allow for determination of the energy of the neutron. Neutron spectroscopy is beyond the scope of this text, but the interested reader is encouraged to consult [7] for additional information.

58.5.1 Thermal Neutron Detection

At thermal energies there are only a few absorption reactions with a high enough cross section to allow for a reasonable detector efficiency. These are summarized in Table 58.6. The materials summarized in this table can be incorporated into any of the aforementioned types of detectors (gas, scintillator, semiconductor).

The most common type of thermal neutron detectors are proportional counters incorporating ^{3}He, ^{10}B, or ^{6}Li. The most common fill gases are ^{3}He and ^{10}BF$_3$. In these detectors, the thermal neutron is captured through the process outlined in Table 58.6. From there, the daughter products ideally deposit all of their energy within the gas, resulting in Q energy to be measured. However, depending on where the absorption occurs, it is possible that some value less than Q will be deposited in the detector if one of the daughter products hits the wall of the detector. The deficit in energy deposited in the gas is a function of how far from the wall the daughter product is created—the closer the product is to the wall, the less energy it will deposit. The result is a

TABLE 58.6 Thermal Neutron Reaction Data[a]

Reaction	Cross Section for 0.025 eV Neutrons (Barns)	Q-Value (MeV)
^{10}B$(n,\alpha)^7$Li	3840	2.792 (6%)
^{10}B$(n,\alpha)^7$Li*	3840	2.310 (94%)
^{3}He$(n,p)^3$H	5400	0.764
^{6}Li$(n,\alpha)^3$H	937	4.78

[a] From Ref. 7.

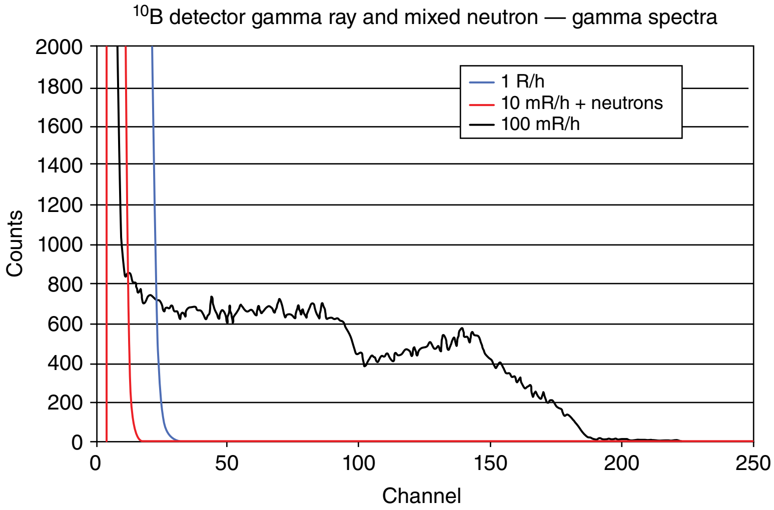


FIGURE 58.21 Example of a spectrum from a ^{10}B-lined proportional counters in the presence of increasing gamma-ray background. Note the two plateaus correspond to the wall effect from the two reaction products, ^{7}Li and an alpha particle, from left to right. Source: Reproduced with permission of GE Reuter-Stokes.

continuum of possible values of deposited energy for that product ranging from near the product's full value to close to zero. This is schematically represented in Figure 58.21 with the corresponding features in the spectrum present.

It is also possible to deposit the neutron-absorbing material on the wall of the detector; ^{10}B-lined proportional counters are a good example of this. This design has the advantage that the aforementioned gases, while good at absorbing neutrons, are not well suited for use in proportional counters due to low *W* values or poor multiplication properties. Lined proportional counters, on the other hand, are designed to have the neutron interact in the wall and then have one of the reaction products enter the gas volume where a more suitable gas is used for proportionality. In this case, it is not possible to have the full *Q* value deposited in the detector since only one of the two reaction products will enter the gas. Additionally, depending on how much wall thickness must be penetrated by that product before it reaches the gas, the amount of energy the product has available to deposit will range from nearly zero (if the absorption occurs far from the gas) to nearly the full product energy (if the absorption occurs very near the surface of the lining). An example spectrum illustrating this is shown in Figure 58.22.

In addition to incorporating ^{3}He, ^{10}B, or ^{6}Li into a variety of different detector configurations, the process of fission also can be used for detection through what is called a fission chamber. In these types of detectors, the wall of the detector is lined with a fissionable material, usually ^{235}U or ^{239}Pu, and a thermal neutron causes a fission in this lining. The result of this reaction are fission products, many of which carry a significant charge to them. Further, the *Q* value associated with fission can be extremely

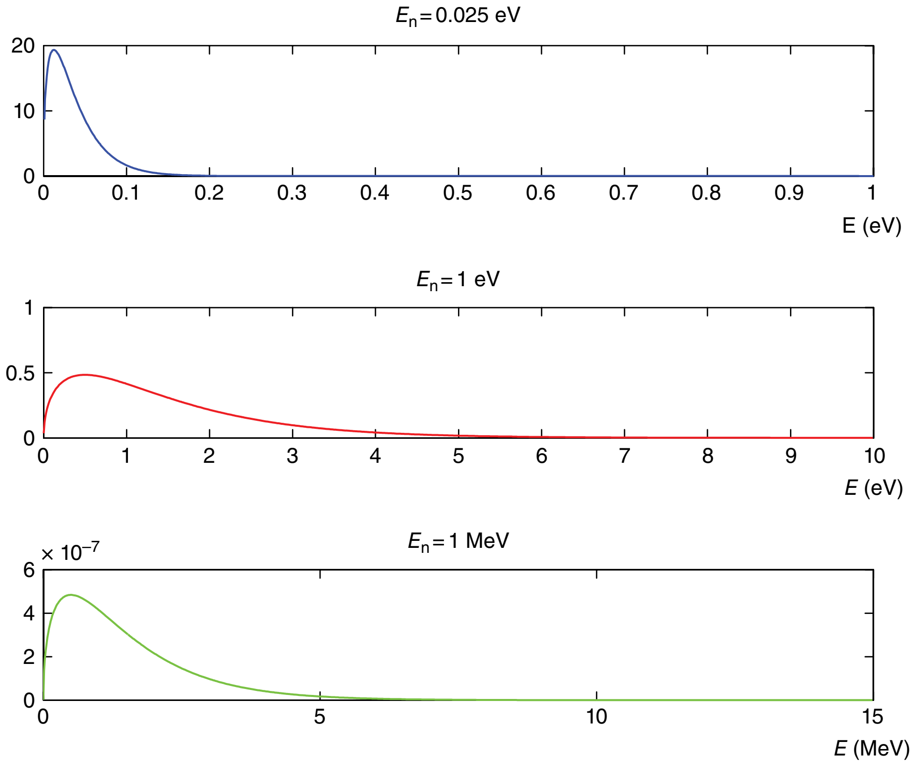


FIGURE 58.22 Example probability distribution function of the Maxwell–Boltzmann distribution for neutron sources of different energies. Note the variations in the axes values.

large—150 MeV or greater. So each fission product carries with it a great deal of energy that can be deposited in the detector. This is advantageous because of the impact on the spectrum of gamma rays. This is shown in Figures 58.21 and 58.22. With increasing gamma-ray dose rates, the lower region of the spectrum begins to overwhelm the region associated with neutron counts. This makes it difficult to set a lower energy threshold for neutron counting. However, if we consider that the peak in the spectrum is associated with full Q deposition, then it is clear that having a larger Q value results in improved threshold setting capabilities to discriminate the noise associated with gamma rays. Additionally, because the Q value is so large, these detectors can be operated as ionization chambers since the resulting charge from ionization is so large that no gas multiplication is required.

58.5.2 Fast Neutron Detection

There are a few options when fast neutrons must be detected. One option is to surround one of the aforementioned thermal neutron detectors with sufficient moderator to slow a fast neutron down to the thermal range and then detect the slower neutrons. However,

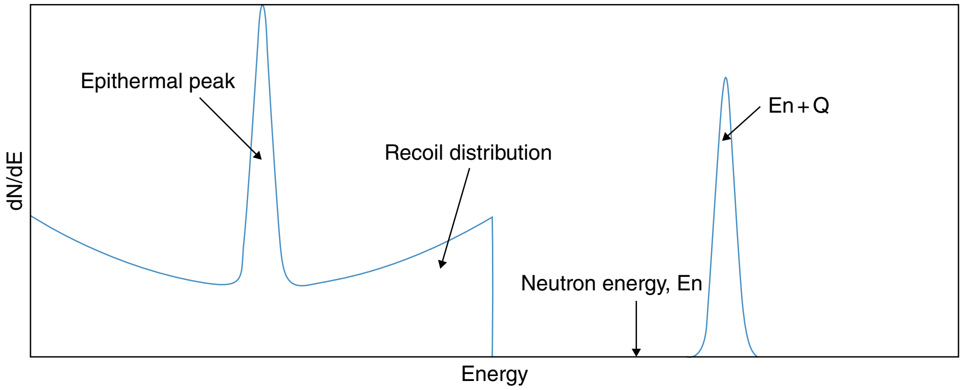


FIGURE 58.23 Sample fast neutron spectrum showing full energy deposition at $E_T = E_n + Q$ and epithermal peak at Q.

there are options for the detection of fast neutrons more directly. Note that like thermal neutrons, fast neutrons are not directly detected but are measured through their reaction products. Some of the aforementioned neutron absorption mechanisms can still be employed for fast neutron detection, albeit with significantly smaller cross sections. However, a new mechanism for fast neutron detection exists based on the increased importance of elastic scattering at fast energies.

When the energy of the neutron is comparable to or greater than the Q value of the material, it is possible to measure the energy of the incident neutron by measuring the energies of the reaction products. Hence, the information on the energy of the neutron is not lost, as it is in thermal neutron detection.

In scattering off of a light nuclei, the energy of the recoil nucleus, E_R, is given by

$$E_R = \frac{4A}{(1+A)^2} E_n \cos^2\theta$$

where A is the mass of the target nucleus, E_n is the energy of the incident neutron, and θ is the scatter angle in the lab reference frame. There are clearly maxima in E_R depending on A and θ, which governs how to properly choose detecting material—materials with small mass have a higher amount of energy imparted to the recoil nuclei, which is ideal for detection.

The most common detectors in this energy range are proportional counters made from hydrogen, ^{3}He, or methane or organic scintillators, either as solid plastic or as liquid. Each different type of detector functions as described in previous sections, but they are measuring the energy deposited by the reaction products or recoil nucleus. Since the energy distribution of the recoil nuclei can be predicted, we can draw the expected energy spectrum for a fast neutron detector. When the full energy of all of the reaction products is deposited, this results in a peak in the spectrum at energy $E_n + Q$, as shown in Figure 58.23. When a portion of the recoil nucleus is measured, a continuum is present representing the multiple possible values of energy deposited by the recoil nucleus. Lastly, a peak is present at Q called the "epithermal peak." This corresponds to the deposition of energy by a thermal neutron, which liberates Q. Based on this, it is clear that the energy of the neutron can be inferred from the full energy peak where the total energy deposited is

$$E_{\mathrm{T}} = E_{\mathrm{n}} + Q.$$

However, it is important to note that this is the energy of the neutron when it reaches the detector. Neutrons are moderated by many things, especially including interactions with the environment between the source and the detector. So while a neutron source itself might be a fast source, by the time the neutron reaches the detector, it could have easily thermalized, especially if the detector is located a significant distance from the source.

58.6 CONCLUDING REMARKS

Many different types of radiation detectors have been presented in this chapter. It can be confusing to determine what type of detector to choose for a particular application. Ultimately the user needs to determine what type of radiation they are attempting to detect and what type of data they hope to obtain from the detector, be it an average dose or high-resolution spectroscopy. Detector selection is usually based on a number of different factors in addition to these, including the overall detection efficiency, resolution (if spectroscopic data is required), the detector's physical size and operating constraints (e.g., does the detector require cryogenic cooling, does it need to operate as a handheld device with its own power supply, etc.), and cost. However, at their core, all radiation detectors perform the same function: they convert the incident radiation to charge when that radiation deposits energy in the detector. Accurate measurement of that radiation requires accurate quantification of the charge created. While current research in radiation detection focuses on creating new materials for converting deposited energy to charge or new readout and data processing algorithms, the fundamental concept of charge measurement is still the same.

REFERENCES

1. H. Cember, *Introduction to Health Physics*, 4th ed. New York: McGraw-Hill Medical, 2009.
2. P. W. Frame, "A history of radiation detection instrumentation," *Health Physics*, vol. 88, no. 6, pp. 613–637, June 2005.
3. G. F. Knoll, *Radiation Detection and Measurement*, 4th ed. Hoboken, NJ: John Wiley & Sons, Inc., 2010.

4. C. Leroy, *Principles of Radiation Interaction in Matter and Detection*, 3rd ed. Singapore: World Scientific, 2012.
5. W. J. Price, *Nuclear Radiation Detection*. New York: McGraw-Hill, 1958.
6. D. Reilly, N. Ensslin, H. Smith, and S. Kreiner, *Passive Nondestructive Assay Manual*. Springfield, VA: National Technical Information Service, 1990.
7. N. Tsoulfanidis, *Measurement and Detection of Radiation*, 2nd ed. Washington, DC: Taylor & Francis, 1995.
8. D. J. Griffiths, *Introduction to Electrodynamics*, 4th ed. Englewood Cliffs, NJ: Prentice Hall, 2013.
9. J. D. Jackson, *Classical Electrodynamics*, 3rd ed. New York: John Wiley & Sons, Inc., 1999.
10. S. Derenzo, M. Boswell, M. Weber, and K. Brennan, "Scintillation Properties." [Online]. Available: http://scintillator.lbl.gov/ (Accessed: February 14, 2014).
11. D. Alexiev, L. Mo, D. A. Prokopovich, M. L. Smith, and M. Matuchova, "Comparison of $LaBr_3$:Ce and $LaCl_3$:Ce with NaI(T1) and cadmium zinc telluride (CZT) detectors," *IEEE Transactions on Nuclear Science*, vol. 55, no. 3, pp. 1174–1177, June 2008.
12. R. González, J. M. Pérez, O. Vela, and E. de Burgos, "Performance comparison of a large volume CZT semiconductor detector and a $LaBr_3$(Ce) scintillator detector," *IEEE Transactions on Nuclear Science*, vol. 53, no. 4, pp. 2409–2415, August 2006.
13. W. M. Higgins, J. Glodo, E. Van Loef, M. Klugerman, T. Gupta, L. Cirignano, P. Wong, and K. S. Shah, "Bridgman growth of $LaBr_3$:Ce and $LaCl_3$:Ce crystals for high-resolution gamma-ray spectrometers," *Journal of Crystal Growth*, vol. 287, no. 2, pp. 239–242, January 2006.
14. K. S. Shah, J. Glodo, M. Klugerman, L. Cirignano, W. W. Moses, S. E. Derenzo, and M. J. Weber, "$LaCl_3$:Ce scintillator for γ-ray detection," *Nuclear Instruments & Methods in Physics Research Section A*, vol. 505, no. 1/2, p. 76, June 2003.
15. B. S. Budden, L. C. Stonehill, J. R. Terry, A. V. Klimenko, and J. O. Perry, "Characterization and investigation of the thermal dependence of Cs_2LiYCl_6:Ce^{3+} (CLYC) waveforms," *IEEE Transactions on Nuclear Science*, vol. 60, no. 2, pp. 946–951, April 2013.
16. J. Glodo, R. Hawrami, and K. S. Shah, "Development of Cs_2LiYCl_6 scintillator," *Journal of Crystal Growth*, vol. 379, pp. 73–78, September 2013.
17. J. Glodo, E. van Loef, R. Hawrami, W. M. Higgins, A. Churilov, U. Shirwadkar, and K. S. Shah, "Selected properties of Cs_2LiYCl_6, $Cs_2LiLaCl_6$, and $Cs_2LiLaBr_6$ scintillators," *IEEE Transactions on Nuclear Science*, vol. 58, no. 1, pp. 333–338, February 2011.
18. C. Kittel, *Introduction to Solid State Physics*, 3rd ed. New York: John Wiley & Sons, Inc., 1966.
19. G. Lutz, *Semiconductor Radiation Detectors Device Physics*. Berlin: Springer, 2007.
20. S. M. Sze, *Physics of Semiconductor Devices*, 3rd ed. Hoboken, NJ: Wiley-Interscience, 2007.
21. A. Owens and A. Peacock, "Compound semiconductor radiation detectors," *Nuclear Instruments and Methods in Physics Research Section A*, vol. 531, no. 1–2, pp. 18–37, September 2004.

22. M. Harrison, “Fano factor and nonuniformities affecting charge transport in semiconductors,” *Physical Review B*, vol. 77, no. 19, 2008.

23. F. Zhang, Z. He, D. Xu, G. F. Knoll, D. K. Wehe, and J. E. Berry, “Improved resolution for 3-D position sensitive CdZnTe spectrometers,” *IEEE Transactions on Nuclear Science*, vol. 51, no. 5, pp. 2427–2431, October 2004.

24. Z. He and B. W. Sturm, “Characteristics of depth-sensing coplanar grid CdZnTe detectors,” *Nuclear Instruments & Methods in Physics Research Section A*, vol. 554, no. 1–3, pp. 291–299, December 2005.

25. A. Owens, *Compound Semiconductor Radiation Detectors*. Boca Raton, FL: Taylor & Francis, 2012.

26. P. J. Sellin, “Recent advances in compound semiconductor radiation detectors,” *Nuclear Instruments and Methods in Physics Research Section A*, vol. 513, no. 1–2, pp. 332–339, November 2003.

27. Y. Zhu, S. E. Anderson, and Z. He, “Sub-pixel position sensing for pixelated, 3-D position sensitive, wide band-gap, semiconductor, gamma-ray detectors,” *IEEE Transactions on Nuclear Science*, vol. 58, no. 3, pp. 1400–1409, June 2011.

59

MEASURING TIME AND COMPARING CLOCKS

JUDAH LEVINE
Time and Frequency Division and JILA, NIST and the University of Colorado, Boulder, CO, USA

59.1 INTRODUCTION

Time and time interval have played important roles in all societies since antiquity. The original definitions were based on astronomy: the solar day and year and the lunar month were widely used as measures of both time and time interval. As I will show in Section 59.15, the strong connection between astronomy and time persists even today, when both time and time interval are measured by means of clocks. I will begin by describing a generic clock, and I will then discuss various means of comparing these devices and characterizing their performance using a combination of deterministic and stochastic parameters. I will conclude with a short description of calibrating them in terms of international standards of time and frequency.

59.2 A GENERIC CLOCK

All clocks consist of two components: a device that produces or observes a series of periodic events and a counter that counts the number of events and possibly also interpolates between consecutive events to improve the resolution of the measurement. The choice of which periodic event to use as the reference period for the clock plays a

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

fundamental role in determining its performance so that it is natural to characterize a particular clock design based on an evaluation of the reference period that it uses to drive its counter.

In addition to the two components discussed in the previous paragraph, real clocks and time scales have a time origin that is derived from some external consideration. As a practical matter, the time origin is generally chosen to be sufficiently far in the past so that most epochs of interest have positive times with respect to the origin.

In addition to a time origin, real time scales are used to construct a calendar—an algorithm that assigns names to clock readings. These considerations are very important but are mostly outside of the scope of this discussion. Although I will not discuss the methods used to implement a calendar, I will discuss the methods that are currently used to define Coordinated Universal Time (UTC) and the discussions that are currently underway (as of 2016) about possibly modifying the definition of this time scale.

Two distinct parameters are important in characterizing the frequency reference of any clock: (i) The accuracy of the reference period—how closely does the period conform to the definition of the second. (ii) The stability of the reference period over both the short and long terms. (Stability is a necessary prerequisite for an accurate device but is not sufficient, and it is quite common for real oscillators to have a stability that is significantly better than the accuracy.) A number of methods have been developed for characterizing the stability of periodic events, and I will briefly describe the tools that implement these methods in the next section. I will discuss the question of accuracy in a subsequent section.

59.3 CHARACTERIZING THE STABILITY OF CLOCKS AND OSCILLATORS

The methods that are used for these purposes fall into two general classes: methods that characterize the worst-case performance of a clock and methods that characterize the average performance using statistical parameters derived from a root-mean-square (RMS) calculation. In both cases, the analysis is based on a finite-length data set.

A worst-case analysis is usually sensitive both to the length of the data set that is analyzed and to the exact interval of the observation because large glitches usually occur sooner or later, and a data set either includes a large glitch or it doesn’t. We might expect that the results of a worst-case analysis would show a large variation from sample to sample for this reason.

A statistical analysis, on the other hand, assumes implicitly that the data are stationary so that neither the interval of observation nor the length of the data set is important in principle. A statistical analysis tends to attenuate the effect of a glitch, since even a large glitch may have only a small impact on an ensemble-average value. More generally, a statistical analysis is not suitable if the data are not stationary, since ensemble-average values will exist in a formal sense but will not be very useful in understanding the performance of the actual device.

In order to characterize the stability of a device under test, we can imagine that it is compared to a second clock that is perfect. That is, the perfect clock produces "ticks" that are exactly uniform in time. The interval between ticks of the perfect clock is τ so that the ticks occur at times 0, τ, $k\tau$, $(k+1)\tau$, etc., where k is some integer. (As I have discussed in Section 59.2, the times are relative to some origin that is defined outside of the measurement process, and the time of the first tick is 0 with respect to that origin. This is not a limitation in the current discussion, since the origin is simply an additive constant that is not important when discussing time stability.) The time of the clock under test is read each time the standard clock emits a tick, and the time differences are x_k, x_{k+1}, …, where x_k is a short-hand notation for the time-difference reading at time $k\tau$, etc. In general, the units of time are seconds and fractions of a second. The "frequency" of a clock is the fractional frequency difference between the device under test and a perfect clock operating at the same nominal frequency. For example, the frequency of the device under test, f, which generates a signal at a physical frequency of F, is

$$f = \frac{F - F_o}{F_o}, \tag{59.1}$$

where F_o is the output frequency of the perfect device used in the comparison. With this definition, the frequency of a clock is a dimensionless parameter, and the time difference after a time τ has changed by $f\tau$.

59.3.1 Worst-Case Analysis

Analyzing the time-difference data from the perspective of a worst-case analysis sounds easy. We simply look for the largest absolute value of x in the data set, assuming that the device under test and the standard reference device were set to the same time at the start of the measurement. A statistic that realizes this idea is the maximum time-interval error (MTIE) [1], which is usually calculated as the difference between the largest and smallest time differences in the ensemble of measurements. (With this definition, a device that has an arbitrarily large and constant time difference has an MTIE value of 0, because MTIE is a measure of the evolution of the time difference, not the magnitude of the time difference itself. In this respect, the MTIE statistic is really a measure of the frequency offset between the device under test and the standard reference.) It is commonly used to characterize the oscillators in telecommunications networks.

The MTIE statistic depends both on frequency accuracy and frequency stability, since a clock with a frequency offset with respect to the reference device will eventually produce unbounded time differences even if the frequency offset is absolutely stable. As we mentioned before, the results of a worst-case analysis can show large variations from one set of data to another one so that the MTIE statistic is normally defined as the largest MTIE value over all data sets of some fixed length within some larger time interval, T [2]. In an extreme case, the value of T is the life of the device so that a device that exhibited one glitch after years of faithful service might not satisfy a specification based on MTIE in this case. This form of the definition is therefore unambiguous but not very useful.

An alternative view is to consider the MTIE value obtained from a single data set to be an estimate of the underlying "true" value of MTIE, which is characterized by the standard statistical parameters of a mean and a standard deviation. The standard statistical machinery can then be used to provide an estimate of the probability that the observed value is (or is not) consistent with a mean value for MTIE that is specified by some required level of performance. Since this analysis method characterizes MTIE as a statistical parameter, it usually requires some ancillary assumption about handling measurements that are not consistent with the mean and standard deviation of the distribution of the remainder of the observations. In other words, is a large outlier (i) treated as an error that should be ignored, (ii) accepted as a low-probability event that is consistent with the mean and standard deviation deduced from previous data, or (iii) an indication that the mean or the standard deviation should be updated by including this new observation? One solution is to completely reject data that differ from the mean by more than four standard deviations and to provisionally accept data that differ by more than three but less than four standard deviations. The data in the provisional category are compared to subsequent values, and all of these newer data may be used to provide an update to the mean or to the standard deviation as appropriate. The specific algorithm that is used is based more on administrative considerations and experience than on a rigorous statistical analysis, since errors are generally not statistical events by definition.

The evolution of the time-difference data that are the input to an MTIE calculation is sensitive both to the frequency stability of the device under test and to the sampling interval, τ, and the data become increasingly insensitive to fluctuations in the frequency of the device under test with respect to the reference device that are much shorter than the sampling period. A fluctuation in the offset frequency whose period is an exact submultiple of the sampling period has no effect on MTIE. Therefore, the sampling period must be short enough so that these shorter-period fluctuations either are not important in the application supported by the device or are known to be small a priori. Since the length of a data set is limited in practice, this consideration implies a trade-off between the shortest and longest frequency fluctuations that can be estimated from a set of measurements. However, a measurement that uses a longer sampling interval to detect longer-period fluctuations must guarantee (by means of digital or analog filtering) that the shorter-period fluctuations are not aliased by the longer sampling period.

Another way of addressing the aliasing problem is to use a very rapid sampling period (so as to minimize or eliminate the impact of aliasing) but to acquire these measurements in blocks separated by dead time in which the clock is not observed. This method will also have a potential aliasing problem for frequency fluctuations that are synchronous with the sum of the sample period and the dead time. This problem can be addressed by varying the dead time in a pseudorandom way, but this complicates the analysis somewhat, since the blocks are no longer equally spaced in time.

I will now describe the statistical estimates of time and frequency that are commonly used to characterize clocks and oscillators outside of the telecommunications domain. I will limit my discussion to the original Allan variance, since its significance can be explained intuitively. The more complicated versions of the Allan variance and the characterizations in the frequency domain using Fourier analysis are described in the literature. (see Ref. [3].)

59.3.2 Statistical Analysis and the Allan Variance

The statistical analysis starts with the same time differences that we discussed in the previous sections. The average frequency of the clock under test with respect to our perfect clock over the time interval τ between measurements is estimated as

$$y_k = \frac{x_k - x_{k-1}}{\tau}. \tag{59.2}$$

The numerator and denominator on the right-hand side of Equation 59.2 have the units of time so that the frequency defined by this equation is a dimensionless quantity. If the device under test had a frequency that was constant with respect to the perfect clock, then Equation 59.2 would give the same result for any value of k. (Note that the device under test need not have the *same* frequency as the perfect clock. We would get the same result for every value of k even if the frequency difference was *any* constant value.)

Real event generators are not perfect, and it is useful to characterize their performance by means of the estimator

$$y_{k+1} - y_k = \frac{x_{k+1} - 2x_k + x_{k-1}}{\tau}. \tag{59.3}$$

Equation 59.3 gives the difference in the frequency of the device under test between two consecutive, equal measurement intervals with no intervening dead time. This estimator provides an estimate of frequency stability—not frequency accuracy. From the perspective of the measurement at time τ_k, this statistic is an estimate of the time difference that will be observed at the next measurement time with index $k+1$ based on the evolution of the time difference in the time interval ending at index k. Its magnitude is not sensitive to a constant time difference or frequency difference between the device under test and the perfect reference device. (Compare this to MTIE, which was discussed in the previous section and which is sensitive to a constant frequency difference but not to a constant time difference.)

The final step in the definition of the estimator is to assert (or to hope) that the variations estimated by Equation 59.3 are stationary. That is, the computation does not depend in a systematic way on the value of the index *k*—any choice of *k* would produce a value that is consistent (in a statistical sense) with the result for any other choice of *k*. Then, the RMS of Equation 59.3 has a well-defined value and that RMS value has an associated, well-defined, standard deviation. When various normalizing constants are added, the mean square value of Equation 59.3 estimated over all possible values of *k* is the two-sample or Allan variance for an averaging time of τ, and the RMS value is the two-sample or Allan deviation for that averaging time. If there are *N* time-difference data with indices 1, 2, …, *N*, then the Allan variance at averaging time τ is defined as the average of the $N-2$ calculations as

$$\sigma_y^2(\tau)=\frac{1}{2(N-2)\tau^2}\sum_{k=2}^{N-1}\left(x_{k+1}-2x_k+x_{k-1}\right)^2, \tag{59.4}$$

and the Allan deviation is the square root of this value. Since the time-difference data are equally spaced in time, we can use the same data to compute the estimate of the Allan variance for any multiple of the sampling interval, τ:

$$\sigma_y^2(m\tau)=\frac{1}{2(N-2m)(m\tau)^2}\sum_{k=2}^{N-2m+1}\left(x_{k+2m-1}-2x_{k+m-1}+x_{k-1}\right)^2. \tag{59.5}$$

The normalization is defined so that the Allan variance has the same value as the classical variance in the case of a random time-difference noise process, which we will discuss later. This type of process is often called “white phase noise.” The number of terms that contribute to the sum in Equation 59.5 decreases as *m* is made larger so that the estimates for large values of *m* are likely to exhibit more variation from one set of observations to another one. The maximum value of *m* that is used in Equation 59.5 is often limited to *N*/3 for this reason.

It is important to emphasize that the two-sample Allan deviation is a measure of frequency *stability*—not frequency accuracy. Frequency stability is obviously a very desirable quality for a clock, and many real-world devices are characterized in this way. It is clearly not sufficient to characterize devices that are used to provide standards of time, time interval, or frequency.

A very powerful technique is to examine the dependence of the Allan variance on the averaging time, τ, because this dependence can provide insight into the noise processes that drive the magnitude of the Allan deviation. To take a simple example, suppose that the device under test has a true constant frequency offset with respect to the perfect device used for the calibration. If this constant fractional frequency offset is *f*, then the measured time differences at times $k\tau$ in the absence of any noise processes would be

$$x_k = fk\tau + x_0, \tag{59.6}$$

where x_0 is the time difference between the device under test and the perfect clock when $k=0$. Then Equation 59.2 would estimate the average frequency as

$$y_k = \frac{x_k - x_{k-1}}{\tau} = f, \tag{59.7}$$

which is independent of *k* so that the estimate of the Allan variance is 0 for all averaging times. As expected, the Allan variance is 0 for a clock with a constant offset frequency, and it provides no information on the magnitude of this frequency.

A more interesting example is to suppose that the device under test had a constant frequency offset as in the previous example but that the time-difference measurements are affected by a random noise process that might originate in the measurement hardware and not in the clock itself. The time-difference measurements in this case would be given by

$$x_k = fk\tau + x_0 + \varepsilon_k, \tag{59.8}$$

where the noise contribution is characterized by a zero-mean signal with a well-defined variance:

$$\begin{gathered}\langle \varepsilon_k \rangle = 0 \\ \langle \varepsilon_j \varepsilon_k \rangle = \sigma^2 \delta(j-k)\end{gathered} \tag{59.9}$$

and δ is the Dirac delta function, which is one if its argument is zero and zero otherwise. The deterministic contributions to the summation in Equation 59.4 or 59.5 cancel as in the previous example, and what is left is a sum that is proportional to the variance of the noise process but independent of the summation index, *k*. The Allan variance therefore decreases as the reciprocal of the square of the sampling interval, and this conclusion does not depend on the magnitudes of either the deterministic or stochastic contributions, provided only that the data satisfy Equations 59.8 and 59.9. A plot of the logarithm of the Allan variance as a function of the logarithm of the measurement interval would have a slope of −2. If the frequency of the oscillator was not exactly constant but varied with the index *k*, then the summations in Equation 59.4 or 59.5 will contain a contribution from the deterministic terms in the time differences that is some function of the interval between the measurements. The log–log plot will have a slope that is greater than −2, and the exact value of the slope will depend on the details of the noise process that is driving the frequency fluctuations. In general, the slope of the log–log plot of the Allan variance as a function of the measurement interval is an indicator of the type of noise process that is contributing to the measured time differences. The details of this relationship and the usefulness of other variances that are related to the simple Allan variance discussed here are described in the literature [3].

The frequency dispersion estimated by the Allan deviation generates a corresponding time dispersion. The time dispersion is usually called $\sigma_x(\tau)$, and it is formally defined in terms of the modified Allan variance, which is described in Ref. [3].

We can provide a simple estimate of the time dispersion in the case of the simple white phase noise example that we considered in the previous paragraph. If we correct the measured time differences by the deterministic Equation 59.6, the residual time dispersion for any measurement is simply ε_k, a random process defined by Equations 59.8 and 59.9. If we substitute Equations 59.8 and 59.9 into Equation 59.4, then the summation is simply a sum of $N-2$ identical terms and

$$\sigma_y^2(\tau) = 3\frac{\langle \varepsilon_k^2 \rangle}{\tau^2} \tag{59.10}$$

or

$$\sqrt{\langle \varepsilon_k^2 \rangle} = \frac{\sigma_y(\tau)\tau}{\sqrt{3}}, \tag{59.11}$$

which relates the statistical RMS time dispersion to the Allan deviation for the case of white phase noise. Since the Allan deviation for white phase noise varies as the reciprocal of the measurement interval, the estimate of the RMS time dispersion does not depend on the interval between measurements. This is not a surprising result, since it follows directly from the assumption of the statistics of the noise process defined in Equation 59.9. A more conservative estimate is to take the RMS time dispersion as simply the product of the time interval between two measurements and the Allan deviation at that time interval, and I will generally use this more conservative estimate in the following discussion, since the constant in the denominator of Equation 59.11 is valid only for white phase noise.

59.3.3 Limitations of the Statistics

Both MTIE and the Allan variance (including a number of variants that we have not discussed) are measures of stability, but they define stability in different ways. Neither statistic is sensitive to a constant time offset, but the value of MTIE is sensitive to a constant frequency offset, whereas the value of the Allan variance is not. Therefore, the Allan variance is a measure of the predictability of the future time difference of a clock based on its past performance, and arbitrarily large *constant* frequency offsets do not degrade the prediction.

However, a clock with some other *deterministic* (*but not constant*) frequency offset produces time-difference values that are just as well determined as a clock with only a constant frequency offset, but the Allan variance treats the two very differently. The frequencies of many types of oscillators (e.g., hydrogen masers) can be approximated as varying linearly with time, and the Allan variance of the time-difference measurements from such devices (which have a quadratic dependence on the time) does not give a realistic estimate of their stability if by stability we mean how well can a future time difference be estimated based on previous performance. For this reason, it is common to estimate and remove a quadratic function of the time from these data before they are analyzed to compute the Allan variance. This process of "prewhitening" the data is well known in the statistical literature and is often implemented using a Kalman filter [4] and in power-spectral analysis [5].

Many oscillators have frequencies whose fluctuations can be more easily (and intuitively) characterized in the Fourier frequency domain rather than as a stationary process in the time domain, which is a basic assumption of the Allan variance. For example, an oscillator whose frequency was sensitive to ambient temperature could be expected to exhibit a diurnal frequency variation if it was operated in an environment without tight control of the ambient temperature. The Allan variance calculation will model these time differences as a stationary noise process, and it usually reports a large value at a time interval corresponding to roughly one-half of the period of the driving process. This is not too difficult to interpret correctly if there is only one such contribution, but it can be ambiguous if there are several "bright lines" in the power spectrum of the time differences, and prewhitening the data is particularly important in this case.

Finally, it is important to keep in mind that the calculation of the Allan variance is based on a particular method for averaging the time differences. Therefore, while the *slope* of a log–log plot of the Allan deviation as a function of averaging time provides insight into the underlying noise processes that drive the time differences, the *value* of the Allan deviation at any averaging time is useful as an indicator of the performance to be expected from the device only if the clock is used in a manner that is consistent with the averaging procedure that is part of the definition. For example, the definition of the Allan variance that we have used is based on data that are equally spaced in time with no dead time between the measurements. There are other statistics that are related to the simple Allan variance we have discussed (the modified Allan variance (Mod Avar), as an example, and its close relative the time variance (TVAR)), and these statistics have more complex averaging procedures, which are less likely to be used in a real application. Although the simple Allan variance has some limitations in identifying the underlying noise type in some circumstances, its definition is often closer to the way a device will actually be used, and it is often the preferred analysis tool for this reason.

59.4 CHARACTERISTICS OF DIFFERENT TYPES OF OSCILLATORS

The frequency stability of an oscillator can be realized either actively, where the discriminator is actively oscillating at a resonant frequency derived from its characteristics, or passively, where the frequency of an oscillator is locked to the resonant response of a passive discriminator.

A quartz crystal oscillator is generally an active device, because the frequency is determined by the mechanical resonance in the quartz that is excited by an external power source. Atomic frequency standards fall into both categories. Atomic clocks that use a transition in cesium or rubidium as the frequency reference generally fall into the second category, where the discriminator is passive and is interrogated by a separate oscillator that is locked to the peak of the transition probability of the clock transition. Hydrogen masers can be either active or passive. In both types of devices, the actual output frequency is generally a function both of the resonant frequency of the discriminator and the method that is used to interrogate it. The parameters of the interrogation method may have a dependence on ambient temperature or may vary with time so that even nominally identical oscillators generally have different output frequencies. (A primary frequency standard is designed so that these perturbing influences are minimized, and the residual perturbations are estimated by means of various ancillary measurements.)

The vast majority of oscillators currently in use are stabilized using a mechanical resonance in a quartz crystal. Newer devices are stabilized using a microelectromechanical (MEMS) device [6]. The quartz crystals used as the frequency reference in inexpensive wristwatches can have a frequency accuracy of 1 ppm and a stability of order 10^{-7}. (A frequency accuracy of 1 ppm translates into a time dispersion of order 0.09 s/day.) These devices are generally sensitive to the ambient temperature so that substantially better performance can be realized using active temperature control. The moderate stability of the frequency is exploited in many control applications and in the operation of Internet time servers, as we will discuss later. In spite of some very clever techniques that have been developed to improve the long-term stability of the frequency of these devices, none of the methods can totally eliminate the sensitivity of the frequency to environmental perturbations, to stochastic frequency fluctuations that are hard to model, and to a dependence on the details of manufacture that are hard to replicate.

Atomic frequency standards use an atomic or molecular transition as the frequency discriminator in an attempt to address the limitations of the frequency stability of mechanical devices that I discussed in the previous paragraph. The atoms can be used in a passive or active configuration. In the passive configuration, the atoms are prepared in the lower state of the clock transition and are illuminated by the output from a separate variable-frequency oscillator. The frequency of the oscillator is locked to the maximum in the rate of the clock transition. There are a number of different methods for preparing the atoms in the lower state and for detecting the transition to the upper state. The details are described in the literature [7].

In the active configuration, the atoms are pumped into the upper state, and the radiation emitted when they decay to the lower state stimulates other atoms to decay by stimulated emission. The oscillating frequency is determined by the atomic transition frequency and by the properties of the cavity that is used to trap the radiation and provide the ambient field that induces stimulated emission. Most lasers and some hydrogen masers work this way. The cavity of an active hydrogen maser is generally tuned to improve the stability of the output frequency [8].

In both active and passive hydrogen masers, the output signal is generated by an oscillator (typically a quartz crystal device) whose frequency is locked to the transition frequency of the atoms. For sufficiently short averaging times (typically less than about 0.1 s), the stability of the output frequency is determined by the properties of the quartz oscillator and by the process used to generate the output ticks. The spectrum is generally white phase noise. At longer averaging times (greater than about 10 s), the stability of a maser is generally limited by the thermal noise in the oscillating field in the cavity and in the control loop that is used to lock the output oscillator. This is usually white frequency noise. Other types of atomic standards are generally operated in the passive configuration and have similar statistics.

59.5 COMPARING CLOCKS AND OSCILLATORS

Comparing the times of two clocks is a simple process in principle, but the process becomes more complicated as the resolution of the measurement increases. I will discuss two classes of comparisons. In the first situation, the devices to be compared are in the same laboratory so that the signals from both clocks are available locally, and we don't have to consider the characteristics of a transmission network. In the second configuration, one of the clocks is at a remote location so that the statistics of the time comparison must include the effects of the transmission network.

The simplest time comparison is simply a one-time measurement. We read the time difference between the two clocks, using a time-interval counter, for example, and we use the measurement to adjust the time of one of them either by making an adjustment to its physical output or by noting its time offset for future administrative corrections. The implication of the process is that the reference clock is much more accurate than the device whose time difference we are measuring, and we need not consider the possibilities that the reference clock is broken or that the time comparison had a significant measurement error. Furthermore, this simple "set it and forget it" process does not provide any insight into the statistics of the device under test so that we have no way of knowing how rapidly its time will diverge from the correct time after it has been set. Thus, we have no way of knowing how often to repeat the measurement process in order to have the device being calibrated maintain some specified level of accuracy. These considerations lead to a more sophisticated measurement program.

To simplify matters, we will again assume that the clock under test is being compared to a second clock that is so much more accurate that it can be considered as perfect from the perspective of the measurement process. We will model the time differences of the clock under test by a combination of deterministic and stochastic parameters. The deterministic time differences are generally specified in terms of three parameters, the initial time difference, the frequency offset, and the frequency aging.

This formulation leads to a quadratic relationship with constant parameters whose independent variable is the elapsed time since the start of the measurement. This formulation is not adequate for most real devices and measurement processes.

In the first place, the frequency offset and frequency aging are not manifest constant parameters for any real device, and treating them as constants is neither adequate nor optimum. In addition, many applications depend on real-time estimates of the parameters, and it becomes increasingly cumbersome to reevaluate a static quadratic form of the modeled time differences each time a new data point is measured. This type of analysis also requires that we save all of the measurements since the start of the experiment. Therefore, most analyses use an iterative form of the estimate, in which the current time difference is modeled based on the parameters estimated from the previous measurements. The previous measurements themselves are not needed.

In this method, we estimate the time difference at time t_k in terms of the previous data as

$$\hat{x}_k = x_{k-1} + y_{k-1}\Delta t + \frac{1}{2}d_{k-1}\left(\Delta t\right)^2, \tag{59.12}$$

where x_k, y_k, and d_k are the time difference, frequency offset, and frequency aging at time t_k and $\Delta t = t_k - t_{k-1}$. It is generally easiest to use measurements that are equally spaced in time, but the formulation of Equation 59.12 is valid whether or not this is the case.

The measurement process measures the time difference at time t_k and returns the measured value X_k. The difference between the value we predicted and the value we observed is

$$\delta_k = X_k - \hat{x}_k. \tag{59.13}$$

In the absence of any noise contributions and assuming that the initial time difference, the frequency offset, and the frequency aging are perfectly constant values, Equation 59.12 has only three parameters so that the time differences in Equation 59.13 will converge to zero after three measurement cycles even if we are totally ignorant of the initial values of these parameters. But now we return to the real world, where the measurements have a stochastic component and the clock parameters are not absolutely constant. The goal of a real-world measurement process is to partition the time differences obtained in Equation 59.13 into a deterministic portion, which we use to update our estimates of the parameters in Equation 59.12, and a stochastic contribution, which we attenuate by averaging or ignoring completely.

The measurement process in the previous paragraph cannot succeed in the most general case because there is only one observable (Eq. 59.13) and multiple parameters that must be determined. The process can succeed in practice because the deterministic parameters in Equation 59.12 change slowly with time so that it is possible to treat them as substantially constant over short averaging times. Stated another way, the process of modeling the time differences by means of Equation 59.12 will succeed if and only if the time interval between the measurements is short enough to validate this assumption. In the next section, I discuss this requirement quantitatively.

59.6 NOISE MODELS

As I mentioned in the previous section, the limitations of the method come from our ignorance of the partition of the measured data into stochastic and deterministic components. To get insight into the characterization of the noise, it is useful to model an oscillator as a passive resonant system (such as an atomic transition) that is interrogated by a separate oscillator. The frequency of the oscillator is locked to the peak in the resonance response, and the output of the oscillator is used to generate the "ticks" that are used to drive the time display. Most oscillators that are stabilized using a transition in cesium or rubidium are configured this way. Lasers and most crystal oscillators cannot be modeled so easily because the oscillation frequency depends on a complicated combination of the gain of an amplifier and the phase shift in a resonant feedback loop. Nevertheless, even these types of oscillators can be reasonably well characterized using the machinery that I will discuss in the following sections.

59.6.1 White Phase Noise

A common method for generating the ticks is to generate an output pulse each time the sine wave of the oscillator passes through zero with a positive slope. Real zero-crossing detectors have some uncertainty in the exact trigger point, and this uncertainty is often represented as an equivalent noise voltage at the input to the circuit. I will designate this noise voltage as V_n. This noise contribution is modeled as a zero-mean random process that is unrelated to the true input signal. If the output of the oscillator has amplitude A and an angular frequency ω, then the noise voltage, whose amplitude is much smaller than the amplitude of the signal, introduces a time jitter in the determination of the time of the zero-crossing whose amplitude is

$$\Delta t = \frac{V_\mathrm{n}}{A\omega}. \tag{59.14}$$

This fluctuation in the times of the output pulses is not associated with any frequency fluctuations in the oscillator itself. The noise voltage is inherent to the discriminator and has nothing to do with the input signal so that the magnitude of the time fluctuation for any measurement is unrelated to the impact for any other measurement. As we showed in the previous section, the Allan deviation for this type of noise varies as the reciprocal of the interval between measurements. Since the fluctuation does not arise in the oscillator itself, correcting the apparent frequency jitter caused by this noise by steering the parameters of the oscillator (e.g., its frequency) is not the optimum strategy. I will discuss this point in more detail later. For now we note that the time jitter defined by Equation 59.14 has a mean of zero. From the physical perspective, there is an underlying "true" time difference that can be estimated by averaging multiple measurements. The standard deviation of the estimate decreases as more measurements are performed, and the mean value at any time is an unbiased estimate of the true time difference.

59.6.2 White Frequency Noise

We next consider the control loop that locks the oscillator frequency to the peak of the resonance response. A common method of detecting the peak in the response is to lock the oscillator to the zero of the first derivative of the resonance response function. The first derivative signal is generated by applying a small modulation to the oscillator frequency and synchronously detecting the amplitude of the response of the resonant system at the frequency of the applied frequency dither. An alternate method locks the oscillator to a frequency that is between two frequencies above and below the resonance where the upper and lower frequencies are measured as the values where the response has fallen to some specified fraction of the peak. (This is often implemented by means of a zero-mean, bipolar, square wave dither of the oscillator frequency.) The input to the discriminator is the difference between the response at the higher frequency and the response at the lower one. Both methods introduce a deterministic dither in the frequency output of the oscillator, which must be removed before the signal is used.

In either method, the discriminator locks the oscillator to a point where some voltage goes through zero. When the frequency is locked, the magnitude of the voltage specifies how much the frequency differs from the desired lock point, and the sign of the voltage specifies whether the frequency output is higher or lower than the desired operating point.

The response of this discriminator is limited by the same noise problems that I discussed in the previous section, but now it is the frequency of the oscillator rather than the output of a pulse that is affected by the equivalent noise at the input of the discriminator. The same argument as in the previous section shows that the result is a random frequency modulation. As in the previous section, the noise contribution is assumed to be a zero-mean random process so that frequency jitter is about a "true" value. This noise is identified by the slope of –0.5 in the log–log plot of the Allan deviation.

These random frequency fluctuations are integrated to produce time dispersion. Since the frequency fluctuations are characterized by a random process with a mean of zero, the impact on the time differences is a random walk with a variable step size, and so white frequency fluctuations are often described as a random walk in time (or phase, which is the same thing). Unlike the white phase noise case discussed in the previous section, the impact of white frequency noise on the measured time differences depends on the averaging time through Equation 59.12.

Since the impact of white frequency noise on the measured time differences is a function of the averaging time whereas white phase noise is independent of it, at least in principle it is possible to distinguish between the two by an appropriate choice of the averaging time; white frequency noise can be neglected at very short averaging times, and white phase noise becomes negligible as the averaging time is increased.

59.6.3 Long-Period Effects: Frequency Aging

The effects I have discussed in the previous sections are modeled as being driven by stochastic noise processes that are assumed to have zero means. That is, there is an underlying "true" time difference in the white phase noise domain, and there is an underlying "true" frequency offset in the white frequency noise domain. These models are useful because the behavior of many oscillators is well characterized in these terms.

Although there are oscillators that also exhibit a stochastic frequency aging that can also be modeled as a random zero-mean process just as we did previously for time and for frequency, the frequency aging of many oscillators is often a combination of a random function that is only approximately characterized as a zero-mean process combined with an approximately constant aging value.

There are a number of processes that can produce frequency aging that varies only very slowly and is approximately constant over many measurement cycles. For example, the transition frequency that is used as the reference frequency in an atomic clock is generally sensitive to external electric and magnetic fields (the DC Stark and Zeeman effects, respectively), to collisions between the atoms, to frequency shifts that result from the interaction with the probing field (the AC Stark effect), and to many other effects. These perturbing influences may have long-period variations that are translated into long-term frequency aging of the oscillator. The mechanical properties of quartz crystals, which determine the resonant frequency of a quartz crystal oscillator, often have similar long-period changes. Stochastic frequency aging may be caused by more rapid variation in any of these parameters and in other effects whose quantitative driving term admittance is not known.

From Equation 59.12, we can see that a constant frequency aging would produce time dispersion proportional to the square of the time interval between the measurements. Based on the relationship between time dispersion and Allan deviation, the Allan deviation for this type of aging would have a slope of +1 on a log–log plot of Allan deviation with respect to averaging time. The plot of the log–log plot of the Allan deviation with respect to averaging time has a slope of +0.5 for stochastic frequency aging.

Since the time dispersion due to frequency aging varies as the square of the interval between measurements, it is often possible to partition the measured variation in the time differences into three domains—a very short domain where white phase noise dominates the variance, an intermediate domain where frequency noise is the main contributor, and a long-period domain where frequency aging dominates.

59.6.4 Flicker Noise

In addition to the white time, frequency, and aging processes that I discussed in the previous sections, there is another contribution to the variance of time-difference measurements that cannot be easily characterized using a simple model analogous to the ones presented in the previous sections. I will characterize this type of noise process by contrasting it to the white phase noise and white frequency noise processes that I discussed before.

If a time-difference measurement process can be characterized as being limited by pure white phase noise, then there exists an underlying "true" time difference between the two devices. The measurements scatter about the true time difference, but the distribution of the measurements (or at worst the mean of a group of them) can always be characterized by a simple Gaussian distribution with only two parameters: a mean and a standard deviation. We can improve our estimate of the mean time difference by averaging more and more observations, and this improvement can continue forever in principle. There is no optimum averaging time in this simple situation—the more data we are prepared to average, the better our estimate of the mean time difference will be.

The situation is fundamentally different for a measurement in which one of the contributing clocks is dominated by zero-mean white frequency noise. Now it is the frequency that can be characterized (at least approximately) by a single parameter—the standard deviation.

Suppose we measure the time differences between a perfect device and the device under test, where the device under test has the same nominal frequency as the perfect device, but its frequency stability is degraded by white frequency noise. If the time difference between the two devices at some epoch is $X(t)$, then, since the deterministic frequency difference is zero, we would estimate the time difference a short time in the future as

$$X(t+\tau) = X(t) + y(t)\tau, \tag{59.15}$$

where $y(t)$ is the instantaneous value of the white frequency noise of the device under test, and

$$\langle y(t) \rangle = 0, \tag{59.16}$$

$$\langle y^2(t) \rangle = \varepsilon^2. \tag{59.17}$$

Since $y(t)$ has a mean of zero by assumption, Equation 59.15 predicts that the time difference at the next instant will be distributed uniformly about the current value of X, and the mean value of $X(t+\tau)$ is clearly $X(t)$. In other words, for a clock whose performance is dominated by zero-mean white frequency noise, the optimum prediction of the next measurement is exactly the current measurement with no averaging. Note that this does not mean that our prediction is that

$$X(t+\tau)=X(t) \tag{59.18}$$

but rather that

$$\langle X(t+\tau)-X(t)\rangle=0, \tag{59.19}$$

which is a much weaker statement because it does not mean that our prediction will be correct but only that it will be unbiased on the average. This is, of course, the best that we can do under the circumstances. The frequencies in consecutive time intervals are uncorrelated with each other by definition, and no amount of past history will help us to predict what will happen next. The point is that this is the opposite extreme from the discussion earlier for white phase noise, where the optimum estimate of the time difference was obtained with infinite averaging of older data.

Clearly, there must be an intermediate case between white phase noise and white frequency noise, where some amount averaging would be the optimum strategy, and this domain is called the "flicker" domain. Physically speaking, the oscillator frequency has a finite memory in this domain. Although the frequency of the oscillator is still distributed uniformly about a mean value of zero, consecutive values of the frequency are not independent of each other, and time differences over sufficiently short times are correlated. Both the frequency and the time differences have a short-term "smoothness" that is not characteristic of a simple random variable, and this smoothness is often mistaken for a pseudodeterministic variation. Flicker phase noise is intermediate between white phase noise and white frequency noise, and we would therefore expect that the Allan deviation of the time-difference data would have a dependence on averaging time that is midway between white and random walk processes. In fact, the simple Allan variance that we have discussed cannot distinguish between white phase noise and flicker phase noise. The more complicated modified Allan variance (Ref. [3]) is needed for this purpose, and the slope of this variance for flicker phase noise is indeed midway between the slopes for white and random walk processes.

The same kind of discussion can be used to define a flicker frequency noise that is midway between white frequency noise and white aging (or random walk of frequency). The underlying physical effect is the same, except that now it is the frequency aging that has short-period correlations. We could think of a flicker process as resulting from a very large number of very small jumps—not much happens in the short term because the individual jumps are very small, but the integral of them eventually produces a significant effect. The memory of the process is then related to this integration time.

The slopes of the log–log plot of both the Allan deviation and the modified Allan deviation are zero for flicker frequency noise. In other words, the estimate of the frequency does not improve with longer averaging times. The Allan deviation for these averaging times is often called the “flicker floor” of the device for this reason.

Data that are dominated by flicker-type processes are difficult to analyze. They appear deterministic over short periods of time, and there is a temptation to try to treat them as white noise combined with a deterministic signal—a strategy that fails once the coherence time is reached. On the other hand, they are not quite noise either—the correlation between consecutive measurements provides useful information over relatively short time intervals, and short data sets can be well characterized using standard statistical measures. However, the variance at longer periods is much larger than the magnitude expected based on the short-period standard deviation.

The finite-length averages that we have discussed can be realized using a sliding window on the input data set. This is simple in principle but requires that previous input data be stored; an alternative way of realizing essentially the same transfer function is to use a recursive filter on the output values. For example, suppose that the optimum averaging time is T. Then, if Y_{k-1} is an estimate of some parameter at time t_{k-1} and if y_k is the new data point received at time t_k, then we would estimate the update to Y at time t_k by means of

$$Y_k = \frac{wY_{k-1} + y_k}{w+1}, \tag{59.20}$$

where w is a dimensionless parameter given by

$$w = \frac{t_k - t_{k-1}}{T}. \tag{59.21}$$

The time interval between measurements is chosen so that $w \leq 1$—there is at least one measurement in the optimum averaging time. This method is often used in time scales, since these algorithms are commonly implemented recursively. Both the recursive and nonrecursive methods could be used to realize averages with more complicated transfer functions. These methods are often used in the analysis of data in the time domain [9].

59.7 MEASURING TOOLS AND METHODS

The oscillators that we have discussed before generally produce a sine wave output at some convenient frequency such as 5 MHz. (This frequency may also be divided down internally to produce output pulses at a rate of 1 Hz. Crystals designed for wristwatches and some computer clocks often operate at 32,768 Hz—an exact power of 2, which simplifies the design of these 1 Hz dividers.) A simple quartz crystal oscillator might operate directly at the desired output frequency; atomic standards relate these output signals to the frequency appropriate to the reference transition by standard techniques of frequency multiplication and division. The measurement system thus operates at a single frequency independent of the type of oscillator that is being evaluated. The choice of this frequency involves the usual trade-off between resolution, which tends to increase as the frequency is made higher, and the problems caused by delays and offsets within the measurement hardware, which tend to be less serious as the frequency is made lower.

Measuring instruments generally have some kind of discriminator at the front end—a circuit that defines an event as occurring when the input signal crosses a specified reference voltage in a specified direction. Examples are 1 V with a positive slope in the case of a pulse or zero volts with a positive slope in the case of a sine wave signal. The trigger point is chosen at (or near) a point of maximum slope so as to minimize the variation in the trigger point due to the finite rise time of the waveform.

The simplest method of measuring the time difference between two clocks is to open a gate when an event is triggered by the first device and to close it on a subsequent event from the second one [10]. The gate could be closed on the very next event in the simplest case, or the *N*th following one could be used, which would measure the average time interval over *N* events. The gate connects a known high-frequency oscillator to a counter, and the time interval between the two events is thus measured in units of the period of this oscillator. The resolution of this method depends on the frequency of this oscillator and the speed of the counter, while the accuracy depends on a number of parameters including the latency in the gate hardware and any variations in the rise time of the input waveforms. The resolution can be improved by adding an analog interpolator to the digital counter, and a number of commercial devices use this method to achieve subnanosecond resolution without the need for a reference oscillator whose frequency would have to be at least 1 GHz to realize this resolution without interpolation.

In addition to the limit on the resolution of time-difference measurements that results from the maximum rate of the oscillator that drives the time-interval counter, time-difference measurements using fast pulses have additional problems. Reflections from imperfectly terminated cables may distort the edge of a sharp pulse, and long cables may have enough shunt capacitance to round the rise time by a significant amount. In addition to distorting the waveforms and affecting the trigger point of the discriminators, these reflections can alter the effective load impedance seen by the oscillator and pull it off frequency. Isolation and driver amplifiers are usually required to minimize the mutual interactions and complicated reflections that can occur when several devices must be connected to the same oscillator, and the delays through these amplifiers must be measured. These problems can be addressed by careful design, but it is quite difficult to construct a direct time-difference measurement system whose measurement noise does not degrade the time stability of a top-quality oscillator, and other methods have been developed for this reason. Averaging a number of closely spaced time-difference measurements is usually not of much help because these effects tend to be slowly varying systematic offsets, which change only slowly in time, and so have a mean that is not zero over short times.

Many measurement techniques are based on some form of heterodyne system. The sine wave output of the oscillator under test can be mixed with a second reference oscillator that has the same nominal frequency, and the much lower difference frequency can then be analyzed in a number of different ways. If the reference oscillator is loosely locked to the device under test, for example, then the variations in the phase of the beat frequency can be used to study the fast fluctuations in the frequency of the device under test. The error signal in the lock loop provides information on the longer-period fluctuations. The distinction between "fast" and "slow" would be set by the time constant of the lock loop. As usual, we assume that any fluctuations in the reference oscillator are small enough to be ignored.

This technique can be used to compare two oscillators by mixing a third reference oscillator with each of them and then analyzing the two difference frequencies using the time-interval counter discussed earlier. In the "dual-mixer" version of this idea developed at NIST [11], this third frequency is not an independent oscillator but is derived from one of the input signals using a frequency synthesizer. The difference frequency has a nominal value of 10 Hz in this case. The time interval counter runs with an input frequency of 10 MHz and can therefore resolve a time interval of 10^{-6} of a cycle. Since the heterodyne process preserves the phase difference between the two signals, a phase measurement of 10^{-6} of a cycle is equivalent to a time-interval measurement with a resolution of 0.2 ps at the 5 MHz input frequency. It would be very difficult to realize this resolution with a system that measured the time difference directly at the 5 MHz input frequency or by measurements of the 1 Hz pulses derived from this reference frequency by a process of digital division.

All of these heterodyne methods share a common advantage: the effects of the inevitable time delays in the measurement system are made less significant by performing the measurement at a lower frequency where they make a much smaller fractional contribution to the periods of the signals under test. Furthermore, the resolution of the final time-interval counter is increased by the ratio of the input frequencies to the output difference frequencies (5 MHz to 10 Hz in the NIST system). This method does not obviate the need for careful design of the front-end electronics—the increased resolution of the back-end measurement system places a heavier burden on the high-frequency portions of the circuits and the transmission systems. As an example, the stability of the NIST system is only a few ps—about a factor of 10 or 20 poorer than its resolution. Some of the factors that degrade the stability of a channel in a dual-mixer system are common to all of the channels in a single chassis so that the differential stability of a pair of channels (which is what drives an estimate of the Allan variance) can be better than the stability of each channel alone.

Heterodyne methods are well suited to evaluating the frequency stability of an oscillator, but they often have problems in measuring absolute time differences because they usually have an integer ambiguity offset—an unknown integer number of cycles of the input frequencies between the cycles that trigger the measurement system and the cycles that produce the 1 Hz output pulses that are the output "on-time" signals. For example, the time difference between two clocks measured using the NIST dual-mixer system is offset with respect to measurements made using a system based on the 1 Hz pulse hardware by an arbitrary number of periods of the 5 MHz input frequency (i.e., some multiple of 200 ns). To further complicate the problem, this offset generally changes if the power is interrupted or if the system stops for any other reason.

The offset between the two measurement systems must be measured initially, but it is not too difficult to recover it after a power failure, since the time step must be an exact multiple of 200 ns. Using the last known time difference and frequency offset, the current time can be predicted using a simple linear extrapolation. This prediction is then compared with the current measurement, and the integer number of cycles is set (in the software of the measurement system) so that the prediction and the measurement agree. This constant is then used to correct all subsequent measurements. The lack of closure in this method is proportional to the frequency dispersion of the clock multiplied by the time interval since the last measurement cycle, and the procedure will unambiguously determine the proper integer if this time dispersion is significantly less than 200 ns. This criterion is easily satisfied for a rubidium standard if the time interval is less than a few hours; the corresponding time interval for cesium devices is generally at least a day.

59.8 MEASUREMENT STRATEGIES

In the previous section, I discussed a number of methods that can be used to measure the time differences between two devices. All measurement techniques have some residual noise that appears as jitter in the time-difference measurements. I will assume for now that this jitter can be characterized as white phase noise. That is, it is a pure random process, and the impact on any measurement can be fully characterized by a distribution with a mean of zero and a standard deviation that is a property of the measurement system and does not depend on temperature, aging, or any of the other limitations that affect most real-world systems. Specifically, the noise of the measurement process can be characterized by relationships similar to Equation 59.9, previously.

With this assumption, the optimum strategy for estimating the time difference between a device under test and a second device that we think of as perfect (or at least very much better than the device under test) is to make repeated measurement of the time difference and to average the results. This number of measurements that can be combined into a single average will be limited by the assumption that the measurements differ only by the white phase noise of the measurement process, perhaps combined with the white phase noise of the oscillator itself as described earlier.

I will model the time differences between the device under test and our perfect reference device in terms of an initial time difference, a frequency offset, y, and a frequency aging, d:

$$x = x_0 + yt + \frac{1}{2}dt^2, \tag{59.22}$$

and my initial goal is to determine the time difference, x_0, in the presence of the white phase noise of the measurement process, which has a standard deviation ε_m.

Equation 59.22 is not very useful as a tool to estimate the time difference directly. For example, it cannot be used to estimate the parameters x_0, y, and d by applying standard least squares to an ensemble of measurements, because the parameters y and d are not constants but change slowly with time and have both stochastic and deterministic contributions. The least squares analysis will provide numerical estimates in a formal sense, but the estimates are neither physically significant nor statistically stationary, since Equation 59.22 is trying to fit a single quadratic relationship to an ensemble of data in which the parameters of the quadratic vary with the value of the independent variable. A simple least squares analysis does not have the flexibility to provide a robust estimate of parameters that are themselves statistical variables. The only exception might be for a very short data set where the principal contribution to the time dispersion is the white phase noise of the measurement process. The frequency offset and frequency aging can be taken to be approximately constant in this situation. However, we will use the iterative form of this relationship, Equation 59.12, in the more general case.

Since the noise of the measurement process is a random function that depends only on the characteristics of the measurement device and not on time or on any external perturbations, I can average the measured time differences to attenuate the effect of the measurement noise. I can continue to do this as long as the assumption that the time differences are randomly distributed about a "true" value as a result of the measurement noise alone. The duration of my average is time T, where T is given by

$$yT + \frac{1}{2}dT^2 \ll \varepsilon_m. \tag{59.23}$$

That is, the average can continue as long as the deterministic evolution of the time differences is much less than the noise of the measurement process so that the measurements are extracted from an ensemble of measurements that have the same time difference within the measurement uncertainty. An averaging time that satisfies this constraint will usually satisfy the weaker constraint that is driven by the noise in the frequency of the oscillator:

$$\sigma_y(T)T \ll \varepsilon_m, \tag{59.24}$$

where $\sigma_y(T)$ is the two-sample Allan deviation of the device under test for an averaging time of T. Equation 59.24 is a weaker condition because the frequency stability of most oscillators is generally better than the frequency accuracy.

These principles can be illustrated with a numerical example. Suppose that we wish to characterize a rubidium oscillator with a time-difference system that has a measurement noise of 1 ns (10^{-9} s). Based on the generic type of the device, we might estimate that the oscillator has a deterministic fractional frequency offset of 5×10^{-11}. (Recall that fractional frequency offsets are dimensionless.) The deterministic frequency aging is about 4×10^{-18}/s (about 1×10^{-11}/month). If we consider the limit imposed by Equation 59.23, the first term requires that $T \ll 20$ s. The second term is negligible for that value of T so that the requirement of Equation 59.23 is primarily driven by a constant frequency offset with no deterministic aging. This is a common result, which we will discuss in greater detail as follows. The two-sample Allan deviation of a rubidium standard for an averaging time of about 20 s is of order 10^{-12} so that Equation 59.24 is also easily satisfied with this averaging time.

Thus, if we know only generic values of the parameters that characterize the device, we can average the time differences for something less than 20 s. If we decide to use 16 measurements of the 1 Hz pulses from the device, we could average the 16 measurements, and the standard deviation of the measurements would be improved by a factor of 4. This is about the best we can do without knowing more about the device. Note that we cannot make a robust estimate of the frequency yet because our measurements are dominated by the white phase noise of the measurement process.

If we average 16 measurements, the uncertainty in the time difference has been reduced to about 0.25 ns or 250 ps. If we continue to make time-difference measurements, we can no longer average them directly, since the distribution becomes increasingly driven by the deterministic frequency offset of the device. We enter the domain where the deterministic frequency offset is making a contribution to the time differences, and the measurements are now limited by a combination of the white frequency noise of the device and white phase noise of the measurement. Thus, the measurement strategy changes from simply averaging the measured time differences to estimating the average frequency offset (the first derivative of the measurements) as well. This intermediate case is difficult to handle with this simple method, since we do not know how to partition the variance of the time-difference data into a contribution of the phase noise of the measurement process and the frequency noise of the clock itself.

One way to handle this ambiguity is to reverse the inequality in Equation 59.24. For example, if we made a time-difference measurement every 10^4 s, the white phase noise of each of the measurements would still be only 1 ns, but the contribution of the frequency noise of the oscillator would now contribute about 10 ns. We have now moved into the complementary domain, where the variance of the time differences is dominated by the frequency noise of the oscillator and the contribution of the noise of the measurement process is small enough to be ignored. The contribution of the deterministic frequency aging to the time differences is of order $0.5 \times 10^{-18}\text{/s} \times 10^8\,\text{s}^2 = 0.05$ ns so that we can make the reasonable assumption that *all* of the observed variance can be modeled as white frequency noise of the oscillator. (The ensemble algorithm that is used at NIST and other national laboratories often operates in this measurement domain where the data can be modeled as pure white frequency noise to a good approximation.) The optimum strategy in this domain is to average the offset frequency estimates obtained from the first differences of the time-difference data divided by the interval between these values.

This approach can continue as long as the measurements are in the white frequency noise domain, and the strategy will give an increasingly accurate estimate of the deterministic offset frequency of the device. We can derive the lower bound of this domain from the white phase noise of the measurement process as we have done before, but we don't have enough information to specify the upper end of this domain uniquely. We can make a rough estimate by comparing the time dispersion resulting from the frequency fluctuations to the time dispersion driven by the deterministic frequency aging and defining the upper limit of the averaging time as the interval when these two contributions are equal. This upper limit to the averaging time is defined by

$$\sigma_y(T)T = \frac{1}{2}dT^2 \tag{59.25}$$

or

$$T = \frac{2\sigma_y(T)}{d}. \tag{59.26}$$

Equation 59.26 highlights a fundamental problem with the model that we are using. The two-sample Allan deviation of a typical rubidium standard is generally not less than 10^{-12} for intermediate averaging times and tends to increase at longer averaging times because nonwhite frequency fluctuations usually become important there. On the other hand, the deterministic frequency aging, *d*, is of order 10^{-18}/s so that the averaging time predicted by Equation 59.26 (where the contribution of deterministic frequency aging to the frequency fluctuations is equal to the stochastic contribution) is generally longer than 10^6 s. In other words, it is difficult to estimate the deterministic frequency aging of a device because the aging is masked by the stochastic frequency fluctuations for moderate averaging times. Furthermore, nonstatistical considerations are often important at longer averaging times, which makes the numerator of Equation 59.26 larger and the problem of estimating the deterministic aging more difficult. On the other hand, the frequency aging contributes to the time differences as the square of the time interval so that it will almost always dominate the time dispersion at sufficiently long times.

The conclusions of the preceding discussion are more general than the specific case that I discussed, and the result is that it is very difficult to characterize the deterministic component of the long-term performance of any oscillator. (A time scale, which is an ensemble of oscillators, is not immune to this problem.) The only difference among the different types of oscillators is where the "long-term" time domain begins. For example, the deterministic frequency fluctuation of a cesium standard is masked by the stochastic frequency fluctuations for almost any averaging time out to the life of the device (years), and cesium standards are generally modeled with no deterministic frequency aging for this reason. A hydrogen maser, on the other hand, has a deterministic frequency aging of order 10^{-16}/day (~10^{-21}/s), but its stochastic frequency fluctuations are small enough so that the frequency aging must be included in any model of the time differences. The frequency aging of a rubidium standard is often of order 10^{-11}/month (~3.9×10^{-18}/s) and can be ignored only for short averaging times.

The basis of this discussion is that it is possible to find measurement domains where only one type of noise process dominates the variance of the measured time differences. This idea is widely used in modeling oscillators that are members of a time scale ensemble. For example, the AT1 algorithm used to estimate the average time of an ensemble of clocks at the NIST laboratory in Boulder, Colorado, is designed based on this principle [12]. The measurement system used at NIST has a measurement noise on the order of 1 ps and a time interval between measurements of 720 s, and the ensemble algorithm is based on the premise that the variance of the time differences can be modeled as white frequency noise. The hydrogen masers that are members of the ensemble have a constant frequency aging that is determined outside of the ensemble algorithm. This parameter is treated as a constant by the algorithm because it is difficult to compute a statistically robust estimate of the aging because of the problems discussed previously.

The algorithm used by the International Bureau of Weights and Measures (the BIPM in French) to compute International Atomic Time (TAI) and UTC is also based on these principles. The measurement noise of the time differences of the clocks located at the various national timing laboratories is much larger than the local time-difference measurements at NIST so that the interval between measurements has to be increased proportionally to guarantee that the contribution of the measurement noise to the time differences is smaller than the contribution of the frequency variations of the clocks. However, this increase in the time interval between measurements increases the impact of the frequency aging of the masers that contribute to TAI and UTC, and the algorithm used by the BIPM has been modified to recognize this effect [13]. (see also [14].) As we would expect from the previous discussion, including an explicit frequency aging term improves the long-term stability of the time scale by bringing the model closer to the actual behavior of the clocks that are members of the ensemble.

59.9 THE KALMAN ESTIMATOR

The preceding discussion depended on the assumption that the time interval between measurements was a free parameter that could be adjusted at will based on statistical considerations. In each case, it was chosen so that the variance of the time-difference measurements could be modeled as arising primarily from a single source. However, there can be systems where the time interval between measurements is constrained by other factors to values that do not support this simplifying assumption. The variance of the time-difference measurements must be apportioned to more than one source in these configurations, but there is no way to do this using the machinery that we have developed so far. The Kalman estimator is one way of partitioning the variance in these situations, and I will discuss the general method in this section.

The Kalman estimator starts from the same recursive relationship for the time differences that we presented in Equation 59.12, but it adds two additional recursive relationships describing the evolution of the offset frequency and the frequency aging. The "Kalman state" of the clock is characterized by the values of these three parameters at any instant. The three equations that characterize the evolution of the state as a function of time are

$$\begin{aligned} x_k &= x_{k-1} + y_{k-1}\Delta t + \frac{1}{2}d_{k-1}\left(\Delta t\right)^2 + \xi \\ y_k &= y_{k-1} + d_{k-1}\left(\Delta t\right) + \eta \\ d_k &= d_{k-1} + \zeta . \end{aligned} \tag{59.27}$$

The time and offset frequency components are defined recursively with a deterministic contribution and a stochastic contribution (ξ and η, respectively). The frequency aging might have an initial constant value but usually is assumed to start at zero with only a stochastic variation, ζ. In each case, the stochastic contribution to the corresponding parameter is assumed to be a noise process that has a mean of zero and a variance that is initially known from other considerations, at least to first order. The noise contributions are assumed to be uncorrelated both in time and with each other. In other words, all three noise parameters satisfy relationships of the form

$$\begin{aligned} \left\langle \xi(t) \right\rangle &= 0 \\ \left\langle \xi(t)\xi(t') \right\rangle &= \xi^2\delta(t-t') \\ \left\langle \xi(t)\eta(t) \right\rangle &= 0 \end{aligned} \tag{59.28}$$

for all combinations of the parameters and for all t and t'.

The deterministic terms in Equation 59.27 describe how the state of the system evolves between measurements. The Kalman formalism can support measurements of any of the components of the state, but the most common arrangement is a measurement of the time difference of the clock with respect to a "perfect" reference as we discussed previously. In general, the measurement of the state component (the time difference in our discussion) will not agree with the value predicted from the previous state values by Equation 59.27, and the Kalman formalism provides a method of assigning the causes of this residual partially to updating the values of the deterministic parameters and partially to the noise parameters. The details of how to do this are in the literature [15]. Practical realizations of the Kalman algorithm applied to estimating the parameters of clocks often have the same difficulty in estimating the frequency aging term that we discussed earlier for basically the same reason—it is difficult to calculate a robust estimate of the frequency aging in the presence of stochastic frequency noise that is generally larger even at moderately large averaging times.

In addition, a Kalman algorithm can be no better than the accuracy of the model Equations 59.27 and 59.28, which are used to define the evolution of the clock state. The assumption that the stochastic inputs to the frequency and frequency aging are random, uncorrelated, zero-mean processes is often not an accurate description of the true state of affairs. Even when the model equations accurately describe the steady-state evolution of the state parameters, Kalman algorithms often have start-up transients that can be troublesome in some real-time applications.

59.10 TRANSMITTING TIME AND FREQUENCY INFORMATION

I now consider the problem of comparing the time difference between two clocks that are not at the same location so that the time difference must be measured by means of a channel that connects the two devices. There is no difference between this configuration and the one we described previously in principle, but there are a number of practical differences that make this type of measurement significantly more complicated.

The first issue that I will consider is the transmission delay that is introduced by the channel. The delay is at least 3 μs/km so that it must be measured in all but the simplest measurement programs of time differences. (If the goal of the measurement program is a comparison of the frequency difference between the two devices, then the magnitude of the channel delay is not important, provided only that it remains constant to the level required by the measurement process. Although this sounds like an easier requirement to satisfy, designing a channel that satisfies this requirement and verifying that it does so is often not significantly easier than going the whole way and measuring the delay itself.)

There are many applications where the delay is small enough to be ignored. For example, there are a very large number of wall clocks, wristwatches, and some process-control devices that are calibrated and set on time by means of the radio signals transmitted by the NIST radio station WWVB in Fort Collins, Colorado. The transmission delay, which can be on the order of milliseconds, is simply ignored in these devices, since the required accuracy is generally only on the order of 1 s. Simple devices that are synchronized using the signals from the Global Positioning System (GPS) often work the same way—the unmodeled portion of the transmission delay (e.g., due to the refractivity of the ionosphere and the troposphere) is of order 65 ns in this case, but it is still much smaller than the required accuracy, which may be only to the nearest millisecond or even only to the nearest second. I will not consider these applications here, and I will focus on the applications where some measurement of the channel delay is needed to satisfy the demands of the application.

I will describe three methods that are currently used to estimate the channel delay: (i) modeling the delay by means of ancillary parameters and measurements, (ii) the common-view method and its "melting-pot" variant, and (iii) two-way methods. I will also discuss the "two-color" method that is often used as an adjunct to one of the other methods when at least some of the path is through a medium that is dispersive. That is, the speed of the signal is a function of the frequency used to transmit the message. I will first discuss the general characteristics and assumptions of each of the methods, and I will then illustrate them with a more detailed discussion where the method is applied to a specific system.

The assumption that is implicit in this discussion is that the channel delay is not an absolute, unvarying parameter that can be measured once at the beginning by a process that may be complicated but needs to be done only once. There are some simple channels that satisfy this requirement—a coaxial cable between two parts of a building, for example.

A measurement of the delay of a long coaxial cable will often have some engineering complexity because coaxial cables are dispersive and have a frequency-dependent attenuation. In general, the attenuation, which is a result of the series inductance and shunt capacitance of the cable, increases with increasing frequency. Therefore, the rise time of a pulse transmitted on such a cable, which is a function of the high-frequency components of the signal, is increased as the signal propagates through the cable. The delay measurement is often further complicated by small impedance mismatches at the end of the cable, which cause reflections that interfere with the primary pulse used to measure the delay. These reflections can be exploited in the measurement process by leaving the far end of the cable unterminated and measuring the round-trip travel time of a pulse sent from the near end and reflected back from the open remote end. (The measurement of the travel time is not sensitive to the details of the signal that is transmitted along the cable, and a measurement that transmits a pseudorandom code instead of a pulse can have some technical advantages because it can be less affected by the attenuation of the high-frequency components of the test signal.) Whichever method is used, the delay is normally considered to be a constant that is a characteristic of the cable so that the measurement is normally a one-time effort. I will not consider this situation in detail, and I will focus on measurements where the delay cannot be measured once as a calibration constant.

59.10.1 Modeling the Delay

This method is based on the assumption that the channel delay can be estimated by means of some parameters that are known or measured from some ancillary measurement. For example, the geometrical path delay between a satellite and a receiver on the ground is estimated as a function of the position of the receiver, which has been determined by some means outside of the scope of the timing measurement, and the position of the satellite, which is transmitted by the satellite in real time.

There are only a few situations where the channel delay estimated from a model is sufficiently accurate to satisfy the requirements of an application. In most cases, the delay estimate has significant uncertainties, even if the model of the delay is well known. For example, the delay through the troposphere is a known function of the pressure, the temperature, and the partial pressure of water vapor, but these parameters are likely to vary along the path so that end-point estimates may not be good enough. This limitation is bad enough for a nearly vertical path from a ground station to a satellite, but the uncertainties become much larger for a lower-elevation path between two ground stations or from signals from a satellite that is near the horizon.

59.10.2 The Common-View Method

This method depends on the fact that there are two (or more) receivers that are equally distant from a transmitter. Since the two path lengths are the same, any signal sent from the transmitter arrives at the same time at both receivers. Each receiver measures the time difference between its local clock and the received signal, and these two measurements are subtracted. The result is the time difference between the clocks at the two receivers. In this simple arrangement, the accuracy of the time difference does not depend on the characteristics of the transmitted signal or the path delay.

In the real world, it is difficult to configure the two receivers so that they are exactly equally distant from the transmitter, and some means must be used to estimate the portion that is not common to the two paths. This estimate is not as demanding as estimating the full path delay so that the common-view method attenuates any errors in the estimate of the path delay.

There are also a number of subtle effects that we must consider when the two path delays of a common-view measurement are not exactly equal. If a single signal is used to measure the time difference, then the signal does not arrive at the two receivers at the same time, since the path delays are somewhat different. Therefore, any fluctuation in the characteristics of the receiver clocks during this time difference must be evaluated. On the other hand, if the measurement is made at the same instant as measured by the receiver clocks, then the signals that are measured by the two receivers did not originate from the satellite at the same instant of time. Therefore, the fluctuations in the characteristics of the satellite clock during this time interval must be evaluated.

Finally, a common-view measurement algorithm does not support casual associations among the receivers. The stations that participate in the measurement process must agree on the source to be observed and the time of the observation. There must also be a channel between the two receivers to transmit the measured time differences. On the other hand, there needs to be no relationship between the receivers and the transmitter, and the transmitter need not even know that it is being used as part of a common-view measurement process. Signals from commercial analog television stations have been used as common-view transmitters, and the zero-crossings of the mains voltage can also be used to compare clocks with an uncertainty on the order of a fraction of a millisecond.

59.10.3 The “Melting-Pot” Version of Common View

The previous discussion of common view focused on a number of cooperating receivers, where each one measured the time difference between a physical signal and the local clock. However, there can be some situations where there is no single transmitter that can be observed by the receivers at the same epoch. For example, if the common-view method is implemented by means of signals from a GPS navigation satellite, then receivers on the surface of the Earth that are sufficiently far apart cannot receive signals from any one satellite at the same time.

However, determining the position of a receiver by means of signals from the GPS satellites depends on the fact that the clock in each satellite has a known offset in time and in frequency from a system-average time that is computed on the ground and transmitted up to the satellites. Each satellite broadcasts an estimate of the offset between its internal clock and this GPS system time scale. By means of this information, two stations that observe two different satellites can nevertheless compute the time difference between the local clock and GPS system time, rather than computing a time difference between the physical signal transmitted by the satellite and the time of the local clock. The common-view time difference in this case is not with respect to a physical transmitter but rather with respect to the computed paper time scale, GPS system time.

In general, a receiver may be able to compute the time difference between its clock and GPS system time by means of the signals from several satellites, which explains the origin of the term “melting-pot” method. All of these measurements should yield the same time difference in principle, but this is not the case in practice for a number of reasons.

In the first place, the path delays between the receivers and the various satellites are not even approximately equal so that any error in computing the path delays is not attenuated as it is in the common-view method described in the preceding text. In addition, the method depends on the accuracy of the offset between each of the satellite clocks and the system time. As a practical matter, the full advantage of the melting-pot method is realized only when the orbits of the satellites and the characteristics of the onboard clocks have been determined using postprocessing—the values broadcast by the satellites in real time are usually not sufficiently accurate to be useful for this method.

On the other hand, the melting-pot method can usually use the observations from several satellites at the same time so that the random phase noise of the measurement process can be attenuated by averaging the data from the multiple satellites. Therefore, a comparison between the simple two-way method and the melting-pot version depends on a comparison between the noise of the measurement process, which would favor a melting-pot measurement using multiple satellites, and the uncertainties and residual errors in the orbital parameters of the satellites and the offset between the clock in each satellite and GPS system time. The accuracy of the melting-pot method improves as more accurate solutions for the orbits and satellite clocks become available [16].

59.10.4 Two-Way Methods

There are a number of different implementations of the two-way method, but all of them estimate the one-way delay between a transmitter and a receiver as one-half of the round-trip delay, which is measured as part of the message exchange. The accuracy of the two-way method depends on the symmetry of the delays between the two end points. The accuracy does not depend on the magnitude of the delay itself; although the magnitude of the delay can be calculated from the data in the message exchange, the accuracy of the time difference does not require this computation.

There are generally two aspects of the delay asymmetry that must be considered. The first is a static asymmetry—a difference in the delays between the end points in the opposite directions. In general, this type of asymmetry cannot be detected from the data exchange, and it places a limit on the accuracy that can be realized with any two-way implementation. The second type of asymmetry is a fluctuation in the symmetry that has a mean of zero. In other words, the channel delay is symmetric on the average, but this does not guarantee the symmetry of any single exchange of data. The impact of this type of fluctuating asymmetry can be estimated with enough data. As I will show in more detail later, a smaller measured round-trip delay is generally associated with a smaller time offset due to any possible asymmetry.

The transmitter and receiver at the end points of the path are often sources of asymmetry. These hardware delays are often sensitive to the ambient temperature. However, the admittances to temperature fluctuations may be different at the two end stations, and the temperatures at the two end points may be quite different. Finally, there are often components of the measurement process that are outside of the two-way measurement loop, and any delays originating in these components must be measured on every message exchange or measured once and stabilized.

59.10.5 The Two-Color Method

Suppose that there is a portion of the path that has an index of refraction that is significantly different from the vacuum value of one. This difference of the index of refraction relative to its vacuum value is the *refractivity* of the path. Suppose also that the refractivity is dispersive. That is, it depends on the frequency that is used to transmit the message. If the length of the path is measured using the transit time of an electromagnetic signal, the refractivity will increase the transit time so that the effect of the refractivity will be to make the path length appear too long. If the length of the true geometric path is D, then the measured length will be L, where L is given by

$$L = nD = D + (n-1)D. \tag{59.29}$$

I will now consider the special case where the refractivity can be expressed as a product of two functions: $F(p)G(f)$. That is,

$$n-1 = F(p)G(f). \tag{59.30}$$

The first function, F, includes parameters that characterize the transmission medium, including any dependence of these parameters on the environment such as the ambient temperature, relative humidity, etc. The second function, G, describes the dispersive characteristics of the path. Both of these functions can be arbitrarily complex and nonlinear—the only requirement is that the separation be complete. The function F cannot depend on the frequency that is used to transmit the signal, and the function G cannot depend on the parameters that describe the characteristics of the path.

If I measure the apparent length of the path using two frequencies, f_1 and f_2, I will obtain two different values for the apparent length because the index is dispersive. These two measured values are L_1 and L_2, respectively. Since the geometrical path length is the same for the measurements at the two frequencies, the difference between the two measurements can be used to solve for the value of the function $F(p)$:

$$\begin{aligned} L_1 - L_2 &= F(p)\left(G(f_1) - G(f_2)\right)D \\ F(p) &= \frac{L_1 - L_2}{D}\frac{1}{\left(G(f_1) - G(f_2)\right)} \\ n_1 - 1 = F(p)G(f_1) &= \frac{L_1 - L_2}{D}\frac{G(f_1)}{G(f_1) - G(f_2)}. \end{aligned} \tag{59.31}$$

If I substitute the last relationship of Equation 59.31 into Equation 59.29 evaluated for frequency f_1, I obtain

$$L_1 = D + (n_1 - 1)D = D + (L_1 - L_2)\frac{G(f_1)}{G(f_1) - G(f_2)}, \tag{59.32}$$

which allows me to find the geometrical path length, D, in terms of L_1, the length measured using frequency f_1, and the difference in the lengths measured at the two frequencies f_1 and f_2 multiplied by a known function of the two frequencies. If I call this function of the two frequencies H, then

$$\begin{aligned} H(f_1, f_2) &= \frac{G(f_1)}{G(f_1) - G(f_2)} \\ D &= (1-H)L_1 + HL_2. \end{aligned} \tag{59.33}$$

The details of the function *G* are not important, provided only that it is known and that the medium is dispersive. (The denominator of the fraction on the right-hand side of Eq. 59.31 or 59.32 is zero for a nondispersive medium. The difference in the apparent lengths in Eq. 59.31 will also be zero in this case.) Note that the second term on the right-hand side of Equation 59.32, which is the correction to the geometrical length *D* due to the dispersive medium, does not depend on *L*, the extent of that medium, but only on the apparent difference in this length for the measurements at the two frequencies. Thus this relationship is equally valid if only a portion of a geometric path is dispersive, and the correction term specifies the apparent change in the length of only that portion of the path. Note also that I do not have to know anything about the function *F*(*p*)—only that the separation into two terms expressed by Equation 59.30 represents the dispersion.

The measurements of both L_1 and L_2 will have some uncertainty in general so that the two-color determination of the geometrical length, *D*, will have an uncertainty that is greater than it would have been if the medium were nondispersive so that a measurement at one frequency would have been adequate. The magnitude of the degradation depends on the details of the function *H*, and I will discuss this point again when I describe measurements using navigation satellites such as those of the GPS.

59.11 EXAMPLES OF THE MEASUREMENT STRATEGIES

In the following sections I will describe systems that use the various measurement strategies that I have outlined earlier. I will begin by describing the characteristics of the satellites of the GPS, since the data transmitted by these satellites are widely used for timing applications in the one-way, common-view, and melting-pot modes. The Russian GLONASS system, the European Galileo system, and the Chinese BeiDou system are different in detail, but the following discussion describes the general features of all of them. In general, the differences between the systems are hidden from the general user and are a concern only of the receiver designer.

59.11.1 The Navigation Satellites of the GPS

The GPS system uses at least 24 satellites in nearly circular orbits whose radius is about 26,600 km. (The number of satellites in the constellation that are active at any time is generally >24.) The orbital period of these satellites is very close to 12 h, and the entire constellation returns to the same point in the sky (relative to an observer on the Earth) every sidereal day (very nearly 23 h 56 m).

The satellite transmissions are derived from a single oscillator operating at a nominal frequency of 10.23 MHz as measured by an observer on the Earth. In traveling from the satellite to an Earth-based observer, the signal frequency from every satellite is modified by two effects that are common to all of them—a redshift due to the second-order

Doppler effect and a blueshift due to the difference in gravitational potential between the satellite and the observer. These two effects produce a net fractional blueshift of about 4.4×10^{-10} (38 μs/day), and the proper frequencies of the oscillators on all of the satellites are adjusted downward to compensate for this effect, which is a property of the orbit and is therefore common to all of them. In addition to these common offsets, there are two other effects—the first-order Doppler shift and a frequency offset due to the eccentricity of the orbit, which vary with time and from satellite to satellite. The receiver computes and applies the corrections for these effects.

The primary oscillator is multiplied by 154 to generate the *L*1 carrier at 1575.42 MHz and by 120 to generate the *L*2 carrier at 1227.6 MHz. (The newer GPS satellites will transmit signals on additional frequencies, and the Galileo, GLONASS, and BeiDou systems transmit signals at slightly different frequencies.) These two carriers are modulated by three signals: the precision "P" code, a pseudorandom code with a chipping rate of 10.23 MHz and a repetition period of 1 week; the "clear access" or coarse acquisition "C/A" code with a chipping rate of 1.023 MHz and a repetition rate of 1 ms; and a navigation message transmitted at 50 bits/s. The codes are derived from the same 10.23 MHz primary oscillator. Under normal operating conditions, the C/A is present only on the *L*1 carrier. Many timing receivers process only the C/A code. Although the P code is normally encrypted with an encryption key that is not available to unclassified users, many receivers can operate in a "semi-codeless" mode where the P code data can be decoded with some increase in the noise of the process.

Each GPS satellite transmits at the same nominal frequencies but uses a unique pair of C/A and P codes. The codes are constructed to have very small cross-correlation at any lag and a very small autocorrelation at any nonzero lag (code division multiple access (CDMA)). The receiver identifies the source of the signal and the time of its transmission by constructing local copies of the codes and by looking for peaks in the cross-correlation between the local codes and the received versions. Since there are only 1023 C/A code chips, it is feasible to find the peak in the cross-correlation between the local and received copies using an exhaustive brute-force method. When this procedure succeeds, it locks the local clock to the time broadcast by the satellite modulo 1 ms, the repetition rate of the whole C/A code. The procedure locks the local clock to the satellite time with a time offset due to the transmission delay (about 65 ms) and allows the receiver to begin searching for the 50 bits/s navigation message.

The navigation message contains an estimate of the time and frequency offsets of each satellite clock with respect to the composite GPS time, which is computed using a weighted average of the clocks in the satellites and in the tracking stations. This composite clock is in turn steered to UTC(USNO), which is in turn steered to UTC as computed by the BIPM. The time difference between GPS system time and UTC(USNO) is guaranteed to be less than 100 ns (modulo 1 s), and the estimate of this offset, which is transmitted as part of the navigation message, is guaranteed to be accurate to 25 ns (also modulo 1 s). In practice, the performance of the system has almost always substantially exceeded its design requirements.

The UTC time scale includes leap seconds, which are added as needed to keep UTC with ±0.9 s of UT1, a time scale based on the position of the Earth in space. The GPS time scale does not incorporate additional leap seconds beyond the 19 that were defined at its inception; the time differs from UTC by an integral number of additional leap seconds as a result. This integer-second difference, GPS time—UTC, is currently (December, 2015) 17 s and will increase as additional leap seconds are added to UTC. The number of leap seconds between GPS time and UTC is transmitted as part of the navigation message but is not used in the definition of GPS time itself. Advance notice of a future leap second in UTC is also transmitted in the navigation message.

Most modern receivers can observe several satellites simultaneously and can compute the time differences between the local clock and GPS system time using all of them at same time. In each case, the time of the local clock that maximizes the cross-correlation with the signal from each satellite is the *pseudorange*—the raw time difference between the local and satellite clocks. (It is related to the geometrical time of flight with an additional time offset since the clock in the receiver is generally not synchronized to GPS system time.)

Using the contents of the navigation message, the receiver corrects the pseudorange for the travel time from the satellite to the receiver, for the offset of the satellite clock from satellite system time, etc. (If the receiver can process both the $L1$ and $L2$ frequencies, then the receiver can also estimate the additional delay through the ionosphere due to its refractivity by applying the two-color method described before to the difference in the pseudoranges observed using the $L1$ and $L2$ frequencies. If the receiver can process only the $L1$ signal, then it usually corrects for the ionospheric delay using a parameter transmitted in the navigation message.) The result is an estimate of the time difference between the local clock and GPS system time.

In principle, the time difference between the local clock and GPS system time should not depend on the specific satellite whose data are used for the computation. In practice, the time differences computed using the data from different satellites will differ because of the noise in the measurement processes and because of errors in the broadcast ephemerides and the parameters of the satellite clocks. The group of time differences with respect to GPS system time, computed from the different satellites, forms a "redundant array of independent measurements" (RAIM), and some analysis methods compare these time differences in an attempt to detect a bad satellite. These "T-RAIM" algorithms succeed when the time difference computed using the data from one satellite differs from the mean of the differences computed from the other satellites by a statistically significant amount. The T-RAIM algorithm can be used in the one-way, common-view, or melting-pot algorithms that I discuss in the next sections. The same idea is used in the Network Time Protocol (NTP) to identify a bad server. I will discuss this point in greater detail as follows.

59.11.2 The One-Way Method of Time Transfer: Modeling the Delay

This method is most often used with time transfer by means of signals from navigation satellites because they are the only systems that transmit enough information to support an accurate delay estimate. The estimate of the transit time of a message from a navigation satellite to a receiver on the ground can be divided into a number of components that are increasingly difficult to estimate.

The largest single estimate to the propagation delay is the delay resulting from the geometric path length. The magnitude of this delay depends somewhat on the position of the satellite in the sky but is typically approximately 65 ms. The path length is computed from the position of the satellite, which is estimated from the orbital parameters transmitted as part of the navigation message, and the position of the ground receiver. In pure timing applications, the position of the ground receiver is assumed to be known from other data, and I will assume that this is the case in the current discussion. (If the position of the receiver is not known *a priori*, it can be estimated by computing the distances from the receiver to multiple satellites and solving for the four unknowns: the three Cartesian coordinates of the position of the receiver and the time offset of its clock with respect to satellite system time.) In a real-time application the accuracy of the estimate of the geometric path delay is limited by any uncertainty in the position of the receiver (the vertical coordinate usually has the largest uncertainty) and by errors in the broadcast ephemeris parameters. These combined uncertainties are generally on the order of a few meters, which is equivalent to an uncertainty in the time delay of about 10 ns or less. Thus the uncertainty in the correction is much smaller than the magnitude of the correction itself.

The additional delay due to the passage of the signal through the ionosphere adds approximately 65 ns to the geometric delay. A receiver that can process both of the frequencies transmitted by a satellite can estimate the effect of the ionosphere using the two-color method that I have described previously. Simpler, single-frequency receivers can use an estimate of the effect of the ionosphere that is broadcast by the satellite as part of the navigation message. This is a globally averaged prediction and is therefore less likely to be accurate at any specific location.

The additional delay due to the passage of the signal through the lower atmosphere (the troposphere) is much smaller than the additional delay through the ionosphere, but there is no easy way of estimating it because the refractivity does not depend on the carrier frequency so that the two-color method cannot be used. Some more sophisticated analyses estimate this delay by means of local measurements of atmospheric pressure, temperature, and water vapor content, but these data are not always available. Even when they are available at a site, these parameters often have significant azimuthal variation, which is generally not easily estimated. (Boulder, Colorado, is potentially particularly bad in this respect, since the mountains to the west and the plains to the east would be expected to have quite different temperature profiles.) There are also models of the refractivity of the troposphere, which estimate this parameter as a function of the day of the year and, possibly, the coordinates of the receiving station.

The magnitude of this delay is typically on the order of 6 ns at the zenith, and it increases for satellites at lower elevation by a factor that is roughly proportional to the increase in the slant path through the troposphere relative to the zenith path length. The increase in the slant path delay relative to the zenith delay is usually estimated as proportional to the reciprocal of the sine of the elevation angle.

If an analysis assumes the slant path model of the variation of the delay, it is possible to solve for the zenith delay by observing the apparent variation in the time difference estimates obtained from satellites at very different elevation angles. This estimate does not work as well as we might like because the slant path model is only an approximation, because the tropospheric refractivity often has a significant azimuthal variation, and because the variation from one satellite to another is also affected by measurement noise and by errors in the broadcast ephemerides or any error in the coordinates of the receiver.

Many time-difference measurements ignore the effect of the troposphere altogether. This introduces a systematic error of order 10 ns in the time-difference estimates; as I mentioned in the preceding text, the magnitude of this error depends on the elevation of the satellites that are being observed.

The final contributions to the model of the delay are effects that are local to the receiver: the delay through the hardware and the motion of the station due to the Earth tides and other geophysical effects. The delay through the receiver hardware is normally assumed to be a constant that varies only very slowly over periods of years. The delay is often dominated by the delay through the antenna and the cable from the antenna to the receiver, and a value on the order of 100 ns is typical. (Delays through coaxial cables are of order 5 ns/m.)

It is possible to calibrate the delay through a receiver using a special signal generator that mimics the signals from the real satellite constellation. This type of equipment is not widely available, and most timing laboratories perform a differential calibration in which the delay through the receiver under test is compared to the delay of a "standard" receiver. This method is obviously not adequate for a one-way measurement but is widely used because most timing laboratories use the satellites in common view, which I discuss in the following section.

The motion of the station and other geophysical effects contribute 1–2 ns to the overall delay. The magnitude of these effects can be calculated and included in more complicated postprocessed analyses but are generally ignored for real-time applications.

59.11.3 The Common-View Method

I have already described most of the important features of the common-view method. It is most often used with signals from the navigation satellites, but it is more general than this and can be used with other sources as well. Signals from LORAN transmitters and even from television stations have been used in this way. There have even been some experiments to use the zero-crossings of the power line in common view within a building or over a small area.

The method has two principal limitations:

1. It is difficult in practice to configure the measurements so that the receivers are all equidistant from the source. Therefore, some correction is almost always necessary to model the differential delay. The differential delay is much smaller than the delay itself in most configurations so that the required accuracy of the model of the delay is correspondingly easier to satisfy. However, ignoring the difference in the path delays is often not sufficiently accurate.
2. The common-view method (and its melting-pot variant) cannot provide any help in mitigating the effects of delays that are local effects at a site. The differential effects of the ionosphere can be significant and are usually estimated using the two-color method. The differential effects of the troposphere cannot be estimated in this way and are often ignored. Ignoring the differential effects of the troposphere is often justified because the total contribution is relatively small and the differential contribution is correspondingly smaller.

Multipath is a more serious local effect that is often too large to ignore. The effect is caused by copies of the signal that reach the antenna after they have been reflected from some nearby object. These signals always travel a longer distance than the primary one, and they arrive later than the primary signal as a result. A simple omnidirectional antenna typically responds to these signals, and the receiver computes a correlation that is a complicated sum of the direct and reflected signals.

The multipath effect is a complicated function of the position of the satellite with respect to the antenna and the local reflectors, and it is therefore periodic with the orbital period of the satellite. From this perspective, the orbital periods of the GPS satellites are all very close to one sidereal day (23 h 56 m) so that the multipath reflections have this periodicity. They can usually be estimated by comparing the time differences measured from any satellite at the same sidereal time on consecutive days.

The BIPM has exploited this sidereal-day periodicity in defining the common-view tracking schedules that are used by timing laboratories and National Metrology Institutes to compare time scales and to facilitate the computation of TAI and UTC. The observation time for each satellite is advanced by 4 min every day in the BIPM schedule so that every satellite returns to the same point in the sky on every track each day relative to the antenna and to any multipath reflectors. Thus, the multipath environment is a constant for each track, although it generally varies from track to track. This has the advantage of converting the varying effects of multipath to systematic offsets that are approximately constant for each track. The assumption of a sidereal-day periodicity is not exact so that the offset due to the multipath contribution changes slowly with time. These long-period effects can be hard to distinguish from the contributions due to the random walk of frequency and frequency aging that I discussed previously.

Locating the antenna far away from reflecting surfaces can help minimize the impact of multipath reflections; adding choke rings and a ground plane to an antenna, which attenuate signals arriving from the side or from below, can also help.

Another strategy to mitigate the impact of multipath is to exploit the sidereal-day periodicity and compute the average frequency of the local clock with respect to the GPS system time as an average over a sidereal day. The multipath contribution cancels in the sidereal-day time difference so that the frequency estimated in this way is almost insensitive to multipath effects.

59.11.4 Two-Way Time Protocols

In the following sections, I will describe three two-way time protocols that are commonly used to compare clocks at remote locations. The list is intended to be descriptive rather than exhaustive. For example, I do not discuss time transmission using optical fibers because this method is generally too expensive to be used for long distances and because the underlying physics is basically the same as the other methods that I do describe. I also do not discuss the Precise Time Protocol (PTP, often called IEEE 1588) in any detail for much the same reasons. Its capabilities are very similar to a hardware-assisted version of NTP, and it is generally not well suited to long-distance time comparisons because it assumes that the delay is nearly constant so that it does not have to be measured on every message exchange.

59.11.4.1 The NTP The NTP is widely used to transmit time and compare clocks that are linked together by a channel that is based on a packet-switched network such as the Internet. The NTP message format is based on the User Datagram Protocol [17] (UDP). The UDP message exchange is not sensitive to the details of the physical hardware that is used to transmit the packets. However, as with all two-way protocols, the accuracy of the NTP message exchange depends on the symmetry of the inbound and outbound delays, and this symmetry is often limited by the characteristics of the physical layer used to transmit the messages. In the following discussion I will focus on the time-difference accuracy of the message exchange; I will defer the question of how often a system should initiate such an exchange (the "polling interval") to a later section, and I will discuss only briefly the question of how a client system should discipline its local clock based on the exchange of messages with a server.

The protocol is initiated when station "A" sends a request for time information to station "B." The two stations might have a client–server relationship, in which the client intends to adjust its clock based on the results of the exchange, or it could be a peer-to-peer exchange in which two systems exchange timing information with the goal of setting the times of both systems to agree with each other.

The message is sent at time T_{1a} as measured by the clock on system A. The transmission delay from station A to station B is δ_{ab} so that when the message arrives at station

B, the time at station A is $T_{1a}+\delta_{ab}$. The time of arrival at station B, measured by the local clock at that station, is T_{2b}, and the time difference between stations A and B is

$$(\Delta T)_{ab}=(T_{1a}+\delta_{ab})-T_{2b}. \tag{59.34}$$

The B system responds by sending a message back to A. The message leaves the B system at time T_{3b} and arrives back at the A system at time $T_{3b}+\delta_{ba}$, and the time at the A system at that instant is T_{4a}.

The total round-trip transit time is measured at station A as the time that has elapsed during the message exchange as measured by the clock on station A, less the time between when station B received the request and when it replied, as measured by the clock on station B:

$$\Theta=\delta_{ab}+\delta_{ba}=(T_{4a}-T_{1a})-(T_{3b}-T_{2b}). \tag{59.35}$$

We now assert that the path delay is symmetric so that the inbound and outbound delays are equal. Then the path delay from A to B, δ_{ab}, is simply one-half of the expression on the right-hand side of Equation 59.35. If we substitute one-half of the right-hand side of Equation 59.35 into Equation 59.34, the time difference between stations A and B is

$$(\Delta T)^{s}_{ab}=\frac{T_{1a}+T_{4a}}{2}-\frac{T_{2b}+T_{3b}}{2}. \tag{59.36}$$

The superscript s indicates that the time difference is computed using a symmetric path delay. If the path delay is not symmetric, then the inbound and outbound delays are not equal. We can parameterize this asymmetry as

$$\delta_{ab}=(0.5+\varepsilon)\Theta. \tag{59.37}$$

The asymmetry parameter ε can take values from +0.5 to −0.5. The positive limit indicates that the path delay from A to B dominates the round-trip delay and the delay in the other direction is negligibly small, while the negative limit specifies the inverse: the delay from A to B is negligible compared to the reverse delay from B to A.

If we substitute Equation 59.37 into Equation 59.34, then the first term on the right-hand side of Equation 59.34 reproduces the time-difference expression of Equation 59.36, and the second term adds a correction to the time difference:

$$(\Delta T)^{a}_{ab}=(\Delta T)^{s}_{ab}+\varepsilon\,\Theta. \tag{59.38}$$

Since we model the measurement based on the assumption of a symmetric delay (Eq. 59.36), the time difference that we estimate is in error. The magnitude of the error is given by $\varepsilon\Theta$, the second term in Equation 59.38. This term is proportional both to the magnitude of the asymmetry and to the round-trip delay. Thus a smaller round-trip delay guarantees a smaller error due to any asymmetry. The lesson is that NTP servers should be widely located so that the round-trip delay to any user is minimized.

The round-trip delay is often of order 100 ms (0.1 s), and typical asymmetries are on the order of a few percent of the delay. Therefore, we might expect that a typical NTP message exchange would have an error on the order of 5 or 10 ms due to the asymmetry of the path delay, and errors of this order should be considered as routine for a server and a client on a wide-area network.

In addition to possible asymmetries in the network delay, there may also be additional asymmetries in the client system. For example, if the process that manages the NTP message exchange runs in a standard user environment, then it must compete for processor cycles with all of the other processes that may be active on the system. In addition, it must issue a request to the system to retrieve the system time each time a message is sent or received in order to have the values to compute the time differences described in the preceding text. All of these effects add to the network delay measurement; depending on the details of the system and the processes that are active, it may also add to the asymmetry.

In order to minimize these effects, the NTP process can be moved into the system space where it runs at much higher priority as a system service. The ultimate version of this idea would be to move the NTP process into the network driver that receives and transmits the network packets, and some version of NTP and its cousin PTP (also called IEEE 1588) operate in this mode.

Although moving the NTP process into the system space or into the network driver itself will make the NTP process appear more stable and more accurate, the overall timing accuracy of an application that uses the system time may be degraded. This application normally runs as a standard user process, and it therefore experiences the same jitter as a user-level NTP process would experience when it issues a request to the system for the current time. This jitter is inside of the measurement loop when the NTP process runs as a user process, and the time-difference calculation therefore takes it into account (at least to some extent even if it also contributes to the asymmetry). However, this delay is completely outside of the measurement loop if the NTP process is pushed down to the system or driver levels so that the application process experiences the full impact of the delay jitter in requesting system services. Thus, while the NTP statistics improve, the accuracy realized by a user process may be degraded, and this problem is not reflected in any of the NTP statistics.

59.11.4.2 The ACTS Protocol The NIST (and a number of other timing laboratories) operate a time service that transmits the time in a digital format by means of standard dial-up telephone lines. The NIST system is called the Automated Computer Time Service (ACTS) [18], and it corrects for the transmission delay using a variant of the two-way protocol I have described.

The ACTS servers transmit a text string each second with the time derived from the NIST clock ensemble and an on-time marker (OTM) character. This character is initially transmitted using a default advance. If the user echoes the OTM back to the server, the server measures the round-trip delay, estimates the one-way delay as one-half of this value, and adjusts the advance of the next OTM transmission so that it will arrive at the user's system on time. The server changes the OTM character from "*" to "#" to indicate that it has entered this delay-calibrated mode. In every case, the server includes the estimate of the one-way delay in the message so that the client can determine the advance that was used. In the context of the previous discussion of the NTP protocol, the ACTS protocol assumes that $(T_{3b} - T_{2b})$ is essentially zero. That is, the client echoes the OTM character back to the server with only negligible delay.

This process continues on every transmission. The OTM character is advanced based on the one-way delays estimated from the average of the round-trip measurements of the previous seconds. The details of the averaging process are determined dynamically by the server based on the measured variation of the round-trip delay from second to second.

The original ACTS system was designed to minimize the complexity of the code in the receiving system. The receiver needs only to echo the OTM back to the server, and the next OTM will be transmitted so that it arrives on time. The receiver did not have to perform any calculations at all. The assumption of this design was that the delay variations were not accompanied by variations in the symmetry, and this assumption was largely confirmed by the original design, which could transmit time messages with an accuracy of order 0.5 ms RMS.

Placing the delay calculation in the server simplifies the design of the client system, but it has the unfortunate side effect that the server cannot detect a change in the symmetry of the delay whether or not this change in symmetry is accompanied by a change in the total round-trip delay. However, the client system is in a better position to determine what is really going on.

Since the true time difference between the client and the server changes by less than 1 ms from second to second, any significant change in the time difference measured from one second to the next one indicates that the advance algorithm in the server has been fooled by a change in round-trip delay that was accompanied by a change in symmetry. For example, it is possible that the change in the measured round-trip delay was really largely confined to either the inbound or outbound paths. The client can detect this possibility by noting the change in the measured time difference between the ACTS time and its system clock and the change in the measurement of the round-trip delay that the server has inserted into the transmission. As a simple example, if the change in the delay is confined to the outbound path between the server and the client, the server will see this as a change in the total round-trip delay and will advance the next OTM by one-half of this value. This is exactly one-half of the correct advance change so that the next OTM will not arrive on time by one-half of the change in the advance. The client will detect that the advance parameter has changed and that the measured time difference has changed by one-half of that amount. This more sophisticated algorithm in the client system can almost completely compensate for the degraded stability of the dial-up telephone system, and the more sophisticated algorithm can transmit time over standard dial-up telephone lines with an uncertainty of order 0.5–0.8 ms RMS. This is about a factor of 10 better than the Internet time servers because the delay through the telephone system is more stable and more symmetric than the delay through a wide-area packet network.

59.11.4.3 Two-Way Satellite Time Transfer This method is used to compare the time scales of National Metrology Institutes and Timing Laboratories and to transmit time and frequency information to the International Bureau of Weights and Measures (the BIPM, in French) for the purpose of computing TAI and UTC.

The method uses the same assumption as in the previous discussions: the one-way delay can be estimated as one-half of the measured round-trip value. The configuration of the message exchange is similar to the NTP exchange discussed previously.

Each station encodes the 1 Hz tick of its local time scale using a pseudorandom code and a subcarrier whose frequency is about 70 MHz. The subcarrier is transmitted up to a communications satellite, which is in a geostationary orbit. (The satellite is located above the equator. The radius of its orbit is ~40,000 km, and its orbital period is 24 h. It therefore appears to be stationary with respect to an observer on the Earth.) The satellite retransmits the modulated signal down to the receiver, where the 1 Hz tick is recovered by a cross-correlation of the received pseudorandom code with a copy of the code generated in the receiver. The time difference between the recovered 1 Hz tick and the local clock is then stored. The message exchange is full duplex, and the time differences at each station are combined to estimate the time difference of the clocks at the two sites.

The uplink and downlink typically use different frequencies in the Ku band. The uplink frequency is nominally 14 GHz, and the downlink is nominally 11 GHz. Both frequencies are used on a portion of the path in each direction, but the paths are not the same so that the delays in the two directions may not be exactly equal. The dispersion of the refractivity of the ionosphere and the troposphere is small at these frequencies so that this asymmetry is generally not an important limitation. Balancing the transmit and receive delays in the ground station hardware is a more difficult problem, especially because these delays are often sensitive to fluctuations in the ambient temperature.

The transit time from one station up to the satellite and down to the other station is about 0.25 s so that the rotation of the Earth during this time must be taken into account. This is called the Sagnac [19] effect. The magnitude of this effect is $2\omega A/c^2$, where c is the speed of light, ω is the angular velocity of the Earth, and A is the area defined by the triangle formed by the satellite and the two stations projected onto the equatorial plane. The effect is positive for a message traveling eastward and negative for messages in the opposite direction.

The design of this method treats both stations as equal partners in the exchange. The transmissions at each end are generated by the local time scale and are not a response to a message received from the other end as with NTP. In principle, this is an important difference between this method and the NTP and ACTS methods described earlier, but the underlying assumptions of all of the methods are the same, and only the details of the analysis are somewhat different.

Based on the notation of the discussion of the NTP message exchange, a well-designed NTP system will have a very small time delay, T_{3b}–T_{2b}, between when a system receives a query and when it responds; the ACTS protocol assumes that this difference is negligible; the two-way system makes no assumptions about this difference except that it is accurately measured on each message exchange.

The time differences measured with the two-way satellite method are much more accurate than the measurements made with either of the systems discussed previously because the delays are much more stable and the assumption of the very small delay asymmetry is more accurate. The RMS uncertainty of the measurements is of order 0.1 ns. The spectrum of the noise is approximately white phase noise for short averaging times, but there are less favorable variations at longer periods; some links have a quasiperiodic approximately diurnal variation in the time-difference data. The source of this variation is not understood at present.

59.12 THE POLLING INTERVAL: HOW OFTEN SHOULD I CALIBRATE A CLOCK?

There are three different considerations that should be used to determine the interval between time-difference measurements. The first consideration is based on a statistical estimate of the noise processes, the second is derived from concerns about nonstatistical errors, and the third is based on a cost–benefit analysis. I will begin by considering choosing a polling interval based on a statistical analysis.

From the statistical perspective, the goal of the time-difference measurements is to improve the accuracy or the stability of the local clock once its deterministic properties have been determined and used to adjust the time, frequency, and aging (if appropriate) of the clock under test. The deterministic parameters can be applied directly to the physical device, or they can be used to adjust the readings of the clock administratively. (In general, timing laboratories usually do not adjust the physical parameters of a clock but rather apply the deterministic offsets administratively.)

Once the deterministic parameters have been included in the readings of the device under test, I assume in a statistical analysis that its time dispersion can be determined solely from its statistical characterization. At least at the beginning, I will assume that both the deterministic and stochastic parameters are constants that do not depend on time or on perturbations such as fluctuations in the ambient temperature.

In this model, the design of the calibration procedure is driven by the requirement that the accuracy or stability of the remote clock as seen through the channel should be better than the corresponding parameters of the local clock for the same query interval. (The channel includes the physical medium used to transmit the time signal and any measurement hardware at the end stations.) As a practical matter, it often turns out that the statistical characteristics of the channel are much poorer than the statistics of the remote clock itself. In this situation, which is quite common, improving the accuracy or the stability of the remote clock will have almost no effect on the performance of the synchronization process, which will be dominated by the statistics of the channel. The following discussion depends only on being able to characterize the combination of the remote clock and the channel by means of the two-sample Allan deviation. The analysis is not sensitive to whether the source of the fluctuations is in the clock or the channel connecting it to the device under test.

Suppose that the statistical characteristics of the remote clock seen through the channel can be described as white phase noise for all averaging times. This is the best that we can hope for—the measurement process using this channel is degraded by a noise process that has a mean of zero and a standard deviation that does not depend on the time the measurement was performed or on any external parameter such as the ambient temperature. The magnitude of the two-sample Allan deviation (the square root of the variance) that describes the statistics of the remote clock seen through this channel varies as the reciprocal of the averaging time. The time dispersion in this configuration is independent of the averaging time. (see Equations 59.10 and 59.11.) This is not a surprising result. If the measurement noise is characterized as white phase noise, then the fluctuations of every measurement are derived from the same distribution with a mean of zero and a fixed standard deviation, and there is no relationship between one measurement and any other one so that the time between measurements is irrelevant to the statistics of the time differences.

Now consider that the stability of the local clock is characterized as pure white frequency noise for moderate averaging times. Again, this is not a surprising result and is about the best we could ever hope to see; once the white phase noise of the measurement process has been accounted for, the next effect is the noise of the frequency control loop, which we also take to be a process with a mean of zero and a known standard deviation. (There are almost always longer-period effects that modify the assumption of pure white frequency noise, but I will assume for now that the averaging times that will be used will not be large enough to make this consideration important.) The two-sample Allan deviation for white frequency noise varies as the reciprocal of the square root of the averaging time so that time dispersion due to white frequency noise increases as the square root of the averaging time. (I will use the conservative estimate that I discussed in the preceding text for Eq. 59.11.)

We can now combine these two results to define two measurement strategies. The noise of the time-difference measurement process is characterized by a standard deviation, M, which does not depend on averaging time. The time dispersion of the local clock due to its white frequency noise is characterized as a function of averaging time by $C\tau^{1/2}$. From the statistical perspective, the goal of the first synchronization procedure is to set the averaging time so that the remote clock seen through the channel is more stable than the local clock. In other words, the averaging time should be chosen so that the free-running time dispersion of the local clock due to its frequency noise is greater than the time dispersion of a time-difference measurement with respect to the remote clock seen through the channel:

$$\begin{aligned} C\tau^{1/2} &\geq M \\ \tau &\geq \left(\frac{M}{C}\right)^2. \end{aligned} \tag{59.39}$$

The result may be surprising but is easily explained. As the remote clock seen through the network becomes less stable (increasing M), the crossover between its stability and the stability of the local device, which increases as the square root of the averaging time, moves to longer and longer averaging times. Thus we would expect that the optimum polling interval for a time transfer that used the wide-area Internet to communicate between the local and remote devices would be longer than the optimum polling interval for the same devices that exchanged messages on a local network connection because the stability of the transmission delay in a wide-area network would be poorer.

Since the channel connection back to the remote clock is characterized by white phase noise, a second strategy would be to make measurements of the time difference as rapidly as possible and average the data. For example, suppose we could make measurements every second for a time interval of T seconds. If we averaged these T measurements, the standard deviation of the mean would be reduced from M to $M/\sqrt{T}$. The time dispersion due to the white frequency noise of the local device would be the same as before so that the comparison of Equation 59.39 becomes

$$\begin{aligned} CT^{1/2} &\geq \frac{M}{T^{1/2}} \\ T &\geq \frac{M}{C}, \end{aligned} \tag{59.40}$$

where the value of T in Equation 59.40 specifies the point at which the noise of the local clock is greater than or equal to the standard deviation of the average so that the averaging algorithm can improve the stability of the local clock starting at an averaging time of T. In this simple model, once this time is reached, additional averaging only makes things better because the standard deviation of the remote clock seen through the channel improves without bound, while the stability of the local clock degrades without bound. In the limit of very large T, we are not using the data from the local clock at all.

This simple model will break down at some point for one of two reasons. The first possibility is that channel noise is not purely white phase noise starting at some averaging time—longer-period fluctuations become important, and they are not zero-mean random processes. These fluctuations can be incorporated by modeling the time dispersion of the remote clock, *M*, as constant at short times but increasing as some power of the averaging time starting at some averaging time, T_{m}. The second possibility is that the requirements of the application limit the averaging time—we cannot average forever because we need to use the time difference for some application. The assumptions that I used in this discussion are somewhat artificial in that they are often better than real-world devices and channels. Therefore, these calculations are more illustrative of the method than rigorous derivations with very general applicability.

59.13 ERROR DETECTION

Any measurement protocol that receives data from a remote device over a noisy channel should be prepared to consider the possibility that the received data are in error, either because the remote clock has failed or the channel characteristics have changed suddenly. A purely statistical analysis cannot be the whole story here, since there is generally no objective way of distinguishing between an error and a very low-probability event that conforms to the statistical description. One method that is commonly used is to regard a measurement that differs from the mean (or from the predicted value) by more than three standard deviations as an error.

The machinery that I developed in the previous section can also be used to detect possible errors. For example, consider the averaging strategy presented in the discussion for Equation 59.40. Instead of waiting until all of the measurements have been completed to evaluate the average time difference, we could construct a running mean with an update each time a new time difference was acquired. The estimate of the mean at the *k*th step, $\bar{X}_k$, after receiving the time difference x_k can be calculated iteratively based on the estimate of the mean at the previous step,

$$\bar{X}_k = \frac{(k-1)\bar{X}_{k-1} + x_k}{k}, \tag{59.41}$$

where the estimate of the mean is initialized to zero.

One possibility is to ignore the *k*th estimate as having a one-time error if it differs from the running mean by more than three times the running estimate of the standard deviation computed from the current average or from a previous one:

$$\left|\left(x_k - \bar{X}_{k-1}\right)\right| > 3\sigma_{k-1} \tag{59.42}$$

The assumption that the difference is a one-time error would be confirmed if the next reading was consistent with the running mean value.

The situation becomes more complicated if the next measurement does not conform to the running mean either, and it may not be possible to distinguish between a problem with the remote clock, the local clock, or the channel. It is sometimes possible to decide this question if a second independent calibration source or an independent channel is available. This solution must be considered in the cost–benefit analysis that I will discuss in the next section.

59.14 COST–BENEFIT ANALYSIS

In this section I will consider the situation where a time-difference measurement has a finite cost in terms of computer cycles, network bandwidth, or some other finite resource. The trade-off between the accuracy of a time-difference measurement and the cost needed to realize it becomes important in this situation.

For example, consider again the simple case where the time-difference measurements are characterized as white phase noise with a mean of zero. The standard deviation of the mean of N measurements decreases as $1/\sqrt{N}$, and this improvement can continue without bound in principle. On the other hand, the cost of the measurement process increases linearly with N, assuming that each measurement has the same cost. In this situation, the cost–benefit analysis is always unfavorable—the cost of the measurements always increases faster than the standard deviation improves, and the best cost–benefit strategy is to make the minimum number of measurements consistent with the standard deviation that is required to meet the needs of the application.

More generally, I assume that the total cost of a measurement procedure, C, is given by the cost of a single measurement, c; the interval between measurements, τ; and the total measurement time, T:

$$C = c\frac{T}{\tau}. \tag{59.43}$$

I take the benefit of the procedure, B, as the time dispersion of the device for an averaging time, τ, where the time dispersion is calculated from the two-sample Allan deviation for that averaging time:

$$B = \sigma_y(\tau)\tau. \tag{59.44}$$

The goal is then to minimize the product BC, possibly with some additional constraint that the time dispersion must be less than some value required by the application.

Apart from the constants, the product BC is a function only of the two-sample Allan deviation so that it will always improve with increasing averaging time as long as the slope of the Allan deviation is negative, and the best strategy will be the longest averaging time that satisfies the time dispersion estimate in Equation 59.44. The slope of the two-sample Allan deviation is negative in white phase noise and white frequency noise domains so that a pure cost–benefit analysis will always favor an averaging time that is at the onset of flicker processes where the slope of the Allan deviation approaches zero. The cost–benefit analysis product is a constant independent of averaging time in the flicker domain, but the time dispersion increases with averaging time (Eq. 59.44) so that the dispersion may not satisfy the requirements of the application in this domain. The cost–benefit analysis becomes unfavorable in the random walk of frequency domain where the slope of the two-sample Allan deviation is positive. The time dispersion of the local clock is increasing faster than the cost is decreasing in this region. This region might still be an acceptable choice if the accuracy requirement is very modest.

The method used for detecting errors also has a cost–benefit aspect. For example, if an Internet client queries *N* Internet servers on every measurement cycle in an attempt to detect an error, then the cost of the synchronization process has increased by a factor of *N*; the benefit will depend on how often this procedure detects a problem. Shortening the polling interval to detect a problem with the local clock more quickly is subject to the same considerations. That is, do problems happen often enough to justify the increased cost of the algorithm? In general, comparing the measured time difference with a prediction based on the statistics of the local clock (e.g., Eq. 59.42) and querying multiple servers only when that test fails are a better strategy because it exploits the statistics of the local clock as a method for detecting a possible error with the remote clock or with the channel.

A cost–benefit analysis is very important from the perspective of the operators of the network and the public time servers—increasing the polling interval and reducing the number of servers queried on each measurement cycle translates directly into the number of users that can be supported with available, scarce resources.

59.15 THE NATIONAL TIME SCALE

The official time in the United States (and in most other countries) is UTC. The length of the UTC second is defined by the frequency of the hyperfine transition in the ground state of cesium. The frequency of this transition is defined to be 9,192,631,770 Hz, and counting this number of cycles defines the length of the second. The other time units (minutes, hours, …) are multiples of this base unit.

The length of the day, computed as 86,400 Cs, is somewhat shorter than the length of the day in the UT1 time scale, which is a time scale based on the rotation of the Earth. The accumulated time difference is currently somewhat less than 1 s/year. In order to maintain a close connection between atomic time, defined by the cesium transition frequency, and the UT1 time scale, which characterizes the position of the Earth in space, additional seconds are added to UTC whenever the difference between UTC and UT1 approaches 0.9 s. The decision to add these "leap seconds" is made by the International Earth Rotation and Reference System Service, and all national timing laboratories incorporate the leap second into their time services.

Leap seconds are normally added as the last UTC second of the last day of June or December. In the vicinity of a leap second, the time stamps are 23:59:58, 23:59:59, 23:59:60, and then 00:00:00 of the next day. Digital time services and most clocks cannot represent the leap second time of 23:59:60 and stop the clock for 1 s at 23:59:59. The time services operated by the National Institute of Standards and Technology implement the leap second by transmitting a time value equivalent to 23:59:59 twice, and most other time services do the same thing. Assigning the same time stamp to two consecutive seconds is ambiguous and has obvious difficulties with respect to causality, and the question of continuing the leap second procedure is currently (as of 2015) being discussed.

The details of the leap second procedure are important for users who must synchronize a clock to the official time scale in the vicinity of a leap second. Unfortunately, some time services implement the leap second in different ways. One method adds the leap second by duplicating the time 00:00:00 of the next day. This eventually results in the same time as the NIST method, but it adds the leap second in the wrong day and has time errors on the order of 1 s in the immediate vicinity of a leap second.

A more troubling method implements the leap second as a frequency adjustment during the last few minutes of the day. The clock is slowed down for some period of time until the additional second has been added. This method has both a time error and a frequency error during the time the leap second is being inserted. The clock is never stopped in this implementation, but both the time and the frequency are not correct with respect to national time standards during this interval. In addition, there is no generally accepted method for implementing the frequency adjustment so that different implementations of this method will also have errors with respect to the national standards of time and frequency in this vicinity of a leap second.

In addition to the two proposals, (i) not to make any changes or (ii) to stop adding future leap seconds to UTC but to continue the number of leap seconds that have already been added, a number of other alternatives have been suggested. One proposal would be to switch to TAI as the legal time scale, which would effectively reset the leap second count to zero in a single step. Another proposal would be to stop adding leap seconds to UTC and to change the name of the time scale to reflect this change in its implementation.

59.16 TRACEABILITY

There are many applications that require time stamps that are traceable to a national time scale, and realizing this requirement requires clocks that are synchronized to a national or international standard of time. A clock is *traceable* to a national time scale if there is an unbroken chain of time-difference calibration measurements between the clock and the reference time scale by means of any of the methods that I have described before. Each one of these measurements must have an uncertainty estimate.

It can be difficult to establish the chain of measurements that is required for traceability. For example, the signal *in space* transmitted by a GPS satellite is traceable to UTC, the national and international reference time scale, through the US Naval Observatory, which monitors the time signals broadcast by the GPS constellation and computes the offset between GPS system time and UTC as maintained by the Naval Observatory. This offset is uploaded into the satellites and is transmitted as part of the navigation message.

However, the traceability of the signal in space does not necessarily extend to the timing signals produced by a GPS receiver unless the receiver, the antenna, and the connecting cable have been calibrated. The traceability almost certainly does not automatically extend to the application that uses the timing signals to apply time stamps as part of some application. This discussion does not suggest that these links in the chain are known to be inadequate or in error, but rather that they do not satisfy the strict definition of traceability without some sort of calibration procedure.

There are some situations where the requirements of strict traceability can be satisfied without a complex calibration procedure. For example, if an application requires that time stamps be traceable to a national time standard with an uncertainty of less than 1 s, then simply certifying that the satellite timing equipment is working properly is likely to be good enough. The uncertainty of the time signals produced by a receiver synchronized using signals from a GPS satellite is several orders of magnitude smaller than the requirements of the application so that the overall system is surely traceable at the level of 1 s if it is working at all. (Verifying that a GPS receiver is working properly may or may not be an easy job—it depends on the specific receiver that is used.)

A second aspect of traceability is *legal traceability*, by which I mean being able to establish in a legal proceeding that a time stamp was traceable to a national time scale. In this situation, “doing the right thing” might not be adequate if you can’t prove it to a judge and jury.

Given that the technical aspects of traceability that I discussed previously have been satisfied, establishing legal traceability is generally a matter of documentation—maintaining log files that show that the system was calibrated with an uncertainty consistent with the requirements of the application and that it was operating normally at the time in question. A log file that has entries only when there is a problem is unlikely to be adequate—it will have no entries when the system is working properly, and an empty log file is ambiguous and may not be of much help.

59.17 SUMMARY

I have discussed a number of methods for synchronizing a clock using a reference device that can be located either in the same facility or remotely and linked to the device under test by a communication channel. I have discussed a number of methods of synchronizing a remote clock and the statistical considerations that characterize the accuracy of the procedure and how often to request a calibration. An important tool in these discussions is the two-sample Allan variance, and I have presented a simple introduction into how this estimator is calculated and used.

59.18 BIBLIOGRAPHY

The following list contains a few of the very large number of publications that contain additional information on time and frequency standards and distribution methods:

1. The publications of the Time and Frequency Division of the National Institute of Standards and Technology, generally available online at tf.nist.gov.
2. "Encyclopedia of Time." Edited by Samuel L. Macey, New York, 1994, Garland Publishing, Inc.
3. Special issue on Time Scale Algorithms, Metrologia, Vol. 45, Number 6, December, 2008.
4. "Time and Frequency Measurement." Edited by Christine Hackman and Donald B. Sullivan, College Park, MD, 1996, American Association of Physics Teachers. This was also published as resource letter TFM-1 in the American Journal of Physics, Vol. 63, Number 4, pages 306–317, April, 1995.
5. "Computer Network Time Synchronization," 2nd edition. David L. Mills, New York, 2011, CRC Press.
6. Special Issue on Time and Frequency, Proceedings of the IEEE, Vol. 79, Number 7, July, 1991.
7. "Understanding GPS: Principles and Applications," 2nd edition. Edited by Elliott D. Kaplan and Christopher J. Hegarty, Norwood, MA, 2006, Artech House, www.artechhouse.com.
8. "Global Positioning System: Signals, Measurements and Performance," 2nd edition. Edited by Pratap Misra and Per Enge, Lincoln, MA, 2006, Ganga-Jamuna Press, GPStextbook@G-JPress.com.

REFERENCES

1. S. Bregni, "Measurement of the Maximum Time Interval Error for Telecommunications Clock Stability Characterization," *IEEE Trans. Instrum. Meas.*, Vol. 45, pages 900–906, October 1996.
2. K. Biholar, T1.101-199X, Synchronization Interface Standard, Draft Standard of the American National Standards Institute, Inc.
3. S. R. Stein, "Frequency and Time—Their Measurement and Characterization," *Precision Frequency Control*, Vol. 2, New York, Academic Press, 1985. This paper and a number of

others on the same topic are reprinted in NIST Technical Note 1337, edited by D. B. Sullivan, D. W. Allan, D. A. Howe, and F. L. Walls, published by the US Department of Commerce, March 1990.

4. B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, New York, Dover Publications, 1979, pages 54, ff.
5. R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*, New York, Dover Publications, 1959, pages 174, ff.
6. Y. Lin, S. Lee, S. Li, Y. Xie, Z. Ren, and C. Nguyen, “Series-Resonant VHF Micromechanical Resonator Reference Oscillators,” *IEEE J. Solid State Circuits*, Vol. 39, No. 12, pages 2477–2491, December 2004. See also Ville Kaajakari, *Practical MEMS*, Las Vegas, Small Gear Publishing, 2009. www.smallgearpublishing.com (accessed November 11, 2015).
7. C. Audoin and B. Guinot, “Atomic Frequency Standards,” *The Measurement of Time: Time, Frequency and the Atomic Clock*, Cambridge, Cambridge University Press, 2001.
8. H. G. Andresen and E. Pannaci, “Servo controlled hydrogen maser cavity tuning,” Proceedings 20th Annual Frequency Control Symposium, Atlantic City, NJ, Fort Monmouth, NJ, Electronic Components Laboratory, U.S. Army Electronics Command, 1966, pages 402–415. See also, C. Audoin, “Fast Cavity Auto-Tuning System for Hydrogen Maser,” *Rev. Phys.*, Vol. A16, pages 125–130, 1981.
9. G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco, CA, Holden-Day, 1970.
10. D. A. Howe, D. W. Allan, and J. A. Barnes, “Properties of signal sources and measurement methods,” Proceedings 35th Annual Symposium on Frequency Control, Washington, DC, Electronic Industries Association, 1981, pages A1–A47. This is document AD-A110870 available from National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.
11. S. Stein, D. Glaze, J. Levine, J. Gray, D. Hilliard, D. Howe, and L. Erb, “Performance of an automated high accuracy phase measurement system,” Proceedings 36th Annual Symposium on Frequency Control, Fort Monmouth, NJ, U.S. Army Electronics Research and Development Command, Electronics Technology and Devices Laboratory, 1982, pages 314–320. This is document AD-A130811 available from National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.
12. J. Levine, “Introduction to Time and Frequency Metrology,” *Rev. Sci. Instrum.*, Vol. 70, No. 6, pages 2567–2596, 1999.
13. G. Panfilo, A. Harmegnies, and L. Tisserand, “A new prediction algorithm for EAL,” Proceedings of the 2011 Joint European Time and Frequency Forum and International Frequency Control Symposium, Piscataway, NJ, IEEE, 2011, pages 850–855.
14. J. Levine, “The Statistical Modeling of Atomic Clocks and the Design of Time Scales,” *Rev. Sci. Instrum.*, Vol. 83, page 021101, 2012.
15. R. H. Jones and P. V. Tryon, “Estimating Time From Atomic Clocks,” *J. Res. Natl. Bur. Stand.*, Vol. 88, pages 17–28, 1983. See also A. Gelb, *Applied Optimal Estimation*, Cambridge, MA, MIT Press, 1974, Chapter 4.
16. Postprocessed orbits and clock solutions are computed by the International GPS Service and are available online at www.igs.org and https://www.igscb.jpl.nasa.gov/components/prods.html (accessed December 15, 2015).

17. Many references. See, for example, D. E. Comer, “Internetworking with TCP/IP,” *Principles, Protocols, and Architecture*, Vol. 1, Englewood Cliffs, NJ, Prentice-Hall, 1991.

18. J. Levine, M. Weiss, D. D. Davis, D. W. Allan, and D. B. Sullivan, “The NIST Automated Computer Time Service,” *J. Res. NIST*, Vol. 94, pages 311–322, 1989.

19. N. Ashby and M. Weiss, *Global Positioning Receivers and Relativity*, National Institute of Standards and Technology Technical Note 1385, Boulder, CO, U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology, 1999.

60

LABORATORY-BASED GRAVITY MEASUREMENT

CHARLES D. HOYLE, JR.
Department of Physics and Astronomy, Humboldt State University, Arcata, CA, USA

60.1 INTRODUCTION

Gravitational experiments invariably inspire images of planetary and galactic measurements, black hole mergers, large-scale gravitational wave detectors, and other massive undertakings. However, laboratory-scale tests of gravity involve relatively small masses and introduce many systematic effects not present in astronomical-scale investigations. In this regime, it is appropriate to discuss gravity as the Newtonian (weak-field) limit of General Relativity. Thus, we expect the gravitational force to obey the inverse-square law (ISL) if General Relativity is valid when small test masses are in close proximity. The higher order effects of General Relativity are not relevant at this scale because they are miniscule for any test masses that can be used in the laboratory. Therefore, tests of the ISL are effectively measurements of the validity of General Relativity at these scales. In addition, by selection of appropriate test masses and test mass geometries, the weak equivalence principle (WEP), a central feature of General Relativity, may also be effectively tested at short distances. Difficulty arises because the relative weakness of gravity compared to the other fundamental forces makes precision measurements of gravitational physics at laboratory scales precise and challenging work.

In addition to testing General Relativity itself, the motivation for improving tests of gravity at the laboratory scale spans many areas ranging from particle physics and cosmology to metrology and precision measurement. Tests of gravitational physics in this regime fall broadly in to five categories: tests of the Newtonian ISL, investigation of

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

the WEP, measurements of the gravitational constant (G), searches for new long-range forces that may couple to a variety of physical quantities or natural symmetries, and searches for effects that may yield insight into the nature of dark matter and/or dark energy. Many experiments are designed to test multiple areas at once by carefully choosing particular test masses and experimental configurations.

Due to the weakness of gravitational force between small test masses, it is necessary to eliminate any effects due to the Earth's gravitational field and other disturbances due to electromagnetic, thermal, and seismic effects in these experiments. Traditionally, torsion pendulums and other precision oscillators have been employed to decouple the experiments from environmental and systematic effects. However, novel atomic, molecular, and nuclear techniques are creating a new experimental arena for laboratory gravitational tests.

The question naturally arises, "What is meant by 'laboratory-scale'?" In the context of this chapter, "laboratory scale" means any experimental apparatus that is self-contained within the confines of a standard laboratory space and employs test masses small enough that the general relativistic corrections to the Newtonian ISL are smaller than the experimental resolution or uncertainty. This does not mean, however, that the effects under investigation are limited to operate only on this distance scale. For example, a long-range deviation from the ISL or WEP can be detected in certain laboratory-scale experimental configurations.

60.2 MOTIVATION FOR LABORATORY-SCALE TESTS OF GRAVITATIONAL PHYSICS

In recent years, tests of gravity have experienced a resurgence due in a large part to models of string or M-theory that predict modification of the gravitational force on short distance scales due to the influence of macroscopic extra spatial dimensions (for a review, see Ref. 1). However, many other theoretical scenarios predict possible violations of the ISL and WEP due to a variety of phenomena.

One of the most definitive string theory predictions in recent years suggests that gravity's strength increases at distances comparable to the size of proposed compactified spatial dimensions [2]. Other models predict weakening of gravity at small scales [3]. Such scenarios are proposed in an attempt to solve the gauge hierarchy problem (the discrepancy in energy scales between the Planck mass and the electroweak scale).

Attempts to explain the observed cosmic distance scale acceleration have also shown that the data would be consistent with a theory that predicts gravity to "turn off" at distances less than about 0.1 mm [4]. The observed value of the vacuum energy density, $\rho_{\text{vac}} \approx 3.8\text{keV/cm}^3$, corresponds to a length scale of $R_{\text{vac}} = \sqrt[4]{\hbar c/\rho_{\text{vac}}} \approx 85\mu\text{m}$, which may also have fundamental significance [5].

Finally, unobserved particles predicted by string theories, such as the dilaton and moduli, may also produce new short-range forces operating through the "chameleon mechanism" [6] that could be observed in short-range tests of gravity.

Most alternative models of gravity predict a violation of the WEP at some level due to interactions coupled to quantities other than mass or modifications of gravity itself. Scalar or vector boson exchange produces forces that inherently violate the WEP over a range determined by the Compton wavelength of the exchange particle, $\lambda = h/m_b c$. The WEP has been tested with incredible precision over distance scales from 1 cm to ∞ [7–9] but has never been subjected to a dedicated test in the subcentimeter regime (corresponding to exchange boson masses $\gtrsim 0.1\,\text{meV}$).

Measurements of the gravitational constant, G, have yielded widely varying results in recent decades, making it one of the least well-measured fundamental constants [10, 11]. Efforts to understand the discrepancy and determine an accurate value G are at the forefront of precision measurement research.

60.3 PARAMETERIZATION

A deviation from ISL behavior is generally modeled using a Yukawa addition to the classical Newtonian potential energy [1]. For point masses m_1 and m_2 separated by distance r, the modified potential energy becomes

$$V(r) = -\frac{Gm_1m_2}{r}\left(1+\alpha \mathrm{e}^{-r/\lambda}\right), \tag{60.1}$$

where G is the Newtonian gravitational constant, α is a dimensionless scaling factor corresponding to the strength of any deviation relative to Newtonian gravity, and λ is the characteristic length scale of the deviation.

It is generally assumed that a WEP violation would result in a coupling to some “charge” that is related to the seemingly conserved quantities of baryon number, B (atomic number, Z, plus neutron number, N), or lepton number, L ($L = Z$ for electrically neutral materials). A general scalar (−) or vector (+) Yukawa coupling would result a potential energy of the form [8]

$$V(r) = \mp\frac{1}{4\pi}\tilde{q}_1\tilde{q}_2\frac{e^{-r/\lambda}}{r}, \tag{60.2}$$

where $\tilde{q}_i$ are the “charges” of the test masses and λ is the Compton wavelength of the exchange boson. A common parameterization assumes $\tilde{q} = \tilde{g}\left[Z\cos\tilde{\psi} + N\sin\tilde{\psi}\right]$ where $\tilde{g}$ is a coupling constant and $\tilde{\psi}$ determines the type of charge. Recasting Equation 60.2 in a form similar to Equation 60.1 yields

$$V(r) = -\frac{Gm_1m_2}{r}\left(1+\tilde{\alpha}\left[\frac{\tilde{q}}{\tilde{g}\mu}\right]_1\left[\frac{\tilde{q}}{\tilde{g}\mu}\right]_2 e^{-r/\lambda}\right), \tag{60.3}$$

where μ is the mass of objects 1 or 2 in units of atomic mass units, u, and $\tilde{\alpha} = \pm\,\tilde{g}^2/(4\pi G u^2)$.

60.4 CURRENT STATUS OF LABORATORY-SCALE GRAVITATIONAL MEASUREMENTS

As discussed in the following sections, there is an ever-increasing variety of experimental techniques for exploring gravity at laboratory distances. The workhorse of gravitational measurements is the torsion pendulum [8], although techniques such as high-frequency oscillators and atomic/molecular interferometry are promising new techniques in this field [12, 13]. The following sections summarize the current level of experimental results obtained from laboratory-scale experiments.

60.4.1 Tests of the ISL

Tests of the ISL can be considered as measurements of the parameters α and λ of Equation 60.1. Previous experiments utilizing various types of torsion pendulums have eliminated large portions of the Yukawa potential α–λ parameter space [8, 14–18]. The shaded region of Figure 60.1 shows the current constraints in the α–λ plane for λ between a few micrometers up to 1 cm. Laboratory tests at larger ranges have been performed [1, 12] and have also discovered no deviation from Newtonian behavior.

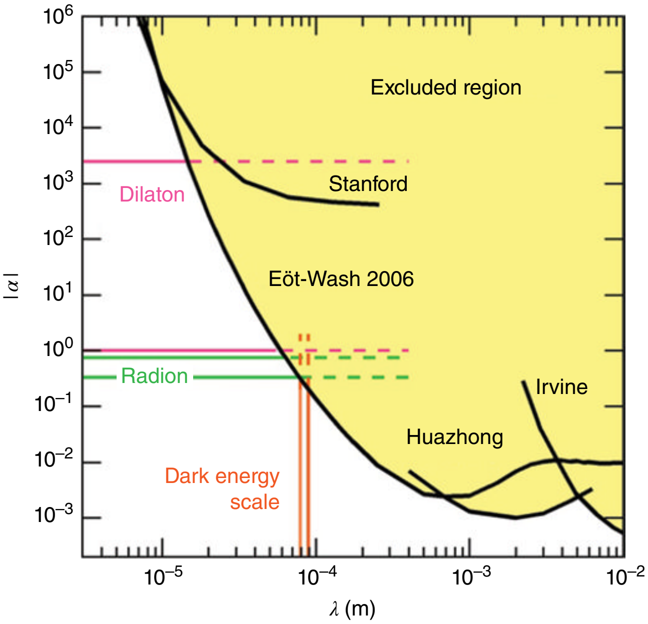


FIGURE 60.1 Current short-range experimental constraints in the $|\alpha|$–λ Yukawa parameter space and theoretical predictions. The shaded region is excluded at the 95% confidence level. Results from previous experiments are shown by the curves labeled Stanford [18], Eöt-Wash [15], Huazhong [16], and Irvine [17]. For a more detailed discussion of theoretical predictions, see Ref. 1.

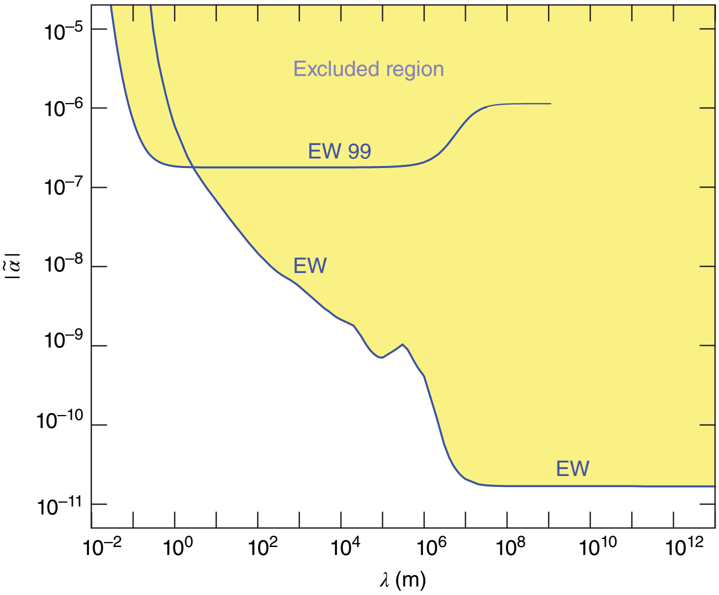


FIGURE 60.2 Best 95% confidence level constraints on violations of the WEP for interactions coupling to $\tilde{q} = B - L$. The curve labeled EW is from reference [7], while EW 99 is from reference [9]. Both limits were obtained with torsion pendulum experiments performed by the Eöt-Wash group at the University of Washington.

60.4.2 WEP Tests

The WEP is well characterized over distances from 1 cm to ∞ [7–9], but it is essentially untested below the centimeter scale. Various laboratory torsion pendulum experiments have set these constraints as summarized in Figure 60.2.

60.4.3 Measurements of *G*

The gravitational constant was one of the first to be measured with some precision by Cavendish (using a torsion balance) in 1798 [19]. It is surprising, therefore, that measurements performed over the last few decades involving a variety of techniques have yielded inconsistent results. In fact the published CODATA values of *G* have been essentially unchanged for decades even though there have been many precision measurements performed [11]. The inconsistency of these measurements has left the uncertainty in *G* unchanged. For a summary of recent measurements, see Figure 3 of Refs. 10 and 20.

60.5 TORSION PENDULUM EXPERIMENTS

Torsion pendulums (or torsion balances) have been used for gravitational measurements since at least the time of Cavendish [19]. A torsion pendulum is only sensitive to torque about the vertical axis and is therefore insensitive to the large vertical component of the Earth's gravitational field or any uniform horizontal forces. Modern torsion pendulums have incredibly high torque sensitivity and are the most sensitive devices for measuring feeble forces between macroscopic objects. For a nice review of torsion pendulums used in fundamental physics experiments, see Ref. 8.

60.5.1 General Principles and Sensitivity

Torsion pendulums, when operated in a vacuum, behave as lightly damped harmonic oscillators in a single rotational dimension whose rotation axis is local vertical. Measurements of gravitational effects are performed by carefully choosing the mass distribution of the pendulum and applying a torque via a specially designed attractor mass. Measurement of the applied torque can be obtained through analysis of the pendulum's twist angle (time series analysis) or free oscillation period (torsion period analysis). For experiments where the applied attractor mass torque is time varying, the time series analysis is advantageous, whereas static attractor configurations sometimes measure the torsion period and the harmonic content of the motion to extract the gravitational influence on the pendulum [8]. Other tests have successfully employed feedback loops to maintain the pendulum stationary [21]. Using this technique, the torque is inferred from the applied feedback signal. An ideal torsion pendulum experiment is null; that is, one in which Newtonian gravity has no effect on the pendulum. Most modern experiments make an effort to achieve an experimental configuration that approximates a true null as much as is realistically possible.

Because torsion pendulums are resonant systems, the angular response to applied torques is frequency dependent. Torque applied on resonance yields the largest angular response; however, noise sources and false effects are also enhanced the most on resonance, and the best results to date have used off-resonance driving techniques [15]. Precision angular measurement is therefore paramount in torsion pendulum experiments. Optical autocollimators have been developed to measure small angular deflections of order 1 nrad or less. Recent developments have begun to push this sensitivity even further [22]. An example of one such system is the one currently in use by the author's group at the Humboldt State University (HSU) Gravitational Research Laboratory. A sample noise spectrum for this autocollimator is shown in Figure 60.3. The autocollimators typically employ position-sensitive photodiodes to measure the deflection of a laser diode beam used as an optical lever. To increase sensitivity, the beam is typically chopped and recorded using lock-in techniques.

60.5.2 Fundamental Limitations

60.5.2.1 Statistical Noise Sources Thermal and readout noise set the sensitivity limit for any torsion balance experiment. In the nonviscous vacuum regime (typically below 10^{-4}–10^{-5} Torr), thermal noise within the torsion fiber itself is a limiting factor.

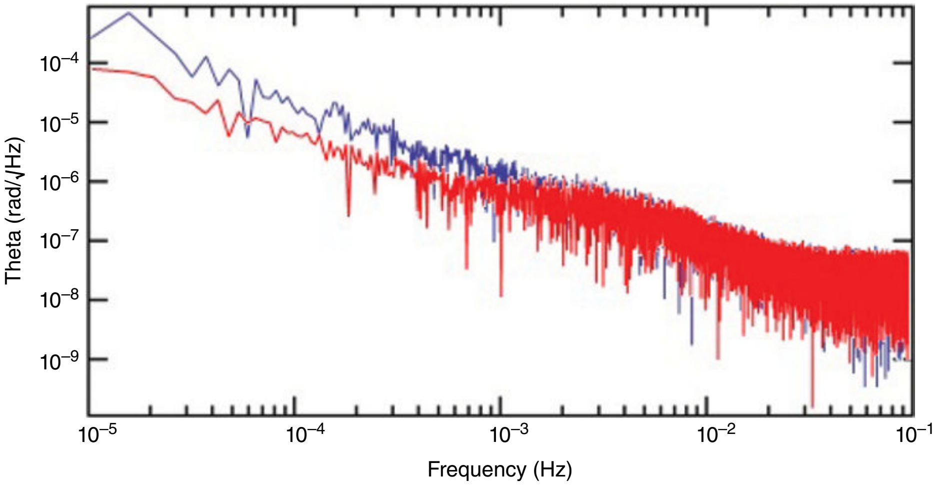


FIGURE 60.3 Sample noise spectrum of the Humboldt State autocollimator twist data taken off of a stationary mirror. The noise level at 5 mHz is 0.3 μrad/√Hz, which yields an uncertainty in angle measurement of 1 nrad in roughly 1 day of integration time. The lower curve was taken with the laser off (electronic noise only), while the upper is with the laser pulsing and incident on the position-sensitive detector.

Internal damping in the fiber material produces frequency-dependent torque noise with amplitude spectral density

$$S_{\mathrm{th}}^{1/2} = \sqrt{\frac{4k_{\mathrm{B}}T\kappa}{2\pi fQ}}, \tag{60.4}$$

where k_{B} is Boltzmann's constant, T is the temperature, f is frequency, and κ and Q are the torsion constant and quality factor of the torsion pendulum, respectively. Typical values of Q for a tungsten fiber are approximately 4000, while higher values have been reported for silicon fibers; however, a challenge associated with the use of silicon fibers is the necessity for the fiber to be conducting so that the pendulum does not accumulate an excess electrical charge [8].

Another noise source arises in the optical readout and the associated electronics. This effective torque noise increases with angular frequency, ω, as the torsion pendulum's resonant motion tends to hide driving signals with frequencies much greater than its resonant frequency [14]:

$$S_{\mathrm{ro}}^{1/2} = S_{\theta}^{1/2}\kappa\sqrt{\left(1-\frac{\omega^2}{\omega_0^2}\right)^2 + \left(\frac{1}{Q}\right)^2}. \tag{60.5}$$

In Equation 60.5, $S_{\theta}^{1/2}$ is the angular noise spectrum of the autocollimator, ω is angular frequency, and ω_0 is the pendulum's resonant angular frequency.

Reduction of statistical noise can be attained by improving autocollimators and lowering the running temperature. Attempts to perform torsion pendulum experiments at cryogenic temperatures are underway; however it remains to be seen whether improved noise levels can be obtained given the extra complication of running a low temperature apparatus [12, 23].

Other broadband sources of torque noise include fluctuating electric and magnetic fields, temperature variations, gravity gradient fluctuations, and changes in apparatus tilt. Typically these environmental sources are only of concern at the specific signal frequency and are therefore classified as sources of “false effects.”

60.5.2.2 False Effects False effects are any coherent effects that could mimic a violation of the ISL or WEP or otherwise contaminate the signal of interest for a given experiment. Due to the relative weakness of gravity, all environmental disturbances must be characterized and/or suppressed to ensure any observed signal is in fact due to a modification of gravity or other new physics. To do so, sources of false effects are generally greatly exaggerated and the response of the pendulum is measured to establish a “feedthrough” for the given source. The source is then measured in a running configuration and, with the known feedthrough, its effect on the pendulum during data acquisition can be established. If a coherent effect is resolved, an effort can be made to suppress it or otherwise it may be subtracted from the data in some cases. If there is no resolved effect, an upper limit can be placed using the measured feedthrough and noise level of the source. A discussion of specific sources of false effects follows in the following text.

Gravity gradient fluctuations due to large- or small-scale mass distribution variations can directly cause torque on the pendulum that could mimic a detection of new physics. Typically efforts are undertaken to vary the attractor mass configuration to be able to subtract any spurious gravitational effects. Reversal or rotation of attractor masses are routinely employed as well as measurement of the local gravitational field using specially designed gradiometer pendulums. Multipole techniques may be used to assess long-range gravity gradient effects on a torsion pendulum [24, 25].

While electromagnetic interactions can contribute to false effects, their contribution to noise is small due to the fact that torsion pendulums are typically made from non-magnetic materials and reside in gold-coated conducting enclosures. Large fluctuating magnetic fields are typically applied to observe the response of the pendulum and determine its magnetic moment. Measurement of ambient field fluctuations then yields a measure of any magnetically induced false effects. Reversal of the attractor masses also reverses the phase of, or may otherwise change, any magnetic false effects. Characterization of patch charge effects and other electrostatic phenomena is difficult. Efforts are made to ensure that the pendulum is electrically grounded to its surroundings and is contained in a perfectly conducting enclosure, including a conducting shield that prohibits any electrostatic influence of the attractor mass (Casimir or otherwise).

Temperature fluctuations are reduced by shrouding torsion pendulum apparatuses in a thermal enclosure. An additional thermal mass also surrounds the torsion fiber inside the vacuum chamber to provide further isolation. Intentional thermal variations are applied to observe the response of the pendulum. The observed level of temperature fluctuations during normal operation multiplied by any observed feedthrough will yield a measure of any temperature-related false effect. Temperature-related effects are notoriously difficult to quantify given the variety of mechanisms by which thermal expansion can affect the twist of the pendulum, so a large effort is typically devoted to keep the apparatus temperature as constant as possible (typical temperature stability at the level of milliKelvin is desired).

Apparatus tilt variations can cause twist of the pendulum due to the well-documented tilt–twist effect [14]. Precision tilt sensors are employed to measure both the purposeful exaggeration of the apparatus tilt (to measure the tilt–twist feedthrough) and the changes in tilt during running configuration.

Radiation pressure noise from the optical angle sensor is not typically a limiting factor in torsion pendulum experiments due to the placement of the beam on the torsion fiber axis and low optical power used by the autocollimator [9].

The ultimate goal of most torsion pendulum experiments is to run at thermal limit, at which the torque resolution is only limited by statistical sources and not by environmental disturbances and false effects. Although not easy, experiments do reach this limit [14, 15].

60.5.3 ISL Experiments

Torsion pendulums have provided the best tests of the ISL at short distances to date. A plot summarizing recent short-range results is shown in Figure 60.1. Deviations from the ISL have been excluded at the 95% confidence level from distances of 55 μm to beyond planetary scales. The following summarizes the specific experiments shown in Figure 60.1.

60.5.3.1 University of Washington Eöt-Wash Experiment At short distances, the torsion pendulum used by the University of Washington's Eöt-Wash group has provided the most stringent constraints on gravitational strength ($\alpha = 1$) effects [15]. The pendulum used a 21-fold azimuthally symmetric mass distribution that was hung over a similar attractor mass that rotated at angular frequency ω. As the attractor rotated, the oscillation of the pendulum's twist angle was recorded, and the distance between the pendulum and attractor was changed. Analysis of the harmonic content of the pendulum's twist at 21ω, 42ω, and 63ω was compared to a Newtonian prediction to extract limits in the α–λ parameter space. No deviation from Newtonian behavior was found. The group is currently improving the design to incorporate the 120-fold "wedge" symmetry shown in Figure 60.4. A second experiment under development and spearheaded by C. Hagedorn of this group utilizes planar pendulum and attractor mass geometry.

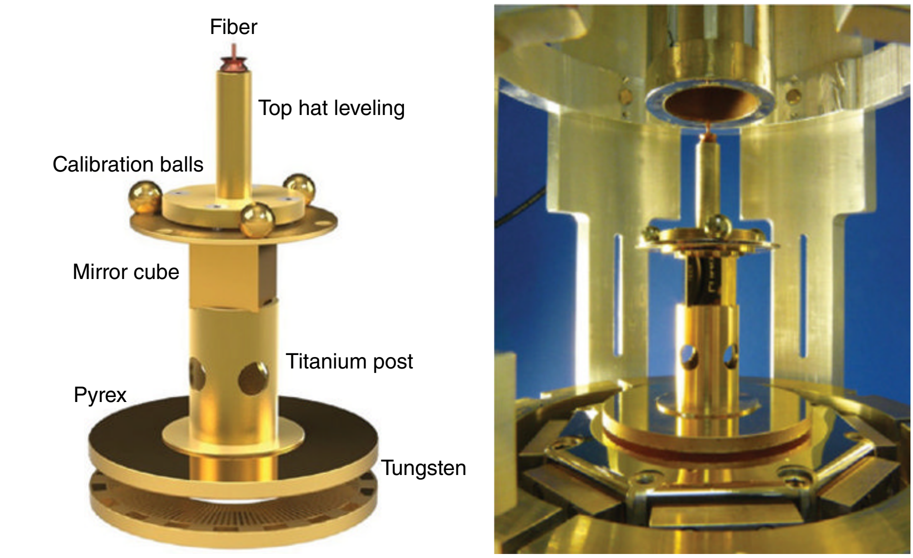


FIGURE 60.4 Left: diagram of next-generation Eöt-Wash torsion pendulum used for probing the ISL. The pendulum has a 120-fold symmetric "wedge" test mass design that will interact with a similarly shaped attractor mass that rotates beneath it. Right: picture of the pendulum suspended above the conducting membrane that electrostatically shields it from the attractor mass. Source: Reproduced with permission of Ted Cook.

60.5.3.2 Wuhan The Wuhan experiment employs a planar torsion pendulum design to achieve the best limits on the ISL in the 0.7–5 mm range [16]. The compensation masses and parallel-plate geometry used in this test provide an essentially null experiment.

60.5.3.3 Longer-Range Tests The UC Irvine/UW group has a well-established history of intermediate-range ISL tests [12]. The group is exploiting a torsion pendulum with cylindrical geometry that will have maximal sensitivity at $\lambda = 12\,\text{cm}$. The specially designed pendulum and attractor mass will also yield an essentially null test.

60.5.4 Future ISL Tests

Several experimental efforts are underway to improve tests of the ISL at various distance scales. We describe here one approach in detail that is being developed by the author's group at HSU.

60.5.4.1 HSU Experiment The HSU Gravitational Research Laboratory is developing a novel parallel-plate torsion pendulum and attractor configuration that proposes to obtain the best sensitivity down to $\lambda \approx 20\,\mu\text{m}$ with a nearly perfect null experiment. The experiment is run primarily by undergraduates.

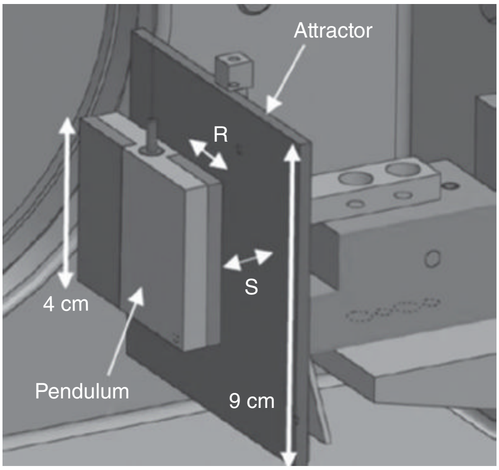


FIGURE 60.5 Basic geometry of the pendulum and attractor plate used in the Humboldt State experiment. When the pendulum/attractor separation, *s*, is modulated, the pendulum's stepped design results in potentially different short-range torques applied to each side, which in turn would result in twist motion at harmonics of the chosen attractor drive frequency. The pendulum is made from aluminum (light) and titanium (dark), and the attractor is copper. A conducting membrane that will separate the pendulum and attractor is not shown. The use of two different materials produces a composition dipole that provides sensitivity to violations of the WEP as well as the ISL.

The nearly null experiment is achieved by noting that the gravitational force does not depend on distance for a test mass interacting with an infinite plane of matter. The proposed tests exploit this fact by using a parallel-plate configuration with planar pendulums and a comparatively large attractor plate; a simplified geometry is shown in Figure 60.5, while Figure 60.6 shows an overview of the laboratory layout.

The proposed design is an aluminum/titanium pendulum with rectangular "steps" and the dimensions shown in Figure 60.5. The width of each step is R = 2 cm, while the thickness of each Al step is 6.25 and 3.75 mm for the Ti steps, ensuring equal mass in all steps. The total mass of the pendulum is 58 g. The attractor is a 9 cm × 9 cm × 0.3 cm-thick rectangular copper plate. The total mass of the plate is approximately 220 g.

The attractor–pendulum separation, *s*, is modulated by moving the attractor sinusoidally at angular drive frequency ω. In the ideal case of an infinite attractor plate, the Newtonian torque on the pendulum does not vary with attractor position, while any short-range interaction produces more torque on the closer, high-density "step" when *s* is smaller than the range of the interaction. This potential short-range torque modulation would cause a variation of the pendulum's twist at harmonics of the attractor modulation frequency (1ω, 2ω, 3ω, etc.). Note that Newtonian signal is only present due to the finite size of any *real* attractor mass and predominantly occurs at 1ω.

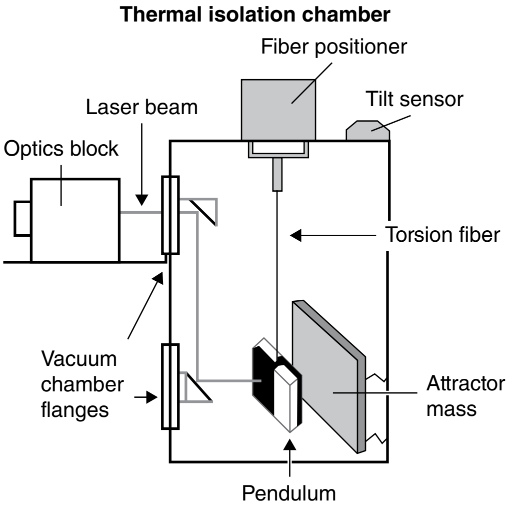
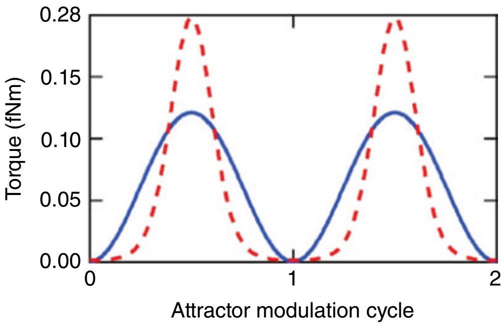


FIGURE 60.6 Overview of the existing laboratory setup at Humboldt State University.



Harmonic torque amplitudes (fNm)

	Newtonian	Yukawa
1ω	0.0605	0.0816
2ω	0.0004	0.0415
3ω	0.0000	0.0154
4ω	0.0000	0.0005
5ω	0.0000	0.0001

FIGURE 60.7 Left: calculated Newtonian and possible Yukawa torques on the Humboldt State pendulum as a function of time for two complete attractor modulation cycles. The peak-to-peak distance modulation amplitude is 0.5 mm and the minimum separation is 100 μm. Any Yukawa torque (dashed curve) with $\alpha = 1$ and $\lambda = 100\,\mu\text{m}$ would be clearly evident and larger than the Newtonian background (solid curve). Right: table of harmonic torque amplitudes for the times series shown on the left. Notice that for the chosen parameters, the 2ω and 3ω Yukawa signals are clearly different from the tiny Newtonian torque amplitudes. This difference in frequency dependence can be used to place constraints on Yukawa parameters, while systematic effects that will be largest at 1ω can be largely avoided.

Systematic false effects are also generally largest at 1ω. Thus the higher harmonics provide means to distinguish Newtonian and systematic false effects from short-range interactions. In fact, for an interaction with very small λ, the higher harmonics are similar in magnitude to one another because the applied torque is essentially a delta function at the closest pendulum/attractor separation (although the value of the torque diminishes rapidly as λ decreases). The harmonic content of a hypothetical interaction is shown in Figure 60.7.

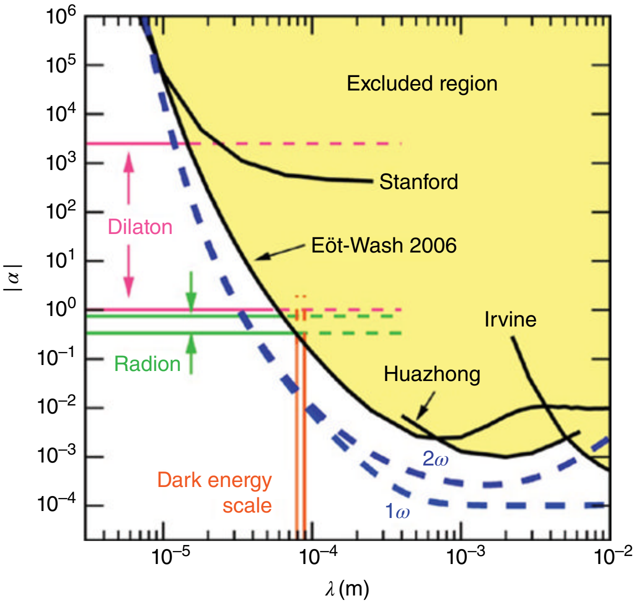


FIGURE 60.8 Reprint of Figure 60.1 including the predicted sensitivity for the HSU experiment. Each dashed line shows the predicted sensitivity of this apparatus for analysis of a single harmonic torque amplitude. Improved constraints may be obtained by analyzing multiple harmonics together. Note that for some values of λ, an improvement by a factor of approximately 50 is obtained over previous efforts.

The digitized angle signal is processed via Fourier techniques to determine the twist amplitude of the pendulum at harmonics of the attractor drive frequency. The amplitude of the twist oscillations will be compared to a detailed model of the expected pendulum/attractor Newtonian and possible Yukawa twist amplitudes. Limits in the α–λ parameter space will be obtained from this comparison, and any deviation from expected ISL (or WEP) values may be an indication of new physics. A prediction of the sensitivity for this design is presented in Figure 60.8

The surface of the pendulum will be polished to an optical finish and gold coated. The polish is necessary to reflect the autocollimator beam from the back side of the pendulum (side facing away from the attractor mass).

The two different materials naturally form a “composition dipole” and provide sensitivity to test the WEP. The materials will be joined first with an adhesive and subsequently machined and lapped to the appropriate dimensions. Fabrication of pendulums with different composition dipole pairs such as copper/titanium or molybdenum/aluminum will ensure that the WEP test is sensitive to all values of $\tilde{\psi}$.

60.5.4.2 Other Tests As previously noted, the Eöt-Wash and UC Irvine/UW group are developing several new ISL tests at short distances. Other research efforts utilizing torsion pendulums are also under way. Clive Speake’s group at the University of

Birmingham is implementing a superconducting torsion pendulum for gravitational tests [26]. The Cowsik group at Washington University is developing a pendulum supported by a torsion ribbon that uses a novel autocollimator [12].

60.5.5 WEP Tests

The best limits on violations of the WEP spamming the distance scale from 1 cm to ∞ have been obtained by composition dipole torsion pendulums in both rotating and nonrotating experimental configurations.

60.5.5.1 Eöt-Wash Rotating Torsion Balance Tests The best limits on a violation of the WEP over distance scales from roughly 3 m to ∞ (curve labeled EW in Fig. 60.2) have been determined with a rotating torsion balance used by the University of Washington Eöt-Wash group [7]. The rotating torsion balance used for WEP tests employs a composition dipole pendulum in a vacuum chamber that is rotated uniformly with frequency on the order of millihertz. Any differential acceleration resulting from the composition dipole would produce a twist of the pendulum at the rotation frequency. Thus, the experiment is sensitive to sources of WEP violations at any length scale greater than approximately the distance from the pendulum to the nearest stationary source mass.

The most recent results from this group were obtained using a torsion pendulum with two different composition dipole configurations: Be-Ti and Be-Al. A picture of this pendulum with its gold-coated test bodies is shown in Figure 60.9.

60.5.5.2 Eöt-Wash Short-Range WEP Test A separate experiment was conducted by the same University of Washington group that instead used a stationary Cu–Pb torsion pendulum interacting with a 3-ton rotating attractor mass made of ^{238}U [9]. The result from this experiment is labeled as EW 99 in Figure 60.2. The neutron-rich attractor allowed expanded exploration of the WEP parameter space; in particular, sensitivity was greatly enhanced for violations of the WEP that arise from neutron excess ($B-L$). The limits obtained represent the best from roughly 1 cm to 3 m. Constraints on violations of the WEP at shorter distances are inferred from ISL experiments. However, the previously mentioned experiment under development by the author's group at HSU is specifically designed to yield measurements of the WEP at subcentimeter distance scales with the possibility to probe diverse composition dipole mass pairings.

60.5.6 Measurements of *G*

Torsion pendulums have been widely used for measurement of the gravitational constant (see Ref. 10). The measurement of G with the smallest uncertainty was obtained by a torsion pendulum used in angular acceleration feedback mode [21]. Several torsion balance experiments are ongoing; however, atom interferometry and other new and promising techniques are on the horizon.

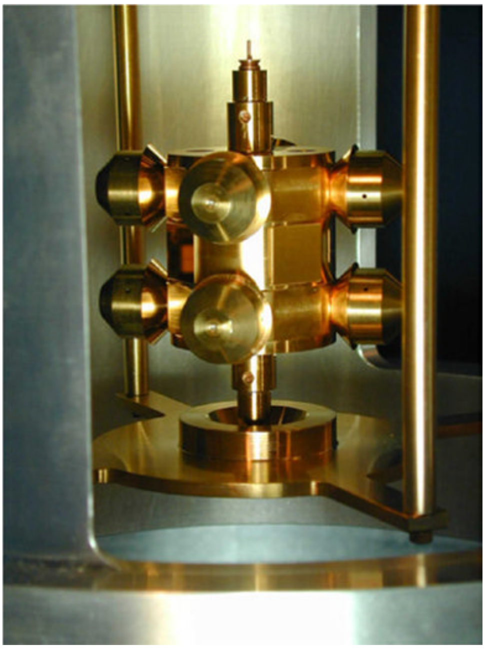


FIGURE 60.9 Eöt-Wash torsion pendulum used in the rotating torsion balance WEP tests of reference [7]. Source: Wagner et al. [7]. Reproduced with permission of IOP.

60.6 MICROOSCILLATORS AND SUBMICRON TESTS OF GRAVITY

For the 10 μm scale and below, the practicality of suppressing electrostatic backgrounds in torsion pendulum experiments becomes insurmountable. At these scales, high-frequency oscillator techniques have been used to place constraints on the ISL; however, no experiment has yet achieved the sensitivity to measure gravitational strength effects.

60.6.1 Microcantilevers

Josh Long at Indiana University is continuing an experimental effort started by the John Price group at the University of Colorado that uses a resonant microcantilever system to search for violations of the ISL [27].

The Kapitulnik group at Stanford University has performed an extended series of measurements also employing a microcantilever system [18, 28].

60.6.2 Very Short-Range ISL Tests

To measure violations of the ISL at very short ranges (<10 μm), experimenters must compete directly with the Casimir force. Most tests in this regime perform direct comparison of the measured force to the Casimir prediction to derive limits on deviations from the ISL. Several recent Casimir measurements have yielded limits on Yukawa-type deviations from the ISL at short distances [29, 30].

60.7 ATOMIC AND NUCLEAR PHYSICS TECHNIQUES

The most recent measurement of the gravitational constant was performed using cold atom interferometry [10]. This novel technique along with other proposed atomic, nuclear, and molecular methods [13, 31–35] have the possibility to open up a wide range of new gravitational measurements in the near future.

ACKNOWLEDGEMENTS

The author's work and the HSU Gravitational Research Laboratory are supported by the National Science Foundation under grants PHY-1065697 and PHY-1306783. The author would like to thank Todd Wagner and Ted Cook for providing figures and images.

REFERENCES

1. E.G. Adelberger, B.R. Heckel, and A.E. Nelson, *Ann. Rev. Nucl. Part. Sci.* 53, 77 (2003).
2. N. Arkani-Hamed, S. Dimopoulos, and G.R. Dvali, *Phys. Lett. B* 436, 257 (1998).
3. G. Dvali, G. Gabadadze, M. Kolanovic, and F. Nitti, *Phys. Rev. D* 65, 024031 (2001).
4. R. Sundrum, *Phys. Rev. D* 69, 044014 (2004).
5. S.R. Beane, *Gen. Relativ. Gravit.* 29, 945 (1997).
6. J. Khoury and A. Weltman, *Phys. Rev. Lett.* 93, 171104 (2004).
7. T.A. Wagner, S. Schlamminger, J.H. Gundlach, and E.G. Adelberger, *Class. Quantum Gravit.* 29, 184002 (2012).
8. E.G. Adelberger, J.H. Gundlach, B.R. Heckel, S. Hoedl, and S. Schlamminger, *Prog. Part. Nucl. Phys.* 62, 102 (2009).
9. G.L. Smith, C.D. Hoyle, J.H. Gundlach, E.G. Adelberger, B.R. Heckel, and H.E. Swanson, *Phys. Rev. D* 61, 22001 (2000).
10. G. Rosi, F. Sorrentino, L. Cacciapuoti, M. Prevedelli, and G.M. Tino, *Nature* 510, 518–521 (2014).
11. P.J. Mohr, B.N. Taylor, and D.B. Newell, *Rev. Mod. Phys.* 84, 1527–1605 (2012).
12. R.D. Newman, E.C. Berg, and P.E. Boynton, *Space Sci. Rev.* 148, 175–190 (2009).
13. A.A. Geraci, S.B. Papp, and J. Kitching, *Phys. Rev. Lett.* 105, 101101 (2010).
14. C.D. Hoyle, D.J. Kapner, B.R. Heckel, E.G. Adelberger, J.H. Gundlach, U. Schmidt, and H.E. Swanson, *Phys. Rev. D* 70, 042004 (2004).

15. D.J. Kapner, T.S. Cook, E.G. Adelberger, J.H. Gundlach, B.R. Heckel, C.D. Hoyle, and H.E. Swanson, *Phys. Rev. Lett.* 98, 021101 (2007).
16. S.-Q. Yang, B.-F. Zhan, Q.-L. Wang, C.-G. Shao, L.-C. Tu, W.-H. Tan, and J. Luo, *Phys. Rev. Lett.* 108, 081101 (2012).
17. J.K. Hoskins, R.D. Newman, R. Spero, and J. Schultz, *Phys. Rev. D* 32, 3084 (1985).
18. A.A. Geraci, S.J. Smullin, D.M. Weld, J. Chiaverini, and A. Kapitulnik, *Phys. Rev. D* 78, 022002 (2008).
19. H. Cavendish, *Philos. Trans. R. Soc. Lond.* 88, 469–526 (1798).
20. C. Speake and T. Quinn, *Phys. Today* 67, 7, 27 (2014).
21. J.H. Gundlach and S. Merkowitz, *Phys. Rev. Lett.* 85, 2869 (2000).
22. M.D. Turner, C.A. Hagedorn, S. Schlamminger, and J.H. Gundlach, *Opt. Lett.* 36, 1479–1481 (2011).
23. F. Fleischer, E.G. Adelberger, M. Bassan, and B. Heckel, American Physical Society April Meeting/AAPT Meeting, February 13–17, Washington, DC (2010).
24. E.G. Adelberger, N.A. Collins, and C.D. Hoyle, *Class. Quantum Gravit.* 23, 125 (2006).
25. C. D’Urso and E.G. Adelberger, *Phys. Rev. D* 55, 7970 (1997).
26. G.D. Hammond, C.C. Speake, A.J. Matthews, E. Rocco, and F. Peña-Arellano, *Rev. Sci. Instrum.* 79, 025103 (2008).
27. J.C. Long, H.W. Chan, A.B. Churnside, E.A. Gulbis, M.C.M. Varney, and J.C. Price, *Nature* 421, 922–925 (2003).
28. S.J. Smullin, A.A. Geraci, D.M. Weld, J. Chiaverini, S. Holmes, and A. Kapitulnik, *Phys. Rev. D* 72, 122001 (2005).
29. V.M. Mostepanenko, R.S. Decca, E. Fischbach, G.L. Klimchitskaya, D.E. Krause, and D. López, *J. Phys. A: Math. Theor.* 41, 164054 (2008).
30. R.S. Decca, D. López, H.B. Chan, E. Fischbach, D.E. Krause, and C.R. Jamell, *Phys. Rev. Lett.* 94, 240401 (2005).
31. F. Sorrentino, M. de Angelis, A. Bertoldi, L. Cacciapuoti, A. Giorgini, M. Prevedelli, G. Rosi, and G.M. Tino, *J. Eur. Opt. Soc.* 4, 1990 (2009).
32. A. Peters, K.Y. Chung, and S. Chu, *Metrologia* 38, 25 (2001).
33. T. Jenke, G. Cronenberg, J. Burgdörfer, L.A. Chizhova, P. Geltenbort, A.N. Ivanov, T. Lauer, T. Lins, S. Rotter, H. Saul, U. Schmidt, and H. Abele, *Phys. Rev. Lett.* 112, 151105 (2014).
34. V.V. Nesvizhevsky, V.V. Nesvizhevsky, H.G. Borner, A.K. Petukhov, H. Abele, S. Baessler, F.J. Ruess, T. Stoferle, A. Westphal, A.M. Gagarski, G.A. Petrov, and A.V. Strelkov, *Nature* 415, 297–299 (2002).
35. S. Dimopoulos, P.W. Graham, J.M. Hogan, and M.A. Kasevich, *Phys. Rev. Lett.* 98, 111102 (2007).

61

CRYOGENIC MEASUREMENTS

Ray Radebaugh
Applied Chemicals and Materials Division, National Institute of Standards and Technology[1], *Boulder, CO, USA*

61.1 INTRODUCTION

Cryogenics usually refers to temperatures less than about 120 K, but in this chapter we use such a definition rather loosely. Typically there is a gradual change in the type of sensors used or measurement methodology as temperature is lowered rather than an abrupt change as the temperature is lowered below 120 K. In some cases, a sensor appropriate for 100 K may also be the best for a temperature as high as 300 K. In this chapter we will specify the temperature range appropriate for a particular sensor or measurement methodology. Instrumentation for processes or experiments involving cryogenic temperatures often requires the use of sensors that must operate at these low temperatures. Certainly the measurement of temperature can only be done with a thermometer at the temperature of interest. However, certain other parameters, such as pressure, flow, liquid level, and magnetic field, have often been made with the active sensor located at room temperature but which could infer the property at cryogenic temperatures. This procedure usually resulted in a loss of accuracy, particularly under dynamic conditions. When cryogenic sensors were required in the early years of cryotechnology, they were usually constructed in the laboratory. The demand for cryogenic

[1] Contribution of the United States Government; not subject to copyright in the United States.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

sensors has grown sufficiently that commercial sensors are often available for use at cryogenic temperatures. In some cases, simple modifications of commercial sensors suffice to make them adaptable for use at cryogenic temperatures. In this chapter we review the availability and properties of commercial sensors and discuss the necessary modifications to make them useful for cryogenic temperatures.

The temperature range of primary interest here is between 4 and 300 K. Except for the case of some thermometers, a sensor that functions at 4 K will usually continue to function at lower temperatures as long as the power input is not too great. One of the main reasons commercially available sensors or transducers cannot be used at cryogenic temperatures is because of the choice of materials. In some cases, a material (e.g., rubber) undergoes a brittle transition at some low temperature that prevents its use at cryogenic temperatures. In other cases the differential contraction of different materials may be great enough at cryogenic temperatures to cause too high stresses or interference with moving components. Sensors with moving parts (such as flow sensors) are particularly difficult to operate at cryogenic temperatures because of the need for dry lubrication. Often electrical power inputs that are satisfactory for operation at room temperature can cause a sensor to self-heat or interfere with the overall experiment at cryogenic temperatures. Thus, commercial sensors can often be adapted for use at cryogenic temperatures by reducing the power input and/or changing a few key materials. Calibration of the sensor at the temperature it is to be used is nearly always necessary. Such calibrations often involve a comparison with a standard at room temperature.

61.2 TEMPERATURE

61.2.1 ITS-90 Temperature Scale and Primary Standards

The temperature scale in use today is known as the International Temperature Scale of 1990 (ITS-90), which extends from 0.65 to 1358 K [1]. It is a very close approximation to a true thermodynamic temperature scale. The scale is established by use of physical phenomena known so well that temperature can be calculated without any unknown quantities. Examples include equation of state of a gas, the velocity of sound in a gas, the thermal voltage or current noise in a resistor, blackbody radiation, and the angular anisotropy of gamma-ray emission from some radioactive nuclei in a magnetic field. A provisional extension covering the range from 0.9 mK to 1 K was established in 2000 and is known as PLTS-2000. It uses the melting curve of ^{3}He as the defining scale [2]. The ITS-90 is defined through a set of fixed points, interpolating primary thermometers, and interpolating equations [1, 3–5]. Fixed points are triple points and superconducting transition points. A standard platinum resistance thermometer (SPRT) is an example of an interpolating primary thermometer for temperatures between the triple point of equilibrium hydrogen at 13.8033 K and the freezing point of silver at 961.78 K. Use of fixed points and primary thermometers is a complex and expensive undertaking, which limits their use mostly to national standards institutions. The primary standards are transferred to secondary standards, such as high-purity platinum or Rh—Fe alloy resistance thermometers. The secondary standards are then used to calibrate commercial (industrial) thermometers for customer use.

61.2.2 Commercial Thermometers

Thermometry at cryogenic temperatures has become very well developed with a wide range of commercially available thermometers for the measurement of temperatures from the millikelvin temperatures up to room temperature and above. Excellent reviews of cryogenic thermometry have been published [6–12]. A new class of thermometers not discussed in these previous reviews (except Ref. 7) is the zirconium oxynitride ceramic film thermometers, which have low magnetoresistive effects. Most commercial thermometers for cryogenic temperatures are resistors, diodes, thermocouples, or capacitors. The change of their electrical characteristic with temperature determines their suitability as a thermometer. A good thermometer should have high sensitivity and be stable over time. For dynamic measurements it should also have a fast response time.

61.2.2.1 Metallic Resistance Thermometers The resistance of most pure metals varies roughly linearly with temperature until at some low temperature the scattering of electrons by impurities dominates the resistance, which leads to a lower limit to the resistance. In most cases, this limit is reached by 4 K, which means it can no longer function as a thermometer. For standard-grade platinum the ratio of resistance at 4.2 K to that at 273.16 K is usually less than 4×10^{-4}. More impurities cause this ratio to increase and the lower limit to be reached at higher temperatures. Platinum is the most widely used metallic resistance thermometer because it is so reproducible over long periods of time. It is often used as a secondary standard to calibrate other commercial thermometers. The platinum wire used in standard thermometers is of very high purity and care is taken in the construction of the thermometer to eliminate strains in the wire, which can affect the resistance. Figure 61.1 shows the typical construction of a capsule type of SPRT. The capsule type is most commonly used for cryogenics as opposed to the long-stem type in which heat conduction in the stem from ambient temperature to low temperature can cause unacceptable heat leaks. A helix of platinum wire (about

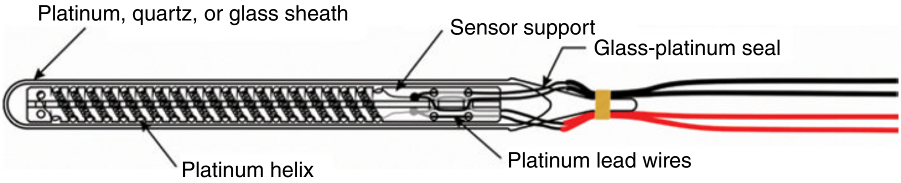


FIGURE 61.1 Cross section of a standard platinum resistance thermometer (SPRT).

75 μm diameter) is bifilarly wound on a notched mica or ceramic cross and placed inside the sheath, after which it is annealed at about 600°C to remove all strains. The capsule is filled with helium gas to enhance heat transfer between the platinum element and the sheath. Platinum, Inconel, glass, and quartz are common sheath materials. Dimensions of the capsule including the platinum-glass electrical feedthrough are about 5.8 mm in diameter by 56 mm long, with a resistance of 25.5 Ω at 0°C. The ITS-90 temperature range for these thermometers is from 13.8 K up to about 250°C. Special high-temperature versions are used for temperatures up to the silver triple point at 961.93°C. The reproducibility of these standard-grade thermometers is about 1 mK. Miniature capsule-type SPRTs have been developed recently that are about 3.2 mm in diameter and 9.7 mm long that can be used down to 13.8 K with some sacrifice in reproducibility [13].

Figure 61.2 shows how resistance varies with temperature for the most common metallic resistance thermometers. These thermometers have a positive temperature coefficient of resistance. A capacitance thermometer characteristic is also shown in Figure 61.1, but it will be discussed later. Two curves are shown for platinum thermometers, one for laboratory-grade SPRTs and one for PRTs made with lower-purity platinum, referred to as an industrial-grade platinum resistance thermometer (IPRT or just PRT). The resistance of the SPRT follows the ITS-90 definition down to 13.8 K, whereas the PRT meets the ITS-90 definition only down to about 70 K through the use of a slightly different resistance versus temperature curve. Their reproducibility is about 5 mK or higher. The distinction between the two grades is based on the purity of

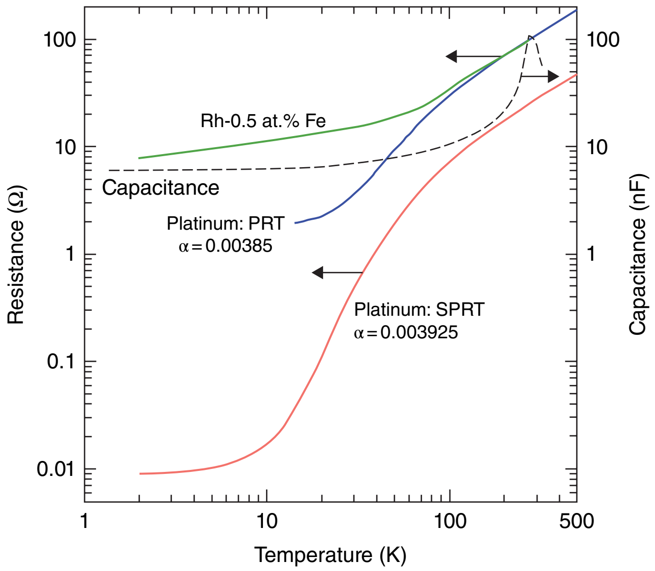


FIGURE 61.2 Characteristics of metallic resistance thermometers and capacitance thermometers.

the platinum and how the active element is supported to eliminate strain. The distinction is quantified by use of the resistance ratio given by

$$W(T) = \frac{R(T)}{R(273.16\text{K})}. \tag{61.1}$$

To be suitable for a SPRT, the platinum must be of sufficient purity to satisfy at least one of the two following relations [1]:

$$W(302.9146\text{K}) \geq 1.11807 \tag{61.2a}$$

$$W(234.3156\text{K}) \leq 0.844235. \tag{61.2b}$$

The defining equations that relate temperature to the ratio W are given by [1]. The constants in the equations are determined from calibrations. An alternate specification of platinum purity used in thermometers is the temperature coefficient of resistance α defined by

$$\alpha = \frac{R(100°\text{C}) - R(0°\text{C})}{100\,R(0°\text{C})}, \tag{61.3}$$

which technically has units of ohms/(ohm·°C) because the 100 in the denominator represents 100°C temperature change, but conventionally α is considered dimensionless. To meet the ITS-90 conditions given by Equations 61.2a and 61.2b, the temperature coefficient α should satisfy

$$\alpha \geq 0.003925. \tag{61.4}$$

This high value of alpha is achieved only with expensive reference-grade platinum (99.999% purity) wound in a strain-free manner and used in laboratory-grade SPRTs. The resistance at 0°C is normally 25.5 Ω. With reference-grade platinum in industrial thermometers, the temperature coefficient is 0.003920. Lower-purity platinum is used in most industrial-grade thermometers. Different standards organizations have adopted different temperature coefficients as their standard. The most widely used standard is the European standard (also widely used in the United States and elsewhere), designated by DIN IEC 60751 and ASTM E-1137 in which $\alpha = 0.0038500$ and the resistance at 0°C is 100 Ω. The resistance of industrial-grade PRTs follows a standard curve down to about 70 K. They are usable for lower temperatures but require an individual calibration. There are three tolerance grades (A, B, and C) for the DIN standard and two for the ASTM standard. For the ASTM standard the grade A has a tolerance ranging from ±0.47 K at 73 K to ±0.13 K at 273 K, whereas the grade B tolerance is ±1.1 K at 73 K and ±0.25 K at 273 K. The tolerance indicates the level of interchangeability for the thermometer.

Resistance thermometers made with Rh-5 at.% Fe have a resistance that continues to change even below 1 K. They are useful for the temperature range of 0.65–500 K, with a linear response above 100 K, as shown in Figure 61.2. RhFe thermometers are not interchangeable like that of Pt thermometers, but their reproducibility of about 0.2 mK in models fabricated like that of the SPRT makes them a candidate for an interpolating standard below 25 K for the ITS-90 temperature scale [14]. Other metals and alloys are sometimes used in resistance thermometers for special reasons. Resistance thermometers that use pure metals are not very sensitive for temperatures of 4 K and below. For these lower temperatures, thermometers made with semiconductors or other negative temperature coefficient materials become a better choice.

61.2.2.2 Semiconductor-Like Resistance Thermometers Figure 61.3 shows the resistance response curves for several types of semiconductor-like resistance thermometers. As the figure indicates, their resistance is very sensitive to temperature below about 100 K, unlike that of platinum thermometers. The response curves for these semiconductor-like thermometers have a negative temperature coefficient. The disadvantage with these thermometers is that except for the RuO_2 thermometers, they do not follow a standard response curve as do platinum thermometers, so they are not interchangeable and must be individually calibrated. The RuO_2 thermometers are only interchangeable for the same manufacturer. The zirconium oxynitride thermometers, sold under the trade name as Cernox thermometers, are a commonly used type and are

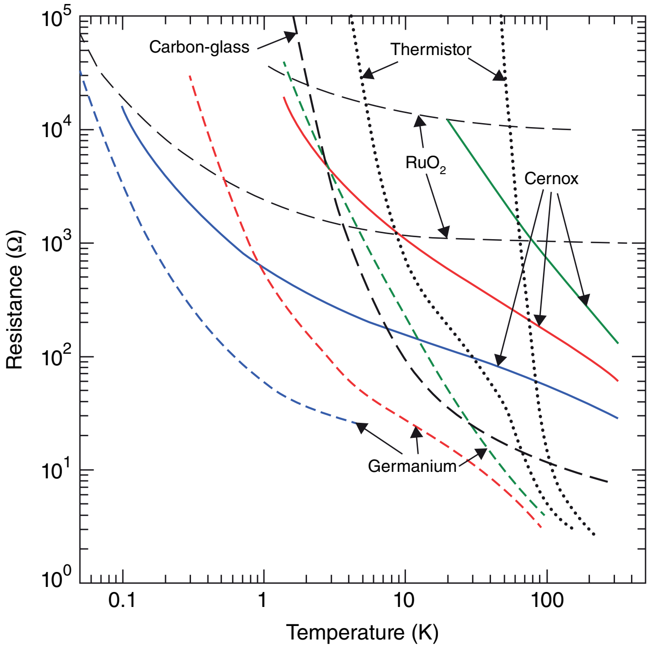


FIGURE 61.3 Characteristics of semiconductor and semiconductor-like thermometers.

available in several versions to cover a wide temperature range. They are a thin film resistor deposited on an alumina substrate. Their reproducibility of 3 mK at 4 K and 15 mK at 77 K allows them to be used in most laboratory settings. Germanium resistance thermometers are generally used for precision measurements below about 80 K. Their reproducibility is about ±0.5 mK at 4 K, but they must be individually calibrated. They can yield accuracies of about 5–15 mK for temperatures between about 1 and 20 K with commercial calibrations. Selected germanium thermometers are useful as thermometers down to about 50 mK and are available with commercial calibrations down to that temperature. Carbon-glass thermometers have a very steep response curve, which gives them high sensitivity. They are made by impregnating porous glass with carbon [15]. They are not interchangeable, but their reproducibility is quite good. Their high sensitivity makes them useful for temperature control. Carbon resistors in the form of electrical circuit resistors have been used often for very low-cost thermometers, but only particular brands have been found to be useful. A useful brand may be discontinued or experience a composition change that greatly affects its low-temperature resistance behavior. Thermistors usually have the steepest response curves of all thermometers, which limits the temperature range for an individual thermistor. Their use is then limited to special applications, such as in precise temperature control.

61.2.2.3 Diode Thermometers The forward voltage of diodes with constant current excitation varies with temperature and makes a good thermometer. Figure 61.4 shows typical voltage curves for Si and GaAlAs diodes with a 10 μA current excitation.

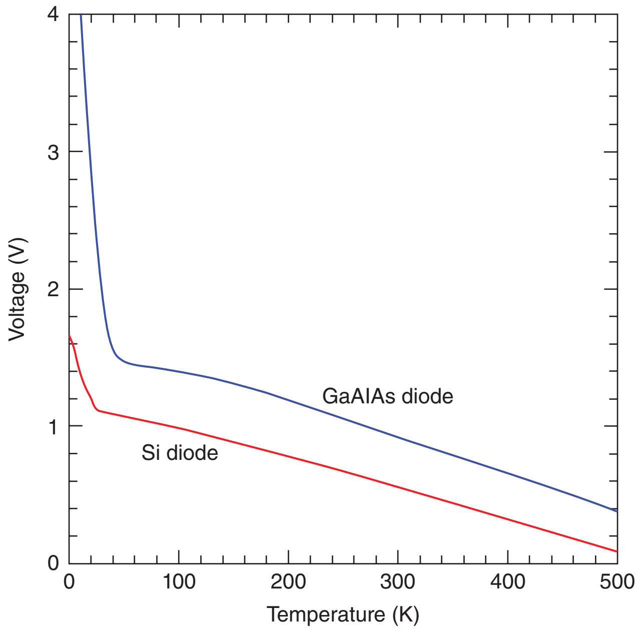


FIGURE 61.4 Characteristics of two types of diode thermometers with 10 μA current.

Special Si diodes make excellent thermometers because of their interchangeability for the same manufacturer and their relatively large voltage signal. They follow a standard curve to within 0.25 K for the "A" grade and 0.5 K for the "B" grade. Magnetic fields have a strong effect on Si diodes but much less effect on GaAlAs diodes. Unfortunately, GaAlAs diodes do not follow a standard curve, so they must be individually calibrated.

61.2.2.4 Thermocouples Thermocouples make use of the thermopower or the Seebeck coefficient in metals. If an electrical conductor is placed in a temperature gradient, current carriers (electrons or holes) will diffuse from the hot to cold end to build up a voltage that prevents further diffusion. The voltage gradient with zero current flow is known as the absolute Seebeck coefficient, which is given by

$$S = -\frac{dV}{dT}, \tag{61.5}$$

where S has units of V/K. Note that it is not possible to measure S directly, since any attempt to measure the voltage with a voltmeter will introduce another conductor in the temperature gradient with its own Seebeck coefficient. What is measured is the relative Seebeck coefficient, which is the difference between the absolute values in the two conductors. The measured Seebeck coefficient of the conductor pair is given as

$$S_{\mathrm{AB}} = S_{\mathrm{A}} - S_{\mathrm{B}} = \frac{dV_{\mathrm{B}}}{dT} - \frac{dV_{\mathrm{A}}}{dT} = -\frac{dV_{\mathrm{AB}}}{dT}. \tag{61.6}$$

The sign convention is quite complicated and will not be discussed here. Because the entropy of charge carriers in a superconductor is zero, S of a superconductor is zero. Thus, by choosing one leg of a pair to be a superconductor, S of the other leg can be determined from the measurement of the pair. Such a technique is useful only up to about 120 K, above which no superconductors exist. For higher temperatures $S(T)$ is found from the difficult measurements of the Thomson coefficient μ and use of the relation

$$S(T) = \int_0^T \frac{\mu(T')}{T'} dT'. \tag{61.7}$$

Figure 61.5 shows the temperature dependence for the absolute Seebeck coefficient for materials commonly used in thermocouples [16]. For a finite temperature difference, the voltage across the conductor is the integral of S over the given temperature range. Though the absolute Seebeck coefficient has no practical use, it aids in understanding voltages developed in thermocouple measurement systems. It shows that voltages are developed along the length of conductors in temperature gradients and not at the junctions. The junction simply ensures that the electrical potential of both legs is the

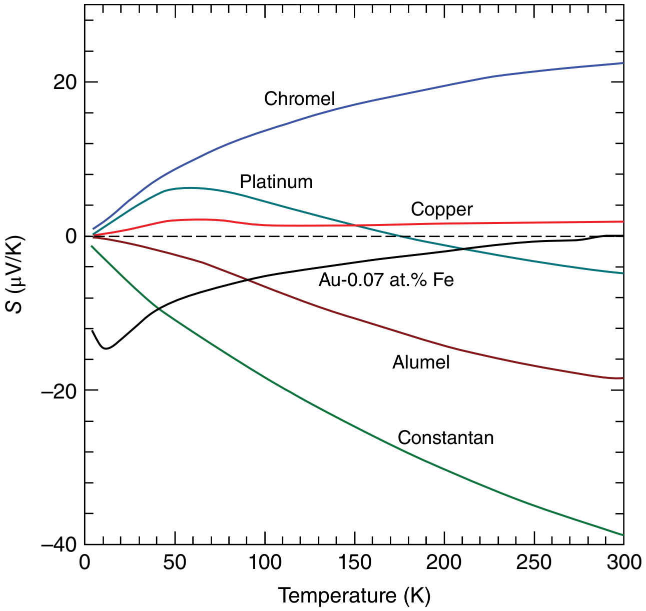


FIGURE 61.5 Absolute Seebeck coefficient of several metals used in thermocouples.

TABLE 61.1 Thermocouples Useful for Cryogenic Temperatures

Type	Positive Wire	Negative Wire	US Color Code	*T* Range (K)	Std. Error
E	Chromel	Constantan	+ purple; − red	3–1173	1.5 K
J	Iron	Constantan	+ white; − red	63–1073	1.5 K
K	Chromel	Alumel	+ yellow; − red	3–1573	1.5 K
T	Copper	Constantan	+ blue; − red	3–673	1% of *T*
Au—Fe	Chromel	Au-0.07 at.% Fe	—	1–573	0.2% of *V*

same at that point. For example, if both legs of a pair are of the identical material and the open ends are at the same temperature, then according to Equation 61.6 there will be no measured voltage difference between the pair. Also, if a third material is introduced in an isothermal part of the thermocouple circuit, then it has no effect on the output because no additional voltage is introduced in the isothermal section. Lastly, Figure 61.5 is useful in understanding the polarity of each leg in a thermocouple.

Several metal combinations are used for thermocouple thermometry, depending on the temperature and other environmental conditions. The most common combinations for use at cryogenic temperatures are given in Table 61.1. Designations for single-leg materials are:

Constantan: EN or TN, nominally 55 wt.% Cu and 45 wt.% Ni

Chromel (Trademark of Concept Alloys, Inc.): EP or KP, nominally 90 wt.% Ni and 10 wt.% Cr

Alumel (Trademark of Concept Alloys, Inc.): KN, nominally 95 wt.% Ni, 2 wt.% Al, 2 wt.% Mn, and 1 wt.% Si

Figure 61.6 shows the Seebeck coefficient S_{AB} (sensitivity) of common cryogenic thermocouple pairs. The integral of S_{AB} gives the emf of the pair with the junction held at 0 K, as shown in Figure 61.7. In practice, the reference junction is usually held at 0°C, so thermocouple tables are usually given with respect to a 0°C reference. The curves in Figure 61.7 with a 0 K reference are converted to a 0°C reference simply by subtracting the voltage at 0°C from the curves. Figure 61.8a shows a schematic of the theoretical wiring arrangement with the reference junction at 0 K and the voltmeter at a temperature T. Figure 61.8b shows the wiring arrangement for the typical scheme of a 0°C reference temperature.

Thermocouples are inherently a sensor for measuring temperature differences, so they are often used to measure small temperature differences. Figure 61.9 shows a schematic for such measurements. In this arrangement most of the temperature difference from some low temperature to ambient is spanned by identical materials on both legs, so no additional voltage is developed over most of the temperature gradient, and both legs can come from the same spool to ensure nearly identical Seebeck coefficients. The reference junctions can be at any temperature, but they must be at the same temperature. Often they are anchored at the low temperature T by potting both of them in a copper piece. With such an arrangement the leads extending to room temperature can be copper or phosphor bronze, which have low values of absolute Seebeck coefficients $S(T)$, so any material variation between the legs has only a minimal spurious

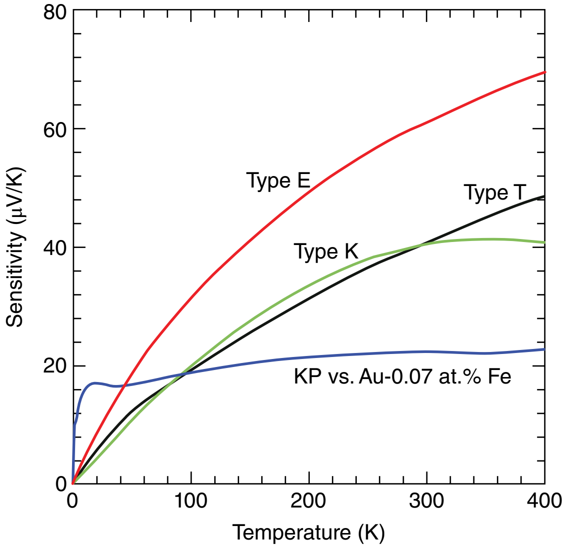


FIGURE 61.6 Relative Seebeck coefficient or sensitivity of common thermocouple types useful for cryogenic temperatures.

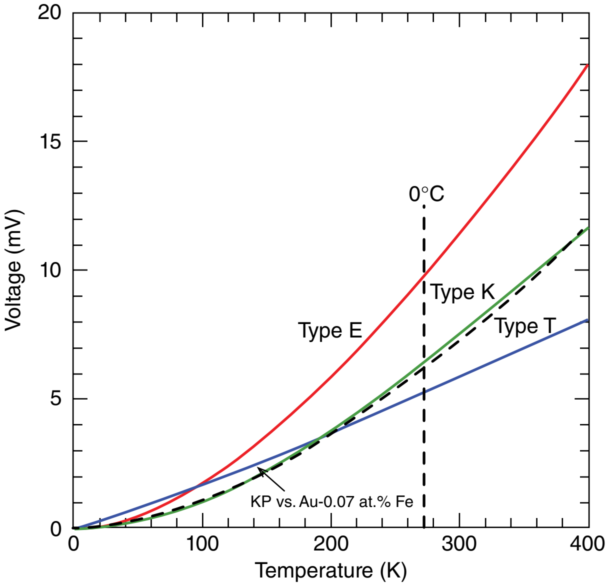


FIGURE 61.7 Voltage output of common cryogenic thermocouples with a 0 K reference temperature. The voltage at 0°C should be subtracted from these readings to convert to a 0°C reference temperature.

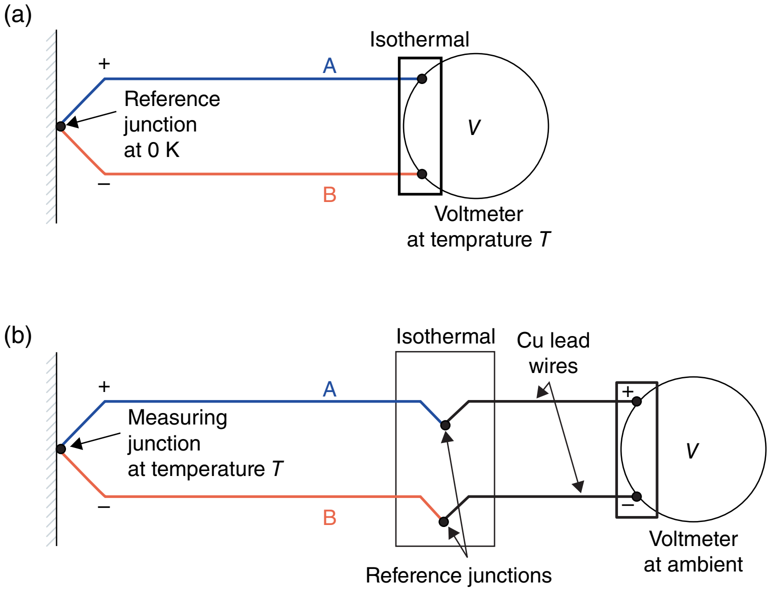


FIGURE 61.8 Schematic of a thermocouple measurement system with (a) 0 K reference temperature and (b) some other reference temperature.

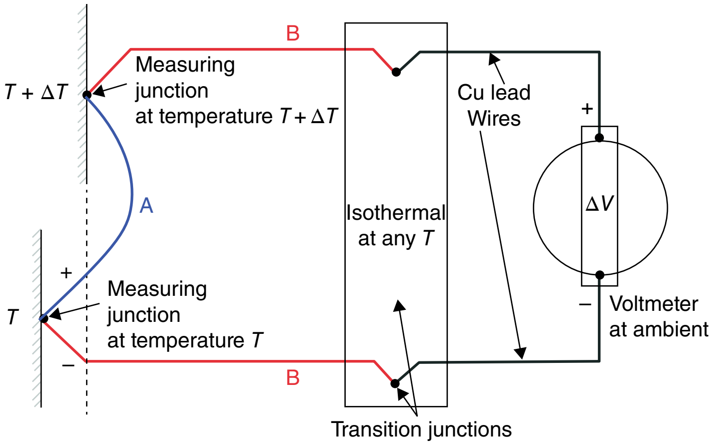


FIGURE 61.9 Schematic of a thermocouple measurement system for measuring small temperature differences.

voltage generation. The reference junctions could also be at the voltmeter connections, but ensuring isothermal conditions there is more difficult and the two leads extending from low T to ambient are thermocouple materials that may have high $S(T)$ and susceptible to larger variations due to material variations. The emf measured at the voltmeter due to a small temperature difference ΔT at the temperature T is given by

$$\Delta V = S_{AB}(T)\Delta T, \tag{61.8}$$

where the relative Seebeck coefficient $S_{AB}(T)$ is evaluated at the temperature T. The output voltage can be increased in such a measurement by the use of a thermopile, in which the thermocouple legs are crisscrossed between the two temperatures to form multiple junctions in series. For extremely high resolution, the thermocouple emf can be measured with a superconducting quantum interference device (SQUID) operating at 4.2 K [17]. Voltages of 10^{-13} V are easily resolved with a SQUID system, and their low-temperature operation eliminates most of the spurious thermal emfs in the system. The Au—Fe thermocouples should not be used in magnetic fields, since they are orientation dependent. The type E thermocouple has a small magnetic field dependence and can be used in moderate fields for temperatures above about 40 K.

61.2.2.5 Capacitance Thermometers The variation of capacitance for typical capacitance thermometers is shown in Figure 61.2 with capacitance shown on the right axis. The capacitance can shift equivalent to a kelvin or more upon thermal cycling, but after about an hour at temperature the drift is only a few tenths of a mK at 4.2 K but a few mK at 300 K. Their strong point is that they are very insensitive to magnetic fields. Their change in a magnetic field is less than 0.05% of the reading over the entire temperature range from 2 to 300 K. Thus, capacitance thermometers are best suited for temperature control during the application of magnetic fields, but a different thermometer must be used to measure the temperature in zero field if that temperature must be known with an uncertainty less than about 1 K.

61.2.3 Thermometer Use and Comparisons

The selection of an appropriate thermometer for cryogenic applications makes use of a comparison of such factors as accuracy, reproducibility, temperature range, temperature resolution, size, cost, magnetic field effect, measuring instrument, and other factors. Accuracy involves how close the true thermodynamic temperature is measured, of which the ITS-90 temperature scale is the latest approximation to such a temperature. How close any thermometer follows the ITS-90 temperature scale depends on the uncertainty of calibration, the temperature resolution of the thermometer and temperature measurement system, and the reproducibility of the thermometer. Typical calibration uncertainties for commercial thermometers range from about ±4 mK at 4 K, ±10 mK at 80 K, and ±25 mK at 300 K. Total thermometer uncertainties must also take into account reproducibility, of which typical values for the various thermometer types are given in Table 61.2. Individual thermometer calibrations over a wide temperature range ensure the lowest uncertainty, but the cost can be quite high. Much lower cost can be achieved by using interchangeable thermometers, such as platinum resistance thermometers, Si diodes, and thermocouples. Their interchangeability is indicated in Table 61.2. Figure 61.1 showed the construction details and size of typical SPRTs. The diameter and length of about 5.8 mm by 56 mm makes them too large for most experimental work. Most industrial thermometers are much smaller and are available in a variety of configurations, as shown in Figure 61.10. For example, wire wound PRTs conforming to IEC 60751 down to 70 K are available in capsules as small as 1.8 mm in diameter and 5 mm long. Thin film platinum RTDs are available in sizes as small as 2 mm × 2 mm × 1 mm thick that conform to IEC 60751 class B down to −50°C and cost about US$1 each in quantities of 100. Quantitative results on the magnetic field effects of many types of thermometers are given by Sample and Rubin [18] and by Rubin et al. [19].

61.2.3.1 Temperature Resolution and Sensitivity The relative temperature resolution is given by

$$\frac{\Delta T}{T} = \frac{\left(\Delta V / V\right)}{S_{\mathrm{d}}}, \tag{61.9}$$

where ΔV is the voltage resolution of the measurement system, V is the voltage, and S_{d} is the dimensionless sensitivity of the thermometer. The expression also applies to a resistor or a capacitor by replacing V with either R or C. For sensors with high-voltage outputs, such as volt-level signals with diode thermometers, a 5½ digit voltmeter can

TABLE 61.2 Characteristics of Various Commercial Thermometers

Sensor	Excitation	Useful Range	Interchangeability	Reproducibility	Long-Term Drift	B Field $\Delta T/T$ in 10 T
Pt (PRT) Pt100	1 mA	70–800 K <70 K cal.	0.5–0.1 K (A) 1–0.25 K (B)	5 mK	10 mK	1% at 77 K 0.1% at 0°C
Rh—Fe	1 mA	0.65–500 K	Poor	0.2 mK	0.2 mK	Poor
Cernox™	10 mV	0.1–420 K	Poor	0.02% of T	25 mK	<0.5%
Ge	1–3 mV	0.05–100 K	Poor	0.5 mK	1–10 mK	Poor
Carbon-glass	1–3 mV	1–325 K	Poor	1 mK@4K	4 mK	<5%
Carbon	1–3 mV	0.1–300 K	5% of T	0.1% of T	8 mK	<5%
Ru—O	10 mV	0.01–40 K	5–10% of T	15 mK	40 mK	<1%
Thermistor			Poor			Good
Si diode	10 µA	1.4–500 K	0.25 K (A) 0.5 K (B)	5–20 mK	10–40 mK	Poor
Thermocouple	—	1.2–1500 K	1 K	20 mK	50 mK	<5%
Capacitance	5 V at 5 kHz	1.4–290 K	Poor	0.3 K	1 K	<0.05%

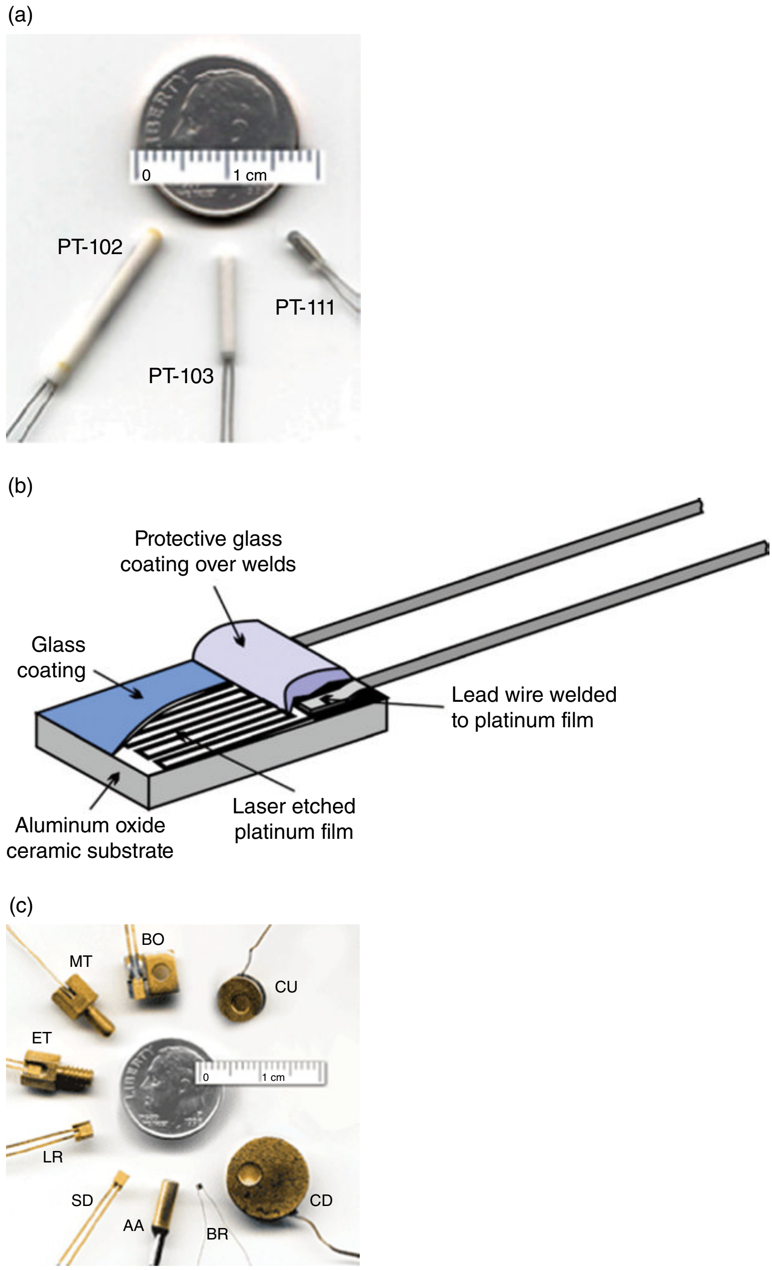

(a)
0
1 cm
PT-102
PT-111
PT-103
(b)
Protective glass coating over welds
Glass coating
Lead wire welded to platinum film
Aluminum oxide ceramic substrate
Laser etched platinum film
(c)
BO
MT
CU
ET
0
1 cm
LR
SD
AA
BR
CD

FIGURE 61.10 (a) Photo of several typical industrial wire wound platinum resistance thermometers. Source: Reproduced with permission of Lake Shore Cryotronics. (b) Drawing of a platinum film resistance thermometer. Typical dimensions are 2 mm × 2 mm. (c) Photo of various types of packages available for many resistance and diode thermometers.

provide a voltage resolution of about 2×10^{-6}. For thermocouples the maximum output voltage may only be about 10 mV, so a voltmeter with 1 μV resolution yields only $\Delta V/V = 1\times10^{-4}$. The dimensionless sensitivity of the thermometer is given by

$$S_{\mathrm{d}} = \left|\frac{T}{V}\frac{dV}{dT}\right|, \tag{61.10}$$

where the voltage V can be replaced with R or C. A comparison of the dimensionless sensitivity for the various thermometer types discussed earlier is shown in Figure 61.11. The parameter O in this figure represents V, R, or C. The dimensionless sensitivity for thermocouples is calculated using the 0 K reference temperature, which can be misleading. For thermocouples it is best to determine the temperature resolution by the equation

$$\Delta T = \frac{\Delta V}{S_{\mathrm{AB}}}, \tag{61.11}$$

where S_{AB} is determined from the sensitivity shown in Figure 61.6. For example, at 100 K the 30 μV/K sensitivity for type E thermocouples results in a temperature

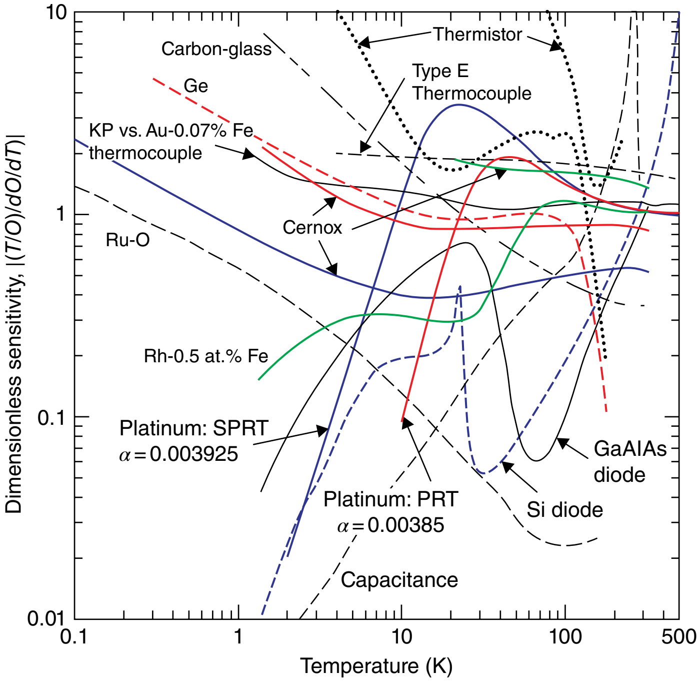


FIGURE 61.11 Dimensionless sensitivity of many types of thermometers. Sensitivity for thermocouples is shown with voltage for 0 K reference temperature. The parameter O can be R, V, or C.

resolution of 0.03 K for a 1 μV voltage resolution. However, the uncertainty of the absolute temperature is best given by the standard error of 1.5 K shown in Table 61.1 that takes into account interchangeability, reproducibility, and long-term drift.

61.2.3.2 Thermometer Electrical Excitation Except for the case of thermocouples, all other thermometers discussed here require some type of electrical excitation. For most accurate measurements, a four-lead measurement should be used right to the thermometer element: two leads for current and two leads for voltage. Such an arrangement eliminates the error caused by voltage drop in a current-carrying lead. The standard calibration curves given for diodes are for an excitation of 10 μA, as listed in Table 61.2. For resistance thermometers, any current or voltage can be used, provided it does not cause self-heating of the thermometer. Thermometers with positive temperature coefficients, such as the metallic resistance thermometers, are best excited with a constant current over a wide temperature range to allow for reduced power at lower temperatures. However, thermometers with negative temperature coefficients, such as semiconductor or semiconductor-like thermometers, are best excited with constant voltage. These typical currents and voltages are listed in Table 61.2. In practice, the maximum excitation can be determined by increasing the current or voltage until a change in resistance is detected. To prevent self-heating errors, the power dissipation should usually be less than about 1–10 μW at 300 K and decrease to about 0.01–0.1 μW at 4.2 K. The excitations listed in Table 61.2 usually yield these power levels. The power dissipation in the diode thermometers is in the range of 20–50 μW at 4.2 K, which will result in self-heating if special care is not taken to thermally anchor them very well.

Resistance thermometers can be excited with either alternating current (AC) or direct current (DC). DC is most commonly used because of the availability of lower-cost instrumentation. However for the most precision work, AC is commonly used to allow for noise reduction with lock-in amplifiers. AC bridge networks with null detectors provide the ultimate in resistance resolution and have the added advantage of eliminating thermal EMFs caused by temperature gradients. A resistance resolution of 1 μΩ is possible, which gives a temperature resolution of 1 μK in a 25 Ω SPRT [20]. A low frequency of about 30 Hz is commonly used to avoid any problems with reactance in the circuit. Precision work using DC excitation must reverse the current and take the average of the resistance from the two current directions to eliminate thermal EMFs.

61.2.3.3 Thermal Anchoring of Thermometers and Leads Any of the thermometers discussed here measure the temperature of the thermometer, so the uncertainties apply only to the thermometer itself and not to the sample to be measured. Because of heat leak through the electrical leads and the self-heating due to thermometer excitation, the thermometer temperature can be higher than the sample unless special care is taken to ensure good thermal contact between thermometer and sample and to minimize heat conduction through the leads. Thermometer manufacturers are careful to provide good thermal contact between the thermometer element and any package. Examples of packages are shown in Figure 61.10. Canister packages must be inserted into a close fitting hole in a high thermal conductivity block or spool. Use of a thermal grease or epoxy ensures good thermal contact between the canister and the block, but the hole must not be blind to allow for air escape and easy thermometer disassembly by pushing on the end rather than pulling on the electrical leads. Figure 61.10 also shows many packages in which the thermometer canister has already been mounted by the manufacturer in a gold-plated copper spool. The gold plating minimizes radiation heating of the spool and provides a high thermal contact conductance when bolted to the sample. For contact areas greater than about 1 cm^2, the use of thermal grease can provide improved thermal contact.

The second consideration in thermal anchoring thermometers is the thermal anchoring of the electrical leads separately from the thermometer. The Wiedemann–Franz law can be used to provide a good rule of thumb to relate heat leak in electrical leads to the electrical resistance of the lead between room and low temperature. For a low temperature of 77 K and a high temperature of 300 K, the Wiedemann–Franz law gives

$$\dot{Q}R \approx 1\,\mathrm{mW}\cdot\Omega, \tag{61.12}$$

where $\dot{Q}$ is the heat flow and R is the electrical resistance. For a two-lead measurement circuit, the lead resistance must be much smaller than that of the thermometer. For a 100 Ω Pt thermometer, the resistance is only about 10 Ω at 80 K, so even a 0.1 Ω resistance in each lead gives a 2% error in resistance and a heat leak of 20 mW in both leads according to Equation 61.12. A typical thermal conductance between a thermometer spool and the sample may be about 1 W/K, which then leads to a temperature difference of 20 mK between the spool and the sample. Another temperature difference will occur between the thermometer and the spool. The problem is not so serious with higher-resistance thermometers or with diodes that have a resistance of about $10^5\,\Omega$. To eliminate the lead conduction problem, electrical leads should be thermally anchored to the sample independently of the thermometer. Often the leads are wrapped around a separate spool that is then bolted to the sample near the thermometer. Typical thermal tempering lengths for various wire sizes and materials are shown in Table 61.3.

Thermocouples are particularly difficult to thermally anchor to a cryogenic sample because of the small tip and the fact that most cryogenic samples will be in a vacuum. A portion of the thermocouple wire can be thermally anchored to the sample by thermal grease, epoxy, or a bolted spool as discussed for resistance thermometers. The tip can also be greased to the surface but care must be taken to ensure electrical isolation unless the measuring instrument has inputs isolated from ground. If that is the case, soldering the tip to the sample provides excellent thermal contact.

TABLE 61.3 Typical Wire Tempering Lengths for Thermometer Leads of Various Sizes and Materials[a]

			Tempering Length for Various Wire Gages (cm)			
Material	T_h (K)	T_c (K)	0.080 mm (#40 AWG)	0.125 mm (#36 AWG)	0.200 mm (#32 AWG)	0.500 mm (#24 AWG)
Copper	300	80	1.9	3.3	5.7	16
	300	4	8.0	13.8	23.3	68.8
Phosphor	300	80	0.4	0.6	1.1	3.2
bronze	300	4	0.4	0.7	1.3	3.8
Manganin	300	80	0.2	0.4	0.4	2.1
	300	4	0.2	0.4	0.7	2.0
Stainless steel	300	80	0.2	0.3	0.6	1.7
304	300	4	0.2	0.3	0.5	1.4

[a] Data from Ekin [7].

61.2.4 Dynamic Temperature Measurements

Most of the thermometers discussed previously have thermal time constants of several seconds. With some exceptions they are not designed for dynamic temperature measurements. Fast response times are achieved with thermometers that have low heat capacity (low mass or low specific heat) and good thermal contact between the sensing element and the object to be measured. The object to be measured can either be a solid object, in which case it can provide support for the thermometer, or a fluid, in which case the thermometer must be supported by some nearby solid and thermally insulated from it. The second case occurs, for example, in the measurement of the instantaneous temperature of the helium working fluid in regenerative cryocoolers. Typical operating frequencies may range from 1 to 60 Hz. To make such measurements at 60 Hz with negligible phase shifts requires a thermal time constant of less than about 300 μs. In the measurement of instantaneous gas temperature, the dominant thermal resistance is often between the sensing element and the gas. As a result, the measurement of dynamic gas temperatures is usually more difficult to measure than that of liquids or solids.

By their nature, thermocouples have a small mass and a potential for fast response times. The use of small diameter thermocouple wire at cryogenic temperatures can often yield response times of a few tenths of seconds. Faster response times are obtained by using commercially available thin foil thermocouples. Foil thicknesses down to 5 μm are available, but considerable care is required in handling the unsupported foil. For measurements of the surface temperatures of solids, a thermocouple film of 3–6 μm thickness can be sputtered on the surface [21]. The internal response time of such a thin film will be about 1 μs or less, but with the thermal resistance at the interface, the response to temperature changes at the surface may be considerably longer. The response time of unsupported 5 μm thick type E thermocouple foil to oscillating helium gas temperatures was measured at NIST and found to be about 10 ms at 80 K.

Thin film platinum or carbon thermometers can also be used for dynamic temperature measurements and have response times comparable to those of the thin film thermocouples. Louie and Steward [22] used an unsupported 4 μm thick Pt foil in the measurement of transient heat transfer to liquid hydrogen for response times down to 10 μs. Carbon films on a quartz substrate have also been used for transient heat transfer experiments to liquid helium [23] and to liquid hydrogen [22]. Giarratano et al. [24] measured a response time of about 50 μs at 77 K for an 18 nm thick platinum film on a quartz substrate.

For high-speed temperature measurements in the range of 1–20 K, a silicon-on-sapphire (SOS) thermometer is the fastest ever reported. The response time of these thermometers in both liquid and gaseous helium was found to be about 300 ns. [25]. These thermometers are made with a 1 μm thick silicon layer on a 0.13 mm thick sapphire substrate. The silicon was ion implanted with phosphorus to give a resistance versus temperature curve similar to germanium resistance thermometers. These thermometers were used to study the temperature oscillations that occur in thermoacoustic oscillations inside small tubes closed at the room temperature end and open to a dewar of liquid helium at the other end [25]. The same thermometers were used for the measurement of instantaneous temperature of the helium gas inside a Stirling cryocooler next to the regenerator. Figure 61.12 shows how these thermometers were suspended from a fiberglass-epoxy support to measure the gas temperatures in a

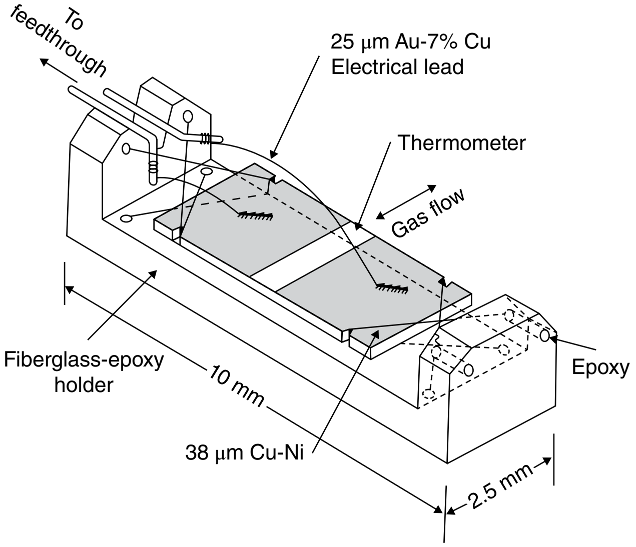


FIGURE 61.12 Drawing of a silicon-on-sapphire (SOS) thermometer for dynamic temperature measurements of flowing fluids at low temperature.

Stirling refrigerator at temperatures of about 10 K. The 38 μm diameter Cu—Ni support wires minimize the thermal contact between the thermometer and the support.

The thermal response times of several carbon, germanium, and diode thermometers at cryogenic temperatures were measured at NIST [26]. The response times reported there for the SOS thermometer were superseded by the measurements of Louie et al. [25]. At 4 K, 1/8 W carbon resistors showed response times as fast as 6 ms. Commercial Si diode thermometers in their basic sensor package have a response time at 4 K of about 10 ms as reported by their manufacturer. A response time of about 6 μs at 4 K has been reported for a miniature silicon diode thermometer [27]. The thin film metal oxy-nitride resistance thermometers have a reported time constant at 4 K of 1.5 ms as an unpackaged chip. When packaged inside a copper canister, the response time increases to about 0.4 s at 4 K.

For high-speed measurements of oscillating gas temperatures, platinum wire of about 5 μm gives a response time in still helium gas at 300 K of about 300 μs. Unfortunately, platinum wire of this small diameter is not strong enough to withstand oscillating mass flows that usually accompany the oscillating temperatures in gas. Instead, 3.8 μm diameter tungsten wire can be used, which is much stronger and has a measured response time of about 260 μs at 300 K in still helium gas [28, 29]. Thermometers made with 2 mm long segments of the 3.8 μm diameter tungsten wire have been used in oscillating gas flows at cryogenic temperatures for hours with no breakage, providing the gas is very clean. Such wire is commercially available from manufacturers of hot-wire anemometers. A description of a small demountable probe using this wire is given in the section on flow. The resistance of the tungsten wire has a linear temperature dependence down to about 77 K. Above that temperature its dimensionless sensitivity is 1.032, which is comparable to that of platinum.

61.3 STRAIN

The measurement of strain has many applications beyond its direct measurement. For example, strain gages are commonly used to measure force and pressure. In turn, pressure transducers are often used in the measurement of flow. Most measurements of strain, including those at cryogenic temperatures, are performed with bonded resistance strain gages. We restrict our discussion to these devices. An excellent review of resistance strain gages is given by Hannah and Reed [30], although their book emphasizes temperatures of 300 K and above. The principles are no different at cryogenic temperatures, but a proper materials selection is important. The resistance strain gage is an element whose resistance is a function of the applied strain. The relative resistance change can be expressed as

$$\frac{\Delta R}{R} = F_{\mathrm{S}}\left(\frac{\Delta L}{L}\right), \tag{61.13}$$

where F_S is the gage factor or strain sensitivity factor and $\Delta L/L$ is the strain. Typical gage factors are about 2 for most of the commonly used metallic alloys. Their gage factors are nearly independent of strain for strain levels up to about ±2000 microstrain (2000×10^{-6}). The metal alloy gages are usable at cryogenic temperatures for strain levels up to about 1.5%. Semiconductors can have gage factors of about 100 or more, but they are very temperature sensitive.

61.3.1 Metal Alloy Strain Gages

Most measurements of strain at temperatures from 4 to 300 K are made with a nickel-chromium alloy or a modified nickel-chromium alloy (73%Ni + 20%Cr + Al + Fe) either in the form of a wire or, more recently, in the form of photoetched foil. The copper-nickel alloy most often used at ambient temperatures has larger temperature and magnetic field effects and is seldom used for cryogenic temperatures. Typical resistance values for these gages are in the range of 60–1000 Ω, although 120 and 350 Ω gages are most commonly used at cryogenic temperatures. The alloy grid is bonded to a carrier matrix (backing) and usually has a geometry like that shown in Figure 61.13. Gage lengths vary from about 0.20 to 100 mm. The large areas in the region of the bends reduce the effects of transverse strain. Typical ratios of gage factors between transverse and longitudinal strains are only a few percent and have negligible effect on most measurements, unless high accuracy is desired. Many other gage geometries are available from gage manufacturers for use in various applications. The most common backing material for use at cryogenic temperatures is glass fiber-reinforced epoxy-phenolic. The polyimide backing commonly used for large strains at room temperature or above is seldom used for cryogenic temperatures. Most gages have a top layer of insulation bonded over the grid and backing. This top layer is known as the overlay or encapsulating layer. It is particularly important for use in liquid cryogens to prevent the formation of bubbles at the surface of the metal caused by self-heating. These bubbles can lead to rapid localized temperature rises, which cause considerable

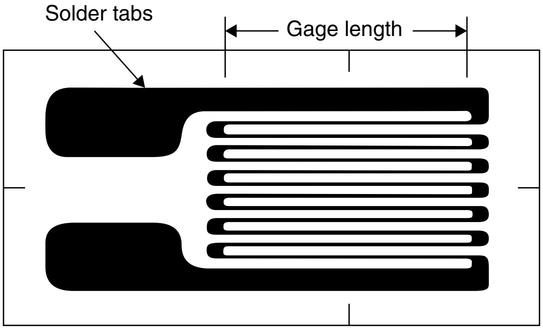


FIGURE 61.13 Geometry of a metal foil strain gage.

noise in the signal. For cryogenic use, the gage is usually bonded to the test specimen with an epoxy recommended by the manufacturer of the gage.

61.3.2 Temperature Effects

Various temperature effects can have a significant impact on the measurement of strain at cryogenic temperatures. There are three different temperature effects that need to be considered. The first is the effect of temperature on the gage factor. The gage factor for the modified nickel-chromium alloy varies linearly with temperature in such a way that the gage factor at 4 K is about 4–5% higher than the gage factor at 297 K [31]. For the copper-nickel alloy, the gage factor decreases by about 3% when cooled to 4 K from 297 K.

The second temperature effect is caused by the change in resistivity with temperature, or the temperature coefficient of resistivity (TCR). With the cryogenic alloys discussed here, the change in resistance of a strain gage when cooled from ambient temperature to 4 K is usually less than 5%. A 5% resistance change is the same change that would occur with a strain of 2.5% in a gage alloy with a gage factor of 2. The resistance change caused only by a temperature change is referred to as apparent strain or thermal output.

The third temperature effect is caused by the strain induced in the gage due to the difference in the thermal expansion between the specimen and the gage. This difference is a function of the specimen material and the gage material. With most metals, the difference in thermal contraction from ambient to 4 K is only a few tenths of a percent. In practice, the second and third temperature effects are combined into one apparent strain (A.S.) or thermal output that is a function only of temperature and the difference in thermal expansion between the gage and the specimen. Strain gage manufacturers can minimize this thermal output for a particular specimen thermal expansion by a proper heat treatment of the gage alloy. That technique is known as self-temperature compensation (STC). Figure 61.14 shows how this apparent strain varies with temperature for the modified nickel-chromium gage bonded to various materials and with the curves normalized at 280 K [32]. Shown for comparison is the dashed curve for a copper-nickel alloy gage bonded to a 304L stainless steel specimen. When the modified nickel-chromium curves are normalized at 4.2 K, all the curves agree with each other up to 20 K and reach a minimum of -700×10^{-6} at a temperature of 15 K.

In order to correct for the thermal output or apparent strain, the temperature of the specimen must be measured. Often it is submersed in liquid cryogens, in which case the barometric pressure defines the bath temperature. When the specimen is not in a liquid bath, its temperature must be measured with a thermometer in thermal contact with a strain-free area of the specimen that is also in good thermal equilibrium with the portion subjected to the strain. Alternatively, the thermal output can be reduced to zero by using a temperature compensating circuit. This circuit is a Wheatstone bridge with two identical resistance strain gages used for the active and

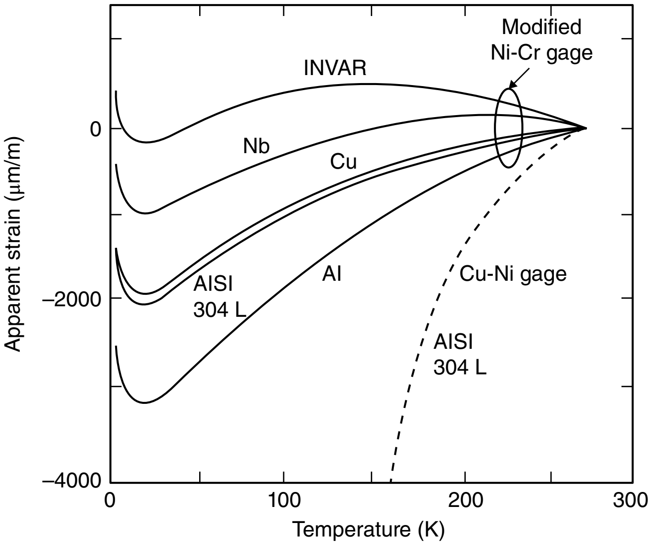


FIGURE 61.14 Apparent strain caused by temperature change from 280 K for Ni—Cr gages and Cu—Ni gages on various test materials.

the compensating (reference) arm. The compensating gage is mounted on a strain-free region of the sample that is at the same temperature as the strained region. If the specimen can remain strain-free until the test temperature is reached, then the thermal output can simply be canceled out by adjusting the reference resistor in the Wheatstone bridge circuit.

61.3.3 Magnetic Field Effects

A magnetic field can cause a change in resistance, which is known as the magnetoresistance effect. This resistance change leads to a strain error. The magnetic field can also affect the gage factor. Walstrom [33] measured the effect of magnetic fields up to 6 T on several strain gage alloys. For the nickel-chromium alloy, a strain error of $+160 \times 10^{-6}$ was found at 4.2 K in a magnetic field of 6 T. The error varied approximately quadratically with magnetic field. There was no detectable magnetoresistance strain error at 296 or 77 K with this alloy. The copper-nickel alloy gage showed much larger magnetoresistance effects. It showed a strain error of -250×10^{-6} at 296 K in a field of 6 T. Presumably, the error would be much larger at cryogenic temperatures, but it was not measured. These results indicate that copper-nickel gages should not be used in magnetic fields. The results of Walstrom also showed that the magnetoresistance effect was independent of field direction and independent of strain for strains up to 10^{-3}. There was about a 1% effect on the gage factor for fields above 3 T with the field perpendicular to the gage surface. No effect on the gage factor was seen with other field orientations.

Freynik et al. [34] extended the magnetoresistance measurements on nickel-chromium alloy gages up to magnetic fields of 12 T at 4.2 K. The strain error was found to be $+400 \times 10^{-6}$ at 4.2 K in a magnetic field of 12 T. Their data were in good agreement with those of Walstrom [33] for fields of 6 T and below.

61.3.4 Measurement System

The measurement of strain with resistance strain gages entails the detection of resistance changes that are often less than 1% of the resistance. Such small changes are best measured with a Wheatstone bridge adjusted for zero output at some known reference condition. An amplifier is used on the output voltage to significantly increase the resolution. Either DC or AC bridge excitation can be used, although most commercial strain gage systems use DC voltage. The use of AC excitation eliminates any thermal EMFs generated in the circuit and allows the use of lock-in amplifiers for enhanced signal-to-noise ratio. Any change in resistance in the electrical leads to the gage can be compensated by a three-wire bridge. As a result, most static measurements of strain should be made with three-wire bridges. The bridge excitation voltage must be kept sufficiently low to prevent self-heating of the gage. Usually, the excitation voltage is determined experimentally. For cryogenic applications, a bridge excitation of 2 V with a 350 Ω gage is typical for use in a liquid cryogen. In vacuum, voltage levels down to 0.5 V may be necessary to prevent self-heating [35].

61.3.5 Dynamic Measurements

The intrinsic frequency response of the resistance strain gage should be in the tens or hundreds of kilohertz range. The adhesive joint will lower this frequency response some, but the resulting frequency response should be well above the maximum frequency used in most measurements. Few measurements are made at frequencies above 100 Hz because of the limitations in the equipment used to apply the dynamic strain [36]. In measurements of dynamic strain, the output of the Wheatstone bridge circuit is coupled to the amplifier through a capacitor (AC coupled) to eliminate any DC component. The AC coupling eliminates all effects associated with slow temperature changes of the specimen. Dynamic strain measurements are typically associated with fatigue measurements. The fatigue life of a properly selected gage can be as high as 10^8 cycles at a strain level of $\pm 2000 \times 10^{-6}$.

61.4 PRESSURE

The easiest and most common method for measuring pressure at cryogenic temperatures is to connect a capillary line between the desired pressure location and a pressure transducer at ambient temperature. In this case, a conventional pressure transducer can be used. This method is limited to measurements of static pressure because of the low-frequency response of the capillary line. This method also has limitations in the low-pressure range because of the thermomolecular pressure correction that occurs in a temperature gradient [37]. The correction becomes particularly large for pressures below about 130 Pa. In this region, the pressure at the warm end of the capillary is higher than that at the cold end. Most commercial pressure transducers are designed for use at ambient temperature and cannot be used at cryogenic temperatures. We discuss some of the exceptions here. There are four types of pressure sensors or transducers that are commonly used at cryogenic temperatures, (i) capacitance, (ii) variable reluctance, (iii) strain gage or piezoresistive, and (iv) piezoelectric.

61.4.1 Capacitance Pressure Sensors

Variable capacitance pressure sensors are one of the most common types of pressure sensors used for precision work at cryogenic temperatures. However, we are not aware of any commercial units that have been used at these temperatures. An excellent review of capacitance pressure sensors used at cryogenic temperatures is given by Jacobs [38]. The sensor consists of a thin, stretched membrane, or for high pressures, a machined diaphragm, which deflects with pressure. It forms one electrode of the capacitor. The other electrode is formed by a stationary disk. The two electrodes must be electrically insulated from each other. The diaphragm and other parts of the sensor are usually of BeCu. A well-constructed capacitance sensor has less than a 5% change in sensitivity when cooled from 300 to 4 K. The capacitance of these sensors is typically in the range of 20–50 pF and can be measured with a capacitance bridge. For better accuracy, a three-lead bridge should be used to eliminate the effects of temperature-dependent lead capacitance. A frequency-to-voltage converter can also be used to measure the capacitance (pressure). Sensitivities of one part in 10^8 have been achieved with some capacitance pressure sensors, although a resolution of one part in 10^5 would be more common with inexpensive electronics. A disadvantage of the capacitance sensors is the requirement for coaxial cables between the sensor and the electronics.

61.4.2 Variable Reluctance Pressure Sensors

The variable reluctance pressure sensor utilizes a magnetically permeable stainless steel diaphragm. Deflection of the diaphragm is sensed by a pair of inductance coils on each side of the diaphragm, as shown in Figure 61.15. The magnetic reluctance of each of the circuits is a function of the gap between the diaphragm and the “E” core. A change in the reluctance on each side of the diaphragm changes the inductance of each of the coils. These two coils are connected in a bridge circuit, with the coils forming one-half of the bridge and a center tapped transformer forming the other half of the four-arm bridge. An AC signal of 3–5 kHz is commonly used in the bridge circuit. A carrier demodulator amplifies the output signal and converts it to a DC voltage

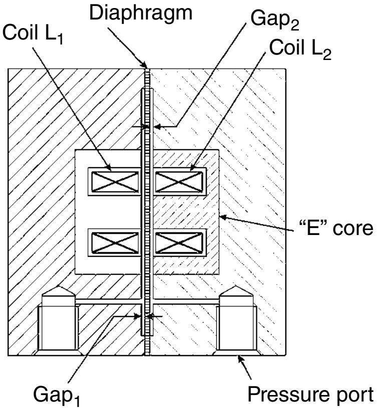


FIGURE 61.15 Cross section of a variable reluctance pressure transducer.

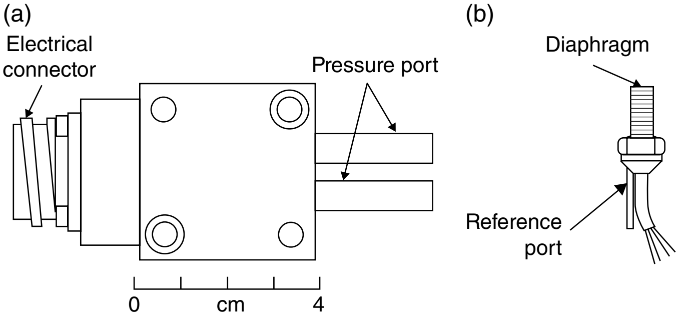


FIGURE 61.16 Two types of pressure transducers adaptable to cryogenic temperatures. (a) Variable reluctance and (b) piezoresistive.

proportional to the pressure. Because the coils are made of copper wire with very low resistance, the power dissipation in the transducer is quite small and acceptable for most cryogenic use. For temperatures of 77 K and below, the power dissipation is about 1 mW with conventional electronics using a 5 V excitation.

Commercial variable reluctance transducers are made with either all-welded construction or with rubber O-ring seals between the diaphragm and the transducer case. For cryogenic applications, the all-welded models must be chosen. Such transducers perform satisfactorily at temperatures down to 2 K. The differential pressure models with relatively low-pressure ranges (below about 100 kPa) remain linear at these low temperatures. Their sensitivity decreases by 12–13% from 300 to 77 K but remains unchanged from 77 to 4 K. They have a natural frequency of about 20 kHz and an internal volume of about 0.07 cm^3 on each side of the diaphragm. Their external dimensions are relatively large, as shown in Figure 61.16a. Unfortunately, we have experienced some large deviations from linearity at 4 K with the high-pressure (5 MPa) models.

The calibration and vibration testing of many of these variable reluctance pressure transducers was reported by Kashani et al. [39] for use at temperatures down to 2.1 K.

The full-scale range of these transducers varied from 0.86 to 138 kPa. They found that the sensitivity of a 0.86 kPa transducer increased by about 4% from 300 to 4.2 K and then decreased by about 2% at 2.1 K. A 138 kPa transducer showed a 6% decrease in sensitivity when cooled from 300 to 4.2 K. There was no significant change in their responses after they were overpressurized. After temperature cycling the transducers two or three times, they exhibited linearity and repeatability to within ±1% of full scale in liquid helium. Daney [40] reported that a 5.5 kPa transducer showed a sensitivity change of only 0.2% between 300 and 4 K. Because of the different values of sensitivity changes reported by different authors on different transducers, it is important that each transducer be calibrated at cryogenic temperatures. Once calibrated, it should be repeatable to within ±1%. The zero reading also shifts with temperature, but the transducer is normally rezeroed electronically after it has reached the desired temperature.

61.4.3 Piezoresistive Pressure Sensors

The piezoresistive pressure sensors use a strain gage to measure the deflection of a diaphragm subjected to a pressure on one side. They are the most common commercial pressure sensor. Some of them can be used at cryogenic temperatures if the materials have been properly selected. One type of sensor uses a metal diaphragm, typically stainless steel, with a thin film strain gage deposited on the diaphragm, or a metal foil stain gage bonded to the diaphragm. Various strain gage materials have been used, but the nickel-chromium alloy discussed in the previous section on strain is particularly good for use at cryogenic temperatures. Cerutti et al. [41] reported on tests with several commercial strain gage pressure sensors at temperatures down to 4.2 K and in magnetic fields up to 6T. As is expected, their behavior is similar to that discussed for strain gages. For example, the calibration factor changes by about 5% for the nickel-chromium gages when cooled from 300 to 4 K. Thermal zero shifts of these same sensors were less than 3% of full scale (F.S.) for a temperature change from 293 to 4.2 K. The apparent pressure error at full scale due to a temperature change from 293 to 4 K was about 4% F.S. The combined nonlinearity and hysteresis was about ±0.2% F.S. at a fixed temperature of 4.2 K. When the sensor was cycled between 293 and 4.2 K, the nonlinearity and hysteresis increased to ±2.3% F.S. A magnetic field caused a maximum signal variation at 5 K of about 0.5% F.S., which occurred at low pressures and a field of 1.6T.

Another type of piezoresistive pressure sensor uses a semiconductor strain gage to measure the deflection of a diaphragm. Doped silicon is nearly always used for these gages because the processing of silicon is a very well-established technology. In some cases, the diaphragm is also made of silicon with the strain gage grid diffused into the diaphragm. The silicon diaphragm is usually bonded to a stainless steel case by epoxy. The silicon diaphragm is often etched (micromachined) into a particular geometry to concentrate the stress at the region where the strain gage is located. This construction has the advantage of nearly eliminating all mechanical hysteresis in the sensor. The epoxy bond has the disadvantage of limiting the maximum negative pressure differential to about 1 MPa before the epoxy bond will crack. The epoxy bond can occasionally develop leaks after rapid cooldown. The sensors that employ a stainless steel diaphragm welded to the case do not suffer from the limited negative pressure differential, and they are less likely to develop small leaks past the diaphragm. The silicon strain gage elements are bonded directly to the back side of the diaphragm to maintain a high-frequency response. The bond does have the disadvantage of slightly higher nonlinearity and hysteresis (0.5–1%) compared with that of the all-silicon construction (about 0.5%). These silicon pressure sensors, with stainless steel or silicon diaphragms, can be made as small as 1.3 mm diameter. A common configuration available in a wide range of full-scale pressure ratings has a diameter of 3.9 mm at the tip where the diaphragm is located. A 10–32 UNF-2A thread (4.8 mm diameter) allows the sensor to be screwed into a small pressure port. Figure 61.16b shows the geometry of this type of pressure sensor or transducer. A rubber O-ring fits in a groove under the head of these transducers to seal against the pressure port. For use at cryogenic temperatures, the O-ring must be replaced with a Teflon gasket about 0.13 mm thick. The gasket is compressed by a circular tongue on the mating assembly that fits closely within the O-ring groove to prevent extrusion of the Teflon. The geometry of this type of transducer makes it easy to place the diaphragm very close to the location where the pressure is to be measured. Their natural frequency is about 500 kHz, which permits pressure measurements at very high frequencies. These transducers can be used in the differential mode by utilizing the reference port, but the permissible pressure on the backside of the diaphragm is limited to about 1 MPa. We have not used these transducers in the differential mode.

The advantage of the silicon strain gage over the metallic strain gage in these pressure transducers is the greatly enhanced sensitivity. The 5 MPa transducers with silicon strain gages have sensitivities in the range of 3600–8700 mV/(V MPa) compared with 0.65 mV/(V MPa) for a transducer with a metal strain gage. The silicon transducers can be used with much cheaper readout electronics. The disadvantage of the silicon device is that it is much more temperature sensitive. Manufacturers of these transducers build in a temperature compensation circuit as part of the Wheatstone bridge that is good for the region of about 260–360 K. The zero shift and the sensitivity shift within this temperature range are less than 4%.

The sensitivity (V/Pa) and the zero reading of these silicon-based pressure sensors will change considerably when used at cryogenic temperatures. Boyd et al. [42] reported on precision calibration measurements of this type of sensor over the temperature range of 78–300 K. A total of 37 sensors were measured. Their sensitivities increased by a factor of 1.7–1.8 as the temperature was decreased from 278 to 78 K. The zero offset changed by about 1% F.S. over this temperature range. The curves for both the sensitivity and the zero offset indicate that they will continue to change as the temperature is reduced below 78 K, but at a slower rate. A thermal hysteresis of about ±0.1% F.S. was reported after many thermal cycles. An excitation of 1 mA was used with these sensors, which had a bridge resistance of about 5 kΩ. The power dissipation was about 5 mW. Hershberg and Lyngdal [43] have shown that for pressure measurements at about 10 K or below, the power dissipation in the sensor should be less than about 5 mW. The normal power dissipation at 300 K in these transducers using the commercial electronics is about 50 mW.

Measurements made in our laboratory with a 5 MPa sensor containing a silicon diaphragm with a diffused silicon strain gage showed a sensitivity increase of 1.9 between 300 and 76 K with almost no change between 76 and 4 K. Such behavior is consistent with results reported by Clark [44] in which the sensitivity increased by about a factor of 1.8 between 300 and 77 K with little change below that temperature. The large change in sensitivity between 300 and 77 K means that the temperature of the pressure transducer must be measured accurately for use in this temperature range. For lower temperatures, the transducer temperature need not be measured accurately. Walstrom and Maddocks [45] tested a series of rather inexpensive semiconductor pressure sensors in the temperature range of 1.6–4.2 K. These sensors had an initial failure rate of about 20% upon cooldown, but those that survived could be cooled repeatedly. They found that the sensitivity at 4.2 K was about a factor of 2.4–2.6 higher than the room temperature value. They also found that the sensitivity at 1.6 K was about 5% higher than the value at 4.2 K and that the sensitivity at high pressure ($P \approx 100$ kPa) was about 5% less than the sensitivity at low pressure ($P < 10$ kPa).

For dynamic pressure measurements in a gas, the temperature of the gas and the silicon strain gage can vary with time. Some corrections may be necessary when using these transducers for dynamic pressure measurements between 77 and 270 K when the dynamic temperature is large. The piezoresistive pressure sensors are commonly used for pressure measurements in cryocoolers because of their small size and fast response.

61.4.4 Piezoelectric Pressure Sensors

Piezoelectric pressure sensors convert the stress applied to the sensing element (typically quartz crystal) to an electrical charge of the order of picocoulombs. A high-impedance charge amplifier converts the charge to a voltage output that will decay with time when the stress remains constant due to charge leakage through resistance in the output leads. These sensors therefore can only measure pressure changes or dynamic pressures. Some commercial piezoelectric sensors have the charge amplifier built into the sensor package to eliminate the need for special low-noise coaxial cables between the sensor and the charge amplifier at room temperature. The charge amplifier converts the high-impedance charge output from the sensor to a low-impedance voltage output. Any low-noise cable between the sensor and the charge amplifier must have insulation resistances as high as $10^{13}\,\Omega$. Sensors with a built-in charge amplifier are powered with a low-cost 24–27 VDC, 2–20 mA constant current supply. The voltage output is usually ±5 V at full scale. A long coaxial or two-conductor ribbon cable can be used to connect the sensor to the room temperature electronics without signal degradation.

Materials and mounting techniques in many of these sensors are compatible with cryogenic use. Some special models designed specifically for cryogenic operation and calibrated at 77 K are commercially available. A low-frequency limit in the range of 0.5 Hz for a 5% error is typical for these piezoelectric pressure sensors due to charge dissipation in the circuit resistance and capacitance. Resonant frequencies are generally greater than 250 kHz. Resolutions of 2×10^{-5} of full scale are possible. Models with full-scale dynamic pressure from 0.3 to 35 MPa are available. They can be used with static pressures much higher than the full-scale dynamic pressure, which makes them very useful for measurements of small dynamic pressure amplitudes superimposed upon a large static pressure. Typical applications are in the measurement of dynamic pressures in regenerative cryocoolers, such as pulse tube cryocoolers. A typical sensor package is about 38 mm long with a 3/8-24 UNF-2A thread or M6 thread for mounting. A brass or copper gasket is available for a leak-tight seal at cryogenic temperatures, but other models have rubber O-ring seals that cannot be used at low temperatures. Miniature models are also available with a 10–32 (4.8 mm) mounting thread and a total length of about 15 mm.

Piezoelectric pressure sensors are usually calibrated by the manufacturer at room temperature. Such a calibration is more complex than that with other pressure sensors because of the need to use a rapid pressure change or dynamic pressure. Calibrations at 77 K are available for some models. The temperature coefficient of sensitivity is about 0.07%/°C. In-house calibrations are often performed by comparison with piezoresistive pressure sensors, which can be calibrated with static pressure from a room temperature sensor, but are also capable of measuring high-frequency dynamic pressure.

61.5 FLOW

Many types of flowmeters have been used successfully at cryogenic temperatures for measurements of gas or liquid flows. Some are mass flowmeters while others are volumetric flowmeters. The determination of mass flow rate from a volumetric flow measurement requires a measurement of the fluid density. Some volumetric flowmeters contain a densitometer to yield an inferential mass flowmeter. The measurement of mass flow of cryogenic fluids is particularly important for custody transfer of cryogens from tank trucks to the customer. Pipe sizes commonly used for this application vary from 3 to 9 cm with volumetric flow rates up to about 20 L/s. For liquefied natural gas (LNG), pipe sizes up to 20 cm are commonly used with flow rates up to about 100 L/s [46]. Due to space limitations, we cannot discuss all the types of flowmeters used for cryogenic service. The flowmeters considered here and their type (M = mass, V = volumetric, N = neither) are positive displacement (V), angular momentum (M), turbine

(V), differential pressure (N), thermal or calorimetric (M), and hot-wire anemometer (M). The differential pressure element can be in the form of an orifice, venturi, packed screens, or laminar flow channels. Other flowmeters not discussed here but used for cryogenic service are ultrasonic (V), vortex shedding (V), dual turbine (M), and Coriolis or gyroscopic (M). Detailed descriptions of many types of flowmeters used in cryogenic service are given by Alspach et al. [47], Brennan et al. [48, 49], and Brennan and Takano [50]. A discussion of the NIST cryogenic flowmeter calibration facility for use with liquid nitrogen and argon is given by Brennan et al. [46]. Most cryogenic flowmeters can be calibrated to an uncertainty of ±0.5% for volume flow and ±0.2% for mass flow in this facility [51]. The repeatability of individual flowmeters may be larger than these uncertainty values.

61.5.1 Positive Displacement Flowmeter (Volume Flow)

The positive displacement flowmeter works on the principle that the flowing liquid must displace some mechanical element. The movement of the mechanical element is then sensed electronically. The various mechanical elements used are screw impeller, rotating vane, and oscillating piston. A detailed description of these various types of positive displacement flowmeters and an evaluation of them for cryogenic service is given by Brennan et al. [48]. They are generally used with moderate flow rates (1–10 L/s) and are capable of being operated over a 5 to 1 flow range, which means the minimum flow is 1/5 of the maximum flow. A pressure drop of about 30 kPa is typical with these meters at maximum flow. With care it is possible to achieve an uncertainty of ±1% with these meters. It is important to subcool the liquid below the saturation curve to prevent the formation of vapor in the flowmeter. A disadvantage of these meters is that they are subject to wear and need to be recalibrated periodically.

61.5.2 Angular Momentum Flowmeter (Mass Flow)

The angular momentum flowmeter has a rotating member with vanes oriented parallel to the axis. The rotating member is driven by an electric motor through a constant torque clutch (hysteresis drive) as shown in Figure 61.17. The liquid enters the meter through a flow straightener and passes by the rotating vanes. The liquid tends to retard the rotational speed of the rotor in a manner that is inversely proportional to the mass flow rate. Rotor speed is sensed by a magnetic pickup, and the resulting signal treated electronically to indicate mass flow rate. A flow range of 8 to 1 is typical with these meters. Maximum flow rates for these meters may vary from about 2 to 15 kg/s. Pressure drops of 20–50 kPa are typical with these flowmeters at maximum flow. They have been tested with liquid hydrogen [47] and liquid oxygen, nitrogen, and argon [49]. Flowmeters of this type are often used as custody transfer flowmeters on delivery vehicles. Uncertainties of ±2% or less are typical for these flowmeters.

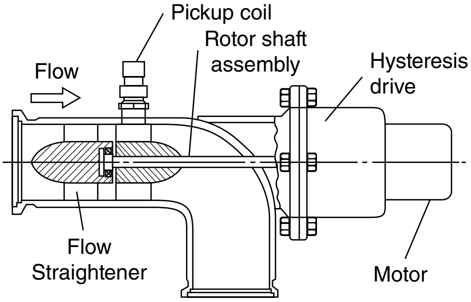


FIGURE 61.17 Angular momentum flowmeter. From Brennan et al. [49].

61.5.3 Turbine Flowmeter (Volume Flow)

The turbine flowmeter consists of a freely rotating bladed rotor, supported by bearings, inside a housing, and an electrical transducer that senses rotor speed. Rotor speed is a direct function of flow velocity. They are used mostly for liquid flows and have a useful range of at least 10 to 1. Calibrations with liquid cryogens may differ from water calibrations by up to ±2% [52]. They are susceptible to errors caused by upstream swirl, so some means of flow straightening is usually required for accurate measurements. An evaluation of several cryogenic turbine flowmeters was reported by Brennan et al. [51]. They ranged in size from 3.2 to 5.1 cm with maximum flow rates between 5 and 14 L/s. Maximum pressure drops ranged from 20 to 100 kPa. Uncertainties were generally less than ±1%.

A commercial turbine flowmeter with ball bearings has even been used for measuring flow in normal and superfluid helium with flow rates between 0.01 and 0.3 L/s [40]. It had a bore diameter of 9.35 mm. The meter output was about 0.5% higher for superfluid helium compared with normal helium. However, care must be taken with superfluid helium to prevent cavitation. A custom-made turbine flowmeter with magnetic bearings has also been used for liquid helium flow [53]. The success of these meters depends very much on maintaining a low drag from the bearings. To insure this, the liquid must be free of solid particles such as frozen air or water. An upstream filter is often used with these flowmeters. Short-term repeatability of ±0.15% was reported for the magnetic-bearing flowmeter. Like other flowmeters with moving parts, they do have a limited lifetime.

Turbine flowmeters have been reported to have response times in the range of 1–10 ms, depending upon blade angle, flowmeter size, and flow rate [52]. These authors also report on the successful use of these meters in reverse flow. As a result, they can be used to a limited extent in transient flow or oscillating flow for frequencies less than about 1–10 Hz.

61.5.4 Differential Pressure Flowmeter

The differential pressure flowmeter can be used with either gas or liquid flows. They operate on the principle that the pressure drop across some flow element is proportional to the flow rate. These meters can be used at cryogenic temperatures if the

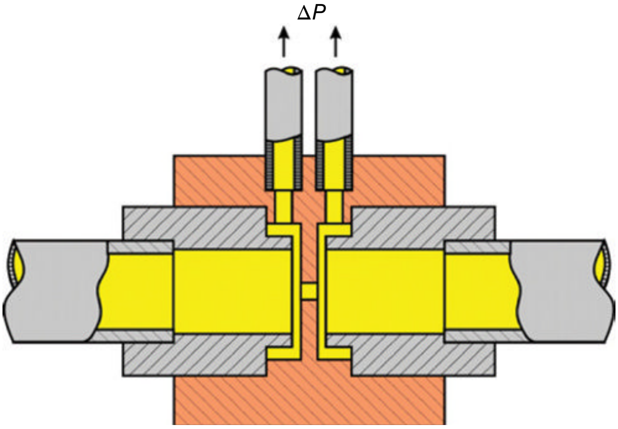


FIGURE 61.18 Orifice flowmeter for oscillating flow.

pressure transducer is located at ambient temperature or if a compatible pressure transducer is used at the cryogenic temperature. These flowmeters have no moving parts; thus, they are desirable for applications that require high reliability. The most common type of flow element is the sharp-edge orifice plate. Usually, the orifice plate is designed for flow in one direction with the sharp edge of the orifice on the entrance side. A symmetric design for the orifice plate as shown in Figure 61.18 has been used in our laboratory for use with oscillating flow [54]. When used for oscillating flows, it is important that the connecting lines are of the same length and that the differential pressure transducer is symmetrical. We have experienced some problems with shifts in the zero reading of differential pressure transducers when the differential pressure changes sign.

The relation between the mass flow rate and the pressure drop ΔP across the orifice is given by

$$\dot{m} = C_{\text{o}} A_{\text{o}} \left[\frac{2\rho \Delta P}{\left(1-\beta^4\right)} \right]^{1/2}, \tag{61.14}$$

where C_{o} is the orifice or discharge coefficient (≈0.6), A_{o} is the cross-sectional area of the orifice, ρ is the fluid density, and β is the ratio of the orifice diameter to the tube inside diameter. Because of the square root dependence of $\dot{m}$ on ΔP, a range of 10 for ΔP yields a range of 3 in $\dot{m}$. This low flow range is a disadvantage of the orifice meter. The orifice coefficient determined from a water calibration can be used for most liquid cryogens with ±2% uncertainty [49]. The uncertainty for use with gas can be somewhat higher. To obtain high accuracy with these meters, it is necessary to have a straight length of tube upstream of the orifice that is at least 20 times the tube diameter, and a length at least five times the tube diameter should be placed downstream of the orifice. Alternatively, flow straighteners in the form of tube bundles can be used if length is

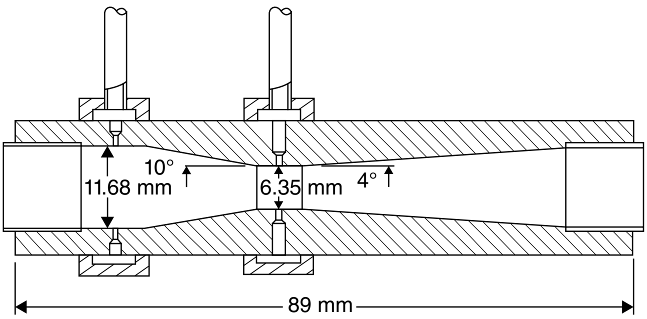


FIGURE 61.19 Cross section of a venturi flowmeter.

restricted. Equation 61.14 shows that this flowmeter is neither an intrinsic mass flowmeter nor an intrinsic volumetric flowmeter because of the square root dependence on the density. The simple design of these meters means that they can be scaled over a very wide range of flow rates. Pressure drops can be made quite small, although for ΔP less than about 1% of the mean pressure, the signal-to-noise ratio may begin to decrease.

A disadvantage of the orifice meter is the large amount of turbulence created by the flow through the orifice. This problem is reduced by using a venturi as the flow element, as shown in Figure 61.19. The throat diameter is usually about one-half the tube diameter. The flow rate through the venturi meter is governed by the same relationship as for the orifice meter (Equation 61.4); however, the discharge coefficient is near unity. Venturi meters have been used in many applications for measuring flow rates of normal, supercritical, and superfluid helium [40, 55]. Short-term repeatability of ±0.5% was reported. Discharge coefficients varied by about 3% over a flow range of 10 to 1 and for temperatures between 1.7 and 4.2 K. The design requirements of the venturi prevent it from being used for reverse flow, such as in fully reversing oscillating flow.

The third type of flow element that can be used in differential pressure flowmeters is the laminar flow element. The laminar flow element gives rise to a linear relationship between mass flow rate and pressure drop. As a result, it can be used over a wider range of flow rates than can the orifice meter and the venturi meter. The governing equation is given by

$$\dot{V} = \frac{\dot{m}}{\rho} = \frac{\Delta P}{Z_{\mathrm{f}}}, \tag{61.15}$$

where $\dot{V}$ is the volumetric flow rate and Z_{f} is the flow impedance of the laminar element. For laminar flow in a gap, the flow impedance is given by

$$Z_{\mathrm{f}} = 12\mu L/wt^{3}, \tag{61.16}$$

where μ is the viscosity, L is the length of the gap, w is the width of the gap, and t is the thickness of the gap. Equation 61.15 shows that the laminar flow element is intrinsically a volumetric flowmeter. In order to achieve laminar flow conditions, the gap thickness must be sufficiently small. As an example, for helium gas with $\dot{m} = 1$ g/s, $P = 2$ MPa, $L = 1$ cm, and $\Delta P = 20$ kPa, the gap thickness must be less than 34 µm at 80 K and less than 9.8 µm at 10 K. The gap width must be 116 mm at 80 K and 240 mm at 10 K. Even though the overall gap width can be achieved with many parallel gaps, the outside dimensions of the laminar flow element will be relatively large. The small gap thickness at cryogenic temperatures makes the laminar flow element difficult to fabricate. No commercial laminar flow elements are available for cryogenic use. A custom-made device has been used with some success to measure oscillating mass flow rate at temperatures of about 10 K in high-pressure helium gas [54]. One caution to note on the use of a gap flow element for measuring oscillating flow is that the high velocity gas flow in the gap can lead to a significant inertance term, which will cause the phase of the pressure drop to lead that of the flow [56]. The inertance I (fluid equivalent of electrical inductance) of a gap is given by

$$I = \frac{\rho L}{wt}, \tag{61.17}$$

and the complex impedance is given by

$$Z_I = j\omega I, \tag{61.18}$$

where $j^2 = -1$ and ω is the angular frequency of sinusoidal oscillation. The presence of j in Equation 61.18 indicates that the component of the dynamic pressure drop due to the inertance leads the flow by 90°. The phase shifting becomes more important at higher frequencies. For oscillating flow the compliance (fluid analog of electrical capacitance) must also be considered to calculate the phase between the pressure drop and the flow.

The fourth type of flow element that can be used in differential flowmeters is that of packed screen or packed spheres. Correlations for the friction factor in such a packing must be used to find the relation between the pressure drop and the flow. With such geometries the relation between the pressure drop and flow is nonlinear. The effect of inertance is generally less in packed screen or packed spheres compared to that in gaps because of slower fluid velocities.

61.5.5 Thermal or Calorimetric (Mass Flow)

In the thermal flowmeter, the flowing fluid is heated with a constant power $\dot{Q}$, which causes its temperature to rise by an amount ΔT. A thermocouple or thermopile measures this temperature difference between the outgoing and incoming fluid flow. The mass flow rate is given by

$$\dot{m} = \frac{\dot{Q}}{C_p \Delta T}, \tag{61.19}$$

where C_p is the specific heat of the fluid. This flowmeter is a true mass flowmeter. A commercial thermal flowmeter was used by Bugeat et al. [57] to measure mass flow rates around 10 mg/s of hydrogen and helium gas at temperatures between 100 and 300 K. They reported a nonlinearity of less than 2% and a response time of about 0.22 s at 158 K in hydrogen gas for a flow of 9 mg/s. This type of flowmeter should work with flow in either direction.

61.5.6 Hot-Wire Anemometer (Mass Flow)

The hot-wire or constant temperature anemometer (CTA) infers mass flow rates from the changing heat transfer rates associated with a heated element [58]. The resistively heated element, often a fine wire, has a large temperature coefficient of resistance and a large length-to-diameter ratio. With feedback electronics, the electrical power to the element is varied automatically to maintain the element at a constant resistance (temperature) as the flow rate varies. The power, or voltage squared, is correlated to the mass flow rate from a calibration of the device against a standard flowmeter at ambient temperature. For an ideal gas, the CTA is a true mass flowmeter [59].

We have used commercial hot-wire anemometer probes successfully for the measurement of helium gas flows at temperatures down to 77 K [54, 59]. The calibration changes in a linear manner with the gas temperature, but it is not a strong function of this temperature. The probes were fabricated with a 3.8 μm diameter tungsten wire about 2 mm long attached to the wire supports. The wire was heated to about 297 K to give high sensitivity and reproducible results. Since the power input with this high wire temperature was about 1 W, the CTA was turn on only briefly in order to make the needed measurements. The response time of the CTA was measured to be less than 15 μs in zero flow. The response time is even faster at finite flow rates. The fast response time of the CTA makes it ideal for measuring turbulence, transient flow, or oscillating flow. Because of the fine wire, it can only be used in very clean gas flows.

The CTA has been used with modified commercial vacuum fittings as shown in Figure 61.20 to measure oscillating mass flow rates in compressed helium gas at temperature down to 77 K within a pulse tube refrigerator [59]. Oscillating frequencies up to 30 Hz could be measured. To provide the necessary temperature correction, an identical tungsten wire probe to serve as an RTD was inserted in the assembly from the opposite side (see Fig. 61.20). The diameter of the tubes extending from the device shown in Figure 61.20 was 3.2 mm. Consequently, this type of flowmeter can be made with very small gas volumes, which is necessary for measurements of the oscillating gas flow within small Stirling or pulse tube refrigerators. Several layers of stainless steel screen on each side of the probes ensured that the flow was uniform in both

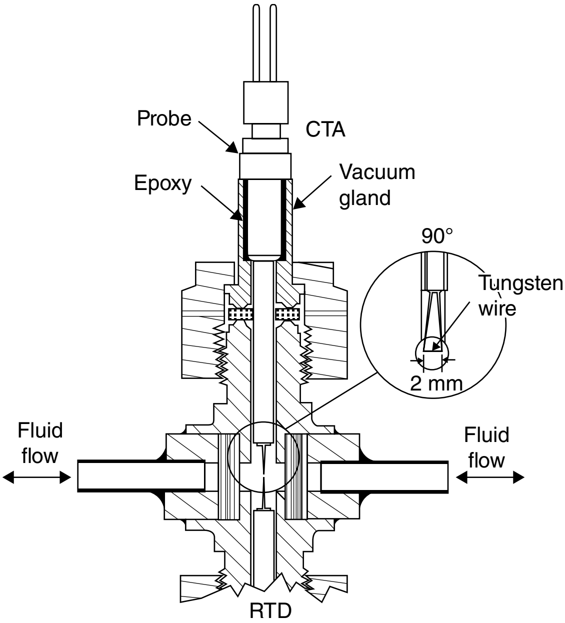


FIGURE 61.20 Modified vacuum assembly with CTA and RTD probes in place.

directions. The integrated mass flows measured for each direction of flow agreed to within 1.6% for measurements near 80 K.

61.6 LIQUID LEVEL

A common measurement problem in cryogenics is that of determining the level of cryogens in storage dewars or within an experimental apparatus. For the heavier cryogens (hydrogen and helium excluded), a measurement of the hydrostatic head with a differential pressure gage can be correlated to the liquid level. This technique is commonly used on storage dewars of liquid nitrogen for approximate (±10%) indications of liquid level. More precise measurements can be made with capacitance liquid-level gages that are available for use with liquid nitrogen from a few different manufacturers. The detection of liquid level at discrete locations (for level control) is often performed in many commercial devices with a self-heated resistance or diode thermometer. The higher heat transfer rate in the liquid phase causes the temperature of the thermometer to decrease when it is immersed in the liquid. When several of these thermometers are located at varying heights, a semicontinuous reading of liquid level is available. For continuous readings, a vertical wire or foil can be resistively heated [60, 61]. The resistance of the wire will be a function of the liquid level. The heat input to the liquid cryogen with this device can be minimized by using a current pulsed periodically to update the previous reading. Modern electronics makes this a simple task.

Greater sensitivity and less heating occurs when the wire is made with a superconducting material that has a critical temperature slightly above the normal boiling temperature of the liquid cryogen of interest. A current very near the critical current is passed through the wire so only the portion in the liquid remains in the superconducting state. Again, intermittent use reduces the total heat dissipation. For liquid nitrogen, the wire can be a high-temperature superconductor. For liquid helium, the wire must be a low-temperature superconductor such as tantalum ($T_c = 4.4$ K). The current through the wire must be varied if the temperature of the helium bath is lowered. These superconducting level sensors are commercially available from several manufacturers.

61.7 MAGNETIC FIELD

Instruments that measure magnetic field are usually called gaussmeters, teslameters, or magnetometers. For measurements of weak magnetic fields, the Superconducting Quantum Interference Device (SQUID) magnetometer is unsurpassed. It can detect changes in magnetic field as small as 10^{-15} T [17]. They can be used for fields up to about 1 mT.

For measurement of high magnetic fields at cryogenic temperatures, magnetoresistive sensors or Hall effect sensors can be used. The principle of magnetoresistive sensors is that the resistance of a metal or semimetal changes with the applied magnetic field. This change can vary from 2 to 3% change in a nickel-iron alloy to as much as 10^6 in bismuth. For small values of magnetic field, the change in resistance is proportional to the square of the magnetic field, whereas for larger fields it may have several higher order terms. Magnetoresistance is also a function of temperature. This complex behavior makes it difficult to use magnetoresistive sensors for measuring magnetic fields over a wide range of temperatures and fields.

The Hall effect sensor works as follows: A current is passed through the sensor in the x direction. With a magnetic field applied in the z-direction, a voltage in the y-direction is generated that is related to the magnetic flux density B. With proper cancellation of offset voltages, the output of the Hall effect element is given by

$$V = \left(2\mu_{\mathrm{H}} \frac{w}{L} R \right) IB, \tag{61.20}$$

where μ_{H} is the electron Hall mobility, w/L is the effective width-to-length ratio for the Hall element, I is the excitation current, R is the resistance of the Hall element, and B is the magnetic flux density. Any temperature dependence comes about as a result of μ_{H} and R. The expression within the parentheses makes up the Hall sensitivity γ. It is a weak function of B. Glowacki and Ignatowicz [62] showed that $\mathrm{Cd}_x\mathrm{Hg}_{1-x}\mathrm{Te}$ ($x = 0.175$) films produced good Hall effect sensors for use between 4.2 and 20 K. There was no temperature dependence in this range and the Hall sensitivity varied by about 20%

between a magnetic flux density of 0.1 and 2 T. The output was very reproducible after many thermal cycles.

Commercial cryogenic Hall effect sensors are available along with the control electronics from a variety of manufacturers. The usual materials are InAs, InSb, and GaAs. Sample and Rubin [18] have measured the temperature dependence and the linearity of several of these Hall sensors. They recommend the Hall sensor over other magnetic field sensors for use at cryogenic temperatures. These sensors are available in axial or transverse field models. Typical diameters are about 6 mm and a length of 5 mm for the axial sensor. The transverse sensor is flat with a width of 5 mm and a length of 16 mm. They can be used in magnetic fields up to ±15 T with deviations from linearity less than 1.5%. Outputs at 4.2 and 77 K are within ±1.5% of the calibration at 300 K. Repeatability after many thermal cycles is within 1%. However, they are susceptible to damage from thermal shock after repeated cycling.

61.8 CONCLUSIONS

We have discussed and compared the various sensors and instrumentation that are commonly used for measurements at cryogenic temperatures. In most cases, commercial products are available for this need. We have reviewed available instrumentation for measurements of temperature, strain, pressure, flow, liquid level, and magnetic field at cryogenic temperatures. The comparisons of various sensors should allow the reader to quickly determine which sensor is best suited for the task at hand. The references cited should be useful if more details are needed.

REFERENCES

1. Preston-Thomas, H., 1990, "The International Temperature Scale of 1990 (ITS-90)," *Metrologia*, Vol. 27, pp. 3–10, and 107.
2. BIPM, 2011, "Supplementary information for the realization of the PLTS-2000", Adopted by the Consultative Committee for Thermometry, International Committee for Weights and Measures, pp. 1–25; slightly modified version of, Rusby, R. L., Fellmuth, B., Engert, J., Fogle, W. E., Adams, E. D., Pitre, L., and Durieux, M., 2007, "Realization of the 3He melting pressure scale, PLTS-2000," *J. Low Temp. Phys.*, Vol. 149, pp. 156–175.
3. Mangum, B. W., Furukawa, G. T., Kreider, K. G., Meyer, C. W., Ripple, D. C., Strouse, G. F., Tew, W. L., Moldover, M. R., Carol Johnson, B., Yoon, H. W., Gibson, C. E., and Saunders, R. D., 2001, "The kelvin and temperature measurements," *J. Res. Natl. Inst. Stand. Technol.*, Vol. 106, pp. 105–149.
4. Strouse, G. F., 2008, *Standard Platinum Resistance Thermometer Calibrations From the AR TP to the Ag FP*, NIST Special Publication SP 250-81, U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology, Gaithersburg, MD.

5. Tew, W. L. and Meyer, C. W., 2003, “Recent Results of NIST Realization of the ITS-90 Below 84 K,” *Temperature: Its Measurement and Control in Science and Technology*, Vol. 7, D. C. Ripple, ed., American Institute of Physics, New York, pp. 143–148.
6. Courts, S. S., Holmes, D. S., Swinehart, P. R., and Dodrill, B. C., 1991, “Cryogenic thermometry,” *Applications of Cryogenic Technology*, Vol. 10, J. P. Kelley, ed., Plenum Press, New York, pp. 55–69.
7. Ekin, J. W., 2006, *Experimental Techniques for Low-Temperature Measurements*, Oxford University Press, Oxford, pp. 185–225.
8. Holmes, D. S. and Courts, S. S., 1992, “Resolution and accuracy of cryogenic temperature measurements,” *Temperature: Its Measurement and Control in Science and Industry*, Vol. 6, J. F. Schooley, ed., American Institute of Physics, New York, pp. 1225–1230.
9. Rubin, L. G., Brandt, B. L., and Sample, H. H., 1982, “Cryogenic thermometry: a review of recent progress, II,” *Cryogenics*, Vol. 22, pp. 491–503.
10. Rubin, L. G., 1997, “Cryogenic thermometry: a review of progress since 1982,” *Cryogenics*, Vol. 37, pp. 341–356.
11. Sparks, L. L., 1983, “Temperature, strain, and magnetic field measurements,” *Materials at Low Temperatures*, R. P. Reed and A. F. Clark, eds., American Society for Metals, Metals Park, OH, pp. 515–571.
12. Yeager, C. J. and Courts, S. S., 2001, “A review of cryogenic thermometry and common temperature sensors,” *IEEE Sensors J.*, Vol. 1, pp. 352–360.
13. Courts, S. S. and Krouse, J. K.; Temperature: Its Measurement and Control in Science and Industry. 2013, “A new capsule platinum resistance thermometer for cryogenic use,” *AIP Conf. Proc.*, Vol. 1552, 8, pp. 168–173.
14. Peng, L., 2012, “A proposal for RhFe thermometer used as the interpolating instrument of temperature scale in the range of 1 K to 25 K,” 26th International Conference on Low Temperature Physics, *J. Phys. Conf. Ser.*, Vol. 400, p. 052016.
15. Lawless, W. N., 1972, “Thermometric properties of carbon impregnated porous glass at low temperatures,” *Rev. Sci. Instrum.*, Vol. 43, pp. 1743–1747.
16. Bentley, R. E., 1998, *Handbook of Temperature Measurement, Vol. 3: The Theory and Practice of Thermoelectric Thermometry*, Springer, Singapore.
17. Fagaly, R. L., 1987, “Superconducting magnetometers and instrumentation,” *Sci. Prog.*, Vol. 71, pp. 181–201.
18. Sample, H. H. and Rubin, L. G., 1977, “Instrumentation and methods for low temperature measurements in high magnetic fields,” *Cryogenics*, Vol. 17, pp. 597–606.
19. Rubin, L. G., Brandt, B. L., and Sample, H. H., 1986, “Some practical solutions to measurement problems encountered at low temperatures and high magnetic fields,” *Advances in Cryogenic Engineering*, Vol. 31, R. W. Fast, ed., Plenum Press, New York, pp. 1221–1230.
20. Cutkowsky, R. D., 1970, “An A-C resistance thermometer bridge,” *J. Res. Natl. Bur. Stand.*, Vol. 74C, pp. 15–18.
21. Kreider, K. G., 1992, “Thin-film thermocouples,” *Temperature: Its Measurement and Control in Science and Industry*, Vol. 6, J. F. Schooley, ed., American Institute of Physics, New York, pp. 643–648.

22. Louie, B. and Steward, W. G., 1990, "Onset of nucleate and film boiling resulting from transient heat transfer to liquid hydrogen," *Advances in Cryogenic Engineering*, Vol. 35A, R. W. Fast, ed., Plenum Press, New York, pp. 403–412.

23. Giarratano, P. J. and Steward, W. G., 1983, "Transient forced convection heat transfer to helium during a step in heat flux," *J. Heat Transf.*, Vol. 105, pp. 350–357.

24. Giarratano, P. J., Lloyd, F. L., Mullen, L. O., and Chen, G. B., 1982, "A thin platinum film for transient heat transfer studies," *Temperature: Its Measurement and Control in Science and Industry*, Vol. 5, J. F. Schooley, ed., American Institute of Physics, New York, pp. 859–863.

25. Louie, B., Radebaugh, R., and Early, S. R., 1986, "A thermometer for fast response in cryogenic flow," *Advances in Cryogenic Engineering*, Vol. 35A, R. W. Fast, ed., Plenum Press, New York, pp. 1235–1246.

26. Linenberger, D., Spellicy, E., and Radebaugh, R., 1982, "Thermal response times of some cryogenic thermometers," *Temperature: Its Measurement and Control in Science and Industry*, Vol. 5, J. F. Schooley, ed., American Institute of Physics, New York, pp. 1367–1372.

27. Rao, M. G., Scurlock, R. G., and Wu, Y. Y., 1983, "Miniature silicon diode thermometers for cryogenics," *Cryogenics*, Vol. 23, pp. 635–638.

28. Rawlins, W., Radebaugh, R., and Timmerhaus, K. D., 1991, "Monitoring rapidly changing temperatures of the oscillating working fluid in a regenerative refrigerator," *Applications of Cryogenic Technology*, Vol. 10, J. P. Kelley, ed., Plenum Press, New York, pp. 71–83.

29. Rawlins, W., Timmerhaus, K. D., and Radebaugh, R., 1992, "Resistance thermometers with fast response for use in rapidly oscillating gas flows," *Temperature: Its Measurement and Control in Science and Industry*, Vol. 6, J. F. Schooley, ed., American Institute of Physics, New York, pp. 471–474.

30. Hannah, R. L. and Reed, S. E., 1992, *Strain Gage Users' Handbook*, Elsevier Applied Science, New York.

31. Starr, J. E., 1992, "Basic strain gage characteristics," *Strain Gage Users' Handbook*, R. L. Hannah and S. E. Reed, eds., Elsevier Applied Science, New York, pp. 1–77.

32. Pavese, F., 1984, "Investigation of transducers for large-scale cryogenic systems in Italy," *Advances in Cryogenic Engineering*, Vol. 29, R. W. Fast, ed., Plenum Press, New York, pp. 869–877.

33. Walstrom, P. L., 1975, "The effect of high magnetic fields on metal foil strain gauges at 4.2 K," *Cryogenics*, Vol. 15, pp. 270–272.

34. Freynik, H. S., Roach, D. R., Deis, D. W., and Hirzel, D. G., 1978, "Evaluation of metal-foil strain gauges for cryogenic application in magnetic fields," *Advances in Cryogenic Engineering*, Vol. 24, K. D. Timmerhaus, R. P. Reed, and A. F. Clark, eds., Plenum Press, New York, pp. 473–479.

35. Ferrero, C. and Marinari, C., 1990, "Strain analysis at cryogenic temperatures: self-heating effect and linearization of the apparent-strain curve," *Advances in Cryogenic Engineering*, Vol. 35B, R. W. Fast, ed., Plenum Press, New York, pp. 1609–1616.

36. Hartwig, G. and Wüchner, F., 1975, "Low temperature mechanical testing machine," *Rev. Sci. Instrum.*, Vol. 46, pp. 481–485.

37. McConnville, G. T., 1969, "Thermomolecular pressure corrections in helium vapour pressure thermometry: the effect of the tube surface," *Cryogenics*, Vol. 9, pp. 122–127.

38. Jacobs, R., 1986, "Cryogenic applications of capacitance-type pressure sensors," *Advances in Cryogenic Engineering*, Vol. 31, R. W. Fast, ed., Plenum Press, New York, pp. 1277–1284.

39. Kashani, A., Wilcox, R. A., Spivak, A. L., Daney, D. E., and Woodhouse, C. E., 1990, "SHOOT flowmeter and pressure transducers," *Cryogenics*, Vol. 30, pp. 286–291.

40. Daney, D. E., 1988, "Behavior of turbine and venturi flowmeters in superfluid helium," *Advances in Cryogenic Engineering*, Vol. 33, R. W. Fast, ed., Plenum Press, New York, pp. 1071–1079.

41. Cerutti, G., Maghenzani, R., and Molinar, G. F., 1983, "Testing of strain-gauge pressure transducers up to 3.5 MPa at cryogenic temperatures and in magnetic fields up to 6 T," *Cryogenics*, Vol. 23, pp. 539–545.

42. Boyd, C., Juanarena, D., and Rao, M. G., 1990, "Cryogenic pressure sensor calibration facility," *Advances in Cryogenic Engineering*, Vol. 35B, R. W. Fast, ed., Plenum Press, New York, pp. 1573–1581.

43. Hershberg, E. L. and Lyngdal, J. W., 1994, "Self heating in piezoresistive pressure sensors at cryogenic temperatures," *Advances in Cryogenic Engineering*, Vol. 39, P. Kittel, ed., Plenum Press, New York, pp. 1123–1130.

44. Clark, D. L., 1992, "Temperature compensation for piezoresistive pressure transducers at cryogenic temperatures," *Advances in Cryogenic Engineering*, Vol. 37B, R. W. Fast, ed., Plenum Press, New York, pp. 1447–1452.

45. Walstrom, P. L. and Maddocks, J. R., 1987, "Use of Siemens KPY pressure sensors at liquid helium temperatures," *Cryogenics*, Vol. 27, pp. 439–441.

46. Brennan, J. A., LaBrecque, J. F., and Kneebone, C. H., 1976, "Progress report on cryogenic flowmetering at the National Bureau of Standards," *Instrumentation in the Cryogenic Industry, Proceedings of the First Biennial Symposium*, Vol. 1, Houston, TX, October 11–14, 1976, Instrument Society of America, Pittsburgh, PA, pp. 621–636.

47. Alspach, W. J., Miller, C. E., and Flynn, T. M., 1966, "Mass flowmeters in cryogenic service," *Flow Measurement Symposium, ASME Flow Measurement Conference*, Pittsburgh, PA, September 26–28, 1966, American Society of Mechanical Engineers, New York, pp. 34–56.

48. Brennan, J. A., Dean, J. W., Mann, D. B., and Kneebone, C. H., 1971, *An Evaluation of Positive Displacement Cryogenic Volumetric Flowmeters*, National Bureau of Standards Technical Note 605, U.S. Department of Commerce, Washington, DC.

49. Brennan, J. A., Stokes, R. W., Kneebone, C. H., and Mann, D. B., 1974, *An Evaluation of Selected Angular Momentum, Vortex Shedding, and Orifice Cryogenic Flowmeters*, National Bureau of Standards Technical Note 650, U.S. National Bureau of Standards, Washington, DC.

50. Brennan, J. A. and Takano, A., 1982, "A preliminary report on the evaluation of selected ultrasonic and gyroscopic flowmeters at cryogenic temperatures," *Proceedings of the Ninth International Cryogenic Engineering Conference*, K. Yasukochi and H. Nagano, eds., Butterworth, Guildford, Surrey, pp. 655–658.

51. Brennan, J. A., Mann, D. B., Dean, J. W., and Kneebone, C. H., 1972, "Performance of NBS cryogenic flow research facility," *Advances in Cryogenic Engineering*, Vol. 17, K. D. Timmerhaus, ed., Plenum Press, New York, pp. 199–205.

52. Alspach, W. J. and Flynn, T. M., 1965, "Considerations when using turbine-type flowmeters in cryogenic service," *Advances in Cryogenic Engineering*, Vol. 10, K. D. Timmerhaus, ed., Plenum Press, New York, pp. 246–252.

53. Rivetti, A., Martini, G., Goria, R., and Lorefice, S., 1987, "Turbine flowmeter for liquid helium with the rotor magnetically levitated," *Cryogenics*, Vol. 27, pp. 8–11.

54. Radebaugh, R. and Rawlins, W., 1993, "Measurement of oscillating mass flows at low temperatures," *Devices for Flow Measurement and Control—1993*, C. J. Blechinger and S. A. Sherif, eds., American Society of Mechanical Engineers, New York, pp. 25–32.

55. Rivetti, A., Martini, G., and Birello, G., 1994, "Metrological performances of Venturi flowmeters in normal, supercritical, and superfluid helium," *Advances in Cryogenic Engineering*, Vol. 39, P. Kittel, ed., Plenum Press, New York, pp. 1051–1058.

56. Yuan, S. W. K., Curran, D. G. T., and Cha, J. S., 2010, "A non-tube inertance device for pulse tube cryocoolers," *Advances in Cryogenic Engineering*, Vol. 55, J. G. Weisend II, ed., American Institute of Physics, New York, pp. 143–148.

57. Bugeat, J. P., Petit, R., and Valentian, D., 1987, "Thermal helium mass flowmeter for space cryostat," *Cryogenics*, Vol. 27, pp. 4–7.

58. Perry, A. E., 1982, *Hot Wire Anemometry*, Clarendon Press, Oxford.

59. Rawlins, W., Radebaugh, R., and Timmerhaus, K. D., 1993, "Thermal anemometry for mass flow measurement in oscillating cryogenic gas flows," *Rev. Sci. Instrum.*, Vol. 64, pp. 3229–3235.

60. Maimoni, A., 1956, "Hot wire liquid-level indicator," *Rev. Sci. Instrum.*, Vol. 27, pp. 1024–1027.

61. Wexler, A. and Corak, W. S., 1951, "Measurement and control of the level of low boiling liquids," *Rev. Sci. Instrum.*, Vol. 22, pp. 941–945.

62. Glowacki, B. A. and Ignatowicz, S. A., 1987, "Hall probe $Cd_xHg_{1-x}Te$ for use in magnetic investigations of Nb_3Sn superconducting layers," *Cryogenics*, Vol. 27, pp. 162–164.

62

TEMPERATURE-DEPENDENT FLUORESCENCE MEASUREMENTS

James E. Parks[1], Michael R. Cates[2], Stephen W. Allison[2],
David L. Beshears[2], M. Al Akerman[2], and Matthew B. Scudiere[2]

[1] *Department of Physics, University of Tennessee, Knoxville, TN, USA*
[2] *Emco-Williams Inc, Knoxville, TN, USA*

62.1 INTRODUCTION

Temperature is one of the most important attributes of physical systems, and its measurement is critical to many aspects of scientific research and development. There are a significant number of circumstances, however, in which temperature is difficult or impossible to measure by standard means such as thermometers, thermocouples, and infrared surface emissions. Most of these circumstances are associated with challenges such as very high temperature, vibrating or moving surfaces, difficulty of access, hazardous locations, and the like. Typical examples include centrifuges, turbine engine components, high-speed motors, and vibrating or moving production machinery. In many situations, it is also necessary to measure temperature remotely, without direct contact, because of difficulties of access, intervening heated air or other gases, or movement of the component to be measured [1–6].

Fluorescent materials (phosphors), bonded to surfaces of interest, provide a very important approach to temperature measurement in many of these difficult circumstances. Most phosphors have characteristic emissions that are affected by temperature, since the phosphor molecular structures are directly correlated to vibrations and rotations associated with temperature. Some phosphors, depending on their molecular structure, tend to have emission bands in the visible and infrared which can

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

be sensed by standard photodetectors of many types. Certain temperature ranges for particular phosphors will show very strong temperature dependence; consequently, monitoring the fluorescence in these ranges can produce a very sensitive measurement of that phosphor's temperature. When the temperature of the phosphor layer is determined, the temperature of the component to which it is bonded can be inferred. The layers required are often no greater than about 50 μm, so they interfere minimally in most situations.

Phosphors that exhibit significant temperature sensitivity in various temperature ranges, and which are reasonably stable chemically, are often called thermographic phosphors (TPs). These TPs will have emissions that can be measured in various ways to determine temperature. The most common method has been to measure the fluorescent intensity as a function of time and extract the characteristic lifetime of the emission. That lifetime, for TPs in particular, is often a strong function of temperature. Another approach was been to select two or more emission wavelength bands, where the intensity of one or more is strongly dependent on temperature and at least one other for which it is not, and to use the ratio of their intensities as a normalized function of temperature. In other specialized situations it is also possible to simply use fluorescence intensity to determine temperature. In yet other circumstances the emission wavelength shift, which can be temperature dependent, can be monitored. Certain emission bands, too, can vary in width as a function of temperature; those variations in width can be used for temperature measurement.

A number of energy sources can be used to stimulate the fluorescence of a TP. The most common source used to date is ultraviolet radiation from a laser, laser diode, or light-emitting diode (LED). The stimulated fluorescence, then, is less energetic, therefore of longer wavelength, typically in the visible region of the electromagnetic spectrum. Other sources, such as X-rays or other high-energy electromagnetic emissions, electrons, protons, and the like will also activate TPs, as indeed, for certain materials, can acoustic impact or related mechanical processes.

In all cases, it is important to recognize that the material measured is the phosphor itself not the surface beneath it. That limitation, however, can be used to advantage when two or more TP layers are used, perhaps with interstitial material layers. In such arrangements heat flux, for example, can be determined by measuring the two or more temperature differences and applying the heat transfer parameters of the interstitial layers. Relatedly, the wear of a surface can be monitored by measuring the emissions from that thinning surface, or the efficiency of a thermal barrier coating can be determined by its fluorescent properties or the emissions of TPs mixed with it or on its surface. TPs in powder form, with particle diameters down to a fraction of micrometer, can also be injected into moving fluids, liquid, or gas, and, when monitored, indicate velocity, temperature, or fluid density. In short, TPs can be used in a variety of ways to make measurements of temperature and related physical properties, often in situations where other methods to make such measurements would be futile or severely limited.

62.2 ADVANTAGES OF PHOSPHOR THERMOMETRY

The use of TPs for temperature measurement has several major advantages over more standard temperature measurement methods. Some of these advantages will be made clear in the following sections, but it is useful to list a number of them here. (i) TP response has no dependence on surface properties such as emissivity or reflectance. The fluorescent characteristic measured is only associated with the molecular conditions of the phosphor layer. (ii) There are TPs covering a vast temperature range, from cryogenic systems up to systems near 2000 K. (iii) Calibrated TP systems do not drift over time or require any kind of reference measurement. (iv) Most TPs are chemically stable and have low electrical and heat conductivity, so their presence on a surface is minimally perturbing. (v) TP emission characteristics have fast time responses, typically on the order of microseconds, so measurement systems with rapid time dependence are straightforward to produce. (vi) TPs can be distributed in very thin layers over significant areas; consequently, they have the promise of effective use in two-dimension temperature measurement.

62.3 THEORY AND BACKGROUND

Many compounds have fluorescent properties, but one important class of these is rare-earth oxides and similar structures that have a small addition of a different rare-earth ion distributed through the molecular lattice. This small addition is called an activator or dopant. Its purpose is to make the fluorescence more likely upon absorption of a stimulating energy source. These are particularly true with rare-earth compounds because the activator ion is relatively isolated from others of its type in the molecular lattice, leaving it with fewer quantum decay routes upon excitation. Many of those routes are photon emitting, hence the fluorescent emission. Rare-earth metals are chemically unique in that the chemical valence shell of the atom is lower in electronic energy than the closed electronic shell above it. Consequently, all rare-earth compounds have very similar chemical behaviors and tend to be stable up to high temperatures. Some refractory materials, such as Y_2O_3 and other similar compounds, can be made into effective TPs by appropriate small additions of other metal ions, such as Eu, Tb, Dy, and others. The use of high-temperature materials that can be applied as powders in binders or sprayed directly on surfaces by other means has made the effective use of TPs for many applications possible.

In this section we will restrict our mathematical analysis to rare-earth phosphors activated by Eu ions, both because they are common and effective and because they are relatively simple in their molecular behavior. For example, there are many europium-doped phosphor compounds in which the lifetime of the fluorescence from certain emission lines is dependent on temperature. These include Y_2O_3 : Eu, La_2O_2S : Eu, $LaPO_4$: Eu, and $LuPO_4$: Eu.

There are a number of mechanisms responsible for phosphor temperature responses in general. Given here is an explanation that often applies to Eu-activated phosphors. The Eu ion replaces a small fraction of the dominant rare-earth ions in the lattice. That relative isolation allows the Eu ion to excite to higher energy levels without significant competition from vibrational states. When those levels deexcite the probability of photon emission is high, thereby making the phosphor more efficient.

Now, we will select one of the very common TPs, especially useful near room temperature, $La_2O_2S:Eu$, to illustrate the theory of temperature-dependent behavior among TPs. Further, we will consider only the fluorescent lifetime method of extracting temperature. This method is not only the most commonly used but also helps clarify the atomic behavior within the molecular structures involved.

An energy level diagram for $La_2O_2S:Eu$ is shown in Figure 62.1. In Figure 62.1a, for clarity, the potential energy diagrams for the lowest ground and excited electronic states are shown. The abscissa corresponds to the position of the activator in the lattice. It will undergo various allowed vibrations. One might picture the europium atom connected by springs to its neighboring oxygen and sulfur atoms and oscillating through the equilibrium center with increasing amplitude as temperature is increased. The horizontal lines within the potential wells illustrate that the vibrational energy is quantized. The higher the temperature, the more vibration states are occupied. The vertical line, arrow #1, corresponds to the energy of an ultraviolet photon, ν_{uv}. Upon

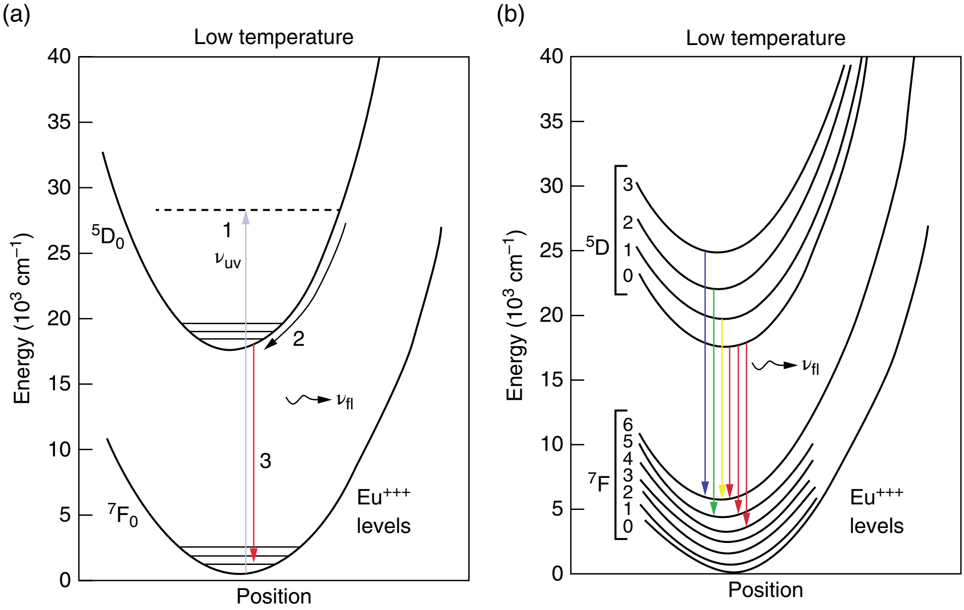


FIGURE 62.1 Energy level diagram of europium phosphor (a) lowest ground and excited electronic states and (b) multiple levels of the 5D and 7F states with fluorescence transitions indicated.

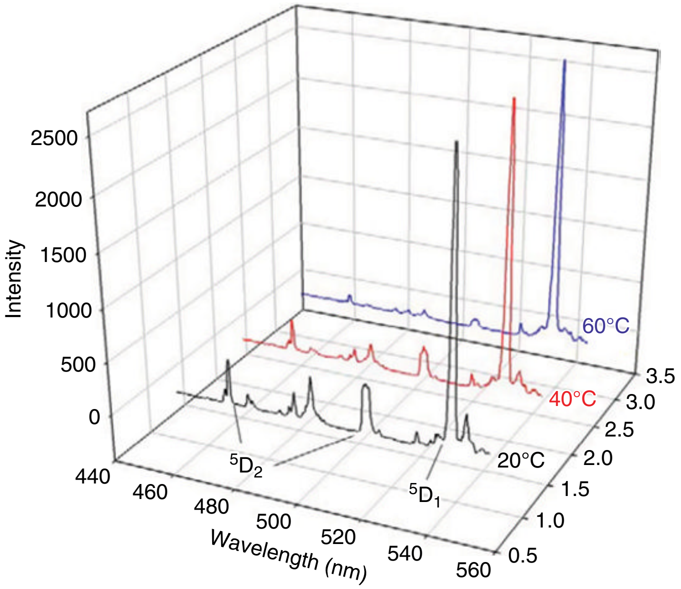


FIGURE 62.2 Fluorescent spectrum of $La_2O_2S:Eu$.

absorption by the host or dopant, almost immediately (i.e., in $< 10^{-9}$ s), excess energy is given to the lattice vibrations following path 2 as seen. The energy is redistributed among the vibration levels as governed by a Boltzmann temperature. At low temperatures, the potential energy in the excited electronic state has only one path for returning to the ground state: to emit the energy in the form of a fluorescence photon, ν_{fl}. The picture is a little more complicated in that, as seen in Figure 62.1b, there are several ground electronic states, denoted by ^7F_j, as well as several ^5D_i excited electronic states that lie sequentially higher in energy. Various fluorescence transitions are possible, as denoted by the arrows, the corresponding transitions emit from red to blue with increasing energy. They occur with differing probabilities and at different energies (or wavelengths). A scan of the fluorescence intensity with wavelength is called a fluorescence spectrum. A portion is shown in Figure 62.2 at three different temperatures. The $^5D_2$ states change in going from 20 to 60°C. The $^5D_1$ state is unchanged.

Another level of complexity has been added to the energy level diagram for $La_2O_2S:Eu$ in Figure 62.3, which depicts a state termed the "charge transfer (CT) state." The CT state explanation, first put forward by Fonger and Struck [7], involves the transfer of an electron from a neighboring atom to the europium. It is indicated qualitatively in the right-hand figure such that only the lower vibrational levels of any excited electronic state may be occupied at low temperatures. At low temperatures, following excitation, the ^5D_i state is populated and fluorescence is essentially the only deexcitation pathway. However, at sufficiently high temperatures, the

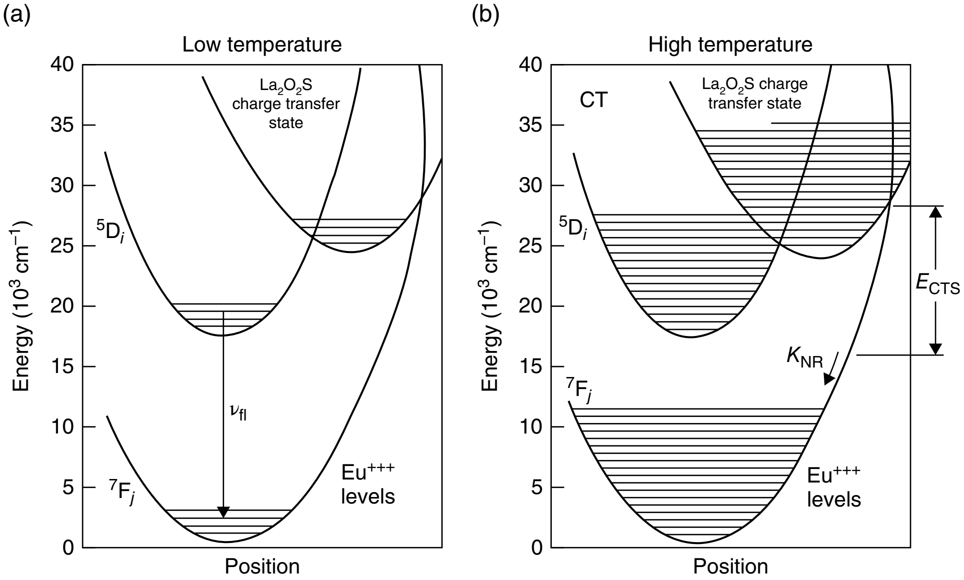


FIGURE 62.3 Energy level diagram for (a) low temperature and (b) high temperature with the charge transfer state.

excited state vibrational levels overlap with those of the CT state. The CT state provides another path for deexciting the phosphor. In this case, cross over to the CT state can occur. The CT state crosses the ground state potential energy curve at E_{CTS}. When this happens, energy is transferred to the host lattice through vibrations (or some energy can leak into a lower ^5D_i state) and not via fluorescence. The higher the temperature, the more likely and the faster this occurs. This depopulation of the electronic state decreases the lifetime and the overall intensity of the emission. This process is called "quenching."

A simple energy level diagram in Figure 62.4 serves as the basis for obtaining a simple rate equation in TPs of this type, illustrating (i) that the fluorescence is characterized by single exponential decay and (ii) the temperature dependence of this decay. It is assumed that the state is excited instantaneously with a population of N_0. The change in number of excited states, dN, is proportional to the number of excited states N and the elapsed time dt such that

$$dN \propto -Ndt \tag{62.1}$$

This model assumes that there is no significant feeding from any other electronic level. A constant of proportionality, κ, is the total decay rate. Therefore,

$$\frac{dN}{N} = -\kappa dt \text{ and } N = N_0 e^{-\kappa t} \tag{62.2}$$

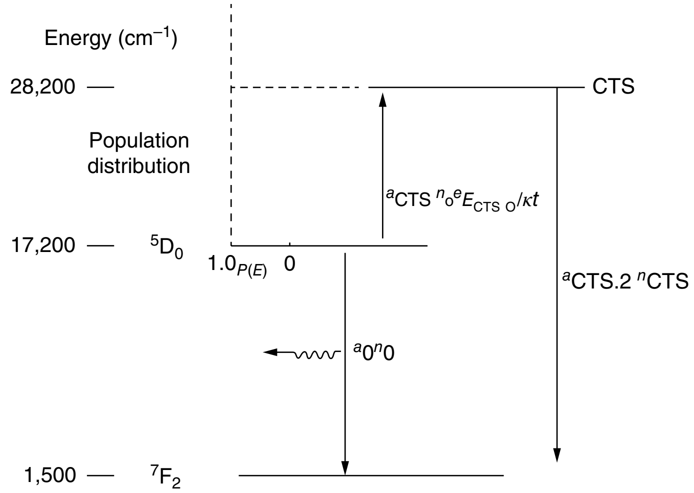


FIGURE 62.4 Simplified energy level diagram for a typical Eu-doped phosphor with Boltzmann distribution, $P(E)$ based on Struck-Fonger Model.

and the decay is exponential. κ is the sum of a radiative component, κ_{rad}, and a nonradiative component, κ_{nonrad}, where energy is lost, in this case, to the CT state. κ_{rad}, is the familiar Einstein coefficient. The expression is

$$\kappa = \frac{1}{\tau} = \kappa_{\text{rad}} + \kappa_{\text{nonrad}} = \frac{1}{\tau_0} + \kappa_{\text{nonrad}} \tag{62.3}$$

where τ is the measured lifetime and τ_0 is the low temperature value of the lifetime before the quenching temperature has been reached. To ascertain the temperature dependence of this non-radiative quenching factor, the relative distribution of vibrational states must be considered. Boltzmann's law, a fundamental law of thermodynamics, shows that the population distribution follows an exponential dependence where, for a given level N_l, the ground state population is N_i,

$$N_l = N_i e^{-(E_l/kT)}. \tag{62.4}$$

Given this functional dependence, the relative population at the vibrational level, whose energy is E_{CTS}, is obtained by substituting into the Boltzmann equation. Figure 62.4 illustrates this distribution qualitatively at low and high temperatures. At sufficiently high temperatures, governed by $e^{-(E_{\text{CTS}}/kT)}$, a significant population exists at an energy E_{CTS}. κ_{nonrad} is therefore proportional to this exponential factor. The constant of proportionality is A, a rate constant typically on the order of 10^{10} or 10^{11} transitions per second. It is related physically to the period of vibration in the lattice and the time it takes for the electron to physically move from the nonmetal ion to the europium ion. We now have

$$\kappa = \frac{1}{\tau} = \frac{1}{\tau_0} + Ae^{-(E_{\text{CTS}}/kT)} \tag{62.5}$$

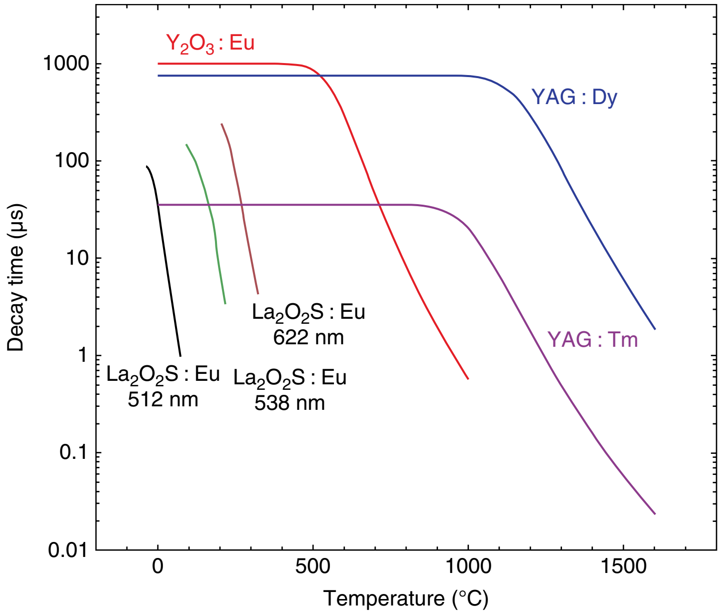


FIGURE 62.5 Fluorescent lifetime versus temperature for various phosphor materials.

or

$$\tau = \frac{\tau_0}{\left(1 + \tau_0 A e^{-(E_{\mathrm{CTS}}/kT)}\right)}. \tag{62.6}$$

As a final result, this expression describes the temperature dependence of fluorescence lifetime for this model.

In order for the lifetime measurements to be a practical indicator of temperature, it is desirable for the lifetime to have a linear or a simple logarithmic dependence on temperature. A semilogarithmic plot of models based on experimental data for lifetime τ versus temperature, T, for several TPs is shown in Figure 62.5. To select Y_2O_3 : Eu, for example, note that in the temperature range from ambient to about 600°C, the logarithm of lifetime is nearly constant with temperature. The logarithmic response permits calibration of the technique with a simple linear relationship over that temperature range.

Returning to our consideration of La_2O_2S : Eu, the functional dependence, for three of the characteristic electronic transition bands in that phosphor, of the lifetime on temperature, illustrated in Figure 62.6, is in agreement with that predicted by Equation 62.6. This agreement and the linear dependence with temperature over the observed range of temperatures can be understood from a simple analysis of Equation 62.6 and

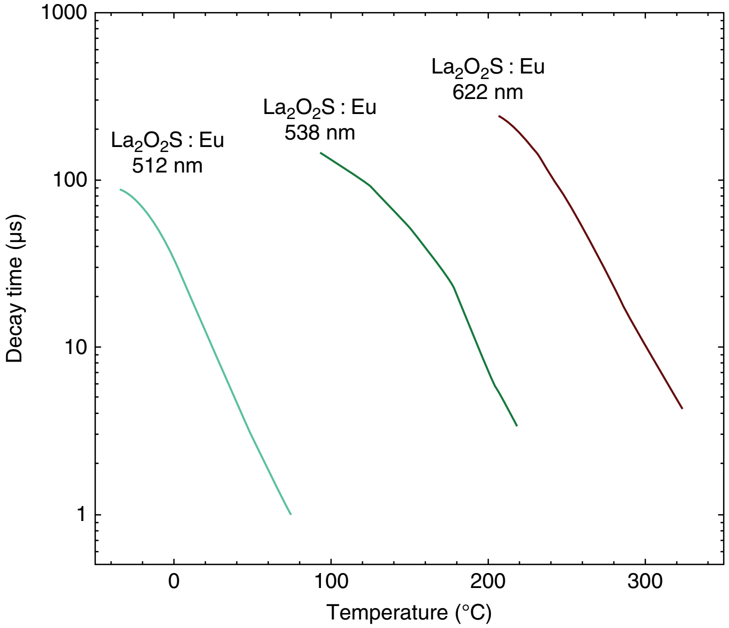


FIGURE 62.6 Decay time (lifetime) versus temperature for three characteristic electronic transition bands in $La_2O_2S:Eu$.

the following considerations. If the natural logarithms of both sides of Equation 62.6 are taken, Equation 62.6 becomes

$$\ln(\tau) = \ln(\tau_0) - \ln\left(1 + \tau_0 A e^{-(E_{\text{CTS}}/kT)}\right). \tag{62.7}$$

Different cases represent three temperature ranges and can help explain that this equation models the experimental data in Figure 62.7 and that the logarithms of the lifetimes have a linear dependence on temperature over a restricted range of temperatures. These three ranges are (i) the range for small temperatures, $E_{\text{CTS}}/kT > 1$; (ii) the range for high temperatures, $E_{\text{CTS}}/kT < 1$; and (iii) the midrange temperatures where the response is useful for making measurements, $E_{\text{CTS}}/kT \cong 1$. These limiting cases are discussed in this order below.

CASE 1 $E_{\text{CTS}}/kT > 1$

For the case of small temperatures, $-(E_{\text{CTS}}/kT)$ is a large negative number so that $e^{-(E_{\text{CTS}}/kT)}$ is a number near zero. In the limit as the temperature, T, approaches zero, $\tau_0 A e^{-(E_{\text{CTS}}/kT)}$ becomes negligible compared to 1, and the value of $\ln\left(1 + \tau_0 A e^{-(E_{\text{CTS}}/kT)}\right)$ approaches zero since $\ln(1) = 0$. Therefore, for the range of small values of temperature, $\ln(\tau) = \ln(\tau_0)$ and the lifetime is a constant, τ_0, as is observed in Figure 62.7.

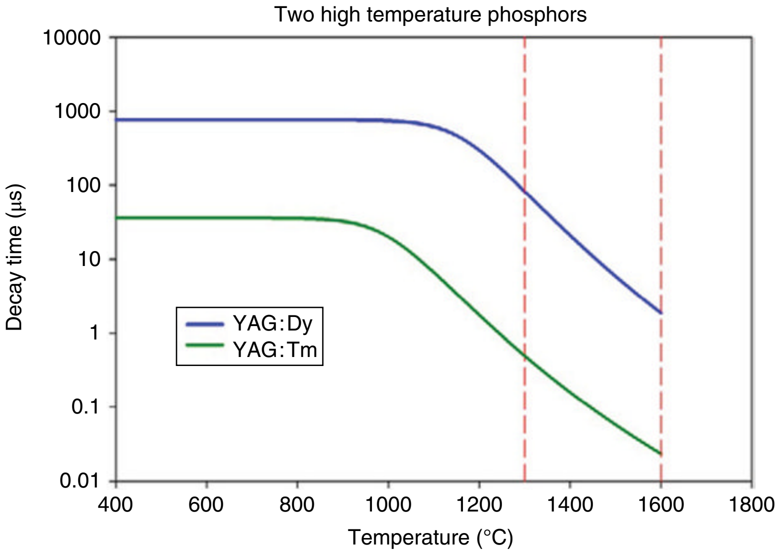


FIGURE 62.7 Temperature dependence of YAG : Dy and YAG : Tm.

CASE 2 $E_{\text{CTS}}/kT \lt 1$

In the high temperature range, E_{CTS}/kT becomes a small number approaching zero and $e^{-(E_{\text{CTS}}/kT)}$ approaches 1 so that Equation 62.6 gives $\tau = \tau_0/(1+\tau_0 A)$, a constant value.

CASE 3 $E_{\text{CTS}}/kT \cong 1$

Since the lifetimes can decrease over four decades, this implies $\tau_0 A$ to be large compared to 1 and that there is a range of values for T in which $\tau_0 A e^{-(E_{\text{CTS}}/kT)}$ is greater than 1. In this range, Equation 62.7 can be approximated by

$$\ln(\tau) = \ln(\tau_0) - \ln\left(\tau_0 A e^{-(E_{\text{CTS}}/kT)}\right). \tag{62.8}$$

or

$$\ln(\tau) = \ln(\tau_0) - \ln(\tau_0) - \ln(A) + \left(\frac{E_{\text{CTS}}}{kT}\right). \tag{62.9}$$

This equation is of the form

$$\ln(\tau) = B + \left(\frac{E_{\text{CTS}}}{kT}\right). \tag{62.10}$$

The linear relationship observed in Figure 62.7 only holds for small changes in temperature, ΔT, about some large temperature T_0. Since the temperature is in degrees

Kelvin, T_0 may be on the order of 1150 K and ΔT may vary ±250 K, as is the case in the example shown in Figure 62.7. As a result, the temperature T can be expressed as $T = T_0 + \Delta T$, and Equation 62.10 may be expanded to yield

$$\ln(\tau) = B + \frac{E_{\text{CTS}}}{k(T_0 + \Delta T)} = B + \frac{E_{\text{CTS}}}{k}(T_0 + \Delta T)^{-1} \tag{62.11}$$

or

$$\ln(\tau) = B + \frac{E_{\text{CTS}}}{kT_0}\left(1 + \frac{\Delta T}{T_0}\right)^{-1}. \tag{62.12}$$

$$\ln(\tau) = B + \frac{E_{\text{CTS}}}{kT_0}\left(1 - \frac{\Delta T}{T_0}\right). \tag{62.13}$$

$$\ln(\tau) = B' + C'\Delta T. \tag{62.14}$$

where $B' = B + (E_{\text{CTS}}/kT_0)$ and $C' = E_{\text{CTS}}/kT_0^2$. This illustrates that the logarithm of the lifetime decreases linearly with small changes in temperature about some moderately large value of temperature.

62.4 LABORATORY CALIBRATION OF TP SYSTEMS

The correlation of TP emissions with temperature requires calibration measurements that associate particular emission properties with the temperature of the TP. This calibration is required since the TP molecular configurations are far too complex in their quantum behavior to allow a theoretical calculation of correspondence with adequate accuracy. However, because these molecular systems, once calibrated, continue to respond with statistical consistency as long as the molecular configuration is not compromised, TP measurement systems require no recalibration or signal drift analysis.

A schematic depiction of a typical TP measurement arrangement is shown in Figure 62.8 (Schematic of a typical TP measurement system). There are three basic components: (i) an interrogation source (UV laser) to stimulate the fluorescence, (ii) a detection system to sense the fluorescence, and (iii) a data analysis system to convert the detector signal to temperature and to estimate accuracy and precision.

Often, TP methods are important to use for very-high-temperature systems. One TP that has been studied for high-temperature use is $Y_3Al_5O_{12}$:Dy (YAG:Dy) [8]. Its characteristic emission spectrum from ultraviolet stimulation is shown in Figure 62.9. Note the emission bands between 400 and 600 nm; these are the bands typically used for high temperature correlation.

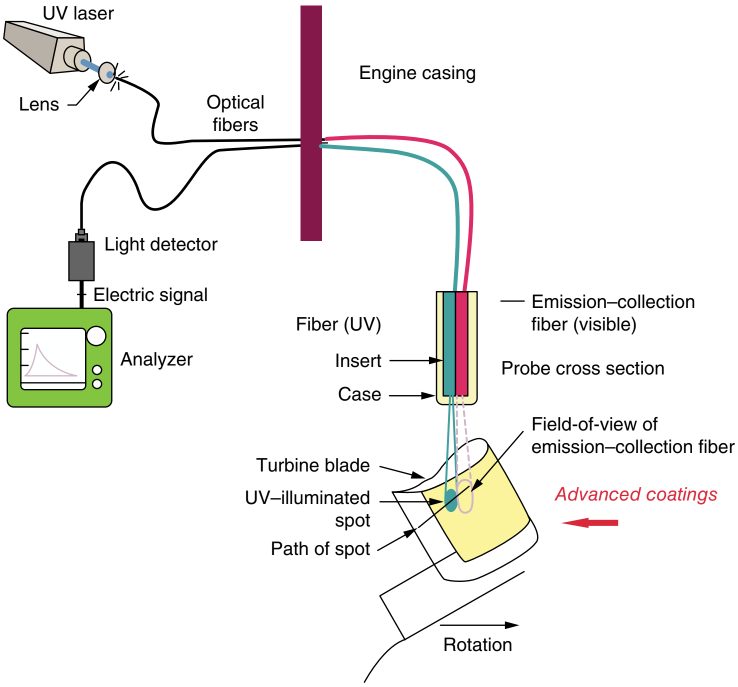


FIGURE 62.8 Schematic of a typical TP measurement system.

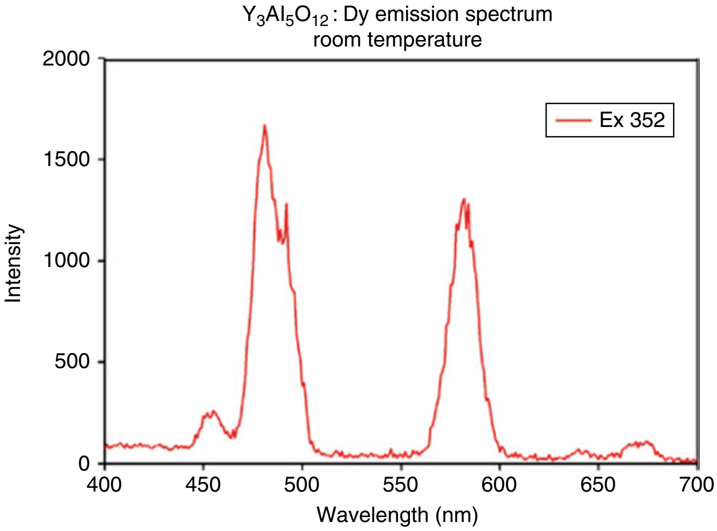


FIGURE 62.9 YAG : Dy emission spectrum.

Figure 62.10 is shown a calibration arrangement to determine the temperature correspondence of the emission of YAG : Dy, a phosphor with excellent high-temperature response. In this arrangement the tripled energy of a YAG laser was used, producing an interrogation wavelength of 355 nm. This beam was optically routed into

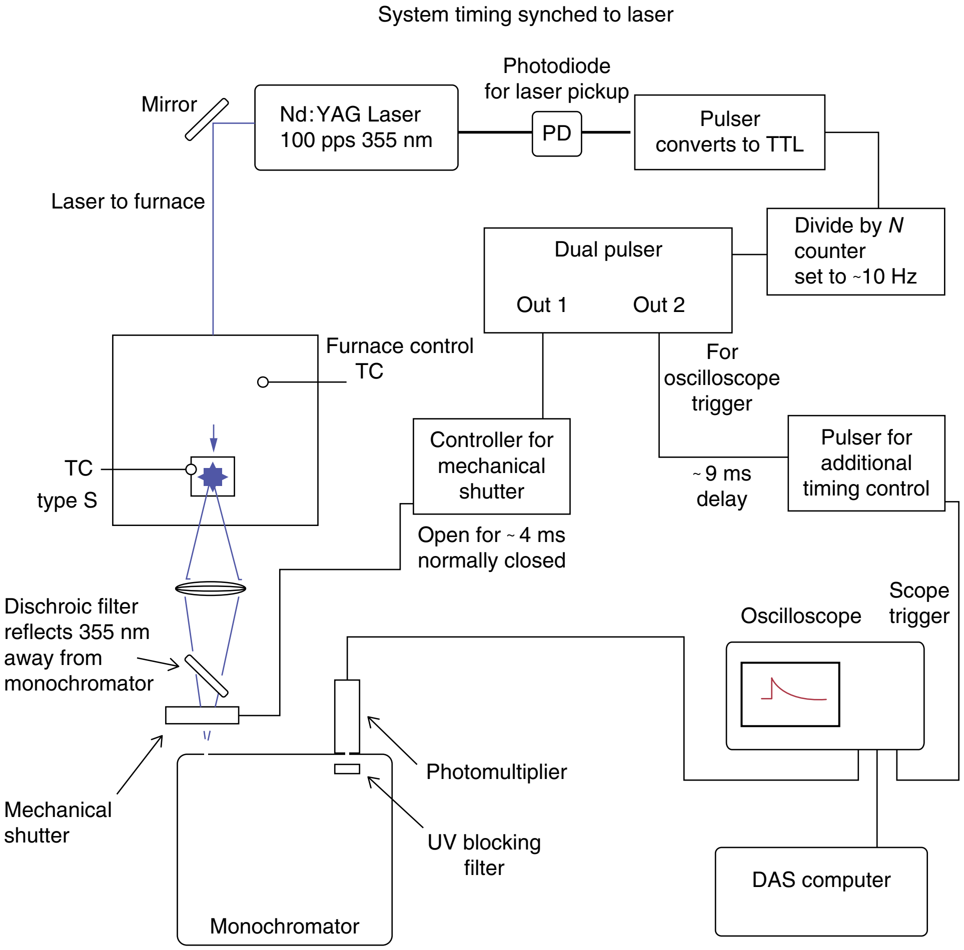


FIGURE 62.10 High-temperature calibration arrangement.

a high-temperature oven, which contained the phosphor sample coated on a ceramic surface. The fluorescent signals were separated optically from the interrogation source, passed through a narrowband filter to collect the appropriate temperature-sensitive emission, and measured by a photomultiplier tube with high gain.

The calibration was done by varying the oven temperature and measuring the phosphor sample temperature with a thermocouple whose junction was in contact with the phosphor surface. Those temperatures were then linked to fluorescent lifetimes, which were measured by repeated laser pulses and signal averaging. Although the blackbody emission background of the sample was minimized by the optical arrangement, those background levels had to be subtracted from the fluorescent signal before the lifetime was determined. With this formulation of YAG : Dy thus calibrated, it could then be applied to various surfaces and measured, yielding the temperature of those surfaces. The a fit of the data for YAG : Dy and YAG : Tm is shown in Figure 62.7.

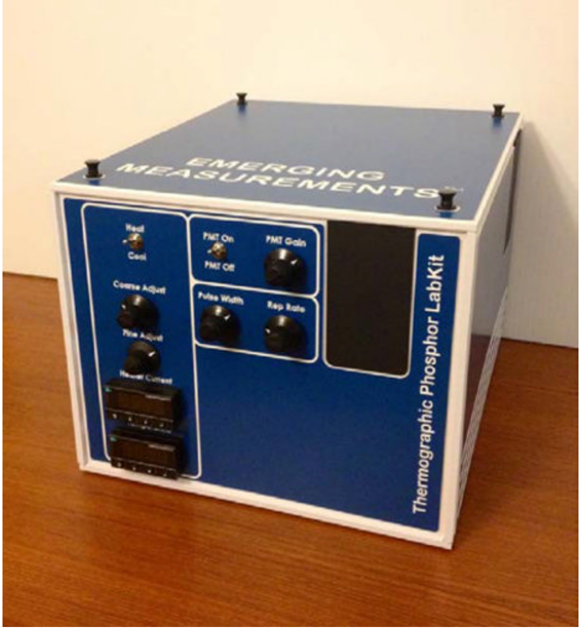


FIGURE 62.11 EMCO thermographic phosphor LabKit.

Other calibration arrangements are much simpler and often use less intense interrogation sources such as LEDs or laser diodes. This more complex arrangement, however, illustrates the feasibility of developing TP systems for very high-temperature measurement in adverse conditions.

For calibrations near room temperature, and to clearly illustrate the PT method, EMCO has developed a Phosphor LabKit containing all the measurement components and including the basic software required to convert fluorescent signals to temperature. Figure 62.11 is a photograph of the LabKit. Figure 62.12 shows how the signal changes versus temperature. This plot shows the decay only. Each curve is the result of 128 averages. Further information can be obtained by contacting EMCO directly.

62.5 HISTORY OF PHOSPHOR THERMOMETRY

While it has been known for many decades [1] that fluorescent materials have temperature dependence, one of the first major examples of its use in practical non-contact applications was developed in the late 1980s at Oak Ridge, Tennessee's Gaseous Diffusion Plant, K-25, to measure the temperature of rotors spinning at high speed to separate isotopes of uranium. Other uses of the technology were quickly added, and many further development studies and measurements involving TPs were done.

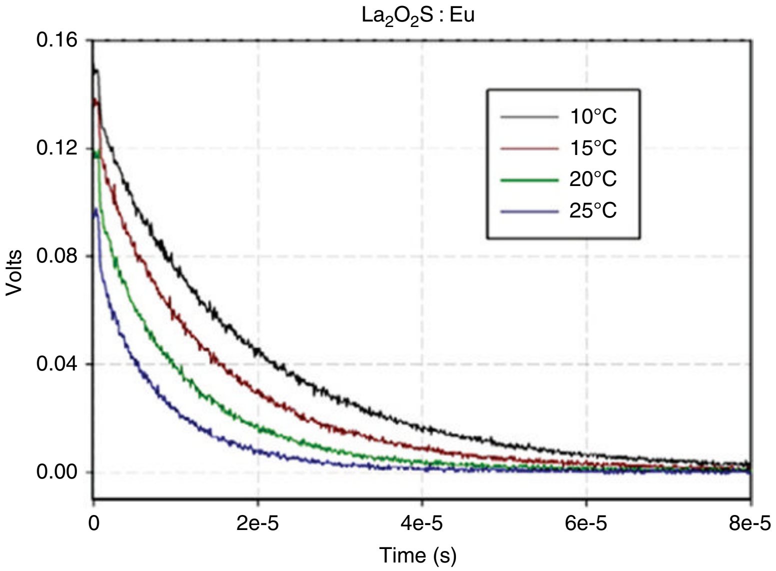


FIGURE 62.12 LED-excited fluorescence versus temperature.

62.6 REPRESENTATIVE MEASUREMENT APPLICATIONS

In this section we will discuss two specific measurement applications to illustrate the utility of TPs in various scenarios where standard temperature measurement approaches would be ineffective or extremely difficult and prone to error. The first is the use of optical fibers and TPs to measure the temperature of a spinning high-speed motor with permanent magnets in the rotor. The second is a temperature study of operating jet turbine components in a measurement test stand.

62.6.1 Permanent Magnet Rotor Measurement

Permanent magnet motors are especially useful for generating high rotational speeds and for being amenable to digital control systems that vary the motor operating parameters. Figure 62.13 is a drawing of the rotor assembly and the optical fiber arrangement used to provide access to the region between the spinning rotor containing permanent magnets and stator below. The underside of the rotor was coated with La_2O_2S:Eu mixed into an epoxy binder. The radial band of phosphor allowed temperature to be measured all around the rotor during operation. Two optical fibers were placed side-by-side, one transporting short pulses from a 337 nm nitrogen gas laser, the other transporting the resultant fluorescence to a photomultiplier filtered to select the appropriate emission band. In this case the filter passed light at about 510 ± 10 nm. The gap between rotor and stator was only about 2 mm, but it was adequate for the fiber optical arrangement.

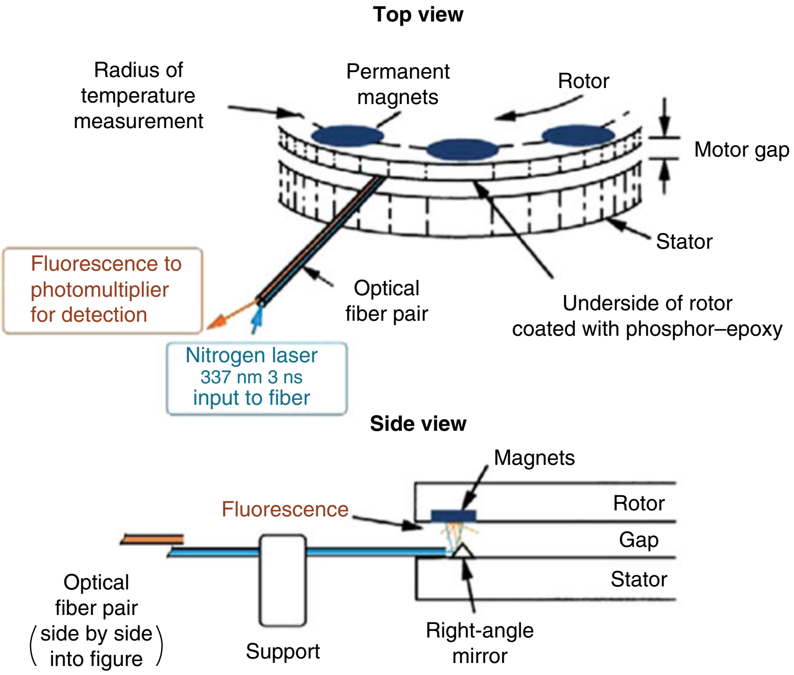

Top view
Radius of temperature measurement
Permanent magnets
Rotor
Motor gap
Stator
Fluorescence to photomultiplier for detection
Optical fiber pair
Underside of rotor coated with phosphor–epoxy
Nitrogen laser 337 nm 3 ns input to fiber
Side view
Magnets
Fluorescence
Rotor
Gap
Stator
Optical fiber pair (side by side into figure)
Support
Right-angle mirror

FIGURE 62.13 Permanent magnet motor experimental arrangement.

To make the measurement it was necessary to pulse the laser several hundred times at the same spot on the rotor in order to accurately determine the temperature of that spot. Consequently, the laser was triggered by a circuit that monitored the rotational speed by reflectance from a reference position, the axial position determined by the fraction of the time between reference pulses. That position was determined by a delay time built into the triggering circuit, and when that delay time was changed the entire rotor surface could be measured in small axial steps. Figure 62.14 shows some typical data for the measurement, data revealing the amount of heating of the permanent magnets for particular operating parameters. The two temperature traces were obtained at different times for the same magnet and shows the influence of eddy current heating.

62.6.2 Turbine Engine Component Measurement

An important TP application has been for measurement of temperature in the hot zones of turbine engines and generators. These zones reach temperatures as high as 1300°C and, are difficult to access, and full of reflected light from the combustion process.

Tables 62.1 and 62.2 show a summary of applications of phosphor thermometry over the history of its use pursuant to turbine and other high-temperature applications. The tables are not intended to be exhaustive rather to be an illustration of the many potential uses of the method.

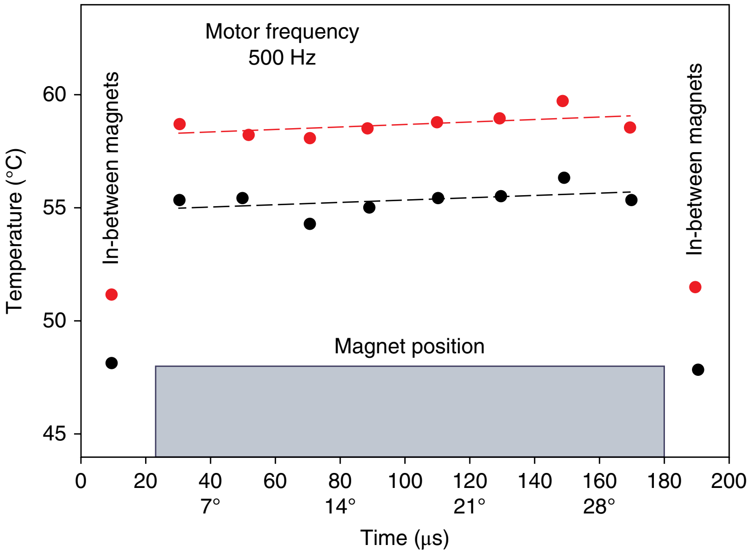


FIGURE 62.14 Heating of magnets in motor.

TABLE 62.1 Various Measurements

Lab Measurements with Flame/Combustion		High Temperatures (Stationary/Slow Surfaces)		High-Speed Surfaces	
Burning wood [9]	840°F	Many tests	To 3100°F	Permanent magnet motor [10]	750 Hz 200°F
Intumescent surfaces [11]	1100°F	Galvanneal steel processing [12]	1300°F	Rail gun armatures [13]	400 m/s 200°F
Pistons intake valves [14]	400°F				

TABLE 62.2 Combustion-Related Measurements

High Temperature and with Combustion	
Burner rig (slow speed rotating) [15]	1650°F
Turbine engine vane [16]	1750°F
Turbine engine afterburner nozzle [17]	1300°F
Turbine engine afterburner flame holder [18]	1100°F

62.7 TWO-DIMENSIONAL AND TIME-DEPENDENT TEMPERATURE MEASUREMENT

TPs can be also be used for complex temperature measurements, such as area temperature mapping and monitoring the temperature of a zone as a function of time. For area measurement the ratio method is often recommended. This approach uses two

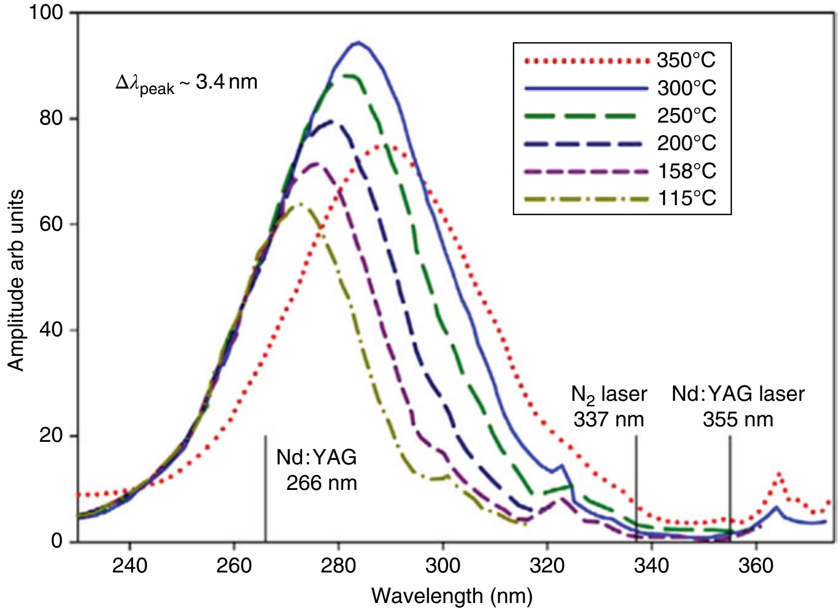


FIGURE 62.15 Excitation band of Y_2O_3 : Eu from 115 to 350°C.

wavelengths of fluorescence measured with a video system, with one image divided by the other. Such arrangements can be calibrated ahead of time and can often use broadband filtering to allow significant light for each image. To increase the measurement sensitivity, image intensifiers can be used for some situations. Another approach is to scan the surface of interest with a one-dimensional system, building the image over multiple scans.

For time-dependent measurements, too, there are various approaches. We discuss a few of them here. A straightforward one is simply to measure fluorescent lifetime as quickly as possible, averaging the measurements in at least one of two ways: extracting temperature from each pulse or maintaining a running average over selected time periods. In any case, the one-sigma limit of time resolution, sometimes called the Sartori limit, is approximately three times the average lifetime measured (Dr. Walter Sartori, personal communication, ca. 1982). Time-dependent temperatures can also be measured by monitoring the intensity of emissions associated with a ratio measurement. The ratio, in this case, would vary as a function of time, with its time resolution associated with the characteristic lifetime of the wavelengths measured. Another approach would take advantage of the shift of an excitation spectrum for a particular phosphor as a function of temperature. Figure 62.15 shows the excitation spectra for Y_2O_3 : Eu at various temperatures for emission in one of strong emission bands. As temperature changes, the strength of the excitation changes. Thus, by measuring a spectrum or the intensity change at a given wavelength, a measurement of temperature versus time can be extracted.

62.8 CONCLUSION

In this article we've discussed the use of fluorescent materials to measure temperature. We have included a theoretical description of the method for certain typical types of TPs and have presented several illustrations to make clear the utility of this powerful physical measurement technology.

REFERENCES

1. S. W. Allison and G. T. Gillies, "Remote thermometry with thermographic phosphors: Instrumentation and applications," *Review of Scientific Instruments*, 68, 2615–2650 (1997).
2. C. Knappe, J. Lindén, F. Abou Nada, M. Richter, and M. Aldén, "Investigation and compensation of the nonlinear response in photomultiplier tubes for quantitative single-shot measurements," *Review of Scientific Instruments*, 83, 034901 (2012).
3. A. H. Khalid and K. Kontis, "Thermographic phosphors for high temperature measurements: Principles, current state of the art and recent applications," *Sensors*, 8, 5673–5744 (2008).
4. M. D. Chambers and D. R. Clarke, "Doped oxides for high-temperature luminescence and lifetime thermometry," *Annual Review of Materials Research*, 39, 325–359 (2009).
5. M. Alden, A. Omrane, M. Richter, and G. Sarner, "Thermographic phosphors for thermometry: A survey of combustion applications," *Progress in Energy and Combustion Science*, 37, 422–461 (2011).
6. J. Brübach, C. Pflitsch, A. Dreizler, and B. Atakan, "On surface temperature measurements with thermographic phosphors: A review," *Progress in Energy and Combustion Science*, 39, 37–60 (2013).
7. W. H. Fonger and C. W. Struck, "Eu^{+3} ^{5}D resonance quenching to the charge-transfer states inY_2O_2S, La_2O_2S, and LaOCl," *Journal of Chemical Physics*, 52(12), 6365 (1970).
8. M. R. Cates, S. W. Allison, S. L. Jaiswal, and D. L. Beshears, "YAG : Dy and YAG : Tm Fluorescence to 1700°C," Proceedings of the 49th International Instrumentation Symposium of the ISA (The International Society of Instrumentation, Systems, and Automation), ISA Vol. 443, Orlando, FL, May 4–8, 2003.
9. A. Omrane, *Thermometry Using Laser-Induced Emission from Thermographic Phosphors: Development and Applications in Combustion*, PhD Dissertation, Lund University, Lund, Sweden, 2005.
10. S. W. Allison, G. T. Gillies, M. R. Cates, and B. W. Noel, "Method for monitoring permanent magnet motor heating with thermographic phosphors," *IEEE Transactions on Instrumentation and Measurement*, 37(4), 637–641 (1988).
11. A. Omrane, Y. C. Wang, U. Göransson, G. Holmstedt, and M. Aldén, "Intumescent coating surface temperature measurement in a cone calorimeter using laser-induced phosphorescence," *Fire Safety Journal*, 42(1), 68–74 (2007).
12. W. W. Manges, S. W. Allison, and J. R. Vehec, "Galvanneal Thermometry with a Thermographic Phosphor System," 1997 AISE Annual Convention and Iron and Steel Exposition, Cleveland, OH, September 29–October 2, 1997.

13. S. W. Allison, M. R. Cates, S. M. Goedeke, A. Akerman, M. T. Crawford, S. B. Ferraro, J. Stewart, and D. Surls, "In-flight armature diagnostics," *IEEE Transactions on Magnetics*, 43(1), 329–333 (2007). (Part II of two parts of Selected Papers from the 13th International Symposium on Electromagnetic Launch (EML) Technology, Berlin, Germany, May 22–25, 2006.)

14. J. S. Armfield, N. Domingo, J. M. Storey, S. W. Allison, D. L. Beshears, and M. R. Cates, "Powertrain Component Temperature Measurements via Phosphor Thermometry," Presented at the World Car Conference, Ref. No. 97WCC018, Riverside, CA, January 19–22, 1997.

15. K. W. Tobin, S. W. Allison, M. R. Cates, G. J. Capps, D. L. Beshears, and M. Cyr, "Remote High-Temperature Thermometry of Rotating Test Blades Using YVO_4:Eu and Y_2O_3:Eu Thermographic Phosphors," AIAA/ASME/SAE/ASEE Proceedings of the 24th Joint Propulsion Conference, AIAA-88-3147, Boston, MA, July 11–13, 1988.

16. B. W. Noel, W. D. Turley, and S. W. Allison, "Thermographic-Phosphor Temperature Measurements: Commercial and Defense-Related Applications," Proceedings of the 40th International Instrumentation Symposium of the ISA, Baltimore, MD, May 1–5, 1994.

17. H. Seyfried, G. Särner, A. Omrane, M. Richter, H. Schmidt, and M. Aldén, "Optical Diagnostics for Characterization of a Full-Size Fighter-Jet Afterburner," Proceedings of the ASME Turbo Expo, Vol. 1, pp. 813–819. ASME Turbo Expo 2005—Gas Turbine Technology: Focus for the Future, Reno-Tahoe, NV: ASME Press.

18. H. Seyfried, M. Richter, M. Aldén, and H. Schmidt, "Laser-induced phosphorescence for surface thermometry in the afterburner of an aircraft engine," *AIAA Journal*, 45(12), 2966–2971 (2007).

63

VOLTAGE AND CURRENT TRANSDUCERS FOR POWER SYSTEMS

Carlo Muscas and Nicola Locci

Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

63.1 INTRODUCTION

Control, management, monitoring, and protection applications on electrical power systems require continuous and trustworthy knowledge of the two fundamental electrical quantities, namely, current and voltage. By considering the paramount importance of these actions, in terms of both safety and economic impacts, when designing a measurement system for such applications, many metrological aspects should be considered carefully so that the overall performance of the system guarantees that the desired parameters are measured with sufficient accuracy.

It is widely acknowledged that, owing to different causes, including faults, power quality (PQ) disturbances, operation of network components, etc., voltages and currents in modern power grids do not usually meet the ideal conditions of rated frequency, rated amplitude, sinusoidal waveform, and positive symmetry in three-phase systems.

All these electrical phenomena, which may have characteristics quite different from each other in terms of amplitude, duration, repetition, and effects on the system elements, as well as perception from users, are nowadays measured by means of digital programmable systems, be they intended for control, management, monitoring, or protection purposes. Digital measurement systems typically receive as input only voltages in the range of a few volts or at most a few tens of volts. Therefore, voltage

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

and current transducers are needed to convert the electric quantities involved in power grids, either currents or higher voltages, into low voltages suitable for the successive measurement steps.

The main functions of the measurement transducers used in electric power systems can be summarized as follows:

- Adapting the level and/or nature of the primary quantity to a value and/or nature that can be dealt with by measurement instrumentation or protection relays
- Ensuring safety of measurement and protection systems, by either insulating such systems from the power grid or using alternative solutions

In power plants, the most commonly used transducers for voltages and currents are the magnetic core instrument transformers, the current transformer (CT) and the voltage transformer (VT), sometimes referred to as potential transformer (PT). Figure 63.1 shows an example of insertion in a single-phase line operating at voltage V_1 where the current I_1 flows. One terminal of the secondary circuit is usually connected to ground for safety reasons.

Commercial VTs and CTs are usually designed to operate with 50 or 60 Hz (depending on the rated frequency of the system) sinusoidal primary quantities, and their accuracy specifications are defined with respect to these rated conditions. In the presence of either distorted waveforms, and thus of harmonics and interharmonics, or transient events, their performance may significantly decay.

Furthermore, it should be considered that, for historical reasons, the rated output value of these transducers is generally 100 V for the VT and 5 A (or 1 A) for the CT, in order to have sufficient energy to drive electromechanical measurement or protection devices.

On the other hand, modern digital measurement systems and electronic relays have low input power requirement. This leads to the need to introduce low-power current transformers (LPCTs) and low-power voltage transformers (LPVTs), which supply a low-voltage signal (from tens of millivolts to a few volts) as secondary output.

To comply with the above requirements, there are on the market voltage and current transducers, for either laboratory or field applications, based on different principles of operation, that can guarantee good accuracy and a wide measuring range, as well as

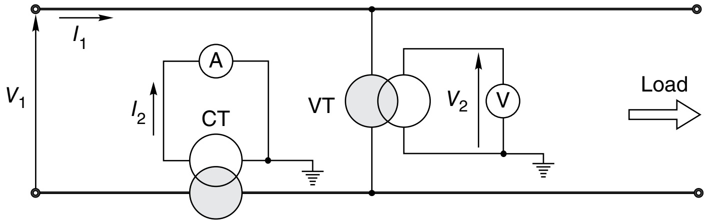


FIGURE 63.1 Insertion of instrument transformers.

a bandwidth extended from a few hertz (or even DC, in some cases) up to the megahertz region. Furthermore, the kind and level of their output quantity are generally compatible with the input of data acquisition systems, and their weight and size are usually lower than for the conventional instrument transformers.

By considering that transducers are often the main source of uncertainty in the entire measurement chain, this chapter is intended to provide the reader with the basic elements to understand the principle of operation and the behavior of the most common transducers for voltage and current in power systems, to assess their metrological characteristics, and finally, to choose the most appropriate devices to be used in the different measurement applications.

Note: In the technical literature about measurement systems, the terms *transducer* and *sensor* are often used as synonymous to define a device that transforms an input physical quantity into a different quantity, with different values and/or different characteristics, which are more suitable for the measurement instrument/system connected at the output of the device itself. In this chapter it has been chosen, only for the sake of clarity and without loss of generality, to use the term transducer for the complete device, which includes all the auxiliary components (power supply, amplifiers, filters, etc.) required to make practically feasible the conversion, while the term sensor will be used sometimes to indicate more specifically the physical "sensing" element, that is, where the physical principle of the conversion acts.

63.2 CHARACTERIZATION OF VOLTAGE AND CURRENT TRANSDUCERS

Choosing the right transducer is one of the most important steps in the design of the entire measurement system. Such choice can be done properly only if the performance of the devices available on the market is known. It is therefore important that manufacturers provide suitable information about the metrological behavior of these devices.

International standards on instrument transformers (e.g., the IEC 61869 series, which will be considered as a reference in the following) provide, for the sinusoidal conditions, the following definitions (in brackets the symbols employed in this chapter for the defined quantities):

- **Actual transformation ratio** (K): the ratio of the actual primary quantity to the actual secondary quantity.
- **Rated transformation ratio** (K_r): the ratio of the rated primary quantity to the rated secondary quantity.
- **Ratio error** (η): the error that a transformer introduces into the measurement of a voltage/current and that arises from the fact that the actual transformation ratio is not equal to the rated transformation ratio.

- **Phase displacement** (ε): The difference in phase between the primary and secondary current vectors, the direction of the vectors being so chosen that the angle is zero for a perfect transformer. It is usually expressed in minutes or centiradians (1 crad = 10^{-2} rad).

Under sinusoidal conditions, if we define **Q** as the phasor of a generic quantity and use the subscripts in and out for transducer's input and output quantities, respectively, it is

$$\mathbf{Q}_{\text{out}} = \mathbf{Q}_{\text{in}} \frac{1}{K_{\text{r}}}\left(1+\eta\right)\cdot e^{j\varepsilon} \tag{63.1}$$

By considering the modulus of both members in the earlier equation, we obtain

$$\eta = \frac{K_{\text{r}}Q_{\text{out}} - Q_{\text{in}}}{Q_{\text{in}}} = \frac{K_{\text{r}}Q_{\text{out}} - KQ_{\text{out}}}{KQ_{\text{out}}} = \frac{K_{\text{r}} - K}{K} \cong \frac{K_{\text{r}} - K}{K_{\text{r}}} \tag{63.2}$$

where Q_{in} and Q_{out} are the root-mean-square (rms) values of the input and output quantities of the transducer. η and ε are defined in steady-state sinusoidal conditions, and their value depends on both the frequency and the amplitude of the voltage (or current) on the primary side.

When measuring the distorted quantities existing in modern power systems, other characteristics of the transducers have an importance that is equal to, or sometimes greater than, the previous parameters. The most significant of these characteristics are linearity and bandwidth.

The importance of linearity, which, in this context, mainly represents the ability to ensure sufficient accuracy in a wide measuring range, is evident since it is desirable, especially in protection applications, that the transducer performance does not change even for large variations of the input quantity.

On the other hand, it is clear, by taking into account the characteristics of the distorted signals to be measured, that it is of great interest defining the frequency range for which the uncertainty stays below prefixed limits. To this purpose, the transducer bandwidth is usually defined as the frequency range where the transformation ratio does not differ from its rated value of more than 3 dB (i.e., ~30%).

63.3 INSTRUMENT TRANSFORMERS

63.3.1 Theoretical Fundamentals and Characteristics

Instrument transformers (VTs or PTs and CTs) are extensively used in production, transmission, distribution, and utilization of electricity, in association with measuring instruments, meters, and protective or control devices.

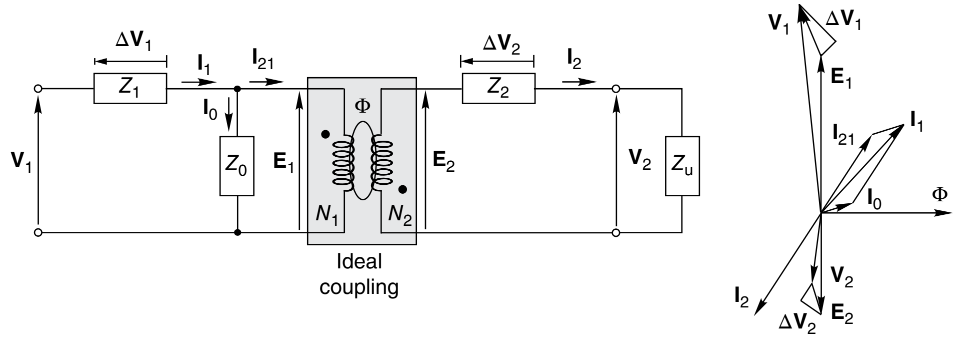


FIGURE 63.2 Equivalent circuit and phasor diagram of a transformer.

Voltage and current instrument transformers consist of a magnetic core on which two windings are wounded. The primary and secondary windings have N_1 and N_2 turns, respectively, and are electrically insulated from the core and from each other. Alternatively, the primary circuit of a current transformer can simply consist of the single conductor where the primary current flows. The secondary circuit is loaded by the secondary burden determined by measurement and/or protection devices.

Instrument transformers are characterized by their rated ratio, that is, the ratio between rated input and output quantities:

$$\text{CT}: \quad K_{Ir} = \frac{I_{1r}}{I_{2r}} \qquad \text{VT}: \quad K_{Vr} = \frac{V_{1r}}{V_{2r}} \tag{63.3}$$

As an example, the rated ratio could be $K_{Ir} = 200{:}5$ (A/A) for a CT and $K_{Vr} = 20.000{:}100$ (V/V) for a VT. An ideal instrument transformer should reduce the amplitude of the input signal according to the rated ratio and let the phases unchanged. Actually, this does not happen, for several reasons, which can be shortly explained by recalling the equivalent circuit of a transformer, shown in Figure 63.2 along with the phasor diagram of the involved electrical quantities.

The behavior of the ideal transformer is summarized by the following relationships between vectors:

$$\mathbf{I}_{21} = -\frac{N_2}{N_1}\mathbf{I}_2 \qquad \mathbf{E}_1 = -\frac{N_1}{N_2}\mathbf{E}_2 \tag{63.4}$$

It is common practice to define a further equivalent circuit of the transformer (Fig. 63.3), where all the quantities, including the ones related to the secondary circuit, such as the impedances of the secondary winding and of the load, are reported to the primary one, by means of the following expressions:

$$\bar{Z}_{21} = \bar{Z}_2\left(\frac{N_1}{N_2}\right)^2 \qquad \mathbf{V}_{21} = -\mathbf{V}_2\frac{N_1}{N_2} \tag{63.5}$$

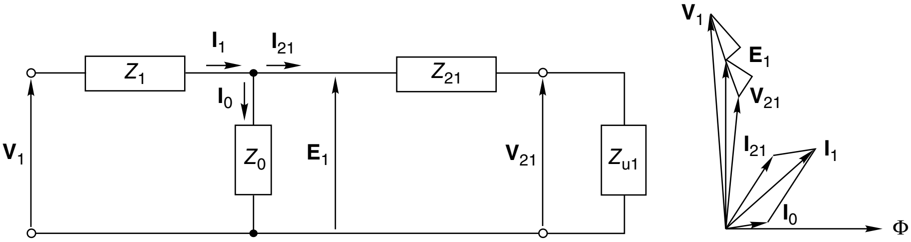


FIGURE 63.3 Equivalent circuit of the transformer referred to the primary side.

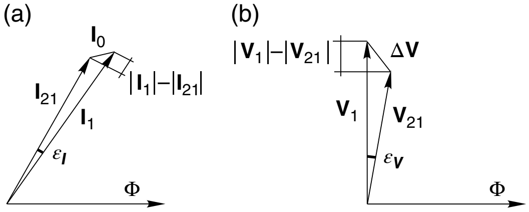


FIGURE 63.4 Phasor diagrams for CTs (a) and VTs (b).

In this way, no ideal transformer is included in the model, and all the quantities refer to the primary voltage V_1.

In this circuit a T-network is present, which is responsible for the different behavior of the real transformer with respect to the ideal one. In order to study these aspects, Figure 63.4 shows the phasor diagrams for CTs (Fig. 63.4a) and VTs (Fig. 63.4b), respectively, while Equation 63.6 shows the relevant mathematical relations:

$$\begin{aligned} &\text{CT}: \quad \mathbf{I}_1 = \mathbf{I}_{21} + \mathbf{I}_0 \cong \mathbf{I}_{21} = -K_{It}\mathbf{I}_2 \qquad \left(K_{It} = \frac{N_2}{N_1} > 1\right) \\ &\text{VT}: \quad \mathbf{V}_1 = \mathbf{V}_{21} + \Delta\mathbf{V} \cong \mathbf{V}_{21} = -K_{Vt}\mathbf{V}_2 \qquad \left(K_{Vt} = \frac{N_1}{N_2} > 1\right) \end{aligned} \qquad (63.6)$$

According to (63.6), an approximated value for the amplitude of the primary quantities (I_1 or V_1) can be achieved by multiplying the amplitude of the secondary ones (I_2 or V_2) by the turn ratio (K_{It} or K_{Vt}).

For a CT with a turn ratio $K_{It}=N_2/N_1$, the vector difference between the currents $\mathbf{I}_1$ and $\mathbf{I}_{21}$ corresponds to the exciting current $\mathbf{I}_0$ flowing in the magnetizing branch (see Fig. 63.4a). This difference is thus minimized if the measured current is much larger than the exciting current. To obtain this, besides a proper sizing of the magnetic circuit, the transformer should work close to short circuit conditions, or in other words, the load on the secondary circuit should be as low as possible.

In any case, even in the presence of a null secondary burden, the unavoidable presence of the exciting current would make it impossible to have the actual ratio K equal to the theoretical one K_t. For this reason, in the common praxis, the rated transformation ratio attributed to the CT is larger than the ratio between the numbers of turns (i.e., $K_{Ir} > N_2/N_1$) so that the natural ratio error is compensated at least partially. In such way, once the current $\mathbf{I}_2$ has been measured and multiplied for the rated constant K_{Ir}, the obtained value is a better approximation of the actual primary current $\mathbf{I}_1$. In the practice, once the desired rated transformation ratio for a CT has been established, the device is built with a turn ratio slightly smaller.

Analogously, for a VT with a turn ratio $K_{Vt} = N_1/N_2$, the vector difference between the voltages $\mathbf{V}_1$ and $\mathbf{V}_{21}$ corresponds to the voltage drop $\Delta\mathbf{V}$ across the series branches of the equivalent circuit (see Fig. 63.4b). To limit this voltage drop, besides a proper design of the windings, in order to the have low resistances and low leakage reactance, a very high impedance should be connected at the secondary side of the VT. However, even assuming an ideal condition of null secondary burden (open circuit), the unavoidable exciting current causes a voltage drop on the primary side, thus affecting both the amplitude and the phase of the actual primary voltage. The presence of a nonzero burden implies additional voltage drops on both sides of the transformer, thus making the actual ratio more different from K_t. This also implies that the actual ratio K varies for varying burden. The rated transformation ratio assigned to the VT, K_{Vr}, is therefore slightly larger than N_1/N_2 so that the voltage drop in the series branches is at least partially compensated.

Clearly, in both cases, this compensation, whose effects on the ratio error vary with the operative conditions, is definitely ineffective with respect to phase displacement.

The standards about instrument transformers usually define the accuracy of these devices in terms of precision class, which represents the maximum value allowed (in percent) for the ratio error defined in Section 63.2. For each class, the standards also define limits for the phase error. As an example, for a class 0.3 VT, according to standard IEC 61869-3, the maximum ratio error is $\eta = \pm 0.5\%$ and the maximum phase error is $\varepsilon = \pm 0.6$ crad. These maximum errors should be ensured for voltages included in the range 80–120% of the rated value (see Fig. 63.5), when the burden is between 25 and 100% of the rated value.

Similar specifications exist for CTs, but in this case different error limits are defined for different values of the input current to take into account the high variability of the current absorbed by both domestic and industrial loads. As an example, IEC 61869-2 considers a current range between 5 and 120% of the rated value (Fig. 63.5).

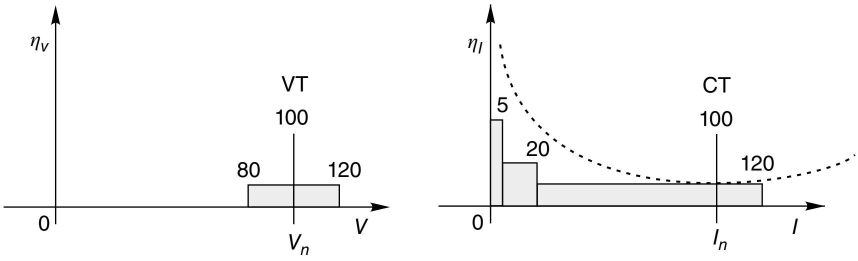


FIGURE 63.5 Error limits for VTs and CTs.

Ratio and phase errors in instrument transformers also vary when the load on the secondary side varies. From this point of view, a limit value for the secondary burden is defined in terms of the maximum apparent power that can be requested to the secondary circuit when the rated quantity is applied on the primary side. As an example, according to IEC 61869-3, for measuring VTs the preferred burden ranges are from 1 to 10 VA (with power factor of 1) or from 10 to 100 VA (with power factor of 0.8, lagging), while IEC 61869-2 specifies for measuring CTs a burden range from 2.5 to 30 VA.

As far as safety is concerned, if the secondary circuit of a CT was opened, all the primary current would become an exciting current (see Fig. 63.3), and the voltage between the terminals at the primary side could reach high values so that a risk for the operator or for the device itself could arise. For this reason, an overvoltage protection may be included in CTs. For dual reasons, overcurrent protection may be instead used in VTs to avoid the problems that could arise if the secondary terminals were connected to a low impedance.

63.3.2 Instrument Transformers for Protective Purposes

When instrument transformers are used with protection relays, their working conditions can be very different from the normal operation. Thus, their metrological characteristics should include some additional specifications.

In the case of VTs, the given accuracy should be ensured for voltage values much higher (typically from 1.2 to 1.9) than the rated one.

The case of current transformers should be analyzed more carefully. Indeed, one of the most common protections in electric grids is the one against overcurrent (arising from either short circuit or overload), usually performed through circuit breakers, whose operation is based on relays supplied by CTs. Thus, the correct and prompt intervention of the protection systems firstly depends on the behavior of the CT. During faults, the transient current is composed of a steady-state component superimposed to a DC decaying component, whose value depends on both the circuit parameters and the instant in which the faults begins. When this fault current is much higher (up to tens of times) than the CT rated current, then saturation occurs and the secondary current may be heavily distorted.

Figure 63.6 shows, for two qualitative examples, the primary current and the secondary one multiplied by the rated transformation ratio: in case (b) the saturation is higher than in case (a). These highly distorted waveforms of the secondary current may represent an issue for the correct operation of the protection relays.

In these situations the usual vector representation cannot be adopted. Thus, besides ratio and phase errors, a new error term can be introduced in the standards for protective CTs (e.g., IEC 61869-2), named the composite error (ε_c), which is usually provided in percent and defined according to the following expression:

$$\varepsilon_c = \frac{100}{I_1}\sqrt{\frac{1}{T}\int_0^T \left(K_{Ir} i_2 - i_1\right)^2 dt} \quad (\%) \tag{63.7}$$

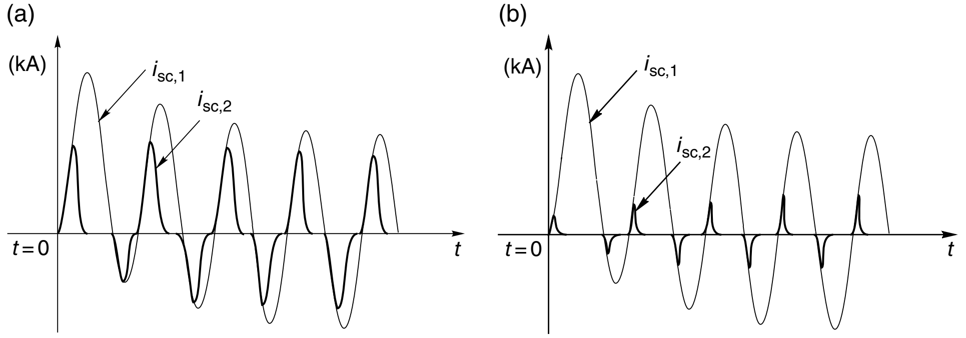


FIGURE 63.6 Examples of possible effects of saturation on the secondary current of a CT: in case (b) the saturation is higher than in case (a).

where i_1 and i_2 are the instantaneous values of the primary and secondary current, respectively, while T is the period of these quantities.

For instance, in a protective CT with a precision class 10P, the maximum allowed composite error is 10%, for an input maximum current that is much higher (typically from 5 to 30 times) than the rated current.

63.3.3 Instrument Transformers under Nonsinusoidal Conditions

In Section 63.1 it has been recalled that voltages and currents in power systems may be characterized by a waveform that is different from the sinusoidal one traditionally considered for such systems, due to both possible steady-state disturbances (e.g., harmonics or interharmonics) and events, such as voltage dips, rapid transients, etc.

Measurement transducers used to measure these quantities must guarantee sufficient accuracy while reproducing the distorted input waveform. As an example, when harmonic and interharmonic components are concerned, it is necessary to accurately transduce the amplitude of each harmonic but also, when power related terms have to be evaluated, its phase.

Magnetic core instrument transformers usually have a limited bandwidth. In order to explain this behavior, let us consider the equivalent circuit of Figure 63.7, which differs from the classical representation of a transformer at low frequency for the presence of the stray capacitances: C_1 and C_{21} refer to the capacitive coupling in the turns of the primary and secondary windings, respectively, while C_{ps} represents the capacitive coupling between the two windings.

At industrial frequency these capacitances are negligible, but their influence increases for increasing frequency. This leads to the possibility of creating resonant circuits, which would result in significant decay of the instrument accuracy.

This problem is particularly evident in instrument transformers for medium- and high-voltage systems. Indeed, in such systems the insulation requirements are stronger and impose the use of larger sizes, thus determining larger stray capacitances and leakage inductances to appear and causing the risk of resonance at lower frequencies.

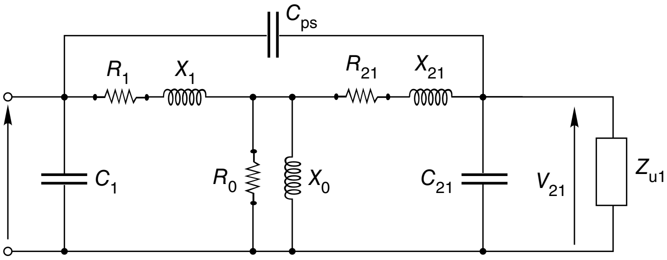


FIGURE 63.7 Equivalent circuit of the transformer for higher frequencies.

As a consequence, if no solutions are adopted to compensate such effects, instrument transformers (especially for medium- and high-voltage systems) are not suitable to measure voltages and currents with significant high-frequency components.

Generally, low-voltage CTs and VTs have a ratio error less than a few percent and phase displacement less than a few degrees for frequency up to a few thousand hertz. For medium-voltage VTs the same errors occur at frequencies lower than 1 kHz, while in high-voltage VTs this happens at about 500 Hz. CTs, owing to their different characteristics, have slightly better performance.

In any case, magnetic core transformers are not appropriate to measure quantities with DC components. Indeed, such components are not transferred to the secondary side, and in addition, they may lead to core saturation, thus affecting significantly also the measurement of alternate components.

63.3.4 Capacitive Voltage Transformer

Insulation issues impose the use of sufficient distances between the elements of a VT. When these devices are used for high voltages, insulation requirements determine large sizes and weights, thus implying difficulties in building them and producing significant impact on the costs, which can be predominant with respect to the accuracy needs. For the preceding reasons, in power systems operated at voltages higher than 150 kV, VTs are replaced by capacitive voltage transformers (CVTs). According to Figure 63.8, in such devices a capacitive divider (composed by capacitances C_1 and C_2) reduces the primary voltage $\mathbf{V}_{in}$ to an intermediate level $\mathbf{V}_{out,C}$, which is then applied to a transformer that, besides providing the required insulation, further reduces the voltage to the desired secondary value $\mathbf{V}_{out}$. The Thevenin equivalent circuit for the capacitive divider has the equivalent voltage $\mathbf{V}_{eq}=\mathbf{V}_{out,C}=\mathbf{V}_{in}\cdot C_1/(C_1+C_2)$ and the equivalent series capacitance $C_{eq}=(C_1+C_2)$.The voltage drop in the capacitive element C_{eq} can be compensated by an opposite voltage drop across a reactive inductance (totally or partially contained in the transformer), whose value makes the circuit resonant for the working frequency. Since these optimal compensation conditions are verified only for the rated frequency, the nominal accuracy of the CVT is ensured only for frequency variations in a very limited range (e.g., about ±0.5 Hz) around the rated

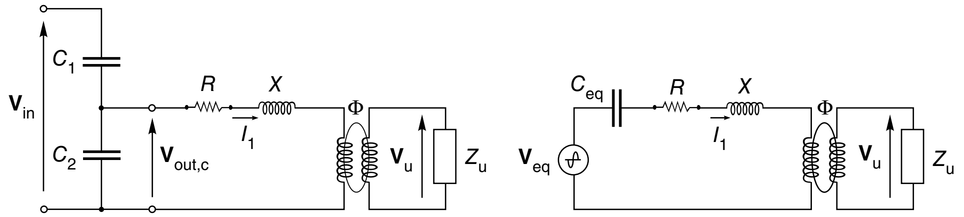


FIGURE 63.8 Capacitive voltage transformer and its equivalent circuit.

value. Out of this range, the lack of the resonance condition between inductance and capacitance may introduce significant errors. As a further consequence of this consideration, in the presence of distorted voltages, the errors in measuring the harmonic components could become unacceptable.

63.4 TRANSDUCERS BASED ON PASSIVE COMPONENTS

Using passive elements (resistors and capacitors) is one of the oldest and most reliable methods to convert currents into voltages and to reduce the amplitude of a primary voltage to a value suitable for measurement instrumentation. These sensors are generally characterized by low cost, easy use, and good accuracy. On the other hand, a major drawback in the use of such devices is that they do not guarantee the insulation between primary and secondary circuits. Thus, alternative solutions are needed to ensure safety of the measurement system.

63.4.1 Shunts

Resistive shunts are based on the Ohm's law to convert the current $i(t)$ flowing into the resistance R_s into the voltage $v(t)$ across its terminals: $v(t)=R_s i(t)$.

On the other hand, this linear relationship is only ideal, whereas in the reality it is affected by a number of influence factors, like variability of the parameters with time, signal frequency, environmental conditions, in particular the temperature, etc.

Shunts can be built with many different techniques, including wound conductors, coaxial resistors, thick film and thin film resistors, metallic plates, etc. Each one of these typologies privileges one or more aspects (robustness, long-term stability, circuit miniaturization, thermal exchange, etc.) over the others, and thus the choice of the most suitable shunt should be done by taking into account the requirements of the specific application.

All resistive shunts have limited bandwidth, owing to the effects of parasitic inductances and capacitances. The inductive reactance is the largest problem in the design of these devices, especially for low resistances. Indeed, when the frequency of the signal components increases, a more appropriate circuital model of the resistive shunt includes at least a series (undesired) inductance, as shown in Figure 63.9.

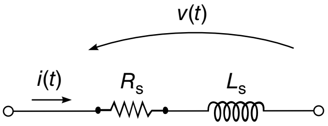


FIGURE 63.9 Equivalent model of a shunt resistor with residual inductance.

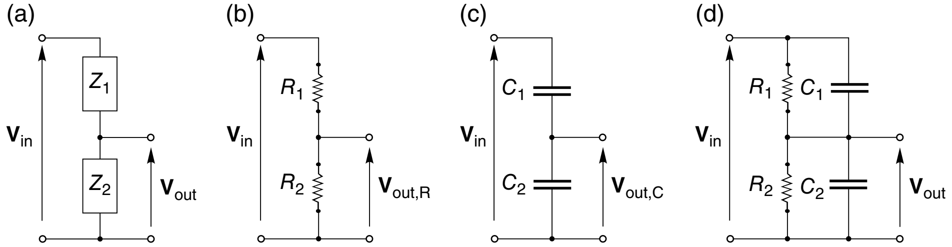


FIGURE 63.10 Voltage dividers: (a) general scheme; (b) resistive divider; (c) capacitive divider; (d) RC divider.

Thus, the relation between current and voltage is actually $v(t) = R_s i(t) + Ldi(t)/dt$. The relative error caused by the presence of the inductance is higher for higher frequency and for lower value of the resistance.

63.4.2 Voltage Dividers

A voltage divider reduces the voltage applied to its primary terminals of a prefixed ratio. The simplest divider consists of the series of two impedances $\bar{Z}_1$ and $\bar{Z}_2$ (Fig. 63.10a). Under sinusoidal conditions and with no load connected on the output terminals, it is

$$\mathbf{V}_{\text{out}} = \mathbf{V}_{\text{in}} \frac{\bar{Z}_2}{\bar{Z}_1 + \bar{Z}_2} = \frac{1}{\bar{K}_r} \mathbf{V}_{\text{in}} \tag{63.8}$$

The divider is said to be compensated when the ratio K_r is a real number independent of the signal frequency. This can be obtained by using either impedances of the same kind, for example, two resistances or two capacitances, as shown in Figure 63.10b and c, respectively, or series/parallel combinations of resistances and capacitances, properly chosen so that the two impedances $\bar{Z}_1$ and $\bar{Z}_2$ have the same time constant (in Fig. 63.10d: $\tau_1 = R_1C_1 = \tau_2 = R_2C_2$). Inductive components are generally not used for these purposes.

In practical cases, the secondary terminals of the divider are always loaded by a burden, which affects the validity of Equation 63.8. The higher the load impedance, with respect to the equivalent divider's impedance, the lower the influence of the load itself on the transducer behavior.

The divider's components are never pure elements but contain parasitic elements, such as internal or link inductances, capacitive couplings between parts of the device or between the device and neighborhood objects, series or leakage resistances in the capacitors, etc. These elements make the behavior of the transducer dependent on frequency. For this reason, often resistive and capacitive dividers contain suitable compensation circuits.

The main advantage of voltage dividers, besides the reduced size and weight, is their good linearity. Thus, the same device can be used in wide voltage ranges, for instance, from tens to hundreds kilovolt, thus allowing proper measurements under both normal operating conditions and faults. This makes these devices suitable to be used in MV and HV systems, instead of magnetic VTs.

Compensated dividers can guarantee very high bandwidth (from DC to tens of megahertz).

On the other hand, they have the same drawback seen for the shunts, since they do not guarantee insulation between high-voltage and low-voltage terminals.

However, often the impedance of the elements that compose the divider (which, for instance, in some commercial devices for medium-voltage systems is in the order of $10^8\,\Omega$) represents by itself a sufficient level of insulation, thus making voltage dividers the most common alternative to VTs in MV systems. Of course, in this case, for safety reasons, one of the divider’s terminals should be connected to ground, and this means that floating voltages, or line-to-line voltages in three-phase systems, cannot be measured with this solution.

In some other cases, the safety of the measurement or protection system may be ensured by either introducing additional devices (e.g., the isolation amplifiers that will be presented in the next section) or implementing alternative solutions, like the ones based on spark gaps or surge arresters.

63.4.3 Isolation Amplifiers

Isolation amplifiers provide the electric insulation between input and output, by means of either magnetic or optical coupling. Therefore, they can be combined with either the shunts or the voltage dividers discussed earlier to realize complete current and voltage transducers, including insulation between primary and secondary circuit. Of course, the overall metrological characteristics of these transducers (accuracy, linearity, bandwidth) are significantly affected by the presence of this new element.

Figure 63.11 shows, as an example, the scheme of an isolation amplifier based on magnetic coupling. Two distinct circuital areas can be noticed, for input and output, respectively. A third section provides power supply for both input and output circuits. All these three sections are mutually insulated through magnetic couplings.

The input signal (which can represent either the voltage drop across the terminals of a current shunt or the output voltage of a voltage divider) is applied to the input buffer.

The useful signal travels across the insulation barrier by means of modulation and demodulation technique. The modulator translates the original baseband signal spectrum to high frequencies. The modulated signal reaches the secondary side by means of the magnetic coupling and is then demodulated into its original baseband. Finally, the output voltage is provided by means of a second buffer. The use of

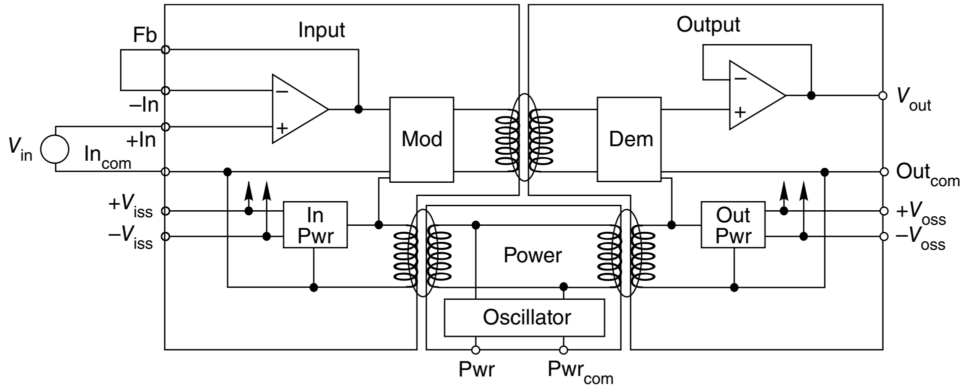


FIGURE 63.11 Isolation amplifier based on magnetic coupling.

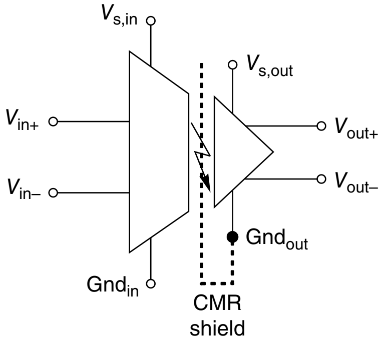


FIGURE 63.12 Isolation amplifier based on optical coupling.

modulation and demodulation allows also DC components to be transferred to the secondary side of the transformer.

A different solution can be implemented by using an optical coupling (Fig. 63.12). In this case, the low voltage at the output of the passive sensor is transformed, by an analog-to-digital converter, into sequence of bits, which is then transmitted across the optical barrier and then, if needed, reconverted in an analog signal.

In many commercial products, the sensing passive element (shunt or divider) and the isolation amplifier are contained in a single device. These transducers are mainly used in low-voltage power electronic applications.

63.5 HALL-EFFECT AND ZERO-FLUX TRANSDUCERS

63.5.1 The Hall Effect

As it is well known, when an electrical current I_p passes through a conducting slab placed in a magnetic field with induction B (Fig. 63.13), a force acts on the charged particles in motion, and consequently, a potential v_H proportional to the current and to

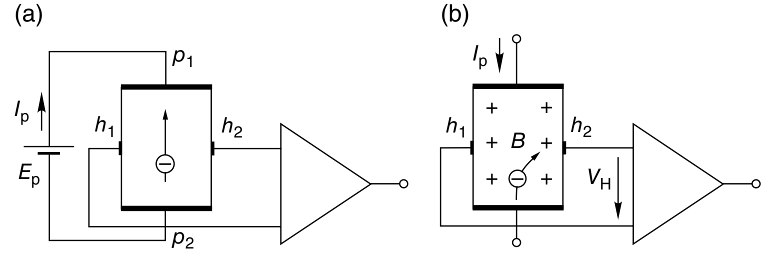


FIGURE 63.13 The Hall effect: path of the electrons without (a) and with (b) magnetic induction.

the magnetic field is developed across the slab in a direction perpendicular to both the current and the magnetic field. This property, known as the Hall effect, can be suitably exploited to measure magnetic induction and related quantities.

The use of Hall-effect transducers is one of the most popular solutions to perform measurements of electric quantities, both in DC and AC, with insulation between input and output. The basic configuration is suitable for current measurements, but as it will be seen, a simple calibrated resistance can allow also voltages to be measured.

63.5.2 Open-Loop Hall-Effect Transducers

In order to explain the behavior of an open-loop Hall-effect current transducer, let us refer to Figure 63.14.

A conductor passes through the hole of a magnetic core, where the Hall sensor is placed. A magnetic field proportional to the current flowing in the conductor is produced around the conductor itself. More in general, a winding composed of a given number of turns may be present in the primary circuit to increase the sensitivity. The lines of the magnetic field concentrate in the core and excite the Hall sensor, giving rise to a voltage proportional to the primary current. This very low voltage is amplified to obtain the voltage at the output terminals.

In the open-loop configuration, accuracy, linearity, and bandwidth of the transducer depend directly on the characteristics of the single components. Furthermore, the saturation of the magnetic core must be avoided in the normal operation of these devices. As a consequence, their metrological performance is usually limited, especially as far as the frequency behavior is concerned. It should be however taken into account that also DC currents can be measured with these devices.

On the other hand, their cost is generally low, and the low power consumption makes them suitable for portable devices where the supply voltage is provided by batteries.

63.5.3 Closed-Loop Hall-Effect Transducers

Closed-loop Hall-effect transducers use feedback technique to improve the performance, by nullifying the magnetic flux in the core. To do this, a secondary winding is added to the components of the open-loop transducer (see Fig. 63.15).

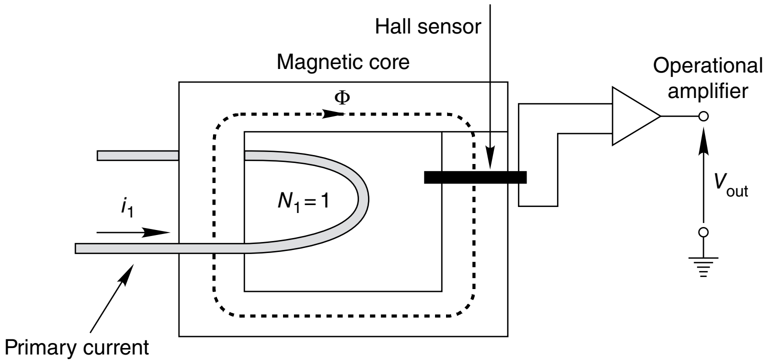


FIGURE 63.14 Open-loop Hall-effect current transducer.

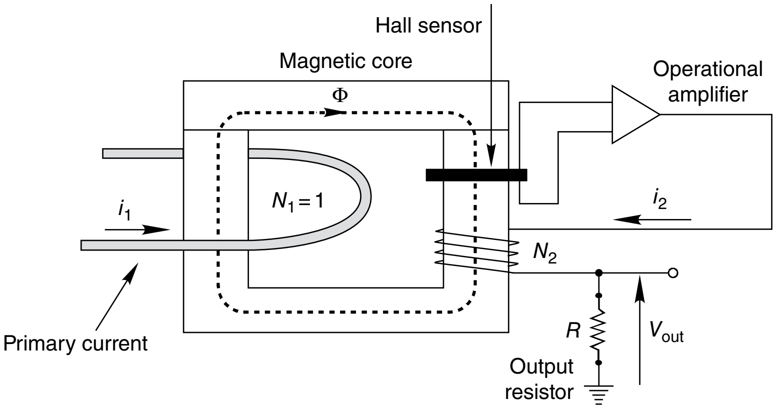


FIGURE 63.15 Closed-loop Hall-effect current transducer.

This compensation winding is so placed that the current flowing in it generates a magnetic field opposite to that produced by the current in the primary conductor. The Hall sensor actually operates as a "zero detector," insofar as it amplifies the "error signal" (a nonzero flux in the core) to drive the feedback action. The component is therefore designed to have the highest sensitivity. By combining the Hall sensor with a high-gain amplifier, a current is generated in the compensation winding to contrast the magnetic field produced by the primary current. This compensating current can be directly used as output quantity (current-to-current transducer). The transformation ratio is determined by balancing the ampere-turns in the core. As an example, in order to have a 1000:1 ratio in a transducer where the primary current passes directly across the window of the magnetic core ($N_1 = 1$), the secondary circuit will have $N_2 = 1000$ turns. Alternatively, the secondary current can be converted into a proportional voltage through a calibrated resistor (current-to-voltage transducer).

The frequency response can be analyzed by considering two partially overlapped regions (Fig. 63.16). In the first one, from DC to low frequencies, the behavior depends mainly on the electronic operation of the Hall sensor. In the second region the

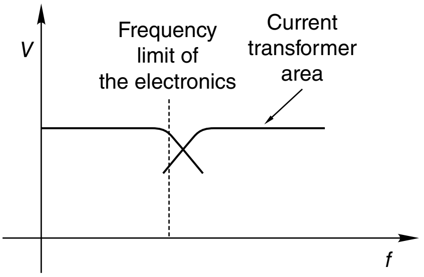


FIGURE 63.16 Bandwidth of a closed-loop Hall-effect current transducer.

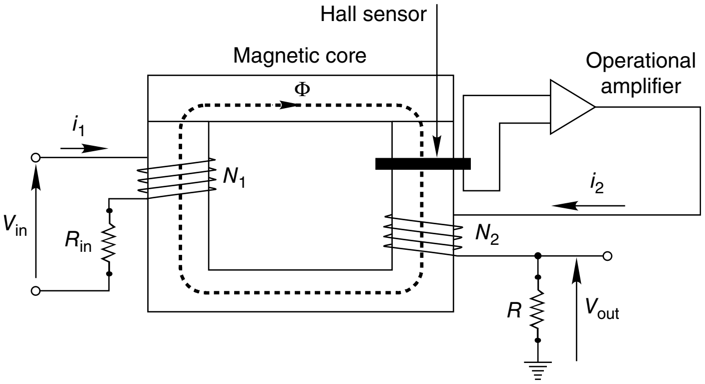


FIGURE 63.17 Hall-effect closed-loop voltage transducer.

compensating winding works practically as the secondary winding of a current transformer, thus extending the frequency range of the device. Obviously, the device should be so designed that the transition between the two regions is gradual to guarantee that the frequency response is sufficiently flat in the overall bandwidth of the transducer (which can be up to hundreds of kilohertz).

As emphasized before, Hall-effect transducers are intrinsically suitable for current measurement. However, voltage transducers can be also built, by adding a calibrated resistance R_{in} across the terminals, which draws a current that can then be measured with a Hall-effect sensor. Figure 63.17 shows the case of a closed-loop voltage transducer. In this case, the additional resistance should be sufficiently low so that, when the measured voltage is applied to its terminals, the current absorbed can be appreciated by the Hall sensor, avoiding sensitivity issues, but also sufficiently high so that the system under test is not subject to excessive loading effects. A trade-off is therefore required.

Note that the voltage drop v_{in} in the primary winding is in phase with the current i_1, due to resistive part of the input impedance, being the inductive one absent, owing the zero flux in the core.

Hall-effect transducers are usually designed for low-voltage systems (below 1000 V), even though devices for medium-voltage systems exist. Their field of application is mainly in industry automation (variable speed drives, power supplies, filters, etc.), traction systems, and, more in general, in those situations where their ability to measure both DC and AC components can be of fundamental importance. Their use in substations or in switchboards of electric distribution grids is much less frequent.

63.5.4 Zero-Flux Transducers

The principle of operation of the closed-loop Hall-effect transducer can be generalized by employing different systems to detect the flux in the magnetic core and obtaining the so-called electronically compensated current transformers (ECCTs) or zero-flux transducers.

These transducers are conceptually and physically similar to current transformers. Indeed, the sensing element is a magnetic circuit where the flux generated by the measured current is nullified by the flux generated by the secondary current. However, if we take into account that in traditional current transformers the main source of uncertainty is the exciting current needed to magnetize the magnetic core, the accuracy of such transducers can be improved by reducing to zero the flux. This can be achieved by supplying the secondary winding through an amplifier that receives at its input a signal proportional to the core flux. In this way, the voltage across the core's magnetizing impedance is reduced and so is for the exciting current, thus improving the transducer's accuracy.

This approach also eliminates the problems related to the saturation of the magnetic core.

The different typologies based on this principle use different magnetic flux detectors, which can be realized by means of additional windings, Hall sensors (Fig. 63.15), or more complex solutions.

As for the Hall effect, the zero-flux technique can be used also to build voltage transducers by adding a calibrated resistor that absorbs the current to be measured by the ECCT.

Compensated transducers have reduced size, with respect to traditional instrument transformers, bandwidth up to hundreds of kilohertz, and very good accuracy. They are mainly used in special applications or in laboratories, even though on the market some solutions for MV systems exist.

Finally, it should be mentioned that in the last years, several digital compensation techniques for instrument transformers have been introduced. These techniques require a mathematical or experimental model of the transformer to be known. On this basis, once the secondary quantity has been acquired, the primary quantity is known with better accuracy by implementing suitable compensation routines on a microprocessor.

63.6 AIR-CORE CURRENT TRANSDUCERS: ROGOWSKI COILS

The possibility of measuring alternate currents by means of air-core transducers (often referred to as Rogowski coils) has been known since the beginning of the twentieth century. However, such devices are receiving great popularity only in these decades,

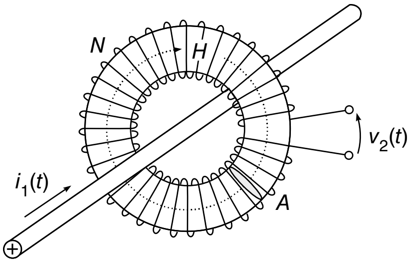


FIGURE 63.18 Rogowski coil.

thanks to the developments of the electronics, which is a necessary complement of these transducers.

Besides ensuring insulation between input and output circuits, Rogowski coils offer wide measurement range (up to hundreds of kiloamperes), wide bandwidth (from a few hertz up to the megahertz region), and excellent linearity.

The Rogowski coil is essentially a coil made of conducting material, uniformly wounded around a ring made of a nonferromagnetic material ($\mu \approx \mu_0$), in whose central window the conductor bringing the measured current is placed (see Fig. 63.18).

The current flowing in the conductor produces a magnetic field around it. According to the Ampere's law, the integral of the magnetic field *H* around a closed path *L* equals the surrounded current i_1:

$$\int_L \overline{H} \cdot d\overline{l} = \int_L H \cdot dl \cos\alpha = i_1 \tag{63.9}$$

where α is the angle between the direction of the magnetic field *H* and the infinitesimal element of length *dl*.

If *n* is the number of turns for unity of length, the number of turns in a portion of the coil having length *dl* is *ndl*. Assuming that the cross area *A* of the turns is constant, the magnetic flux concatenated with such portion is

$$d\Phi = \mu_0 HndlA\cos\alpha \tag{63.10}$$

Therefore, the flux concatenated with the entire coil is

$$\Phi = \int_L d\Phi = \mu_0 nA \int_L H\cos\alpha dl = \mu_0 nAi_1 \tag{63.11}$$

According to the Faraday's law, the voltage $v_2(t)$ at the coil's terminals is given by

$$v_2(t) = -\frac{d\Phi(t)}{dt} = -\mu_0 nA \frac{di_1(t)}{dt} \tag{63.12}$$

An example, if a current $i_1(t)$ with a triangular waveform is considered, as in Figure 63.19, the output of the Rogowski coil is a square wave voltage $v_2(t)$.

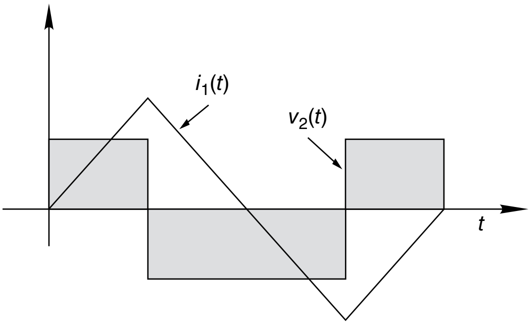


FIGURE 63.19 Output voltage of a Rogowski coil with a triangular wave input current.

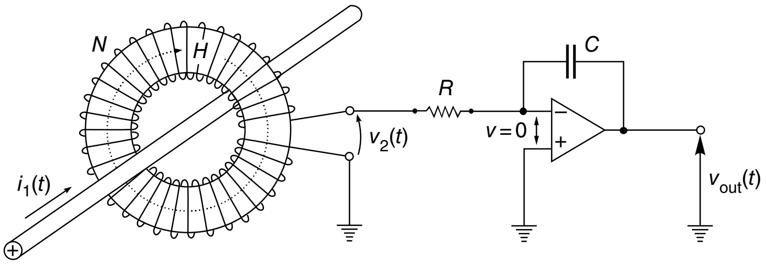


FIGURE 63.20 Rogowski coil with integrator.

Obviously, since the Rogowski coil is sensitive to the current's derivative, it cannot be used to measure DC currents. However, differently from magnetic CTs, the presence of a possible DC component in the primary current does not adversely affect the accurate measurement of the alternate components superimposed to it.

As seen before, the voltage at the coil's output terminals is proportional to the current's derivative. Thus, in order to have an actual transducer's output proportional to the primary current, such voltage must be integrated. This can be done by means of either active or passive integrating circuits. In Figure 63.20 the analog integrator is built through an operational amplifier with a capacitance C in its feedback path. As an example, if the triangular current $i_1(t)$ of Figure 63.19 is considered, the integration of the square wave output voltage $v_2(t)$ provides a voltage $v_{\text{out}}(t)$ that reproduces, for each time instant, the triangular wave of the current $i_1(t)$.

As an alternative solution, the voltage $v_2(t)$ could be acquired and digitized, and the integral could be calculated by means of digital signal processing.

The Rogowski coil is an extremely versatile current transducer. Its sensitivity can be regulated in a wide range by acting on both the turn density and their cross area. In this way, currents from a few milliamperes to several megaamperes can be measured.

Two possible physical typologies exist: flexible coil and rigid coil. The former is more versatile and more suitable for high-frequency currents but is less accurate than the latter. Rigid coils are more suitable for low-frequency and low currents.

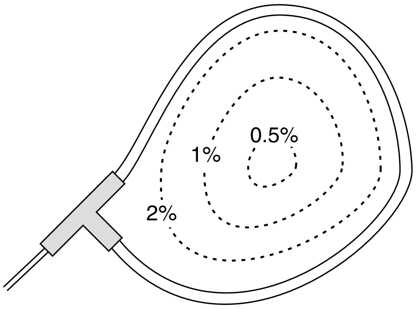


FIGURE 63.21 Possible impact of the conductor placement on the accuracy of a Rogowski coil.

The main uncertainty sources of the Rogowski coils are:

- *Assembly tolerance*:

  Theoretically, the transducer's output should not depend on the position of the conductor with respect to the coil. Actually, the nonperfect setting up of windings and insulating supports gives rise to errors that can be up to some percent. The largest error arises when the conductor passes close to the junction between the two extremes of the coil. The smallest error is found, obviously, when the conductor passes near the center of the coil (Fig. 63.21).
- *Temperature variations*:

  The dependency on the temperature can be minimized by using special materials with very low temperature coefficient. As an alternative, the temperature can be measured and its effects compensated.
- *Cross talk*:

  The cross talk effect, due to the influence of currents in neighboring conductors, mainly depends on the phase-to-phase distance. Such disturbance can be minimized in the design stage. A fundamental solution is that the return link should be placed internally to the nonmagnetic core of the support where the turns are wounded (Fig. 63.22). If this was not done, possible magnetic fields having direction perpendicular to that of the coil plane would be concatenated with the coil and would affect seriously the measurement results.

Differently from current transformers and other transducers, Rogowski coil does not need magnetic cores, and thus no saturation occurs. This intrinsically linear behavior makes this device suitable to measure high fault currents. Thanks to this wide measurement range (Fig. 63.23), a single current transducer can be used for both measurement and protection purposes. On the contrary, when traditional CTs are used, different windings should be considered for the different measurement ranges.

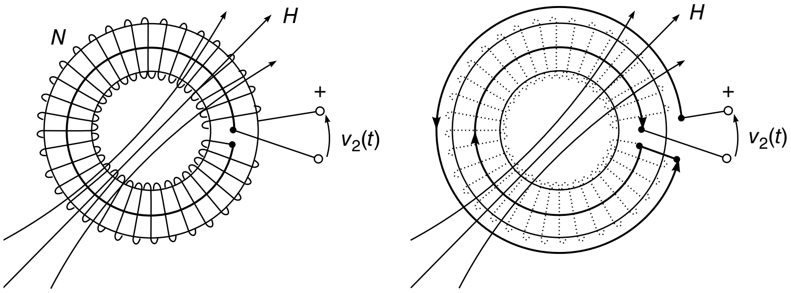


FIGURE 63.22 Compensation of cross talk in a Rogowski coil: position of the return path (left) and equivalent circuit (right).

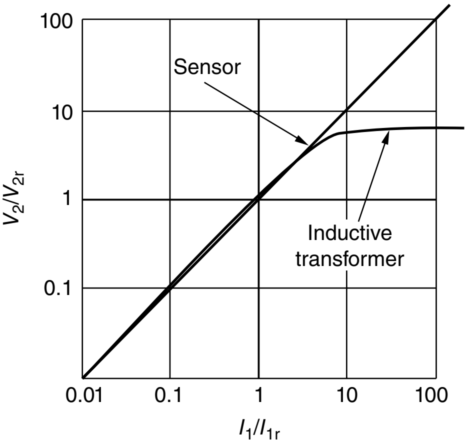


FIGURE 63.23 Input–output characteristic of air-core sensor and inductive CT.

Furthermore, the possible safety issues that could occur in CTs if the secondary circuit was opened (recalled in Section 63.3.1) are not present in Rogowski coils, because in this cause no overvoltage would be generated.

The main limitations to the current range arise from the integrator. The measurement range depends on both amplitude and frequency of the measured current. Low currents at low frequency would produce very low voltages at the coil's terminals, thus leading to sensitivity problems. On the other hand, high currents rapidly changing (i.e., with high time derivative) can lead to high voltages that the electronics of the integrator could not tolerate.

The integrator also limits the bandwidth, which, in practical cases, varies for a few hertz to hundreds of kilohertz, which is enough for most applications in power systems. Special productions allow wider bandwidth to be reached, up to hundreds of megahertz.

Thanks to their attractive practical and metrological features, Rogowski coils are used in several applications on power grids, such as fault current detection, PQ monitoring, power and energy measurements, measurement of small currents superimposed on large DC (e.g., capacitor ripple), large AC currents (e.g., arc furnaces), measurement of bearing and shaft currents in large machines, etc. All these possible applications make this device the most valid alternative solution to CTs, in particular in low- and medium-voltage systems. Some applications for high-voltage systems have been also proposed.

63.7 OPTICAL CURRENT AND VOLTAGE TRANSDUCERS

Optical methods for measuring voltages and currents have received increasing attention in the last years, especially for high-voltage systems. This is due to the advantages they allow with respect to instrument transformers: high immunity to electromagnetic interferences, excellent insulation, lightweight and reduced size, no saturation, wide measurement range, and wide bandwidth.

An optical transducer for voltage or current is a complex system that includes an optical sensing element, an optical fiber communication path, and an electronic module for signal processing and interfacing with measurement and protection equipment.

The sensing element is placed close to the quantity to be measured and generally produces a light modulation.

Some fundamental concepts of optics should be known to understand the principle of operation of optical voltage and current transducers. In order to make this section "self-consistent," in the following the most important concepts will be recalled in an extremely concise and descriptive way, which must be considered neither complete nor scientifically rigorous. For deeper analysis the reader should refer to specialized texts.

Light *polarization* is a phenomenon specific of the propagation of the light waves, characterized by the oscillation direction of the waves in a given plane, known as polarization plane. This is the plane defined by the oscillation direction (conventionally determined by the direction of the electric field) and the propagation direction (see, for instance, Fig. 63.24, where the shaded gray area defines the polarization plane and *c* is the propagation speed). When the polarization plane holds a fixed direction,

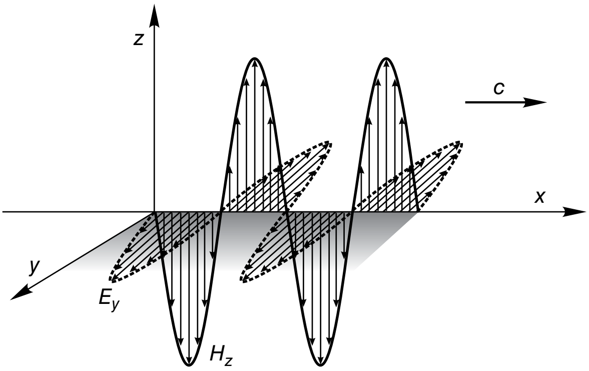


FIGURE 63.24 Propagation of an electromagnetic wave (horizontal *E* polarization).

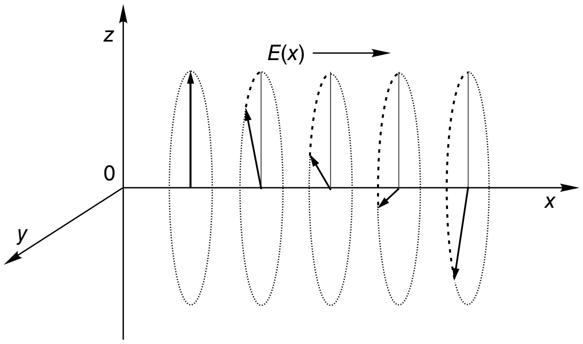


FIGURE 63.25 Rotation of a linearly polarized wave.

the light waves are said to be linearly polarized. This occurs when the electric field components with respect to the axis of a plane orthogonal to the propagation direction are in phase to each other. If the polarization plane rotates around the propagation direction with constant angular speed, the polarization is said to be elliptical (electric field components are not in phase) or circular (when the phase shift between the electric field components is $\pi/2$).

Optical transducers used in power systems are mainly based on two characteristics of optical materials: *optical activity* and *birefringence*.

The *optical activity* is a property of some materials in which, when a light wave passes through them, its polarization plane rotates (Fig. 63.25). If a linearly polarized light beam crosses an optically active matter, the transmitted wave is still linearly polarized but, on a different plane, shifted of a given angle with respect to the incidence polarization plane.

Birefringence is a property exhibited by some materials, in which the index of refraction depends on the propagation direction of light. In particular, the index has two different values for mutually orthogonal light polarizations. Birefringence can be intrinsic, in anisotropic materials like crystals, or induced by either mechanical, electrical, or magnetic stimulus. The different propagation speed of the light in the materials introduces a phase shift between the two orthogonal components. Thus, for instance, a linearly polarized light becomes elliptically polarized.

63.7.1 Optical Current Transducers

The principle of operation of optical current transducers (OCTs) is based on the magneto-optic Faraday's effect, which is essentially a modulated optical activity. When an optical material, subjected to a magnetic field, is crossed in the field direction by a linearly polarized light beam, the polarization plane rotates with an angle proportional to the field intensity:

$$\theta = \mu V \int_L H \, dl \qquad (63.13)$$

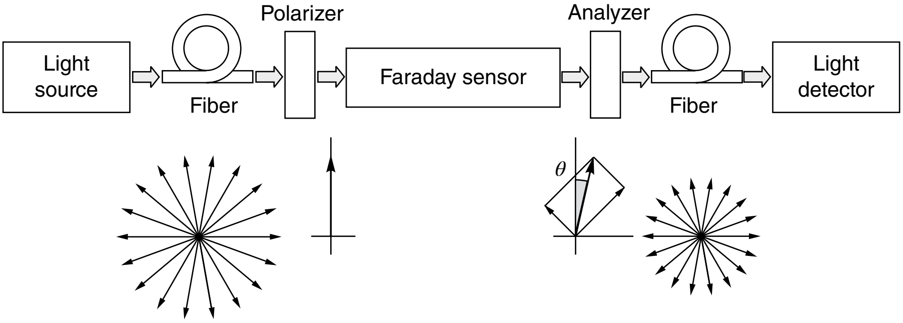


FIGURE 63.26 Main components of an optical current transducer and relevant light polarization status.

where θ is the rotation of the polarization, μ is the magnetic permeability of the material, H is the magnetic field component parallel to the propagation direction of the beam, L is the path length of the light, and V is the Verdet constant, which depends on material, wavelength, and temperature.[1]

An OCT generally consists of the following elements (Fig. 63.26):

- Light source (usually a diode)
- Optical fibers, which provide the link between the sensing element and the electronic components
- Polarizer, which could be thought as a device that selects only one kind of light polarization of all the incident optical energy
- Sensing element, built with either silicon crystals or optical fibers
- Photodetector (photodiode), which converts optical signals into electrical signals.
- Digital signal processor (DSP), which implements the function that relates the physical phenomena to the measured parameters. In addition, the processor allows performing some compensation (temperature, linearity, etc.) and establishing digital communication with other devices.

In most recent and accurate OCTs, the sensing element has an optical path surrounding the conductor. In this way, according to the Ampere's law, if the light is uniformly sensitive to the magnetic field H in the closed path L around the conductor, then the rotation θ of the polarization plane is proportional to the current flowing in the conductor:

$$\theta = \mu V \int_L H\, dl = \mu V i \qquad (63.14)$$

This solution can be practically implemented in different ways.

[1] In some texts the Verdet constant includes the magnetic permeability μ, which, consequently, does not appear in the expression of the rotation. The substance does not change, but it is necessary to consider carefully the definition used, in order to have a correct dimensional interpretation of Equation 63.13.

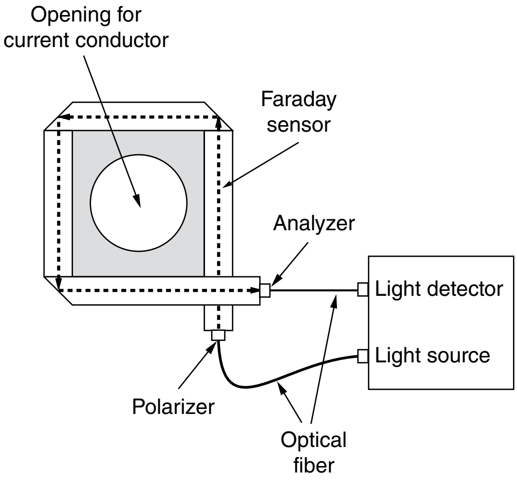


FIGURE 63.27 Example of OCT.

As an example, the sensing element can consist of crystals linked to each other to form a closed ring that surrounds the conductor (Fig. 63.27).

In a second solution the path of the light is developed in an optical fiber wounded around the conductor. The use of optical fibers allows more flexibility to be achieved. Since optical fibers have usually a low Verdet constant, the required sensitivity is obtained by increasing the number of turns *n* around the conductor, according to the expression:

$$\theta = \mu n V i \tag{63.15}$$

Actually, especially when fiber optic OCTs are concerned, the Faraday's magneto-optic effect may be exploited in a different way, that is, involving different polarization conditions of the light, with respect to the previously described solution.

As far as the analysis of the output signal of the sensor is concerned, it should be considered that the variation of the polarization status, for example, the rotation of the polarization plane in the scheme of Figure 63.26, cannot be measured directly, since photodetectors are not sensitive to the light polarization but to the optical power, which is proportional to the square of the electric field. Therefore, suitable methods should be implemented to measure indirectly the rotation of the polarization.

As an example, by referring again to the solution described in Figure 63.26, one of the simplest methods consists of processing the light leaving the sensing element with a second polarizer, named analyzer, whose axis forms typically an angle of $\pi/4$ with respect to the axis of the first polarizer.

If α is the angle between the direction of light polarization at the sensor output and the polarization axis of the analyzer, under the assumption of no losses, the optical power P_{det} received by the detector can be expressed as a function of the input power P_{in}:

$$P_{\text{det}} = P_{\text{in}} \cos^2 \alpha = \frac{1}{2} P_{\text{in}} \left[1 + \cos(2\alpha)\right] \tag{63.16}$$

where $\alpha = \pi/4 + \theta$.

This means that

$$P_{\text{det}} = \frac{1}{2} P_{\text{in}} \left[1 - \sin(2\theta)\right] = P_{\text{dc}} - P_{\text{ac}} \tag{63.17}$$

The rotation θ varies with time following the time variability of the input current. Thus, the signal at the detector has a constant term $P_{\text{dc}} = 1/2P_{\text{in}}$ and a variable term $P_{\text{ac}} = 1/2P_{\text{in}} \sin(2\theta)$, which represents the modulation caused by the Faraday effect. By taking into account that the rotation angle θ is small, the variable component can be expressed as

$$P_{\text{ac}} = \frac{1}{2} P_{\text{in}} \sin(2\theta) = P_{\text{in}}\theta + \text{higher-order terms} \tag{63.18}$$

By neglecting the higher-order terms, normalizing the alternating component of the P_{det} with respect to the constant one and making explicit the time variability of the quantities, it results to

$$\frac{P_{\text{ac}}}{P_{\text{dc}}} = 2\theta(t) = Ki(t) \tag{63.19}$$

where $i(t)$ is the measured current and K is a constant term that depends on the characteristics of the sensor. It is evident that this method does not allow the DC components of the current to be measured. To overcome this limit, different solutions, often involving more than one analyzer, can be implemented.

As far as the metrological behavior is concerned, OCTs today available can have good performance under both sinusoidal and distorted conditions. Accuracy specifications are ensured in a wide dynamic range, from 1% up to 200% of the rated current. This allows the same transducer to be used for both measurement and protection purposes, thus avoiding the need to have two instrument transformers.

As for the frequency response, that of the sensing element could be very high, but that of the complete transducer is limited to some tens of kilohertz, mainly because of the presence of sampling and conversion electronic devices.

Further advantages of the OCTs are reduced weight and size, which lead to reduction of transportation and installation costs and low maintenance costs.

63.7.2 Optical Voltage Transducer

Most optical voltage transducers (OVTs) base their operation on the electro-optic Pockels effect: when a crystal is subjected to an electric field parallel to the light direction, a birefringence proportional to the electric field, and thus to the applied voltage, is induced.

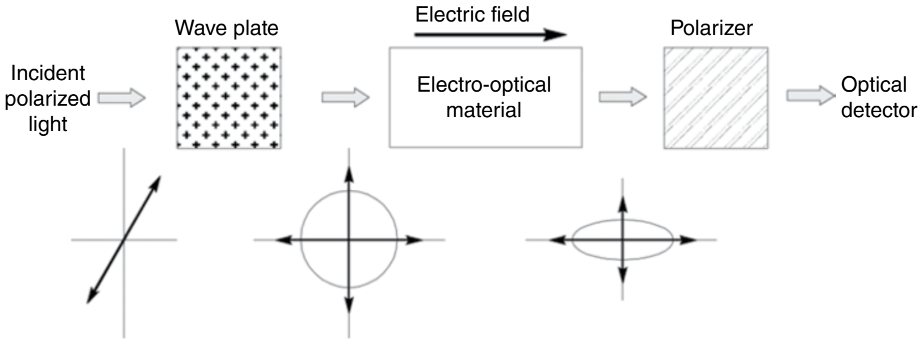


FIGURE 63.28 Main components of an optical voltage transducer and relevant light polarization status.

Figure 63.28 shows the main optical elements that compose an OVT. After a polarizer (not shown in Fig. 63.28) has converted the incident nonpolarized light into a linearly polarized beam, which can be considered composed by two in-phase orthogonal components, the light enters the quarter wave plate, which introduces a $\pi/2$ shift between the two components, thus giving rise to a circularly polarized light beam.

This beam enters the Pockels cell, where an electric field in the light direction is generated by the voltage applied to a couple of electrodes. The electric field induces the birefringence, which causes a further phase shift between the two light components, thus transforming the circular polarization into an elliptical one. The phase difference induced by the birefringence, which is proportional to the applied voltage, is converted by an analyzer into a modulation of the optical power. This power is measured by means of a photodetector, which converts the optical signal into an electrical one, and suitable data acquisition and processing systems.

Some commercial OVTs employ capacitive dividers to reduce the voltage applied to the optical sensor.

As for the metrological behavior, most of the considerations done for OCTs hold for OVTs. Their accuracy class is comparable to that of instrument transformers, and in addition, thanks to wide bandwidth and excellent linearity, these instruments can accurately measure both DC and AC voltages in a range from 20 to 200% of the rated value.

63.7.3 Applications of OCTs and OVTs

Manufacturers often propose commercial solutions that include in a single device the functionalities of optical voltage and current transducers, thus resulting in more compact units that can be advantageous in those installations where space is a critical issue.

As far as the field of application is concerned, it has been already stated that optical current and voltage transducers have been originally designed and built for high-voltage systems. However, the optical technology has evolved so that more recently, optical transducers have become available also for medium-voltage grids. Their characteristics may be in fact very useful in a measurement system designed to monitor, control, and protect active distribution networks. The high dynamic range allows measuring very accurately both low and high currents, thus eliminating the need for parallel current transformers. At the same time, this flexibility may allow reducing inventory and simplifying maintenance. Their bandwidth extends to several kilohertz, thus allowing a correct reproduction of the distorted quantities present in these systems. The possibility of a digital output means a direct compatibility with digital measurement and protection devices. The lightweight design and intrinsic insulation enable the sensors to be easily and economically installed along feeders. For all the previous reasons, optical transducers are considered among the most promising solutions for next-generation distribution grids.

REFERENCES AND FURTHER READING

Journal and Conference Papers

Emerging Technologies Working Group and Fiber Optic Sensors Working Group: “Optical current transducers for power systems: a review,” *IEEE Transactions on Power Delivery*, Year: 1994, Volume: 9, Issue: 4, Pages: 1778–1788, DOI: 10.1109/61.329511.

Kojovic, L.: “Rogowski coils suit relay protection and measurement,” *IEEE Computer Applications in Power*, Year: 1997, Volume: 10, Issue: 3, Pages: 47–52.

Kucuksari, S.; Karady, G.G.: “Experimental comparison of conventional and optical current transformers,” *IEEE Transactions on Power Delivery*, Year: 2010, Volume: 25, Issue: 4, Pages: 2455–2463.

Locci, N.; Muscas, C.: “Comparative analysis between active and passive current transducers in sinusoidal and distorted conditions,” *IEEE Transactions on Instrumentation and Measurement*, Year: 2001, Volume: 50, Issue: 1, Pages: 123–128.

Locci, N.; Muscas, C.; Sulis, S.: “Experimental comparison of MV voltage transducers for power quality applications,” *IEEE Instrumentation and Measurement Technology Conference*, Year: May 5–7, 2009, Pages: 92–97, DOI: 10.1109/IMTC.2009.5168422.

Minkner, R.; Schweitzer, E.O. III: “Low Power Voltage and Current Transducers for Protecting and Measuring Medium and High Voltage Systems,” 26th Annual Western Protective Relay Conference, Washington State University, October 1999.

Oates, C.D.M.; Burnett, A.J.; James, C.: “The design of high performance Rogowski coils,” *International Conference on Power Electronics, Machines and Drives*, Year: June 4–7, 2002, Pages: 568–573, DOI: 10.1049/cp:20020179.

Ray, W.F.; Hewson, C.R.: “High performance Rogowski current transducers,” *IEEE Industry Applications Conference*, Year: 2000, Volume: 5, Pages: 3083–3090.

Sawa, T.; Kurosawa, K.; Kaminishi, T.; Yokota, T.: “Development of optical instrument transformers,” *IEEE Transactions on Power Delivery*, Year: 1990, Volume: 5, Issue: 2, Pages: 884–891.

Ward, D.A.: “Measurement of current using Rogowski coils,” *IEE Colloquium on Instrumentation in the Electrical Supply Industry*, Year: 1993, Volume: 1, Pages: 1–3.

Xiao, C.; Zhao, L.; Asada, T.; Odendaal, W.G.; van Wyk, J.D.: “An overview of integratable current sensor technologies,” *38th IAS Annual Meeting. Conference Record of the Industry Applications Conference*, Year: 2003, Volume: 2, Pages: 1251–1258.

Application Notes and Technical Brochures

ABB: "*Instrument Transformers—Application Guide*," 4th edition, ABB, Pinetops, NC, 2015.

LEM Components: "*Isolated Current and Voltage Transducers: Characteristics—Applications—Calculations*," 3rd edition, LEM Corporate Communications, Geneva, Switzerland, 2005.

Analog Devices: "Analog Isolation Amplifiers", MT-071 Tutorial, 2009.

International Standards

IEC 61869-1:2007:
Instrument transformers—Part 1: General requirements

IEC 61869-2:2012:
Instrument transformers—Part 2: Additional requirements for current transformers

IEC 61869-3:2011:
Instrument transformers—Part 3: Additional requirements for inductive voltage transformers

IEC 61869-4:2013:
Instrument transformers—Part 4: Additional requirements for combined transformers

IEC 61869-5:2011:
Instrument transformers—Part 5: Additional requirements for capacitor voltage transformers

IEC 60044-7:1999:
Instrument transformers—Part 7: Electronic voltage transformers

IEC 60044-8:2002:
Instrument transformers—Part 8: Electronic current transformers

IEEE Std 3004.1-2013:
IEEE Recommended Practice for the Application of Instrument Transformers in Industrial and Commercial Power Systems
Year: 2013, DOI: 10.1109/IEEESTD.2013.6512522

IEEE Std 1601-2010:
IEEE Trial-Use Standard for Optical AC Current and Voltage Sensing Systems
Year: 2010, DOI: 10.1109/IEEESTD.2010.5674139

IEEE Std C57.13-2008:
IEEE Standard Requirements for Instrument Transformers
Year: 2008, DOI: 10.1109/IEEESTD.2008.4581634

IEEE Std C37.235-2007:
IEEE Guide for the Application of Rogowski Coils Used for Protective Relaying Purposes
Year: 2008, DOI: 10.1109/IEEESTD.2008.4457884

64

ELECTRIC POWER AND ENERGY MEASUREMENT

Alessandro Ferrero and Marco Faifer
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

64.1 INTRODUCTION

Power and energy measurements are probably the most important and critical measurements in power systems, since they are used to quantify every energy transactions and assign an economical value to energy flowing through a given section of the grid.

No wonder then that measurement methods and instruments aimed at quantifying energy flow and power have represented a challenge since the very beginning of the commercial exploitation of electricity. At first, when DC systems were the only available systems to distribute electricity, and the distance between generators and loads was so short that voltage was assumed to be constant, only current was integrated by ampere hour meters. It was soon realized, however, that the assumption of constant voltage was not correct and that voltage variations had a significant impact on the amount of delivered energy. The first DC energy meters were then designed and installed.

Things became even more complex when the first AC distribution systems began to be used as an alternative and more efficient way to distribute electricity and, in the end, replaced the DC systems almost completely. Not only voltage magnitude contributed to the amount of useful energy transfer but also the current phase shift with respect to

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

voltage. Methods and instruments to measure active and reactive power and energy had to be developed.

Today instruments are significantly different from the early ones and also from those who had been in use since about a decade ago. Therefore, the following sections will focus only on the modern ones.

On the other hand, the theoretical concepts that are behind the definition of electric power and energy are always the same and often not fully perceived. Since they may have a nonnegligible impact on the correct interpretation of the measurement results, the next section will briefly recall these concepts.

64.2 POWER AND ENERGY IN ELECTRIC CIRCUITS

It is generally thought that power and energy definitions are the same for DC and AC circuits. This assumption is only an approximation of a stricter derivation. To prove this, the DC and AC conditions will be covered separately.

64.2.1 DC Circuits

Let us consider a constant electric field, represented by a constant electric field vector $\vec{E}$. Let us also suppose that a constant free electric charge q is present in the field. The field will manifest itself with a force acting on charge q, such as

$$\vec{F} = q\vec{E} \tag{64.1}$$

Since the charge is free, it will move in the electric field along the same direction x as the orientation of the electric field, and the elementary work done by the field on the charge will be

$$dW = Fdx \tag{64.2}$$

According to (64.1), this elementary work can be expressed in terms of the electric field and charge values as

$$dW = qEdx \tag{64.3}$$

It is well known that, according to Stokes theorem, the infinitesimal electric potential difference is related to the electric field as $dV = Edx$. It is also known that power is defined as the first time derivative of work. Therefore, differentiating (64.3) and considering the electric potential yields

$$\frac{dW}{dt} = q\frac{dV}{dt} \tag{64.4}$$

Taking into account that the electric current I is defined as the flux of an electric charge in time, (64.4) becomes

$$\frac{dW}{dt} = IdV \tag{64.5}$$

Equation 64.5 is the well-known definition of power in a section of a DC electric circuit:

$$P = VI, \tag{64.6}$$

where I is the current flowing through the section and V is the potential difference (voltage) across the section.

The energy flowing through the same section is, of course, the integral of P over time and is given, therefore, by

$$W = \int_{t_0}^{t} VIdt = VI\left(t - t_0\right) \tag{64.7}$$

being V and I constant.

It can be readily perceived, from (64.6) and (64.7), that power and energy, in a DC circuit, can be measured by measuring V and I and the elapsed time.

64.2.2 AC Circuits

64.2.2.1 General Case It is generally thought that (64.5) can be applied also to circuits where voltages and currents vary with time (such as the AC circuits) by simply replacing the constant values of V and I by their functions of time. However, this is incorrect because (64.3) assumes that only the electric field exists, and this is not true if the field is not constant in time. Under time-varying conditions, electric and magnetic fields exist in the same point of space and are related by Maxwell equations.

Under these time-varying conditions, the power associated with the transverse electromagnetic wave has to be considered. Without entering into too many theoretical details, for which the reader is addressed to [1–3], let us only remind that the instantaneous power of the transverse electromagnetic wave is given by the flux of the Poynting vector through a given closed surface Σ. In mathematical terms, the Poynting vector is given by

$$\vec{\wp} = c^2\varepsilon_0\left(\vec{E} \times \vec{B}\right) \tag{64.8}$$

where

c is the speed of light

ε_0 is the permittivity of free space

$\vec{E}$ is the electric field vector
$\vec{B}$ is the magnetic field vector

Its flux through a given closed surface Σ (i.e., the power associated with the electromagnetic wave) can be decomposed as

$$\oint_{\Sigma} \vec{\wp} \cdot \vec{u}_n \, d\Sigma = \oint_{\Sigma} v\vec{j} \cdot \vec{u}_n \, d\Sigma + \oint_{\Sigma} v \frac{\partial \vec{d}}{\partial t} \cdot \vec{u}_n \, d\Sigma + \oint_{\Sigma} \vec{B} \times \frac{\partial \vec{A}}{\partial t} \cdot \vec{u}_n \, d\Sigma \tag{64.9}$$

where
$\vec{u}_n$ is the orthogonal versor to surface Σ
v is the electric potential
$\vec{j}$ is the current density vector
$\vec{d}$ is the displacement current vector
$\vec{A}$ is the magnetic potential vector

Let us assume, without losing generality, that surface Σ is a spherical surface with diameter d, and let us assume that the wavelength $\lambda|_{\max(f)}$[1] of the highest-frequency component in the electromagnetic quantities is such that $d \ll \lambda|_{\max(f)}$. Under these conditions, it can be proved that the last two integrals in the right side of (64.9) become negligible with respect to the first one. Therefore

$$\frac{dW}{dt} = \oint_{\Sigma} \vec{\wp} \cdot \vec{u}_n \, d\Sigma \cong \oint_{\Sigma} v\vec{j} \cdot \vec{u}_n \, d\Sigma = v(t)i(t) \tag{64.10}$$

This last equation proves that the generally used definition of electric power, under variable conditions,

$$p(t) = v(t)i(t) \tag{64.11}$$

obtained as the product of the instantaneous voltage and current in a section of the electric circuit is only an approximation and provides correct results if and only if the dimensions of the circuit are negligible with respect to the wavelength of the electromagnetic quantities.

Of course, this is true for AC circuits operated under sinusoidal conditions at mains (50 or 60 Hz) fundamental frequency. It still holds when the signals are distorted, provided that the frequency bandwidth of the signal remains well below a few hundreds of kilohertz. Should this be not the case, the second and third integrals in (64.9) might become significant enough to make (64.11) not sufficiently accurate.

[1] It is worth reminding that the wavelength of a waveform is related to its frequency by $\lambda = c/f$, c being the speed of light.

64.2.2.2 The Sinusoidal Conditions Nowadays, sinusoidal AC systems are the most widely used to transport and distribute electric energy. It is therefore important to analyze (64.11) under these conditions.

Let us assume that the voltage and current waveforms are sine waves and are described, respectively, by

$$v(t)=\sqrt{2}V\sin(2\pi ft) \tag{64.12}$$

$$i(t)=\sqrt{2}I\sin(2\pi ft+\varphi), \tag{64.13}$$

where

V and I are the rms values of voltage and current, respectively

f is the frequency at which the AC system operates

φ is the phase displacement between the current and voltage waveforms and is considered conventionally positive if the current is leading and negative if the current is lagging

According to (64.11), the instantaneous power is given by

$$p(t)=v(t)i(t)=2VI\sin(2\pi ft)\sin(2\pi ft+\varphi) \tag{64.14}$$

Taking into account that $\sin\alpha\sin\beta=\frac{1}{2}\left[\cos(\alpha-\beta)-\cos(\alpha+\beta)\right]$, (64.14) becomes

$$p(t)=2VI\left\{-\frac{1}{2}\left[\cos(4\pi ft+\varphi)-\cos\varphi\right]\right\}=VI\cos\varphi-VI\cos(4\pi ft+\varphi) \tag{64.15}$$

$v(t)$, $i(t)$, and $p(t)$ are plotted in Figure 64.1 for $V=1\,\text{V}$, $I=1\,\text{A}$, $f=50\,\text{Hz}$, and $\varphi=-\pi/6$.

It can be readily perceived, from (64.15) and Figure 64.1, that the instantaneous power shows an average value $P=VI\cos\varphi$ that differs from zero if $\varphi\neq\pm\pi/2$ and shows a variable term that oscillates about P with a frequency that is twice the voltage and current frequency. The average power P represents the useful power transfer.

The energy flow in a circuit section is given by

$$W=\int_{t_0}^{t_0+t}v(\tau)i(\tau)d\tau \tag{64.16}$$

It can be immediately recognized that, when the integration time t is equal to the signal period T or an integer multiple of T ($t=kT$, k an integer), (64.16) becomes

$$W=\int_{t_0}^{t_0+kT}Pd\tau \tag{64.17}$$

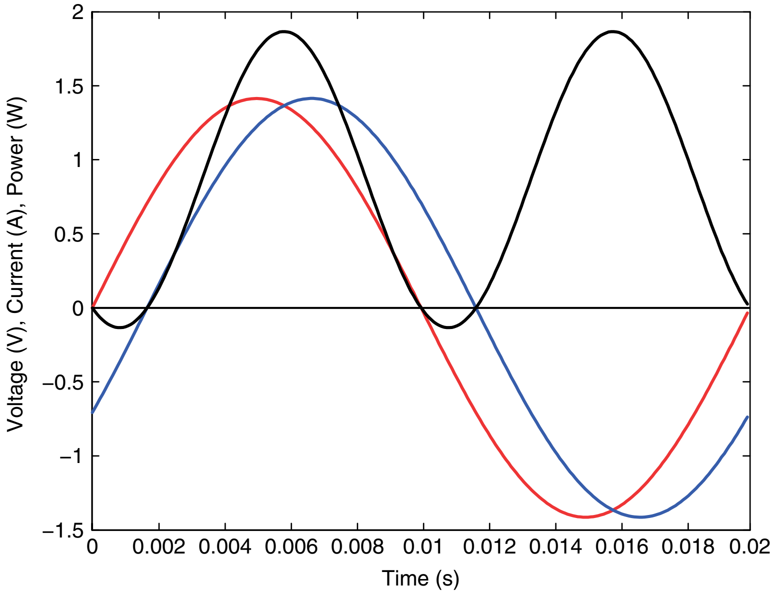


FIGURE 64.1 Voltage (dark gray), current (light gray), and instantaneous power (black) waveforms in a sinusoidal AC system. $V = 1\,\text{V}$, $I = 1\,\text{A}$, $\varphi = -\pi/6$.

It is now interesting to analyze the oscillating part of the instantaneous power to understand whether it shows interesting properties. To do so, let us decompose the current waveform into a component in phase and a component in quadrature ($\pi/2$ phase shift) with the voltage waveform. It is expanding (64.13):

$$\begin{aligned} i(t) &= \sqrt{2}I\sin(2\pi ft)\cos\varphi + \sqrt{2}I\cos(2\pi ft)\sin\varphi \\ &= \sqrt{2}I\sin(2\pi ft)\cos\varphi + \sqrt{2}I\sin\left(2\pi ft + \frac{\pi}{2}\right)\sin\varphi \end{aligned} \tag{64.18}$$

Using the first line of (64.18) in (64.14), we get

$$p(t) = 2VI\cos\varphi\sin^2(2\pi ft) + 2VI\sin\varphi\sin(2\pi ft)\cos(2\pi ft) \tag{64.19}$$

Taking into account that

$$\sin^2\alpha = \frac{1}{2} - \frac{1}{2}\cos 2\alpha$$

and

$$\sin\beta\cos\beta = \frac{1}{2}\sin 2\beta,$$

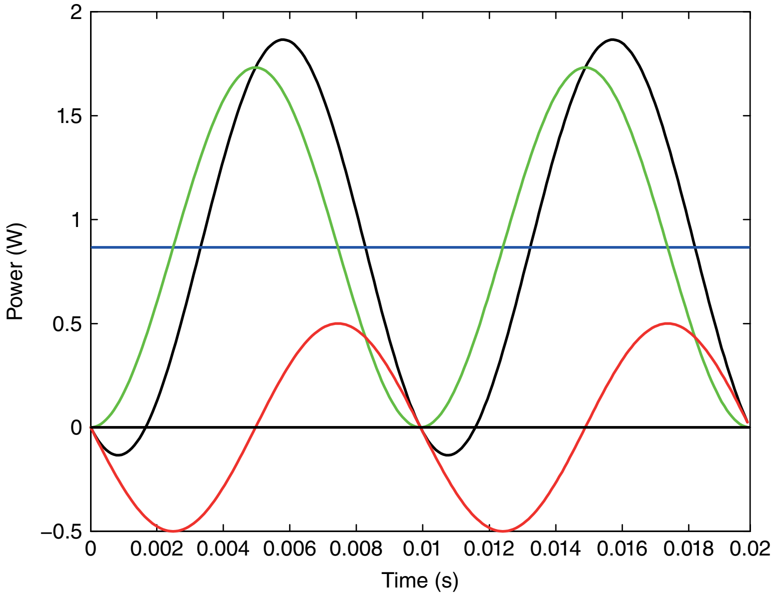


FIGURE 64.2 Instantaneous power components in a sinusoidal AC system: instantaneous power (black), average power (constant gray line), component originated by the in-phase current summed to the average power (light gray), component originated by the quadrature current (dark gray). $V=1\,\text{V}$, $I=1\,\text{A}$, $\varphi=-\pi/6$.

(64.19) becomes

$$p(t)=VI\cos\varphi-VI\cos\varphi\cos(4\pi ft)+VI\sin\varphi\sin(4\pi ft) \qquad (64.20)$$

The three components of power evidenced by (64.20) are plotted in Figure 64.2, again for $V=1\,\text{V}$, $I=1\,\text{A}$, $f=50\,\text{Hz}$, and $\varphi=-\pi/6$. In particular, the black line shows the instantaneous power $p(t)$, the constant line shows the average power P, the light gray line shows the sum of the first two components in (64.20), and the dark gray line shows the third component.

It can be readily checked that if $\varphi=0$, the third term in (64.20) is nil, the average power becomes $P=VI$, and the amplitude of the second oscillating term in (64.20) is VI. This is the situation of a single purely resistive load, supplied by a sinusoidal voltage. For this reason, and since it represents useful power transferred from the supply to the load, the average power P is also called *active power* [4].

On the other hand, it can be readily checked that if $\varphi=\pm\pi/2$, the first and second term in (64.20) are nil and the total instantaneous power is given by the third term and oscillates with zero mean value and peak value equal to VI. This is the situation of a single purely reactive load (inductor or capacitor) supplied by a sinusoidal voltage. For this reason, the peak value $Q=VI\sin\varphi$ of this component of the instantaneous power is called *reactive power*. Under sinusoidal conditions the reactive power has the property of quantifying the amount of useless power transferred due to the presence of reactive elements. It can be used to design and size passive reactive compensators to minimize this power component [4].

Therefore, the power properties of an electric device can be fully represented by its active and reactive powers:

$$P = VI\cos\varphi \tag{64.21}$$

$$Q = VI\sin\varphi \tag{64.22}$$

The following quantity can be also defined:

$$S = \sqrt{P^2 + Q^2} = VI \tag{64.23}$$

S is called *apparent power* and represents the maximum active power that can be transferred under $\varphi = 0$ conditions. It is often referred to as the *design power*, since it can be obtained as the product of the rated voltage by the rated current.

P is measured in *watts* (W), Q in *reactive volt-amperes* (VA_r), and S in *volt-amperes* (VA). It is worth noting that the dimension is always that of a power, but different names have been given to the units to reinforce the different physical properties of the considered power components.

The following quantity is also defined:

$$\lambda = \frac{P}{S} = \cos\varphi \tag{64.24}$$

and is called *power factor*. It can be readily checked that it is always $\lambda \le 1$ and that λ represents an index of how efficiently power is transferred for given values V and I of voltage and current.

64.3 MEASUREMENT METHODS

64.3.1 DC Conditions

64.3.1.1 Measurement Method According to (64.6), power in DC circuits can be measured by measuring voltage and current by means of a voltmeter and an ammeter, as shown in Figure 64.3. Under ideal conditions, the two instruments can be connected, as shown in Figure 64.3, with the voltmeter connected before the ammeter, or with the voltmeter connected after the ammeter, directly in parallel with the load, indifferently.

However, real instruments feature an internal resistance that, depending on the employed connection, yields a measured value of power different from the expected one. If the connection shown in Figure 64.3 is considered again, the load resistance and those of the employed instruments are connected as shown in Figure 64.4.

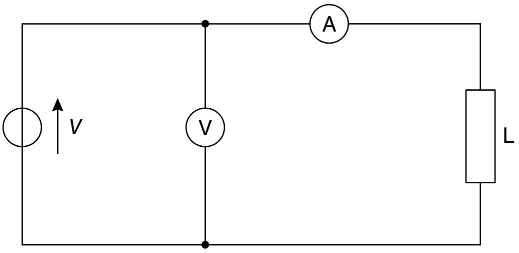


FIGURE 64.3 Instrument connection to measure the DC power drawn by a DC load L supplied by a DC voltage *V*. Power is measured by means of a voltmeter (V) and an ammeter (A).

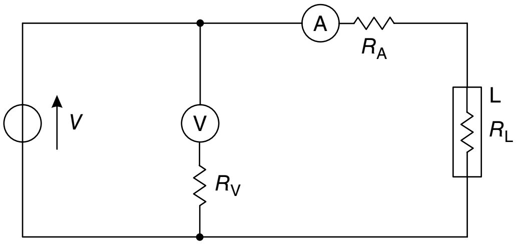


FIGURE 64.4 Instrument connection to measure the DC power drawn by a DC load L supplied by a DC voltage *V*. The ammeter is connected in series with the load. The load resistance, R_L, and the ammeter and voltmeter internal resistances, R_A and R_V, respectively, are also shown.

It can be readily perceived that the ammeter internal resistance R_A is connected in series with the load resistance R_L. Therefore, the current I_m measured by the ammeter is still the current I flowing in the load, but the voltage V_m measured by the voltmeter differs from the voltage V across the load and is given by

$$V_m = V + R_A I \tag{64.25}$$

Therefore, the measured power P_m is given by

$$P_m = V_m I_m = VI + R_A I^2 = P + P_A, \tag{64.26}$$

that is, the measured power is the actual power taken by the load plus the power dissipated by the internal circuits of the ammeter. This means that the internal resistance of the ammeter is responsible for a systematic contribution that becomes significant when this internal resistance is not much lower (some orders of magnitude) than that of the load. If this is the case, a correction must be applied for this systematic contribution.

On the other hand, if the voltmeter is connected directly in parallel with the load, as shown in Figure 64.5, voltage V_m measured by the voltmeter is voltage V on the load. The current measured by the ammeter (I_m) is given by

$$I_m = I + \frac{V}{R_V} \tag{64.27}$$

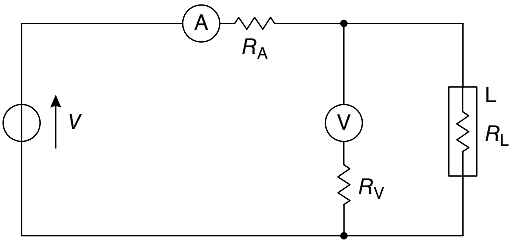


FIGURE 64.5 Instrument connection to measure the DC power drawn by a DC load L supplied by a DC voltage *V*. The voltmeter is connected in parallel with the load. The load resistance, R_L, and the ammeter and voltmeter internal resistances, R_A and R_V, respectively, are also shown.

Therefore, the measured power P_m is given by

$$P_m = V_m I_m = VI + \frac{V^2}{R_V}, \tag{64.28}$$

that is, the measured power is the actual power taken by the load plus the power dissipated by the internal circuits of the voltmeter. This means that the internal resistance of the voltmeter is responsible for a systematic contribution that becomes significant when this internal resistance is not much higher (some orders of magnitude) than that of the load. If this is the case, a correction must be applied for this systematic contribution.

64.3.1.2 Uncertainty Evaluation After having applied the required corrections, if needed, for the above systematic effects, uncertainty evaluation is a typical example of combined uncertainty evaluation.

Let us assume that the standard uncertainty values *u*(*V*) and *u*(*I*) have been evaluated for the measured values of voltage (V_m) and current (I_m) respectively, following either a type A or a type B evaluation method, as suggested by the Guide to the Expression of Uncertainty in Measurement (GUM) [5]. According to the GUM, the combined standard uncertainty *u*(*P*) is given by [5]

$$u(P) = \sqrt{\left(\left.\frac{\partial P}{\partial V}\right|_{V_m}\right)^2 u^2(V) + \left(\left.\frac{\partial P}{\partial I}\right|_{I_m}\right)^2 u^2(I) + 2\left.\frac{\partial P}{\partial V}\right|_{V_m}\left.\frac{\partial P}{\partial I}\right|_{I_m} u(V)u(I)r(V,I)}, \tag{64.29}$$

where *r*(*V*, *I*) is the correlation coefficient that takes into account the possible correlation among the voltage and current measurements. By developing the partial derivatives in (64.29), we get

$$u(P) = \sqrt{I_m^2 u^2(V) + V_m^2 u^2(I) + 2 I_m V_m u(V)u(I)r(V,I)}. \tag{64.30}$$

64.3.2 AC Conditions

64.3.2.1 Single-Phase Circuits According to (64.21), (64.22), and (64.23), the power exchange in a section of a circuit operated under sinusoidal AC conditions cannot be fully defined by measuring only the rms values of voltage and current. Dedicated instruments must be employed, and the most important is the wattmeter.

The schematics of the instrument connections are shown in Figure 64.6, where a wattmeter W is used to measure the active power (64.21). This instrument has a current port, connected in series with the load, and a voltage port, connected in parallel. Like the ammeters and voltmeters, its internal circuits feature a finite impedance, different from the ideal one, which is zero for the ampermetric circuit and infinite for the voltmetric circuit. Therefore a correction should be applied, as seen in the case of DC power measurement, to subtract the power drawn by the wattmeter internal circuit to the measured one. In general, the internal impedance of the voltmetric circuits is almost purely resistive, so that its effect can be more easily compensated. For this reason, the connection shown in Figure 64.6, where the voltmeter and the voltmetric circuits of the wattmeter are connected in parallel with the load, is preferred to the connection with the voltmetric circuits connected before the ampermetric ones.

According to Figure 64.6, the wattmeter provides the measured value P_m of the active power, the voltmeter the measured value V_m of the voltage, and the ammeter the measured value I_m of the current. Therefore, the measured value of the apparent power is obtained as $S_m = V_m I_m$, and the absolute value of the measured reactive power is obtained as $Q_m = \sqrt{S_m^2 - P_m^2}$. If the nature of the load (inductive or capacitive) is known, the sign of Q_m is also known. Otherwise, another instrument, called *varmeter*,[2] has to be connected, in the same way as the wattmeter, to measure the reactive power in a direct way.

The voltmeter and ammeter in Figure 64.6 are used not only to measure the apparent power but also to control the values of voltage and current in order not to overload the internal circuits of the wattmeter.

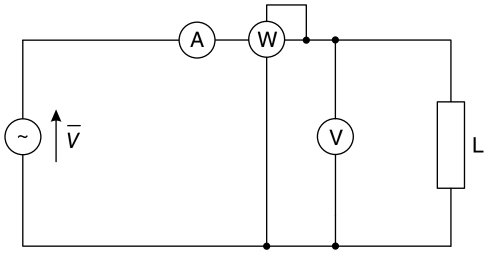


FIGURE 64.6 Instrument connection to measure the AC power drawn by a load supplied by a sinusoidal AC voltage $\bar{V}$. The active power is measured by a wattmeter (W).

[2] When sinusoidal AC conditions are considered, varmeters are generally wattmeters with an additional circuit that delays the input voltage by $\pi/2$.

64.3.2.2 Three-Phase Circuits Nowadays, most electric systems operated under sinusoidal AC conditions are three-phase systems. They can be three-wire systems, such as the one represented in Figure 64.7, or four-wire systems. The system in Figure 64.7 can be fully characterized by the three line-to-line voltages $\bar{V}_{ab}, \bar{V}_{bc}, \bar{V}_{ca}$ and the three line currents $\bar{I}_a, \bar{I}_b, \bar{I}_c$.

The most efficient way to operate such systems is when the three line-to-line voltages have the same amplitude and are phase-shifted by $\pi/3$ with respect to each other and the load is composed by three equal single-phase loads *Y* or Δ connected, so that the three line currents have the same amplitude and are phase-shifted by $\pi/3$ with respect to each other, as shown in the phasorial diagram of Figure 64.8. This situation is called *symmetrical* and *balanced*, and, under this condition, three-wire and four-wire systems behave in the same way, since the current in the fourth wire is always zero.

The diagram in Figure 64.8 shows also the line voltages $\bar{V}_a, \bar{V}_b, \bar{V}_c$, referred to the theoretical neutral point *O* (the center of gravity of the triangle of the line-to-line currents). The sum of these three line voltages, in phasorial terms, is nil, and it is $V_{ab} = \sqrt{3}V_a$.

It can be readily proved [4] that, under symmetrical and balanced conditions, the active power of a three-phase system is given by

$$P = 3V_a I_a \cos\varphi = \sqrt{3}V_{ab} I_a \cos\varphi, \tag{64.31}$$

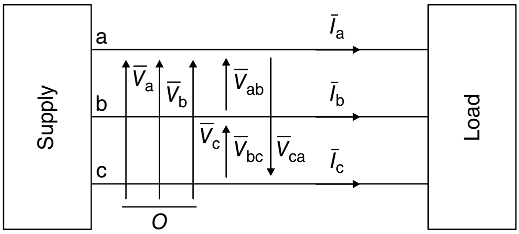


FIGURE 64.7 Three-wire, three-phase system.

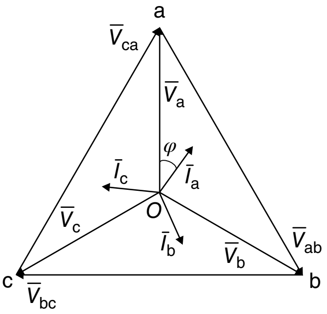


FIGURE 64.8 Phasorial diagram of voltages and currents in a symmetrical and balanced three-wire, three-phase system.

where φ is the phase angle between the line voltage phasor $\bar{V}_{\mathrm{a}}$ and the line current phasor $\bar{I}_{\mathrm{a}}$. In phasorial terms, (64.31) can be rewritten as

$$P = 3\left(\bar{V}_{\mathrm{a}} \cdot \bar{I}_{\mathrm{a}}\right). \tag{64.32}$$

It can be also proven that, if the symmetrical and balanced conditions are not met, as it may happen under practical situations, the active power is given by

$$P = \bar{V}_{\mathrm{a}} \cdot \bar{I}_{\mathrm{a}} + \bar{V}_{\mathrm{b}} \cdot \bar{I}_{\mathrm{b}} + \bar{V}_{\mathrm{c}} \cdot \bar{I}_{\mathrm{c}} \tag{64.33}$$

Since the system is a three-wire system, it is also $\bar{I}_{\mathrm{a}} + \bar{I}_{\mathrm{b}} + \bar{I}_{\mathrm{c}} = 0$, and, hence $\bar{I}_{\mathrm{b}} = -\bar{I}_{\mathrm{a}} - \bar{I}_{\mathrm{c}}$. Equation 64.33 can be, therefore, written as

$$P = \bar{V}_{\mathrm{a}} \cdot \bar{I}_{\mathrm{a}} - \bar{V}_{\mathrm{b}} \cdot \left(\bar{I}_{\mathrm{a}} + \bar{I}_{\mathrm{c}}\right) + \bar{V}_{\mathrm{c}} \cdot \bar{I}_{\mathrm{c}} = \left(\bar{V}_{\mathrm{a}} - \bar{V}_{\mathrm{b}}\right) \cdot \bar{I}_{\mathrm{a}} + \left(\bar{V}_{\mathrm{c}} - \bar{V}_{\mathrm{b}}\right) \cdot \bar{I}_{\mathrm{c}}. \tag{64.34}$$

According to the diagram in Figure 64.8, it can be seen that, under whatever conditions, it is $\bar{V}_{\mathrm{ab}} = \bar{V}_{\mathrm{a}} - \bar{V}_{\mathrm{b}}$ and $\bar{V}_{\mathrm{cb}} = \bar{V}_{\mathrm{c}} - \bar{V}_{\mathrm{b}}$. Therefore, (64.34) becomes

$$P = \bar{V}_{\mathrm{ab}} \cdot \bar{I}_{\mathrm{a}} + \bar{V}_{\mathrm{cb}} \cdot \bar{I}_{\mathrm{c}}. \tag{64.35}$$

This last equation shows that the total power across a section of a three-wire, three-phase system can be measured by means of only two wattmeters, connected as shown in Figure 64.9. This connection is generally known as the Aron connection and can be generalized to the four-wire systems, where three wattmeters are needed, with the current circuit connected to three wires and the voltage circuits connected across the same wires as those of the current circuits and the fourth wire, to measure the total power.

A similar theorem can be proven, under sinusoidal conditions, for the total reactive power. It is

$$Q = \left\|\bar{V}_{\mathrm{a}} \times \bar{I}_{\mathrm{a}}\right\| + \left\|\bar{V}_{\mathrm{b}} \times \bar{I}_{\mathrm{b}}\right\| + \left\|\bar{V}_{\mathrm{c}} \times \bar{I}_{\mathrm{c}}\right\| = V_{\mathrm{a}} I_{\mathrm{a}} \sin\varphi_{\mathrm{a}} + V_{\mathrm{b}} I_{\mathrm{b}} \sin\varphi_{\mathrm{b}} + V_{\mathrm{c}} I_{\mathrm{c}} \sin\varphi_{\mathrm{c}}, \tag{64.36}$$

where φ_{a}, φ_{b}, φ_{c}, are the phase angles between the line voltages and currents of phases a, b, and c, respectively. Therefore, the total reactive power can be measured with two varmeters connected in the same way as the wattmeters in Figure 64.9.

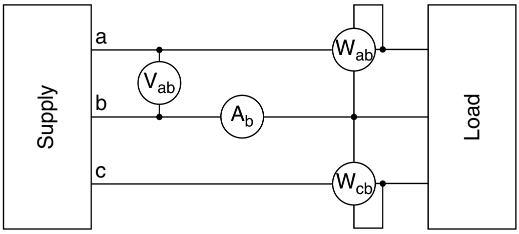


FIGURE 64.9 Aron wattmeter connection in a three-wire, three-phase system.

64.4 WATTMETERS

64.4.1 Architecture

Wattmeters and energy meters are probably among the oldest instruments built since the industrial exploitation of electric energy. The first wattmeters and energy meters were electromechanical instruments.

Electrodynamical actions were exploited to realize wattmeters [6], and this operating principle remained the same until the electronic circuits replaced, in the second half of the twentieth century, the electromechanical structures. Wattmeters based on thermal principles and analog multiplier structures have replaced the electrodynamic ones, until the more modern architectures based on digital signal processing (DSP) techniques have replaced also those instruments.

Energy meters, on the other hand, were based on the same principle as the magnetic rotating field obtained by Galileo Ferraris and used to design and realize AC electric machines. In the energy meters based on this principle, a thin metallic disk acts as the motor rotor and rotates at a speed proportional to the electric active power. The number of turns, counted by a mechanical system, is proportional to the active energy flowing in the metering sections [6]. These energy meters, known as *Ferraris meters*, were first produced at the very beginning of the twentieth century, represented the most used and widespread instruments all over the world, and millions of them are still in use, though they are now being slowly replaced by more modern instruments based on DSP structures.

Since the electromechanical and analog electronic structures are slowly disappearing, only the modern DSP-based structures will be briefly discussed in the following. The block diagram of their architecture, for a single-phase instrument, is shown in Figure 64.10.

The voltage $v(t)$ and current $i(t)$ signals are sensed by transducer circuits and signal conditioning circuits. The transducers play the important role of adjusting the signal

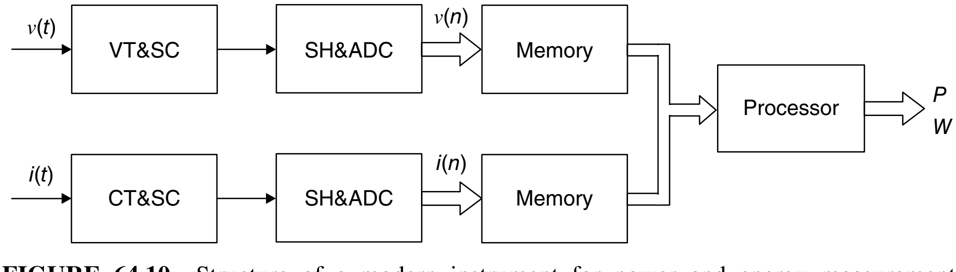


FIGURE 64.10 Structure of a modern instrument for power and energy measurement. CT&SC, current transducer and signal conditioning; SH&ADC, sample and hold and analog-to-digital converter; VT&SC, voltage transducer and signal conditioning. $v(t)$ and $i(t)$ voltage and current time signals, respectively. $v(n)$ and $i(n)$ sequences of the voltage and current samples, respectively.

level to that of the instrument's electronics and guarantee the necessary galvanic insulation with given accuracy over a given frequency band. Since their performance can be critical in ensuring the desired accuracy of the whole instrument, the next paragraph will be specifically dedicated to the voltage and current transducers and their structure and performance.

The signal conditioning circuits are needed to further adjust the input signals to the features of the subsequent blocks of the instrument. Among these circuits an important role, also in determining the final accuracy is played by the antialiasing filters [7].

The voltage and current signals are then sampled and converted into digital by means of sample and hold devices and analog-to-digital converters [7]. These two devices convert the original, continuous-time signals into sequences of samples, that is, the discrete-time signals $v(n)$ and $i(n)$ [7] that can be stored into the instrument memory and processed by the processor unit, according to dedicated signal processing algorithms. These algorithms are the core of the modern instruments and allow one to implement different measurement functions on the same instrument, such as active, reactive, and apparent power measurements, as well as their integration over time, so that active and reactive energy can be measured for billing purposes.

Three-phase instruments can be readily obtained from this structure simply by adding more input channels, to acquire the voltage and current signals as shown in previous Section 64.3.2.2, and by implementing the required algorithm to process them.

64.4.2 Signal Processing

As mentioned in the previous section, the signal processing algorithms represent the core of the modern instruments for power and energy measurement. Therefore, they are worth a dedicated analysis, since an incorrect choice of these algorithms or an incorrect choice of the sampling strategy may affect the accuracy of the obtained measurement results dramatically.

Let us start from the correct selection of the sampling frequency, since this is the most important point to ensure correct results. It is well known that, to be correctly sampled, the input signals must be upper limited in frequency [7]. Let us suppose, accordingly, that the voltage and current signals are periodic in time over a period T; that the voltage signal $v(t)$ is upper limited, in frequency, by the Mth harmonic component; and that the current signal $i(t)$ is upper limited, in frequency, by the Nth harmonic component.

According to the mathematical derivations reported in the previous sections, all relevant power quantities are obtained from the instantaneous power $p(t)=v(t)i(t)$,[3] whose samples are obtained from the samples of voltage and current as $p(n)=v(n)i(n)$. The

[3] This holds in general and especially under nonsinusoidal conditions. Since the modern instruments are more and more used for power quality measurements, when the signals are distorted, correct sampling of the instantaneous power is a critical issue with these instruments and must be carefully considered.

employed sampling frequency must therefore satisfy the sampling theorem for the instantaneous power $p(t)$.

In the frequency domain, the instantaneous power is obtained as the convolution of the frequency-domain versions $V(f)$ and $I(f)$ of $v(t)$ and $i(t)$, respectively. Therefore, the maximum harmonic component of $p(t)$ has order $M+N$. The ideal method for obtaining the correct number of samples of $p(t)$ is that of sampling $v(t)$ and $i(t)$ with a sampling period T_s so that [7]

$$T=\left[2(M+N)+1\right]T_s \tag{64.37}$$

where $2(M+N)+1$ is the minimum number of samples required to sample $p(t)$ correctly.

Assuming $K=M+N$, the active power can be obtained as the average value of the instantaneous power as

$$P=\frac{1}{j(2K+1)}\sum_{k=0}^{j(2K+1)-1} v(k)i(k) \tag{64.38}$$

$j=1, 2, 3, \ldots$ being an integer.

If $j=1$ is chosen, the obtained active power value is averaged over a single period. If higher values are taken for j, a longer averaging period can be considered. In particular, if $j=50$ is taken for systems operated at 50 Hz fundamental frequency and $j=60$ is taken for systems operated at 60 Hz fundamental frequency, the active power value P_1 is obtained, averaged over a 1 s interval of time.

It is then possible to easily obtain the total active energy flowing through the metering section in a time interval of k seconds as

$$W=\sum_k P_{1k}\cdot \tag{64.39}$$

Apparent and reactive powers can be also obtained from the samples of voltage and current by implementing one of the definitions given in Section 64.2.

It is worthwhile emphasizing the high flexibility of this modern DSP-based structure, which allows the implementation of different possible definitions. This is a great advantage, especially when power components under nonsinusoidal conditions have to be measured. On the other hand, since, as it will be shown later in this chapter, different definitions may yield, under nonsinusoidal conditions, different measured values for the same conditions, it is important that the adopted algorithm is always declared to avoid gross mistakes.

64.5 TRANSDUCERS

Power and energy measurements are performed in power systems at all voltage levels, from low-voltage to ultrahigh-voltage systems. It can be immediately perceived that these voltage levels are not compatible with the input dynamics of the employed instruments, especially the electronic ones. Moreover, galvanic insulation must be ensured, for obvious safety reasons, between the power system and the instruments.

This requires the use of voltage and current transducers that adjust the voltage and current values to those compatible with the instrument dynamics and ensure the required insulation level.

The impact of these transducers on the metrological performance of the whole measuring system is so significant that it is worth covering this topic in a dedicated section of this chapter.

64.5.1 Current Transformers

The most common current transducer in power applications is the current transformer (CT). This transducer performs a reduction of the value of the AC current to be measured in order to adjust it to the input dynamics of the employed ammeter. Another important feature of this transducer is that it provides galvanic insulation between the power network and the measurement equipment, thus granting operator safety.

From an ideal point of view, the CT is an ideal transformer whose primary is series connected to the line carrying the current to be measured. Its secondary winding is short-circuited by means of an ideal ammeter as shown in Figure 64.11.

The current measured by the ammeter, $\bar{I}_A$, depends on the primary current of the transformer, $\bar{I}_1$, according to the theoretical ratio K_T:

$$\bar{I}_A = -\bar{I}_2 = -\left(-\frac{N_1}{N_2}\bar{I}_1\right) = \frac{1}{K_T}\bar{I}_1 \tag{64.40}$$

where N_1 and N_2 are the numbers of turns of the primary and secondary windings, respectively.

In ideal CTs, K_T is a real constant. Consequently, (64.40) shows that the measured current $\bar{I}_A$ is in phase with current $\bar{I}_1$ and is simply scaled by a constant factor.

Since the CT is usually employed to perform a reduction of the current value, K_T is a number much greater than one. This implies that if the secondary winding is left

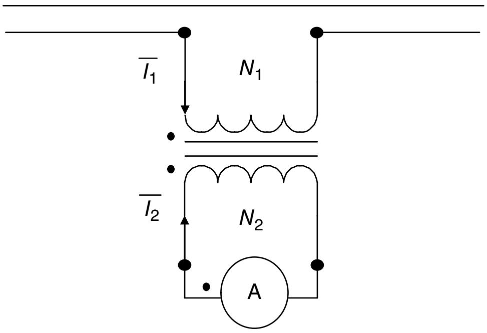


FIGURE 64.11 Ideal CT.

open, the voltage induced at the secondary terminal will be $\overline{V}_2 = K_T \overline{V}_1 \gg \overline{V}_1$. That is a very dangerous condition that must be avoided since it can result in electric arcs or explosions. Always for safety reasons, one of the terminal of the secondary winding must be connected to the ground in order to force its potential to zero. The absence of this connection can result in the presence of a dangerous voltage at the secondary terminals due to the capacitive coupling between the primary and secondary windings. The rated secondary output of CTs is 1 A or 5 A.

64.5.1.1 Measurement Errors Ideal CTs feature, of course, zero measurement errors. This is not the case for real CTs, and, in order to evaluate their errors and, hence, their contribution to the uncertainty of the whole measurement chain, let us now take into account the model of a real CT (Fig. 64.12).

This model considers the resistance and leakage reactance of the two windings, as well as the magnetizing branch. An impedance $\overline{Z}_b$ representing the electric load of the CT has been also added. In order to evaluate the actual relationship between the currents at the primary and secondary windings of the CT, a phasor analysis shall be done. For the sake of simplicity, let us suppose that the CT operates with a load Z_b equal to zero. Under these conditions the phasor graph representing the CT is reported in Figure 64.13. In the graph, current $\bar{I}_0$ flowing in the magnetizing branch has been splitted into the magnetizing current $\bar{I}_m$, and the core and iron losses current $\bar{I}_c p$.

Figure 64.13 shows that (64.40) is no longer valid. In fact, it can be noticed that there are a phase displacement and a difference in amplitude between $\bar{I}_1$ and $-\overline{K_T}\,\bar{I}_2$.

For this reason two errors can be defined: the phase error ε and the ratio error η.

By considering that the phase error is generally small, it can be written as

$$\varepsilon \cong \sin\varepsilon = \frac{\overline{AB}}{\overline{OA}} = \frac{I_0}{I_1}\sin\theta \tag{64.41}$$

$$\theta = \varphi_0 - \varphi_2 \tag{64.42}$$

where φ_2 is the angle between $\bar{I}_2$ and $\overline{E}_2$, the secondary no-load voltage, and φ_0 is the angle between $\bar{I}_0$ and $\overline{E}_1$, the voltage applied to the magnetizing branch.

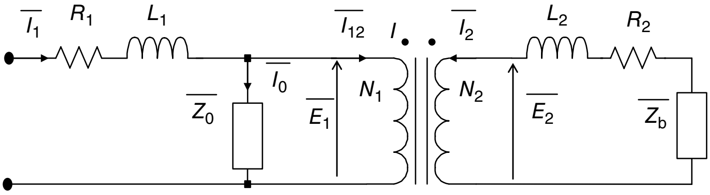


FIGURE 64.12 Model of a real CT.

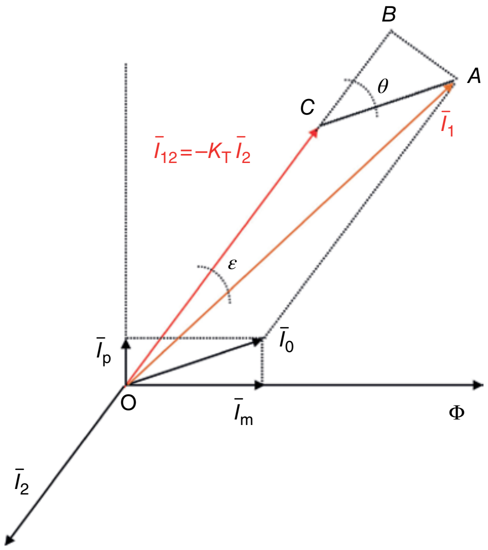


FIGURE 64.13 Phasor graph of a real CT.

In a similar way the ratio error of the CT can be defined. Let us define the rated ratio K_r as the ratio between the rated primary current I_{1r} and the rated secondary current I_{2r}. Again under the assumption of small errors, the ratio error can be written as

$$\eta = \frac{I_{1r} - I_1}{I_1} \cong \frac{K_r I_2 - K_T I_2 - I_0 \cos\theta}{I_1} \cong \frac{K_r I_2 - K_T I_2}{K_r I_2} - \frac{I_0}{I_1}\cos\theta = \frac{K_r - K_T}{K_r} - \frac{I_0}{I_1}\cos\theta \tag{64.43}$$

Equations 64.41 and 64.43 show that the CT errors depend on its magnetizing current. Therefore the improvement of the CT performances requires the optimization of the magnetic core, in terms of employed material and dimensions, and the numbers of turns of the transformer windings. Equation 64.43 also shows that the ratio error can be set to zero for a chosen current by properly defining the rated ratio of the CT.

It is important to notice that also the series parameters of the transformer play a role in the error definition. In fact, considering the model of Figure 64.12, for a given current $\bar{I}_2$, voltage $\bar{E}_2$ depends on the series parameters R_2, L_2, and $\bar{Z}_b$:

$$\bar{E}_2 = \left(\bar{Z}_b + R_2 + j\omega L_2\right)\cdot \bar{I}_2 \tag{64.44}$$

$$I_0 = \frac{E_1}{Z_0} = \frac{E_2}{K_T Z_0} \tag{64.45}$$

From (64.44) and (64.45), the role of the load becomes evident: the higher is the load, the higher is I_0 and, consequently, the errors.

For this reason, CTs are characterized by a rated load, expressed in VA at a given rated current, that represents the maxim value of the impedance that can be connected to the CT in order to guarantee the errors for the rated accuracy class. The common accuracy classes for CTs employed in power and energy measurements are 0.1, 0.2, 0.5, and 1. The accuracy specifications for these classes are listed in Table 64.1.

64.5.1.2 Saturation A typical problem of CTs is the saturation due to the nonlinearities of the magnetic core. Two kinds of saturations can be defined. The first one can be defined as symmetrical or AC saturation. This saturation occurs when the value of the AC current to be transduced forces the core to work with high values of magnetic flux so that it leaves the linear range of the *B*–*H* curve (Fig. 64.14).

An example of AC saturation is reported in Figure 64.15. The primary current is a sine wave at the mains frequency, but its amplitude results in a too high magnetic flux that, due to the nonlinearity of the *B*–*H* curve, is distorted in the peak area. Since

TABLE 64.1 Accuracy Specifications for CTs

Accuracy Class	± Percentage Ratio Error				± Phase Error (Minutes)			
% of rated the current	5	20	100	120	5	20	100	120
0.1	0.4	0.2	0.1	0.1	15	8	5	5
0.2	0.75	0.35	0.2	0.2	30	15	10	10
0.5	1.5	0.75	0.5	0.5	90	45	30	30
1	3	1.5	1	1	180	90	60	60

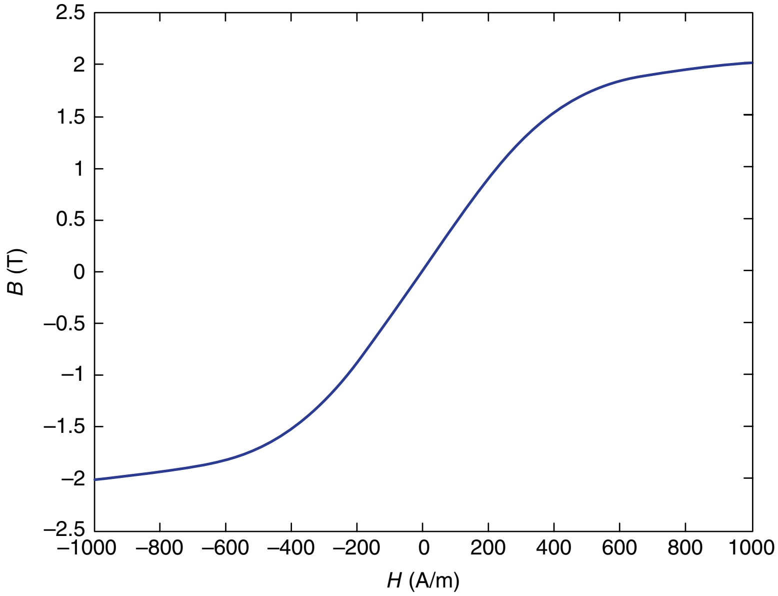


FIGURE 64.14 *B*–*H* characteristic of the core.

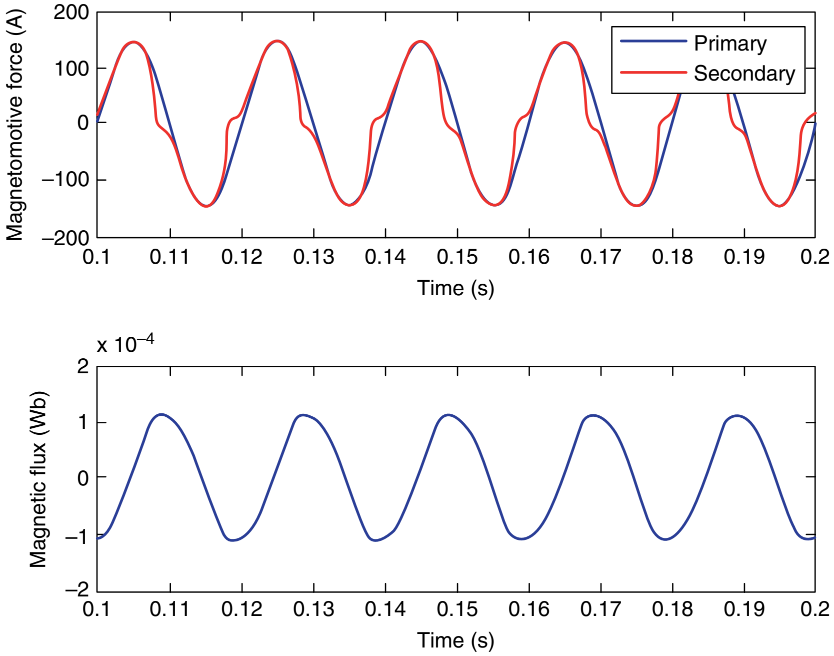


FIGURE 64.15 AC saturation.

current and flux are almost in quadrature, the effect of this distortion can be seen in the area of the zero crossing of the secondary current.

The second kind of saturation can be defined as asymmetrical saturation and is due to the presence of unidirectional components in the primary current or to the remanence in the magnetic core. In this condition, the magnetic core works on an asymmetrical loop in the *B–H* plane. Because of this, if the unidirectional component is not too high, the AC component of the current will cause a magnetic flux that is distorted only for positive or negative values. An example of asymmetrical saturation due to a unidirectional transient of the primary current is reported in Figure 64.16.

64.5.1.3 Bandwidth A CT is usually designed in order to properly work at the mains frequency. The definition of its errors, as previously reported, was derived by considering the CT working with a sinusoidal current at the mains frequency. It is evident that by changing the frequency of analysis, the values of the CT parameters change, so its behavior. In particular the core losses will change, as well as the *B–H* magnetization loop. The amount of these variations strongly depends on the CT design, in particular on materials and geometry. Moreover it has to be considered that a CT is a nonlinear system. The definition of its transfer function, which could be theoretically used in order to compensate for the errors, suffers from model uncertainty. In fact it does not take into account the effect of the intermodulation due to the CT nonlinearities. This effect can be significant and may lead the specifications for the rated accuracy class to be exceeded. In general, the −3 dB bandwidth of a CT does not exceed 1 kHz.

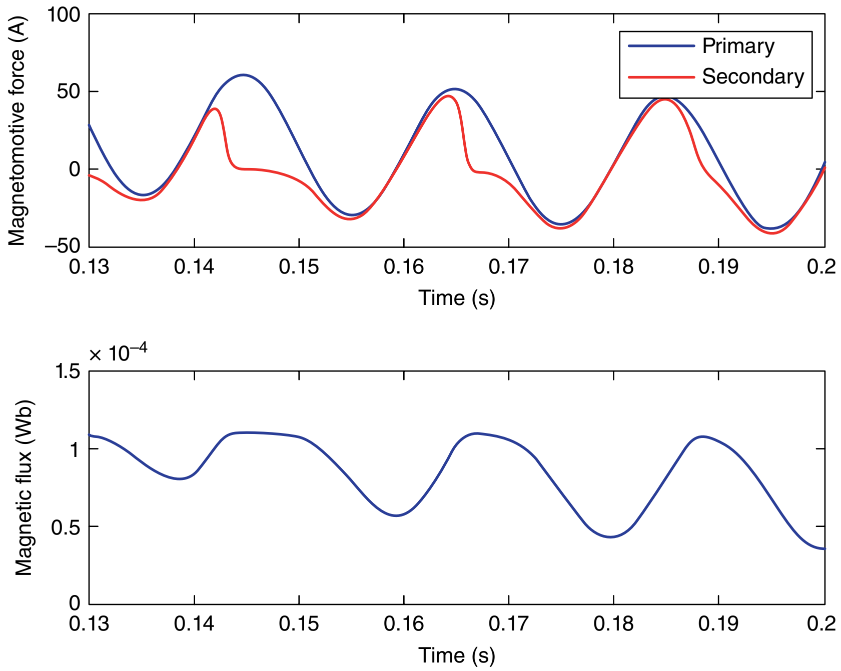


FIGURE 64.16 Asymmetrical saturation.

64.5.2 Hall-Effect Sensors

In order to overcome the typical problems of traditional CTs, a good solution is represented by the zero-flux transducers and in particular those based on Hall-effect sensors. A Hall-effect sensor is a transducer that converts a magnetic field B into a voltage V_H. In particular, due to the Hall effect, a voltage difference V_H is generated at the terminals of a semiconductor[4] plate, when a current I flows in the plate and the plate itself is placed inside a magnetic field B, according to the topology shown in Figure 64.17. It is

$$V_H = B \cdot I \cdot K_H \tag{64.46}$$

where K_H is a constant that depends on the sensor dimensions and material.

This sensor is used to implement a zero-flux transducer in a closed-loop configuration. The typical solution is that shown in Figure 64.18. The Hall-effect sensor is placed inside an air gap in the magnetic core and senses the field generated by the current I_1. The sensor output voltage controls a current generator that provides the

[4] Actually, the Hall effect is present also in conductor plates, but it is generally much lower than in semiconductors, so that only semiconductors are employed in Hall-effect sensors.

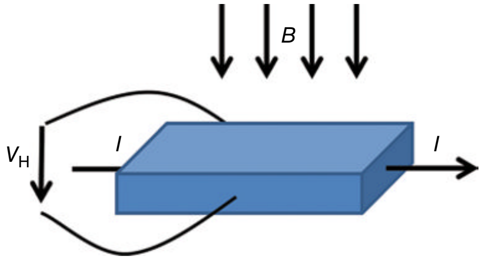


FIGURE 64.17 Hall effect.

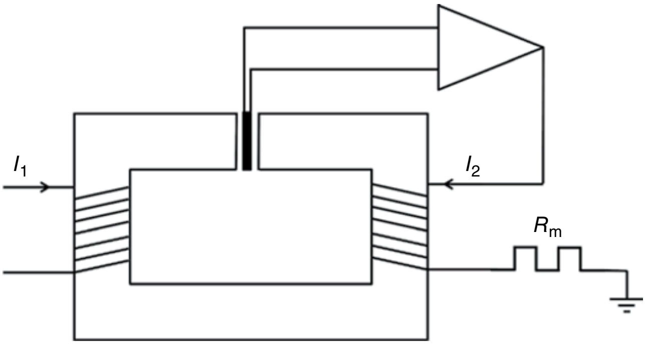


FIGURE 64.18 Zero-flux configuration based on a Hall-effect sensor.

current I_2 required to zero the magnetic field in the core. In this condition the voltage drop V_m on resistance R_m is

$$V_m = I_1 \cdot R_m \cdot \frac{N_1}{N_2} \tag{64.47}$$

The advantage of this transducer, with respect to the traditional CTs, is that it does not suffer from saturation problems, the dynamic performances are basically those of the current generator, and it has a much wider bandwidth. In general, a bandwidth from DC up to 100 kHz can be obtained with these transducers, with an accuracy of 0.1% of the rated current for normal applications.

64.5.3 Rogowski Coils

A Rogowski coil is a simple coil wound on a toroidal nonferromagnetic core. The working principle of Rogowski coils is based on the Ampere law (Fig. 64.19):

$$\oint_L \vec{H} \cdot d\vec{l} = \iint_S \vec{J} \cdot d\vec{s} \tag{64.48}$$

where $\vec{H}$ is the magnetic field, L is a closed line, S is the surface enclosed in L, and $\vec{J}$ is the current density vector.

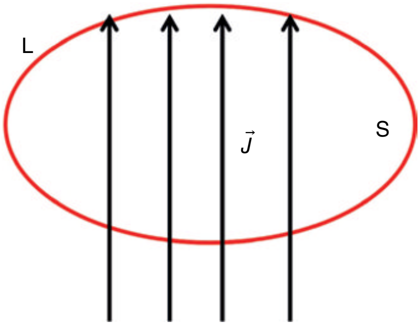


FIGURE 64.19 Ampere law.

Equation 64.48 can be rewritten as

$$\oint_{L} \frac{\vec{B}}{\mu_0} \cdot d\vec{l} = I_S \tag{64.49}$$

where I_S is the current flowing through surface S. Therefore by integrating the magnetic field along a closed line around a conductor, it is possible to evaluate the current flowing in that conductor.

Let us now consider a toroidal winding wound along line L. The voltage induced at the winding terminals will be (Fig. 64.20)

$$v_c = M\frac{di}{dt} - R_c i_c - (M + L_c)\frac{di_c}{dt} \tag{64.50}$$

where M is the mutual inductance between the winding and the conductor carrying current i, R_c and L_c are the resistance and the inductance of the coil, and i_c is the current flowing in it (Fig. 64.20).

Therefore if voltage v_c is measured with an ideal voltmeter, having infinite impedance, (64.50) can be rewritten as follows:

$$v_c = M\frac{di}{dt} \tag{64.51}$$

Hence by integrating the voltage induced in the coil, it is possible to evaluate the current i flowing in a conductor crossing the toroid.

The Rogowski coil can suffer from crosstalk due to a magnetic field produced by currents not crossing the coil. In fact an external magnetic flux can couple the coil through the central hole of the toroid. In order to reduce this effect, a compensating turn is placed into the toroid and connected in series with the main coil to compensate for the voltage induced by external fields (Fig. 64.21). Another, more accurate method to reduce crosstalk requires to wound an even number of layers of turns and connecting them in series.

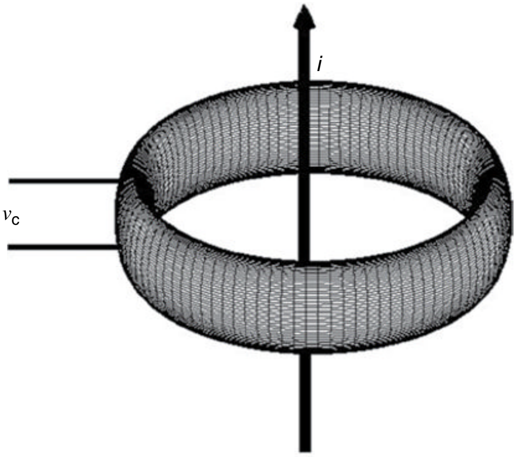


FIGURE 64.20 Rogowski coil.

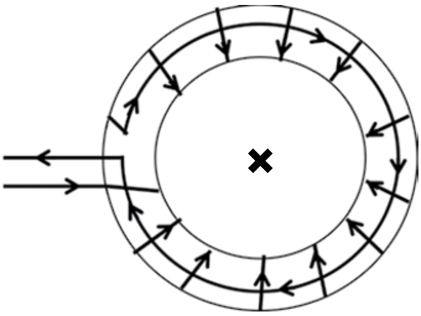


FIGURE 64.21 Compensation loop.

The Rogowski coil does not saturate and is characterized by wide bandwidth and range of measurement. Moreover it guarantees galvanic insulation. Typical values are a bandwidth up to 1 MHz, primary currents up to MA, and insulation levels up to several hundreds of kilovolt.

64.5.4 Voltage Transformers

The most widely used voltage transducer is the voltage transformer (VT). Similarly to the CT, this transducer is basically a transformer that performs a reduction of the value of the quantity to be measured. In particular it performs a known reduction of the amplitude of an AC voltage, in order to adjust it to the input dynamics of the employed voltmeters. Like the CT, the VT provides galvanic insulation between the power network and the measurement equipment.

The primary winding of the VT is connected in parallel to the power line, and its secondary winding, in ideal conditions, is left open (Fig. 64.22).

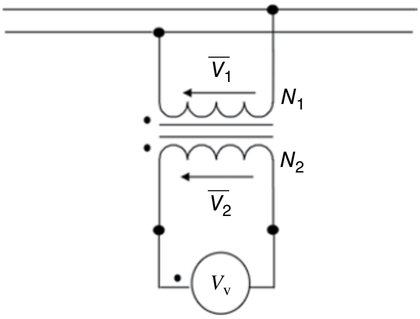


FIGURE 64.22 Ideal VT.

The voltage V_v measured by an ideal voltmeter is

$$\overline{V}_v = \overline{V}_2 = \overline{V}_1 \frac{N_2}{N_1} = \frac{1}{K_T} \overline{V}_1 \tag{64.52}$$

where N_1 and N_2 are the numbers of turns of the primary and secondary windings, respectively. In ideal VTs, the theoretical constant K_T is a real constant, and, therefore, the secondary voltage is a scaled replica of the primary voltage. For safety reasons, VTs require the connection to ground of one of the secondary terminals to avoid the presence of dangerous common mode voltages on the secondary windings. The rated secondary output of VTs is 100V.

64.5.4.1 Measurement Errors The analysis of the VT measurement errors can be done by considering the electrical model of a real VT under no-load conditions (Fig. 64.23).

In a similar way as that followed for the CTs, two kinds of errors can be defined by means of the phasor analysis: the phase error ε and the ratio error η. In fact, by composing the voltage drops characterizing the VT, the graph reported in Figure 64.24 can be easily plotted.

By analyzing the graph, under the condition of small phase error ε, this same error can be defined as

$$\varepsilon \cong \sin\varepsilon = \frac{\left|\overline{Z}_1 \cdot \overline{I}_0\right|}{V_1} \sin\theta \tag{64.53}$$

$$\theta = \varphi_0 - \varphi_1 \tag{64.54}$$

$$\overline{Z}_1 = R_1 + j\omega L_1 \tag{64.55}$$

where φ_1 is the phase angle of $\overline{Z}_1$ and φ_0 is the angle between $\overline{I}_0$ and $K_T\,\overline{V}_2$.

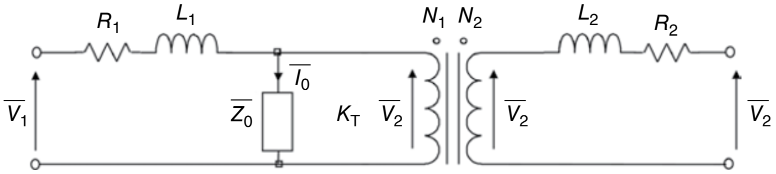


FIGURE 64.23 Electrical model of a VT.

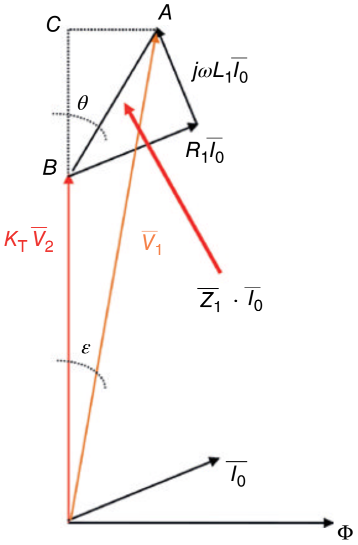


FIGURE 64.24 Phasor graph of a VT.

As for the CT, also the ratio error can be defined as

$$\eta \cong \frac{K_r - K_T}{K_r} - \frac{|\overline{Z}_1 \cdot \overline{I}_0|}{V_1}\cos\theta \tag{64.56}$$

$$K_r = \frac{V_{1r}}{V_{2r}} \tag{64.57}$$

Equations 64.53 and 64.56 show that the VT errors depend on its magnetizing current and the primary series parameters of the transformer. Therefore the improvement of the VT performances requires the optimization of the magnetic core, the numbers of turns of its windings, and the primary winding parameters. Equation 64.56 also shows that the ratio error can be zeroed, for a given voltage, by properly defining the rated ratio K_r of the VT.

In the performed analysis, the ideal no-load condition was considered. However, the measurement of voltage $\overline{V}_2$ implies the connection of a voltmeter, which can be represented by its internal impedance $\overline{Z}_b$ (Fig. 64.25).

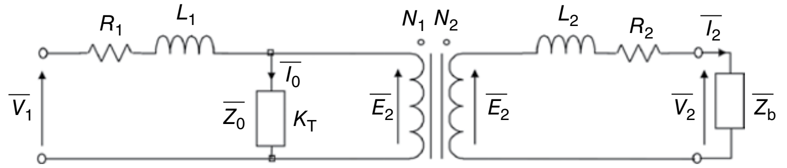


FIGURE 64.25 Electrical model of a loaded VT.

It is clear that the presence of a load modifies the working condition of the VT. In particular there will be now a secondary current $\overline{I}_2$ that causes a voltage drop on the secondary series parameters, thus increasing the error:

$$\overline{V}_2 = \overline{E}_2 - \overline{I}_2 \cdot \left(R_2 + j\omega L_2\right) = \overline{E}_2 \cdot \left(\frac{\overline{Z}_b}{\overline{Z}_b + \left(R_2 + j\omega L_2\right)}\right) \tag{64.58}$$

In addition the secondary current causes an increment in the primary current, thus increasing the voltage drop on the primary series parameters.

Similarly to CTs, also VTs suffer from the problems of the core saturation and limited bandwidth. In general, their bandwidth is upper limited to 1 kHz.

64.5.5 Electronic Transformers

In the last decade, stricter requirements in terms of bandwidth, related to power quality measurements, as well as the need of new functions, have pushed toward the development of new voltage and current transducers. Nowadays, thanks to the use of electronics, it is possible to guarantee the performances in terms of insulation and accuracy avoiding the use of the traditional CTs and VTs. Moreover the introduction of electronics opens the way to the implementation of functions such as measurement synchronization, diagnostics, measurement preprocessing, etc. For these reasons, two standards have been issued: IEC 60044-8 for electronic current transformers (ECTs) [8] and IEC 60044-7 for electronic voltage transformers (EVTs) [9]. These standards provide the requirements and the characteristics that the new transducers must satisfy, regardless of the employed technology.

One of the main differences between CTs/VTs and ECTs/EVTs is related to their output signals. Electronic transformers (ETs) can have two kinds of output signals: analog and digital. Moreover the analog output can have a rated value different from that of the classical CTs and VTs.

Another significant difference is related to the introduction of the concept of delay. The ET output can have a rated time delay different from zero. The ET errors are computed taking into account the rated delay. This new concept is related to the digitalization and transmission of the information. The maximum phase and amplitude errors are specified not only for the mains frequency but also for its harmonics. The errors are defined according to the aimed function: measurement or protection.

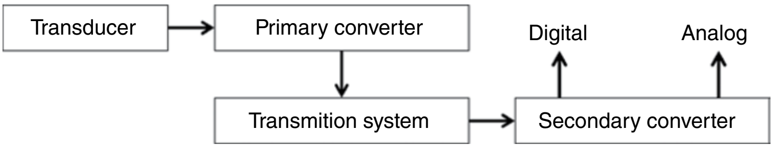


FIGURE 64.26 ET general architecture.

The typical architecture of ETs is shown in Figure 64.26.

By means of this architecture many solutions can be obtained, according to the chosen technology. As an example, by choosing as primary converter an analog-to-digital converter, the transmission system can be based on fiber optics, thus assuring the required insulation in a rather immediate way. In the secondary converter, a digital-to-analog converter can be present, as well as a digital interface. Computation capabilities can be placed in different parts of the ET, in order to perform signal preprocessing as, for instance, filtering, parameter computation, etc.

The architecture of Figure 64.26 can be modified by adding as many transducers as necessary, for instance, for implementing a combined three-phase measurement transformer. In this case the transformer will be able to measure simultaneously three currents and voltages as well as to compute active power, power factors, etc.

The kinds of transducers that can be used for the implementation of an ET are significantly higher than those used with the traditional transducers, thanks to the described architecture. In fact, by using a primary converter based on electronics characterized by stable and easily tunable input impedance, it is possible to use wideband sensors such as resistive shunts and Rogowski coils for the current and capacitive voltage dividers for the voltage. The insulation is ensured by the communication channel.

ETs can easily feature bandwidth up to 100 kHz, rated input voltage up to 500 kV and rated input current up to 10 kA, accuracy of 0.1% of the rated primary value, and an insulation level up to hundreds of kilovolts.

64.6 POWER QUALITY MEASUREMENTS

The previous sections have mostly dealt with power and energy measurements under AC sinusoidal conditions, since nowadays electric systems are generally operated under these conditions. However, the situation is slowly changing, due to the widespread use of power electronic devices in power systems. High-power power electronic converters are expected to increase in number and power, in the new smart-grid scenario, when energy production from renewables will play a significant role, and, consequently, generation will become more and more distributed.

There are two consequences of this new scenario: the energy flow will become bidirectional in almost every branch of the grid, and the voltage and current signals will become distorted, thus causing the sinusoidal conditions to be abandoned.

The first problem can be easily solved with a structure such as the one shown in Figures 64.10 and 64.26. This structure can be readily equipped with proper algorithms capable of accounting energy flowing in both directions. Moreover, being based on a digital processor, more functions, such as connections to a remote center, can be implemented, thus obtaining the so-called smart meters, which are the natural evolution of wattmeters and energy meters in the smart-grid scenario.

The second problem is definitely more critical and gives rise to the so-called power quality problems.

It is well known that power electronic devices are strongly nonlinear and also, sometimes, time variant. Therefore, they inject periodic disturbances on the line current, and these disturbances can be synchronous with the fundamental frequency (harmonic disturbances) or asynchronous with the fundamental frequency (subharmonic or interharmonic disturbances).

Since the short-circuit power in a given network section is never infinite, current disturbances causes voltage drops on the equivalent source impedance, so that also the voltage, in the same given section, shows the same kind of distortion as the current [10]. This causes power components to be associated with the frequency-domain components (harmonic and nonharmonic) of voltage and current.

While the active power, being defined as the average value of the instantaneous power, keeps its physical meaning and can still be measured by instruments based on (64.38), provided they feature enough bandwidth,[5] all other power components lose the physical properties they show under sinusoidal conditions [10].

The first attempts to extend the reactive power definition to the nonsinusoidal conditions go back to the early decades of the twentieth century and are due to Budeanu [11] and Fryze [12]. These two approaches to power components definition under nonsinusoidal conditions have been widely discussed in the second half of the twentieth century and sometimes strongly criticized [13], but no final word has been said on this issue yet [10, 14]. A recent IEEE standard, the IEEE Std. 1459-2010 [15], gives some interesting definitions, but it still clearly states that “There is not yet available a generalized power theory that can provide a simultaneous common base for: Energy billing, Evaluation of electric energy quality, Detection of the major sources of waveform distortion, Theoretical calculations for the design of mitigation equipment such as active filters or dynamic compensators.”

Unfortunately, most definitions available in the literature[6] try to extend the reactive power concept to the nonsinusoidal conditions and call the newly defined quantities always “reactive power.” The consequence is that different values can be measured,

[5] The bandwidth is not only ensured by a correct sampling frequency, as it is generally thought, but also by a correct choice of the voltage and current transducers. As shown in Section 64.5, transducers may have a dramatic impact on the performances of the instruments connected to their output terminals, both in terms of bandwidth and accuracy.

[6] A short survey of these definitions can be found in [14].

under the same distortion conditions, for, apparently, the same quantity, depending on the adopted definition.

Moreover, all definitions have been derived under the assumption that the instantaneous power can be defined by (64.11). It is worth recalling that this is true, provided that the highest-frequency components in the signal have wavelength far lower than the dimensions of the circuit. Considering that the switching frequency of power electronic devices is ever increasing, the presence of high-frequency components with nonnegligible energy cannot be excluded, in the future, on power systems.

It is then possible to conclude that the modern instruments for measuring electric power and energy, such as those based on the architecture shown in Figures 64.10 and 64.26, can virtually measure every newly defined quantity. The problem, as also reported in [16], is that there is no agreement, yet, on what to measure, on which frequency band, and with which accuracy.

Therefore, we limit this contribution to warn the readers that power quality measurements, unlike voltage and current quality measurements, are still largely undefined and require a significant additional research work to be agreed upon by the scientific and technical community. The key features, in this kind of measurements, are the measurement algorithms and bandwidth and accuracy of the voltage and current transducers. They are an important part of the whole measurement equipment and cannot be neglected when assessing the metrological performance of the employed equipment.

REFERENCES

1. R. A. Serway, J. W. Jewett, *Physics for Scientists and Engineers*, 6th Edition, Thomson Brooks/Cole, Salem, OR, 2004.
2. P. A. Tipler, G. Mosca, *Physics for Scientists and Engineers*, 6th Edition, W.H. Freeman and Company, New York, 2008.
3. M. Alonso, E. J. Finn, *Fundamental University Physics: Vol. 2 Fields and Waves*, Addison-Wesley Publishing Co., Reading, MA,1971.
4. J. Bird, *Electrical Circuit Theory and Technology*, 5th Edition, Routledge, Oxon, 2014.
5. JCGM 100:2008, *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement (GUM 1995 with Minor Corrections)*, Joint Committee for Guides in Metrology, 2008. Available online: http://www.bipm.org/en/publications/guides/gum.html (accessed November 7, 2015).
6. M. B. Stout, *Basic Electrical Measurements*, Prentice-Hall, Englewood Cliffs, NJ, 1960.
7. G. D'Antona, A. Ferrero, *Digital Signal Processing for Measurement Systems: Theory and Applications*, Springer, New York, 2006.
8. IEC, "Instrument Transformers—Part 8: Electronic Current Transformers—Edition 1.0", *IEC Std. 60044-8*, 2002.
9. IEC, "Instrument Transformers—Part 7: Electronic Voltage Transformers—Edition 1.0", *IEC Std. 60044-7*, 1999.

10. A. Ferrero, Measuring electric power quality: problems and perspectives, *Measurement*, Vol. 41, no. 2, 2008, pp. 121–129.
11. C. I. Budeanu, *Puissances reactives et fictives*, Inst. Romain de I'Energie, Bucharest, 1927.
12. (a) S. Fryze, Active, reactive and apparent power in circuits with non-sinusoidal voltage and current, *Przegl. Elektrotech.*, Vol. 7, 1931, pp. 193–203 (in Polish); (b) S. Fryze, Active, reactive and apparent power in circuits with non-sinusoidal voltage and current, *Przegl. Elektrotech.*, Vol. 8, 1931, pp. 225–234 (in Polish); (c) S. Fryze, Active, reactive and apparent power in circuits with non-sinusoidal voltage and current, *Przegl. Elektrotech.*, Vol. 22, 1932, pp. 673–676 (in Polish).
13. L. S. Czarnecki, What is wrong with the Budeanu concept of reactive and distortion power and why it should be abandoned, *IEEE Trans. Instrum. Meas.*, Vol. 36, 1987, pp. 834–837.
14. A. Ferrero, Definitions of electrical quantities commonly used in non-sinusoidal conditions, *Eur. Trans. Electr. Power*, Vol. 8, no. 4, 1998, pp. 235–240.
15. IEEE, "IEEE Standard Definitions for the Measurement of Electric Power Quantities Under Sinusoidal, Nonsinusoidal, Balanced, or Unbalanced Conditions," *IEEE Std 1459-2010 (Revision of IEEE Std 1459-2000)*, 2010.
16. A. Ferrero, "Measurements on Electric Power Systems: Are We Prisoners of Tradition?" *IEEE International Workshop on Applied Measurements for Power Systems (AMPS)*, Aachen, Germany, September 25–27, 2013, pp. 120–125.

PART VIII

CHEMISTRY

65

AN OVERVIEW OF CHEMOMETRICS FOR THE ENGINEERING AND MEASUREMENT SCIENCES

BRAD SWARBRICK[1] AND FRANK WESTAD[2]

[1] Quality by Design Consultancy, Sydney, New South Wales, Australia

[2] CAMO Software AS, Oslo, Norway

65.1 INTRODUCTION: THE PAST AND PRESENT OF CHEMOMETRICS

The term chemometrics was coined by Svante Wold in 1971 in a grant application [1], and soon after, the International Chemometrics Society (ICS) was formed. The ICS defines chemometrics as follows: “Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.” It must be noted here in the definition that chemometrics is an area of chemistry, not a branch of abstract mathematics, and it is by this definition that gives chemometrics a practical nature.

Chemometrics is fast becoming a commonplace set of tools used in many industrial and research environments. Its rise over the past two decades can be attributed to two major factors:

1. The speed of computers has improved exponentially and personal computers can store large quantities of data for analysis.
2. Analytical technologies, such as spectrometers and chromatographs, can generate large volumes of data per measurement that requires highly sophisticated methods of analysis.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

Chemometric methods are based on the larger discipline of multivariate analysis (MVA), also known as multivariate data analysis (MVDA). MVA methods that are most applicable to chemometrics can be divided into three main categories:

1. **Exploratory data analysis** (**EDA**): Mathematical methods used to investigate the natural patterns that exist in a specified data set by the application of typically linear methods of analysis, whereby sample and variable relationships may be established and used for further applications or enhanced data insights. Exploratory methods are also known as unsupervised classification methods.
2. **Regression methods**: These methods aim to define a model that relates one set of variables, known as independent variables to a set of responses, or dependent variables. The quality of the model is determined by its predictive ability, and this is determined based on comprehensive statistics generated by chemometric models. The set of independent variables may itself be a compilation of variables from various types of sensors or those generated by a single multichannel instrument, such as a spectrometer.
3. **Classification methods**: Multivariate classification is also known as supervised classification and requires the definition of classification rules. These rules can be submodels developed in the EDA phase of data collection, and the quality of such a classification schema is determined again on predictive ability.

A key theme of the chemometric methods defined earlier is the model's predictive ability. It must be remembered that chemometric models are not hard models applicable in a general way but are empirical (soft) models. An empirical model is, by definition, an approximation made based on a limited set of available data. This gives rise to two of the most important concepts of chemometric modeling: representation and validation. These two topics will be discussed in great detail in this chapter.

Prior to the widespread availability of fast personal computers, chemometrics was a topic reserved for the academic community. What are simple calculations by today's standards that are complete in a matter of minutes or seconds could have previously taken hours or days to complete. Even in the early days of the personal computer, the analysis of relatively small data sets could take hours to generate results. The advent of the Microsoft® Windows platform also contributed to the acceleration of the usage of chemometrics in the past two decades. This is because of the better graphics capabilities of current computer operating systems. Since chemometrics is a highly visual analysis method, high definition and superior graphical capabilities enhance the interpretability of chemometric models.

The current challenge that needs to be addressed by chemometric methods is how to handle the ever-growing data sets being generated by modern instrumentation and manufacturing systems. Algorithms that can better utilize processor cores and grid computing will better handle large data sets; however, many chemometric methods were initially implemented with algorithms based on deflation. By this it meant that in order to complete an entire analysis, the process must be performed in a sequential manner; thus, the calculation of the next step of the process is dependent on the last one completing. There is no increase in speed obtained as the calculation proceeds, so if the data set is large, the calculation time will increase as a function of size. Over the years, algorithms tailored to specific matrix dimensions have been developed [2, 3].

A related data analysis methodology to chemometrics is design of experiments (DoE), which has received much attention in the literature and is seen as its own discipline distinctly different to chemometrics, although the methods share some commonality. As many DoE studies also involve several responses and other variables (covariates) that may be correlated internally and with the design factors, the generic multivariate methods play a role in a more holistic approach. The interested reader is referred to the extensive literature that is available on DoE [4–7].

Terminology is also a major challenge in chemometrics and Section 65.8 provides a detailed review of the most commonly used terms in chemometrics.

65.2 REPRESENTATIVE DATA

It cannot be stressed more that good data results in good models provided there is information in the data to be modeled in the first place! Put another way, the old adage "garbage in–garbage out" holds well when applied to chemometric modeling. By good data, it is meant that they are representative of the situation to be modeled. Multivariate methods aim to describe the variance that exists in a selected data, and this variation comes from two main sources:

1. **Explainable**: Systematic variations within the data that can be potentially modeled using multivariate (or other) techniques.
2. **Unexplainable**: Random variations in the data that cannot be modeled.

The terms explainable and unexplainable are commonly used in the statistical process control (SPC) literature [8] for detecting common cause events that are different from the random (noise) within the measurement system. This also relates back to a highly important method of statistical analysis called analysis of variance (ANOVA). One-way ANOVA is used to determine whether the between treatment averages are significantly different than the between sampling averages. Put into mathematical terms,

$$SS_{\mathrm{T}} = SS_{\mathrm{Between}} + SS_{\mathrm{Within}}$$

where the following equation terms are defined:

SS_{T} = Total variance within the data set that can be explained (i.e., 100% total)

SS_{Between} = The variance explained by the model that can differentiate between two or more treatment levels

SS_{Within} = The variance between the sample replicates within a treatment level, that is, the precision of the measurements

It naturally follows on from this equation that the total variability in a data set is the sum of what can be explained and what cannot be explained by the model. Therefore, the ANOVA expression can be rewritten in nonmathematical terms as

$$\text{DATA} = \text{INFORMATION} + \text{NOISE}$$

How much information can be extracted from the data is a function of:

1. How representative the data set selected is of the problem to be solved.
2. The quality of the measurement system used to measure the samples in the data set.
3. The form of the model used (i.e., if a linear model is used to measure nonlinear data, then the information will be lower due to what is known as lack of fit).

Sample representation can take on a number of forms when developing a chemometric model. First and foremost, great attention is required when selecting a sample set to model that is going to be characteristic of future samples. If a crystal ball was a reality, then this would be a simple process; however, the development specialist does not have such a luxury, so in many cases, models have to be developed in an iterative manner. In particular, when a model is to be developed on natural samples, such as agricultural products, biological specimens, or soils, these samples must be selected to span a range of one or more properties. This may take some time to build a reliable and robust calibration and validation set of data since new samples are typically found by chance during model development. Another challenge of natural samples is their inherent heterogeneity, that is, the sample has different characteristics depending on where it is measured.

One excellent example of heterogeneous sampling is encountered when developing predictive models of agricultural products using near-infrared (NIR) spectroscopy. In particular, if a model for predicting the protein content of wheat is to be developed, an analyst should aim to collect a number of samples of various protein levels, over a number of growing seasons and regions, etc. Once the pool of samples is collected, they must be split in some way as to obtain a representative split and then each sample must be scanned numerous times to average out packing differences and local heterogeneity. Again, it is stressed that representative sampling is the most critical part of chemometric model building.

In industries where samples can be artificially manufactured, the development of calibration samples that span large regions is possible through careful experimental design. Industries such as the pharmaceutical sector have the ability to develop robust samples that exceed the normal tolerances of typical manufacturing targets, and using pilot scale equipment (or using production equipment at the end of a batch run) can develop representative samples that can be used to develop reliable models.

The next step in model development requires the selection of samples suitable for calibration development. In chemometrics there are two main strategies used to test the reliability of a model for future use on new samples:

1. **Test set validation**: This requires a rational sample selection method to separate the sample pool into a calibration (training) set and a validation (test set). More details on this approach are provided in Section 65.6.1.
2. **Cross validation**: Typically, cross validation is used when there are not enough samples available in the pool to create a robust model and test it against a representative set. There are a number of ways cross validation can be used and these are further discussed in Section 65.6.2.

Finally, when the samples have been found to span a suitable region for model development, the measurement system has been found to provide reliable data, and a method of rational sample selection has been decided upon, the next step is to ensure that the samples represent the entire space of the model developed and this is dependent on the modeling strategy to be employed. In the case of exploratory models (see Section 65.3.3), only the independent *X*-variable space needs to be spanned in order to develop robust models. When regression models are to be developed, the calibration and validation set must representatively span both the *X*- and *Y*-variable (response) space. More details of this will be presented in Section 65.6 on validation.

65.2.1 A Suggested Workflow for Developing Chemometric Models

Representation is not only applicable to sample selection but is also applicable to the measurement systems used to measure the samples. If a sensor or spectrometer response is unreliable, no matter how representative the samples are, the data are bad and therefore a reliable model cannot be developed. Table 65.1 provides a quick checklist to follow when selecting a representative data set and developing a robust chemometric model.

65.2.2 Accuracy and Precision

In the development of any analytical procedure, the concepts of accuracy and precision are extremely important for assessing the model's ability to perform its task. The International Conference on Harmonization (ICH) [9] has developed a document that is a useful guide when developing quantitative methods of analysis. Of the main aspects of model validation, the following are deemed to be critical for reliable model development:

1. **Accuracy**: How close the predicted value of the new method is to the reference value (when a validated reference method exists).

TABLE 65.1 Suggested Workflow for Developing Chemometric Models

Task	Suggested Approach
Select samples to build the model such that they span future sample ranges	Start with a small set, when the nature of the samples is unknown, and use a nondestructive (where possible) method to find samples different from the initial set before much effort is put into model development For situations where artificial samples can be produced, create a small set of samples, measure them, and check that the samples made to target specifications have similar characteristics to samples made using the actual manufacturing process
Ensure measurement systems are reliable	When using measurement systems, perform gage R&R or some form of measurement systems analysis (MSA) to check the quality of measurements made [8] Ensure instruments such as spectrometers and chromatograms have been calibrated and qualified before use
Ensure that measurements made are repeatable and reproducible	Measure a single sample multiple times and visually assess results. Perform ANOVA where needed in order to determine if multiple measurement with averaging is required before a model is developed After a suitable sample averaging method has been defined, compare multiple sample averages for consistency of results If the sample is unstable or requires an exact measurement window for reliable results, automate the sample collection process as much as possible and send the samples for reference analysis as soon as possible (if required)
Reference method analysis	The reference method is the "gold standard" when developing quantitative methods of analysis. Theoretically, the secondary method cannot be more accurate than the reference method; however, the secondary method can be more precise (see Section 65.2.2 for more details) It is important to remember that the sample measured by the secondary method is the one analyzed by the reference method. In cases where the sample measured by the secondary method is too large to be measured by the reference method, a suitable sample splitting method with replicate analysis may be required, depending on the heterogeneity of the sample being measured
Validate the model and interpret it	Depending on the context the model is to be used for, it is always suggested to validate a model using an independent test set. There are a number of sample selection strategies available to create representative training and test sets (see Section 65.6.1.2), and this will provide the most reliable estimate of the models future performance

TABLE 65.1 (Continued)

Task	Suggested Approach
	Parsimony is a key term used in model development, and in general terms, the simpler the model, the more interpretable it is. A model that cannot be interpreted on a physical, chemical, or biological level should not be used for practical purposes
Implement the model	A model is only of use if it can be interpreted and, more importantly, implemented for practical usage in a production or research environment

2. **Precision**: How close a set of replicate measures are to each other. Precision is further subdivided into,
   (a) **Repeatability**: How close a set of predicted values is to each other when the same sample is measured multiple times (with replacement) using the secondary method. Repeatability is a measure of the inherent noise in the measurement system.
   (b) **Intermediate precision**: This is a measure of reproducibility. Intermediate precision is a measure of the sampling error between different analysts or different measurement systems (or a combination of both) and estimates the simplicity of the method usage.
3. **Robustness**: A measure of how sensitive the method is when small but deliberate changes are made to the system. This may include turning on/off a lamp in a spectrometer during analysis or even changing the lamp to assess the impact on predicted values.

Other criteria important to model development are linearity, range, limit of detection (LOD), and limit of quantification (LOQ). These must be assessed based on criticality (particularly LOD and LOQ when predictions are to be made close to the dynamic limits of the measurement system). These points will be expanded upon in Section 65.4.4.7.

ICH Q2(R1) [9] also states that once intermediate precision has been established, then accuracy is inferred.

What does this all mean in terms of chemometric model development? Without accurate and precise reference methodology, it is highly unlikely that the model will be reliable. In business critical operations such as in pharmaceutical or biopharmaceutical applications, the inability to establish accuracy and precision invalidates a model completely. Accuracy and precision are easily visualized using the “dartboard” principle. Figure 65.1 shows the principles of accuracy and precision and why they are important for reliable model development.

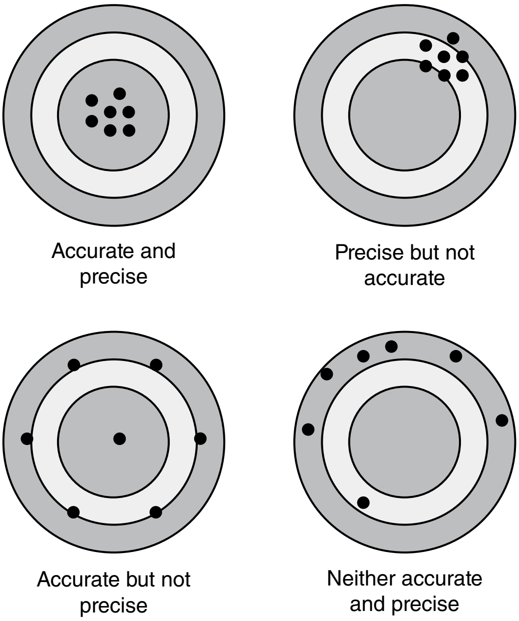


FIGURE 65.1 Diagrammatic representation of accuracy and precision.

65.2.3 Summary of Representative Data Principles

It was the intent of Section 65.2 to stress the importance of sound sampling and preparation before any attempt is made to develop reliable chemometric models. Ninety percent of the effort required to build a model was described in this section and a proven approach to development was presented.

Overall, reliable models are simple models, are easy to interpret, and can be easily root cause analyzed in the event of an outlying result. This is the result of good planning, understanding of the system being investigated, and putting in the effort to build robustness into the model from the start. As mentioned earlier, the principle of garbage in–garbage out must be acknowledged and this is where subject matter expertise is required to minimize the risk of modeling pitfalls.

The following points summarize the prerequisites of reliable model development:

1. Select a secondary method that is fit for purpose, that is, run a small feasibility set to ensure the system is capable of measuring the properties of interest.
2. Understand the complexity of the sample being measured. Heterogeneous samples will require more replicate measurements to be made for both the reference and the secondary methods in order to generate accurate and reliable results.
3. Ensure all equipment used is calibrated and in good working order. Sound obvious? It is rarely established in practice!

4. Use sound validation principles to generate the simplest models and interpret the model on either a physical, chemical, or biological level (this is the topic of Section 65.6).
5. Implement the model and learn from its usage.

65.3 EXPLORATORY DATA ANALYSIS

EDA is typically the first step in any data analysis problem. It allows an analyst to get an initial feed for a data set, particularly the distribution of the data for each of the variables. Section 65.3.1 discusses the univariate approach to data analysis (and some of its pitfalls) and describes why a multivariate approach is to be preferred. There are a number of methods available for EDA; however, the most commonly used methods are discussed in further depth in the following sections. The key point to remember is that EDA is unsupervised in nature, that is, the analyst is looking for natural patterns in the data. Once these patterns are established, rules can be developed and used to classify new samples. This is called supervised analysis or pattern recognition. Supervised methods are discussed further in Section 65.5.

65.3.1 Univariate and Multivariate Analysis

Prior to the development of many chemometric models, an investigative analysis is usually performed to gain some initial insights into the data structure. Scientists and engineers are typically taught in undergraduate programs to investigate raw data using simple charting techniques to look for "obvious" trends. While these simple plotting tools are powerful when analyzing variable 1 (or maybe 2 and 3) at a time, they become highly cumbersome when dealing with multivariate data.

Analyzing data, one variable at a time has been defined as the "scientific approach" to analysis. This only works if the variables are independent of each other, that is, there is little to no correlation between the variables. In Figure 65.2, two situations are provided that only take into account two variables at a time. This is a classical example and has been presented many times in the literature [8, 10] but serves as the best way of showing that for even the simplest of systems, the failure to analyze a data set multivariately may lead to false and sometimes fatal conclusions. It is assumed in this example that the reader is familiar with control charts used for SPC.

In case 1, two control charts, one for the temperature and one for the feed rate of a particular system, are provided. The control charts show that the variables appear to be in a state of statistical control. When the two sets of data are plotted point for point as a scatter plot, it can be seen that there is no linear (or other) relationship between the points, that is, the correlation is close to zero and therefore the two variables can be considered to be independent. In this case, the two sets of control

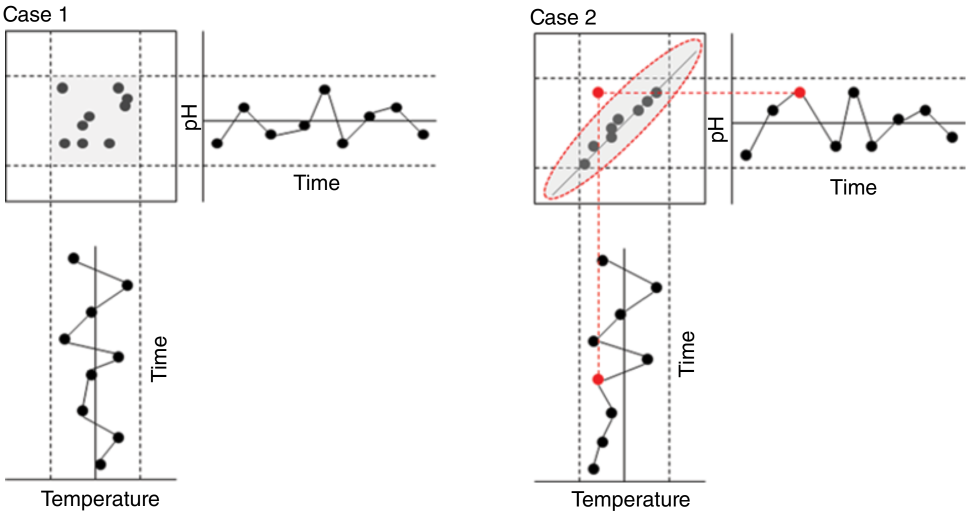


FIGURE 65.2 Simple data representations showing the multivariate nature of data.

limits bound the variability of both variables and anywhere inside the limit box are deemed to be in control.

Looking at case 2, in this case, two control charts, one of temperature and one of pH, are presented. As was the situation for case 1, the variables appear to be in a state of statistical control; however, there is an outlier in the data that cannot be seen univariately. When the data are plotted together as a scatter plot, note there is a linear relationship between the variables. Thinking about this from a chemical point of view, this makes sense as there is a scientific justification for why pH changes with temperature (i.e., use subject matter expertise to interpret the system). Note now in case 2, the limit box is no longer representative of the data and, more importantly, a visual outlier can be detected.

This occurs because the variables cannot be analyzed in isolation of each other. This is a simple case that shows a failure of the scientific approach when the variables are correlated. The control limits now change from being box shaped to be ellipse shaped. This is a key principle in MVA, that is, joint confidence intervals, and for this reason multivariate statistical process control (MSPC) (Section 65.7) is being widely adopted by industry for early event detection.

65.3.2 Cluster Analysis

The main goal of cluster analysis is to detect inherent patterns within a data set in an unsupervised manner. For the purposes of clarification:

1. Unsupervised methods look for natural patterns within the data in order to define possible classes and groups can be assigned based on known characteristics of the data set.

2. Supervised methods use so-called classification rules to separate new samples into predefined groups. Supervised methods are described in more detail in Section 65.5.

There are many types of cluster analysis methods available to the development scientist or engineer. These methods aim to find the similarity (dissimilarity) between samples in a data set, based on the variables used to measure the samples. Two of the simplest methods of cluster analysis available are known as *K*-means and hierarchical cluster analysis (HCA).

65.3.2.1 K-Means *K*-means is one of the most conceptually easy-to-understand approaches to unsupervised data analysis. It aims to separate samples into *K*-predefined classes with grouping based on the points being closest to a class centroid. As new points are added to a class, the centroid is recalculated and the points in the sample set are reassessed for class membership. Therefore, if the nearest neighbor distance is statistically exceeded, the sample is assigned to a new class and so on until all samples have been classified into the number (*K*) of predefined classes [11]. The algorithm works by minimizing the within-cluster sum of squares [12], thus allowing the definition of statistical limits for sample acceptance/rejection from a defined class. In all cases, *K*-means aims to partition all samples into one class only.

Adams [12] describes a four-step process for the *K*-means algorithm as follows:

Step 1: Define *K* clusters (*K* usually being a small integer) to group the data into and define any initial samples per cluster (should such class knowledge exist). Calculate the cluster means and the initial partition error.

Step 2: For the first sample, calculate the increase/decrease of the partition error by moving the sample into the classes defined. If the error is reduced by moving the sample to a particular class, keep it in that class; otherwise, leave it in its original class. Recalculate the class means each time a sample is moved to a new class.

Step 3: Repeat step 2 for all samples in the data set.

Step 4: If no samples have been moved, stop the process; otherwise, go to step 2.

Disadvantages of K-Means Clustering There are a number of disadvantages encountered when using *K*-means (or the related *K*-medians) methods. Firstly, the method requires an analyst to define the number of clusters to partition the samples before analysis begins. This means that if the first analysis is unsuccessful, many iterations of cluster definition may be required. Secondly, the final grouping of samples reflects the initial choice of clusters or initial samples chosen to define the first cluster centroids. Other disadvantages revolve around the distance measures used to determine the similarity/dissimilarity of samples for class assignment. The most commonly used distance metric is the Euclidean distance, but other methods such as the city block or

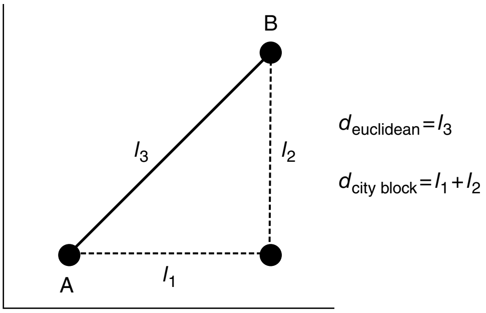

B
l_3
l_2
A
l_1
$d_{\text{euclidean}} = l_3$
$d_{\text{city block}} = l_1 + l_2$

FIGURE 65.3 Euclidean and city block distance metrics.

correlation are available. Figure 65.3 provides a graphical example of the difference between the Euclidean and city block distance metrics.

It is not the intention of this chapter to discuss the algorithm details of the *K*-means algorithm, and the interested reader is referred to the text by Everitt [13] for more details.

An Example of K-Means Clustering One of the classical experiments performed for assessing the ability of clustering methods is Fisher's iris classification data set, first published by Sir Ronald Fisher in 1936 [14]. This is a simple multivariate classification problem and shows in simple terms how the *K*-means algorithm works but, at the same time, shows the limitation of the method, when the number of variables exceeds three.

For the purposes of simplicity, the data set is not repeated here and only a description is provided. The main aim of the experiment was to develop an objective method of classifying three types of iris, namely, *Iris setosa*, *Iris versicolor*, and *Iris virginica*, based on four easy-to-measure variables:

1. Sepal length
2. Sepal width
3. Petal length
4. Petal width

Data was collected on 150 samples (50 of each type) to generate a data table of dimension 150 rows by 4 columns. For each sample, the class name was assigned; therefore, the number of classes to define is $K = 3$. Using the Euclidean distance measure, the classification rate is presented in Table 65.2.

Using the *K*-means method of classification with Euclidean distance, it can be seen from Table 65.2 that:

1. Setosa can be uniquely classified from versicolor and virginica.
2. Versicolor can be uniquely classified from setosa but can be confused 4 times in 100 with virginica.

TABLE 65.2 Confusion Matrix for Iris Classification Using *K*-Means (Euclidean Distance)

Predicted/Actual	Versicolor	Virginica	Setosa	Classification Rate (%)
Versicolor	48	2	0	96
Virginica	14	36	0	72
Setosa	0	0	50	100

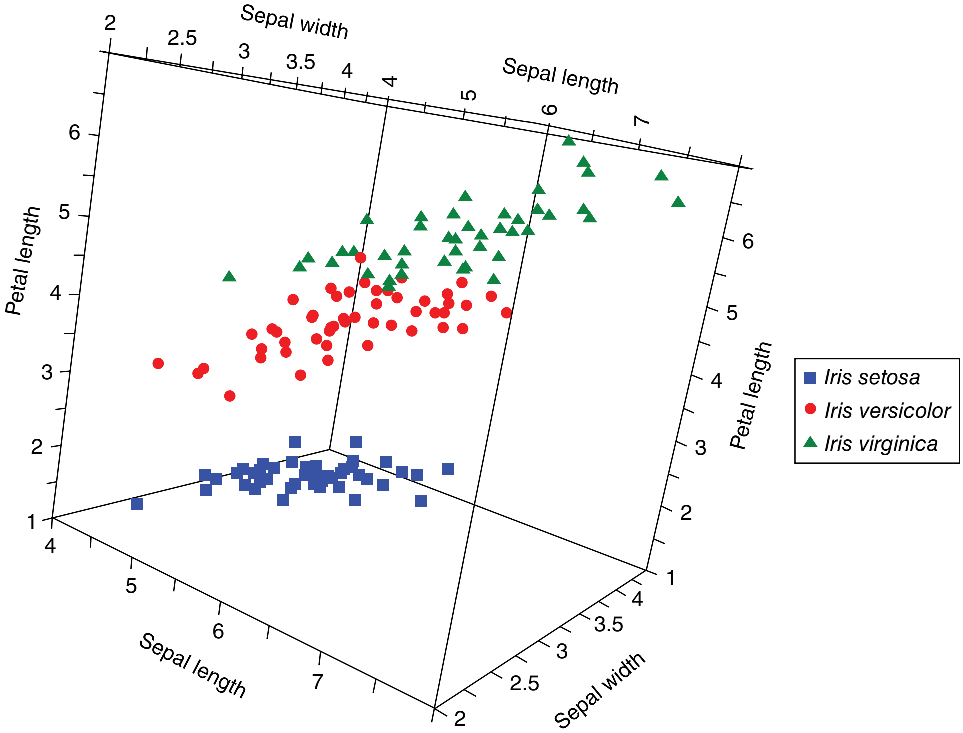


FIGURE 65.4 3D scatter plot of Fisher's iris data grouped by *K*-means clustering.

3. Virginica like versicolor can be uniquely classified from setosa but can be confused 28 times in 100 with versicolor.

The results presented in Table 65.2 can be improved by using correlation as the distance measure; however, these results are not presented here. To visualize the results of a *K*-means analysis, it is possible to plot the complete analysis as long as there are only three variables measured. In the Iris example, there are four variables; therefore, only three variables at a time can be shown. It is possible to plot multiple scatter plots; however, the process becomes extremely cumbersome when the number of variables becomes larger (typically >10). Figure 65.4 provides a three-dimensional (3D) scatter

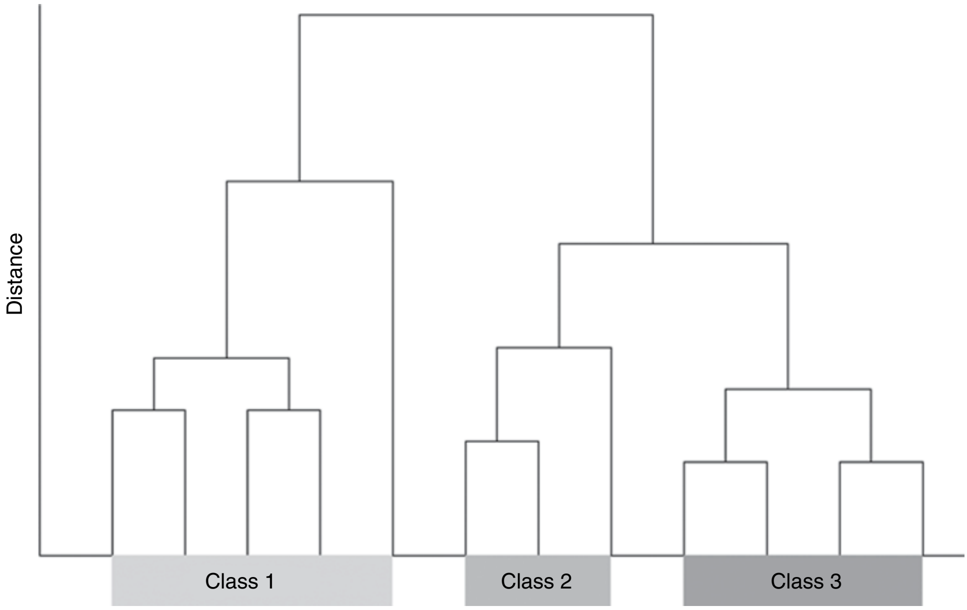


FIGURE 65.5 Example dendrogram.

plot of the variables sepal length, sepal width, and petal length with the samples assigned to their clusters.

It can now been seen in Figure 65.4 that the species versicolor and virginica are close to each other in properties and why the algorithm cannot completely separate the two classes (refer to the confusion matrix in Table 65.2).

65.3.2.2 Hierarchical Cluster Analysis Hierarchical methods aim to separate the original data into a few classes by either agglomerative or divisive methods [12]. Agglomerative methods fuse together smaller subclusters of samples and successively build to larger groups of samples, whereas divisive methods start with a single cluster and divide it into smaller clusters of similar samples.

As with *K*-means, a disadvantage of HCA is that the distance method has to be decided, and there are many to choose from as there are available HCA methods, including the well-known Ward's method [13]. The major advantage of HCA over *K*-means is that it provides a graphical display of the clusters known as a dendrogram. A dendrogram is a tree structure showing the linkages and similarity/dissimilarity of the samples. Figure 65.5 provides an example of a dendrogram.

It is not the intent of this chapter to provide extensive details on the principles of *K*-means and HCA, and the interested reader is referred to the excellent texts available by Adams [12] and Everitt [13] for more detailed discussions of these methods.

65.3.3 Principal Component Analysis

Principal component analysis (PCA) is a bilinear modeling method [14] that provides an interpretable overview of the main information contained in a multidimensional table. It is also known as a projection method, because it takes information carried by the original variables and projects them onto a smaller number of latent (or hidden) variables called principal components (PCs). Each PC explains a certain amount of the total information contained in the original data, and the first PC contains the greatest source of information in the data set. Each subsequent PC contains, in order, less information than the previous one. For PCA, the objective function is to maximize the variance for each subsequent PC. PCA is one of the most powerful EDA methods known, particularly because PCA models can be:

- Easily validated and interpreted
- Investigated using a wide range of graphical and diagnostic tools

The general PCA equation is as follows:

$$X = TP' + E$$

where

X is the original data to be analyzed
T is a matrix of sample structure information known as scores (Section 65.3.3.2)
P is a matrix of variable structure information known as loadings (Section 65.3.3.3)
E is a matrix of residuals that cannot be explained by the PCA model

Therefore, returning to the definition provided in Section 65.2,

X is the data component to be analyzed after centering and application-dependent scaling of the variables.
TP′ is the information part of the analysis (i.e., what can be explained by the model).
E is the noise part of the analysis (i.e., what cannot be explained by the model).

To redefine the definition in Section 65.2, the PCA model becomes

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

This is the reason why PCA is known as a dimensionality reduction (or a decomposition method). It is an unsupervised classification method that aims to take large data sets and break them down into smaller yet more informative components (PCs) that are a collection of the most important sources of variability.

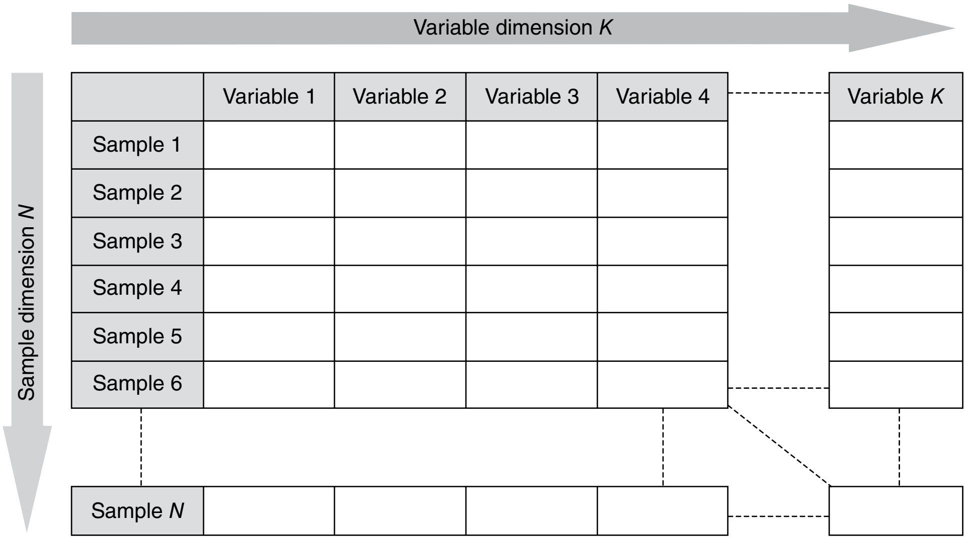


FIGURE 65.6 General table structure for PCA.

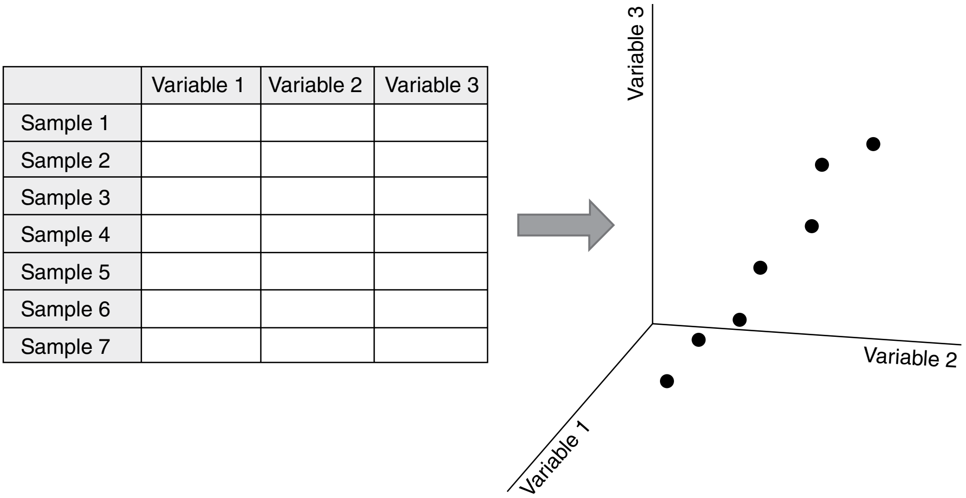
Mathematically, each PC is orthogonal to all other PCs and they can be conveniently plotted against each other on a Cartesian coordinate system. This provides PCA with one of the most comprehensive and powerful range of plotting capabilities available to any MVA method.

Before moving on to how PCA works in practice, the next sections define the terminology used when developing a model and how these terms relate to each other.

65.3.3.1 The PCA Problem and the Dual Nature of Data Samples and variables are not mutually exclusive, that is, a sample is characterized by the variables used to describe it and variables cannot be measured unless a sample is present. A data table suitable for analysis by PCA must be presented as a structured matrix. The general format of the matrix is the rows (N) represent individual samples and the columns (K) represent the variables measured on each table. Figure 65.6 shows the general table structure.

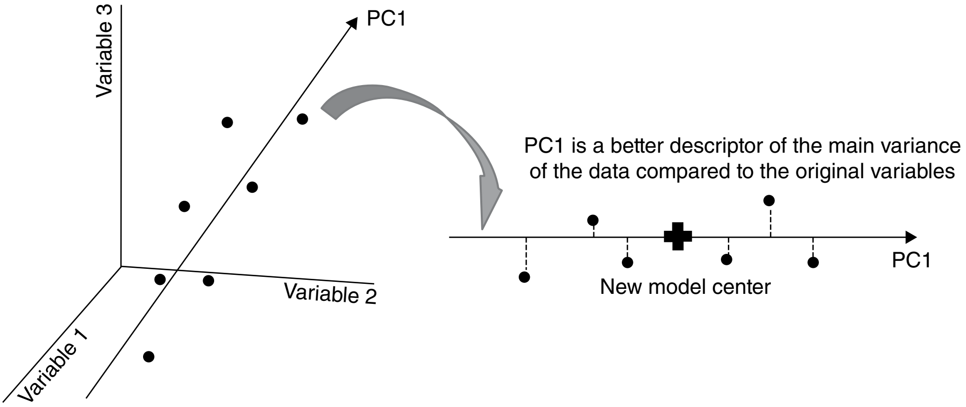
The matrix in Figure 65.6 consisting of N samples and K variables has dimension $(N \times K)$. If $K > 3$, then the entire data set cannot be visualized using standard plotting tools since as humans, visualization in greater than 3-dimensions becomes difficult. What is meant by the dual nature of data is that samples can be plotted in variable space so that sample groupings and other sample relationships can be investigated. Conversely, variables can be plotted in sample space to understand the relationship between variables. The biggest challenge of the aforementioned approach is that it is equivalent to a univariate analysis of multivariate data when N and $K \gg 3$.

When samples and variables are plotted in the ways described previously, an analyst can get a small insight into the correlation structure of the data. PCA extends on this



	Variable 1	Variable 2	Variable 3
Sample 1			
Sample 2			
Sample 3			
Sample 4			
Sample 5			
Sample 6			
Sample 7			

FIGURE 65.7 Plotting samples in variable space to understand the latent structure.



Variable 3
PC1
Variable 2
Variable 1
PC1 is a better descriptor of the main variance of the data compared to the original variables
PC1
New model center

FIGURE 65.8 Fitting the first principal component to a data set.

principle by finding the main sample and variable relationships in multivariate data and reducing them down into simpler and more interpretable PCs. Consider a matrix with $N > 3$ and $K = 3$. This means that the entire table can be plotted in a 3D scatter plot. Consider the hypothetical data set shown in Figure 65.7.

PCs are sometimes referred to as latent variables. By definition, latent means hidden; therefore, PCA aims to find the hidden structure within a data set that may not be obvious from a simple univariate analysis. Note the points plotted in Figure 65.7 are not random but have a distinct structure in 3D space. By fitting the least squares line through the greatest direction of variability shows that there is a better direction in space that describes the data than the original three variables. This is shown in Figure 65.8.

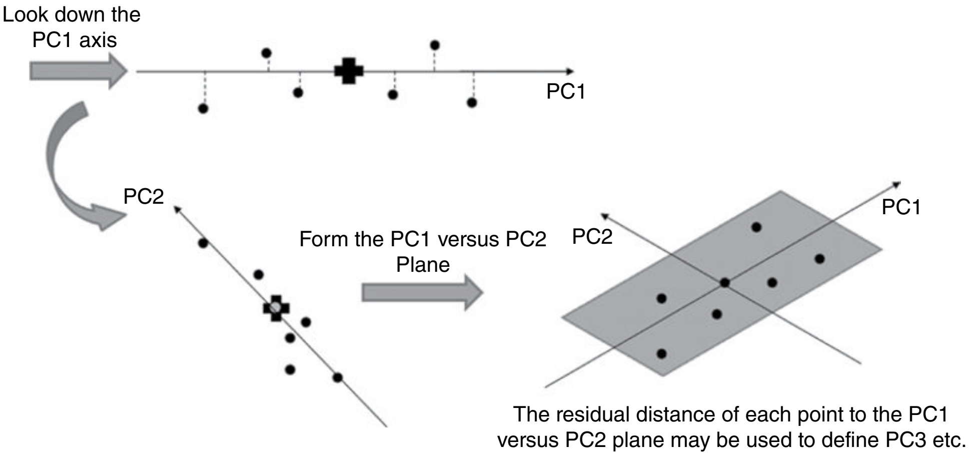

Look down the PC1 axis
PC1
PC2
Form the PC1 versus PC2 Plane
PC2
PC1
The residual distance of each point to the PC1 versus PC2 plane may be used to define PC3 etc.

FIGURE 65.9 Fitting the second principal component to a data set and forming the PC1 versus PC2 plane.

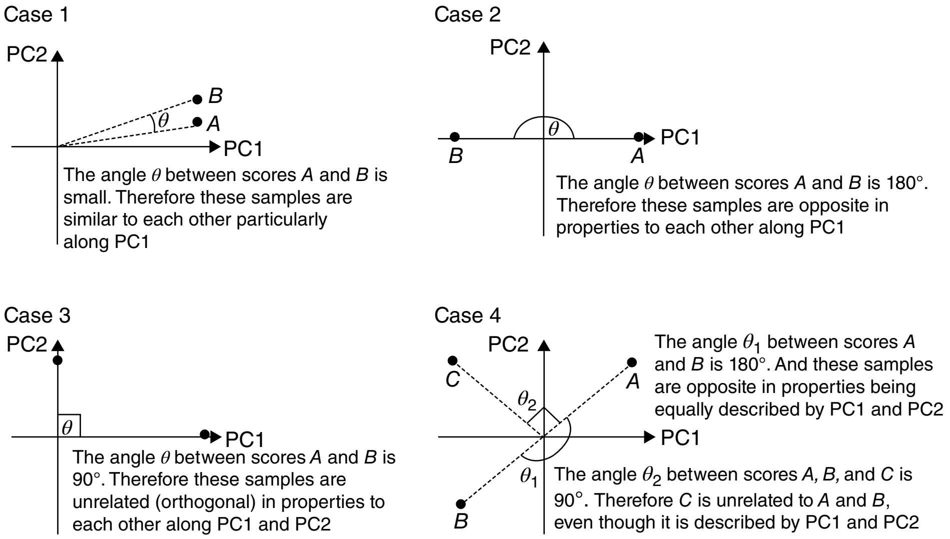
PCA aims to describe the greatest sources of systematic variation in a data set. In the 3D data in Figure 65.8, this describes the direction of greatest elongation in the data. This direction is known as the first PC. When displayed in typical software packages, the first PC is plotted as the new *x*-axis of the data. This is also shown in Figure 65.8.

To find the next greatest source of variability with the constraint that the next PC is orthogonal to the first, the next PC is estimated from the remaining variance. To visualize this, the data must be conceptually viewed down the axis of the first PC. This is shown in Figure 65.9.

The data are now distributed around the first PC axis and a least squares line is fitted through the next greatest source of variability. The combination of PC1 and PC2 now forms a new plane in 2-dimensions. This is also shown in Figure 65.9.

Since the conceptual data set being analyzed has an original dimension of three, in order for PCA to be effective, one or two PCs maximum should describe the data with minimal error, that is, in a 2-PC model, all points should fit closely to the surface of the PC1 versus PC2 plane. If the data still do not fit the model, then the information in all three variables is contributing to the data set. This is called a full rank problem. If, however, two PCs describe the majority of the data and only random fluctuations occur around the PC1 versus PC2 plane, then the rank of the data set is 2 (compared to the dimension of the data $K=3$).

Plotting the sample information in PC space leads to the so-called scores plot of the data. When the variable information is plotted in PC space, this leads to the so-called loadings plot. Scores are discussed in more detail in Section 65.3.3.2 and loadings are discussed further in Section 65.3.3.3.



The angle θ between scores *A* and *B* is small. Therefore these samples are similar to each other particularly along PC1

The angle θ between scores *A* and *B* is 180°. Therefore these samples are opposite in properties to each other along PC1

The angle θ between scores *A* and *B* is 90°. Therefore these samples are unrelated (orthogonal) in properties to each other along PC1 and PC2

The angle θ_1 between scores *A* and *B* is 180°. And these samples are opposite in properties being equally described by PC1 and PC2

The angle θ_2 between scores *A, B,* and *C* is 90°. Therefore *C* is unrelated to *A* and *B*, even though it is described by PC1 and PC2

FIGURE 65.10 Sample relationships along a selected PC axis.

65.3.3.2 PCA Scores The PCA method was first introduced in the early 1900s by Pearson [16] and was later adapted for the analysis of psychology data and the terminology has remained to this date. The term scores relates to the samples' importance in describing the variability of a data set. Consider again the conceptual data set introduced in Section 65.3.3.1. If the first PC is plotted in the original variable space, the samples distribute themselves along this direction as shown in Figure 65.8.

A score (t) is the orthogonal projection of each sample onto the selected PC axis. The length of the distribution of the samples along the axis is proportional to how much information is contained in the PC. When the first PC is fitted to the data, a new model center (or origin) is established. This is the point of minimum variability in the data set, that is, samples at or close to the origin are not well described by the model. PC space covers both positive and negative regions along the PC axis. Along any given PC axis, the following holds. These properties are presented graphically in Figure 65.10 with a short explanation of each case:

1. Samples that group close to each other are similar (i.e., share the same pattern for the variables, case 1 in Figure 65.10).
2. Samples that lie close to the origin show the least variability compared to samples at the extremes of the PC axis (i.e., they are average in characteristics).
3. Samples that lie at one extreme of the PC axis are systematically different for the variables that contribute to this PC compared to samples lying at the other extreme (case 2 in Fig. 65.10).

When the next (and subsequent) PC is added to a model, higher-dimensional scores plots are possible. In the case of the two-dimensional (2D) scores plot, the following relationships between samples hold:

1. Along any selected PC, the rules as aforementioned hold.
2. Any samples lying exactly along one PC (but not at the origin) lie perpendicular to any samples that are exactly on the second PC. Based on simple geometry, the directional cosine between such samples is zero; therefore, it can be concluded that the sample types are independent in character from each other (case 3 in Fig. 65.10).
3. Samples that occupy spaces in between the PC axes are influenced by variables that are important on both axes. Although the PC axes describe independent sources of variability, samples may be influenced by more than one PC. There is one score for each sample as for the individual variables and the scores can be described as "super variables." To understand which variables contribute to individual PCs, interpretation of the loadings is required (case 4 in Fig. 65.10).

Overall, a scores plot is a map of sample relationships plotted in either 1-, 2-, or 3-dimensions. These plots are an excellent way to study sample relationships; however, this is where the dual nature of data becomes important. The samples can only group the way they do, based on the variables measured to characterize the samples. No interpretation of sample grouping is possible without an understanding of the variables and their relationships contributing to describing the samples. This is where the loadings plot described in Section 65.3.3.3 is required.

65.3.3.3 PCA Loadings Loadings (dimension P (where $P \leq K$)) for each PC, relate to the weighting placed on each variable for describing a particular PC. PCA loadings can be described as the individual contributions of the input variables for describing a sample set, and the following equation can be used to describe a particular PC:

$$\mathrm{PC}_A = z_1 x_1 + z_2 x_2 + \cdots + z_p x_p$$

where

PC_A is Ath PC being investigated

z_p is the loading or variable contribution to variable x_p

x_p are the original variables used in the analysis

PCA is an empirical modeling method, which means that the quality of the model is limited by the quality and scope of the data used to construct it. As is the case of any statistical method, PCA is used to describe a smaller sample of a larger population. Consequently, the variable contributions may be completely different for a smaller

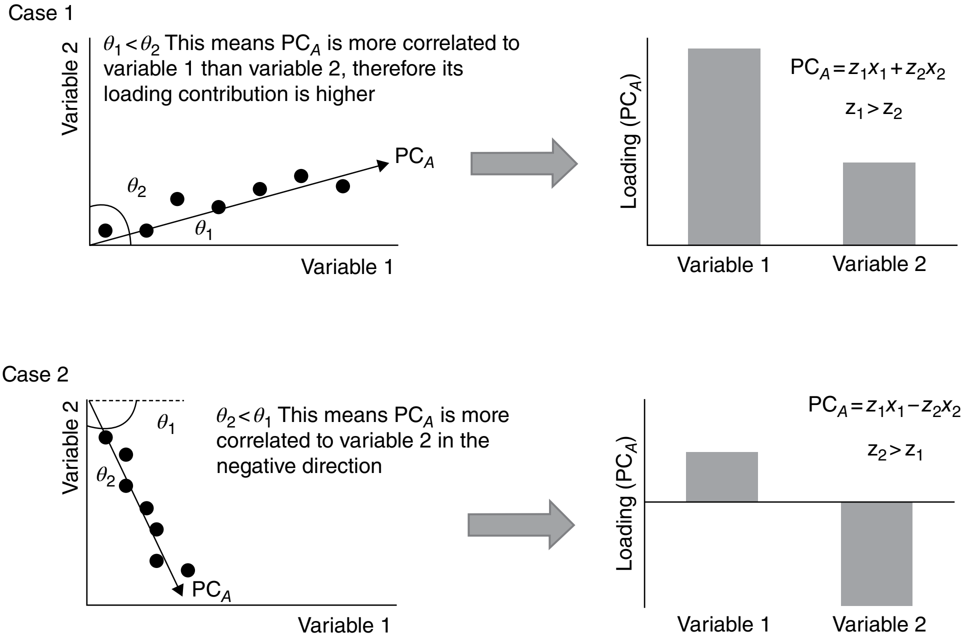


FIGURE 65.11 Definition of PC loadings in variable space.

data set compared to a larger data set (or the entire population). What does this mean then for the loadings? To interpret the variable contributions to a PC, the magnitude of each loading must be taken into account. Loadings are scaled between [−1,1] and the squared sum of the loadings along any PC sum to 1 in most implementations of PCA. This allows a direct comparison of the contribution of each variable to each other. Figure 65.11 shows the conceptual data plotted in the original variable space to describe how PC loadings are calculated.

Using a geometrical interpretation, a PC loading is defined as the correlation (up to a scaling factor) of the PC axis with each of the individual variables used to characterize the data. If, for example, the PC direction is perfectly parallel to variable x_1, then it is also perfectly correlated to this variable and its directional cosine is either −1 or 1. This means (for a three-variable situation) that the PC equation can be represented as

$$\mathrm{PC}_1 = 1 \times x_1 + 0 \times x_2 + 0 \times x_3$$

Or in other words, the entire system is described by the variable x_1 only and that the other variables do not contribute to this PC at all. If, for example, the weightings in subsequent PCs were all zero, the dimensionality of the problem has been reduced from 3 to 1. This is a special case and there are other ways of reducing the problem to a one-dimensional problem as shown in the following:

$$\mathrm{PC}_1 = 0.333 \times x_1 + 0.333 \times x_2 + 0.333 \times x_3$$

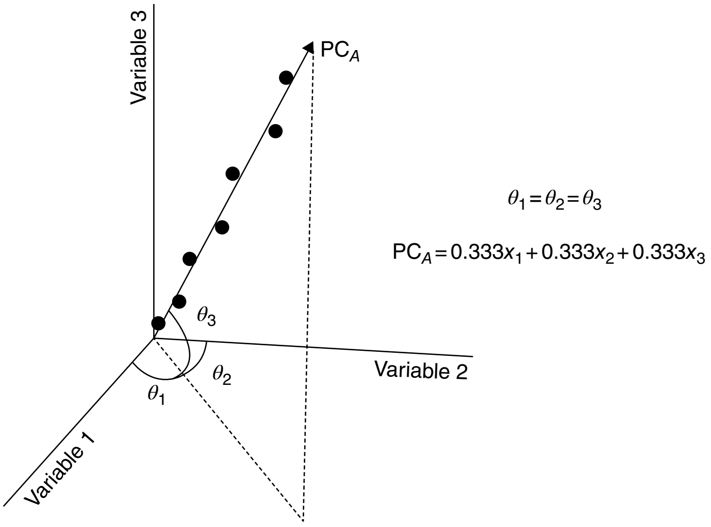

Variable 3
PC$_A$
$\theta_1 = \theta_2 = \theta_3$
PC$_A$ = 0.333x_1 + 0.333x_2 + 0.333x_3
θ_3
θ_2
θ_1
Variable 2
Variable 1

FIGURE 65.12 Equally contributing variables in the $K = 3$ variable situation.

In this case, all variables contribute equally to describing the PC. This situation is shown graphically in Figure 65.12 where the PC intersects all three variable axes equally.

Loadings can be plotted in a number of ways and some of these are defined as follows:

1. Loadings line plots: Commonly used with spectroscopic or chromatographic data to show the importance of spectral bands or peak elutions for interpreting such data.
2. Loadings scatter plots: Commonly used in the interpretation of process variables or sensory data to understand discrete variable relationships.

Figure 65.13 provides some examples of the loadings plots used in describing PCA models.

In section, "Singular Value Decomposition" (SVD) algorithm is defined where it will be seen that although the scores vectors are also orthogonal, they are weighted by their importance in the model, that is, the variance explained by each component. This means that a direct comparison of scores and loadings on a single plot, known as the biplot, may not be entirely reliable for interpretation and it is suggested here that the interested analyst use caution when comparing scores and loadings in this manner. However, like scores, loadings with values close to zero (or close to the origin) contribute little to describing the samples in the PCs under investigation.

Since the loadings are scaled, the most important variables will have the highest absolute values and these are the ones that should be interpreted. Section 65.3.3.4 provides a brief introduction to variable scaling and its importance when developing a PCA model, particularly when taking into consideration the magnitude of the loadings.

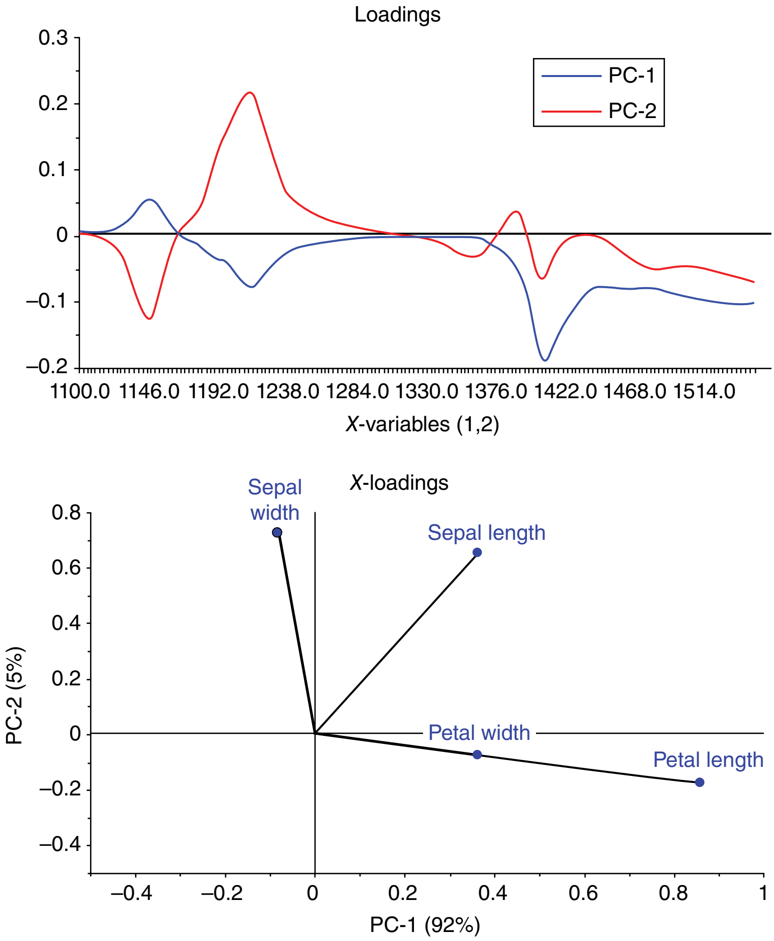


FIGURE 65.13 Examples of loadings plots used for different kinds of analysis situations.

In some cases, when the scaling is performed correctly, even though some variables do not contribute to the model, they may still be highly correlated to a particular PC. To view such correlations, the loadings are calculated to be scale invariant and this allows the generation of the correlation loadings plot. Figure 65.14 provides an example of a loadings plot and its corresponding correlation loadings plot showing how unimportant variables can be highly correlated to a PC.

The essential features of the correlation loadings plot are as follows:

1. The outer ellipse shows the points where variables are perfectly correlated to one or more PCs.
2. The inner ellipse shows the points where variables are, in sum, 50% explained by one or more PCs that are plotted. Since variance is equal to correlation, the explained variance can be calculated directly from the correlation loadings.

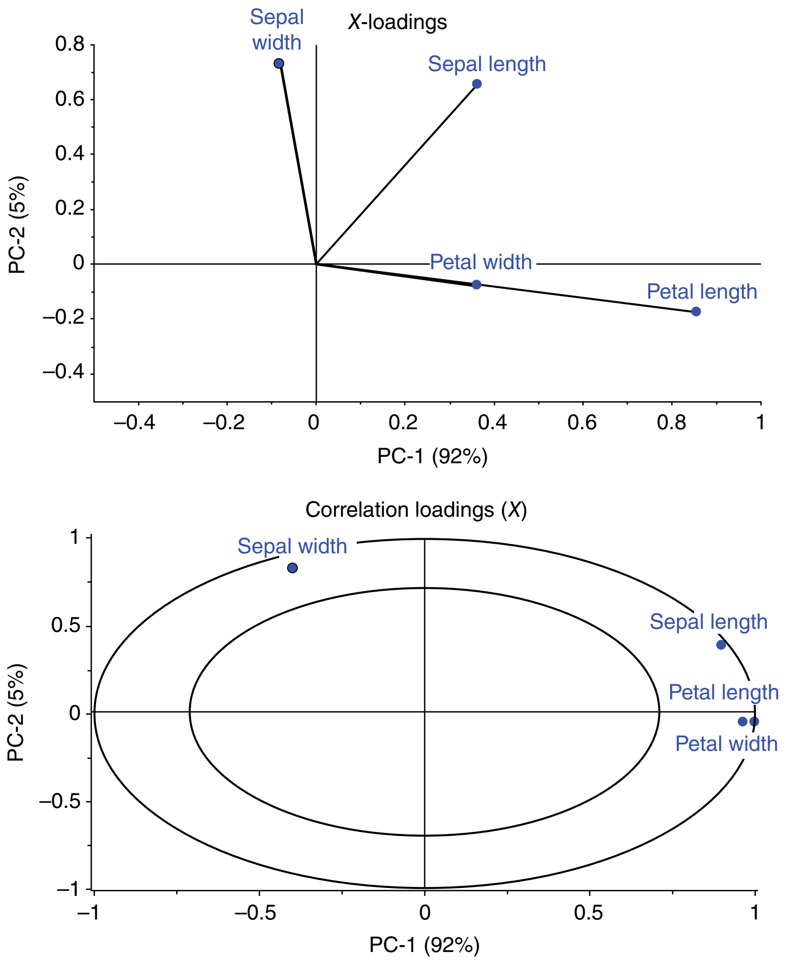


FIGURE 65.14 Example loadings plot and its corresponding correlation loadings plot.

Generally speaking, if a loading value is high and its correlation loading is also high, then the variable contributes highly to describing the PC. If the loading value is low for a variable and its correlation loading is high, then the variable is not contributing much to describing the PC, but its correlation structure should be noted for interpretation purposes. Finally, if a variable correlation loading lies within the 50% ellipse and the origin, for example, with a correlation loadings of 0.5 (25%), then most likely, this variable does not contribute at all to describing the PC. This rule of thumb, as a cutoff value in percentage, is highly dependent on the type of data. Spectroscopic data may hold important information in PCs that explain below 1% of the total variance, whereas for process data this will, in general, not be the case.

Overall, like scores, PC loadings provide a map of the variable relationships based on the characteristics of the samples measured. Loadings cannot be interpreted without scores and vice versa. The process of interpreting scores and loadings will be provided by example later in this chapter.

65.3.3.4 A Short Introduction to Variable Scaling and Preprocessing The two common types of data analyzed using PCA are data generated by spectrometers/chromatograms or discrete process variables; however, many other data types exist. When dealing with these data types, variable scaling or preprocessing is an important first step prior to data analysis. Typically there are two separate approaches used to scale spectra and process data. This section is not meant to be an exhaustive description of data preprocessing, and there are many excellent subject-specific references in the literature [11, 12, 15–18]; it is meant only to be used as a guide for preprocessing data and interpreting it.

Preprocessing of Spectral Data The data typically generated by spectrometers (or chromatographs) is usually scaled to a set axis, that is, in absorption spectroscopy, the scale for each variable measured is between 0 and 5. It is usually not the intent in the analysis of such data to weight individual variable for two main reasons (although exceptions can and do occur):

1. A typical spectrometer/chromatograph can generate many points and individual variable scaling can be highly tedious and unnecessary.
2. Since the variables are all measured on a common scale, it is usually the intent to use PCA to find regions of the spectra that contribute to the variability in the data that are distinct from other regions.

In the analysis of spectral data, an analyst is trying to extract some chemical or biological information from the data. If the data contain unwanted physical effects, and these are the dominant source of variation in the data, then the PCA model will focus on these effects in the first one or first few PCs, rendering the chemical/biological information less important. Preprocessing is used to minimize such physical effects before analysis such that the desired information can be found in the first few PCs.

Depending on the type of spectroscopic or chromatographic method used, the most common preprocessing methods are presented in Table 65.3 with a brief listing of common methods and applications.

It is noted here that each preprocessing method has its own purpose for a particular effect. In some cases, correction of multiple effects can be performed by using combinations of preprocessing techniques; however, the overuse of preprocessing can distort the data leading to potentially false data interpretations after chemometric analysis.

Scaling of Discrete Data Sources Unlike spectroscopic or chromatographic data, process data usually comes from multiple sources with each source having its own variable scale range. As discussed in Section 65.3.3, the purpose of PCA is to find the

TABLE 65.3 Some Common Preprocessing Techniques Used for Spectroscopic and Chromatographic Data

Physical Effect	Preprocessing Methods	Common Applications
Baseline offset	Derivatives, baseline correction, detrending	Near- and mid-infrared, Raman, UV–visible
Scatter effects	Standard normal variate (SNV), multiplicative scatter correction (MSC)	Near-infrared, diffuse reflectance mid-infrared
Scale shift	Correlation optimization warping (COW)	Chromatography, nuclear magnetic resonance (NMR) spectroscopy

greatest sources of variability in a data set. Consider, for instance, a chemical reactor being measured by three discrete variable sources:

1. pH on a scale from 1 to 14
2. Temperature on a scale from 100 to 200°C
3. Pressure on a scale from 1000 to 2000 psi

This data was deliberately chosen with an order of magnitude step between the variables. Based on a straight univariate statistical analysis, the variability in the highest magnitude variable (i.e., pressure) may far exceed even the entire scale of the lowest magnitude variable (i.e., pH). It may be the case that small changes in pH provide the highest contributions to a quality parameter of the reaction, for example, final yield; however, a straight PCA on unscaled data would indicate that pressure has the greatest influence on data variability. To overcome this issue, the use of variable scaling is required.

Unlike univariate analysis, MVA techniques are typically not concerned with the distributional problems of the individual variables but more concerned with the overall variability and its relation to all variables being analyzed. The most common approach to variable scaling is autoscaling, where the variables are first mean centered and then they are divided by their standard deviations.

Figure 65.15 shows the difference between the original data, the mean centered data, and the autoscaled data.

Importantly to note, when developing a process control model using PCA, the values used to autoscale new data should be typical of the variability commonly experienced in normal process operations. This will ensure that abnormal situations are detected due to better understanding of the system.

The autoscaled data in Figure 65.15 now show that each variable can contribute to a PCA model on an equal scale in order to determine variable importance when assessing the loadings. Table 65.4 provides some examples of common scaling options used for the analysis of discrete source multivariate data.

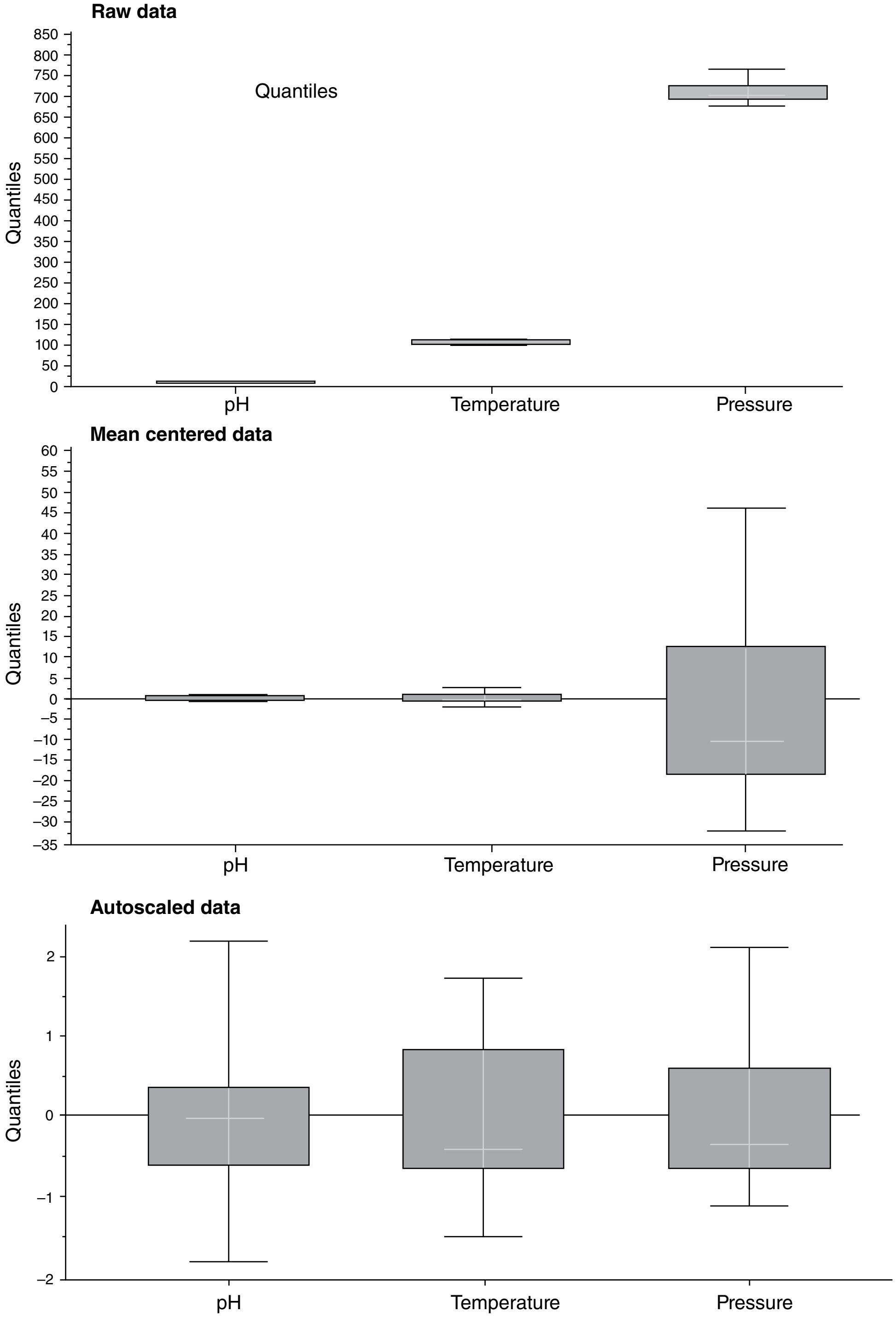


FIGURE 65.15 Examples discrete variable scaling.

TABLE 65.4 Some Common Preprocessing Techniques Used for Discrete Source Process Data

Preprocessing Methods	Common Applications
Autoscaling	Chemical reactors, pharmaceutical processes
Sphering	General multivariate data (of nonspectroscopic origin)
Quantile normalization	Metabolomic and other systems biology data

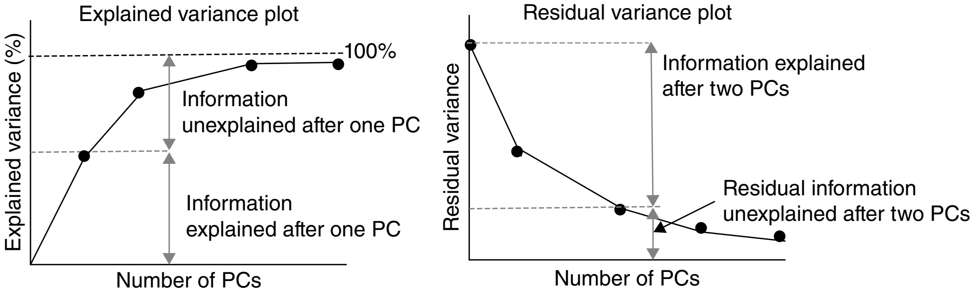


FIGURE 65.16 Explained and residual variance plots for well-behaved models.

As was the case with spectral/chromatographic preprocessing, subject matter expertise is the best way of deciding how to scale a data set. It is important to keep in mind that scaling is used to enhance information extraction during chemometric analysis and for no other purpose. Only use a transformation where it is warranted; otherwise, the model can become overcomplex for the wrong reasons.

65.3.3.5 Explained/Residual Variance There are a number of steps involved in the validation and interpretation of a PCA model. Once a data set has been loaded into a suitable chemometric software package, preprocessed, or scaled, the appropriate validation method has been selected (refer to Section 65.6), and the analysis completed, the first step is to look at the complexity of the model.

The principle of parsimony is very important in chemometric modeling, that is, the simplest models are usually the most robust and easiest to interpret. Most chemometric programs provide a user with the explained or residual variance plots. These plot how much information is captured by each PC, either as:

1. **Explained variance**: This has a maximum value of 100% and plots how much each PC contributes to explaining the information in the data set.
2. **Residual variance**: This has a minimum value of zero and plots the residual variance in the original data after each PC has been added to the model.

Figure 65.16 provides examples of both plots for well-behaved data.

The faster these plots converge toward a plateau (i.e., toward 100% for explained variance or zero for the residual variance), the more systematic information is contained

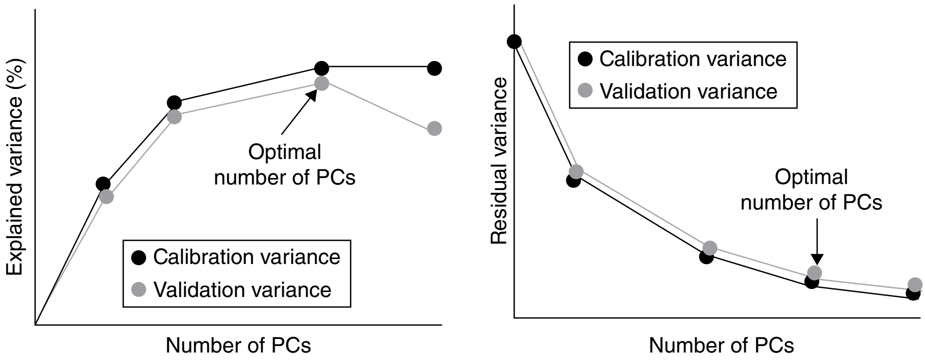


FIGURE 65.17 Explained and residual variance plots calibration and validation data.

in the model. Before model interpretation is attempted, an analyst should have some basic idea as to the sources of variation present in the data. For example, if the explained variance plot converges to 100% after only two PCs, it can safely be assumed that there are only two major sources of variation in the data to be interpreted.

Good chemometric software packages allow a user to view the calibration and validation variance of a model in one plot. Calibration variance is calculated by adding a component to a model and using the model to predict the calibration samples. Although this may seem invalid statistically, the calibration variance is used to determine the point where all information has been captured by the model and therefore provides a baseline for determining model complexity. Validation variance is used to assess the model performance using some form of validation (either cross validation or independent test set). Since the noise characteristics of the validation set may differ from the calibration set, comparison of the two curves allows an analyst to determine how many components to safely interpret. Figure 65.17 provides two examples of explained and residual variance curves for calibration and validation data.

In the case of the explained variance curve, the calibration curve continues to converge toward 100%, while the validation curve diverges at the thirrd PC. In the case of the residual variance plot, the calibration and validation curves follow each other closely and converge together with no signs of diverging. In this case, the point where the two curves plateau is the best place to start interpreting the model.

These variance plots are also useful diagnostic tools for detecting outliers in the data set. This is discussed in more detail in Section 65.3.3.5. Once the number of PCs to interpret in a model has been established, a detailed assessment of scores and loadings can be performed.

65.3.3.6 Residuals and Diagnostic Tools in PCA One of the most powerful aspects of PCA is its comprehensive graphical and diagnostic toolkits for validating and interpreting a model. This section provides a brief description of some of the methods used to detect outliers in a model and also effectively interpret it for use in practical applications.

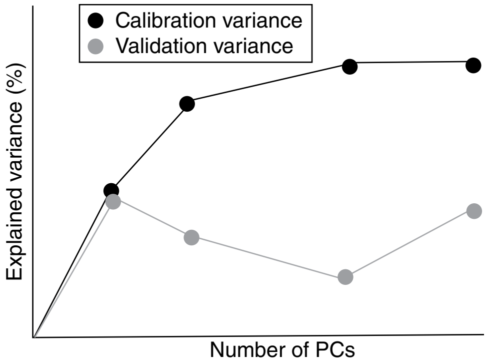


FIGURE 65.18 Outlier detection in the explained variance plot.

Outliers in PCA The explained/residual variance plots introduced in Section 65.3.3.5 are useful in determining model complexity, but they can also be used as a first check for the presence of outliers. By definition, an outlier can be classified into one of the following situations:

1. Measurement error
2. Wrong labeling
3. Deviating sample
4. Noise
5. Extreme/interesting sample

From the aforementioned definition, an outlier does not necessarily have to be a bad measurement. For example, if a group of 10 samples were measured and 1 had different natural characteristics than the other 9, this is not a measurement error or noise, but an interesting sample (or possibly the discovery of a new sample class).

A gross outlier can usually be detected in the explained/residual variance plot by an abnormal structure in the validation curve. Figure 65.18 provides an example of such a curve and the erratic behavior of the validation curve profile.

When such behavior is observed, an analyst must then look at the scores or influence plots (section "Influence Plots") to look for the samples causing the deviation.

Hotelling's T^2 The Hotelling's T^2 statistic [19] is a multivariate generalization of the Student *t*-test. It is similar to the Mahalanobis distance [20] reported by texts on chemometrics but has the advantage that statistical limits can be placed on the distances. The form of the Hotelling's T^2 statistic is as follows:

$$T^2 = (X - \bar{X})W^{-1}(X - \bar{X})$$

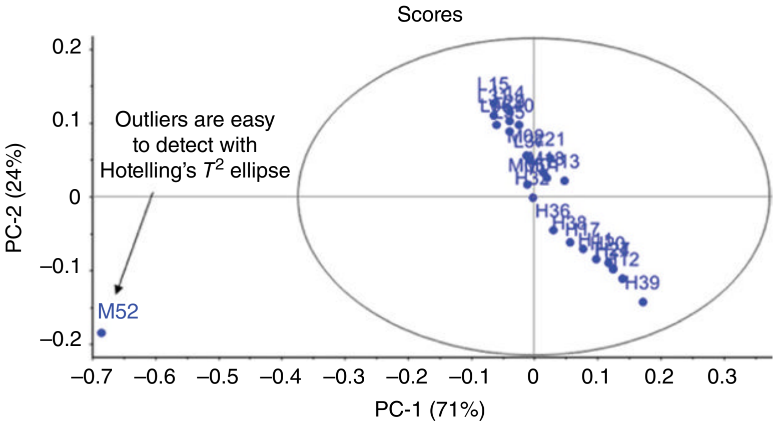


FIGURE 65.19 Using Hotelling's T^2 ellipse in scores space to detect outliers.

where

X is the original data matrix

$\bar{X}$ is the mean of the data set

W is the covariance matrix of X

The Hotelling's T^2 statistic is approximately F-distrubuted as follows:

$$F_{p,n,\alpha} \sim T^2 \frac{(n-a)}{a(n-1)}$$

Any sample that has a calculated F-value that exceeds the critical F-value can be considered for investigation as an outlier. In chemometrics the scores rather than the original variables are used in the calculation of Hotelling's T^2. A convenient way to display the Hotelling's T^2 statistic is by displaying it on a scores plot, where it can be shown at a number of levels of significance. Figure 65.19 shows how the T^2 ellipse can be used to detect an outlier in scores space.

Hotelling's T^2 is one of the most important statistics in MSPC. This topic is discussed in more detail in Section 65.7 where uses of the T^2 chart will be provided.

Leverage A closely related diagnostic tool to Hotelling's T^2 for detecting outliers is leverage (h_i) and is defined as the potential of a sample to be influential [15]. Leverage is calculated as follows:

$$h_i = \frac{1}{N} + x_i'(X'X)^{-1} x_i$$

where

N is the number of samples in the model
x_i is the centered x-vector for sample i

This equation defines an ellipse and samples with equal leverage lie at equal distances around the center of the model. There are no strict rules for setting up when a sample is a leverage outlier and a general rule of thumb is to investigate a sample whose leverage is 2–3 times larger than $(1 + A)N$, with A being the number of PCs used in the model. There is a one-to-one relationship between Hotelling's T^2 and leverage where leverage typically is used as outlier criteria when the number of samples is limited.

X-Residuals An X-residual is defined as that part of the data that has not been modeled. Expanding on the principle that data is equal to information plus noise, the X-residual is a measure of the noise after a PCs have been taken into account. Expanding on the form of the PCA model, the following equation can be derived:

$$X = TP' + E = \sum_{a=1}^{a} t_a p'_a + E$$

This equation can be rearranged as follows:

$$E = X - \sum_{a=1}^{a} t_a p'_a$$

where the residual E is what remains after a PCs have been subtracted from the original data set. This can be simplified down to individual sample residuals as follows:

$$e'_i = x'_i - t_a p'_a$$

A desirable property of residuals is that they should be randomly distributed without any systematic structure. If a residual has any form of interpretable structure, there are two main causes:

1. Not enough PCs have been added to the model to account for the systematic variation still remaining in the data.
2. The sample is an outlier and its structure is not well described by the model.

X-residuals in PCA find most use for the analysis of spectral outliers; this is because regions where the spectra are not being adequately modeled can be visualized and interpreted. Figure 65.20 provides an example of a single outlier and the diagnostics available to detect the outlier. Note the structure remaining in the X-residuals.

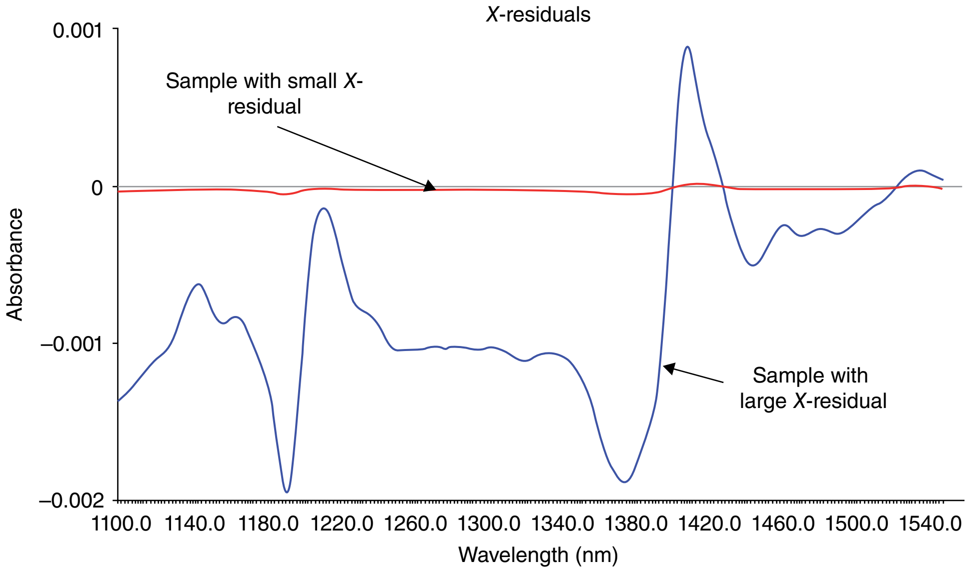


FIGURE 65.20 Using *X*-residuals to detect spectral outliers.

X-residuals alone can be subjective when using them as a visualization tool. Taking the sum of squared residual values can reduce the *X*-residual to a single point for each sample. To be able to assess these residuals objectively, the use of *Q*-residuals and *F*-residuals is discussed in the next sections.

Q-Residuals *Q*-residuals were first introduced by Jackson and Mudholkar [21] to detect the so-called type B outliers. These outliers can be the result of:

1. Too few PCs used to adequately describe the original data *X*.
2. The samples are truly outliers from the model.

The *Q*-residual is calculated from the regular *X*-residual as a squared sum:

$$Q_i = e_i^t e_i = \left(x_i - t_a p_a'\right)'\left(x_i - t_a p_a'\right)$$

The critical value for *Q* can be obtained from the following formula:

$$Q_\alpha = \theta_1\left[\frac{c_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0\left(h_0 - 1\right)}{\theta_1^2} + 1\right]^{\frac{1}{h_0}}$$

where Q_α is the critical value for the *Q*-distribution and the remainder of the terms can be found in the text by Jackson [22]. The use of the *Q*-residual in MSPC applications is very important for detecting the onset of failure before it becomes a critical issue. Refer to Section 65.7 on MSPC for more details.

F-Residuals F-residuals are calculated from Q-residuals and are considered more conservative by many with respect to Q-residuals. The F-residuals are calculated as follows:

$$F_i = \frac{Q_i}{(K - A)}$$

where

F_i is the F-residual
Q_i is the Q-residual
K is the number of variables
A is the number of PCs used in the model

The F-residual is compared for significance against the standard F-test hypothesis:

$$F_{\text{crit}} \sim (\alpha, 1, i)$$

where i is the number of samples in the model. The major advantages of the F-residual over Q-residuals include the following: the F-test is a more established test compared to the Q-test, the calculation of F-residuals is computationally faster, and, most importantly, F-residual analysis can be applied to both calibration and validation residuals, where Q-residuals are only applicable to calibration residuals.

Influence Plots A convenient way to visualize the presence or absence of outliers is to plot leverage on the x-axis and X-residual on the y-axis. This is known as an influence plot. The idea behind the influence plot is shown in Figure 65.21.

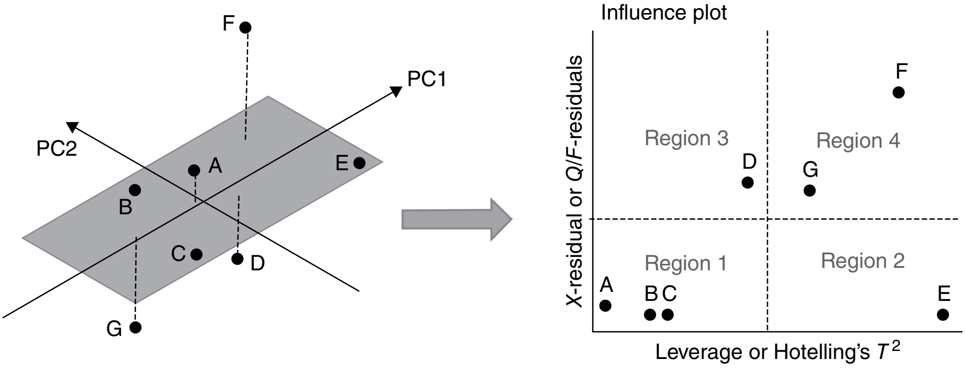
Consider the plane formed by two PCs in Figure 65.21. Samples that lie close to the plane but are extreme along the PC axes are leverage samples, that is, they can influence the orientation of the plane if they become too extreme with respect to the center of the model.

Samples that do not lie on the plane in an orientation perpendicular to the plane show some form of X-residual. As samples lie further away from the plane, they do not fit the model well and are therefore X-residual outliers. When a sample is extreme in both X-residual and leverage, it can in most cases be concluded that the sample is a true outlier.

There are a number of variants of the influence plot; the most common forms are:

1. X-residual versus leverage
2. Q-/F-residual versus Hotelling's T^2

Any combination of the aforementioned variants is possible. The advantage of using Q-/F-residuals and Hotelling's T^2 in influence plots is such plots allow the placement



Region 1: Samples similar to the majority of the calibration population.
Region 2: Samples fit model but are extreme in properties.
Region 3: Samples differ from the average model population.
Region 4: Samples are different and extreme (most likely outliers).

FIGURE 65.21 The idea behind the influence plot.

of statistical limits on the plots, thus providing objective evidence for the presence or absence of outliers. Figure 65.21 shows an influence plot with statistical limits. Such a plot forms the basis of MSPC control charting introduced in Section 65.7.

65.3.3.7 Application of PCA to Fisher's Iris Data To demonstrate the graphical power and diagnostic capability of the PCA method, the iris data set introduced earlier will be used. Preliminary investigation of the data shows that only four variables were measured, so a maximum of three PCs should be interpreted. If less PCs are needed, the more similar is the information contained in the original variables. Figure 65.22 shows the PCA overview for the iris data analysis.

Explained Variance The plot of explained variance shows that PC1 describes approximately 92% of the total data variability and PC2 a further 5% for a total of 97% explained in two PCs. PC3 only contributes slightly and can be excluded from the analysis. The calibration and validation curves follow each other closely; therefore, a 2-PC model can be validated and possibly interpreted.

Scores Plot It is now justified to interpret a 2-PC model only. The scores plot in Figure 65.22 has been grouped based on iris type. It can be interpreted from this plot that:

1. Setosa can be uniquely distinguished from versicolor and virginica.
2. Versicolor and virginica are slightly distinct from each other; however, the potential for overlap can be seen, therefore leading to ambiguous interpretation of the data.

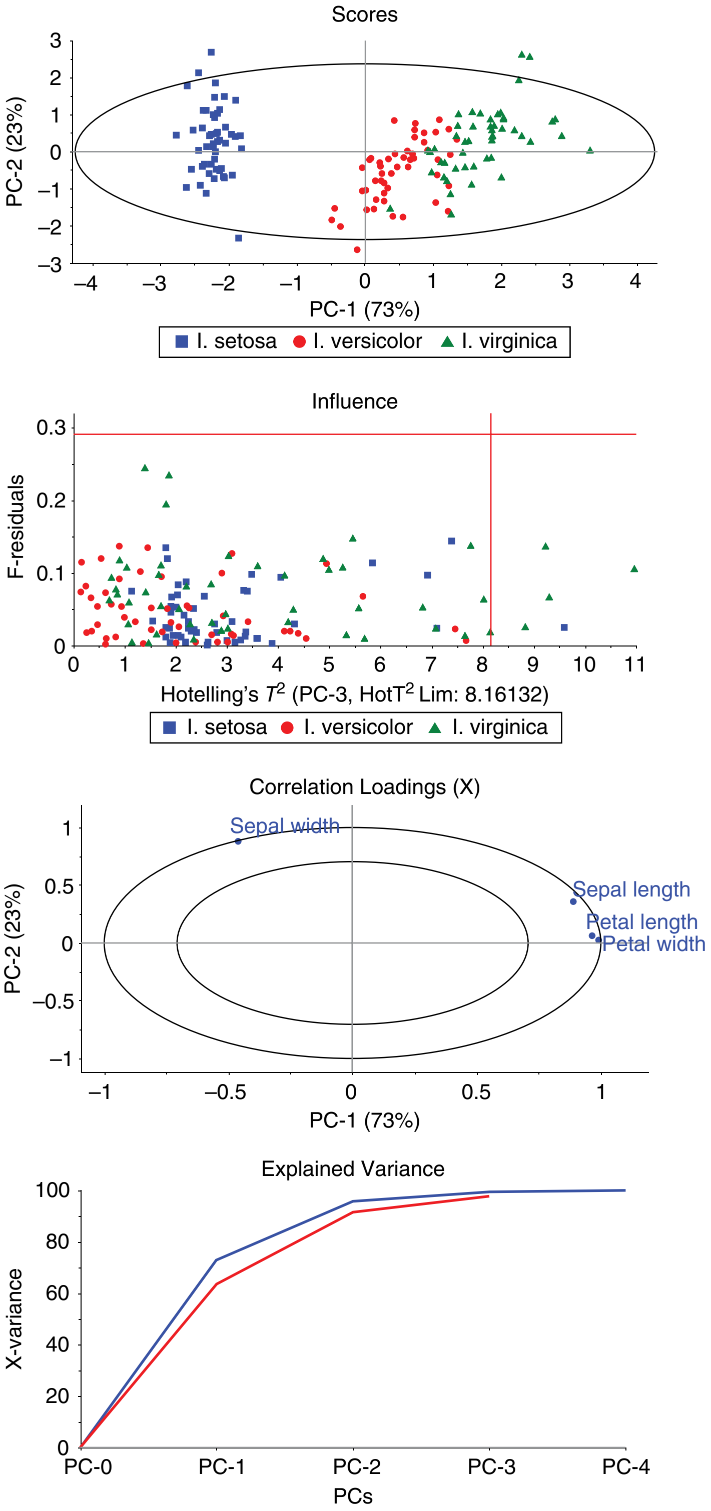


FIGURE 65.22 PCA overview of Fisher's iris data.

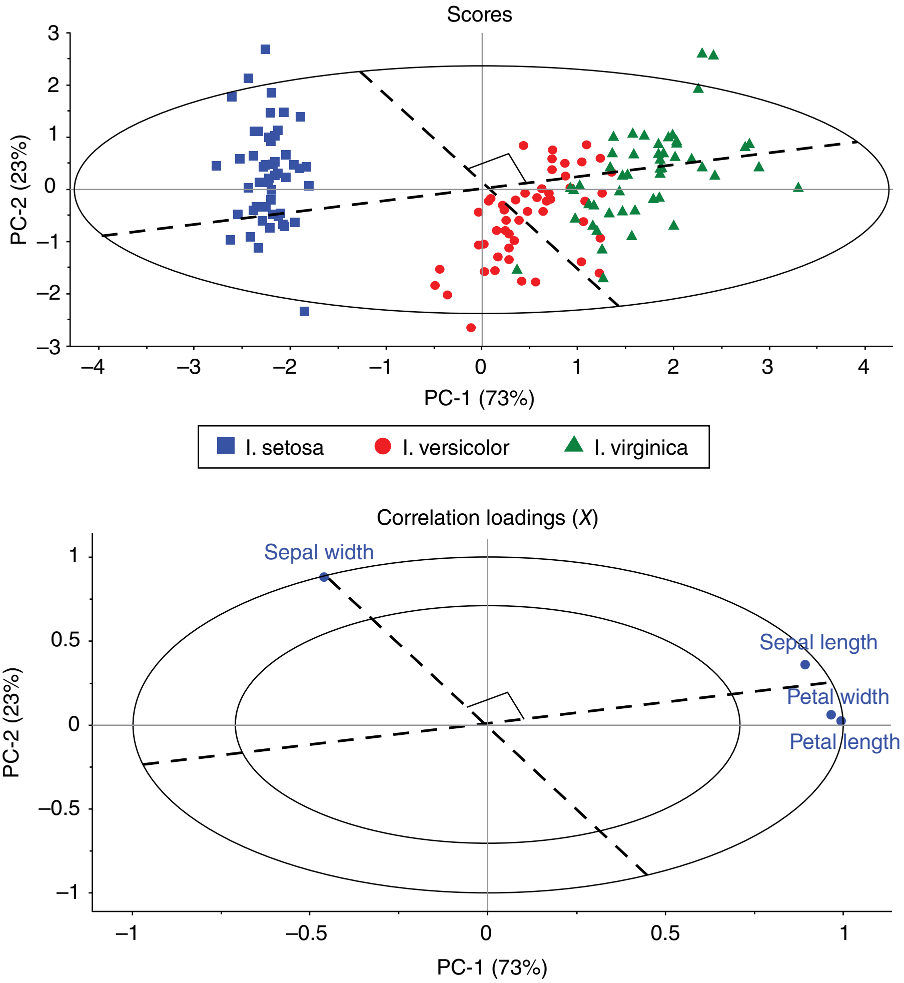


FIGURE 65.23 Geometrical interpretation of PCA scores and loadings for Fisher's iris data.

To understand why the samples group and spread the way they do, an interpretation of the loadings plot is required.

Loadings Plot For the purposes of clarity, the correlation loadings plot is shown in Figure 65.22. Along PC1, petal length and petal width almost lie exactly on the positive PC1 axis. This indicates that these two variables contribute most to PC1 and since they lie on top of each other, it can be assumed that they are highly correlated.

Sepal length also lies in the positive PC1 direction, but since it also occupies some of the PC2 space, it contributes to the spread of the samples along the PC2 direction in the scores plot. The largest contributor to the spread in the PC2 direction is sepal width. Figure 65.23 shows the scores and loadings plot for the iris data, this time showing the geometrical interpretation of the data.

Starting with the loadings, the direction that captures the joint petal length and width and the sepal length describes why setosa is different from versicolor and virginica; in particular, versicolor and virginica are distinguished from setosa based on larger values of petal length and width and sepal length.

Two new lines were plotted in the scores and loadings plot, and it can be seen that they are nearly orthogonal to each other. This indicates that the effect of sepal width is independent of the effect of the other variables. As PC2 only describes approximately 5% of the variability, sepal width is the minor contributor to distinguishing between the three classes of iris. Removal of this variable actually has no effect on the discrimination power of the PCA model. The recalculated PCA without sepal width is shown in Figure 65.24.

This ability to remove variables and recalculate a model based on observation and the interpretation of diagnostics is what makes PCA one of the most powerful data analysis methods available.

Outlier Analysis Figure 65.22 shows the F-residuals versus Hotelling's T^2 influence plot for the iris data. The plot shows that there are no observable samples grouping plot and no X-residual outliers, and therefore the data set is representative and indicates that all samples fit the model well. There are, however, six potential leverage outliers, primarily from virginica and one from setosa but one from versicolor. Figure 65.25 shows the scores plot with the 95% Hotelling's T^2 ellipse drawn and the influence plot with leverage outliers marked. The leverage samples are those that exceed the Hotelling's T^2 boundaries in PC1 and PC2. These samples are extreme within the data set, but they are close to the boundaries. These samples were not considered to highly influence the model.

PCA Summary for Fisher's Iris Data Overall, the objective of the analysis was achieved; the original four-dimensional data set (i.e., four variables) was reduced to a 2D problem. The information in petal length and width was found to be highly correlated, so there is no advantage to having both measures in the analysis.

It was also found that sepal width did not contribute to the separation of the classes and after elimination, there was no change in the data structure. The ability to justifiably remove unimportant variables is a powerful tool of PCA. PC1 was found to be the main contributor to distinguishing between setosa, versicolor, and virginica. The variables most responsible for this distinction were petal length and width and sepal length. Sepal length was responsible for the spread of the sample data in PC2.

An outlier analysis detected no X-residual outliers. There were some leverage outliers detected; however, when assessed with the scores plot with Hotelling's T^2 ellipse at 95% confidence, these extreme samples were found not to be too extreme and did not influence the model.

65.3.3.8 Algorithms for PCA There are two main algorithms commonly used in software packages for calculating PCs. These are the SVD algorithm first introduced

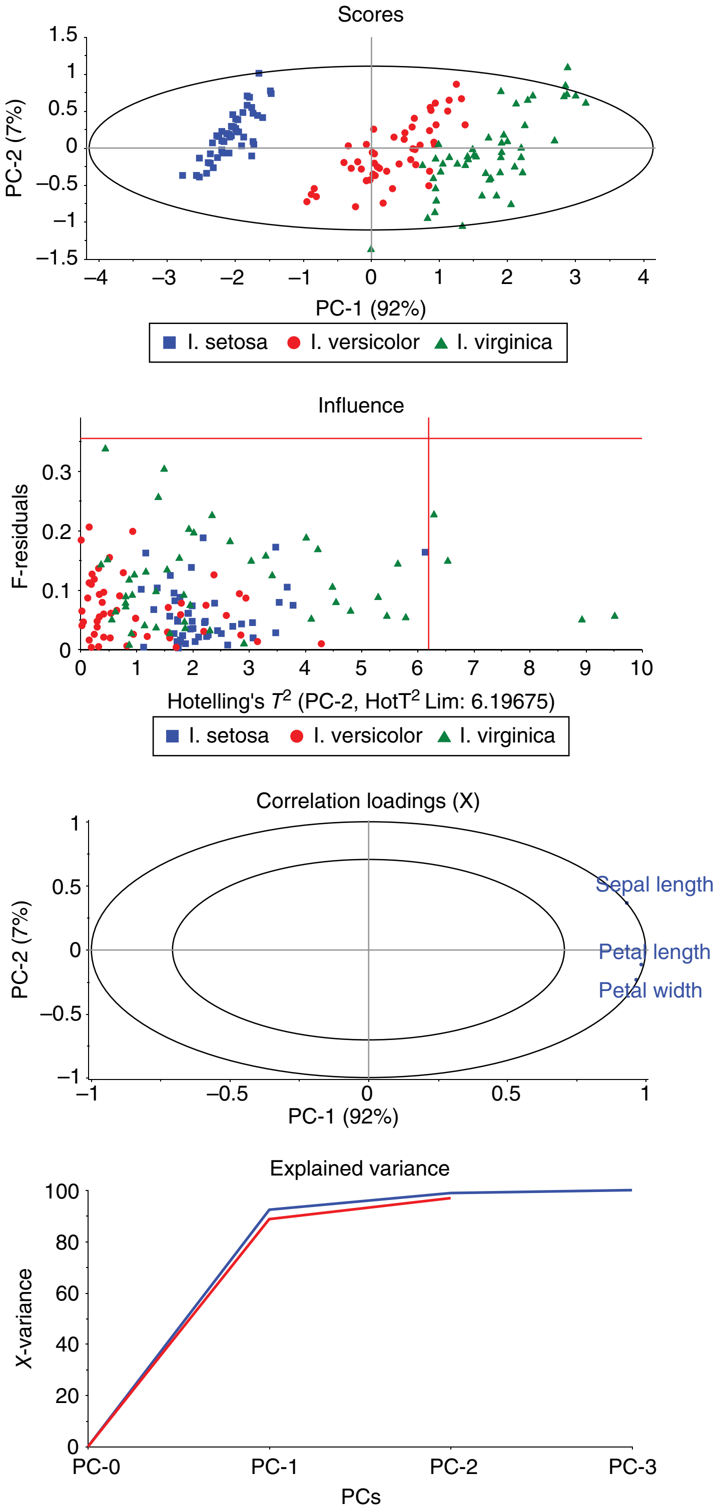


FIGURE 65.24 PCA overview of Fisher's iris data after the removal of an unimportant variable.

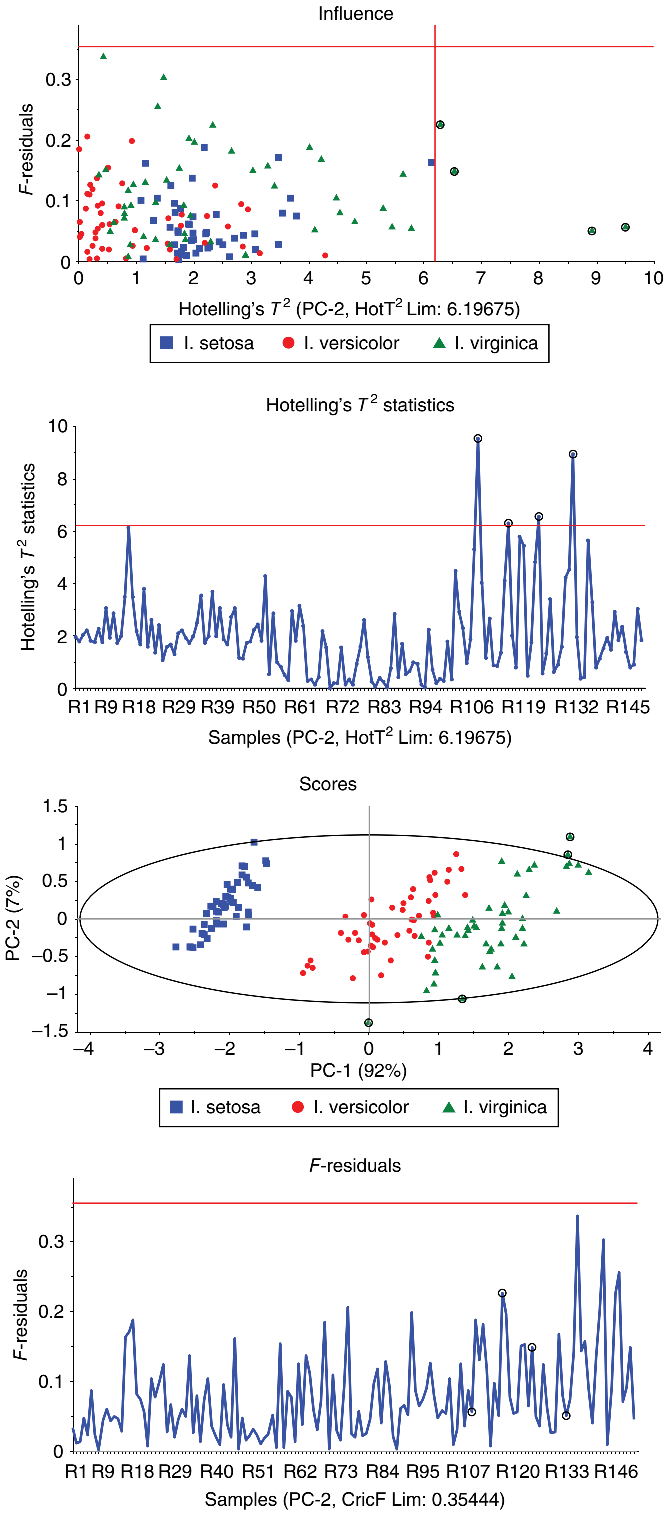


FIGURE 65.25 Outlier analysis of Fisher's iris data.

by [22] and the noniterative partial alternating least squares (NIPALS) algorithm first introduced by Wold [23].

Singular Value Decomposition In SVD, a matrix X is decomposed into a product of the so-called characteristic vectors of $X'X$, the characteristic values of XX', and a function of their characteristic roots [22]. Now, if all of the PCs are used to describe the data, the original matrix X can be regenerated as follows:

$$X = TP'$$

SVD calculates all of the characteristic vectors of a data set in one operation. The following describes in general how the algorithm works.

The general form of the SVD model is defined as

$$X = U\Lambda P'$$

where

X is the original data to be analyzed

U is an $(N \times K)$ orthogonal matrix containing the so-called left singular vectors (N is the row dimension of U)

Λ is a symmetrical matrix of diagonal elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ (p being the column dimension of both U and P) (the diagonal elements of Λ are called the singular values and describe the importance of each PC being calculated; singular values are the square roots of the eigenvalues calculated from the covariance matrices of the original data X)

P is a $(p \times p, p \leq K)$ orthogonal matrix containing the so-called right singular vectors

The matrix U is calculated by eigenanalysis from the matrix XX' whose dimension is $(N \times N)$, that is, U describes sample characteristics and correlations. Correspondingly, the matrix P has dimension $(p \times p)$ and describes variable characteristics and correlations.

In order to relate U to P, the singular value matrix Λ describes the importance of each row of U based on the variable contributions in P. It is common practice in PCA to combine the matrix product $U\Lambda$ and define it as the vector T, that is, the scores vector. The common PCA model definition can now be stated as

$$X = TP' + E$$

where the matrix E is the residual that cannot be explained by the model. By combining $T = U\Lambda$ it can now be seen how the explained variance of each PC can be calculated. If X contains 100% unexplained variability, then each singular value λ_a describes a certain proportion of the original variability, with the proviso that the scores vector t_1 contains the most information followed in order by the remaining scores vectors. The trace of Λ provides an estimate of the total variability described by the PCA model, and its dimension can be used to determine the rank ($p \leq K$) of the original data (i.e., the number of important features in the data).

For a more detailed description of the algorithm details of SVD, the interested reader is referred to the text by Jackson [22] and Hastie [24].

Noniterative Partial Alternating Least Squares The algorithm extracts one component at a time, with each component obtained iteratively by repeated regressions of X onto the scores $\hat{t}$ to obtain improved loading vectors $\hat{p}$ and then repeated regressions of $\hat{p}$ on X to obtain improved $\hat{t}$ [15, 18].

The following algorithm assumes that all preprocessing and centering of the data have been performed prior to analysis:

Step 1: Choose an initial scores vector $\hat{t}_a$ in X as the column with the highest remaining sum of squares.

Step 2: Improve the estimate of the calculated loading vector $\hat{p}_a$ by projecting the matrix X_{a-1} onto $\hat{t}_a$, that is,

$$\hat{p}'_a = \left(\hat{t}'_a \hat{t}_a\right)^{-1} \hat{t}_a X_{a-1}$$

Step 3: Scale the length of $\hat{p}_a$ to 1.0 to avoid scaling ambiguity,

$$\hat{p}_a = \hat{p}_a \left(\hat{p}'_a \hat{p}_a\right)^{-0.5}$$

Step 4: Improve the estimate of score $\hat{t}_a$ for the component by projecting the matrix X_{a-1} onto $\hat{p}_a$,

$$\hat{t}_a = X_{a-1} \hat{p}_a \left(\hat{p}'_a \hat{p}_a\right)^{-1}$$

Step 5: Improve the estimate of the eigenvector $\hat{\tau}_a$,

$$\hat{t}_a = \hat{t}'_a \hat{t}_a$$

Step 6: Check convergence by subtracting the $\hat{\tau}_a$ from the previous iteration. If the difference is smaller than some predetermined tolerance, the method has converged; otherwise, perform another iteration (in most software applications, a user will be allowed to specify a maximum number of iterations to converge).

Step 7: Once convergence has been reached for a particular component, subtract this from the starting X-data,

$$X_a = X_{a-1} - \hat{t}_a \hat{p}'_a$$

Step 8: Go to step 1 and repeat to step 7 for the deflated matrix to calculate the next (and successive) components.

It can be seen from the aforementioned algorithm description why it is referred to as an iterative algorithm as it aims to improve the scores and loadings precision through convergence criteria. This is the algorithm used in most software packages as it is relatively fast, but it has the distinct advantage over SVD that it can handle missing values in the original *X*-matrix. On the other hand, SVD is based on algorithms that give exact results from back-substitution or equivalent procedures and are regarded as numerically superior.

65.3.3.9 Summary of PCA PCA is known as the "workhorse" of MVA methods [18]. As an EDA method, PCA is unrivaled with its ability to provide a highly graphical environment along with many diagnostic tools that can help interpret and improve a model.

PCA separates a data set into sample information (scores) and variable information (loadings). Used in combination, scores and loadings can provide extensive insights into complex data sets. Not only does PCA allow understanding of sample and variable relationships, but it also provides a means to deciding the complexity of the model using explained or residual variance plots. PCA models are validatable using either external data sets or internal segregation of samples for training and validation purposes (refer to Section 65.6 for more details on validation).

The incorporation of statistics into MVA methods allows the definition of multivariate confidence intervals. This provides PCA with the capability of objective outlier detection. Diagnostics such as Hotelling's T^2 ellipses and intervals and *Q*- and *F*-residuals are available at a number of statistical confidence levels. This is highly important for detecting and justifiably removing outliers from a data set. The properties have given rise to MSPC applications further discussed in Section 65.7, where PCA models can supplement traditional SPC models for enhanced understanding of many processes in many industries.

PCA allows visualization of clusters in data sets containing many classes of samples. If separation of the classes can be established visually, individual PCA models can be developed and used in multivariate classification (Section 65.5). Even if complete separation of all classes is not possible, the joining of classes into a single PCA model is possible.

The next section of this chapter focuses on multivariate regression. It will be seen in this section that PCA also forms the basis of many of the methods discussed. Overall, PCA is a recommended first approach to the analysis of any multivariate data set. After the application of an appropriate preprocessing method (Section 65.3.3.4), PCA will provide greater insights into many data sets and avoids the need for the common "one-variable-at-a-time" approach that may lead to false or no conclusions at all.

65.4 MULTIVARIATE REGRESSION

This section provides an overview of some of the most common multivariate regression method currently used in research and industrial practice.

65.4.1 General Principles of Univariate and Multivariate Regression

Regression is a mathematical approach for relating two or more sets of variables to each other [15, 25]. In regression modeling, a representative set of *X*-variables (where *X* is multivariate) is used for modeling one or several *Y*-variables, also known as response variables. Regression methods have played an important role as a tool in analyzing a multitude of samples of various kinds from the pharmaceutical, food and beverage, agricultural, and many other industries.

The ideal situation in a regression context is that the change in the *X*-variable(s) is 100% related to the change in the *Y*-variable, thus requiring high sensitivity of at least one *X*-variable in the set to the response. In many cases it is the simultaneous contribution from several *X*-variables that enables a multivariate modeling of the property under investigation.

As was the case in PCA (Section 65.3.3), the general regression model has the following form (repeated here for clarity):

$$\text{DATA} = \text{INFORMATION} + \text{NOISE}$$

The main difference between PCA and regression is simple; PCA models the internal structure of the *X*-variables in one table, where regression models the relationship between *X*- and *Y*-variables in two (or more) tables. Mathematically, the regression model is represented by

$$Y = XB + E$$

where

Y is the variable to be predicted (response or dependent variables)

X is the variables measured to predict the response (predictors or independent variables)

B is a matrix of regression coefficients (i.e., the model that relates the *X*-variables to the *Y*-variables)

E is a matrix of residuals representing the lack of fit of the model

To introduce some key concepts related to the construction of regression models, the simplest case occurs when a single response has to be predicted based on the measurement of a single independent variable. This case occurs frequently in the physical and chemical sciences and engineering. In particular, a common application in chemistry is where the concentration of an analyte must be quantified based on the measurement of a univariate physical property or of an individual instrumental signal (e.g., absorbance at $\lambda_{\max}$ in Beer's law [11] or diffusion limit current in Ilkovič Equation [26]). In mathematical terms, the univariate regression model can be stated as

$$y_i = f\left(x_i\right) + e_i = \hat{y}_i + e_i$$

where, for each object, the predicted response value $\hat{y}_i = f(x_i)$ represents an approximation of the true value y_i, the difference between the two being the residual e_i. Although the general functional relation $f(x)$ is used in this equation, it is assumed that the most common mathematical form of the regression models is a linear model. Therefore, under the assumption of a linear relationship between the dependent and the independent variables, the model becomes

$$y_i = \hat{y}_i + e_i = b_1 x_1 + b_0 + e_i$$

where b_1 is the slope and b_0 is the intercept of the model (or simply put, the straight line model of common form $y = mx + b$). The generation of a regression model requires finding the optimal values of the model parameters b_1 and b_0. In this context, the criterion that is normally used is the so-called least squares criterion [27, 28], which is based on a sum of squared error loss function. In particular, the concept behind approximating the functional relation between X and Y via a least squares model is to look for the value of the parameters that allows fitting the data with the minimum possible error [29]:

$$\min_{b_0, b_i} \sum_{i=1}^{r} e_i^2 = \min_{b_0, b_i} \sum_{i=1}^{r} \left(y_i - \hat{y}_i\right)^2 = \min_{b_0, b_i} \sum_{i=1}^{r} \left(y_i - b_1 x_1 - b_0\right)^2$$

r being the number of X-/Y-pairs used to build the model (training set) and the rest of the symbols as previously defined. The optimal value of the parameters is the one that minimizes the sum of squares of the model residuals, known as least squares fitting. The interested reader is referred to the literature for the derivation of the normal equations used to solve the least squares problem [25].

The simple univariate model described previously is easily generalized to the multivariate situation. The general form of the multivariate regression model is as follows:

$$y_i = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + b_0 + e_i = \hat{y}_i + e_i$$

This is the general form of the multiple linear regression (MLR) model discussed in Section 65.4.2.

65.4.2 Multiple Linear Regression

With reference to the simple univariate least squares approach discussed in Section 65.4.1, MLR is relatively simple and straightforward. However, the mathematical structure of the model can severely limit the possibility of its practical application to many real cases, where a large number of variables are measured on a relatively small number of samples. The general model for finding the regression coefficients in the MLR model is shown as follows:

$$\mathbf{b}_{\mathrm{MLR}} = X^{+}y = (X'X)^{-1}X'y$$

where the symbol + indicates Moore–Penrose pseudoinverse [30, 31].

Estimation of the optimal value of the regression coefficients relies on the inversion of the matrix ($X'X$), and for many experimental data this inverse doesn't exist or it is ill-conditioned (i.e., it is unstable when the model is applied in practice).

In particular, the conditions that have to be satisfied in order for $(X'X)^{-1}$ to be estimated in a reliable way are that the columns of X are linearly independent (meaning that the predictors are uncorrelated) and that the number of training samples r is greater than the number of independent variables p. From a practical standpoint, the latter condition could be, at least in principle, met by increasing the samples to variables ratio either by measuring more samples or by variable selection. However, the former is rarely satisfied, especially when signals coming from modern instrumentation are involved, as the variables are quite often correlated by nature or by sampling. The direct consequence of the matrix X being ill-conditioned is that the coefficients are not stable and are characterized by high variance, since the solution is mostly affected by the noise part of the data [15]. Dependent on the underlying structure of X, this may also give higher prediction error. This must be investigated by proper validation. It should be mentioned that this instability may also lead to false interpretation of the coefficients as well as the interpretation of results from ANOVA (p-values) and that these pitfalls occur long before there are numerical problems in inverting ($X'X$). For this reason many implementations of MLR check the so-called condition number to give warnings about the rank of X. However, to take out some variables because they happen to be correlated to others due to the fact that one is not applying a suitable method to handle this situation is scientifically unsatisfactory.

To deal with these drawbacks, different methods have been proposed in the literature, most of them based on the concept of bilinear modeling, already introduced in Section 65.3.3. Indeed, when the description of the data set using the experimentally measured variables is substituted by a more parsimonious one, relying on the concept of latent variables, then it is often possible to capture the essential structure of the data with a very limited number of descriptors. It is then evident that these two characteristics (low number of mutually orthogonal predictors) allow overcoming all the limitations described previously and make multivariate calibration applicable to a wider host of real-world problems. In this framework, the most commonly used latent variable-based methods are principal component regression (PCR) and partial least squares regression (PLSR), which will be described in Sections 65.4.3 and 65.4.4, respectively.

65.4.3 Principal Component Regression

As the name suggests, PCR [15, 32] is based on the use of PCA [22] (see Section 65.3.3 to produce a parsimonious description of the independent matrix *X*).

Indeed, as projection of the samples onto the first PCs constitutes the best low-dimensional approximation of the original data matrix, the natural extension is the use of PCA scores as the independent variables in the MLR problem, which may overcome the limitations of the method when dealing with ill-conditioned experimental matrices. Therefore, PCR modeling is a two-step process firstly involving PCA decomposition of the *X*-variables and successively the generation of an MLR model on the scores [32].

In matrix notation, the independent matrix *X* is described by the bilinear model already discussed in Section 65.3.3 and repeated for clarity here:

$$X = TP' + E$$

where *T* and *P* are as previously defined as the scores and loadings matrices, respectively, while *E* are the *X*-residuals of the model. Based on this decomposition, the PCR method proceeds by building an MLR model on the scores computed in the PCA step. Accordingly, the regression model for PCR is

$$Y = \hat{Y} + E_Y = TB + E_Y$$

where the subscript *Y* was added to the *Y*-residual matrix to differentiate it from that of the *X*-block and *B* is the matrix of regression coefficients for the MLR model relating the dependent variables *Y* to the PC scores of the independent block *T*. Extending the definition of the MLR model regression coefficients, the matrix *B* can be computed as

$$\mathbf{B}_{\mathrm{PCR}} = T^{+}Y = \left(T'T\right)^{-1} T'Y$$

where the regression coefficient matrix *B* relates the dependent matrix *Y* to the *X*-scores (*T*).

With respect to MLR, as PCR involves a projection step, where the data are represented on a low-dimensional latent variable space, there is the need of deciding what the complexity of this space should be or, in other terms, how many PCs are needed (refer to Section 65.3.3.5). In general, there is a trade-off in selecting the optimal number of components: including too few components, which could lead to models not able to fit *X* well and to predict *Y* accurately, whereas the use of too many components can result in overfitting Y and X. As a consequence, this may result in unreliable predictions on new samples. Therefore, the choice of model complexity is normally accomplished through some sort of validation procedure (see Section 65.6), in which the optimal number of PCs is selected as the one leading to the lowest prediction error on validation estimates.

One possible drawback of PCR modeling is that it relies on using the PCs as predictors for the responses, but PCs do not necessarily correlate with Y. Indeed, the main characteristic of PCA is to extract features that capture as much as possible the variation in X; however, in cases where many sources of uninformative variation and/or a high level of noise are present, they can be poorly related to the Y (and, hence, not predictive). To overcome this problem, some authors suggest to only choose those latent variables (PCs), correlating maximally with the responses [33].

65.4.4 Partial Least Squares Regression

As discussed in Section 65.4.3, PCR is a two-step process, in which the projection stage is separated and independent from the regression one and this can lead to the drawback that the components that are extracted in the decomposition step, based only on the information about the X-matrix, can be poorly predictive for the Y block. Starting from these considerations, an alternative method was proposed called PLSR [15, 34, 35], in which information in Y is actively used for the definition of the latent variable space. PLSR extracts components (known as PLS factors), which compromise between explaining the variation in the X-variables and predicting the responses in Y. This corresponds to a bilinear model, whose mathematical structure is summarized as

$$X = TP' + E$$
$$Y = \hat{Y} + E_y = UQ' + F$$

This definition is formally identical to that for PCR, although the calculated components and the model coefficients are not the same, as the two projections are governed by different criteria. In particular, a major difference revolves around the way scores T are defined for PLSR, that is, they are defined in a way such that they are relevant both for interpretation and prediction, through the statistical concept of covariance. Accordingly, PLSR is a sequential algorithm: the PLS latent variables are computed so that the first PLS component is the direction of maximum covariance with the dependent variables, the second PLS component is orthogonal to the first and has maximal residual covariance, and so on.

65.4.4.1 PLS Scores As was the case in PCA (Section 65.3.3.2), the typical scores plot as a 2D scatter plot provides a map of the objects where similarity between objects and groups of objects can be interpreted. In the case of PCR, the scores are computed from PCA on the X-data.

The major difference between PCA/PCR and PLS scores is that in PLSR the scores are estimated from X and the loading weights and are thus based on the covariance between X and Y. This means that PLS scores capture the part of the structure in X, which is most predictive for Y. Therefore, although PCA and PLS scores can be visualized and interpreted in a similar way, they are not the same. PCA scores model variations in X only, whereas PLS scores model variations in X most related to Y. This is an important concept that required this further elaboration.

65.4.4.2 X-Loadings in PLSR Assume the model $X = TP^T + E$ and then the loadings reflect the importance of all X-variables for each component/factor. For spectral data a plot of the loadings as a line plot may indicate if the factor carries information: if the vector looks like random numbers, it should not to be included as structure that can be modeled.

As the loadings in PCR are normally scaled to unit variance, there can be no ad hoc rule set if a loading value above a certain value is important. As an alternative the correlation loadings can be plotted, which are simply the correlation between the original variable and the scores vectors (refer to Section 65.3.3.3). For PLSR the loadings do not exactly have length = 1.0, but correlation loadings are still valuable for interpretation about how the variance in X is modeled from the scores.

65.4.4.3 Y-Loadings in PLSR The Y-loadings express the importance of the individual Y-variables for the factors $1:F$. The following equation depicts how Y is decomposed in the U scores and Y-loadings Q:

$$Y = UQ' + F$$

Refer to Section 65.4.4.8 regarding more details on the algorithmic details of the PLSR method.

In the version of the PLSR algorithm where the vectors w_a are scaled to unity, the inner relation coefficients are 1.0 and thus

$$Y = TQ' + F$$

65.4.4.4 Loading Weights Loading weights are specific to PLSR (they have no equivalent in PCR) and express how the information in each x-variable relates to the variation in Y summarized by the u-scores. They are called loading weights because they also express, in the PLSR algorithm, how the T-scores are to be computed from the X-matrix to obtain an orthogonal decomposition. The loading weights are typically normalized to 1.0. Variables with large loading weight values are important for the prediction of Y. The first loading weight vector in case of only one response variable is the covariance (or correlation if the variables are scaled to unit variance) between the individual x-variables and y.

X-loadings and *Y*-loadings or loading weights and *Y*-loadings are often shown together in 2D plots and interpreted similarly to loadings from PCA. Figure 65.26 provides an excellent example of the comparison of loadings and loading weights using a set of NIR spectra of gasoline samples with some samples containing an additive and the majority containing no additive.

In the top left plot of Figure 65.26, the NIR spectra of the samples are shown. Two classes of samples are present, those with an additive (yes) and those without an additive (no). The spectra are grouped based on this classification, and it can be seen that above 1380 nm, there are spectral differences between the two sample types.

The scores plot for the PLSR is shown in the bottom left quadrant of Figure 65.26. It can be seen that the samples with an additive separate from those without the additive, primarily along score factor 1. The top right-hand plot is the PLS loadings for the analysis. Since the loadings represent the structure contained in *X*, the spectral features greater than 1380 nm are found to be important. However, when the loading weights for these samples (shown in the bottom right quadrant), this region is weighted less, as they do not contribute to predicting *Y*. This example provides an excellent overview of the difference between loadings and loading weights, and this type of analysis must be performed as part of any PLSR model interpretation.

The scores and loadings may also be visualized together in a biplot [36], but there is no “truth” when it comes to scaling of the axes in a biplot and caution should be used in the interpretation of the relative position of the objects and variables [37].

65.4.4.5 Regression Coefficients Regardless of the type of regression, the model can be represented in terms of regression coefficients (*B*). The ideal situation occurs when the individual elements in the regression vector are directly interpretable as to provide the true model of the system. This is however only the case if the *x*-variables are orthogonal as in a factorial design [38]. When some of the *x*-variables are correlated, the situation of indeterminacy due to collinearity is encountered. As already discussed in this chapter for the latent variable methods, they handle collinearity from a numerical point of view but not necessarily from an interpretational point of view. Let a model for body weight be a function of two *x*-variables: height and shoe size. In the case of MLR, the model will use all the variance in *X* and the coefficient for shoe size may be negative although it clearly is a positive correlation between shoe size and body weight. In this case the underlying dimensionality is one and the regression vector from PLS regression will in this case reflect the true relationships between *X* and *Y* in the first PLS factor. Nevertheless, with latent variable models and a correct assessment of the dimensionality of the model from proper model validation, the regression vector may give valuable information about the underlying phenomena, be they physical, chemical, or biological in nature.

Assume that in a system there is one response variable *Y* of interest and other sources of systematic variance (e.g., other constituents, properties) that give rise to signals in *X*. Under noise-free conditions the regression vector estimated by PLSR is, up to

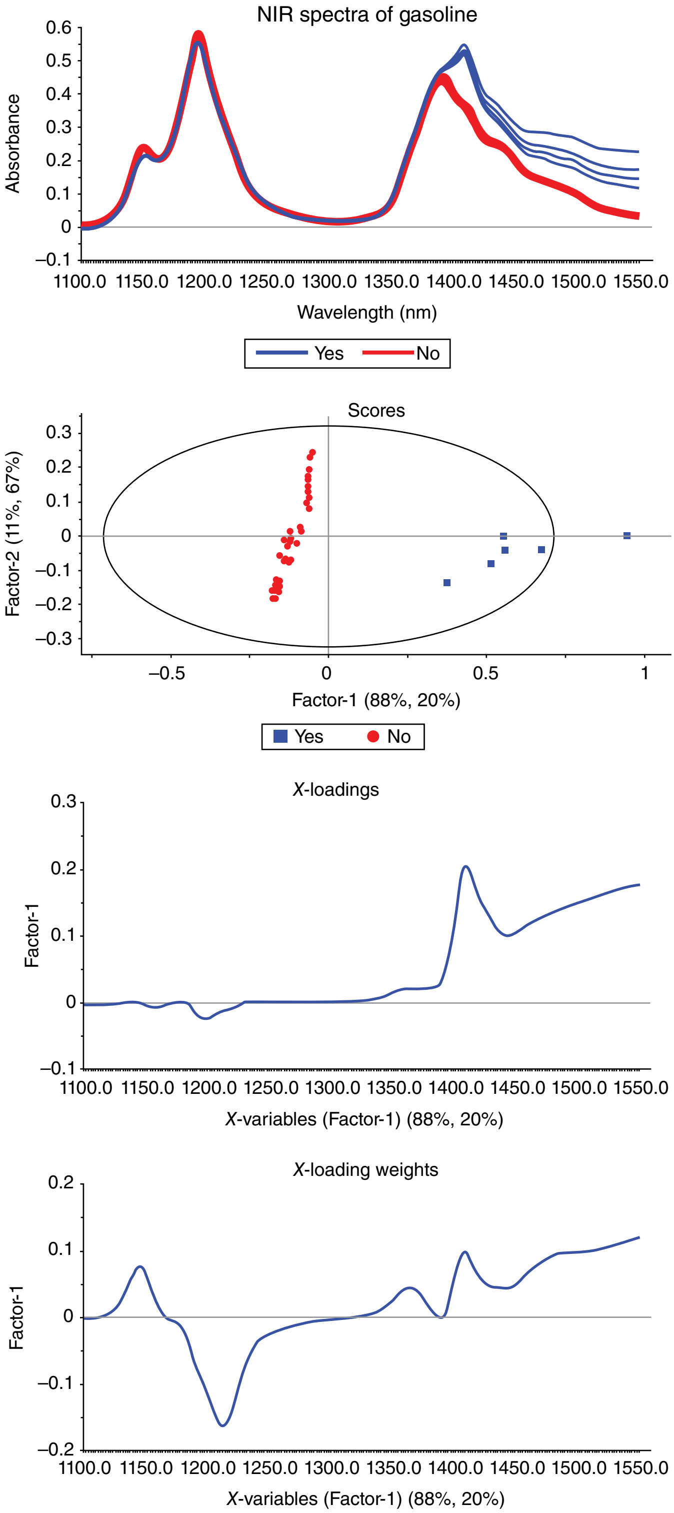


FIGURE 65.26 Comparison of PLS loadings and loading weights for the NIR spectra of gasoline samples.

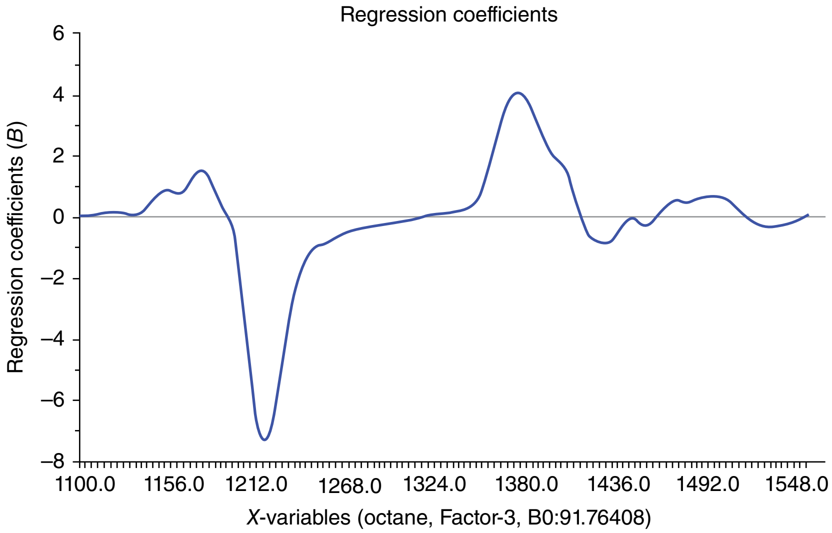


FIGURE 65.27 Regression coefficients for the prediction of octane number in gasoline samples using PLSR.

normalization, the net analyte signal. This vector is defined as the part of the response *Y* that is orthogonal to the response vectors of all other constituents/properties. In the case of unstructured noise, PLSR computes a final regression vector that is not in general purely proportional to the net analyte signal vector but has the important property of being optimal under a mean squared error of prediction criterion [39].

For the gasoline example introduced previously, the regression coefficients for the prediction of octane number are shown in Figure 65.27. It is noted in this plot that three PLS factors were used to generate the model (based on validation diagnostics). It can be seen in Figure 65.27 that the region above 1380 nm does not contribute to modeling octane number. The important regions can be found between 1120–1220 nm (interpreted as aromatics and straight chain hydrocarbon content) and 1350–1380 nm (hydrocarbon content).

65.4.4.6 Predicted versus Reference Plot The predicted versus reference plot should show a straight line relationship between predicted and reference values, ideally with a slope of 1 and a correlation close to 1. The predicted versus reference plot for the gasoline example is shown in Figure 65.28 for a three-PLS factor model.

From a philosophical point of view, in some software packages and in the chemometric literature, this predicted versus reference plot is sometimes called the predicted versus actual plot. The use of the term “actual” gives the indication that the reference value is actually the absolute truth, where in fact it has some form of error associated with it. The term “reference” indicates that the value was generated from a reference method and therefore carries the connotation that it has error associated with it.

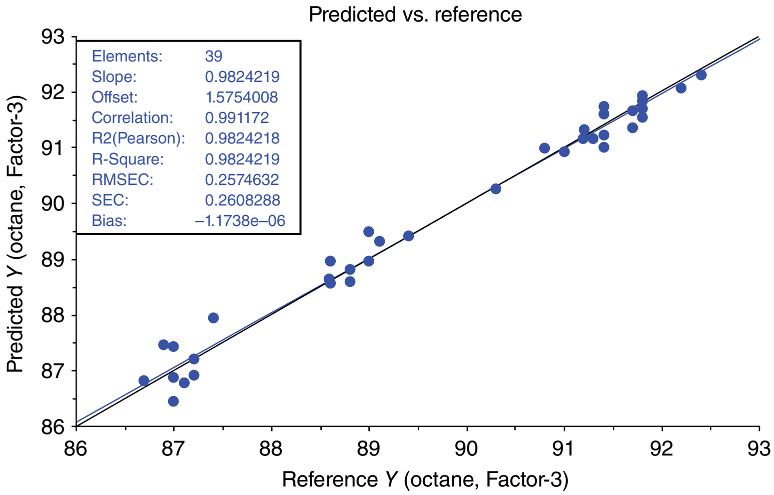


FIGURE 65.28 Predicted versus reference plot for octane number in gasoline analysis.

65.4.4.7 Residuals and Diagnostic Tools in PLSR

X- and Y-Residuals All the bilinear models described in this chapter operate by fitting a portion of the variation in the *X*- and the *Y*-blocks and can be summarized as

$$X = \hat{X} + E_x$$
$$Y = \hat{Y} + E_y$$

where the aim of the modeling phase is normally to find estimates that fit as well as possible the corresponding block matrices. However, inspection of the residual matrices E_X and E_Y can provide useful information about the model quality. In this respect, residuals can be investigated at different levels, and different information can be obtained depending on the level considered. Indeed, residual analysis can be carried out for the detection of outliers, for the identification of systematic variation that was not accounted for by the model (especially when variables are homogeneous, such as in a spectrum or a chromatogram), for the detection of drifts or trends in the data, or in general to define a distance to the model. From a practical standpoint, each of these tasks is better accomplished by adopting a proper representation of the information contained in the residuals. In this framework, the first way of looking at the residuals is to consider the distribution of $e_{x_{i,j}}$ and $e_{y_{i,j}}$, which are the elements of the matrices E_X and E_Y, respectively. Many models assume random Gaussian noise or, in general, symmetrically distributed residuals. Therefore, plotting the residuals or verifying whether the distributional assumptions are met (e.g., by means of normality tests) can provide a good diagnostics of the model. As an example, in Figure 65.29a the distribution of the *X*-residuals for a situation where no anomalies are present is shown: the histogram shows almost perfect symmetry and assumes a Gaussian-like

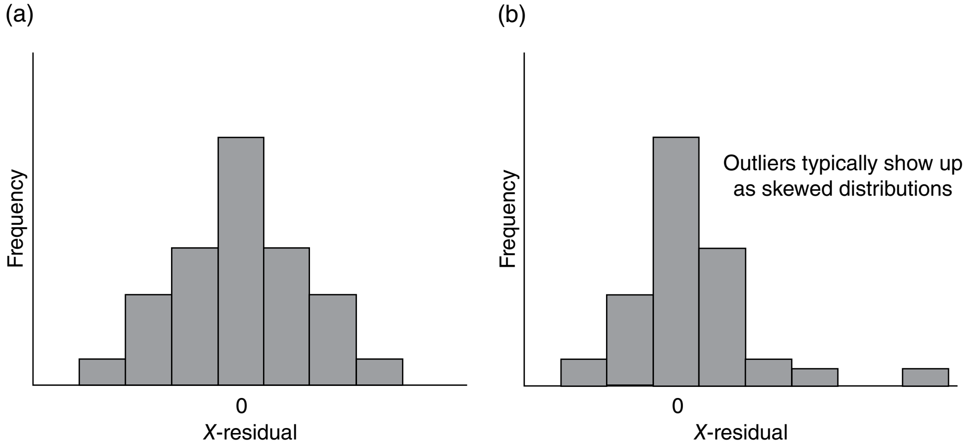


FIGURE 65.29 Distribution of *X*-residuals for a well-behaved model (a) and a model with an outlier (b).

shape, as expected. On the other hand, the distribution of the *X*-residuals for a case where outlying observations are present in the data set is reported in Figure 65.29b. It is evident from this figure that the histogram is no longer symmetric and that there is an increased probability associated to high values of the residuals, indicating that some anomalies are present in the data.

A second set of diagnostic measures can then be inspected when considering that the residual matrices E_X and E_Y have the same dimensions as the fitted matrices $\hat{X}$ and $\hat{Y}$ and therefore it is possible to extract rows and columns to investigate the residual variances associated with one particular sample or variable compared to the rest of the model. It is customary to summarize the variation in one direction or the other by calculating the sum of squares of the vectors corresponding to the individual samples (or variables). In particular, the sum of squares residuals of the i^{th} sample can be expressed, for the *X*- and the *Y*-blocks, as

$$e_{X,i}^2 = \left\| e_{X,i} \right\|^2 = \sum_{j=1}^{v} e_{X,ij}^2$$

$$e_{Y,i}^2 = \left\| e_{Y,i} \right\|^2 = \sum_{k=1}^{r} e_{Y,ik}^2$$

where υ is the number of predictors, r is the number of responses, and e_{X_i} and e_{Y_i} are the ith rows of the matrices E_X and E_Y, respectively, while $e_{x_{i,j}}$ and $e_{y_{i,j}}$ are the corresponding elements. When plotting the values of the sum of squared residuals for the samples, different situations can occur, the most frequent of which are shown in Figure 65.30. In particular, Figure 65.30a shows a random distribution of the summed squared residuals for the different samples, as expected when no outlying observations occur. On the

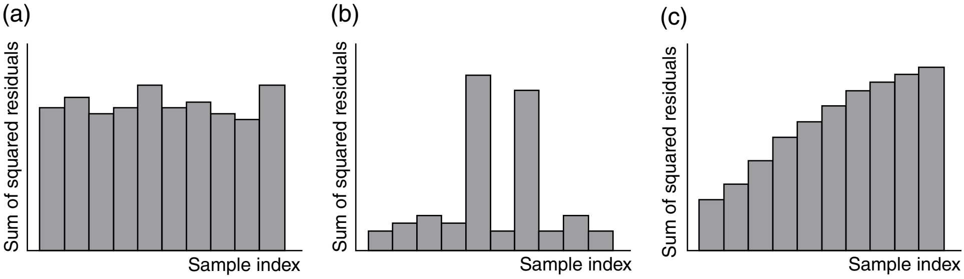


FIGURE 65.30 Example situations of residual patterns in well behaved and models that are not well behaved. (a) Well behaved model, (b) two outliers in the model, and (c) trending in residuals.

other hand, the situation in which two samples are anomalous with respect to the others is plotted in Figure 65.30b; in particular, these two samples are characterized by containing an additional interferent, which is not present in the rest of the objects, and this, in turn, results in a significantly higher value of the residuals. Lastly, the situation where there is a trend along the objects, which is not captured by the model, is depicted in Figure 65.30c.

Besides plotting the sum of squares for each sample, sometimes it can be more useful also to plot, sample-wise, the whole vector of residuals, in order to evidence the presence of unmodeled systematic structure (especially when the variables are homogeneous), or to identify blocking effects. Indeed, when there are sources of systematic variation that are not explained by the model, the residuals for that particular sample are no longer randomly distributed and present a structured shape.

Analogously, the sum of squared residuals for the individual predictors or response variables can be obtained by summing over the analyzed samples, according to

$$e_{X,j}^2 = \left\| e_{X,j} \right\|^2 = \sum_{i=1}^{m} e_{X,ij}^2$$

$$e_{Y,k}^2 = \left\| e_{Y,k} \right\|^2 = \sum_{i=1}^{m} e_{Y,ik}^2$$

where m is the number of samples and $e_{X,j}$ and $e_{Y,k}$, are the jth column of E_X and the kth column of E_Y, respectively, while $e_{X,ij}$ and $e_{X,ik}$ have the same meaning as previously defined.

Error Measures The simplest and most efficient measure of the uncertainty on future predictions is the root mean square error (RMSE). This value (one for each response) is a measure of the average uncertainty that can be expected when predicting Y-values for new samples, expressed in the same units as the Y-variable. The results of future predictions can then be presented as "predicted values ±2×RMSE" (at ~95% confidence).

This measure is valid provided that the new samples are similar to the ones used for the development of the calibration model; otherwise, the prediction error might be much higher. For an MLR calibration model, the RMSE from calibration (i.e., RMSEC) is expressed by

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N - df}}$$

where N is the number of samples and df is the number of variables −1. For test set validation for any regression method, the formula for the root mean square error of prediction (RMSEP) is

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}}$$

For cross validation the formula is as for test set validation but should formally be reported as RMSECV. Validation residual, explained variances and RMSEP are also computed in exactly the same way as calibration variances, except that prediction residuals are used instead of calibration residuals.

Plots of RMSE as a function of number of factors (for latent variable methods) are also used to find the optimum number of model components (similar to the residual variance plots discussed in Section 65.3.3.5). When validation residual variance is minimal, RMSEP is also minimized, and the model with an optimal number of components will have the lowest expected prediction error.

RMSEP can and should be compared with the precision of the reference method (refer to Section 65.2). The error of the reference method is sometimes referred to as the standard error of laboratory (SEL). It is of utmost importance to have an estimate of this precision to evaluate to what extent the model has a sufficiently good predictive ability given the actual application. It cannot be expected that RMSEP be any lower than 1.4 times SEL [40]. The RMSEP is a sum of the sampling error, measurement error, model error, and reference method error.

An alternative error measure is the predicted residual sums of squares (PRESS):

$$\text{PRESS} = \sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

As PRESS is reported as a square number, it does not directly relate to the values (or range) of the response variable Y.

Some other useful statistics available when assessing regression models are described as follows:

1. **Bias**

   The bias is the average value of the difference between the reference and predicted values:

$$\text{BIAS} = \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2}{N}$$

2. **Standard error of prediction (SEP)**

   SEP is the standard deviation of the prediction residuals:

$$\text{SEP} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i - \text{BIAS}\right)^2}{N-1}}$$

3. **Ratio of standard error of prediction to sample standard deviation (RPD)**

   RPD is the ratio of the standard deviation for the response variable *Y* and the RMSE. There exist some ad hoc rules regarding how RPD relates to a "good," "fair," or "bad" model:

$$\text{RPD} = \frac{s_y}{\text{SEP}}$$

   0–2.3 very poor

   2.4–3.0 poor

   3.1–4.9 fair

   5.0–6.4 good

   6.5–8.0 very good

   8.1+ excellent

However, it must be stated that RPD depends on the range used for *Y*. In this respect RMSE is a more generic error measure. Similarly, this is why the correlation coefficient R^2 does not necessarily provide a good indication about a model's predictive ability.

The least squares effect also needs to be taken into consideration when discussing the range of *Y*. For all regression methods with least squares as the objective, overprediction of the low values and underprediction of the high values for *Y* can occur. Thus, if it is expected that many future samples will lie far from the mean of the model, the model can be corrected to give a bias of 0 and slope of 1. This will now not give the least squares solution but avoid over- and underprediction. This principle is illustrated in Figure 65.31.

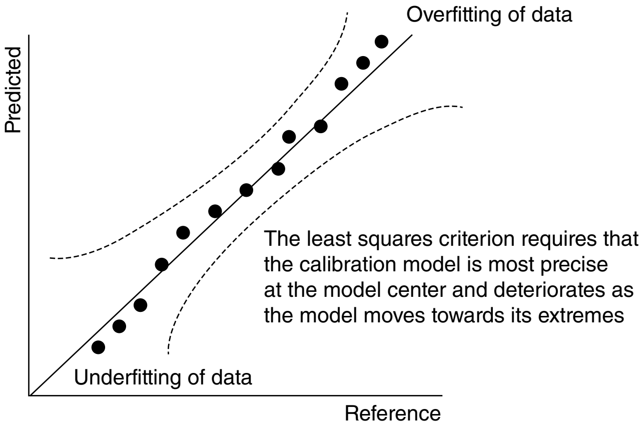

Overfitting of data
Predicted
The least squares criterion requires that the calibration model is most precise at the model center and deteriorates as the model moves towards its extremes
Underfitting of data
Reference

FIGURE 65.31 Error distribution of the least squares fit.

65.4.4.8 Algorithms for PLSR Although there are a number of PLSR alternatives [2, 3, 41], the NIPALS algorithm is still the most commonly used algorithm. The next sections describe the single variable variant (PLS-1) and the multiple response variant (PLS-2) algorithms.

PLS-1 Algorithm

Step 1: Scale and center the X- and Y-variables

$$X_0 = X - 1\bar{x}' \text{ and } Y_0 = Y - 1\bar{y}$$

Step 2: Use the variability remaining in y to find the loading weights w_a where a represents the selected number of PLS factors to calculate,

$$X_{a-1} = y_{a-1}w'_a + E$$

Scale the vector to length 1:

$$\hat{w}_a = cX'_{a-1}y_{a-1}$$

where c is the scaling factor that makes the length of the final $\hat{w}_a$ equal to 1, that is,

$$c = \left(y'_{a-1}X_{a-1}X'_{a-1}y_{a-1}\right)^{-0.5}$$

Step 3: Estimate the scores $\hat{t}_a$ using the local model

$$X_{a-1} = t_a\,\hat{w}'_a + E$$

Since $\hat{w}'_a w_a = 1$ the least squares solution is

$$\hat{t}_a = X_{a-1}\hat{w}_a$$

Step 4: Estimate the *X*-loadings *p* using the local model

$$X_{a-1} = \hat{t}_a p'_a + E$$

which gives the least squares solution

$$\hat{p}_a = \frac{X'_{a-1}\hat{t}_a}{\hat{t}'_a \hat{t}_a}$$

Step 5: Estimate the *Y*-loadings *q*, using the local model

$$y_{a-1} = \hat{t}_a q_a + F$$

where *F* is the residual term after fitting the model to *Y*.

This gives the solution

$$\hat{q}_a = \frac{y'_{a-1}\hat{t}_a}{\hat{t}'_a \hat{t}_a}$$

Step 6: Create new *X* and *y* residuals by subtracting the new PLS factor as follows,

$$\hat{E} = X_{a-1} - \hat{t}_a \hat{p}'_a$$

$$\hat{F} = y_{a-1} - \hat{t}_a \hat{q}_a$$

Replace the former X_{a-1} and y_{a-1} by the new residuals $\hat{E}$ and $\hat{F}$ and increase *a* by 1.

Step 7: Determine *A* the number of valid PLS factors to retain in the calibration model.

Step 8: Compute $\hat{b}_0$ and $\hat{b}$ for *A* PLS factors to be used in the prediction model,

$$\hat{b} = \hat{W}\left(\hat{P}'\hat{W}\right)^{-1}\hat{q}$$

$$b_0 = \bar{y} - \bar{x}'\hat{b}$$

PLS-2 Algorithm The PLS-2 algorithm is almost identical to the PLS-1 algorithm, except the *Y* vector is replaced by a matrix $Y_{i,j}$ where the dimension *j* represents the number of *y*-variables to be modeled by the algorithm. In the early days of chemometrics, there was a distinction between PLS-1 and PLS-2 because the PLS-2 is an iterative algorithm and this increased the computational time. Today there is really no need to make this distinction apart from a conceptual point of view.

Step 1: Define a temporary *y*-score $\hat{u}_a$, for example, the column in *Y* with the largest variance.

Step 2: Use $\hat{u}_a$ that summarizes the remaining variability in *Y* to find the loading weights $\hat{w}_a$ by least squares fitting using the local model

$$X_{a-1} = \hat{u}_a w'_a + E$$

Scale the vector to length 1. The least squares solution is

$$\hat{w}_a = cX'_{a-1}\hat{u}_a$$

where *c* is the scaling factor that makes the length of the final $\hat{w}_a$ equal to 1, that is,

$$c = \left(\hat{u}'_a X_{a-1} X'_{a-1} \hat{u}_a\right)^{-0.5}$$

In the first iteration $\hat{u}_a$ is given a starting value, typically the column in *Y* with the largest sum of squares.

Step 3: Estimate the scores $\hat{t}_a$ using the local model

$$X_{a-1} = t_a \hat{w}'_a + E$$

Since $\hat{w}'_a w_a = 1$ the least squares solution is

$$\hat{t}_a = X_{a-1}\hat{w}_a$$

Step 4: Estimate the *X*-loadings *p* using the local model

$$X_{a-1} = \hat{t}_a p'_a + E$$

which gives the least squares solution

$$\hat{p}_a = \frac{X'_{a-1}\hat{t}_a}{\hat{t}'_a \hat{t}_a}$$

Step 5: Estimate the *Y*-loadings *q*, using the local model

$$\hat{u}_a = \hat{t}_a q_a + F$$

where *F* is the residual term after fitting the model to *Y*.

This gives the solution

$$\hat{q}_a = \frac{\hat{u}_a \hat{t}_a}{\hat{t}'_a \hat{t}_a}$$

Step 6: Test for convergence to see that all elements no longer change significantly from the last iteration.

Step 7: If convergence has not been achieved, estimate temporary factor scores u_a using the model,

$$Y_{a-1} = u_a \hat{q}'_a + F$$

Giving the least squares solution

$$\hat{u}_a = Y_{a-1}\hat{q}_a \left(\hat{q}'_a \hat{q}_a\right)^{-1}$$

Return to step 2.

Step 8: If convergence has been reached, create new X and y residuals by subtracting the new PLS factor as follows,

$$\hat{E} = X_{a-1} - \hat{t}_a \hat{p}'_a$$
$$\hat{F} = y_{a-1} - \hat{t}_a \hat{q}_a$$

Replace the former X_{a-1} and y_{a-1} by the new residuals $\hat{E}$ and $\hat{F}$ and increase a by 1.

Step 9: Determine A the number of valid PLS factors to retain in the calibration model.

Step 10: Compute $\hat{B}$ and b'_0 for A PLS factors to be used in the prediction model,

$$\hat{b} = \hat{W}\left(\hat{P}'\hat{W}\right)^{-1}\hat{Q}'$$
$$b'_0 = \bar{y}' - \bar{x}'\hat{B}$$

For more details on the algorithms for PLS-1 and PLS-2, the interested reader is referred to the literature [15, 18].

65.5 MULTIVARIATE CLASSIFICATION

In Section 65.3 the topic of EDA was discussed in great detail. The methods of cluster analysis discussed were defined as unsupervised methods, that is, they looked for the natural patterns in the data without being guided by an external classification rule.

The method of PCA discussed in Section 65.3.3 provides a highly graphical environment for detecting clusters within data set of samples. By making a separate class model for each cluster, a library of clusters can be developed and thus a classification rule can be established in order to group new samples into known classes. This is known as supervised classification, or pattern recognition, and this topic is an important area known generally as multivariate classification.

Multivariate classification methods take full advantage of the multivariate nature of the data. Although there are many multivariate classification methods available, this section will only look at four commonly used methods:

1. Linear discriminant analysis (LDA)
2. Soft independent modeling of class analogy (SIMCA)
3. Partial least squares discriminant analysis (PLS-DA)
4. Support vector machine classification (SVMC)

The aforementioned classification methods will be discussed in order and comparisons will be made to their regression counterparts already discussed in the previous sections.

65.5.1 Linear Discriminant Analysis

By definition, discriminant analysis aims to find discriminating features that separate samples into different data classes. In the case of the classical LDA, also known as Fisher's LDA [42], the algorithm aims to find the discriminating axes, that is, linear combinations of the original p variables that optimally separate two or more classes.

In order to separate samples into classes, a training set of two or more known samples must be available to develop the classification rule. There are a number of distance measures that can be used to optimally separate classes including:

1. Linear separators
2. Quadratic separators
3. Mahalanobis distance

Figure 65.32 shows diagrammatically how the linear and quadratic methods separate samples into classes.

When separating the two-class problem, the axes used to discriminate between the classes may be viewed as a projection onto $A = 1$ dimensions and the discrimination axis could be viewed as a component vector separating the two classes (refer to Fig. 65.32).

The regression equivalent of LDA is MLR (Section 65.4.2). LDA suffers from the same collinearity effects that MLR does and also requires more samples than variables being measured. To overcome these issues, a version of LDA, known as PCA–LDA, is

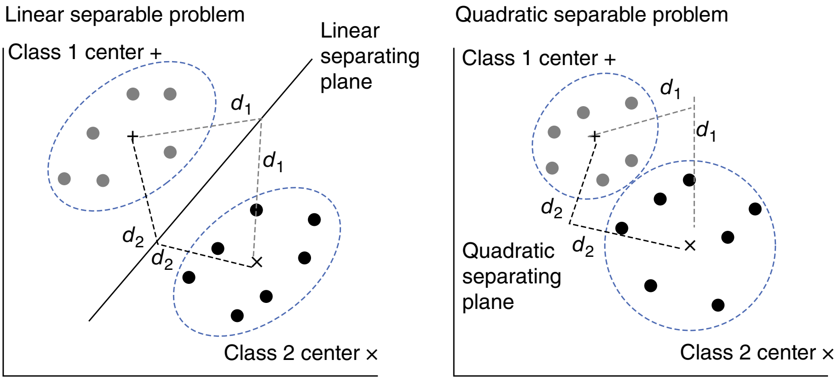


FIGURE 65.32 Linear and quadratic separation in LDA.

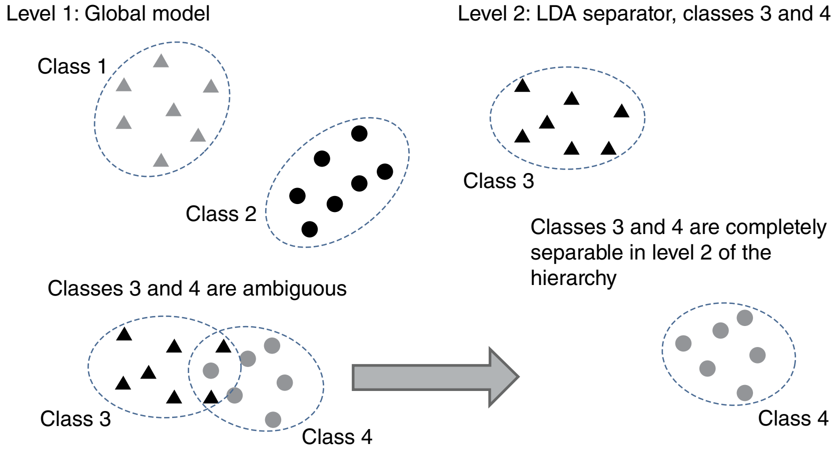

Level 1: Global model
Class 1
Class 2
Classes 3 and 4 are ambiguous
Class 3
Class 4
Level 2: LDA separator, classes 3 and 4
Class 3
Classes 3 and 4 are completely separable in level 2 of the hierarchy
Class 4

FIGURE 65.33 Resolving classification ambiguities using hierarchical models.

available to overcome collinearity issues when higher-dimensional data (particularly spectral data) are being analyzed.

Two other major limitations of LDA are that the assumption of a common covariance structure in the classes exists, which is very rare in practice, and LDA is usually only suited to the two-class separation problem. The latter point is further exacerbated by the limitation that if LDA does not uniquely classify a sample into a unique class, it puts the sample into the class that has the closest distance to the class center. This limitation can have serious ramifications, particularly in applications such as pharmaceutical raw material identification or antiterrorist hazardous material applications.

For these limitations alone, LDA is best suited when used in hierarchical classification schemes. In many instances, when a classification model contains many classes, ambiguities may arise due to limitations in the measurement system(s) used to separate the classes. Figure 65.33 shows graphically the concept of ambiguity and how a hierarchical model can be used to resolve the two-class problem.

The main model is commonly known as the global model. Here, a new sample is assessed by each class library in the model and the highest match is found based on some predefined statistical limits. If a sample is classified into two or more classes, the sample is said to be ambiguous with the global model, that is, a unique separation is not possible. To overcome this problem for the two-class ambiguity, LDA is ideal for the following reasons:

1. It is ideal for separating the two-class problem based on its simplicity.
2. Since the hierarchical model has already determined that the sample is either one of two classes, the use of LDA is more reliable since it will classify the sample into one of the two classes without the risk of the sample being mistaken for something not in the global model.

Due to the reasons provided earlier, a classification method that can detect a null class, that is, detect that a new sample is a complete unknown, is much more useful than LDA. This method known as SIMCA is discussed in the next section.

65.5.2 Soft Independent Modeling of Class Analogy

PCA allows a user to develop class models based on their clustering in PC scores space. Once this library of PCA models is developed, a rule must be established that can direct new samples to the class(es) they belong to. The method of SIMCA provides such a rule.

SIMCA was first introduced by Wold in 1976 [43] and it allows new samples to display their uniqueness as well as their common patterns, which provides the advantage of SIMCA over LDA to be able to reject samples as not belonging to any class, rather than put the sample into the closest class.

65.5.2.1 Practical Steps for Building a SIMCA Model As with any empirical modeling strategy, a training phase is required. The term soft in SIMCA relates to the model being empirical and therefore a representative set of samples must be modeled to understand the natural class variability. It is important to note here that since PCA (or PLSR) is used to develop class models in SIMCA, if all samples in the set have identical variability, then no PCA/PLSR model is possible. This is one of the potential drawbacks of the SIMCA methodology.

By representative (refer to Section 65.2) each class must consist of samples that will represent future sample variability. This variability is considered to be the natural variability to be expected for this sample type. Once the data are collected, a PCA/PLSR model is calculated for all classes to understand the variability within and between classes. This represents the ANOVA problem defined in Section 65.2, and in general, PCA is a multivariate version of ANOVA and can be described as a visual ANOVA. Figure 65.34 shows a PCA scores plot with a number of different classes identified.

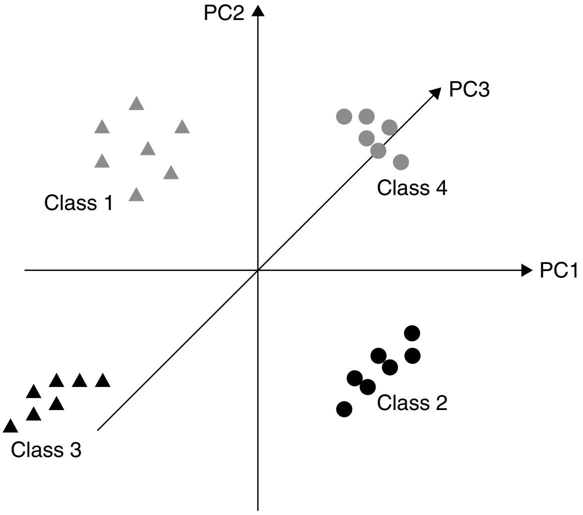


FIGURE 65.34 PCA scores plot showing class separation of data.

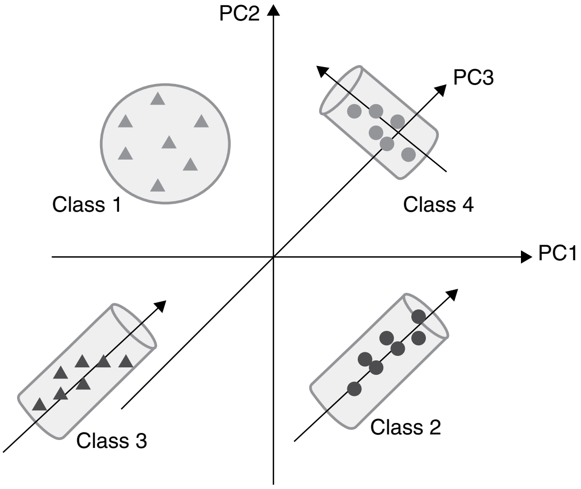


FIGURE 65.35 PCA scores plot showing class separation of data with class limits.

To place objective limits around each class, the diagnostic tools associated with PCA/PLSR are used to determine confidence intervals. In particular, the statistics, sections "*X*-Residuals" and "Leverage" are used. Figure 65.35 shows the data in Figure 65.34, this time with statistical limits around the classes.

To develop a SIMCA training model, each class must be saved as an individual PCA model and validated, preferably with a test set (see Section 65.6.1). The development of individual PCA models results in what is known as a disjoint class model, since the individual libraries retain their uniqueness and must be joined using an external rule.

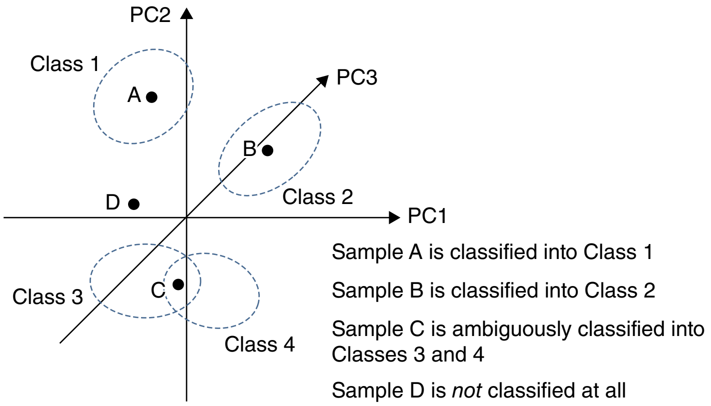

PC2
Class 1
A
PC3
B
Class 2
D
PC1
Class 3
C
Class 4
Sample A is classified into Class 1
Sample B is classified into Class 2
Sample C is ambiguously classified into Classes 3 and 4
Sample D is *not* classified at all

FIGURE 65.36 Three cases of classification outcome when using SIMCA.

It is important to note here that even though SIMCA uses individual PCA models as class rules, there must be complete commonality of the PCA models to each other in terms of number of variables and data preprocessing, that is, each PCA library model must have identical variables and must be preprocessed exactly the same way. This is to ensure complete statistical representation of the results. It is also a potential drawback to the SIMCA method; however, these issues can be overcome in the case of ambiguities by using a hierarchical model. Also, the trick of weighting down variables for different ranges of the total number of variables in the individual models may be used.

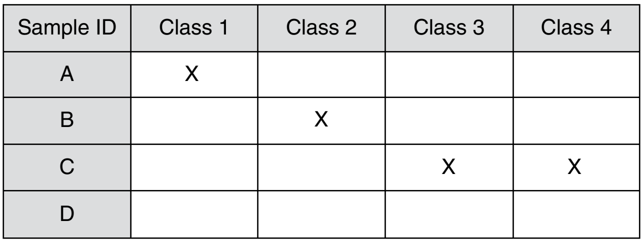
The next step in SIMCA library development is to select the validated PCA class models and enter them into a scheme that allows new samples to be assessed by each model for class assignment. In SIMCA, there are three possible outcomes:

1. Unique classification: The new sample resides in one class only.
2. Ambiguous classification: The sample could be a member of two or more classes simultaneously.
3. No classification: The new sample is not a member of any of the class models present in the SIMCA scheme.

It is point three that provides SIMCA with the most versatility compared to most other classification methods. Figure 65.36 provides a diagram of all three cases described earlier.

When validating a SIMCA model for practical usage, there are two approaches that should be used:

1. Use the model to predict the training set: Although this is highly biased, it should be the first step to provide assurance that the library is capable of predicting itself. During this stage, the confidence intervals for the model can be fine-tuned.



Sample ID	Class 1	Class 2	Class 3	Class 4
A	X			
B		X		
C			X	X
D				

FIGURE 65.37 Classification table showing all three classification scenarios for the data shown in Figure 65.34.

2. Use the model to classify a separate test set of samples to understand how the library will perform on new samples in the future.

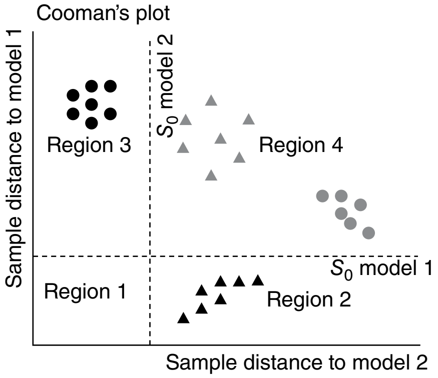
The next sections describe the diagnostic and graphical tools available to assess the performance of SIMCA models.

65.5.2.2 Diagnostic Tools for SIMCA The SIMCA approach generates a multitude of statistics relevant for validating the performance of the classification model. Effectively, the class models are wrapped into an envelope that defines the class membership boundaries. The major tools and statistics to be described for SIMCA are listed as follows:

1. The classification table
2. The Coomans plot
3. The distance versus leverage plot
4. Model distance
5. Variable discrimination power
6. Modeling power

The Classification Table To provide an overview of the classification results, the classification table is the first diagnostic tool to check. It is a standard table with sample ID's listed as the rows and class model names as the columns. Each time a sample is classified into a class, it is marked in the table by some form of symbol. If the sample simultaneously belongs to two or more classes, there will be a symbol shown for each class the sample belongs to. In the event of no classification, the table entry for the particular sample will remain blank. Figure 65.37 shows an example classification table for the example shown in Figure 65.36, that is, with the three possible classification results shown.

Coomans Plot Named after the Belgian chemometrician Prof. Danny Coomans [44], this plot shows the sample to model distances for the training and validation sets two



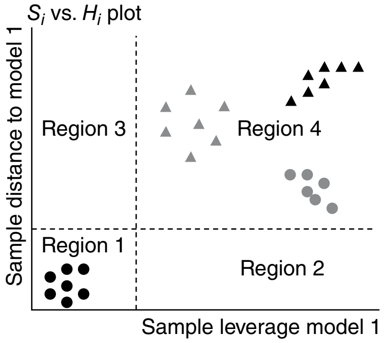
Region 1: Samples that lie in this space cannot be uniquely classified by models 1 and 2.

Region 2: These samples are uniquely classified by model 2.

Region 3: These samples are uniquely classified by model 1.

Region 4: These samples are not classified by either model 1 or model 2.

FIGURE 65.38 The Coomans plot and its interpretation.



Region 1: Samples that lie in this space are classified by model 1. It is expected that only samples from class 1 will occupy this space

Region 2: Samples that lie in this space are extreme with respect to model 1

Region 3: Samples that lie in this space are not modeled well by model 1

Region 4: Samples that lie in this space are not classified by model 1

FIGURE 65.39 Construction of the S_i versus H_i plot and how it is interpreted.

models at a time, that is, it shows the orthogonal distances between models. Two sets of limits (S_0) are shown on the horizontal and vertical scales of the plot showing where the statistical limits are located for the two models under investigation.

Since the Coomans plot is a pairwise model comparison tool, it is used to assess the degree of model overlap in the case of ambiguity. To interpret the plot, if any samples lie in the boundary between the origin and where the two limits cross, then the individual library models are not capable of uniquely separating the classes. If the samples from both models lie orthogonally separate in their own classes, then the models are capable of uniquely separating the two classes under investigation. Figure 65.38 provides an overview of the Coomans plot and its interpretation.

The Distance versus Leverage (S_i vs. H_i Plot) This plot is also called the membership plot [18] because it shows the statistical limits used to envelop the class model. The statistic S_i is similar to X-residual discussed in section "X-Residuals" and is the residual standard deviation of the sample to the model. H_i is the leverage and shows the distance to model center for ach samples. Figure 65.39 shows diagrammatically how S_i and H_i are represented in scores space and how this translates to the S_i versus H_i plot.

Interpretation of the S_i versus H_i plot is simple. It compares a selected class model to all other class models in the library. If the model is able to uniquely classify samples of its own class, the samples will all lie between the S_i and H_i limits for the class at a specified confidence interval. The leverage limit is determined by the number of PCs used in the model. This plot is very similar to the influence plot discussed in section "Influence Plots." There is from the conceptual point of view nothing wrong by applying the Hotelling's T^2 statistics as the limit of distance within the model. As the number of samples is often limited for the individual classes, the leverage-based rule of 3× the average leverage was historically assumed to be better than using a distribution-based metric (Hotelling's T^2 is based on an F-distribution).

Samples close to the origin of the S_i versus H_i plot are considered to be real members of that class. As the sample moves toward the leverage limit, the sample is considered to be extreme in characteristics for that class. If it exceeds the boundary but is within the S_i boundary of the model, then the sample is too extreme to be part of the model. An analogous situation is described as follows. Consider the origin representing an espresso coffee and as the leverage moves toward the limits, the coffee is becoming more latte in characteristics, then the sample exceeds the limit, and it is more latte in character than espresso. In general, leverage samples represent the same class but only extremes within that class.

Continuing with the same analogy, as a sample moves along the S_i axis, these samples are not as well modeled as those closer to the origin. Therefore, the espresso changes from coffee to being more tealike in character. If the sample lies outside of the S_i axis, it is now characteristic of tea rather than coffee. A sample that exceeds both boundaries simultaneously is not a member of that class and in terms of the analogy is now a milky tea rather than an espresso.

Another variant of the S_i versus H_i plot is the S_i/S_0 versus H_i plot. The main difference is that the value S_i is now scaled to the average distance to the model S_0. This plot is interpreted in the same way as the S_i versus H_i plot.

Overall, the S_i versus H_i plot is a one-stop diagnostic for determining the ability of each model to uniquely classify samples. The diagnostics presented in the next section are used to support the findings made in the classification table, the Coomans plot, and the S_i versus H_i plot.

Model Distance Model distance is usually plotted as a bar chart and shows the relative distances of all models in the library to each other with the proviso that a model's distance to itself is 1. In general, a large intermodal distance indicates clearly separated models. Model distance is calculated as the pooled residual standard deviations by fitting samples from two different classes to their own models and every other model in the library. Figure 65.40 provides an example of a model distance plot.

A general rule of thumb when interpreting a model distance plot is a model distance greater than three indicates models that are significantly different, although there are

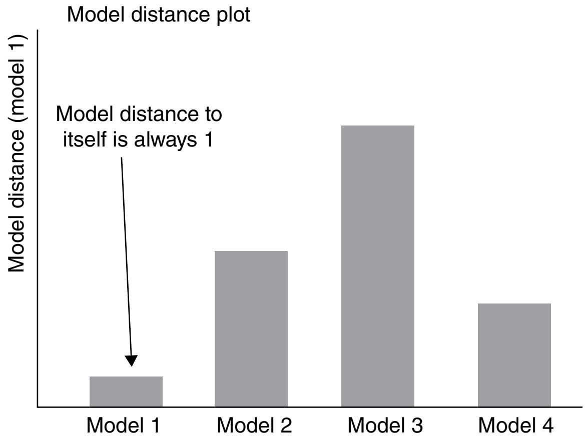


FIGURE 65.40 Example model distance plot.

exceptions to this rule. In most cases, when the model distance is less than three, there is no discrimination power of the models for the classes under investigation.

Variable Discrimination Power The discrimination power of a variable provides information about its ability to discriminate between any two-class models. It is calculated by fitting the samples from one model to all other models in the library and to its own class model.

As it is a pairwise model comparison tool, eliminating variables from one model pair results in removing these variables from all model classes, as based on the rules of SIMCA defined in Section 65.5.2.1.

Therefore, if two models show ambiguities and the removal of specific variable improves the results for the two classes but is detrimental to all other classes, this is where the hierarchical approach becomes of high value. The ideal situation is that by removing noncontributing variables from one model will result in an improvement in all models. Figure 65.41 provides an example of a variable discrimination power plot.

Modeling Power Modeling power quantifies the importance of a particular variable for modeling a particular class. It is a measure of the variables variance that is used to describe the class model. With any variable elimination step in SIMCA, it is important to assess the impact on other models in the library.

Variables with large modeling power have a large influence on the model. As a general rule of thumb, if the modeling power is less than 0.3, this variable could be detrimental to the model. This statistic should also be used in combination with the variable discrimination power to make sure that the discriminating power of the variable is also low before it is considered a candidate for elimination. Figure 65.42 provides an example of modeling power plot.

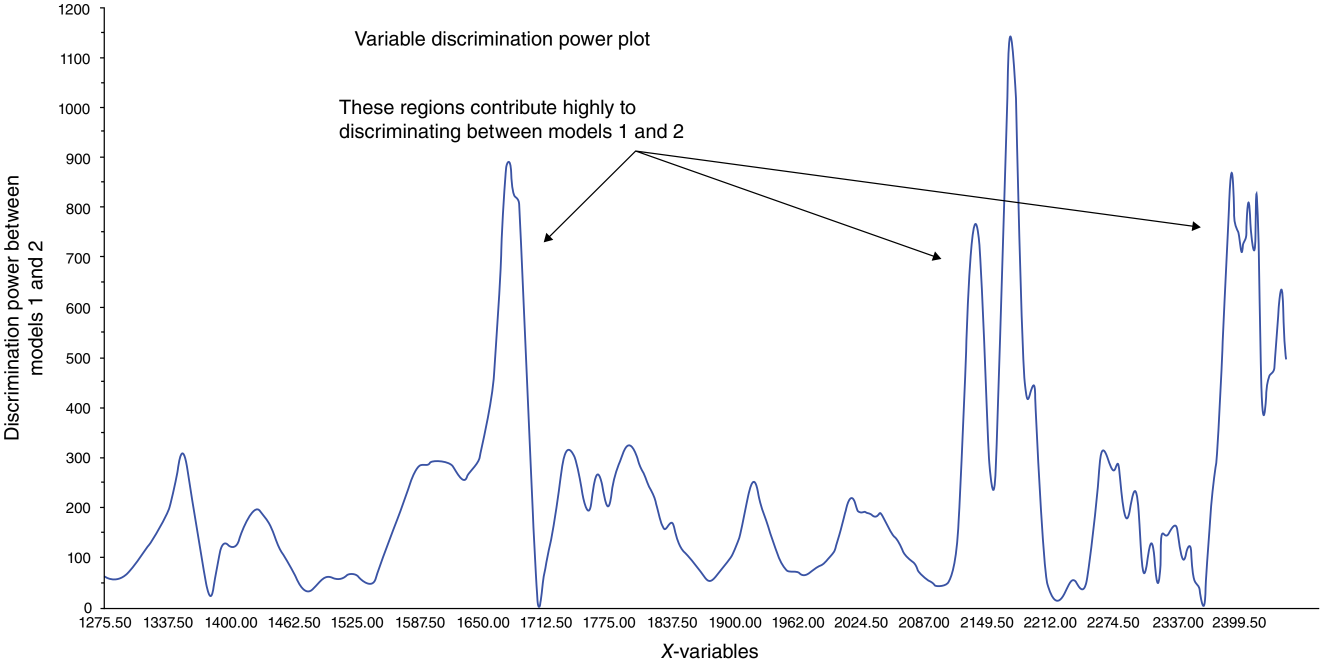


FIGURE 65.41 Example variable discrimination plot.

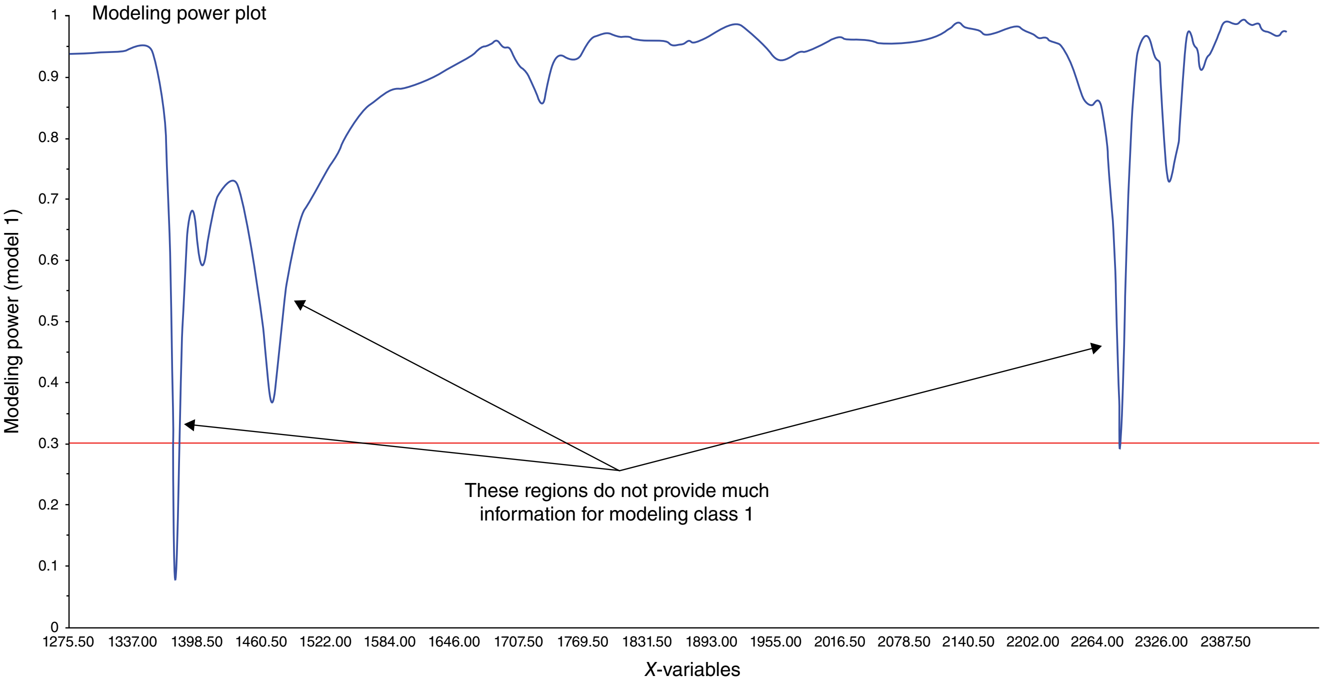


FIGURE 65.42 Example modeling power plot.

65.5.2.3 Summary of SIMCA The discussion presented on SIMCA has shown it to be one of the most versatile and powerful multivariate classification models available. It can not only utilize PCA models, but PLSR models can also be utilized for their scores structure.

The method is based on disjoint modeling, meaning that SIMCA uses individual models to form a library where each individual model defines a classification rule and the SIMCA architecture defines how to apply the library models to each sample. Because of the highly graphical and diagnostic rich nature of PCA and PLSR models, there are a multitude of diagnostic tools available for training and applying SIMCA models to new samples. SIMCA also has the major advantages that it can detect ambiguities and nonclassification of samples in the model.

Some drawbacks to the method include:

1. All PCA/PLSR class models must have the same number of variables and the same preprocessing applied to them.
2. If an unimportant variable is removed from one model, it has to be removed from all library models. This could be beneficial to one model but detrimental to the classification ability of other models in the library.

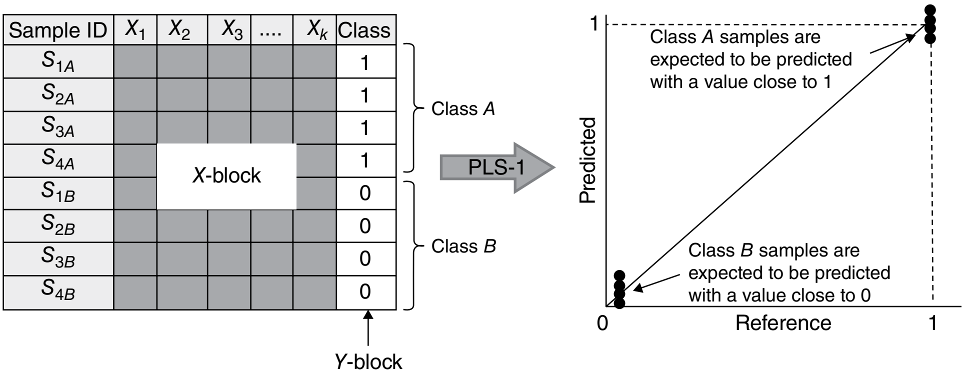
These disadvantages are only minor though compared to classical LDA, and hierarchical modeling approaches can be used to avoid these issues. As the basis for the classification is individual PCA models, there is no direct objective to separate the classes. Thus, the assumption is that the main components from PCA also are suited for classification. This is also related to the discrimination and modeling power as described previously.

65.5.3 Partial Least Squares Discriminant Analysis

PLS-DA [45] is a method that utilizes the PLSR algorithm for classifying samples into distinct classes by using a binary class separator. In the simplest case, a binary class variable where class A is designated 1 and class B is designated 0 is defined. The PLSR algorithm is applied to determine if a model can be generated that predicts close to 1 for class A and close to 0 for class B. This is shown diagrammatically in Figure 65.43.

When there are more than two classes to be separated, the form of the PLSR model becomes the complete PLS-2 algorithm and the Y-variable structure must also be changed. This is shown in Figure 65.44.

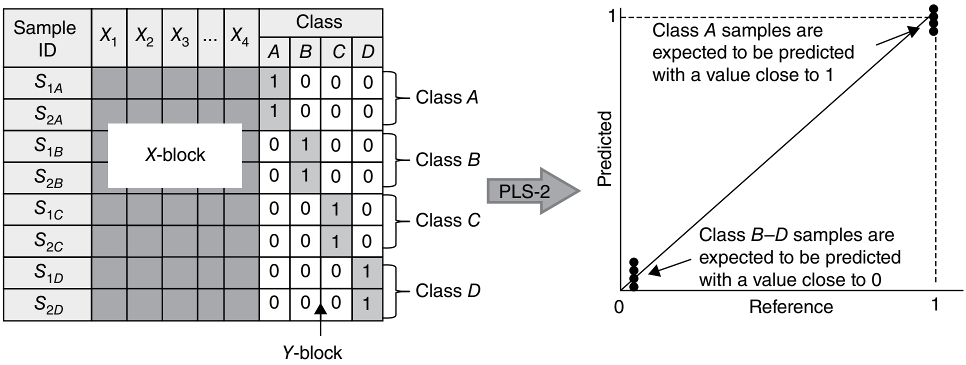
The structure of the Y-variable is now arranged by adding a new column for each class. For the class column, all samples of that class are given a 1 designation while all other classes are given a 0 designation. The PLS-2 algorithm then makes a model for each class setting the class to be modeled as 1 and all other class predictions to 0. In this way, each PLS model is similar to a PCA library in SIMCA, only it is held together in the PLS-2 architecture.



Sample ID	X_1	X_2	X_3	....	X_k	Class
S_{1A}						1
S_{2A}						1
S_{3A}						1
S_{4A}						1
S_{1B}						0
S_{2B}						0
S_{3B}						0
S_{4B}						0

X-block
Class A
Class B
Y-block
PLS-1
1
Class A samples are expected to be predicted with a value close to 1
Predicted
Class B samples are expected to be predicted with a value close to 0
0
Reference
1

FIGURE 65.43 PLS-DA for the two-class discrimination problem.



Sample ID	X_1	X_2	X_3	...	X_4	Class			
						A	B	C	D
S_{1A}						1	0	0	0
S_{2A}						1	0	0	0
S_{1B}						0	1	0	0
S_{2B}						0	1	0	0
S_{1C}						0	0	1	0
S_{2C}						0	0	1	0
S_{1D}						0	0	0	1
S_{2D}						0	0	0	1

X-block
Class A
Class B
Class C
Class D
Y-block
PLS-2
1
Class A samples are expected to be predicted with a value close to 1
Predicted
Class B–D samples are expected to be predicted with a value close to 0
0
Reference
1

FIGURE 65.44 PLS-DA for the three (and higher) class discrimination problem.

The main advantage of PLS-DA over SIMCA is that the underlying factors as represented by the scores and loading weights/loadings are guided by the *Y*-variables. This can add more specificity to the training stage. It also allows the full usage of the diagnostic tools available in PLSR. Overall, there are similar advantages and disadvantages as to whether to use PLS-DA over SIMCA; however, like all classification methods, they should be viewed as complementary and again, in a hierarchical modeling sense, the combination of both approaches to solve specific ambiguity situations should be considered.

When interpreting the results of a PLS-DA, the predicted versus reference plot (Section 65.4.4.6) is one of the primary tools for assessing model quality. Based on the principles of least squares modeling, the distribution of results around the regression line is assumed to be normal. This means that for accurate estimations of class membership, a suitable number of samples should be used in the training model to provide an estimate of precision around the predicted values 0 and 1. The predicted results should ideally be distributed as a *t*-distribution depending on the number of samples used to train each class. From there, statistical limits can be put around 0 and 1 so as to

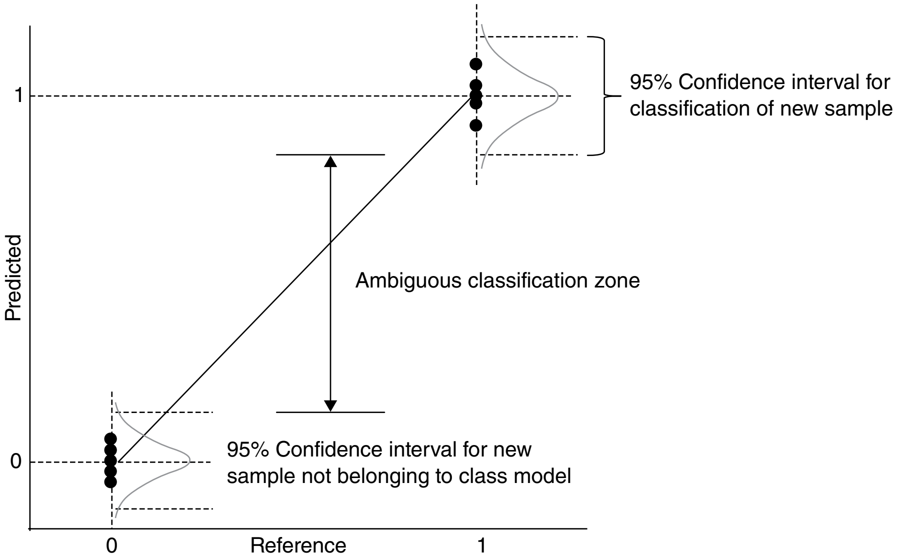


FIGURE 65.45 Predicted versus reference plot for PLS-DA with class membership limits.

determine whether a sample is part of the class or not. This is shown diagrammatically in Figure 65.45.

Overall, PLS-DA is a useful tool for solving multivariate classification problems. It provides the powerful graphical and diagnostic tools associated with the PLSR method and can be used for classification problems where more specificity is required for relating the *X*-variables to the *Y*-variables.

65.5.4 Support Vector Machine Classification

SVMC is a pattern recognition method that is used widely in data mining applications and provides a means of supervised classification, as do SIMCA and other linear discriminators. SVM was originally developed for the linear classification of separable data but is applicable to nonlinear data with the use of kernel functions. SVMs are used in machine learning, optimization, statistics, bioinformatics, and other fields that use pattern recognition. In this section a minimum of mathematics will be used to define the SVMC class separation problem. The mathematical principles behind SVM are outside of the scope of this book, and the interested reader is referred to the literature for more information [46].

65.5.4.1 The Idea behind SVMC SVM is a classification method based on statistical learning that fits a hyperplane to a multidimensional data set for optimal separation of classes. As linear functions are not always able to model complex separation problems, in SVM, data are mapped into a new feature space and a dual

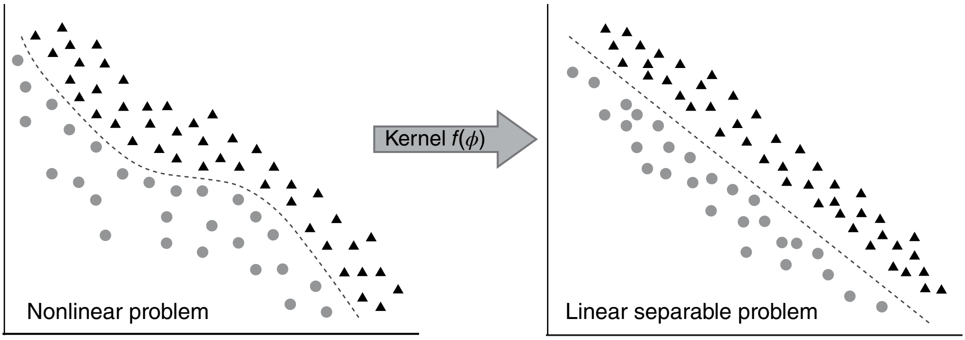
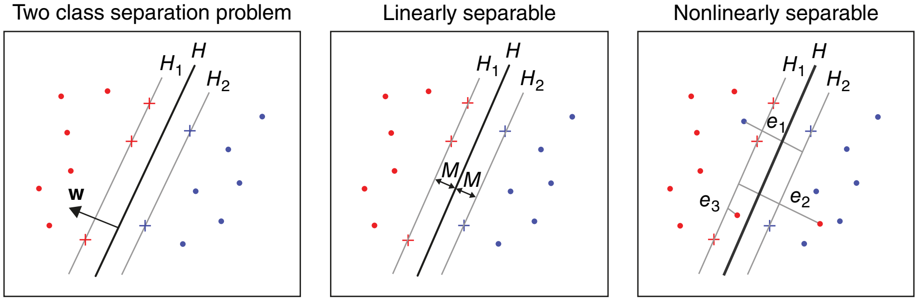


FIGURE 65.46 Using a kernel to map a high-dimensional space to a simpler feature space.



Two class separation problem

Example: Two classes {–1, 1}
+/+ denote support vectors
Hyperplanes H_1 and H_2 define class borders
Decision boundary H: $x^T w + b = 0$

Linearly separable

Classification given by $sign[x^T w + b]$
Many choices of H would give perfect classification
Define $M = 1/\|w\|$
Choose a H that maximizes M (minimizes $\|w\|$)

Nonlinearly separable

Define slack variables (penalties) $e_i \geq 0$
Each e_i given by distance to sample on wrong side of class border
Choose a H that minimizes $\sum e_i$

FIGURE 65.47 Samples that define the support vectors.

representation is used with the samples represented by their dot product. A kernel function is used to map from the original space to the feature space and can be of many forms, thus providing the ability to handle nonlinear classification cases. The kernels can be viewed as a mapping of nonlinear data to a higher-dimensional feature space while providing a computation shortcut by allowing linear algorithms to work with higher-dimensional feature space. The support vector is defined as the reduced training data from the kernel. Figure 65.46 illustrates the principle of applying a kernel function to achieve class separability.

In this new space SVM will search for the samples that lie on the borderline between the classes, that is, to find the samples that are ideal for separating the classes; these samples are named support vectors. Figure 65.47 shows this principle where samples marked with + for the two classes are used to generate the rule for classifying new samples.

A situation where SVM performs well is when some classes are inhomogeneous and partly overlapping, that is, where classical methods, such as SIMCA, will result in ambiguities and thus not be effective as a classification rule. SVM will in this case find a set of the most relevant samples in terms of discriminating between the classes and is invariant to samples far from the discrimination line.

SVM has advantages over classification methods such as neural networks, as its outputs are more transparent, and has less tendency of overfitting when compared to other nonlinear classification methodologies. For finding an optimal model, a grid search is often applied to investigate various combinations of the parameters governed by the type of SVM and kernel. In all cases, model validation is the critical aspect in avoiding overfitting for any method. Not only are SVMs effective for the modeling of nonlinear data, but they are also relatively insensitive to variation in these model parameters. SVM uses an iterative training algorithm to achieve separation of different classes. In the case of similar classification rate for the training data, the model with the most linear parameters and minimum number of support vectors should be chosen.

There are two general SVM classification types that are based on different means of minimizing the error function of the classification:

1. c-SVC: also known as classification SVM type 1
2. nu-SVC: also known as classification SVM type 2

In the c-SVM classification, a capacity factor, *C*, can be defined. The value of *C* should be chosen based on knowledge of the noise in the data being modeled. Its value can be optimized through cross validation procedures. When using nu-SVM classification, the nu value must be defined. Nu serves as the upper bound of the fraction of errors and is the lower bound for the fraction of support vectors. Increasing nu will allow more errors while increasing the margin of class separation.

In practice, there are four main kernel types that can be used to separate classes; these are:

1. Linear
2. Polynomial
3. Radial basis function
4. Sigmoid

The detailed explanation of kernel functions and feature spaces is outside of the scope of this chapter, and the interested reader is referred to the article by Luts et al. [47] for an excellent discussion of the various kernel types and their application in chemistry.

65.6 TECHNIQUES FOR VALIDATING CHEMOMETRIC MODELS

Probably the single most important issue of chemometric model development is model validation [48]. The objective of validation is to evaluate the performance of a multivariate model, be this related to modeling and interpretation, discrimination, or prediction. Whether exploratory or regression methods are being developed, validation is concerned about prediction performance.

As validation has been the topic in many publications, this section is more of a discussion of the principle rather than presenting specific applications.

A distinction between data-driven (internal) and hypothesis-driven (external) validation may be drawn. The latter is focused on confirming the known structure in a system under observation such as to find the true signals of, for example, chemical compounds. Another aspect to consider, when building empirical models based on multi-channel spectroscopic data, is to observed whether the model highlights any known chemical groups based on a-priori assumptions of chemical absorbances in specific regions of the spectrum. This may also be confirmed using existing chemical group information in the literature. This is again related to interpretation, and a good rule is "no prediction without interpretation, no interpretation without evaluating prediction ability." The following will focus on data-driven validation.

Validation of multivariate (or any model in general) is essential in order to make sure that the model will work in the future for new, similar data sets, and indeed do this in either a qualitative or a quantitative way. In regression models, this can be viewed as prediction error estimation (section "Error Measures"), while in classification models, this can be viewed as misclassification rates. Validation is often also used in order to find the optimal dimensionality of a multivariate model, that is, to avoid either overfitting or underfitting or incorrect interpretation [49].

65.6.1 Test Set Validation

When the objective is to establish a calibration model for predicting quantities such as concentration or determining classification models selectivity and specificity, the most conservative validation is to test the model on a representative, independent test set of sufficient size. This has been discussed in length by Esbensen and Geladi [50].

It may be debated what is meant by a test set given a specific situation. Question such as, can the set be used to test extrapolation of the calibration set, or will changes in sample matrix, with respect to the calibration set, invalidate the model? These sources of variation that are, in principle, unknown for future samples, may be quantified by several approaches.

When a test set is used, the model error will be expressed as the RMSEP (section "Error Measures"). This is the most reliable measure of the models future performance on new samples, provided the sample is similar to the calibration set samples. The diagnostic tools discussed in Section 65.4.4.7 can also be used during prediction to assess the quality of the *X*-data such that the predicted value(s) can be assured of being reliable.

In order to create a representative test set for model evaluation, there are a number of general approaches and two of these approaches will be discussed as follows.

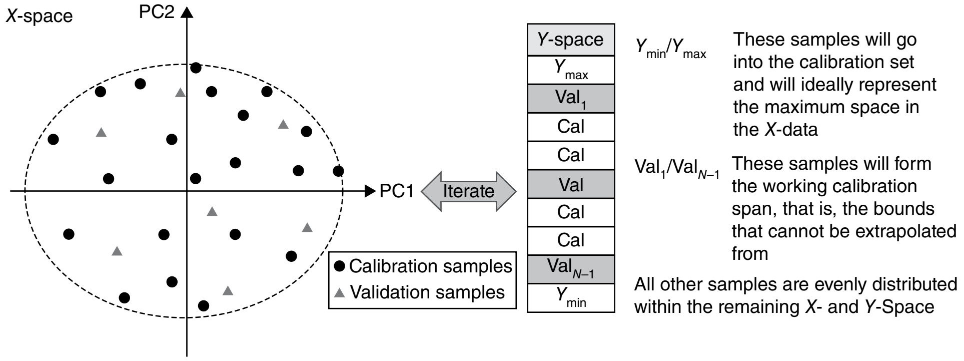

X-space
PC2
PC1
Iterate
● Calibration samples
▲ Validation samples
Y-space
Ymax
Val1
Cal
Cal
Val
Cal
Cal
ValN–1
Ymin
Ymin/Ymax
These samples will go into the calibration set and will ideally represent the maximum space in the X-data
Val1/ValN–1
These samples will form the working calibration span, that is, the bounds that cannot be extrapolated from
All other samples are evenly distributed within the remaining X- and Y-Space

FIGURE 65.48 The process of maximum space sample selection.

65.6.1.1 Maximum Space Sample Selection Maximum space sample selection is a manual process that requires the visualization of the scores space of the entire sample pool and selects samples at the extremes of scores space for all interpretable components/ factors in the model. These samples will form the first part of the calibration set. An even set of well-distributed samples can then be selected from scores space to define the most even coverage of *X*-space. The next step in the process is to order the samples based on their *Y*-values and first select the highest and lowest value to go into the calibration set. Ideally these samples will match those already selected in *X*-space.

The calibration to validation sample ratio is usually 2 : 1; therefore, an even 2 : 1 split of samples based on *Y*-values will yield an even distribution of samples (provided the underlying sample population is not normal or skewed). It is stated here that a boxcar distribution of samples is to be preferred when developing calibration models. Simple row exchange based on the most even distribution of *X*- and *Y*-space is performed such that the entire validation set span in *X*- and *Y*-space is encapsulated by the calibration sample space. This process is shown in Figure 65.48.

65.6.1.2 The Double Kennard–Stone Method The double Kennard–Stone method [51] starts by first finding the two most different samples, often applied on the scores from PCA rather than the original data. Thereafter the next sample is found from the largest distance to these, and this continues until a given number of samples is selected, eventually until the distance is lower than a preset limit. The Euclidean or Mahalanobis distance is normally used. This ensures an even coverage of the samples in the multivariate space. An even better approach is to for every second sample selected put this in a validation set, the so-called double Kennard–Stone or duplex.

Though the objective is to have enough samples in the available sample pool to put a reasonable number aside as a test set, this is not always possible due, for example, to the cost of samples or reference testing. The best alternative to an independent test set for validation is to apply cross validation when a suitable number of samples is not available.

65.6.2 Cross Validation

With cross validation [52], the same samples are used both for model development and testing. A few samples are left out from the calibration data set (based on some predefined criteria) and the model is calibrated on the remaining objects. Then the values for the left-out objects are predicted and the prediction residuals are computed. The process is repeated with another subset of the calibration set, using a systematic way until every object has been left out once; then all residuals are combined to compute the validation residual variance and root mean square error of cross validation (RMSECV) in prediction or classification rates for classification models. It is of utmost importance that the user is aware of which level of cross validation to use. For example, if one physical sample is measured three times, and the objective is to establish a model across samples, the three replicates must be held out in the same cross validation segment. If the objective is to validate the repeated measurement, keep out one replicate for all samples and generate three cross validation segments. The calibration variance is always the same; it is the validation variance that is the important figure of merit.

Kos et al. [53] make a general comment that for sample sets greater than 50 test set validation is preferred whereas cross validation is best for small to medium data sets. Given a specific stratification of the samples in a data set, the level of validation in cross validation should reflect the objective, for example, is the model to be used for other batches of raw materials?

Some general rules and practical considerations for applying cross validation are provided in the following, given the level of validation:

1. Full cross validation, also known as leave-one-out (LOO), leaves out only one sample at a time. If the number of objects is less than 20, this may be a viable option.
2. Segmented cross validation. There are theoretical and practical results indicating that, for example, 10 random segments give a good estimate of the prediction error. Or in general terms, if the model changes considerably when 10% of the samples are taken out, it means the model is not stable. However, it is only in the case where there is no stratification of the samples based on the underlying sampling strategy or the origin of the samples that a random segment CV is justified.
3. Systematic segmented cross validation leaves out a whole group of samples at a time. A typical example is when there are replicated measurements of one physical sample. Depending on the objective either take out all replicates for each physical sample or replicate *n* for all samples.
4. Validating across categorical information about the samples. This enables the analyst to validate across the model and evaluate the robustness across season, raw material supplier, location, operator, etc.

The main purpose of establishing a model may not in itself be for predicting or classifying new objects but to understand the inherent structure in the system under observation. In chemometrics this relates to latent variables that may convey the basic chemical, physical, or biological phenomena. The interpretation of such models is highly dependent on the number of latent variables, and therefore it is vital to assess the correct dimensionality of the model, that is, in more mathematical terms the model rank. It is important to distinguish between numerical rank, statistical rank, and the application-specific rank. Note that even though a representative test set is present, it is nevertheless important to find the correct model rank in the calibration model for predicting the test set.

Both test set validation and cross validation can be applied to any regression model made by either PCA, MLR, PCR, PLS, or other methods. These validation methods are equally applicable to augmented regression models like nonlinear regression, including support vector machine (SVM) models and neural networks, for example, and are perhaps even more important for methods that involve estimates of many parameters as these imply even greater risks of overfitting.

65.7 AN INTRODUCTION TO MSPC

In traditional SPC applications, variables are measured one at a time on control charts where each variable is assumed to be independent of each other. In most process applications, this is not always the case and in many situations, even though individual control charts show the variables to be in control, the overall process is out of control. This is because, as discussed in great detail in this chapter, of the collinear nature of data. Refer to Figure 65.2, case 2, for an excellent example.

Methods such as PCA and PLS are mature and have been used in many applications for quality control using instruments such as spectrometers. Recently, they have gained more interest for modeling the data generated by manufacturing process, and therefore, they form the basis of MSPC methods. Before a detailed discussion of MSPC tools is provided, a short section on multivariate projection will be provided.

65.7.1 Multivariate Projection

The method of PCA can be used effectively as an MSPC tool for monitoring multiple variables simultaneously. In order to be useful as a monitoring tool, it uses the method of projection to achieve this objective. Consider the validated PCA model, where only the informative part of X is retained in terms of scores and loadings:

$$X = TP'$$

This model can be rearranged in terms of the scores T to yield

$$T = XP$$

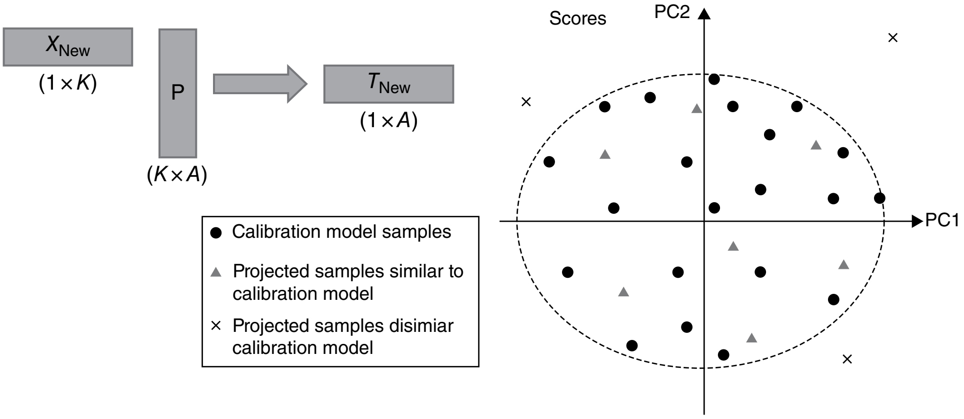

X_{New}
(1 × K)
P
(K × A)
T_{New}
(1 × A)
● Calibration model samples
▲ Projected samples similar to calibration model
× Projected samples disimiar calibration model
Scores
PC2
PC1

FIGURE 65.49 Some common situations in PCA projection.

This means that, given a set of PCA loadings, new samples X can be projected onto these loadings to provide new scores ($\hat{T}_{\text{New}}$). Given a model with established Hotelling's T^2 limits and Q-/F-residual limits, the projected sample can be assessed as either belonging to the population the model was developed from or not. This is expressed mathematically as

$$\hat{T}_{\text{New}} = X_{\text{New}} P$$

Some common situations arising from projection are provided in Figure 65.49.

65.7.2 Hotelling's T^2 Control Chart

Monitoring PCA/PLS scores in control charts or as scatter plots is effective when the number of PCs in a model is at maximum 2. Beyond 2, 3D charts, or multiple 2D charts may be required to visualize the entire process. As was discussed in section "Hotelling's T^2," Hotelling's T^2 is closely related to leverage and is a measure of the distance of a sample from the center of a model. The center of a model in PCA is analogous to the center line of a control chart; therefore, the further away the sample is from the model center, the more likely it represents an out-of-control situation.

The Hotelling's T^2 chart provides a convenient summary of all variables simultaneously no matter how many PCs/factors are in the model. The minimum value of Hotelling's T^2 is zero; this means that when using this chart in a process monitoring application, only an upper limit is required. Figure 65.50 provides an example of a Hotelling's T^2 chart.

In the event where an outlier is detected in the Hotelling's T^2 chart, in most software applications, a user is able to drill down to the variables contributing to why the sample is an outlier. This plot is called a contribution plot [54] and is a weighted

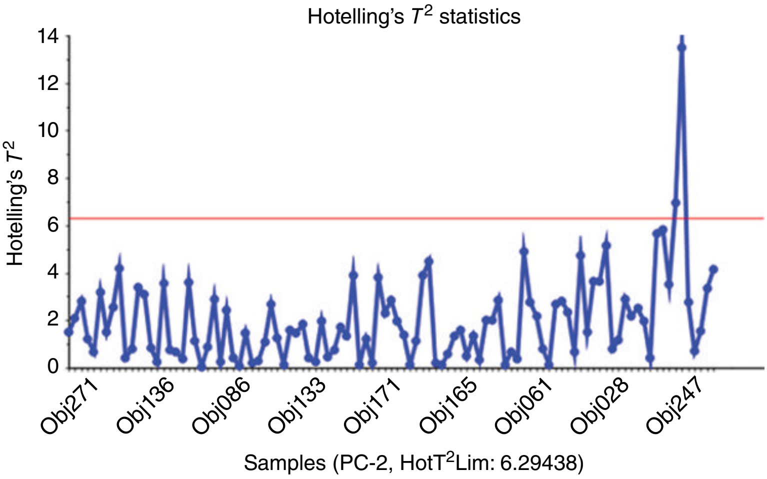


FIGURE 65.50 Example Hotelling's T^2 chart.

loadings plot specific to the sample. The calculation of contribution is provided in the following:

$$c_{nk} = \sum_{a=1}^{A} S_{aa}^{-1} t_{\text{new},a} x_{\text{new},nk} p_{ak}$$

where S_{aa}^{-1} is a square matrix with the inverse of the eigenvalues on the diagonal.

The contribution plot shows which variables contribute most to the sample being an outlier, with respect to the loadings of the model. The variables that are most weighted can be used to provide feedback to a control system for making control decisions. This is known as advanced process control (APC) [55].

65.7.3 *Q*-Residuals

Like the Hotelling's T^2 plot, the *Q*-residuals can be plotted as a line plot to detect those samples that look dissimilar to those used to construct the model. *Q*-residuals also only have an upper limit with the minimum residual being zero (since *Q*-residuals are based on a sum of squares of the residuals). Contribution plots can also be generated from a *Q*-residuals plot to better understand variable contributions to outlying samples.

65.7.4 Influence Plot

The influence plot is fast becoming the standard MSPC plot because it captures all of the diagnostics information for each sample in one plot. The MSPC influence plot is typically constructed as *Q*-residuals versus Hotelling's T^2 plot with statistical limits typically set at 95%. The space bounded by the *Q*-residuals and Hotelling's T^2 limits for a validated number of PCs/factors represents the situation where all variables are

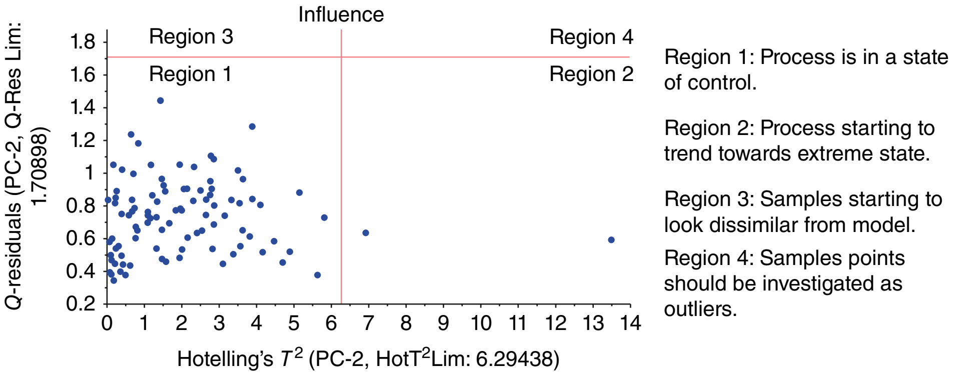


FIGURE 65.51 Example influence plot used for MSPC.

simultaneously in control. Samples that exceed a Q-residual or Hotelling's T^2 boundary only are interpreted as per the charts acting independently. When a sample exceeds both boundaries simultaneously, this situation is indicative of a true outlier and such as situation will require a root cause investigation to resolve the problem. Figure 65.51 provides an example of MSPC influence plot with all of the regions marked on the plot.

65.7.5 Continuous versus Batch Monitoring

In a typical manufacturing environment, there are two main types of process:

1. Continuous manufacturing: This is characterized by processes that have reached a steady state and process models are concerned with detecting out-of-control situations typically based on static models.
2. Batch manufacturing: Typically characterized by processes exhibiting some form of process signature (or trajectory). These processes require a monitoring strategy that assesses whether the process data remain in an envelope defined by the process signature.

65.7.5.1 Continuous Processes Common examples of continuous processes include water quality monitoring, gasoline blending, and many other examples where the process is maintained in a steady state. The variables measured are looking to keep the process in a target range that is static over the entire time the process is being run.

Consequently, the MSPC models used to monitor the process are simpler than those used in batch processes. Section 65.7.6 defines the requirements for variable measurement and alignment systems required for implementing successful process monitoring strategies.

65.7.5.2 Batch Modeling Only a brief discussion on batch modeling will be provided here as a full discussion is outside of the scope of this chapter. In a batch process, raw materials are combined in a suitable batch vessel before chemical, physical, or biological transformation takes place, resulting in an end product. In many cases the control of the batch process is recipe driven and the operations are not adjusted to accommodate raw material variation, changes in uncontrollable factors, and other changing circumstances. The best possible end product quality is achieved by adapting batch operations according to any detectable changes during processing, thus providing a control mechanism to drive a product toward what is known as its desired state. Optimal run settings and the ability to control them within a design space [56] leads to reduced rework and rejects, and improved end product quality, which has the major benefit of saving industry money and resources and, more importantly, increased consumer trust in the product name.

There are a number of batch modeling approaches and the most common assume equal lengths of batches, that is, the batch is expected to start at the same chemical or biological time t_0 and has the same number of time points for all batches. This leads to problems during model building if the data set has uneven rows and ultimately during monitoring if new batches do not meet these criteria. Numerous approaches to handle uneven batch lengths exist, including replacing time with a maturity index [57], dynamic time warping (DTW) [58], time linear expanding/compressing [44], etc. Complications can occur in all of these methods if the first measurement does not coincide with the true t_0, that is, the new batches do not start at the same chemical/biological state. The PARAFAC [59] approach models the data as a true three-way model, which has a possible advantage that the time is modeled as a separate dimension and not connected to ether samples or variables as in the unfolding case. However, the challenges with unequal batch length and chemical time still need to be addressed. Also, the monitoring phase requires dynamic recalculating of models up to the current point of time [60].

An improved batch modeling approach accommodating uneven batch lengths, unknown true t_0, phase changes, and uneven residence times has been proposed by Westad et al. [61]. This is achieved by a true multivariate, feature-based approach that does not make any assumptions about the synchronization and duration of batches. Instead the so-called relative time is estimated by the method itself. Relative time is here used in a broad sense for any transient process including nonlinear behavior, and it is often found to correspond with the underlying chemical, biological, or physical changes during the process.

Figure 65.52 provides an example of how a batch modeling MSPC approach looks like and shows how the process trajectory, representing the process signature, can be displayed with an envelope around the trajectory.

The main difference between continuous and batch modeling strategies is that in continuous, the process is considered to have reached steady state; therefore, univariate control charts can be used to monitor individual variables and MSPC charts will be used to ensure that the process remains within the limits of the static model developed for the process. In batch modeling, the limits are dynamic and thus require a different

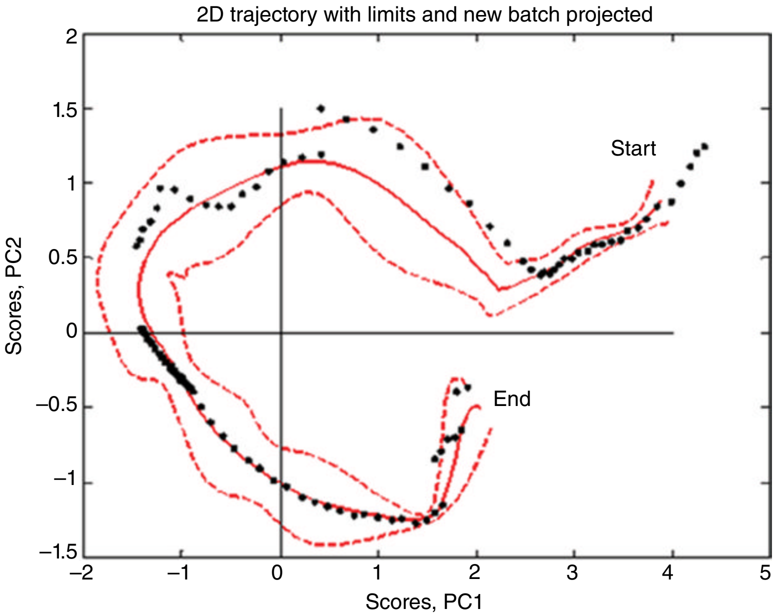


FIGURE 65.52 Example batch process MSPC display.

approach to modeling and monitoring. For more details on batch modeling, the interested reader is referred to the literature cited [55–61].

65.7.6 Implementing MSPC in Practice

The implementation of MSPC requires an intimate knowledge of the process under investigation and a strong data management system. Consider the generic process and data management system shown in Figure 65.53. There are a number of variables being measured and typically these are being measured at different frequencies. This is where subject matter expertise is essential. The time interval for monitoring the process must be set in order to reliably detect a critical change in the process.

If a temperature sensor measures a reading every 100 ms and a critical change in a process is detectable in a time frame of tens of minutes, there is no logic to monitoring the process at such a high sampling rate. Conversely, if a process is rapid and critical changes are picked up in seconds, this would justify a much higher sampling rate.

In today's world of rapid sensors, there is an abundance of information that can be obtained. IBM is quoted as saying that 90% of the world's data has been captured in the past 2 years [62]. With big data, business intelligence, and manufacturing intelligence becoming buzz words in all industries, there will be a much greater expectation on future systems to provide as much data as possible; however, the key question and challenge is, what to do with the data?

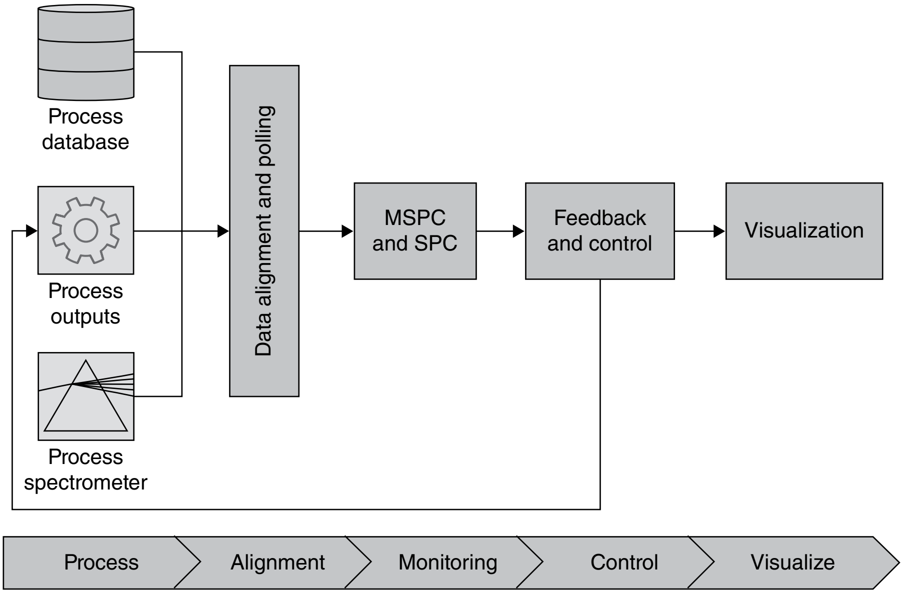


FIGURE 65.53 Generic process and data management system example.

The multivariate methods discussed in this chapter form the basis of what to do with the data, the fundamental challenge if the agglomeration of many data sources into a representative (meaningful) array that can be modeled and analyzed. The first step to a successful process monitoring systems is therefore a data management system capable of importing, in run time, many different formats of data and then compiling them into a representative array. Until the data is time aligned and searchable, the modeling phase becomes a manually driven, painstaking exercise. Figure 65.54 provides a simple situation where four sensors are feeding data at various rates from a process into the data management system.

As the data are being fed at their own natural frequencies (ν_i), a filtering and agglomeration system can be implemented primarily to remove mechanical shock data points from the individual data and to compile the data until they are ready to be polled. Below the agglomeration system is a polling layer. The polling layer is the critical part of representative data collection. It is set, based on subject matter expertise, to poll the data currently sitting in the agglomerator. Whenever the system is polled, the average of the filtered data is used as a single value to build a fused data array that can either be used for multivariate modeling or can be passed onto a multivariate model for evaluation of process state. This data model works for both continuous and batch modeling strategies.

The model shown in Figure 65.54 can be expanded to include vector inputs, such as the data generated by spectrometers, chromatograms, etc. In this case, the whole vector, parts of it, or even the scores obtained by applying multivariate models to the

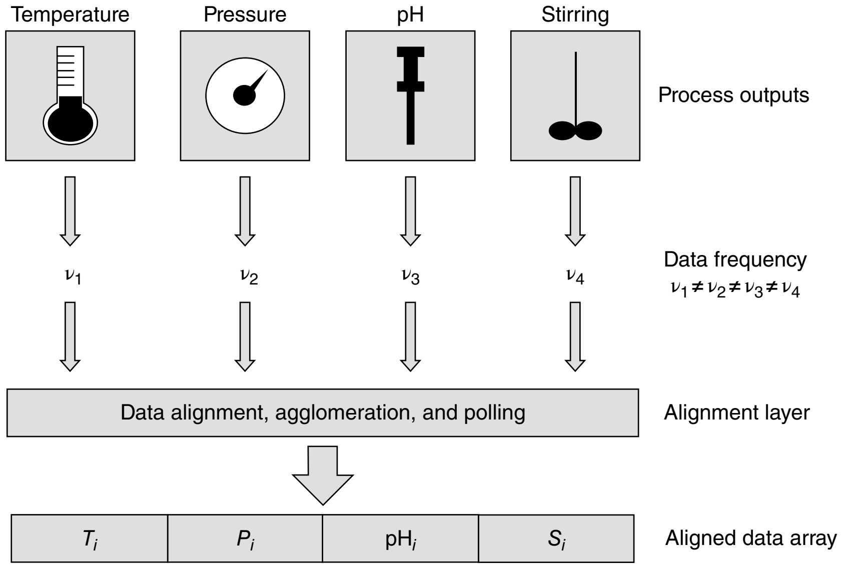


FIGURE 65.54 Process data being fed into a data management system.

data can be used to build a fused data array. The objective of many process control strategies is to achieve feed forward/feedback control loops. By using a modern data management system approach combined with data fusion from process sensors and MVA, the following goals can be achieved:

1. Both SPC and MSPC monitoring approaches can be implemented simultaneously.
2. The alarms generated from SPC/MSPC can be cross-referenced to each other to determine whether an issue is arising from one variable or a combination of variables (refer to Figure 65.2).
3. The polling rate of the process data can be set over multiple operations in order to allow the control system to adapt accordingly to the process.
4. Predictions performed or scores extracted from the fused data can be used as inputs to other parts of the process to provide feed forward/feedback capability to the system. This is particularly important in continuous manufacturing or processes where 100% inspection of units is required.
5. It allows innovation and learning such that the process can be optimized in a shorter time compared to traditional approaches and the predictive ability of the models can allow forward projection (using a time series approach) to warn a system that a process failure may occur.

It is the last point that makes the modern data management system approach most attractive to industry. The ability to predict an event before it occurs allows for a proactive approach, not a reactive approach, to quality, and therefore, this approach should be considered by any manufacturer looking to improve quality, reduce costs through less scrap and better equipment management, use less energy, and improve brand recognition through better and possibly greener approaches.

65.8 TERMINOLOGY

It must be noted here that the usage of terminology has traditionally been a matter of preference; however, terminology standardization has been a subject of debate in the chemometric community in recent times. Wherever possible, this chapter uses only standardized terminology as used in the wider community and avoids the use of proprietary terms. To add more confusion to the matter, similar methods to chemometrics exist in other disciplines, such as engineering, and even though a chemist and an engineer are talking about the same thing, terminology is the killer of interpretation!

This section aims, as far as possible, to introduce the key terms most commonly used in chemometrics.

Accuracy The closeness of agreement between a predicted result and an accepted reference value.

Bias The arithmetic average difference between the reference values and the values produced by the analytical method under test for a set of samples.

Calibration The stage of data analysis where a model is fitted to the available data set and assessed for quality of fit by a proper validation method.

Category variable This is typically a noncontinuous class variable without any quantitative equivalent. Used to group samples into predefined categories and for cross validation to assess the stability of the model during chemometric modeling.

Classification A systematic approach used to sort a set of objects/samples into a set of distinguishable classes. Rules can then be put in place to direct the classification of new objects into these classes. The first step is called unsupervised classification, where classes are defined, and the second stage is supervised, where the rules are used to classify new samples.

Cluster analysis A group of mathematical methodologies used to find sample patterns in complex data sets with the intent of interpreting the groups based on prior subject knowledge.

Collinearity The linear relationship between variables. Two variables are collinear if the value of one variable can be computed from the other using a linear relation.

Continuous variable Any variable measured on an infinitely divisible scale.

Correlation A unitless measure of the amount of linear relationship between two variables.

Covariance A measure of the linear relationship between two variables of whose scale is dependent on the magnitude of the two variables being analyzed.

Cross validation A simulated version of test set validation that creates independent calibration and validation sets in a predefined manner. The selected validation sets are left out of the model calculations and are used to validate the submodel generated. The validation samples are then returned to the sample pool and new calibration and validation samples are selected and the entire process is repeated until all samples have been used for calibration and validation purposes.

Dependent variables Also known as *Y*-variables or responses are a set of variables collected on a representative data set as a reference set such that they can be modeled by a set of independent variables.

Explained variance The proportion of the total variance in a predefined data set that is accounted for by a model generated from that data.

Independent variables Also known as *X*-variables or predictors are a set of variables collected on a representative data set to be modeled and used to either gain insights into the variable and sample relationships or to be used to predict a dependent variable.

Influence A measure of how much impact a single object/sample (or a single variable) has on a developed model. The influence depends on the leverage and the residuals.

Latent variable A variable that is not directly observed but rather inferred (through a mathematical model) from other variables that are observed and directly measured.

Leverage A measure of how extreme an object/sample or a variable is with respect to the center of the model space. Leverage has a one-to-one relationship to the Hotelling's statistic.

Linear discriminant analysis (**LDA**) LDA is the simplest of all possible classification methods that are based on Bayes' formula. The objective of LDA is to determine the best fit parameters for classification of samples by a developed model.

Loading weights These are generated in PLSR models and show how much each *X*-variable (predictor) contributes to explaining the *Y*-variable's (responses) variation for each model factor.

Loadings Loadings are a concentration of the main information carried by variables onto a few components. Each variable has a loading along each model component and its magnitude is a representation of how important that variable is. The corresponding correlation loadings simplify the interpretation as it is independent of the number of variables.

Model A mathematical equation summarizing variations in a data set.

Multiple linear regression (**MLR**) A method for relating the variations in a response variable (*Y*-variable) to the variations of several predictors (*X*-variables). An important assumption for the method is that the *X*-variables are linearly independent, that is, that no linear relationship exists between the *X*-variables. When the *X*-variables carry common information, problems can arise.

Multivariate analysis (MVA) MVA is based on a set of methodologies used for analyzing more than one variable at a time. It encompasses but is not limited to data mining and predictive analytic applications.

Multivariate statistical process control (**MSPC**) A complementary method to traditional statistical process control (SPC) that overcomes the one-variable-at-a-time issues by utilizing multivariate methods that provides information on not only the main variables but also their interactions.

Orthogonal Two variables are said to be orthogonal if the angle between them is 90° (i.e., at right angles to each other).

Outlier An object/sample or variable that shows abnormal characteristics with respect to the rest of the data set analyzed.

Partial least squares regression (**PLSR**) A method for relating the variations in one or several response variables (*Y*-variables) to the variations of several predictors (*X*-variables). This method performs particularly well when the various *X*-variables express common information, that is, when there is a large amount of correlation between the variables.

PLS factor The equivalent of a PC in PLSR only the information in the factor is most related to the response.

Precision A measure of the closeness of repeat predictions made on either the same sample (repeatability) or true replicates (reproducibility).

Prediction Estimating response values from predictor values, using a regression model.

Principal component (**PC**) A PC is a condensation of sample and variable information that describes a particular source of variability in complex data sets. Each PC describes a certain proportion of the total variability of a system, and the first PC (PC1) describes the greatest source of variability each successive PC describing less information than the previous one. Each PC describes an independent source of variability in the data, and therefore, they can be plotted as *X*–*Y* scatter plots for increased interpretability. In mathematical terms the direction of the loading vector is the *eigenvector*.

Principal component analysis (**PCA**) An exploratory data analysis method used to understand complex sample and variable relationships in multivariate data. The aim of PCA is to isolate those variables that most contribute to sample patterns observed. PCA models can be further used for developing classification models or predicting the state of new samples with respect to the original model.

Principal component regression (**PCR**) PCR is a method for relating the variations in a response variable (*Y*-variable) to the variations of several predictors (*X*-variables). This method performs particularly well when the various *X*-variables express common information, that is, when there is a large amount of correlation. PCR is a two-step method. First, a principal component analysis is carried out on the *X*-variables. The principal components are then used as predictors in a multiple linear regression.

Projection In multivariate methods such as PCA and PLSR, each object/sample can be considered as a single point in multivariate space. When a sample is projected onto a model, this results in a score, while variable projections are called loadings.

Regression Generic name for all methods relating the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables). Regression can be used to describe and interpret the relationship between the X-variables and the Y-variables and to predict the Y-values of new samples from the values of the X-variables.

Regression coefficient In a regression model equation, regression coefficients are the numerical coefficients that express the link between variation in the predictors and variation in the response.

Repeated measurement Measurements performed several times on a single sample in a short time period. Repeated measures are used for the estimation of measurement error.

Replicate Measurements carried out several times of different preparations of the same sample. Replicate measures are used for the estimation of experimental error.

Residual A measure of the variation that is not taken into account by a model. The residual for a given sample or a given variable is computed as the difference between observed value and fitted (or predicted) value of the sample.

Residual variance The mean square of all residuals sample- or variable-wise. The complement of residual variance is explained variance.

Root mean square error of calibration (**RMSEC**) A measurement of the average difference between predicted and reference values for calibration samples only.

Root mean square error of cross validation (**RMSECV**) A measurement of the average difference between predicted and reference values for validation segments when cross validation is the method used for validating the model.

Root mean square error of prediction (**RMSEP**) A measurement of the average difference between predicted and reference values for validation samples when the validation method used is test set validation.

Sample Object or unit on which variables are measured and which builds up a row in a data table.

Scores Scores carry information on several variables and are concentrated onto a few underlying variables. Each sample has a score along each model component. The scores show the locations of the samples along each model component and can be used to detect sample patterns, groupings, similarities, or differences.

Standard error of calibration (**SEC**) Variation in the precision of calibration sample predictions over several samples. SEC is computed as the standard deviation of the prediction residuals and is dependent of the validation method used.

Standard error of cross validation (**SECV**) Variation in the precision of predictions over several samples. SECV is computed as the standard deviation of the prediction residuals when cross validation is the validation method used.

Standard error of prediction (**SEP**) Variation in the precision of predictions over several samples. SEP is computed as the standard deviation of the prediction residuals when test set validation is the validation method used.

Support vector The data points that lie closest to the decision surface and are typically the most difficult to classify.

Support vector machine (SVM) SVM is a classification (or regression) method formally defined by a separating hyperplane. In other words, given a known training set of data the algorithm outputs an optimal hyperplane that categorizes or predicts new examples.

Test samples A representative set of samples that are independent of the calibration set used to validate the model using the method of test set validation.

Test set validation A model validation method using a separate test set of samples, providing the most reliable estimate of model performance on future samples.

Validation Validation means checking a models fitness for purpose. In regression, validation allows for estimation of the quality of future predictions. The validation variance can be used as a way to determine how well a single variable is taken into account in an analysis. A variable with a high explained validation variance is reliably modeled, whereas a variable with a low explained validation variance is not explained very well by the model.

Variable Any measured or controlled parameter that has varying values over a given set of samples.

Variance A measure of a variable's spread around its mean value and is computed as the mean square of deviations from the mean. It is equal to the square of the standard deviation.

65.9 CHAPTER SUMMARY

Chemometrics is the area of MVA applied to chemical data; however, this definition can be extended to the physical and biological worlds while still maintaining the same definition. Chemometrics aims to provide insights into complex and sometimes large data sets. Traditional methods of plotting one variable at a time fail very quickly when the number of variables in a data table becomes large and the tools of chemometrics can be used to overcome these obstacles.

Chemometrics has traditionally been associated with spectroscopic or chromatographic data where typical profiles can consist of hundreds to thousands of points per sample. To investigate the finer or hidden details in the data, chemometrics is used. This is why the methods used in chemometrics are sometimes called latent methods (i.e., they look for the hidden structure in the data). In today's world of process sensors, multivariate data arrays can also be generated from many sensor values fused together. Thus, the data move from being homogeneous (such as a spectrum) to heterogeneous (i.e., temperature, pressure, and pH being simultaneously monitored). For heterogeneous data, weighting strategies must be employed that allow individual variables to contribute to a model on an equal basis.

There are three main approaches to chemometric modeling, these being EDA, regression modeling, and multivariate classification. EDA should be the first method applied to a multivariate data set in order to identify any patterns in the data in an unsupervised manner, that is, the data should be analyzed without any preassumptions applied. From there, an analyst can make certain decisions regarding the homogeneity of the data, whether multiple class models should be developed or whether the natural.

Although there are a number of EDA methods available, the most powerful and the workhorse of chemometric methods is PCA. PCA provides a map of sample relationships (known as the scores plot) and a map of variable correlations (known as the loadings plot). When interpreted together, samples groupings observed in the scores can be interpreted based on how the variables measured correlate to each other. Another powerful property of PCA is the complete range of diagnostic tools available for the interpretation and validation of models. This provides PCA not only with a complete range of data modeling tools but also allows PCA models to be used in real-time process monitoring applications. In the case where a new sample sits outside of the calibration population, it can be investigated and the variables that cause the sample to deviate can be investigated and corrected. This is the basis of what is known as MSPC.

Regression modeling aims to make a model that relates a set of independent (*X*-variables) to a set of dependent (*Y*-variables). Multivariate regression methods utilize latent variable approaches to model the structure in *X* and relate to the structure in *Y*. Although MLR is not a latent variable approach, if the scores from PCA are used as variables in the MLR model, the method of PCR results. PCR models the *X*-variables independent from the *Y*-variables at first via PCA. It then regresses the *Y*-variables against the scores to create a prediction model. One of the downsides of PCA is that in some cases, the first score in PCA may not be relevant to modeling *Y*; therefore, an inflation of variance occurs. As subsequent PCs containing chemical (or other) information are added to the model, the model is better able to describe the variability in *Y*.

To overcome the limitations of PCR, the PLSR method was developed. Unlike PCR, PLSR models the *X*- and *Y*-data simultaneously, finding the factors in *X* most correlated to *Y*. This means that the PLSR algorithm will in general converge faster than the PCR algorithm, however, generally yielding the same solution. The PLSR algorithm contains a number of useful diagnostic tools for interpreting and validating the model. From a purist's point of view, PCR is fundamentally simpler and it forces an analyst to better understand their data and the preprocessing used to make a model such that the first component is related to chemical (or other) information. PLSR is the algorithm of choice in most software packages and has therefore become the de facto multivariate regression method. Other regression methods exist, for example, support vector machine regression (SVMR), which are nonlinear approaches to multivariate regression but are outside of the scope of this chapter.

Multivariate classification is the qualitative counterpart of multivariate regression. In multivariate classification, rules are developed, typically during the EDA stage of analysis, and are used to assign new samples into existing classes. This area is known as supervised methods or pattern recognition. There are a number of multivariate classification approaches available, including LDA and SVMC; however, the most effective methods are again based on latent variable methods. These are SIMCA and PLS-DA.

SIMCA utilizes PCA or PLS models to project new samples onto the PCA or PLS loadings to generate new scores values. These scores are compared to each model in the SIMCA library, and there are three possible outcomes: unique classification, ambiguous classification, and no classification.

A unique classification is based on a predefined statistical confidence interval, thus allowing SIMCA to be validated on a purely statistical basis. An ambiguous classification situation occurs when a sample lies in the same space as two or more classes simultaneously. In order to resolve ambiguities, another model has to be generated and the use of SIMCA in a hierarchical approach can be very useful for the classification of complex systems. SIMCA has many diagnostic tools for the interpretation and validation of a model; however, its greatest strength is its ability to classify a sample into a null class. Methods such as LDA will attempt to put a sample into the nearest class, independent of how far away the sample is from a class. SIMCA will reject such as sample based on the measures of *X*-residual and leverage used to define a class model.

PLS-DA utilizes the PLSR algorithm but instead of having continuous *Y*-variables, it uses a binary system to define classes in a data set. For a two-class problem, class *A* can be designated as a zero (0) and class *B* designated as a one (1). The PLS algorithm is applied and the predicted versus reference plot can be used to assess the model's ability to classify new samples. When the problem extends to three or more classes, the form of the PLSR model becomes the PLS-2 model and each class will have its own *Y*-column. Where the class is represented, it has a designation of one (1) in the columns and all other classes are designated zero (0). The same diagnostic and interpretation tools for regular PLSR can be used for PLS-DA.

Until a model is used for a practical application, it is of little value. This is where the area of MSPC is gaining more attention in industry because it allows the combination of traditional SPC and the multivariate methods in unique ways such that process control and fault detection becomes proactive, not reactive. MSPC approaches can be applied to continuous (single or multiple stage) processes or to batch processes. The model development aspects of continuous and batch models vary greatly and both require unique methods to model the system under investigation.

In order for MSPC to be effective, a robust data management system is required to collect data from multiple sources (either scalar sensors or vector sensors such as spectrometers) and fuse the data together in order to generate representative data arrays. These arrays can be used for modeling or for predictive purposes and are sure to find widespread usage in all industries moving forward, particularly because of their ability to be used in closed-loop control systems.

Chemometrics is wide and diverse. It had its infancy over 30 years ago and has now matured into a scientific approach for a number of research and industrial applications. It utilizes as much data as can be analyzed and provides insights into sample patterns, variable relationships, outliers, and the overall importance of variables being used to analyze a system.

REFERENCES

1. Wold, S., “Chemometrics; what do we mean with it, and what do we want from it?”. *Chemometrics and Intelligent Laboratory Systems* 30 (1): 109–115 1995.
2. Dayal, B. S. and MacGregor, J. F., “Improved PLS algorithms”. *Journal of Chemometrics* 11: 73–85 1997.
3. Rannar, S. Lindgren, F. Geladi, P., and Wold, S., “A PLS kernel algorithm for data sets with many variables and fewer objects, Part 1: theory and algorithm”. *Journal of Chemometrics* 8: 111–125 1994.
4. Box, G. E. P. Hunter, J. S., and Hunter, W. G., “*Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building*”, John Wiley & Sons, Inc., New York 1978.
5. Montgomery, D. C., “*Design and Analysis of Experiments*”, 6th Edition, John Wiley & Sons, Inc., Hoboken, NJ 2004.
6. Montgomery, D. C. and Myers, R. H., “*Response Surface Methodology, Process and Product Optimization Using Designed Experiments*”, 2nd Edition, John Wiley & Sons, Inc., New York 2002.
7. Whitcomb, P. J. and Anderson, M. J., “*RSM Simplified, Optimizing Processes Using Response Surface Methods for Design of Experiments*”, Productivity Press, New York 2005.
8. Montgomery, D. C., “*An Introduction to Statistical Quality Control*”, 5th Edition, John Wiley & Sons, Inc., Hoboken, NJ 2005.
9. ICH Harmonized Tripartite Guideline Q2(R1), “Validation of analytical procedures: text and methodology”. *Federal Register* 62(96): 27463–7 1997.
10. Swarbrick, B., “*Multivariate Analysis for Dummies*”, John Wiley & Sons, Inc., Hoboken, NJ 2012.
11. Miller, J. N. and Miller, J. C., “*Statistics and Chemometrics for Analytical Chemistry*”, 5th Edition, Prentice Hall, New York 2005.
12. Adams, M. J., “*Chemometrics in Analytical Spectroscopy*”, The Royal Society of Chemistry, Cambridge 1995.
13. Everitt, B. S., Landau, S., and Leese, M., “*Cluster Analysis*”, 4th Edition, John Wiley & Sons, Inc., New York 2001.
14. Fisher, R. A., “The use of multiple measurements in taxonomic problems”. *Annals of Eugenics* 7 (2): 179–188 1936.
15. Martens, H. and Naes, T., “*Multivariate Calibration*”, John Wiley & Sons, Inc., New York 1989.

16. Pearson, K., “On lines and planes of closest fit to systems of points in space”. *Philosophical Magazine* 2 (11): 559–572 1901.
17. Naes, T. Issakson, T. Fearn, T., and Davies, T., “*A User Friendly Guide to Multivariate Calibration and Classification*”, NIR Publications, Chichester 2002.
18. Esbensen, K. H., “*Multivariate Data Analysis in Practice*”, 5th Edition, CAMO Software, AS., Oslo, Norway 2012.
19. Hotelling, H., “Analysis of a complex of statistical variables into principal components”. *Journal of Educational Psychology* 24: 417–441 1933.
20. Mahalanobis, P. C., “On the generalised distance in statistics”. *Proceedings of the National Institute of Sciences of India* 2 (1): 49–55 1936.
21. Jackson, J. E. and Mudholkar, G. S., “Control procedures for residuals associated with principal component analysis”. *Technometrics* 21 (3): 341–349 1979.
22. Jackson, J. E., “*A User Friendly Guide to Principal Components*”, John Wiley & Sons, Inc., New York 1991.
23. Wold, S. and Esbensen, K., “Principal component analysis”. *Chemometrics and Intelligent Laboratory Systems* 2: 37–52 1987.
24. Hastie, T. Tishbirani, R., and Friedman, J., “*The Elements of Statistical Learning, Data Mining, Inference and Prediction*”, 2nd Edition, Springer Science and Business Media, New York 2009.
25. Draper, N. R. and Smith, H., “*Applied Regression Analysis*”, 3rd Edition, Wiley-Interscience, New York 1998.
26. Daintith, J., “*A Dictionary of Chemistry*”, 6th Edition, Oxford University Press, Oxford 2008.
27. Legendre, A. M., “Sur la Méthode des moindres quarrés”. In: *Nouvelles méthodes pour la determination des orbites des comètes*, Firmin Didot, Paris, 72–80 1805.
28. Gauss, C. F., “*Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*”, Henrich Dieterich, Göttingen 1823
29. Sharaf, M. A. Illman, D. L., and Kowalski, B. R., “*Chemometrics*”, John Wiley & Sons, Inc., New York 1986.
30. Moore, E. H., “On the reciprocal of the general algebraic matrix”. *Bulletin of the American Mathematical Society* 26: 394–395 1920.
31. Penrose, R., “A generalized inverse for matrices”. *Proceedings of the Cambridge Philosophical Society* 51: 406–413 1955.
32. Jolliffe, I. T., “A note on the use of principal components in regression”. *Journal of the Royal Statistical Society, Series C* 31: 300–303 1982.
33. Mason, R. L. and Gunst, R. F., “Selecting principal components in regression”. *Statistical and Probability Letters* 3: 299–301 1985.
34. Wold, S., Martens, H., and Wold, H., “The multivariate calibration problem in chemistry solved by the PLS methods”. In: Ruhe, A. and Kågstrøm, B. (eds.), *Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad, Sweden, March 22–24, 1982*, Springer Verlag, Heidelberg, 286–293 1983.
35. Geladi, P. and Kowalski, B. R., “Partial least squares regression: a tutorial”. *Analytica Chimica Acta* 185: 1–17 1986.

36. Gower, J., “A general theory of biplots”. In: Krzanowski W. J. (ed.), *Recent Advances in Descriptive Multivariate Statistics*. Royal Statistical Society Lecture Notes, 2, Oxford University Press, Oxford, 283–303 1995.

37. Kjeldahl, K. and Bro, R., “Some common misunderstandings in chemometrics”. *Journal of Chemometrics* 24: 558–564 2010.

38. Seasholtz, M. B. and Kowalski, B. R., “Qualitative information for multivariate calibration models”. *Applied Spectroscopy* 44: 1337–1348 1990.

39. Nadler, B. and Coifman, R. R., “Partial least squares, Beer’s law and the net analyte signal: statistical modeling and analysis”. *Journal of Chemometrics* 19: 45–54 2005.

40. Broad, N. Graham, P. Hailey, P. Hardy, A. Holland, S. Hughes, S. Lee, D. Prebble, K. Salton, N., and Warren, P. Guidelines for the Development and Validation of Near-Infrared Spectroscopic Methods in the Pharmaceutical Industry”, In: Chalmers, J. M. and Griffiths, P. R. (eds.), *Handbook of Vibrational Spectroscopy*, John Wiley & Sons, Inc., New York 2002.

41. Trygg, J. and Wold, S., “Orthogonal projections to latent structures (O-PLS)”. *Journal of Chemometrics* 16 (3): 119–128 2002.

42. Fisher, R. A., “The use of multiple measurements in taxonomic problems”. *Annals of Eugenics* 7 (2): 179–188 1936.

43. Wold, S., “Pattern recognition by means of disjoint principal components model”. *Pattern Recognition* 8: 127–139 1976.

44. Eriksson, L. Johansson, E. Kettaneh-Wold, N. Trygg, J. Wikstrøm, C., and Wold, S., “*Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications*”, Umetrics Inc, Umeå, Sweden 2006.

45. Vong, R. Geladi, P. Wold, S., and Esbensen, K., “Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS)”. *Journal of Chemometrics* 2: 281–296 1988.

46. Cristianini, N. and Shawe-Taylor, J., “*An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*”, Cambridge University Press, New York 2000.

47. Luts, J. Ojeda, F. Van de Plas, R. De Moor, B. Van Huffel, S., and. Suykens, J. A. K., “A tutorial on support vector machine-based methods for classification problems in chemometrics”. *Analytica Chimica Acta* 665: 129–145 2010.

48. Harshman, R. A., “How can I know if it’s real?” A catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling”. In: Low, H. G., Snyder, Jr., C. W., Hattie J. and McDonald R. P. (eds.), *Research Methods for Multi-Mode Data Analysis*, Praeger, New York, 566–591 1984.

49. Bro, R. Kjeldahl, K. Smilde, A. K., and Kiers, H. A. L., “Cross-validation of component models: a critical look at current methods”. *Analytical and Bioanalytical Chemistry* 390: 1241–1251 2008.

50. Esbensen, K. H. and Geladi, P., “Principles of proper validation: use and abuse of re-sampling for validation”. *Journal of Chemometrics* 24: 168–187 2010.

51. Kennard, R. W. and Stone, L. A., “Computer aided design of experiments”. *Technometrics* 11 (1): 137–148 1969.

52. Stone, M., “Cross-validatory choice and assessment of statistical predictions”. *Journal of the Royal Statistics Society* 36: 111–147 1974.

53. Kos, G. Lohniger, H., and Krska, R., “Validation of chemometric models for the determination of deoxynivalenol on maize by mid-infrared spectroscopy”. *Micotoxin Research* 19: 149–153 2003.

54. MacGregor, J. and Kourti, T., “Statistical process control of multivariate processes”. *Control Engineering in Practice* 3 (3): 403–414 1995.

55. Zhao, C. Zhao, Y. Su, H., and Huang, B., “Economic performance assessment of advanced process control with LQG benchmarking”. *Journal of Process Control* 19 (4): 557–569 2009.

56. ICH, “Pharmaceutical development”. ICH Harmonized Tripartiate Guideline Q8(R2), Federal Register, Vol. 71 (98) 2009.

57. Nomikos, P. and MacGregor, J. F., “Monitoring of batch processes using multi-way principal component analysis”. *AIChE Journal* 40: 1361–1375 1994.

58. Kassidas, A. MacGregor, J., and Taylor, P., “Synchronization of batch trajectories using dynamic time warping”. *AIChE Journal* 44: 864–875 1998.

59. Smilde, A. Bro, R., and Geladi, P. “*Multi-Way Analysis, Applications in the Chemical Sciences*”, John Wiley & Sons, Inc., Hoboken, NJ 2004.

60. Meng, X. Morris, A. J., and Martin, E. B., “On-line monitoring of batch processes using a PARAFAC representation”. *Journal of Chemometrics* 17: 65–85 2003.

61. Westad, F. Gidskehaug, L. Swarbrick, B., and Flaaten, G.R., “Assumption free modeling and monitoring of batch processes”. *Chemometrics and Intelligent Laboratory Systems* 149: 66–72 2015.

62. IBM, “Accelerate delivery of pervasive analytics with a big data platform”. http://www.ibm.com/analytics/in/en/what-is-smarter-analytics/innovate-with-analytics-tools.html (Accessed December 10, 2015).

66

LIQUID CHROMATOGRAPHY

ZHAO LI, SANDYA BEERAM, CONG BI, ELLIS KAUFMANN,
RYAN MATSUDA, MARIA PODARIU, ELLIOTT RODRIGUEZ,
XIWEI ZHENG, AND DAVID S. HAGE
Department of Chemistry, University of Nebraska, Lincoln, NE, USA

66.1 INTRODUCTION

Chromatography is a separation method in which chemicals or sample components are separated by their different rates of travel through a system that contains a stationary phase and a mobile phase (see Fig. 66.1) [1–5]. The mobile phase acts to transport the sample components through the system. The stationary phase is used to interact with some or all of these components. The stationary phase is held in place by a solid support and may consist of the surface of this support, a coating on the support, or a bonded layer on the support. If the sample components have different degrees of interactions with the stationary phase, they will spend different amounts of time in the mobile phase and, thus, travel at different rates through the chromatographic system. The result is a separation of these chemicals based on their interactions with the stationary phase and mobile phase [3–6].

Liquid chromatography (*LC*) is a type of chromatography in which the mobile phase is a liquid [1–6]. The use of LC with a support held in a column, as is shown in Figure 66.1, was first described by the Russian botanist Mikhail Tswett in 1903 who used this technique to separate plant pigments [5, 7]. Since that time, LC has become an important separation component in many analytical assays and chemical purification methods. The applications of this method range from small ions and molecules up to large biological molecules and polymers. LC is also used in fields that span from biomedical research, pharmaceutical science, and clinical chemistry to environmental

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

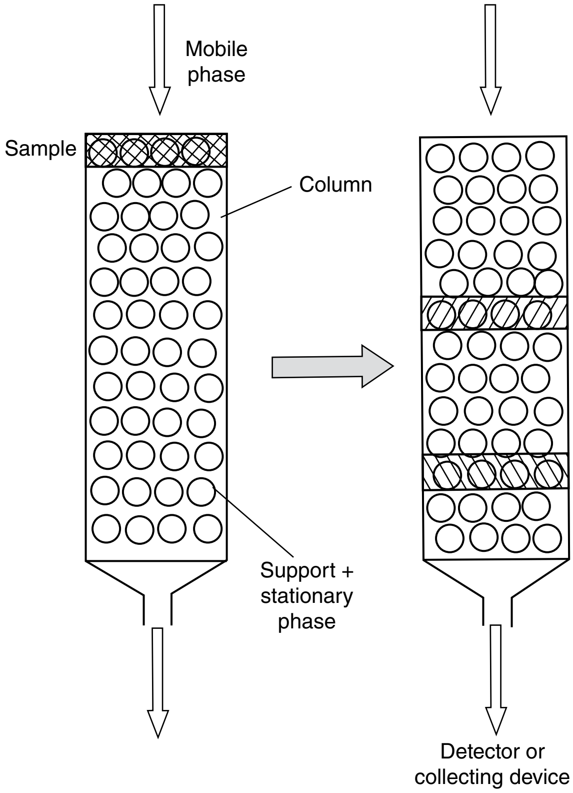


FIGURE 66.1 A typical separation by chromatography, as carried out by using a simple column LC system.

testing, food testing, quality control, and the large-scale purification of chemicals and biochemicals [1–6].

LC can be used in a relatively simple, manual form for chemical isolation and purification, as shown in Figure 66.1 [3–5]. However, instrumental forms of LC are frequently used in many modern applications. A general design of an instrument for conducting modern LC is shown in Figure 66.2. This instrument is known as a *liquid chromatograph* [5, 9]. In this type of system, a pump is used to apply the mobile phase to the column, and samples are injected into the mobile phase stream by means of an injection valve, autoinjector, or similar device. The mobile phase and sample are then passed through a column, which contains the support and stationary phase. A detector is often used to monitor and measure the sample components as they exit the column. Alternatively, a fraction collector may be utilized to collect portions of the eluted components for later detection or use in other methods [5, 9].

A typical separation that is obtained in modern LC is shown in Figure 66.2 [8]. This separation is often represented by a *chromatogram*, which is a plot of the response of the detector or of the measured amount of the eluting compounds as a function of either

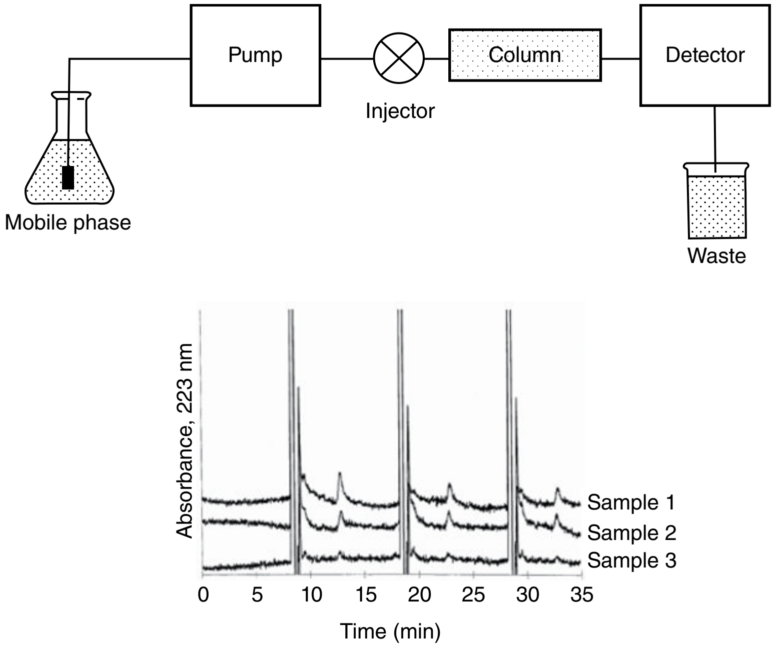


FIGURE 66.2 General design of a modern liquid chromatograph (top) and a set of separations (each done in triplicate) obtained by high-performance liquid chromatography for the analysis of herbicides in three water samples (bottom). The bottom example is adapted with permission from Ref. 8.

the time that has elapsed since the sample was applied to the column (as given by a compound's retention time, t_R) or the volume of mobile phase that has been applied to the column up to this point in time (as given by the compound's retention volume, V_R) [1, 2, 5]. The values of the retention time and retention volume for a chemical on a column are related to a compound's interactions with the stationary phase versus the mobile phase and can be used to identify this substance. The height or area of the peak that is obtained for the same compound can be related to the amount of that substance that was present in the original applied sample [5].

The fact that LC makes use of a liquid as the mobile phase means that the sample and compounds to be separated must be soluble to some extent in this liquid [4–6]. This requirement is needed so that the sample and its components can be applied to and eluted from the column through the use of this mobile phase. The need for solubility in a liquid mobile phase is a much less stringent requirement than what is required in the related technique of gas chromatography (GC), which instead needs sample components to be sufficiently volatile to enter the gas phase for their separation and injection. This difference in requirements, which is due to the use of a liquid instead of a gas as the mobile phase, is a major advantage for LC in terms of its range of applications. For instance, LC can be used with even relatively large biological compounds, such as proteins or DNA, that are difficult to place into the gas phase. The use of a liquid mobile phase also means that LC is usually operated at a much lower temperature than GC, which makes LC more applicable for work with thermally unstable chemicals [4, 5, 10].

66.2 SUPPORT MATERIALS IN LC

The type of support that is used in LC is often used to characterize this method based on its efficiency or "performance." For instance, the diameter of a particulate support in LC will affect the plate height (H) and number of theoretical plates (N) of such a system, with smaller diameters leading to systems with higher resolution due to their small values for H and large values for N [4–6].

The first supports that were used in LC were relatively large and nonrigid materials. Examples of these materials included large particles of silica, alumina, agarose, and cellulose [3–6, 11–13]. These supports are inexpensive and have relatively low back pressures, allowing their use with flow based on gravity or a peristaltic pump. However, these same materials tend to have slow mass transfer properties, which lead to poor efficiencies and large plate heights or small plate numbers. This, in turn, often results in separations with broad peaks, only moderate detection limits, and relatively long analysis times. Such a method is sometimes known as "column chromatography" or *low-performance liquid chromatography* [3–6].

High-performance liquid chromatography (*HPLC*) is a form of LC that utilizes smaller diameter supports (e.g., porous supports with typical diameters of 10 µm or less) or materials with better mass transfer properties when compared to the supports used in low-performance methods [3–6, 9, 11, 12]. These features provide HPLC with a much better efficiency and resolution than traditional LC, which also make it possible to separate a greater number of substances in a shorter period of time and to obtain sharper peaks with lower limits of detection. The use of more efficient support materials in HPLC requires the use of special pumps and instrumental components that can be used at mobile phase pressures ranging from a few hundred to a few thousand psi (e.g., see Fig. 66.2) [3–6]. Higher pressure systems involving even smaller support particles and pressures of 5000–6000 psi or higher are used in an extension of HPLC known as *ultra performance liquid chromatography* (*UPLC*) [5, 14, 15].

There are many materials that can be used as supports in traditional LC and in HPLC. Particle-based supports have been used in all of these methods, with pellicular supports, perfusion supports and monolithic supports also being of interest in HPLC-related methods [5, 6, 12, 16–22]. Traditional porous particles contain pores throughout their structure, while pellicular supports are made from nonporous particles that are coated with a thin layer of a porous material [12]. The use of small porous particles or pellicular supports helps to provide good efficiency by minimizing the distance that solutes have to travel as they move through the pores or porous layer. Perfusion particles have both small side pores and large through pores that act in a similar manner to improve the efficiency of a column [12]. Monolithic supports consist of a continuous bed of a porous polymer and also provide improved mass transfer properties compared to traditional particle-based supports [16–18].

Another possible form of LC is *planar chromatography* [2, 3, 5, 23]. This method is carried out by using a stationary phase that is present on a planar support. Examples are *paper chromatography*, which uses paper as the support, and *thin-layer chromatography* (*TLC*), which uses some other material as the support (e.g., silica particles coated on a glass or plastic plate) [5, 23]. While column-based forms of LC separate solutes based on the time or volume of the mobile phase that is required to have these solutes travel a given distance (e.g., through the column), planar chromatography separates solutes based on the distances they travel in a given amount of time [3, 5]. In TLC the use of small particles (e.g., 5–8 μm diameter porous silica) that are coated on a surface is a method known as *high-performance thin layer chromatography* (*HPTLC*) [3, 24, 25]. Recently, the use of even thinner support layers (e.g., based on monoliths or nanomaterials) has led to a method known as *ultrathin-layer chromatography* (*UTLC*) [26, 27].

66.3 ROLE OF THE MOBILE PHASE IN LC

The retention of a solute in LC can depend on the solute's interactions with both the mobile phase and stationary phase. This means the retention of solutes in LC can also be controlled by varying the type of mobile phase that is applied to a column containing a given stationary phase. Mobile phases in LC can be placed into two categories depending on how similar or different they are in their chemical interactions when compared with the stationary phase. A *strong mobile phase* is a mobile phase that causes a solute to have weak retention on a column, which occurs when the solute tends to spend more time interacting and flowing with the mobile phase than it does interacting with the stationary phase. A *weak mobile phase*, on the other hand, is a mobile phase that produces high retention for a solute on a column. In this case, the interactions of the solute with the stationary phase are favored over the solute's interactions with the mobile phase. In either situation, the type of liquid or solution that constitutes a weak or strong mobile phase will depend on the type of stationary phase that is present in the column [5].

Isocratic elution is a term used to describe a chromatographic separation that uses a mobile phase with a constant composition [2, 5]. A change in the composition of the mobile phase during a chromatographic separation is referred to as *solvent programming* [2, 3, 5]. Solvent programming begins with a weak mobile phase, to allow good retention to be obtained for early eluting solutes, and then moves to a strong mobile phase, to allow later eluting solutes to leave the column in a reasonable amount of time [2–5]. The change in the composition of the mobile phase during solvent programming may be done in a linear fashion, by using a step change or by utilizing a nonlinear change in the mobile phase composition over time [4, 5].

66.4 ADSORPTION CHROMATOGRAPHY

There are five types of LC based on the mechanisms by which they separate solutes. These types are adsorption chromatography, partition chromatography, ion-exchange chromatography (IEC), size-exclusion chromatography (SEC), and affinity chromatography [2, 5]. *Adsorption chromatography* is a type of LC that separates solutes based on their adsorption to the surface of the support. This method is also known as liquid–solid chromatography [1, 2, 4, 5].

The process that gives rise to retention in adsorption chromatography is shown in Equation 66.1. This process involves the binding of solute S to the surface of a support (as represented by the subscript "Support") in place of *n* molecules of the mobile phase (M) [3–5]:

$$\mathrm{S} + n\mathrm{M}_{\mathrm{Support}} \rightleftharpoons \mathrm{S}_{\mathrm{Support}} + n\mathrm{M} \tag{66.1}$$

The retention of the solute in this type of column will depend on the binding strength of the solute to the support and the surface area of the support [3, 4]. This retention will also depend on the amount of mobile phase that is displaced from the surface by the solute and the strength with which the mobile phase binds to the support. The strength of the mobile phase in binding to a given support is described by a term known as the *eluotropic strength* ($\varepsilon°$) [2, 4]. A mobile phase with a large eluotropic strength will bind strongly to the support, which will cause a displaced solute to spend more time in the mobile phase and elute more quickly from the column.

Silica and alumina are the most common stationary phases and supports that are employed in adsorption chromatography. Because these stationary phases are polar in nature, they will retain polar compounds to the greatest degree. Carbon-based materials can be used as nonpolar supports in adsorption chromatography and will retain nonpolar solutes the most. Other supports that have been used in adsorption chromatography are florisil, polyamides, and celite. Increasing the surface area for any of these supports will result in stronger solute retention because this increases the amount of stationary phase versus mobile phase that is present [3–5].

A liquid or solution with a low eluotropic strength on a given support will bind weakly to this material and act as a weak mobile phase for the same support in adsorption chromatography. A liquid or solution with a high eluotropic strength for a support will act as a strong mobile phase for this material. For instance, toluene or heptane are weak mobile phases on a polar support such as silica or alumina but are strong mobile phases on a nonpolar support such as charcoal [3–5].

Adsorption chromatography is a relatively inexpensive and general method for the purification or isolation of organic compounds. For instance, this approach is often used to separate starting materials from products following an organic synthesis [3–5]. This method is especially useful in separating geometrical isomers, such as *cis*/*trans*-isomers or *ortho*/*meta*/*para*-isomers. Adsorption chromatography also forms the basis of many TLC methods. In this type of application, adsorption chromatography is used for the screening and semiquantitative analysis of chemicals such as drugs of abuse and amino acids [3–5, 23–25].

66.5 PARTITION CHROMATOGRAPHY

Partition chromatography is a type of LC in which solutes are separated based on the degree to which they partition between the mobile phase and a stationary phase that is coated on or bonded to a support. This method is also sometimes known as liquid–liquid chromatography.

The retention of a solute (S) in partition chromatography can be described by the following reaction [3–5]:

$$S_{\text{Mobile phase}} \rightleftharpoons S_{\text{Stationary phase}} \tag{66.2}$$

The degree of retention for the solute in partition chromatography will depend on the relative solubility of this solute in the stationary phase versus the mobile phase. This retention will also depend on the amount of the stationary phase versus mobile phase that is present in the column [3–5]. Originally in partition chromatography, the support was coated with a liquid stationary phase, which was also immiscible with the mobile phase. However, many modern types of partition chromatography use stationary phases that are chemically bonded to the support [3, 5].

Based on the polarity of the stationary phase, partition chromatography can be divided into two categories: *normal-phase liquid chromatography* (NPLC) and *reversed-phase liquid chromatography* (RPLC) [2–5]. NPLC (or normal-phase chromatography) is a type of partition chromatography that uses a polar stationary phase. Early columns and systems for carrying out NPLC used supports like silica that were coated with liquids such as water, ethylene glycol, dimethyl sulfoxide, or ethylenediamine as the stationary phase. However, these liquid stationary phases could be lost from the column over time in a process known as *column bleed* [1, 5]. To avoid this problem, modern NPLC supports instead often use chemically bonded phases that contain polar functional groups like cyanopropyl, aminopropyl, or diol groups (see Fig. 66.3) [3–5, 23, 28].

The presence of a polar stationary phase in NPLC means that the weak mobile phase is a nonpolar liquid or solution (e.g., hexane, heptane, or octane). The strong mobile phase in NPLC is a more polar liquid or solution (e.g., tetrahydrofuran, ethanol, or 2-propanol) [3–5, 23, 28]. A mixture of a weak mobile phase and a small amount of a miscible strong mobile phase may be used for isocratic elution or a change from a weak mobile phase to a strong mobile phase over time may be used for gradient elution [3–5]. The solutes in NPLC will elute in the order of their polarity. The least polar solutes will have the weakest retention and will elute from the column first, *Normal phase liquid chromatography (NPLC)*

Aminopropyl phase	Support-$CH_2CH_2CH_2\mathbf{NH_2}$
Cyanopropyl phase	Support-$CH_2CH_2CH_2\mathbf{CN}$

Reversed phase liquid chromatography (RPLC)

Octyl phase (C_8)	Support-$\mathbf{(CH_2)_7CH_3}$
Octadecyl phase (C_{18})	Support-$\mathbf{(CH_2)_{17}CH_3}$

FIGURE 66.3 Some common stationary phases that are used in normal-phase liquid chromatography and reversed-phase liquid chromatography.

while more polar solutes will be more strongly retained and elute later. The applications of NPLC are similar to those for adsorption chromatography with a polar support in that both methods are general purpose tools for the isolation of organic compounds [3–5, 23, 28].

RPLC (or reversed-phase chromatography) is a type of partition chromatography that uses a nonpolar stationary phase [1–5]. Coatings of nonpolar liquids like heptane, squalene, and hydrocarbon polymers were originally used as the stationary phases in this method. These liquid coatings, however, were subject to the same issues with column bleed that were noted previously for liquid stationary phases in NPLC. Most current stationary phases that are used in RPLC consist of a support such as silica that contains a chemically bonded phase with an *n*-alkane or some other nonpolar group. The most common bonded phases that are used in RPLC are *n*-octyl (C_8) and *n*-octadecyl (C_{18}) groups (see Fig. 66.3). Other bonded phases that are often used are *n*-butyl (C_4) and phenyl groups [3–5, 23, 28].

Solutes in RPLC elute in order of their decreasing polarity. This means the most polar compounds in a sample will elute from the column first followed by more nonpolar solutes. A weak mobile phase in RPLC is polar and is often water or an aqueous buffer. Less polar liquids such as acetonitrile, methanol, and 2-propranol are often used as strong mobile phases. Agents like triethylamine and trifluoroacetic acid may also be added to the mobile phase to prevent interactions of polar silanol groups on a silica support with the injected solutes [3–5].

An important advantage of RPLC is that water acts as a weak mobile phase for this method. This is useful because it makes RPLC compatible with the injection of aqueous-based samples, such as biological samples, food or agricultural samples, and many environmental samples. The fact that RPLC separates solutes based on polarity is another valuable feature. As a result of these combined advantages, RPLC is a popular method in areas such as biochemical research, pharmaceutical analysis, clinical testing, food analysis, and environmental analysis (see Fig. 66.2). Examples of chemicals and biochemicals that have been separated by this method include drugs, fatty acids, amino acids, peptides, proteins, and nucleic acids, among many others [3–5, 9, 10, 23, 28].

Another type of partition chromatography is *hydrophilic interaction liquid chromatography* (*HILIC*). HILIC uses a support with polar functional groups and a mobile phase that often consists of a mixture of water and a miscible organic solvent like acetonitrile [29]. Solutes are separated in this method as they partition between a region near the surface of the support that contains water and an area in the mobile phase that is enriched in the organic solvent [29, 30]. Like NPLC and RPLC, this method separates chemicals based on their polarity. The use of some water in the mobile phase makes this method more convenient to use than NPLC with polar compounds that may have low solubility in a nonpolar mobile phase. Applications of HILIC have included its use in proteomics and in the separation and analysis of polar compounds such as drugs and drug metabolites in clinical samples [30–32].

66.6 ION-EXCHANGE CHROMATOGRAPHY

IEC is a type of LC that separates charged solutes based on their interactions with a stationary phase containing fixed groups with a charge opposite to that of the solutes [1, 2]. IEC can be divided into two categories based on the charge of the stationary phase: *cation-exchange chromatography* and *anion-exchange chromatography* [2–5].

Cation-exchange chromatography uses a stationary phase that has negatively charged groups. This method is used to retain and separate positive ions (cations). The stationary phase that is used in this type of IEC may be the conjugate base of a strong acid (e.g., sulfonate) or it may be a conjugate base of a weak acid (e.g., a carboxylate group). In anion-exchange chromatography, the stationary phase has positively charged groups and is used to separate negative ions (anions). The stationary phase in this second type of IEC can be a conjugate acid of a strong base (e.g., a protonated quaternary amino group) or a conjugate acid of a weak base (e.g., the protonated form of a diethylaminoethyl group) [3–5]. Examples are provided in Figure 66.4.

Retention and elution in IEC can be described by the competition of a solute ion and a competing ion in the mobile phase for stationary phase sites having the opposite charge. This process is illustrated in Equation 66.3 for the retention of a positively charged solute (S^+) on a negatively charged cation-exchange support (Support^-) and in the presence of a positively charged competing ion (C^+):

$$S^+ + \text{Support}^-(C^+) \rightleftharpoons \text{Support}^-(S^+) + C^+ \tag{66.3}$$

A similar reaction can be written for the retention of a negatively charged solute in anion-exchange chromatography. The degree to which the solute ion will be retained by the stationary phase in IEC will depend on how strongly the solute ion and competing

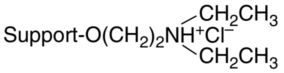
Cation-exchange chromatography	
Sulfonic acid	Support-$SO_3^-H^+$
Carboxylic acid	Support-COO^-H^+
Anion-exchange chromatography	
Quaternary amine	Support-$CH_2N(CH_3)_3^+Cl^-$
Diethylaminoethyl (DEAE)	

FIGURE 66.4 Some common stationary phases that are used in anion-exchange chromatography and cation-exchange chromatography.

ion each bind to the fixed charges on the support, the amount of these charged sites that are present, and the concentration of the competing agent [3–5].

Various types of supports have been used in IEC. Silica that has been modified to contain charged groups is one example. Polystyrene supports that have been modified to contain positively or negatively charged groups are often used in IEC to separate small inorganic and organic ions [3, 4]. Carbohydrate-based supports such as agarose, cross-linked dextran, or cellulose have also been modified to contain charged groups and, due to their relatively large pore sizes and low nonspecific binding, are often used to separate charged biological agents such as proteins and nucleic acids [3–5].

A weak mobile phase in IEC will be a solution that has few or no competing ions present. A weak mobile phase in this method should also have a pH that is optimum for creating the charges needed for binding to occur between the applied solutes and the stationary phase. An increase in the concentration of the competing ion is often used to lower the retention of solute ions in IEC and for gradient elution in this chromatographic method. The pH of the mobile phase can also be used to adjust retention if either the solute or the stationary phase is the conjugate acid or base of a weak base or weak acid. The addition of a complexing agent to the mobile phase is another way of altering the retention of some solute ions in IEC. For example, a sample that contains Fe^{3+}, which would normally be separated by cation-exchange chromatography, can be combined with Cl^- to form $FeCl_4^-$, which could then be separated by anion-exchange chromatography [3–5, 9].

One application of IEC is the use of this technique to remove ionic components from samples and solutions. For instance, cation-exchange and anion-exchange supports are used in water purification systems to produce deionized water by replacing the cations in water with H^+ and the anions with OH^-. In addition, IEC columns and supports are used in biochemistry to concentrate and purify proteins, peptides, and nucleotides based on their charges at a given pH or based on their isoelectric points (i.e., the pH at which a zwitterionic solute has a net neutral charge). Another application of IEC is its use to concentrate and analyze organic or inorganic ions in food, environmental samples, and commercial products [3–5, 9, 33, 34].

Ion chromatography is a special type of IEC that is used with a conductivity detector for chemical analysis. A conductivity detector gives a response that is related to the total ionic content of a solution. Many types of traditional IEC can use a fairly high concentration of a competing ion to elute a solute, which will result in a high background signal on this type of detector. In ion chromatography, the concentration of the competing ion that is needed for solute elution is decreased by using a stationary phase that has a small number of charged sites. Furthermore, the IEC column that is used to separate the desired ions in a sample is combined with a suppressor column or membrane separator, which contains groups that have an opposite charge to those present in the first ion-exchange column. This second column/membrane is used to replace the competing ions with other ions that produce a solution with a much lower conductivity. For example, a system for the analysis of cations would use a cation-exchange analytical column and an anion-exchange suppressor column. A similar strategy can be applied to the analysis of anions by using an anion-exchange analytical column followed by a cation-exchange suppressor column [3, 5].

66.7 SIZE-EXCLUSION CHROMATOGRAPHY

SEC is a type of LC in which the separation of solutes is based on their size [1–5]. In this method, a support with a range of pore sizes is used, in which the pores approach the sizes of the solutes to be separated. Solutes are separated based on their ability to enter various fractions of these pores. Large solutes will be able to enter none or only a few of these pores and will spend most of their time in the mobile phase that is freely flowing outside of the support and through the column. Smaller solutes will be able to enter most or all of the pores and will take longer to pass through the column. The result is a separation based on the size, shape, and molar mass of these solutes [3–5, 34].

In SEC, the volume of the mobile phase that occupies the region outside of the pores of the support is referred to as the excluded volume, V_E. The total volume of mobile phase that is present in both the excluded volume and within the pores of the support is represented by the void volume of the column, V_M. If no other interactions are present between the injected solutes and the support, the retention volume (V_R) for a solute in SEC should be at or between the values for V_E and V_M. Small solutes will elute with a value for V_R that approaches or is equal to the void volume, and large solutes will have a value for V_R that approaches or is equal to the excluded volume. Solutes with intermediate sizes will have retention volumes that have values between V_E and V_M [3–5].

The porous support that is used in SEC should have pore diameters that are in the same general range as the sizes of the solutes that are to be separated. This support should also be inert and not interact directly with the solutes if the separation is to be based only on size. Many of the supports that were discussed in the previous sections can be used in SEC. Biological compounds and aqueous-based samples are usually separated by SEC through the use of agarose, dextrose, and other carbohydrate-based supports. Polystyrene and diol-bonded silica may also be used for work in SEC with samples in organic solvents or aqueous solutions, respectively [3, 33, 34].

Because there are ideally no interactions of the solutes with the support and there is no true stationary phase in SEC, there is also no weak or strong mobile phase in this method. The selection of the mobile phase in SEC is instead made based on the solubility of the chemicals that are to be separated and the stability or properties of the support. Either polar or nonpolar solvents can be utilized as mobile phases in SEC. If the mobile phase is an organic solvent, the term *gel permeation chromatography* is often used to describe the resulting SEC method. *Gel filtration chromatography* is the term used to refer to a SEC technique in which the mobile phase is water or an aqueous solution [1, 2, 5].

SEC has both preparative and analytical applications. As an example, gel filtration chromatography is often used to purify biological samples, such as the removal of small solutes from large biological molecules like proteins. SEC can also be used to characterize the molar mass or distribution in mass for a solute or group of solutes. To determine the molar mass of a solute, standards that are similar to the solute of interest but that have known masses are first injected onto an SEC column. This group of solutes should have a size range that includes some standards that can enter all or most of the pores of the support, some that are completely excluded from the pores, and several that can access intermediate volumes. A plot is then made of the logarithm of the molar mass of each standard versus the solute's measured retention volume, retention time, or some related measure of retention (see Fig. 66.5). This plot can then be used to determine the molar masses of other similar solutes that are injected onto the same column [3, 5].

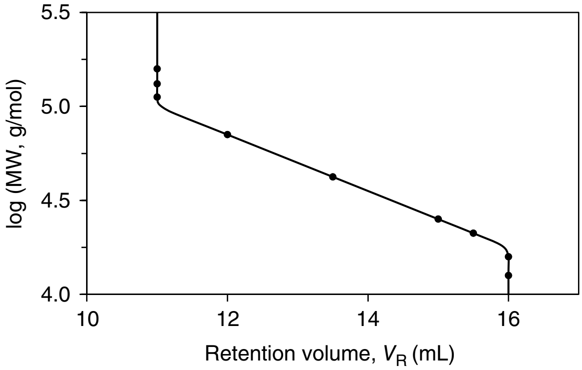


FIGURE 66.5 An example of calibration curve for determination of the molecular weight (MW) of solutes based on size-exclusion chromatography.

66.8 AFFINITY CHROMATOGRAPHY

Affinity chromatography is a type of LC in which solutes are separated based on their binding to a stationary phase that is a biologically related agent [2, 5, 35]. Because of the strong and selective nature of many biological interactions, this method can be a powerful technique for the purification and analysis of solutes that are complementary to the immobilized binding agent. This immobilized binding agent is known as the affinity ligand [35, 36]. The affinity ligand often interacts with its target solute through several interactions, such as dipole–dipole interactions, hydrogen bonding, van der Waals forces, and ionic interactions. The fit of the target at the affinity ligand's binding site may also involve steric effects. The overall result of these various interactions is selective and strong, but usually reversible, binding between the target and the affinity ligand [35, 36].

Figure 66.6 shows a common format for using affinity chromatography [35, 36]. First, a sample containing the target solute is injected or applied in the presence of an application buffer that promotes binding by the target to the immobilized affinity ligand. During this step, the target will be bound by the column while other sample components will tend to be washed away. If the target has strong binding to the affinity ligand, an elution buffer can later be passed through the column to release the bound target. The released target can then be detected or collected for further use. Once the target has eluted, the application buffer can be passed again through the column, and the affinity ligand is allowed to regenerate prior to the next application or injection of a sample [36].

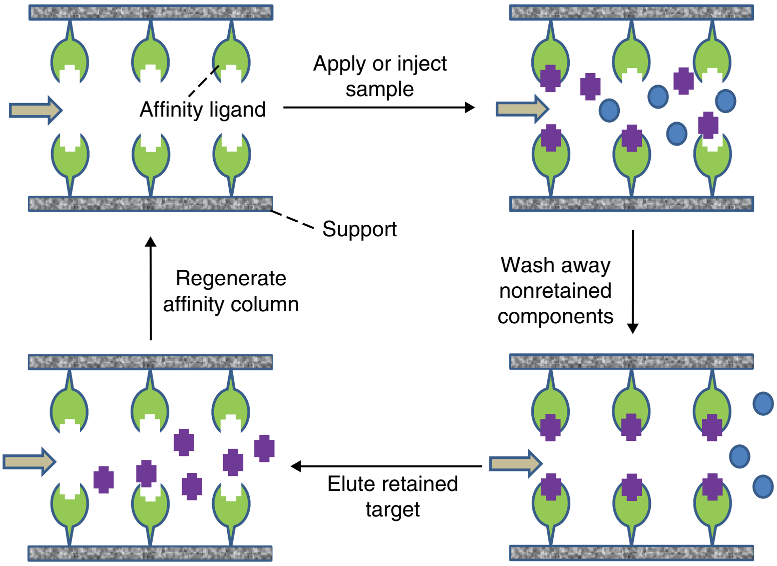


FIGURE 66.6 A typical on/off elution scheme used in affinity chromatography.

The affinity ligand that is used as the stationary phase in affinity chromatography plays a key role in determining which solutes will be retained by the column. This ligand may be a biological agent, a mimic of a biological agent, or even a synthetic compound [35–38]. All of these binding agents can be classified as being either a *high-specificity ligand* or a *general* (or *group specific*) *ligand* [35, 36]. A high-specificity ligand tends to bind to only one target solute or a group of closely related solutes. Common examples of high-specificity ligands are antibodies (which can bind to their corresponding antigens) and enzymes (which can bind to their substrates, cofactors, or inhibitors). A general ligand tends to bind to a set of targets that have a common feature in their structures. Examples of general ligands are immunoglobulin-binding proteins like protein A or protein G, lectins (i.e., nonimmune system proteins that bind sugar residues), boronates, biomimetic dyes, and metal ion chelates [35–38].

An affinity ligand can be immobilized onto the support in various ways. The most common way for immobilizing an affinity ligand is to covalently attach it to the support. This process uses an activated support that can react with functional groups on the affinity ligand, such as amine groups, carboxylic acids, sulfhydryl residues, or aldehyde groups [35–38]. Some affinity ligands can be immobilized through noncovalent adsorption. This approach is often used with antibodies by adsorbing these affinity ligands to immobilized protein A or protein G, and it is used in the adsorption of biotin-tagged affinity ligands on supports that contain immobilized avidin or streptavidin [35, 38]. Another technique for immobilizing large affinity ligands such as liposomes, cells, or proteins is to entrap or encapsulate these binding agents within the support [35–41]. In addition, synthetic affinity ligands against a given target can be generated during the preparation of some polymeric supports in a process known as *molecular imprinting* [35, 42, 43].

The mobile phase will also affect the degree to which a target solute will bind to the affinity ligand. The application buffer is the weak mobile phase in affinity chromatography and is typically a buffer or solution that mimics the natural conditions under which the affinity ligand binds to its target [35, 36]. The strong mobile phase in affinity chromatography is the elution buffer. This buffer may involve a change in the pH, ionic strength, or polarity to weaken the binding of the affinity ligand with its target, giving an approach known as *nonspecific elution*. Another approach for elution is to add an agent to the mobile phase that competes with the target for the affinity ligand or with the affinity ligand for the target. This second approach is known as *biospecific elution* [35, 36]. If the target has sufficiently weak binding to the affinity ligand in the presence of the application buffer, it is possible to even use isocratic conditions to elute the target from the column. This last situation tends to occur with systems that have association equilibrium constants of 10^5–10^6 M^{-1} or less and gives a method known as *weak affinity chromatography* [35].

Probably the most common application for affinity chromatography is the selective purification and isolation of biological compounds [35–38]. This type of application includes the large-scale purification of biopharmaceuticals and enzymes, such as by using biomimetic dyes or immobilized enzyme inhibitors as the affinity ligands [35, 37, 44]. This also includes the isolation of recombinant histidine-tagged proteins through the use of *immobilized metal ion affinity chromatography* (*IMAC*) [35, 45, 46]. Other examples are the use of antibodies as ligands for the isolation of various specific targets in an approach known as *immunoaffinity chromatography* (*IAC*) and the use of lectins to isolate glycoproteins or other carbohydrate-containing targets [35, 38, 47, 48].

There are a variety of analytical applications for affinity chromatography. For instance, many chiral separations are based on the binding of drugs or other chiral solutes to stereoselective affinity ligands such as cyclodextrins, enzymes, or serum transport proteins [35, 49, 50]. Antibodies and immunoaffinity columns have also been used in various formats to analyze specific solutes in a wide range of samples by using chromatographic-based immunoassays [35, 47, 48]. In addition, affinity columns containing ligands such as antibodies and lectins have been combined in both online and offline extraction formats with techniques like RPLC and mass spectrometry for chemical analysis. Columns containing antibodies have further been used to remove possible interfering compounds from samples in a method known as *immunodepletion*. This last approach has been used in proteomics to remove major proteins from samples and to make it easier to detect or measure less abundant proteins [48].

Affinity chromatography can be further used to study biological interactions. This approach is sometimes known as *analytical affinity chromatography* or *biointeraction affinity chromatography* [35, 51, 52]. A variety of data on a biological interaction can be generated through affinity chromatography, such as the strength or rate of the reaction and the number of binding sites that are involved. Information on the location of interaction sites and the binding strength of a target at specific sites on an affinity ligand can also be acquired through this approach [35, 51, 52]. Examples of systems that have been examined by this method are sugar/lectin interactions, protein/protein interactions, and drug/protein interactions [35, 51, 52].

66.9 DETECTORS FOR LIQUID CHROMATOGRAPHY

Many types of detectors can be used to monitor solutes as they elute from modern LC systems. One general detector that can be used is a *refractive index* (*RI*) *detector* [5, 6, 9, 34]. This detector responds to the change in RI that occurs in the mobile phase as solutes elute from the column. An RI detector can detect the presence of a solute by comparing the RI of the mobile/solute mixture to a reference stream or portion of the mobile phase with no solutes present. An RI detector has a moderate limit of detection when compared to other detectors for LC, as shown in Table 66.1. However, this type of detector is valuable for work with solutes that have an unknown composition or that may not contain a chromophore or fluorophore (e.g., as is the case for many carbohydrates and lipids). The response of an RI detector can be affected by changes in temperature,

TABLE 66.1 Common LC Detectors[a]

Detector Type	Compounds Detected	Detection Limits
Refractive index detector	Universal—all compounds	0.1–1 μg
UV/Vis absorbance detector	Compounds with chromophores	0.1–1 ng
Evaporative light scattering detector	Nonvolatile compounds	10 μg
Conductivity detector	Ionic compounds	0.5–1 ng
Fluorescence detector	Fluorescent compounds	1–10 pg
Electrochemical detector	Electrochemically active compounds	0.01–1 ng
Mass spectrometry	Universal (full scan mode) Selective (selected ion monitoring mode)	0.1–1 ng

[a] This table is based on information provided in Refs. [5, 25] and obtained from manufacturers of these detectors. The concentration limits of detection for these devices can be estimated by dividing the above mass values by 10–100 μl, which is a typical range of sample injection volumes that are used in HPLC.

so it is important to keep this device at a constant temperature. A change in mobile phase composition, as occurs during gradient elution, may also cause a change in the background response of an RI detector if the reference solution does not undergo similar changes in composition [5, 6, 9].

Ultraviolet/visible (UV/Vis) absorbance detectors are also commonly found on modern LC systems [5, 6, 9]. These detectors measure the ability of eluting solutes to absorb either UV or visible light. One device in this category is a *fixed wavelength absorbance detector*, which always monitors a specific wavelength (e.g., 254 nm, where many organic compounds with aromatic groups or unsaturated bonds absorb light). Another device in this group is a *variable wavelength absorbance detector*, which can allow absorbance to be measured at a wavelength that is selected from a relatively broad range (e.g., 190–900 nm). A third type of device is a *photodiode array detector* (*PDA*), which uses a detector array to simultaneously measure absorbance at many wavelengths. All of these detectors require that some type of chromophore be present on the solute or that a chromophore first be added to the solute through derivatization. These detectors tend to provide much better limits of detection than an RI detector for such solutes, with detection limits usually in the 10^{-8} M range. Absorbance detectors are also easy to use with gradient elution as long as the weak and strong mobile phases do not have significant absorption of light at the wavelength that is being monitored [5, 6, 9, 34].

Another type of general detector for LC is the *evaporative light scattering detector* (*ELSD*) [5, 9]. This device can be used to monitor solutes that are less volatile than the mobile phase. In this detector, the mobile phase/solute mixture is converted into a spray of small droplets. The solvent is next evaporated away, leaving behind small particles of the less volatile sample components. These particles are detected by examining their ability to scatter a beam of light, where the degree of light scattering is related to both the size and amount of solute particles that are present. An ELSD has a better limit of detection than an RI detector and is easier to use with gradient elution. In addition, the solutes that are detected by an ELSD do not have to have any chromophore present, which is an advantage of this device over absorbance detectors [5, 9].

A *conductivity detector* is a detector that can be used in LC to monitor ionic solutes. These solutes are detected by measuring the ability of the mobile phase and its ionic contents to conduct a current when this mixture is placed in an electrical field. As was mentioned earlier, a conductivity detector is used in ion chromatography to aid in the analysis of ions in various samples. A conductivity detector can be used with gradient elution; however, the background signal will change if the ionic strength (and possibly pH) of the mobile phase are not kept constant. It is also necessary with this type of device to keep the overall ionic composition and background conductance of the mobile phase reasonably low so that eluting solute ions can be detected.

A more specific detector that can be used in LC is a *fluorescence detector* [3, 5, 6]. This type of detector looks at the ability of eluting solutes to both absorb and emit light at a given set of wavelengths. This set of wavelengths allows for more selective detection than occurs when using absorbance measurements and allows for a lower background signal, which provides improved limits of detection. This type of device can be used to look at solutes that are naturally fluorescent or that have been converted into a fluorescent derivative. A fluorescence detector can be used with gradient elution but does require relatively pure mobile phases to help maintain good fluorescence signals and a low background response [3, 5, 6, 9].

Electrochemical detectors can also be used in LC, giving a method referred to as *liquid chromatography/electrochemical detection* (*LC-EC*) [3, 5, 6]. This type of detector is used to monitor solutes that can undergo oxidation or reduction (e.g., aldehydes, ketones, phenols, mercaptans, peroxides, and some carbohydrates). Many of these detectors work by measuring the amount of current that is generated as a solute is oxidized or reduced at a given potential, although other formats are possible as well. The response in this case will depend on the amount of oxidation or reduction that is taking place, which affects the size of the measured current. This type of electrochemical detector can have a quite low limit of detection because of the accuracy and precision with which currents can be measured. Electrochemical detectors and LC/EC can also be used with gradient elution [3, 5, 6, 9, 34].

Liquid chromatography/mass spectrometry (*LC/MS*) is another important combination that is popular in the separation and analysis of chemicals by LC [3, 5, 6, 9]. This method uses mass spectrometry and mass measurements of ions to measure and identify chemicals based on the molecular ions or fragment ions that are generated for these chemicals. Such a system can be used to look at all or most of the ions that are produced in a "full-scan mode," which results in a general detection method. Alternatively, only a few ions that are characteristic of a particular solute or set of solutes can be examined through "selected ion monitoring," providing a more selective method for detection [3, 5, 9].

LC/MS is often carried out by using electrospray ionization (ESI) to generate ions from the eluting solutes [5, 6, 53]. These ions are then examined and separated based on their mass-to-charge ratios by using a quadrupole mass analyzer or other type of mass analyzer. ESI can be used in LC/MS to examine substances that range from small polar compounds to proteins. ESI and LC/MS can also be utilized with gradient elution. The use of LC/MS with ESI is particularly useful in work with proteins and peptides, which tend to give ions with high mass-to-charge ratios when examined by other ionization methods. In ESI, many charges are often placed on one protein or peptide, which provides ions with reasonably low mass-to-charge ratios that are easier to measure by common mass analyzers [5, 6, 53].

66.10 OTHER COMPONENTS OF LC SYSTEMS

In addition to the detector, other important components of LC instruments are the pumps and columns. A pump for HPLC should be able to generate pressures up to 5000–6000 psi and flow rates ranging from 0.1 to 10 mL/min with good reproducibility [3, 5, 9, 53]. Even higher pressures are required in UPLC [5, 14]. The two types of pumps that are used in most LC systems are reciprocating pumps and syringe pumps. A reciprocating pump contains a chamber in which a rotating cam causes a piston to move back and forth. Mobile phase is pumped into and out of this chamber by the movement of the piston, and the direction of this flow is set through the use of check valves. A reciprocating pump works well in the milliliter per minute flow rate range and is easy to use with gradient elution. When a lower flow rate or a smaller column is to be used, a syringe pump is usually employed. In this device, a syringe produces flow of the mobile phase by this liquid or solution being passed out of a chamber as a plunger is slowly depressed into the chamber. A syringe pump can achieve constant flow rates in the microliter per minute range. However, this pump is also less convenient to use than a reciprocating pump in work with large volumes of mobile phase or when changing from one mobile to another during gradient elution [3, 5, 9].

There are various column dimensions and flow rates that are used in LC [3, 5, 9, 25]. In HPLC, the columns usually have a length of 5–30 cm and an internal diameter of 4.1 or 4.6 mm. Such columns are used at typical flow rates of 1–3 mL/min [5, 25]. Longer columns or capillaries with smaller internal diameters, such as microbore columns and packed capillaries, are often needed for separations requiring higher efficiencies. For instance, microbore columns often have lengths of 10–100 cm, internal diameters of 1–2 mm, and are used at flow rates of 0.05–0.2 mL/min [25]. These longer and narrower columns also require smaller sample volumes, to avoid overloading the column, while the use of lower flow rates provides a reasonable operating pressure. These latter features can be advantageous in that they can allow work with small amounts of samples. The use of low flow rates is especially useful in LC/MS in that it means less solvent must be removed from solutes during the ionization process and before the resulting ions can be examined by the mass spectrometer [3, 5, 9, 53].

ACKNOWLEDGEMENTS

Portions of this work were supported by the University of Nebraska–Lincoln (UNL) UCARE program, the NSF (under the REU program and grant CMI 1309806), the NSF/EPSCoR program (grant EPS-1004094), and the NIH (grants R01 GM044931 and R01 DK069629). Additional support for R. Matsuda was obtained through a fellowship from the Molecular Mechanisms of Disease Program at UNL.

REFERENCES

1. R.E. Majors and P.W. Carr, "Glossary of liquid-phase separation terms", *LC-GC*, 19 (2001) 124–162.
2. J. Inczedy, T. Lengyel, and A.M. Ure, *International Union of Pure and Applied Chemistry-Compendium of Analytical Nomenclature: Definitive Rules*, Blackwell Science, Malden, 1997, Chapter 9.
3. C.F. Poole and S.K. Poole, *Chromatography Today*, Elsevier, New York, 1991.
4. B.L. Karger, L.R. Snyder, and C. Horvath, *An Introduction to Separation Science*, John Wiley & Sons, New York, 1973.
5. D.S. Hage and J.D. Carr, *Analytical Chemistry and Quantitative Analysis*, Pearson Prentice Hall, Upper Saddle River, 2011, Chapter 22.
6. D.A. Skoog, F.J. Holler, and T.A. Nieman, *Principles of Instrumental Analysis*, 5th Ed., Brookes Cole, Boston, 1998, Chapter 28.
7. L.S. Ettre, "M.S. Tswett and the invention of chromatography", *LC-GC*, 21 (2003) 458–467.
8. M.A. Nelson, A. Gates, M. Dodlinger, and D.S. Hage, "Development of a portable immunoextraction/RPLC system for field studies of herbicide residues", *Anal. Chem.*, 76 (2004) 805–813.
9. W.J. Lough and I.W. Wainer, *High Performance Liquid Chromatography: Fundamentals Principles and Practice*, Blackie Academic, New York, 1995.
10. K.K. Unger, R. Ditz, E. Machtejevas, and R. Skudas, "Liquid chromatography—its development and key role in life science applications", *Angew. Chem. Int. Ed.*, 49 (2010) 2300–2312.
11. R.E. Majors, "Effect of particle size on column efficiency in liquid-solid chromatography", *J. Chromatogr. Sci.*, 11 (1973) 88–95.
12. R.E. Majors, "A review of HPLC column packing technology", *Am. Lab.*, 35 (2003) 46–54.
13. L.S. Ettre, "Csaba Horvath and the development of the first modern high performance liquid chromatograph", *LC-GC*, 5 (2005) 85–90.
14. J.W. Thompson, J.S. Mellors, J.W. Eschelbach, and J.W. Jorgenson, "Recent advances in ultrahigh-pressure liquid chromatography", *LC-GC*, 24 (2006) 16–20.
15. J.E. McNair, K.C. Lewis, and J.W. Jorgenson, "Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns", *Anal. Chem.*, 69 (1997) 983–989.

16. F. Svec and C.G. Huber, “Monolithic materials: promises, challenges, and achievements”, *Anal. Chem.*, 78 (2006) 2100–2108.

17. G. Guiochon, “Monolithic columns in high-performance liquid chromatography”, *J. Chromatogr. A*, 1168 (2007) 101–168.

18. F. Svec, “Porous polymer monoliths: amazingly wide variety of techniques enabling their preparation”, *J. Chromatogr. A*, 1217 (2010) 902–924.

19. G. Guiochon and F. Gritti, “Shell particles, trials, tribulations and triumphs”, *J. Chromatogr. A*, 1218 (2011) 1915–1938.

20. R. Hayes, A. Ahmed, T. Edge, and H. Zhang, “Core-shell particles: preparation, fundamentals and applications in high performance liquid chromatography”, *J. Chromatogr. A*, 1357 (2014) 36–52.

21. R.W. Brice, X. Zhang, and L.A. Colón, “Fused-core, sub-2 microm packings, and monolithic HPLC columns: a comparative evaluation”, *J. Sep. Sci.*, 32 (2009) 2723–2731.

22. J.J. Kirkland, F.A. Truszkowski, C.H. Dilks, Jr., and G.S. Engel, “Superficially porous silica microspheres for fast high-performance liquid chromatography of macromolecules”, *J. Chromatogr. A*, 890 (2000) 3–13.

23. A. Braithwaite and F.J. Smith, *Chromatographic Methods*, 4th Ed., Chapman & Hall, New York, 1990.

24. R.B. Patel, M.R. Patel, and B.G. Patel, “Experimental aspects and implementation of HPTLC”, in M.M. Srivastava (Ed.), *High Performance Thin Layer Chromatography (HPTLC)*, Springer, Germany, 2011, pp. 41–54.

25. S.K. Poole and C.F. Poole, “High performance stationary phases for planar chromatography”, *J. Chromatogr. A*, 1218 (2011) 2648–2660.

26. H.E. Hauck and M. Schulz, “Ultrathin-layer chromatography”, *J. Chromatogr. Sci.*, 40 (2002) 550–552.

27. H.E. Hauck, O. Bund, W. Fischer, and M. Schulz, “Ultra-thin layer chromatography (UTLC). A new dimension in thin-layer chromatography”, *J. Planar Chromatogr.*, 14 (2001) 234–236.

28. L.R. Snyder and J.J. Kirkland, *Introduction to Modern Liquid Chromatography*, 2nd Ed., John Wiley & Sons, New York, 1979.

29. B. Buszewski and S. Noga, “Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique”, *Anal. Bioanal. Chem.*, 402 (2012) 231–247.

30. P.J. Boersema, S. Mohammed, and A.J.R. Heck, “Hydrophilic interaction liquid chromatography (HILIC) in proteomics”, *Anal. Bioanal. Chem.*, 391 (2008) 151–159.

31. Y. Hsieh, “Potential of HILIC-MS in quantitative bioanalysis of drugs and drug metabolites”, *J. Sep. Sci.*, 31 (2008) 1481–1491.

32. K. Novakova and H. Vlckova, “A review of current trends and advances in modern bio-analytical methods: chromatography and sample preparation”, *Anal. Chim. Acta*, 656 (2009) 8–35.

33. E. Katz, R. Eksteen, P. Schoenmakers, and N. Miller (Eds.), *Handbook of HPLC*, Marcel Dekker, New York, 1998, Chapter 10.

34. B. Ravindranath, *Principles and Practice of Chromatography*, John Wiley & Sons, New York, 1989.

35. D.S. Hage (Ed.), *Handbook of Affinity Chromatography*, 2nd Ed., CRC Press/Taylor & Francis, New York/Boca Raton, 2006.
36. R.R. Walters, "Affinity chromatography", *Anal. Chem.*, 57 (1985) 1099A–1114A.
37. M. Zachariou (Ed.), *Affinity Chromatography: Methods and Protocols*, 2nd Ed., Humana Press, Totowa, 2010.
38. G.T. Hermanson, A.K. Mallia, and P.K. Smith, *Immobilized Affinity Ligand Techniques*, Academic Press, New York, 1992.
39. Q. Yang and P. Lundahl, "Immobilized proteoliposome affinity chromatography for quantitative analysis of specific interactions between solutes and membrane proteins. Interaction of cytochalasin B and D-glucose with the glucose transporter Glut1", *Biochemistry*, 34 (1995) 7289–7294.
40. C.M. Zeng, Y. Zhang, L. Lu, E. Brekkan, A. Lundqvist, and P. Lundahl, "Immobilization of human red cells in gel particles for chromatographic activity studies of the glucose transporter Glut1", *Biochim. Biophys. Acta*, 1325 (1997) 91–98.
41. A.J. Jackson, J. Anguizola, E.L. Pfaunmiller, and D.S. Hage, "Use of entrapment and high-performance affinity chromatography to compare the binding of drugs and site-specific probes with normal and glycated human serum albumin", *Anal. Bioanal. Chem.*, 405 (2013) 5833–5841.
42. M. Komiyama, T. Takeuchi, T. Mukawa, and H. Asanuma, *Molecular Imprinting from Fundamentals to Applications*, Wiley-VCH, Weinheim, 2002.
43. D. Kriz, O. Ramstrom, and K. Mosbach, "Molecular imprinting. New possibilities for sensor technology", *Anal. Chem.*, 69 (1997) 345A–349A.
44. Y.D. Clonis, N.E. Labrou, V. Kotsira, K. Mazitsos, S. Melissis, and G. Gogolas, "Biomimetic dyes as affinity chromatography tools in enzyme purification", *J. Chromatogr. A*, 891 (2000) 33–44.
45. J. Porath, J. Carlsson, I. Olsson, and B. Belfrage, "Metal chelate affinity chromatography, a new approach to protein fraction", *Nature*, 258 (1975) 598–599.
46. H. Block, B. Maertens, A. Spriestersbach, N. Brinker, J. Kubicek, R. Fabis, J. Labahn, and F. Schäfer, "Immobilized-metal affinity chromatography (IMAC): a review", *Methods Enzymol.*, 463 (2009) 439–473.
47. D.S. Hage, "A survey of recent advances in analytical applications of immunoaffinity chromatography", *J. Chromatogr. B*, 715 (1998) 3–28.
48. A.C. Moser and D.S. Hage, "Immunoaffinity chromatography: an introduction to applications and recent developments", *Bioanalysis*, 2 (2010) 769–790.
49. W.J. Lough (Ed.), *Chiral Liquid Chromatography*, Blackie and Son, Glasgow, 1989.
50. S. Allenmark, *Chromatographic Enantioseparations: Methods and Applications*, 2nd Ed., Ellis Horwood, New York, 1991.
51. I.M. Chaiken (Ed.), *Analytical Affinity Chromatography*, CRC Press, Boca Raton, 1987.
52. J.E. Schiel, K.S. Joseph, and D.S. Hage, "Biointeraction affinity chromatography: general principles and recent developments", *Adv. Chromatogr.*, 48 (2009) 146–187.
53. G.W. Ewing (Ed.), *Analytical Instrumentation Handbook*, 2nd Ed., Marcel Dekker, New York, 1997.

67

MASS SPECTROSCOPY MEASUREMENTS OF NITROTYROSINE-CONTAINING PROTEINS

XIANQUAN ZHAN[1] AND DOMINIC M. DESIDERIO[2]

[1] *Key Laboratory of Cancer Proteomics of Chinese Ministry of Health, Xiangya Hospital, Central South University, Changsha, P. R. China*

[2] *The Charles B. Stout Neuroscience Mass Spectrometry Laboratory, Department of Neurology, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, USA*

67.1 INTRODUCTION

67.1.1 Formation, Chemical Properties, and Related Nomenclature of Tyrosine Nitration

Tyrosine nitration is a posttranslational modification (PTM) derived from oxidative/nitrative stress, which results from *in vivo* nitrating agents such as peroxynitrite ($ONOO^-$) and nitrogen dioxide [1–4]. Tyrosine nitration is the addition of a nitro group ($-NO_2$), an electron-withdrawing group, to position 3 of the phenolic ring of a tyrosine residue to form a 3-nitrotyrosine residue in a protein [3]. Nitration of a tyrosine residue significantly changes its physical and chemical properties [3, 5, 6]; for example, (i) the pK_a value of the phenolic hydroxyl group of nitrotyrosine (pK_a = ~7.1) is significantly decreased relative to that of tyrosine (pK_a = ~10), (ii) the electron density of the phenolic ring of nitrotyrosine is significantly decreased relative to that of tyrosine because the nitro group ($-NO_2$) is an electron-withdrawing group, and (iii) nitrotyrosine can be reduced to aminotyrosine. Based on nitration of a tyrosine residue, some related nomenclatures were derived: a nitropeptide is defined as a nitrotyrosine-containing peptide or a nitrated peptide; a nitroprotein is defined as a nitrotyrosine-containing protein or a

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

nitrated protein; a nitroproteome is defined as all nitroproteins in a proteome; and nitroproteomics is defined as the use of proteomics to study the nitroproteome.

67.1.2 Biological Roles of Tyrosine Nitration in a Protein

Tyrosine nitration in a protein in a biological system has extensive biological functions. First, because the nitro group ($-NO_2$) is an electron-withdrawing group, tyrosine nitration, namely, addition of nitro group to the phenolic ring of the tyrosine residue, decreases the electron density of the phenolic ring of a tyrosine residue in a protein [3, 7]. Nitration yields several direct biological consequences: (i) it shifts the phenolic pK_a value (from ~10 for tyrosine) into the physiological pH range (~7.1 for 3-nitrotyrosine) to affect chemical properties of a tyrosine residue [3, 6, 8] and (ii) if the tyrosine nitration occurs exactly within an interacting region between a receptor and ligand or between an enzyme and substrate, then the decreased electron density could negatively affect interaction intensity (receptor–ligand and enzyme–substrate) to hinder the functions of that protein [3, 7]. Second, nitration and phosphorylation would compete for the same tyrosine residue when the tyrosine residue is within a tyrosine phosphorylation motif ([R or K]-x2(3)-[D or E]-x3(2)-[Y]) to affect tyrosine phosphorylation signaling pathways and involve important biological processes [7, 9–11]. Third, some studies have demonstrated that tyrosine nitration in a protein in a biological system might be a reversible and dynamic process between nitration and denitration because of the discovery of a putative denitrase [8, 12, 13]. Therefore, tyrosine nitration is not only a pathological consequence, a marker of oxidative injuries, but also involve multiple biological processes such as neurotransmission and redox signaling [4]. Tyrosine nitration can alter protein functions and associated multiple physiological or pathological processes such as tumorigenesis, inflammatory and neurodegenerative diseases, modification of enzymatic activities, and immunogenicity [2, 3, 14–17].

67.1.3 Challenge and Strategies to Identify a Nitroprotein with Mass Spectrometry

In order to understand the biological functions and roles of tyrosine nitration in a protein, an essential step is to identify endogenous nitroproteins and accurately locate each nitrotyrosine site. Mass spectrometry (MS) is the key technique for these tasks. However, MS identification of endogenous nitroproteins and nitrotyrosine sites is severely challenged because of several factors: (i) endogenous nitroproteins occur with an extreme low abundance (1 in ~10^6 tyrosines), (ii) each MS instrument has its sensitivity limitation that requires a sufficient amount of samples for MS detection, and (iii) various MS behaviors are present in different types of MS analyses; for example, there is a characteristic photodecomposition pattern of a nitro group that is present in UV-laser-based matrix-assisted laser desorption ionization (MALDI)-MS analysis of a nitroprotein [18–20] but not for electrospray ionization (ESI)-MS [17, 20–23]. As a result, photodecomposition of a nitro group decreases signal intensities of a nitropeptide and complicates interpretation of a mass spectrum; in turn, that characteristic photodecomposition pattern can confirm the existence of a nitro group in a peptide [20].

Several strategies must be developed and used before the use of MS to identify endogenous nitroproteins and nitrotyrosine sites: (i) different chemical derivation techniques [20] are used to convert the nitro group (with various MS behaviors) to an amino group (with stable MS behavior) to resolve the varied MS behaviors of a nitro group and (ii) different enrichment techniques [3, 4, 24] are used to preferentially enrich endogenous nitropeptides or nitroproteins to overcome the extreme low abundance of endogenous nitropeptides/nitroproteins in a biological system and sensitivity limitations of a mass spectrometer. Currently, chemical derivation and targeted enrichment prior to an MS analysis [7, 25, 26] mainly include the following: (i) Nitrotyrosine antibody-based immunoaffinity is used to preferentially enrich nitropeptides [27] or nitroproteins [3, 28]. (ii) Conversion of a nitro group to an amino group is coupled with target enrichment [29]. Briefly, all amines are first acetylated, followed by conversion of nitrotyrosine to aminotyrosine and biotinylation of aminotyrosine. (iii) Conversion of a nitro group to an amino group is coupled with derivatization of the amino group [30]. Briefly, alpha- and epsilon-amino groups in a protein or peptide are protected with $^{13}C_0/^{13}C_4$- or D_0/D_6-acetic anhydride, nitrotyrosine is reduced to aminotyrosine with sodium dithionate (also known as sodium hydrosulfite), and aminotyrosine is derivatized with 1-(6-methyl[D_0/D_3] nicotinoyloxy) succinimide. (iv) The nitro group in a nitropeptide is reduced to an amino group and dansylated with dansyl chloride, followed by MS^n analysis [31, 32]. (v) The "light"- and "heavy"-labeled acetyl groups are used to block *N*-terminal and lysine residues of tryptic nitropeptides, followed by reduction of nitrotyrosine to aminotyrosine with sodium dithionite and derivatization of light- and heavy-labeled aminotyrosine peptides with either isobaric tags for relative and absolute quantification (iTRAQ) or tandem mass tags (TMT), respectively [33]. (vi) Selective chemoprecipitation and subsequent release of tagged species (conversion of nitro group to a small 4-formylbenzylamido tag) are used to analyze nitropeptides with liquid chromatography–tandem mass spectrometry (LC-MS/MS) [34]. (vii) iTRAQ quantitative reagents are used to selectively label nitrotyrosine residues (not primary amines) followed by MS analysis [35]. And (viii) use of combined fractional diagonal chromatography (COFRADIC) [36, 37] peptide sorting is based on a hydrophilic shift after reduction of the nitro group to its amino counterpart, followed by ESI-MS [36] and MALDI-MS [37] identification of a nitropeptide. Moreover, except for proteomics with antinitrotyrosine antibodies and gel-based separation, multidimensional chromatography, precursor-ion scanning, and/or chemical derivatization have also emerged to identify and quantify nitroproteins and nitrotyrosine sites [26, 38].

67.1.4 Biological Significance Measurement of Nitroproteins

MS measurement of nitroproteins must finally serve for real application in a biological system. In order to achieve that goal, several biological significance-related measurements of nitroproteins are worth further study: (i) Quantitative proteomics strategies such as iTRAQ-based quantification should be developed and used to quantify a nitroprotein specific to a pathological condition (or called quantitative nitroproteomics), and the degree of nitration should be quantified in a specific biological condition [33]. (ii) Bioinformatics should be developed and used to locate nitrotyrosine sites within corresponding protein domains and motifs [3, 10] in order to understand in depth the effects of tyrosine nitration on the structure and functions of a protein. (iii) Systems biology methods should be developed and used to elucidate protein-network systems that are involved in nitroproteins [7, 39] for clarification of effects of tyrosine nitration on protein-network systems in a biological condition. (iv) Structural biology should be used to reveal three-dimensional (3D) crystal structure and local primary structure that occur with tyrosine nitration of every biologically important nitroprotein to address the impact of tyrosine nitration on that protein's functions to develop a drug against tyrosine nitration [7, 40, 41]. And (v) body fluids such as serum and cerebrospinal fluid (CSF) are important window to predict and diagnose a disease; body fluid nitroproteomics and nitropeptidomics should be developed and used to discover body fluid biomarkers for prediction, diagnosis, and prognosis of a disease [7, 42].

67.2 MASS SPECTROMETRIC CHARACTERISTICS OF NITROPEPTIDES

67.2.1 MALDI-MS Spectral Characteristics of a Nitropeptide

For UV-laser MALDI-MS analysis of a nitropeptide, high-energy UV-laser light (337 nm) can induce photochemical decomposition of nitro group ($-NO_2$) to yield a characteristic photochemical decomposition pattern ($[M+H]^+$, $[M+H-16]^+$, $[M+H-30]^+$, and $[M+H-32]^+$) in an MS spectrum of a nitropeptide [18, 19, 23]. This photochemical decomposition will decrease the intensity of precursor ion from a nitropeptide and complicate its MS spectrum [20]; meanwhile, recognition of this photochemical decomposition pattern will assist in the interpretation of MS data of a nitropeptide [4, 18–20]. Figure 67.1 shows the formation of a nitrotyrosine and its likely products derived from photochemical decomposition [43].

The nitrotyrosine UV-induced photodecomposition pattern induced by MALDI UV laser has been extensively confirmed in several experiments with a synthetic nitropeptide [A-A-F-G-Y($-NO_2$)-A-R; $[M+H]^+=800.4$] [19], tetranitromethane (TNM)-nitrated bovine serum albumin (BSA) [19], TNM-nitrated angiotensin II ($[M+H]^+$, *m/z* 1092.5)

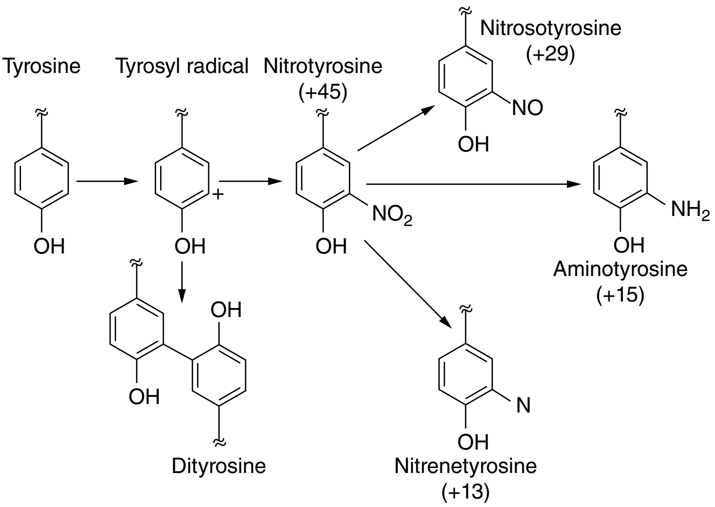


FIGURE 67.1 Formation of dityrosine and nitrotyrosine and photochemical decomposition products of a nitrotyrosine. Source: Turko and Murad [43]. Reproduced with permission of Elsevier.

[18], synthetic leucine enkephalin [LE1: Y-G-G-F-L; molecular weight (MW) = 555.1818 Da], nitro-Tyr-leucine enkephalin [LE2: Y(—NO_2)-G-G-F-L, MW = 600.0909 Da], and d_5-Phe-nitro-Tyr-leucine enkephalin [LE3: Y(—NO_2)-G-G-F(d_5)-L, MW = 605.1818 Da] [20], respectively. Here, UV-laser MALDI-MS analysis of synthetic nitropeptide A-A-F-G-Y(—NO_2)-A-R ($[M+H]^+$ = 800.4) [19] is taken as an example. The MS spectrum shows a typical ion pattern ($[M+H]^+$, $[M+H-16]^+$, $[M+H-14]^+$, $[M+H-32]^+$, and $[M+H-30]^+$) that corresponded to *m/z* 800.4, 784.4, 786.4, 768.4, and 770.4, respectively (Fig. 67.2), which results from photochemical decomposition of the nitro group in the $[M+H]^+$ ion = 800.4. The $[M+H]^+$ ion (*m/z* 800.4) represents the nitrotyrosine (Tyr-NO_2)-containing peptide; $[M+H-16]^+$ ion (*m/z* 784.4) represents nitrosotyrosine (Tyr-NO)-containing peptide after loss of an oxygen atom from the nitro group; $[M+H-14]^+$ ion (*m/z* 786.4) represents hydroxylaminotyrosine (Tyr-NHOH)-containing peptide after reduction of the nitroso (Tyr-NO) group; $[M+H-32]^+$ ion (*m/z* 768.4) represents triplet nitrene-tyrosine (Tyr-N)-containing peptide after loss of two oxygen atoms from the nitro group; and $[M+H-30]^+$ ion (*m/z* 770.4) represents aminotyrosine (Tyr-NH_2)-containing peptide after reduction of the triplet nitrene (Tyr-N) group. Also, intensities of photochemical decomposition products $[M+H-32]^+$ and $[M+H-30]^+$ ions (Tyr-N and Tyr-NH_2) are much lower than those of the protonated molecule ion $[M+H]^+$ (Tyr-NO_2) and photochemical decomposition products $[M+H-16]^+$ and $[M+H-14]^+$ ions (Tyr-NO and Tyr-NHOH).

Furthermore, MALDI UV-laser-induced photochemical decomposition significantly consumes the protonated molecule ion $[M+H]^+$ (Tyr-NO_2) of a nitropeptide to impact detection of endogenous low-abundance nitropeptides/nitroproteins. Figure 67.3 clearly

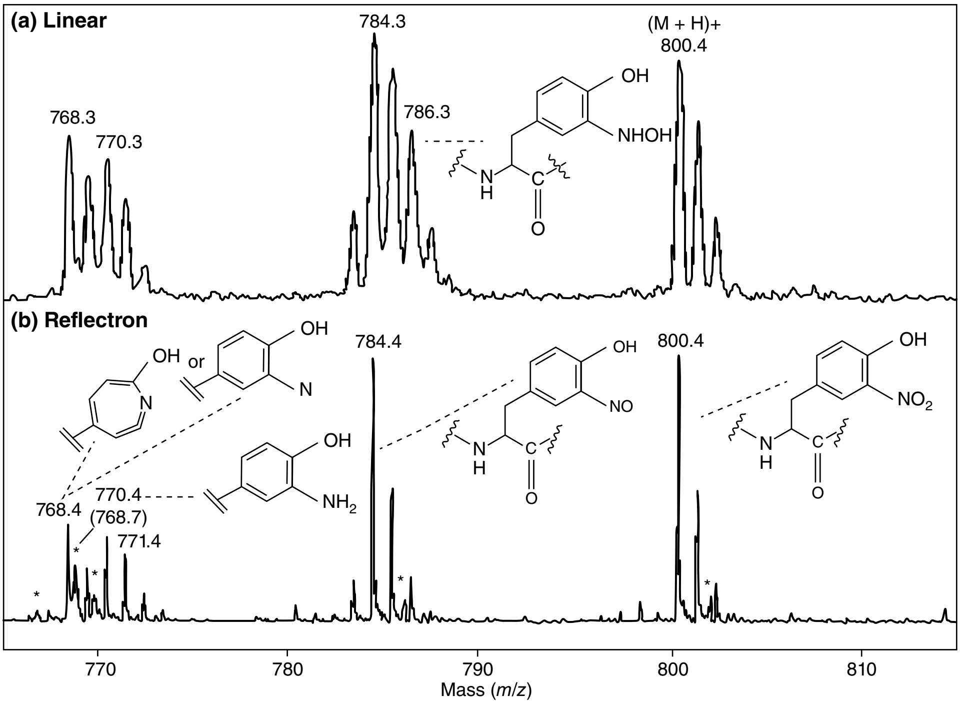


FIGURE 67.2 Photochemical decomposition pattern of synthetic nitropeptide AAFGY(—NO_2) AR in a UV-laser MALDI-TOF spectrum in (a) linear mode and (b) reflectron mode. The structures of 3-nitrotyrosine and proposed photochemical decomposition products are shown in the corresponding ions. Several small ions (asterisk) might represent metastable peaks. A weak increase in the abundance of the ion at *m/z* 771.4 over what would be expected for the ^{13}C isotope peak for the aminotyrosine products at *m/z* 770.4 in the linear and reflectron spectra suggests that a small amount of a catechol product might have formed as well. Source: Sarver et al. [19]. Reproduced with permission of Springer.

demonstrates that the peak intensity of $[M+H]^+$ ion of leucine enkephalin (LE1, NL = 1.01E5) was much higher than that of nitro-Tyr leucine enkephalin (LE2, NL = 3.25E4) and *d*(5)-Phe-nitro-Tyr leucine enkephalin (LE3, NL=9.09E4) and that the MS spectrum of nitropeptide (LE2 and LE3) is much more complicated relative to the nonnitrated peptide (LE1) [20].

However, one should note that infrared light-MALDI-Fourier transform ion cyclotron resonance MS (IR-MALDI-FT-ICR-MS) does not fragment the $[M+H]^+$ of a nitropeptide to produce an efficient approach to identify protein nitration [44]. The exact reason still remains unknown why MALDI with laser light (337 nm) induces photochemical decompositions of a nitro (—NO_2) group in a nitropeptide (but not with infrared light) and the structures of those photodecomposition products [19]. High-energy UV-laser light (337 nm) might induce loss of one (−16 mass units) or two (−32 mass units) oxygen atoms of the nitro group of a nitropeptide [19].

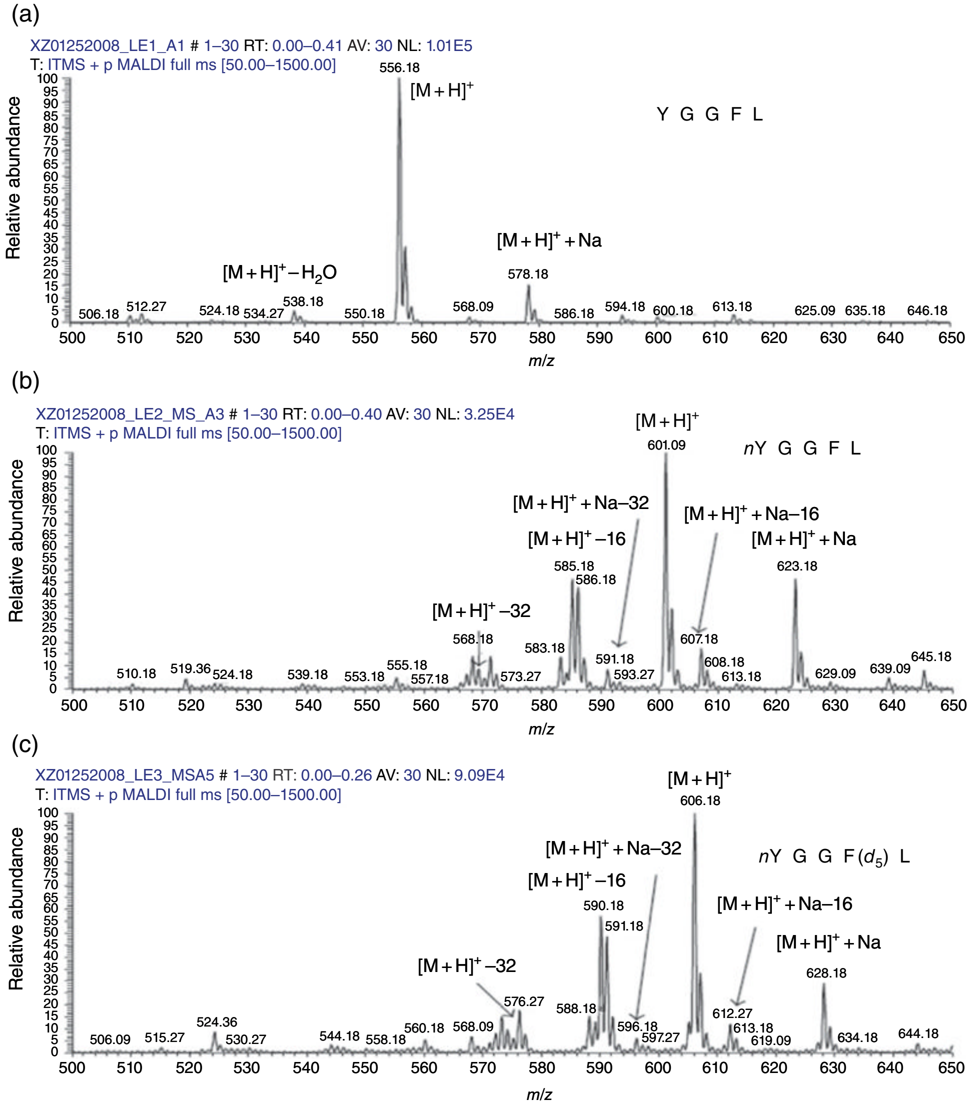


FIGURE 67.3 UV-laser MALDI-MS spectra of LE1 (a), LE2 (b), and LE3 (c). nY = Nitrotyrosine residue. F(d_5) = Phe residue with five ^{2}H (*d*) atoms. Source: Zhan and Desiderio [20]. Reproduced with permission of Elsevier.

67.2.2 ESI-MS Spectral Characteristics of a Nitropeptide

Compared to the UV-laser MALDI-MS spectrum of a nitropeptide, an ESI-MS spectrum of a nitropeptide does not show decomposition of a nitro group [17–23, 45]. However, an ESI-MS/MS spectrum shows a characteristic immonium ion (*m/z* 181.06) that is derived from a nitrotyrosine residue to indicate the presence of a nitrotyrosine

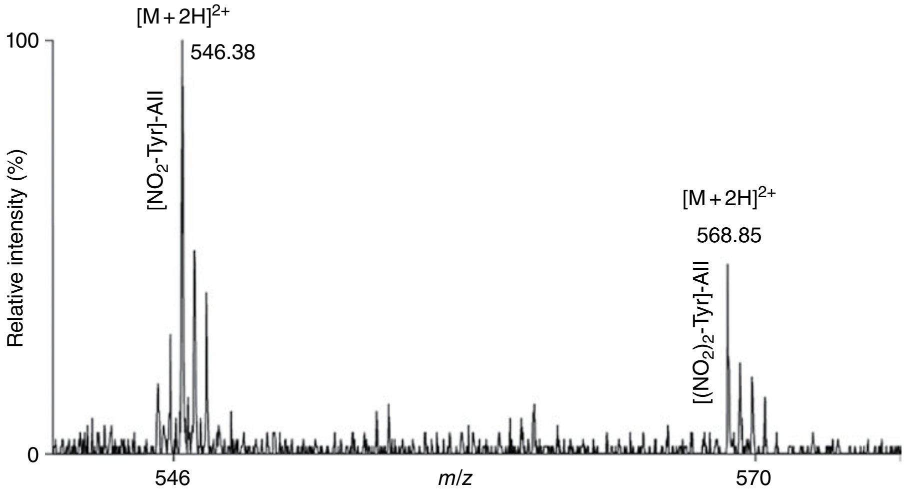


FIGURE 67.4 The ESI-MS spectrum of nitrated angiotensin II to show mononitrated and dinitrated angiotensin II. Source: Petersson et al. [18]. Reproduced with permission of John Wiley & Sons, Inc.

residue. Moreover, precursor-ion scanning based on an immonium ion at *m/z* 181.06 for nitrotyrosine will accurately identify a nitropeptide/nitroprotein [18].

ESI-MS and ESI-MS/MS spectral characteristics of a nitropeptide and precursor-ion scans for an immonium ion at *m/z* 181.06 have been confirmed with TNM-nitrated angiotensin II [D-R-V-Y(—NO_2)-I-H-P-F; MW = 1090.76 Da] and TNM-nitrated BSA [18]. ESI-MS analysis of TNM-nitrated angiotensin II [D-R-V-Y(—NO_2)-I-H-P-F; MW = 1090.76 Da] is taken as an example. First, no chemical composition pattern of a nitro group was found in an ESI-MS spectrum, except for a mononitrated ion ($[M + 2H]^{2+}$ *m/z* 546.38) that represents [NO_2-Tyr]-angiotensin II and a dinitrated ion ($[M + 2H]^{2+}$ *m/z* 568.85) that represents [$(NO_2)_2$-Tyr]-angiotensin II (Fig. 67.4). Second, ESI-MS/MS with collision-induced dissociation (CID) fragmentation was used to analyze doubly charged precursor ions of mononitrated angiotensin II at *m/z* 546.38 and dinitrated angiotensin II at *m/z* 568.85; characteristic immonium ions occurred at *m/z* 181.06 (mononitrated tyrosine) and at *m/z* 226.0 (dinitrated tyrosine) in those ESI-MS/MS spectra (Fig. 67.5). Third, precursor-ion scan spectra based on characteristic immonium ions at *m/z* 181.06 (mononitrated tyrosine) and *m/z* 226.0 (dinitrated tyrosine) accurately identified a nitropeptide in a complicated sample (Fig. 67.6).

67.2.3 Optimum Collision Energy for Ion Fragmentation and Detection Sensitivity for a Nitropeptide

The use of UV-laser vMALDI-MS/MS with CID to analyze synthetic peptides LE1 (Y-G-G-F-L; 555.1818 Da), LE2 [(3-NO_2)Y-G-G-F-L; 600.0909 Da], and LE3 [(3-NO_2)Y-G-G-(d_5)F-L; 605.1818 Da] found that, first, b- and a-ions were the most

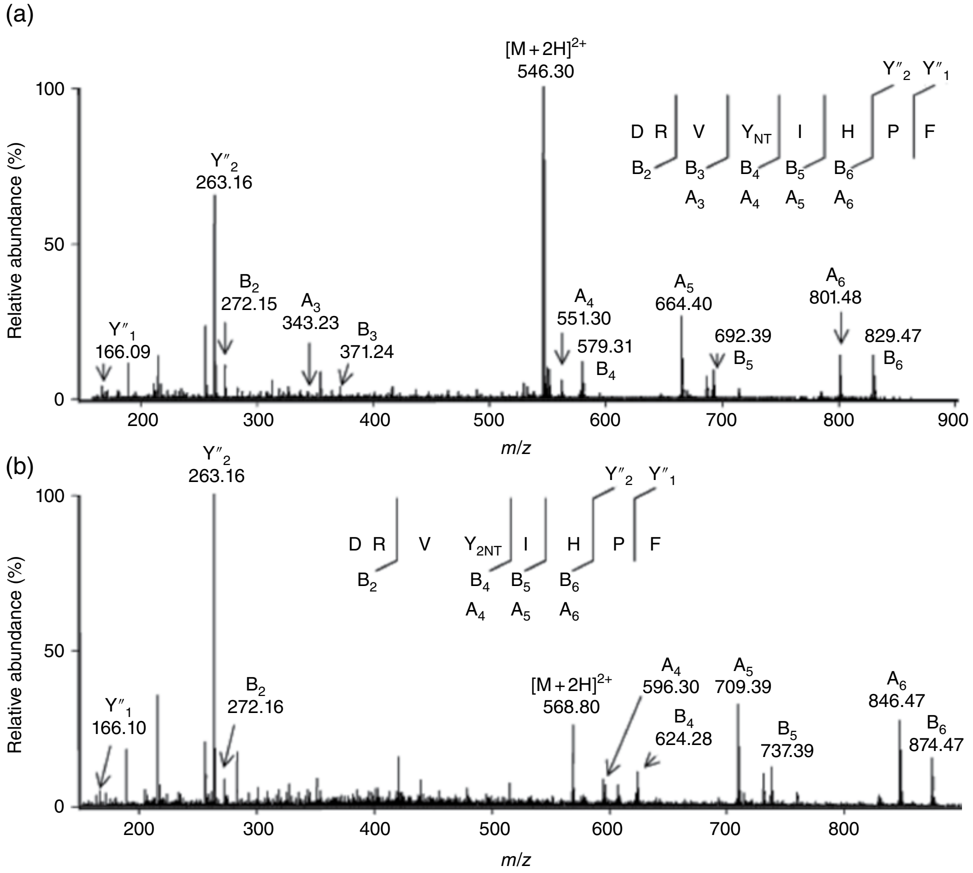


FIGURE 67.5 The MS/MS spectra of mononitrated angiotensin II peptide (precursor ion $[M+2H]^{2+}$ at *m/z* 546.30) (a) and dinitrated angiotensin II peptide (precursor ion $[M+2H]^{2+}$ at *m/z* 568.80) (b). Source: Petersson et al. [18]. Reproduced with permission of John Wiley & Sons, Inc.

intense fragment ions relative to y-ions (Fig. 67.7) [20]; those data were confirmed with UV-laser MALDI-MS/MS analysis of nitrated angiotensin II [18]. Second, relative to unmodified peptide (LE1), more collision energy optimized ion fragmentation of the nitropeptide (Fig. 67.8a) but increased intensity of the a_4-ion and decreased intensity of the b_4-ion (a-ion = loss of CO from a b-ion) (Fig. 67.8b). Third, optimized UV-laser fluence maximized ion fragmentation of the nitropeptide. Fourth, MS^3 analysis confirmed the MS^2-derived amino acid sequence; however, MS^3 analysis required a greater amount of peptides relative to MS^2 [20]. Thus, MS^3 analysis might not be suitable for routine analysis of endogenous low-abundance nitroproteins. Only when a target is determined can MS^3 be used for confirmation. Fifth, to detect a nitropeptide, the amount of peptide must satisfy the sensitivity of a mass spectrometer; for synthetic nitropeptides, the sensitivity of vMALDI-LTQ was 1 fmol for MS detection and 10 fmol for MS^2 detection [20].

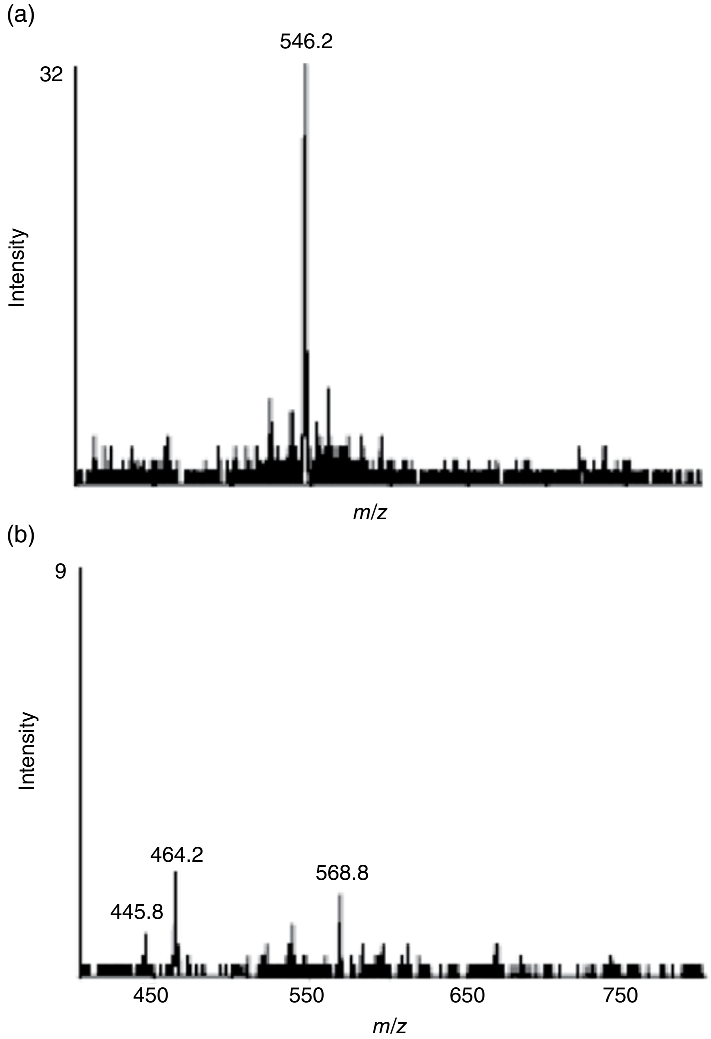


FIGURE 67.6 Precursor-ion scans spectra of nitrated angiotensin II based on immonium ion at *m/z* 181.06 for mononitrated tyrosine (a) and at *m/z* 226.0 for dinitrated tyrosine (b). Source: Petersson et al. [18]. Reproduced with permission of John Wiley & Sons, Inc.

67.2.4 MS/MS Spectral Characteristics of a Nitropeptide under Different Ion-Fragmentation Models

For MS/MS analysis of a nitropeptide, ion-fragmentation behaviors differ significantly among CID-, electron-capture dissociation (ECD)-, electron-transfer dissociation (ETD)-, and metastable atom-activated dissociation (MAD)-MS [46, 47]:

i. CID behavior of a nitropeptide. Studies demonstrated that the presence of nitration did not affect the CID behavior of the peptides.

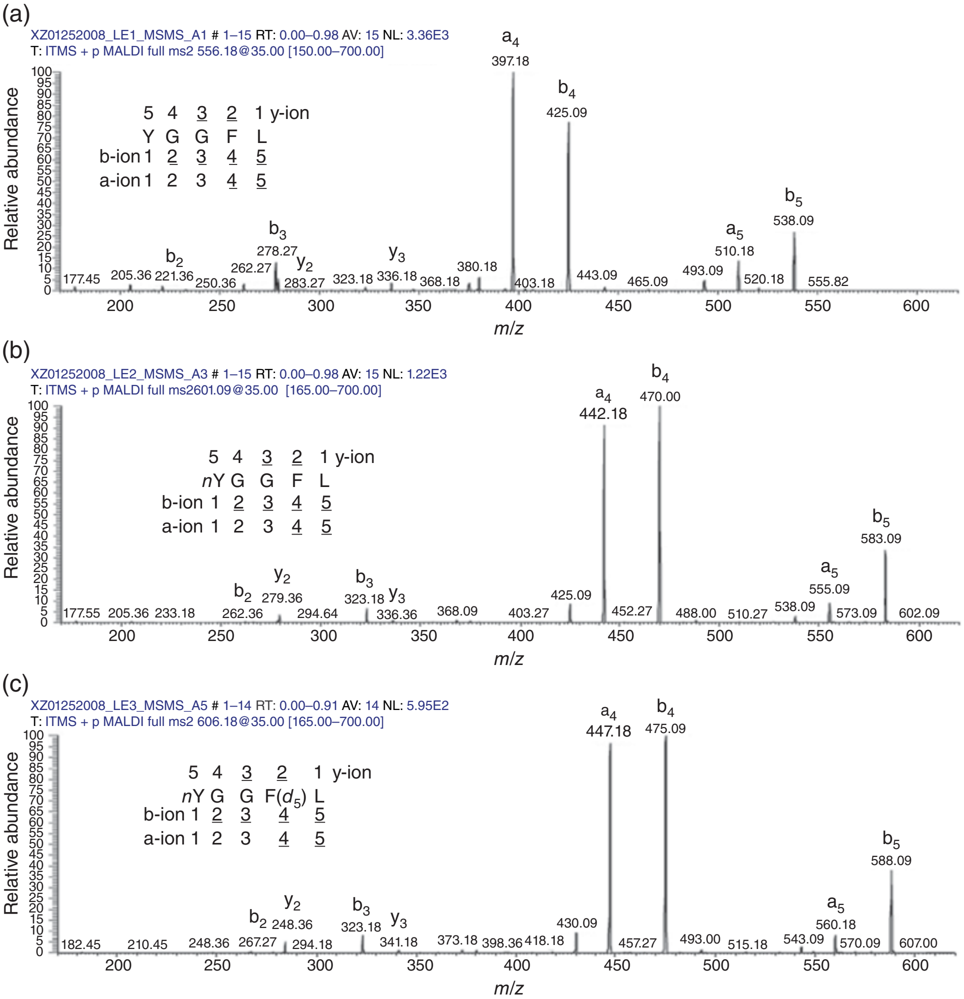


FIGURE 67.7 MS/MS spectra of LE1 (a), LE2 (b), and LE3 (c). nY = Nitrotyrosine residue. F(d_5) = Phe residue with five ^{2}H (*d*) atoms. Source: Zhan and Desiderio [20]. Reproduced with permission of Elsevier.

ii. ECD behavior of a nitropeptide. For doubly charged peptides, production of ECD sequence fragments was severely inhibited with nitration; ECD of triply charged nitropeptides produced some singly charged sequence fragments. ECD of nitropeptides was characterized with multiple losses of small neutral species, including hydroxyl radicals, water, and ammonia. The origin of neutral losses was investigated with activated ion (AI) ECD. Loss of ammonia appears to be the result of noncovalent interactions between a nitro group and protonated lysine side chains [47, 48].

iii. MAD behavior of a nitropeptide. Some studies found that high kinetic energy helium MAD produced extensive backbone fragmentation with significant

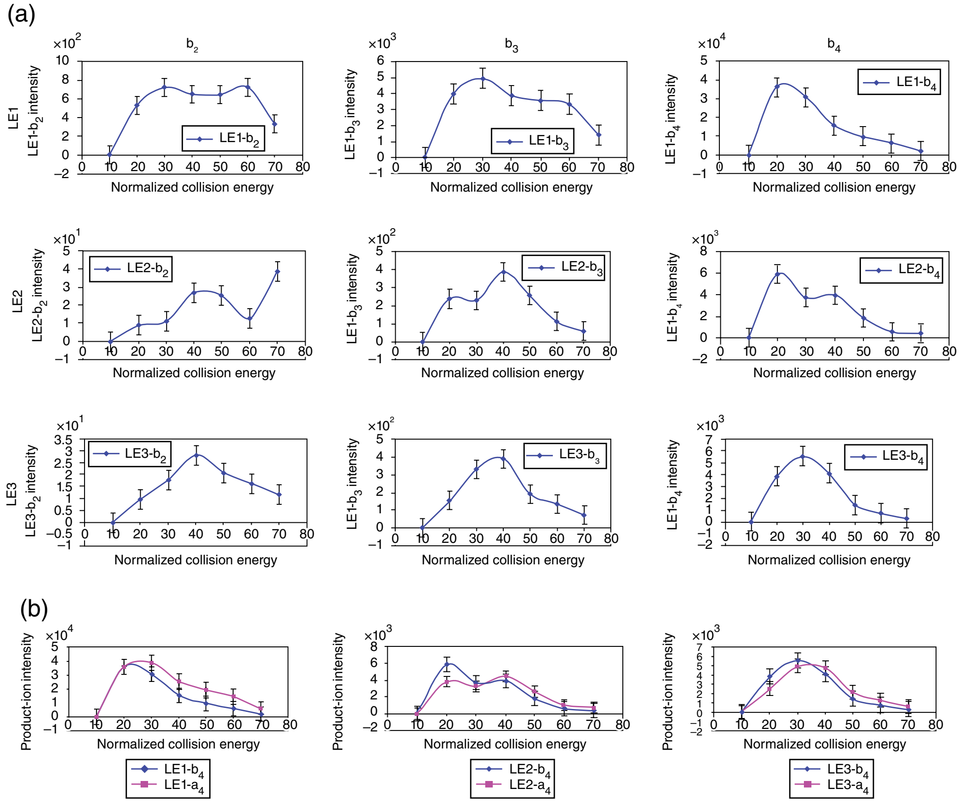


FIGURE 67.8 Effect of collision energy on collision-induced dissociation (CID) fragmentation of nitropeptides. (a) Relationship between collision energy and product-ion intensity ($n=3$). (b) Relationship between collision energy and product ion b_4 and a_4 intensities ($n=3$). Source: Zhan and Desiderio [20]. Reproduced with permission of Elsevier.

retention of PTMs. Although the high electron affinity of a nitrotyrosine moiety quenched radical chemistry and fragmentation in ECD and ETD, MAD does produce numerous backbone cleavages in the vicinity of the nitration. Compared to CID, MAD produced more fragment ions and differentiated I/L residues in nitrated peptides. MAD induced radical ion chemistry even in the presence of strong radical traps and, therefore, offers unique advantages to ECD, ETD, and CID for determination of nitropeptides [46].

iv. The different types of CID-MS/MS have different abilities to identify nitroproteins [49]. For example, for the same samples, 119 nitropeptides and 23 multiply nitrated nitropeptides were studied with a QSTAR Elite (QTOF) with CID, whereas 197 nitropeptides and 36 multiply nitrated nitropeptides were studied with a dual-pressure ion-trap mass spectrometer (LTQ Velos) with CID (Fig. 67.9) [49]. Therefore, it is essential to choose an appropriate mass spectrometer to analyze nitropeptides/nitroproteins.

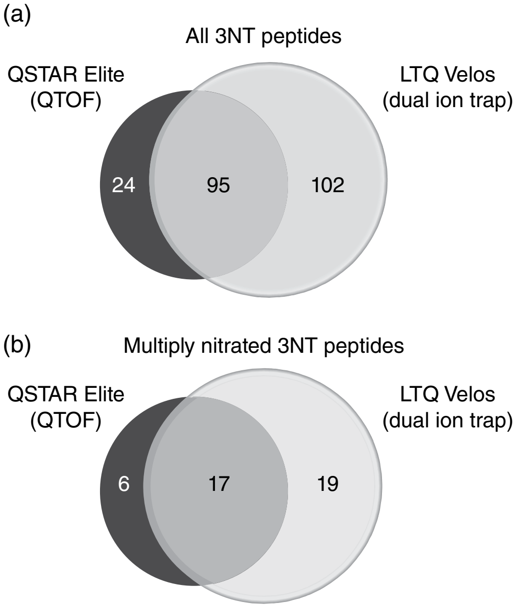


FIGURE 67.9 Overlap analysis of validated 3-nitrotyrosine (3NT)-containing peptides identified with a QSTAR Elite and LTQ Velos. Venn diagram of the overlap of all validated 3NT (a) and multiple 3NT-modified (b) peptides identified with a QSTAR Elite and LTQ Velos. Source: Li et al. [49]. Reproduced with permission of Elsevier.

67.3 MS MEASUREMENT OF *IN VITRO* SYNTHETIC NITROPROTEINS

67.3.1 Importance of Measurement of *In Vitro* Synthetic Nitroproteins

Detection of a nitropeptide is affected by the particular MS characteristics of nitropeptides, the extreme low abundance of *in vivo* nitrotyrosine sites, and MS sensitivity limitations. Moreover, it is much easier to obtain or produce *in vitro* synthetic or nitrated nitropeptides/nitroproteins relative to *in vivo* nitrotyrosine sites. Therefore, it is necessary to use *in vitro* synthetic nitroproteins to develop and establish methods to analyze *in vitro* nitroproteins.

67.3.2 Commonly Used *In Vitro* Nitroproteins and Their Preparation

Some synthetic peptides such as AAFGY(—NO_2)AR [19], leucine enkephalin, nitro-Tyr-leucine enkephalin, and d_5-Phe-nitro-Tyr- leucine enkephalin [20] have been used to study MS characteristics of a nitropeptide. Angiotensin II, BSA, and ovalbumin (OVA) are commonly used standard peptides and proteins that are *in vitro* nitrated with liquid TNM [5, 18, 19, 38, 50], gaseous nitrogen dioxide and ozone (NO_2+O_3) [38], or peroxynitrite [51]. In order to simulate *in vivo* proteome situations, some proteomes such as human plasma were also nitrated *in vitro* with TNM followed by nitroproteomic analysis [34].

67.3.3 Methods Used to Measure *In Vitro* Synthetic Nitroproteins

Several well-established nitroproteomic methods based on anti-3-nitrotyrosine antibody and gel-based separations have been used to study *in vitro* prepared standard nitropeptides, nitroproteins, and nitroproteome samples [18, 26]. Methods that involved multidimensional chromatography, diagonal chromatography [36, 37], precursor-ion scanning [18], and/or chemical derivation can identify and quantify protein nitration sites [26]. Of them, chemical derivation is an important step to establish those methods.

Several chemical derivatization methods have been developed to analyze nitropeptide/nitroprotein prior to MS analysis [25]. Because an amino (—NH_2) group in a nitroprotein/nitropeptide is more stable than a nitro (—NO_2) group during MS analysis, all chemical derivation methods of nitrotyrosine reduce nitrotyrosine to aminotyrosine with reducing agents, including $Na_2S_2O_4$ [5, 19, 52], and derive the generated amino group with specific reagents. Those chemical derivatization methods are presented as follows:

i. A nitrotyrosine residue converted to an aminotyrosine residue *via* reduction can readily discern aminotyrosine peptides in a background of nonnitrated peptides. Thus, aminotyrosine peptides were more stable in a single MS mode and led to easy-to-interpret peptide mass maps [51].

ii. Dansyl chloride was used to label nitration sites. MS/MS and a precursor-ion scan identified the proteins and determined the nitrotyrosine sites [31, 32].

iii. A method that specifically enriches nitropeptides to unambiguously identify nitrotyrosine peptides and nitration sites with LC-MS/MS was used to follow conversion of nitrotyrosine to *N*-thioacetyl-aminotyrosine followed with high-efficiency enrichment of sulfhydryl-containing peptides with thiopropyl sepharose beads [24]. Briefly, the derivatization protocol includes the following: (a) all primary amines were acetylated with acetic anhydride to block those primary amines, (b) nitrotyrosine was reduced to aminotyrosine, (c) aminotyrosine was derivatized with *N*-succinimidyl *S*-acetylthioacetate, and (d) *S*-acetyl on *S*-acetylthioacetate was deprotected to form free sulfhydryl groups [24]. This method has been used to analyze *in vitro* nitrated human histone H1.2, BSA, and mouse brain tissue samples [24].

iv. Although iTRAQ is an effective quantitative proteomics method, it is limited to primary amines. A new strategy based on the use of iTRAQ reagents coupled with MS analysis was developed to selectively label nitrotyrosine residues [35] to simultaneously localize and quantify nitration sites in model proteins and biological systems [35].

v. A strategy that combined precursor isotopic labeling and isobaric tagging (cPILOT) was developed to increase the multiplexing capability to quantify a nitrotyrosine protein to 12 or 16 samples with TMT or iTRAQ, respectively. Light- and heavy-labeled acetyl groups were used to block *N*-terminal and lysine residues of tryptic peptides. Nitrotyrosine was reduced to aminotyrosine with sodium dithionite, and light- and heavy-labeled aminotyrosine peptides were derivatized with either

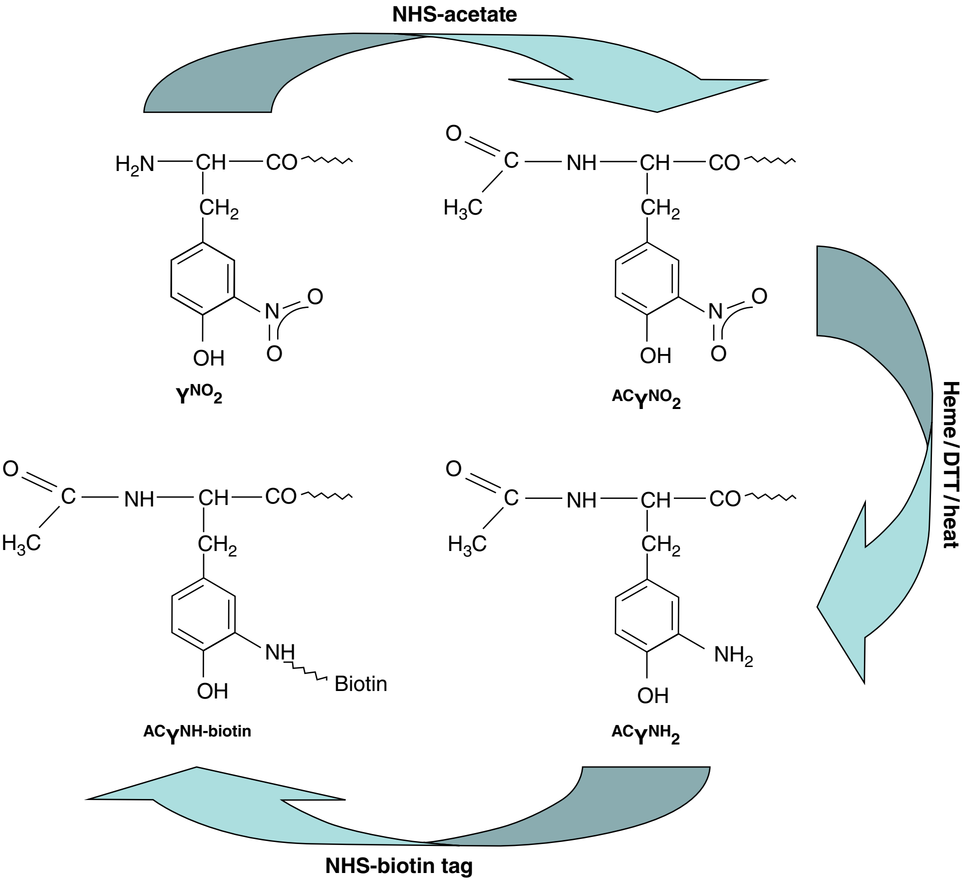


FIGURE 67.10 Reaction scheme of the chemical-labeling method as exemplified with an *N*-terminal nitrotyrosine residue. All amines were blocked with acetylation with acetic acid *N*-hydroxysuccinimide ester (NHS acetate). Nitrotyrosine was reduced to aminotyrosine with heme and DL-dithiothreitol in a boiling-water bath. The reaction sequence was completed with biotinylation of aminotyrosine with NHS-biotin. Source: Abello et al. [29]. Reproduced with permission of Elsevier.

TMT or iTRAQ multiplex reagents [33]. This method demonstrated proof of principle to analyze *in vitro* nitrated BSA and mouse splenic proteins [33].

vi. An improved chemical-labeling method was designed to enrich nitropeptides independent of sequence context (Fig. 67.10). Briefly, all amines were blocked with acetylation, and nitrotyrosine was converted to aminotyrosine, followed by biotinylation of aminotyrosine [29]. Moreover, the entire reaction was carried out in a single buffer without any sample cleanup or pH changes to minimize any sample loss. Also, a strong cation exchanger was used to remove free biotin, and an immobilized avidin column was used to enrich the labeled peptides, followed by analysis of enriched peptides with LC-MS/MS [29]. This method was approved for *in vitro* nitrated samples [29, 53].

vii. Because a MALDI UV-laser causes photochemical decomposition of a nitro group in a nitropeptide, a new strategy includes (a) acetylation of *N*-terminal amines and epsilon amines of lysine residues with acetic anhydride, (b) reduction of nitrotyrosine to aminotyrosine with sodium hydrosulfite, and (c) derivatization of aminotyrosine with 1-(6-methyl[d_0/d_3]nicotinoyloxy) succinimide, followed by MALDI-TOF MS analysis [30]. The optimum matrix was sinapinic acid, not 2,5-dihydroxybenzoic acid, for MALDI-MS analysis of a nitropeptide [54].

viii. Another developed method is the COFRADIC approach [36, 37]. Nitrotyrosine is reduced to aminotyrosine with sodium dithionite and peptides sorted with reversed-phase chromatography based on a hydrophilic shift from nitropeptide (more hydrophilic) to aminotyrosine-containing peptide (more hydrophobic) followed by EDI-MS identification [36] and MALDI-MS [37]. COFRADIC characterized tyrosine nitration in a TNM-nitrated BSA and peroxynitrite-nitrated proteome of human Jurkat cells [36, 37].

Interpretation of MS and MS/MS data of nitropeptides (and especially endogenous) is very challenging. To avoid any risk of linking MS/MS spectra to an incorrect amino acid sequence, the combination of reduction of nitrotyrosine to aminotyrosine and use of the Peptizer algorithm to inspect MS/MS quality-related assumptions [55] has been developed.

The optimal approach to determine the amino acid sequence of an endogenous nitropeptide is a manual approach [3].

67.4 MS MEASUREMENT OF *IN VIVO* NITROPROTEINS

67.4.1 Importance of Isolation and Enrichment of *In Vivo* Nitroprotein/Nitropeptide Prior to MS Analysis

Nitration of tyrosine residue in a protein mainly due to oxidative stress is a low-abundance (1 in ~10^6 tyrosines) modification in an *in vivo* proteome [14, 56]. MS is the key technique to identify nitropeptides/nitroproteins and to accurately determine nitrotyrosine sites in a nitroprotein [2, 3, 57]. MS sensitivity is the high-femtomole/low-picomole level [20]. Therefore, it is essential to isolate and preferentially enrich nitroproteins/nitropeptides before MS analysis [17, 21, 22, 57].

67.4.2 Methods Used to Isolate and Enrich *In Vivo* Nitroproteins/Nitropeptides

Several enrichment protocols have been used for isolation and preferential enrichment of *in vivo* nitropeptides/nitroproteins from a biological proteome: (i) Two-dimensional gel electrophoresis (2DGE) coupled with nitrotyrosine Western blotting was used to

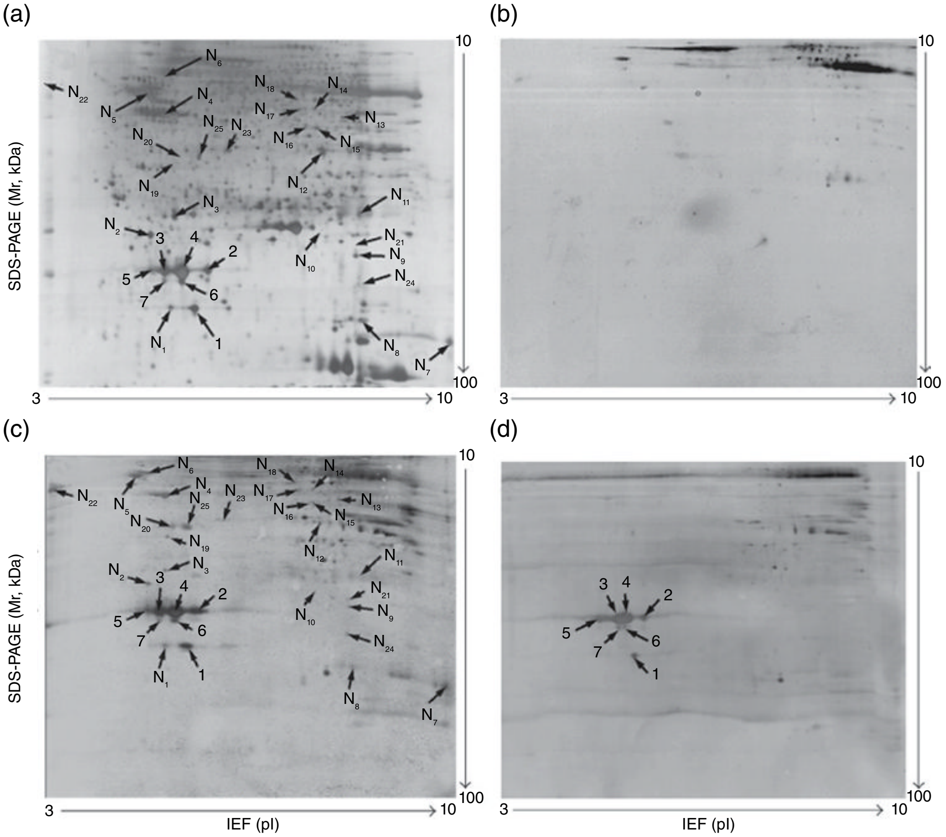


FIGURE 67.11 Two-dimensional Western blotting analysis of anti-3-nitrotyrosine-positive proteins in a human pituitary (70 μg protein per 2D gel). (a) Silver-stained image on a 2D gel before transfer of proteins onto a PVDF membrane. (b) Silver-stained image on a 2D gel after transfer of proteins onto a PVDF membrane. (c) Western blot image of anti-3-nitrotyrosine-positive proteins (anti-3-nitrotyrosine antibodies + secondary antibody). (d) Negative control of a Western blot to show the cross-reaction of the secondary antibody (only the secondary antibody; no anti-3-nitrotyrosine antibody). Source: Zhan and Desiderio [57]. Reproduced with permission of Elsevier.

analyze nitroproteins in a proteome (Fig. 67.11). Briefly, the nitroproteins in a proteome were arrayed and relatively enriched with 2DGE, transferred to PVDF membrane, and detected with antinitrotyrosine antibody, followed by visualization [2, 7, 16, 57–59]. (ii) Nitrotyrosine affinity column (NTAC) (Fig. 67.12) was used to enrich nitroproteins [3, 7, 60] and to enrich nitropeptides [61]. (iii) A method was used to acetylate all primary amines in a nitropeptide, convert easily a nitrotyrosine residue into aminotyrosine, and then enrich with biotinylation of an aminotyrosine (Fig. 67.10) [5, 25, 29]. (iv) A method was used to acetylate all primary amines, reduce nitrotyrosine to aminotyrosine, derivatizate aminotyrosine into a free sulfhydryl group, and then enrich sulfhydryl-containing

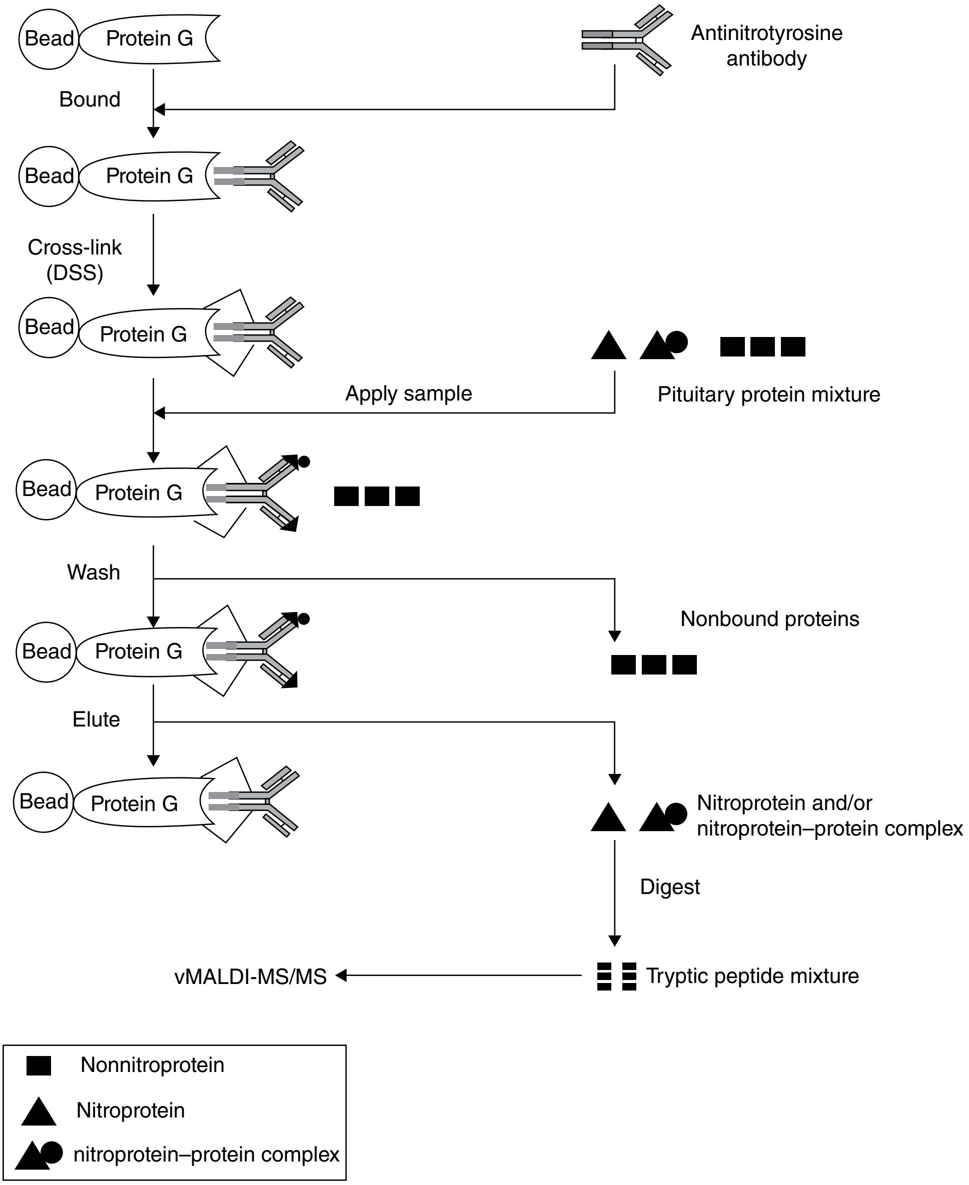


FIGURE 67.12 Experimental flowchart to identify nitroproteins and nitroprotein–protein complexes with NTAC-based MALDI-LTQ MS/MS. The control experiment (without any anti-3-nitrotyrosine antibody) was carried out in parallel with the NTAC-based experiments. Source: Zhan and Desiderio [3]. Reproduced with permission of Elsevier. Zhan, Wang, and Desiderio [7]. CC-BY.

peptides with thiopropyl sepharose beads [24]. (v) A method was utilized for the use of dansyl chloride to label nitration sites, followed with a precursor-ion scan and MS^3 analysis [31, 32]. (vi) A method was utilized for the use of a new tagging reagent, (3R, 4S)-1-(4-(aminomethyl) phenylsulfonyl)pyrrolidine-3,4-diol (APPD), for the selective fluorogenic derivatization of nitrotyrosine residues in peptides (after reduction to aminotyrosine) and boronate affinity enrichment [62]. (vii) COFRADIC was used to sort peptides per the hydrophilic shift after a nitro group was reduced to an amino group and then analyzed with ESI or MALDI-MS [36]. (viii) TMT- or iTRAQ-based quantitative nitroproteomics was used to quantitatively identify nitroproteins/nitropeptides [33, 35]. Briefly, the *N*-terminal and lysine residues of tryptic peptides were blocked with "light-" and "heavy-"labeled acetyl groups, and nitrotyrosine was reduced to aminotyrosine, followed by derivatization of light- and heavy-labeled aminotyrosine-containing peptides with either TMT or iTRAQ multiplex reagents [33, 35]. That method can relatively enrich and quantitatively identify nitroproteins/nitropeptides.

One should note that protocols (i), (ii), and (vii) were used in the identity of *endogenous* nitrotyrosine sites [2, 3, 36, 57, 60]; whereas protocols (iii)–(vii) succeeded mainly with an *in vitro* model peptide or protein and with an *in vitro* nitrated proteome [5, 24, 31, 36, 37]; however, they offer promise to study *in vivo* nitroproteins. Protocols (i)–(vii) mainly focused on the identity of nitropeptides, nitroproteins, and nitrotyrosine sites. However, in order to determine disease-related nitroproteins, it needs to quantitatively identify nitroproteins except for characterization of nitrotyrosine sites and nitroprotein. Protocol (viii) holds promise to achieve that goal because it can relatively enrich nitropeptides, identify nitrotyrosine sites, and quantify nitroproteins; and its sample-multiplexing capabilities has also been enhanced [33, 35].

67.5 MS MEASUREMENT OF *IN VIVO* NITROPROTEINS IN DIFFERENT PATHOLOGICAL CONDITIONS

Tyrosine nitration in a protein is an important oxidative/nitrative stress-mediated modification, which functions in a wide range of cellular, physiological, and pathological processes [63, 64]. Tyrosine nitration can alter activity of a protein and extensively associate pathophysiological conditions. The documented literature demonstrates that endogenous nitroproteins and nitrotyrosine sites have been identified in different types of pathophysiological conditions (Table 67.1) and was summarized here:

i. Tyrosine nitration in inflammation-related diseases. Studies have demonstrated that tyrosine nitration is extensively associated with inflammatory diseases with identification of nitroproteins in a septic patient's rectus abdominis muscle [65], bronchial epithelial cells, and bronchoalveolar lavage with asthma [66], Chagas' disease [67], experimental sepsis [68], and serum sample of a C57BL6/J mouse model with septic shock [36].

ii. Tyrosine nitration in aging and aging-related diseases. Protein nitration is extensively studied in aging and aging-related diseases with discovery of nitroproteins in aging rat heart [69], rat skeletal muscle [70, 71], and mouse liver [72].

TABLE 67.1 Endogenous Nitroproteins Identified from Different Pathological Conditions

Reference	Specimen	Methods	Nitroprotein and Nitrotyrosine Sites	Remark
(i) Inflammation-related disease				
Lanone et al. [65]	Rectus abdominis muscle from the same control and septic patients	Western blot, MALDI-TOF-PMF, and molecular modeling	Inducible nitric oxide synthase (iNOS) was nitrated at Tyr299, Tyr336, Tyr446, and Tyr698	Analysis coupled with iNOS three-dimensional crystal model
Ghosh et al. [66]	Lung tissues from allergen-induced murine model of asthma	2D-Western blot and LC-ESI-MS/MS	Twenty-seven putative nitrated proteins were identified	Inflammation-related disease. No nitrotyrosine site were identified
Dhiman et al. [67]	Plasma from patients with Chagas' disease	1D- and 2D-Western blot, MALDI-PMF, and LC-ESI-MS/MS	Fifty differentially expressed/ nitrated proteins were identified	Inflammation-related disease. No nitrotyrosine site were identified
Chatterjee et al. [68]	Spleens from LPS-induced systemic inflammation model of C57BL6/J mice	1D-Western blot and LC-ESI-MS/MS	Carboxypeptide B1 (CPB1) was nitrated at specific tyrosine sites	Inflammation-related disease
Ghesquiere et al. [36]	Serum proteome from a C57BL6/J mouse with septic shock	COFRADIC and ESI-MS	α2-Macroglobulin, apolipoprotein A-I, haptoglobin, and vitamin D-binding protein were nitrated at six specific tyrosine sites	Inflammation-related disease. Nitrotyrosine sites were identified
(ii) Aging and aging-related diseases				
Kanski et al. [69]	Heart from 5-month-old and 26-month-old Fischer 344/BN F1 hybrid rats	1D- and 2D-Western blot, ESI-MS/MS	Forty-eight putative nitrated proteins. Nitration at Tyr105 of the electron-transfer flavoprotein was identified	Endogenous. Heart homogenate and heart mitochondria. Nitration is effects of biological aging. Not every protein's nitrotyrosine site was identified
Kanski, Hong, and Schöneich [70]	Skeletal muscle from 34-month-old Fischer 344/Brown Norway F1 rats	IEF, 1D-Western blot, and ESI-MS/MS	Eleven nitroproteins and twelve nitrotyrosine sites were identified	Endogenous

Sharov et al. [71]	Skeletal muscle from 6-month-old and 34-month-old Fischer 344/Brown Norway F1 hybrid rats	1D-Western blot and HPLC-ESI-MS/MS	Phosphorylase b was found in the accumulation of 3-nitrotyrosine on Tyr113, Tyr161, and Tyr573. Nitration on Tyr 113 was detected in 6-month-old and 34-month-old rat; nitration on Tyr161 and Tyr573 was detected only in 34-month-old rat	Endogenous. Nitration is accumulated with aging
Marshall et al. [72]	Liver from young (19–22 weeks) and old (24 months) C57/BL6 male mice	1D-Western blot and LC-ESI-MS/MS	Six putative nitrated proteins were identified	Nitration is associated with aging. No nitrotyrosine site was identified
(iii) Tumor				
Fiore et al. [73]	Human glioma tissues	Immunohistochemistry, 1DE-MALDI-PMF	Tublin was nitrated at Tyr224 in glioma grade IV but not in grade I and noncancerous brain tissue	Endogenous. Laser-induced decomposition
Nakagawa et al. [74]	C6 rat glioma cell line	HPLC, MALDI-PMF	Cytochrome c was nitrated at Tyr48, Tyr67, and Tyr74	*In vitro* nitrated with peroxynitrite
Zhan and Desiderio [2]	Human pituitary postmortem tissue	2D-Western blot and vMALDI-MS/MS	Four nitroproteins and four nitrotyrosine sites were identified	Endogenous. Laser-induced decomposition
Zhan and Desiderio [3]	Human nonfunctional pituitary adenoma tissue	NTAC-vMALDI-MS/MS	Nine nitroproteins, ten nitrotyrosine sites, and three nitroprotein-interacting protein were identified	Endogenous. Laser-induced decomposition
Zhan and Desiderio [57]	Human pituitary postmortem tissue	2D-Western blot and vMALDI-MS/MS	Four nitroproteins and four nitrotyrosine sites were identified	Endogenous. Laser-induced decomposition

(*Continued*)

TABLE 67.1 (Continued)

Reference	Specimen	Methods	Nitroprotein and Nitrotyrosine Sites	Remark
(iv) Neurodegenerative diseases				
Casoni et al. [75]	Spinal cord from Tg SOD1 G93A mice and Tg SOD1 WT mice	2D-Western blot and MALDI-PMF	Thirty-two nitroproteins and sixteen nitrotyrosine sites	Endogenous. Laser-induced decomposition. Familial amyotrophic lateral sclerosis
Sacksteder et al. [76]	Brain from C57BL/6J mice	SCX-LC-ESI-MS/MS	Twenty-nine nitroproteins and thirty-one nitrotyrosine sites	Endogenous. Links to neurodegenerative disease
Danielson et al. [77]	Dox-inducible MAO-B PC12 cells	LC-ESI-MS/MS	Alpha-synuclein was nitrated at Tyr39	Model of Parkinson's disease
Yoon et al. [78]	Mouse hippocampal cell line HT22	2D-Western blot and MALDI-PMF	Thirteen nitroproteins were detected	Glutamate-treated HT22 cells
Zhang et al. [79]	Brain from C57BL/6J mice	LC-ESI-MS/MS		Endogenous
(v) Cardiovascular system and related diseases				
Chen et al. [80]	Sprague Dawley rat *in vivo* myocardial regional ischemia–reperfusion model	1DE-LC-ESI-MS/MS	Flavin subunit is nitrated at Tyr56 and Tyr142	Endogenous. Mitochondrial complex II in the postischemic myocardium
Ai et al. [81]	Endothelial cell of human coronary arteries	LC-ESI-MS/MS	LDL was nitrated	
Liu et al. [82]	Male C57BL/6 mice myocardial ischemia–reperfusion (I/R) injury model	1D- or 2D-Western blot, LC-ESI-MS/MS	Twenty-three nitroproteins were identified. Ten of them were from mitochondria	Endogenous. No nitrotyrosine sites were identified
(vi) The neurovisual system				
Palamalai et al. [83]	Photoreceptor rod outer segments of cyclic light-reared rats treated or not with the antioxidant	2D-Western blot and LC-ESI-MS/MS	Ten putative nitroproteins were identified	Endogenous. No nitrotyrosine sites were identified

Justilien et al. [60]	Mouse posterior eyecups	NTAC, SDS-PAGE, LC-ESI-MS/MS	Eight nitroproteins and nine nitrotyrosine sites	Endogenous. SOD2 knockdown mouse model of early AMD
Murdaugh et al. [84]	Human Bruch's membrane	HPLC, ESI-MS/MS	A2E was nitrated	Endogenous. Nitro-A2E is a specific biomarker of nitrosative stress in Bruch's membrane, and its concentration is directly related to tissue age
(vii) Diabetes				
Kato et al. [85]	Healthy and diabetic human urine	LC-ESI-MS/MS	Urine nitrotyrosine	Endogenous
(viii) Kidney disease				
Piroddi et al. [86]	Plasma from kidney disease patients	2DE and LC-ESI-MS/MS	Fourteen tentative nitroproteins and seven nitrotyrosine sites were identified	Endogenous
(ix) Plant diseases				
Chaki et al. [87]	Sunflower hypocotyls	2D-Western blot and LC-ESI-MS/MS	Twenty one putative nitroproteins were identified	No nitrotyrosine sites were identified
(x) Others				
Aslan et al. [88]	Liver and kidney from sickle cell disease mouse	Western blot and precipitation, MALDI-PMF, LC-ESI-MS/MS	Actin was nitrated at Tyr91, Tyr198, and Tyr240	Endogenous
Webster, Brockman, and Myatt [89]	Human placenta	1D-Western blot and MALDI-PMF	p38 MAPK was nitrated	No nitrotyrosine sites were identified

(*Continued*)

TABLE 67.1 (Continued)

Reference	Specimen	Methods	Nitroprotein and Nitrotyrosine Sites	Remark
Ulrich et al. [90]	Human lung tissues and blood samples, aminal granule protein preparation	Western blot and MALDI-PMF	Six nitroproteins and nitrotyrosine sites at Tyr349 in eosinophil peroxidase (EPO) and Tyr33 in other eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN)	Endogenous
Hamilton et al. [91]	Human plasma	LC-ESI-MS/MS	Low-density lipoprotein (LDL) was nitrated at Tyr276, Tyr666, and Tyr720 of LDL-alpha 1, Tyr2524 of LDL-alpha 2, Tyr4141 of LDL-alpha 3, Tyr3139, Tyr3205, and Tyr3489 of LDL-beta 2	
Zhu et al. [92]	Liver from SOD1−/− and WT C57BL/6 mice	1DE, LC-ESI-MS/MS	Ten candidate nitrated proteins were identified	No nitrotyrosine sites were identified
Reed et al. [93]	Traumatic brain injury rats	2D-Western blot and MALDI-PMF	Several nitroprotein such as GSH were identified	
Lee et al. [45]	Hippocampus from smoke inhalation rat model	2D-Western blot and MALDI-PMF or MALDI-MS/MS	Five nitroproteins of mitochondrial proteins were identified	Endogenous. No nitrotyrosine sites were identified
Casanovas et al. [94]	Lipoprotein lipase of bovine and rat	2D-Western blot and LC-ESI-MS/MS	Lipoprotein lipase was nitrated at Tyr95, Tyr164, Tyr 316	Endogenous
Sharov et al. [95]	Rabbit muscle	LC-ESI-MS/MS	Glycogen phosphorylase b was nitrated at 28 nitrotyrosine sites	
Ohama and Brautigan [96]	Human peripheral blood mononuclear cells	LC-ESI-MS/MS	Protein phosphatase 2A was nitrated at Tyr284	

Sekar et al. [97]	Mast cells	2D-Western blot and LC-ESI-MS/MS	Aldolase was nitrated	No nitrotyrosine sites were identified
Redondo-Horcajo et al. [98]	Endothelial cells from bovine aorta and mouse lung	LC-ESI-MS/MS	Manganese superoxide dismutase (MnSOD) was nitrated at Tyr34	*In vitro* nitrated with cyclosporine A
Chen and Chen [99]	Human blood samples from smokers and nonsmokers	LC-ESI-MS/MS under the selected reaction monitoring (SRM) mode	Hemoglobin was nitrated at Tyr24 and Tyr42 (alpha globin) and Tyr130 (beta globin)	Nitration of human hemoglobin is associated with cigarette smoking

iii. Tyrosine nitration in tumorigenesis. Tumor is another important target that tyrosine nitration is involved in with clear evidences that nitroproteins were identified in human gliomas [73] and rat glioma cell lines [74]; moreover, 8 nitroproteins were discovered in human pituitary control tissue [2, 57], and 9 nitroproteins, 3 nitroprotein-interacted proteins, and 10 nitrotyrosine sites were discovered in a pituitary adenoma [3].

iv. Tyrosine nitration in neurodegenerative diseases. Studies have discovered nitroproteins in neurodegenerative-related model or disease such as spinal cords of a mouse model of familial amyotrophic lateral sclerosis [75], Parkinson's disease [77], mouse brain [76, 79, 100], and HT22 hippocampal cells [78].

v. Tyrosine nitration in the cardiovascular system and related diseases. It has been characterized with identification of nitroproteins in the vascularity [81], ischemia–reperfusion injury [80, 82, 101], and mouse heart [79, 100].

vi. Tyrosine nitration in the neurovisual system. For example, nitroproteins were discovered in outer segments of photoreceptor rod [83], the eyecup of SOD_2 knockdown mouse [60], and human Bruch's membrane [84].

vii. Tyrosine nitration in diabetes. Diabetes were found to involve tyrosine nitration modification with the discovery of nitroproteins in diabetic mellitus patients [102], a diabetic patient's urine [85], and diabetic rat models [103].

viii. Tyrosine nitration in kidney disease. It was evidenced with characterization of 14 nitroproteins and 7 nitrotyrosine sites in a kidney disease patient's plasma [86].

ix. Plant disease. Studies discovered nitroproteins in plant disease [87, 104].

x. Others. Tyrosine nitration is also discovered to participate in many other pathophysiological processes with identification of nitroproteins in sickle cell disease [88], placenta/preeclampsia [89], murine liver [92], eosinophil granule toxins [90], traumatic brain-injured rat models [93], rat hippocampus after acute inhalation of combustion smoke [45], rabbit muscle [95], hypertriglyceridemia [94], human plasma [91, 105], mononuclear cells from human peripheral blood [96], mast cells [97], endothelial cells [98], and cigarette-associated human hemoglobin [99].

67.6 BIOLOGICAL FUNCTION MEASUREMENT OF NITROPROTEINS

In order to elucidate in depth the biological functions of tyrosine nitration in a protein and biological system, the MS/MS-identified nitroproteins and nitrotyrosine sites must be further analyzed with other strategies such as literature data-based rationalization of biological function, protein domain/motif analysis, systems pathway network, and structural biology analysis. Here, nitroproteins and nitrotyrosine sites from pituitary control and adenoma (Table 67.2) are taken for an example to address those functional analyses.

TABLE 67.2 Nitroproteins and Unnitrated Proteins Identified from Pituitary Adenoma [3] and Control Tissues [2, 57]

Pituitary Adenoma		Pituitary Control	
Protein Name	nY Site	Protein Name	nY Site
Nitrated protein		*Nitrated protein*	
Rho-GTPase-activating 5 [Q13017] (ARHGAP5)	Y^{550}	Synaptosomal-associated protein (SNAP91)	Y^{237}
Leukocyte immunoglobulin-like receptor A4 [P59901]	Y^{404}	Ig alpha Fc receptor [P24071] (FCAR)	Y^{223}
Zinc finger protein 432 [O94892]	Y^{41}	Actin [P03996] (ACTA2, ACTG2, ACTC1)	Y^{296}
PKA beta regulatory subunit [P31321] (PRKAR1B)	Y^{20}	PKG 2 [Q13237] (PRKG2)	Y^{354}
Sphingosine-1-phosphate lyase 1 [O95470]	Y^{356}, Y^{366}	Mitochondrial cochaperone protein HscB [Q8IWL3]	Y^{128}
Centaurin beta 1 [Q15027]	Y^{485}	Stanniocalcin 1[P52823] (STC1)	Y^{159}
Proteasome subunit alpha type 2 [P25787] (PSMA2)	Y^{228}	Proteasome subunit alpha type 2 (PSMA2)	Y^{228}
Interleukin 1 family member 6 [Q9UHA7] (IL1F6)	Y^{96}	Progestin and adipoQ receptor family member III [Q6TCH7] (PAQR3)	Y^{33}
Rhophilin 2 [Q8IUC4] (RHPN2)	Y^{258}		
Nitroprotein-interacted protein			
Interleukin-1 receptor-associated kinase-like 2 (IRAK-2) [O43187] (IRAK2)			
Glutamate receptor interacting protein 2 [Q9C0E4] (GRIP2)			
Ubiquitin [P62988] (UBB or UBC)			

Source: Zhan and Desiderio [2, 3, 57]. Reproduced with permission of Elsevier.
Note: nY = Nitrotyrosine.

67.6.1 Literature Data-Based Rationalization of Biological Functions

A large volume of literature data analysis of 9 nitroproteins containing 10 nitrotyrosine sites and 3 nonnitrated proteins from a human pituitary adenoma (Table 67.2) [3] demonstrated that 3 nonnitrated proteins (ubiquitin, glutamate receptor-interacting protein 2, and interleukin 1 (IL1) receptor-associated kinase-like 2) interacted with nitroproteins to form 3 nitroprotein–protein complexes, including (i) nitrated beta subunit of cAMP-dependent protein kinase (PKA) complex, (ii) nitrated proteasome–ubiquitin complex, and (iii) nitrated interleukin 1 family member 6–interleukin 1 receptor–interleukin 1 receptor-associated kinase-like 2 (IL1F6–IL1R–IRAK2) [3, 7]. Furthermore, those

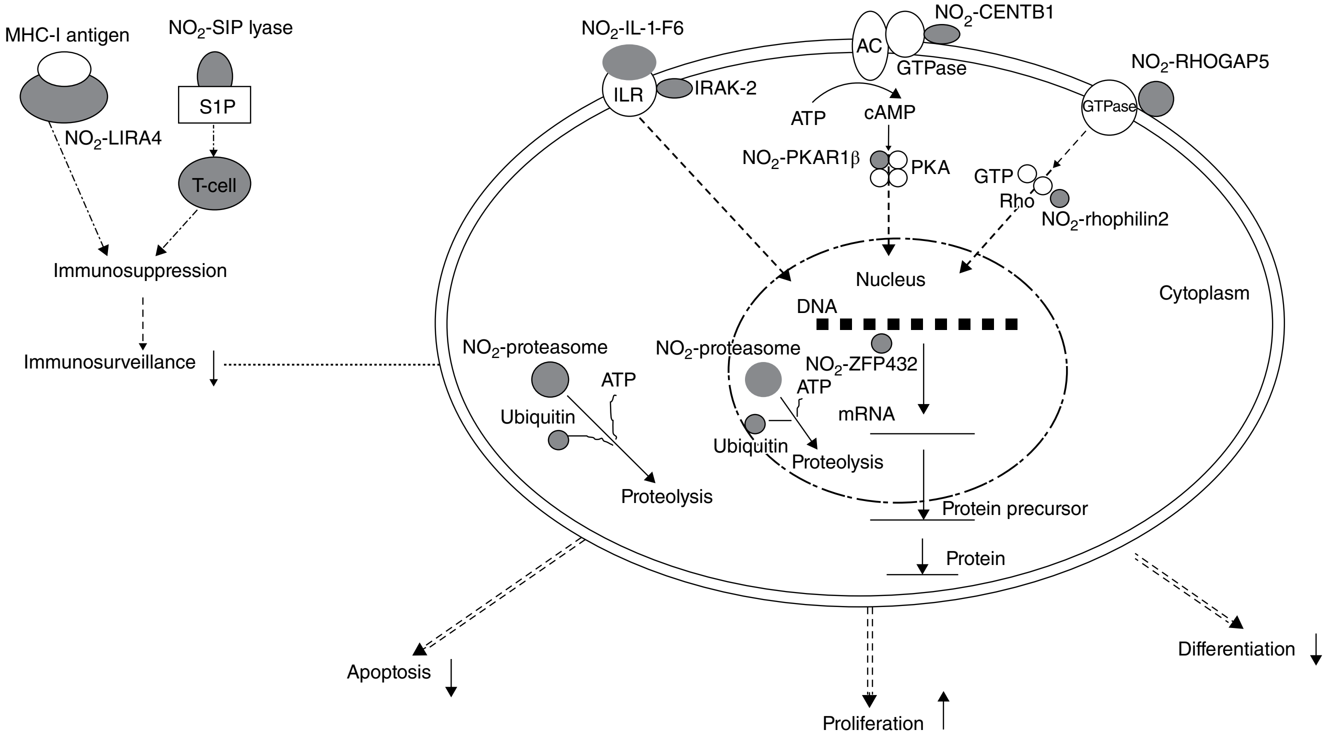


FIGURE 67.13 Experimental data-based model of nitroproteins and their functions in human nonfunctional pituitary adenomas. Source: Zhan and Desiderio [3]. Reproduced with permission of Elsevier.

three nitroprotein–protein complexes and nine nitroproteins were rationalized into an experiment-based biological function system (Fig. 67.13) [3, 7]. Nitrated leukocyte immunoglobulin-like receptor subfamily A member 4 (LIRA4) was associated with the immune system. Nitrated proteasome–ubiquitin complex, an important enzymatic complex, was involved in intracellular nonlysosomal proteolytic pathway [3, 7]. Nitrated sphingosine-1-phosphate lyase 1 (S1P lyase 1) was involved in sphingolipid metabolism to regulate immune system, cell proliferation, survival, and cell death [3, 7]. Nitrated IL1F6 and IRAK2 in the IL1R complex were involved in the cytokine system. Nitrated cAMP-dependent protein kinase type I-beta regulatory subunit (PKAR1-β) and nitrated centaurin beta 1 (CENT-β1) were involved in the PKA signal pathway. Nitrated rhophilin 2 and nitrated Rho-GTPase-activating protein 5 (RHOGAP5) were involved in the GTPase signal pathway. Nitrated zinc finger protein 432 (ZFP432) functioned in transcription regulatory systems [3, 7].

67.6.2 Protein Domain and Motif Analyses

Recognition of protein domains/motifs and location of nitrotyrosine sites into the corresponding domain/motif in a protein will much benefit the accurate clarification of biological activities of tyrosine nitration [7]. The commonly used software programs including Motif Scan (http://myhits.isb-sib.ch/cgi-bin/motif_scan), ProDom (http://prodom.prabi.fr/prodom/current/html/form.php), ScanProsite (http://us.expasy.org/tools/scanprosite/), Pfam (http://www.sanger.ac.uk/Software/Pfam/), and InterProScan (http://www.ebi.ac.uk/InterProScan/) were effective in the determination of significant domains/motifs in a nitroprotein and in location of each nitrotyrosine site within a protein domain/motif, for insights into the effect of tyrosine nitration on protein functions [3, 7]. Our studies demonstrated that most nitrotyrosine sites occurred within important domains/motifs in a protein [3] (Fig. 67.14), which hints that protein functions are altered by tyrosine nitration. For instance, nitrated S1P lyase 1 (Fig. 67.14) in a human pituitary adenoma is a key enzyme that catalyzes decomposition of S1P; two nitrations (NO_2–^{356}Y and NO_2–^{366}Y) were discovered within the enzyme activity region to decrease the interaction intensity of enzyme–substrate (S1P lyase 1/S1P) and to alter enzymatic activities of S1P lyase 1 [3, 7].

67.6.3 Systems Pathway Analysis

Because each protein in a proteome functions in a multiple, complex, and interacting systematic network but does not work alone [39], it is necessary to rationalize tyrosine nitration within those complex pathway system networks and to address pathway network variations due to tyrosine nitration. Currently, lots of pathway network analysis software have been developed such as Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com/) and MetaCore Pathway Analysis programs (http://www.genego.com/metacore.php/). IPA was used to determine signaling networks that involve nitroproteins from a

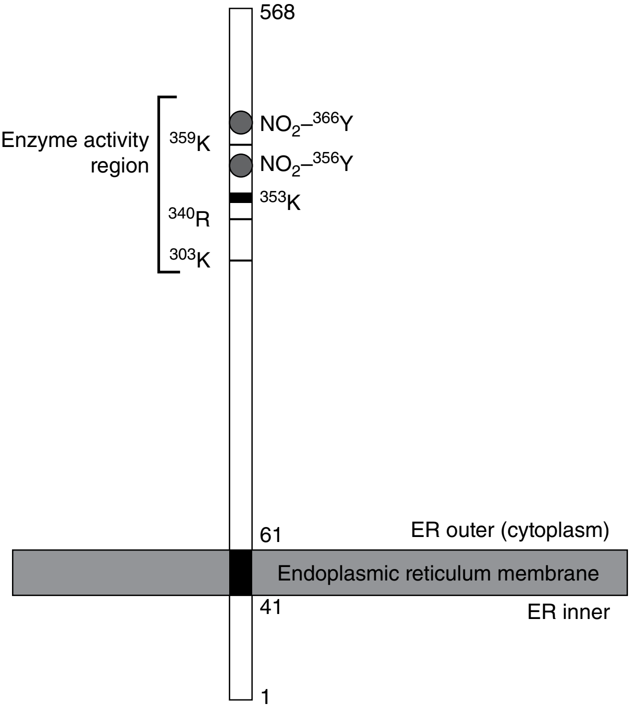


FIGURE 67.14 Nitration site and functional domains of sphingosine-1-phosphate lyase 1. Source: Zhan and Desiderio [3]. Reproduced with permission of Elsevier.

human pituitary control and adenoma tissues (Table 67.2) [3, 7]. As a result, those nitroproteins and their complexes from pituitary adenoma were involved in the IL1 and tumor necrosis factor (TNF) signaling networks (Fig. 67.15a), which function in cancer, cell cycle, and reproductive system diseases [39]; whereas those nitroproteins from control pituitary were involved in actin cellular skeleton and transforming growth factor beta 1 (TGFβ1) signaling networks (Fig. 67.15b), which function in gene expression, cellular development, and connective tissue development. Both adenoma and control networks include a beta-estradiol signal pathway, which hints that hormone metabolism is involved in pituitary normal physiology and adenoma pathology. Furthermore, pathway analysis revealed three important signaling pathway network systems (oxidative stress, cell-cycle dysregulation, and MAPK-signaling abnormality) in a pituitary adenoma that involve tyrosine nitration [7, 39]. Those pathway network data clearly elucidate biological functions and roles of tyrosine nitration in pituitary tumorigenesis.

67.6.4 Structural Biology Analysis

The electron density of a phenolic ring of a tyrosine residue is decreased by tyrosine nitration, which causes diminishing interaction intensity between receptor and ligand or between enzyme and substrate [2]. Thus, the spatial position of a nitrotyrosine in a

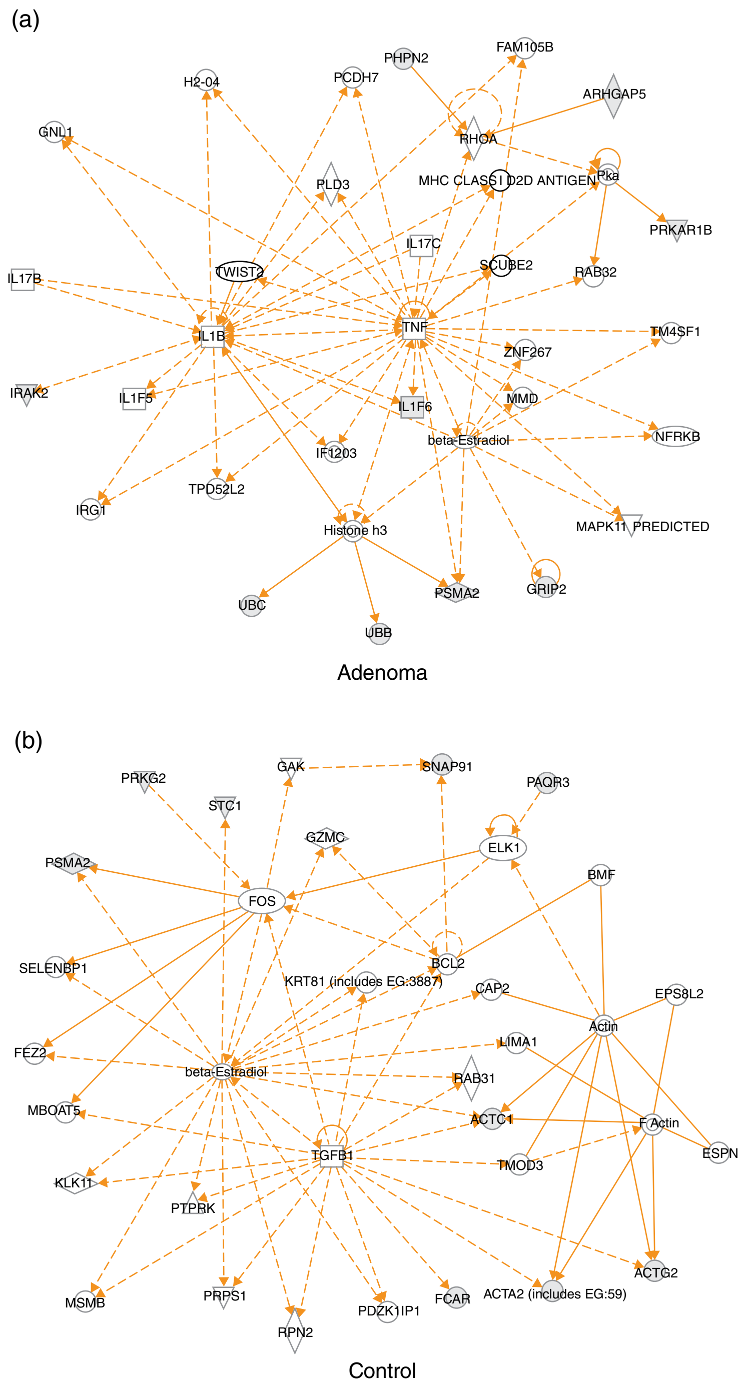


FIGURE 67.15 Significant signaling pathway networks mined from pituitary adenoma and control nitroproteomic data sets. (a) Network was derived from pituitary adenoma nitroproteomic data and function in cancer, cell cycle, and reproductive system diseases. A filled node denotes an identified nitroprotein or protein that interacts with nitroproteins. (b) Network is derived from pituitary control nitroproteomic data and function in gene expression, cellular development, and connective tissue development and function. A filled node denotes an identified nitroprotein. An solid edge denotes a direct relationship between two nodes (molecules: proteins, genes). An nonsolid edge denotes an indirect relationship between two nodes (molecules: proteins, genes). The various shapes of nodes denote the different functions. A curved line means an intracellular translocation; a curved arrow means an extracellular translocation. Source: Zhan and Desiderio [39]. CC-BY 2.0.

protein might obviously influence on biological functions and roles of tyrosine nitration. 3D spatial structure of a protein can clearly determine its biological functions. Therefore, reconstruction of 3D spatial structure of a nitroprotein with X-ray crystallography data would be very easy to clarify effects of tyrosine nitration on 3D structure of a nitroprotein. Also, based on tyrosine nitration site, domain, and 3D structure, it is possible for one to design a small drug toward 3D structure and domain that contains tyrosine nitration [7]. For example, the nicotinamide adenine dinucleotide (NAD^+) binding assay demonstrated that nitrated glyceraldehyde-3-phosphate dehydrogenase (GAPDH) did not bind NAD^+ [40]; while X-ray crystal structure was used to interpret effects of tyrosine nitration on the capability of NAD^+ binding in GAPDH [40]. MS analysis of nitrated GAPDH determined Tyr^{311} and Tyr^{317} were the only sites of nitration; and the distances between Tyr^{311} and Tyr^{317} and the cofactor NAD^+ were revealed by X-ray crystal structure to be less than 7.2 and 3.7 Å, respectively. Those data imply that nitration of these two residues might affect NAD^+ binding [40]. Another example is the use of X-ray crystal structure of mammalian succinate ubiquinone reductase (SQR or complex II) to effectively interpret the association of tyrosine nitration (Tyr^{142}) of the flavin subunit with *S*-glutathionylated cysteine residue Cys^{90} of mitochondrial complex II in a postischemic myocardium [80]. Briefly, X-ray crystal structure of SQR indicates a Rossman-type fold with four major domains in flavin subunit and Tyr^{142} in the major helix (residues 136–158) of a floating subdomain (residues 105–196). Moreover, Tyr^{142} is highly surface exposed and situated in the hydrophilic environment, which suggests that this tyrosine should be susceptible to nitration with $OONO^-$. X-ray crystal structure of SQR also indicates that Tyr^{142} is approximately 20 Å away from isoalloxazine ring of flavin adenine dinucleotide (FAD) and that Cys^{90} is located within the part of the *N*-terminal beta barrel subdomain (residue 53–104) of the large FAD-binding domain, near the AMP moiety of FAD (~7.7 Å), where major catalysis of electron transfer and O_2^- production occurs. Thus, *S*-glutathionylaton of Cys^{90} seems likely to induce a conformational change near the floating subdomain (residues 105–196) to increase shielding effect on Tyr^{142} and to render Tyr^{142} less accessible to $OONO^-$ oxidation [80]. Therefore, 3D structure of a protein can accurately interpret effects of tyrosine nitration on that protein's structure and functions.

67.7 PITFALLS OF NITROPROTEIN MEASUREMENT

Although measurement strategies of nitroproteins have extensively been developed, one must clearly realize that no highly reliable, high-sensitivity, high-throughput, and high-reproducibility approach exists to analyze the extremely challenging endogenous tyrosine nitration in a proteome. Therefore, different approaches are still under development. Those pitfalls include the following: (i) Nitrotyrosine antibody-based immunoaffinity methods such as 2D-Western blotting and NTAC succeeded to identify endogenous nitrotyrosine sites; however, an overwhelming amount of nonnitrated tryptic peptides negatively affects nitropeptide characterization. For that reason, we suggest development of immunoaffinity enrichment of tryptic nitropeptides—not nitroproteins—prior to MS analysis. (ii) Until now, most methods based on chemical derivatization (as described earlier) are used only for *in vitro* experiments and not for endogenous nitrotyrosine sites. Although the COFRADIC-based characterization of nitropeptides succeeded in a serum proteome, sensitivity and throughput were very low, and it has not been used extensively in *endogenous* tissue nitroproteomes. Therefore, development of better nitrotyrosine analysis methods is necessary in the following aspects—alone or in combination: (i) derivatize a nitro to amino group to stabilize MS behaviors, (ii) develop specific amino group tags to enrich nitrotyrosine peptides, (iii) enrich nitrotyrosine—or aminotyrosine peptides are better than nitrotyrosine—or aminotyrosine proteins for sensitivity, (iv) improve liquid chromatography isolation, (v) develop super high-sensitivity mass spectrometers, (vi) choose an appropriate ion source and collision model to fragment nitropeptide or aminopeptides, and (vii) develop reliable software for data analysis. The combined multiple aspects among items (i)–(vii) are recommended to maximize coverage of endogenous nitrotyrosine sites in a proteome.

67.8 CONCLUSIONS

Protein tyrosine nitration is an important oxidative-/nitrative-mediated modification and associates a wide range of pathophysiological conditions [2–4, 7, 20, 57]. Moreover, evidence suggests the presence of a denitrase in mammalian tissues although a denitrase has not been isolated and its enzymatic activity not confirmed. Thus, tyrosine nitration can be considered as reversible; and tyrosine nitration is not only a result from oxidative damage, but it also participates in pathophysiological processes [106]. Nitration dynamically alters protein functions [107], including activation or inactivation [108–110]. MS-based identification of nitroproteins and nitrotyrosine sites is essential to understand biological roles of this modification [111–113]. However, it is analytically very challenging to identify *endogenous* nitroproteins and nitrotyrosine sites due to nitration's low abundance in biological samples and its multiple mass spectrometric behaviors among MALDI UV laser, ESI, CID, ECD, ETD, and MAD. Endogenous nitroproteins/nitropeptides must be enriched prior to MS analysis. Several enrichment methods have been developed, including immunoaffinity enrichment, biotin-affinity enrichment, and COFRADIC. Nitrotyrosine sites have been found in many different pathophysiological conditions. TMT- or iTRAQ-based quantitative nitroproteomics are needed to quantify disease-key nitroproteins/peptides. Protein domain/motif analysis, systems pathway analysis, and structural biological analysis of nitroproteins are significantly needed to elucidate the biological roles of tyrosine nitration.

NOMENCLATURE

1DE	one-dimensional electrophoresis
2D	two-dimensional
2DE	two-dimensional electrophoresis
2DGE	two-dimensional gel electrophoresis
3D	three-dimensional
APPD	(3R, 4S)-1-(4-(aminomethyl)phenylsulfonyl)pyrrolidine-3,4-diol
BSA	bovine serum albumin
CENT-β1	centaurin beta 1
CID	collision-induced dissociation
COFRADIC	combined fractional diagonal chromatography
cPILOT	combined precursor isotopic labeling and isobaric tagging
ECD	electron-capture dissociation
ESI	electrospray ionization
ETD	electron-transfer dissociation
FAD	flavin adenine dinucleotide
FT-ICR	Fourier transform ion cyclotron resonance
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
HPLC	high-performance liquid chromatography
IEF	isoelectric focusing
IL1	interleukin 1
IL1F6	interleukin 1 family member 6
IL1R	interleukin 1 receptor
IPA	Ingenuity Pathway Analysis
IR	infrared
IRAK2	interleukin 1 receptor-associated kinase-like 2
iTRAQ	isobaric tags for relative and absolute quantification
LC	liquid chromatography
LE1	leucine enkephalin
LE2	nitro-Tyr leucine enkephalin
LE3	*d*(5)-Phe-nitro-Tyr leucine enkephalin
LIRA4	leukocyte immunoglobulin-like receptor subfamily A member 4
MAD	metastable atom-activated dissociation
MALDI	matrix-assisted laser desorption/ionization
MAPK	mitogen-activated protein kinase
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MW	molecular weight
m/z	mass-to-charge ratio
NAD^+	nicotinamide adenine dinucleotide
NHS acetate	acetic acid *N*-hydroxysuccinimide ester
NTAC	nitrotyrosine affinity column

$OONO^-$	peroxynitrite
OVA	ovalbumin
PKA	cAMP-dependent protein kinase
PKAR1-β	cAMP-dependent protein kinase type I-beta regulatory subunit
PMF	peptide mass fingerprint
PTM	posttranslational modification
RHOGAP5	Rho-GTPase-activating protein 5
S1P	sphingosine-1-phosphate
SCX	strong cation exchange
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SQR	succinate ubiquinone reductase
SRM	selected reaction monitoring
TGFβ1	transforming growth factor beta 1
TMT	tandem mass tags
TNF	tumor necrosis factor
TNM	tetranitromethane
TOF	time of flight
UV	ultraviolet
ZFP432	zinc finger protein 432

ACKNOWLEDGMENTS

The authors acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 81272798 and 81572278 to X.Z.), the Xiangya Hospital Funds for Talent Introduction (to X.Z.), the Hunan Provincial Natural Science Foundation of China (Grant No. 14JJ7008 to X. Z.), China “863” Plan Project (Grant No. 2014AA020610-1 to X. Z.), and the National Institutes of Health, United States (RR16679; NS 42843 to D.M.D.).

REFERENCES

1. Scaloni A. 2006. Mass spectrometry approaches for the molecular characterization of oxidatively/nitrosatively modified proteins, in *Redox Proteomics: From Protein Modification to Cellular Dysfunction and Diseases* (Dalle-Donne I, Scaloni A, Butterfield DA, Eds.), John Wiley & Sons, Inc., Hoboken, NJ, pp. 59–100.
2. Zhan X, Desiderio DM. 2004. The human pituitary nitoproteome: detection of nitrotyrosyl-proteins with two-dimensional Western blotting, and amino acid sequence determination with mass spectrometry. *Biochem Biophys Res Commun* 325: 1180–1186.
3. Zhan X, Desiderio DM. 2006. Nitroproteins from human pituitary adenoma tissue discovered with a nitrotyrosine affinity column and tandem mass spectrometry. *Anal Biochem* 354: 279–289.

4. Zhan X, Desiderio DM. 2009a. Mass spectrometric identification of in vivo nitrotyrosine sites in the human pituitary tumor proteome. *Methods Mol Biol* 566: 137–63.
5. Ghesquiere B, Goethals M, Van Damme J, Staes A, Timmerman E, Vandekerckhove J, Gevaert K. 2006. Improved tandem mass spectrometric characterization of 3-nitrotyrosine sites in peptides. *Rapid Commun Mass Spectrom* 20: 2885–2893.
6. Yee CS, Seyedsayamdost MR, Chang MC, Nocera DG, Stubbe J. 2003. Generation of the R2 subunit of ribonucleotide reductase by intein chemistry: insertion of 3-nitrotyrosine at residue 356 as a probe of the radical initiation process. *Biochemistry* 42: 14541–14552.
7. Zhan X, Wang X, Desiderio DM. 2013. Pituitary adenoma nitroproteomics: current status and perspectives. *Oxid Med Cell Longev* 2013: 580710.
8. Irie Y, Saeki M, Kamisaki Y, Martin E, Murad F. 2003. Histone H1.2 is a substrate for denitrase, an activity that reduces nitrotyrosine immunoreactivity in proteins. *Proc Nat Acad Sci U S A* 100: 5634–5639.
9. Mallozzi C, D'Amore C, Camerini S, Macchia G, Crescenzi M, Petrucci TC, Di Stasi AM. 2013. Phosphorylation and nitration of tyrosine residues affect functional properties of synaptophysin and dynamin I, two proteins involved in exoendocytosis of synaptic vesicles. *Biochim Biophys Acta* 1833: 110–121.
10. Zhan X, Desiderio DM. 2011. Nitroproteins identified in human ex-smoker bronchoalveolar lavage fluid. *Aging Dis* 2: 100–115.
11. Zhan X, Du Y, Crabb JS, Kern TS, Crabb JW. 2007. Identification of nitrated proteins in diabetic rat retina. *Invest Ophthalmol Vis Sci* 48: E-abstract 4962.
12. Aulak KS, Koeck T, Crabb JW, Stuehr DJ. 2004. Dynamics of protein nitration in cells and mitochondria. *Am J Physiol* 286: H30–H38.
13. Koeck T, Fu X, Hazen SL, Crabb JW, Stuehr DJ, Aulak KS. 2004. Rapid and selective oxygen-regulated protein tyrosine denitration and nitration in mitochondria. *J Biol Chem* 279: 27257–27262.
14. Haddad IY, Pataki G, Hu P, Galliani C, Beckman JS, Matalon S. 1994. Quantitation of nitrotyrosine levels in lung sections of patients and animals with acute lung injury. *J Clin Investig* 94: 2407–2413.
15. Halliwell B, Zhao K, Whiteman M. 1999. Nitric oxide and peroxynitrite. The ugly, the uglier and the not so good: a personal view of recent controversies. *Free Radac Res* 31: 651–669.
16. Miyagi M, Sakaguchi H, Darrow RM, Yan L, West KA, Aulak KS, Stuehr DJ, Hollyfield JG, Organisciak DT, Crabb JW. 2002. Evidence that light modulates protein nitration in rat retina. *Mol Cell Proteomics* 1: 293–303.
17. Yeo WS, Lee SJ, Lee JR, Kim KP. 2008. Nitrosative protein tyrosine modifications: biochemistry and functional significance. *BMB Rep* 41: 194–203.
18. Petersson AS, Steen H, Kalume DE, Caidahl K, Roepstorff P. 2001. Investigation of tyrosine nitration in proteins by mass spectrometry. *J Mass Spectrom* 36: 616–625.
19. Sarver A, Scheffler K, Shetlar MD, Gibson BW. 2001. Analysis of peptides and proteins containing nitrotyrosine by matrix-assisted laser desorption/ionization mass spectrometry. *J Am Soc Mass Spectrom* 12: 439–448.

20. Zhan X, Desiderio DM. 2009b. MALDI-induced fragmentation of leucine enkephalin, nitro-Tyr leucine enkephalin, and *d*(5)-Phe-nitro-Tyr leucine enkephalin. *Int J Mass Spectrom* 287: 77–86.

21. Kim JK, Lee JR, Kang JW, Lee SJ, Shin GC, Yeo WS, Kim KH, Park HS, Kim KP. 2011. Selective enrichment and mass spectrometric identification of nitrated peptides using fluorinated carbon tags. *Anal Chem* 83: 157–163.

22. Lee JR, Lee SJ, Kim TW, Kim JK, Park HS, Kim DE, Kim KP, Yeo WS. 2009b. Chemical approach for specific enrichment and mass analysis of nitrated peptides. *Analyt Chem* 81: 6620–6629.

23. Lee SJ, Lee JR, Kim YH, Park YS, Park SI, Park HS, Kim KP. 2007. Investigation of tyrosine nitration and nitrosylation of angiotensin II and bovine serum albumin with electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom* 21: 2797–2804.

24. Zhang Q, Qian WJ, Knyushko TV, Clauss TR, Purvine SO, Moore RJ, Sacksteder CA, Chin MH, Smith DJ, Camp DG 2nd, Bigelow DJ, Smith RD. 2007. A method for selective enrichment and analysis of nitrotyrosine-containing peptides in complex proteome samples. *J Proteom Res* 6: 2257–2268.

25. Dekker F, Abello N, Wisastra R, Bischoff R. 2012. Enrichment and detection of tyrosine-nitrated proteins. *Curr Protoc Protein Sci* DOI:10.1002/0471140864.ps1413s69.

26. Feeney MB, Schöneich C. 2013. Proteomic approaches to analyze protein tyrosine nitration. *Antioxid Redox Signal* 19: 1247–1256.

27. Dr Gusanu M, Petre BA, Przybylski M. 2011. Epitope motif of an anti-nitrotyrosine antibody specific for tyrosine-nitrated peptides revealed by a combination of affinity approaches and mass spectrometry. *J Pept Sci* 17: 184–191.

28. Sultana R, Reed T, Butterfield DA. 2009. Detection of 4-hydroxy-2-nonenal- and 3-nitrotyrosine-modified proteins using a proteomics approach. *Methods Mol Biol* 519: 351–361.

29. Abello N, Barroso B, Kerstjens HAM, Postma DS, Bischoff R. 2010. Chemical labeling and enrichment of nitrotyrosine-containing peptides. *Talanta* 80: 1503–1512.

30. Tsumoto H, Taguchi R, Kohda K. 2010. Efficient identification and quantification of peptides containing nitrotyrosine by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry after derivation. *Chem Pharm Bull* 58: 488–494.

31. Amoresano A, Chiappetta G, Pucci P, D'Ischia M, Marino G. 2007. Bidimensional tandem mass spectrometry for selective identification of nitration sites in proteins. *Anal Chem* 79: 2109–2117.

32. Amoresano A, Chiappetta G, Pucci P, Marino G. 2008. A rapid and selective mass spectrometric method for the identification of nitrated proteins. *Methods Mol Biol* 477: 15–29.

33. Robinson RA, Evans AR. 2012. Enhanced sample multiplexing for nitrotyrosine-modified proteins using combined precursor isotopic labeling and isobaric tagging. *Anal Chem* 84: 4677–4686.

34. Prokai-Tatrai K, Guo J, Prokai L. 2011. Selective chemoprecipitation and subsequent release of tagged species for the analysis of nitropeptides by liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* 10 DOI:10.1074/mcp.M110.002923.

35. Chiappetta G, Corbo C, Palmese A, Galli F, Piroddi M, Marino G, Amoresano A. 2009. Quantitative identification of protein nitration sites. *Proteomics* 9: 1524–1537.
36. Ghesquiere B, Colaert N, Helsens K, Dejager L, Vanhaute C, Verleysen K, Kas K, Timmerman E, Goethals M, Libert C, Vandekerckhove J, Gevaert K. 2009. In vitro and in vivo protein-bound tyrosine nitration characterized by diagonal chromatography. *Mol Cell Proteomics* 8: 2642–2652.
37. Larsen TR, Bache N, Gramsbergen JB, Roepstorff P. 2011. Identification of nitrotyrosine containing peptides using combined fractional diagonal chromatography (COFRADIC) and off-line nano-LC-MALDI. *J Am Soc Mass Spectrom* 22: 989–996.
38. Zhang Y, Yang H, PoschI U. 2011. Analysis of nitrated proteins and tryptic peptides by HPLC-chip-MS/MS: site-specific quantification, nitration degree, and reactivity of tyrosine residues. *Anal Bioanal Chem* 399: 459–471.
39. Zhan X, Desiderio DM. 2010a. Signaling pathway networks mined from human pituitary adenoma proteomics data. *BMC Med Genomics* 3: 13.
40. Palamalai V, Miyagi M. 2010. Mechanism of glyceraldehyde-3-phosphate dehydrogenase inactivation by tyrosine nitration. *Protein Sci* 19: 255–262.
41. Seeley KW, Stevens SM Jr. 2012. Investigation of local primary structure effects on peroxynitrite-mediated tyrosine nitration using targeted mass spectrometry. *J Proteomics* 75: 1691–1700.
42. Zhan X, Desiderio DM. 2010b. The use of variations in proteomes to predict, prevent, and personalize treatment for clinically nonfunctional pituitary adenomas. *EPMA J* 1: 439–459.
43. Turko IV, Murad F. 2005. Mapping sites of tyrosine nitration by matrix-assisted laser desorption/ionization mass spectrometry. *Methods Enzymol* 396: 266–275.
44. Petre BA, Youhnovski N, Lukkari J, Weber R, Przybylski M. 2005. Structural characterization of tyrosine-nitrated peptides by ultraviolet and infrared matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry. *Eur J Mass Spectrom* 11: 513–518.
45. Lee HM, Reed J, Greeley GH Jr, Englander EW. 2009a. Impaired mitochondrial respiration and protein nitration in the rat hippocampus after acute inhalation of combustion smoke. *Toxicol Appl Pharmacol* 235: 208–215.
46. Cook SL, Jackson GP. 2011. Characterization of tyrosine nitration and cysteine nitrosylation modifications by metastable atom-activation dissociation mass spectrometry. *J Am Soc Mass Spectrom* 22: 221–232.
47. Jones AW, Mikhailov VA, Iniesta J, Cooper HJ. 2010. Electron capture dissociation mass spectrometry of tyrosine nitrated peptides. *J Am Soc Mass Spectrom* 21: 268–277.
48. Jones AW, Cooper HJ. 2010. Probing the mechanisms of electron capture dissociation mass spectrometry with nitrated peptides. *Phys Chem Chem Phys* 12: 13394–13399.
49. Li B, Held JM, Schilling B, Danielson SR, Gibson BW. 2011. Confident identification of 3-nitrotyrosine modifications in mass spectral data across multiple mass spectrometry platforms. *J Proteomics* 74: 2510–2521.
50. Sokolovsky M, Riordan JF, Vallee BL. 1966. Tetranitromethane. A reagent for the nitration of tyrosyl residues in proteins. *Biochemistry* 5: 3582–3589.

51. Fujigaki H, Saito K, Lin F, Fujigaki S, Takahashi K, Martin BM, Chen CY, Masuda J, Kowalak J, Takikawa O, Seishima M, Markey SP. 2006. Nitration and inactivation of IDO by peroxynitrite. *J Immunology* 176: 372–379.

52. Sokolovsky M, Riordan JF, Vallee BL.1967. Conversion of 3-nitrotyrosine to 3-aminotyrosine in peptides and proteins. *Biochem Biophys Res Commun* 27: 20–25.

53. Abello N, Kerstjens HA, Postma DS, Bischoff R. 2009. Protein tyrosine nitration: selectivity, physicochemical and biological consequences, denitration, and proteomics methods for the identification of tyrosine-nitrated proteins. *J Proteome Res* 8: 3222–3238.

54. Sheeley SA, Rubakhin SS, Sweedler JV. 2005. The detection of nitrated tyrosine in neuropeptides: a MALDI matrix-dependent response. *Anal Bioanal Chem* 382: 22–27.

55. Ghesquiere B, Helsens K, Vandekerckhove J, Gevaert K. 2011. A stringent approach to improve the quality of nitrotyrosine peptide identifications. *Proteomics* 11: 1094–1098.

56. Shigenaga MK, Lee HH, Blunt BC, Christen S, Shigeno ET, Yip H, Ames BN. 1997. Inflammation and NO(X)-induced nitration: assay for 3-nitrotyrosine by HPLC with electrochemical detection. *Proc Natl Acad Sci U S A* 94: 3211–3216.

57. Zhan X, Desiderio DM. 2007. Linear ion-trap mass spectrometric characterization of human pituitary nitrotyrosine containing proteins. *Int J Mass Spectrom* 259: 96–104.

58. Aulak KS, Miyagi M, Yan L, West KA, Massillon D, Crabb JW, Stuehr DJ. 2001. Proteomic method identifies proteins nitrated in vivo during inflammatory challenge. *Proc Natl Acad Sci U S A* 98: 12056–12061.

59. Butt YK, Lo SC. 2008. Detecting nitrated proteins by proteomic technologies. *Methods Enzymol* 440: 17–31.

60. Justilien V, Pang JJ, Renganathan K, Zhan X, Crabb JW, Kim SR, Sparrow JR, Hauswirth WW, Lewin AS. 2007. SOD2 knockdown mouse model of early AMD. *Invest Ophthalmol Vis Sci* 48: 4407–4420.

61. Petre BA, Ulrich M, Stumbaum M, Bernevic B, Moise A, Döring G, Przybylski M. 2012. When is mass spectrometry combined with affinity approaches essential? A case study of tyrosine nitration in proteins. *J Am Soc Mass Spectrom* 23: 1831–1840.

62. Dremina ES, Li X, Galeva NA, Sharov VS, Stobaugh JF, Schöneich C. 2011. A methodology for simultaneous fluorogenic derivatization and boronate affinity enrichment of 3-nitrotyrosine containing peptides. *Anal Biochem* 418: 184–196.

63. Dalle-Donne I, Scaloni A, Butterfield DA (Eds.). 2006. *Redox Proteomics: From Protein Modifications to Cellular Dysfunction and Diseases*, John Wiley & Sons, Inc., Hoboken, NJ.

64. Dalle-Donne I, Scaloni A, Giustarini D, Cavarra E, Tell G, Lungarella G, Colombo R, Rossi R, Milzani A. 2005. Proteins as biomarkers of oxidative/nitrosative stress in diseases: the contribution of redox proteomics. *Mass Spectrom Rev* 24: 55–99.

65. Lanone S, Manivet P, Callebert J, Launay JM, Payen D, Aubier M, Boczkowski J, Mebazaa A. 2002. Inducible nitric oxide synthase (NOS2) expressed in septic patients is nitrated on selected tyrosine residues: implications for enzymic activity. *Biochem J* 366: 399–404.

66. Ghosh S, Janocha AJ, Aronica MA, Swaidani S, Comhair SAA, Xu W, Zheng L, Kaveti S, Kinter M, Hazen SL, Erzurum SC. 2006. Nitrotyrosine proteome survey in asthma identifies oxidative mechanism of catalase inactivation. *J Immunology* 176: 5587–5597.

67. Dhiman M, Nakayasu ES, Madaiah YH, Reynolds BK, Wen JJ, Almeida IC, Garg NJ. 2008. Enhanced nitrosative stress during Trypanosoma cruzi infection causes nitrotyrosine modification of host proteins: implications in Chagas' disease. *Am J Pathol* 173: 728–740.

68. Chatterjee S, Lardinois O, Bonini MG, Bhattacharjee S, Stadler K, Corbett J, Deterding LJ, Tomer KB, Kadiiska M, Mason RP. 2009. Site-specific carboxypeptidase B1 tyrosine nitration and pathophysiological implications following its physical association with nitric oxide synthase-3 in experimental sepsis. *J Immunol* 183: 4055–4066.

69. Kanski J, Behring A, Pelling J, Schöneich C. 2005a. Proteomic identification of 3-nitrotyrosine-containing rat cardiac proteins: effects of biological aging. *Am J Physiol Heart Circ Physiol* 288: H371–H381.

70. Kanski J, Hong SJ, Schöneich C. 2005b. Proteomic analysis of protein nitration in aging skeletal muscle and identification of nitrotyrosine-containing sequences in vivo by nanoelectrospray ionization tandem mass spectrometry. *J Biol Chem* 280: 24261–24266.

71. Sharov VS, Galeva NA, Kanski J, Williams TD, Schöneich C. 2006. Age-associated tyrosine nitration of rat skeletal muscle glycogen phosphorylase b: characterization by HPLC-nanoelectrospray-tandem mass spectrometry. *Exp Gerontol* 41: 407–416.

72. Marshall A, Lutfeali R, Raval A, Chakravarti DN, Chakravarti B. 2013. Differential hepatic protein tyrosine nitration of mouse due to aging-effect on mitochondrial energy metabolism, quality control machinery of the endoplasmic reticulum and metabolism of drugs. *Biochem Biophys Res Commun* 430: 231–235.

73. Fiorce G, Di Cristo C, Monti G, Amoresano A, Columbano L, Pucci P, Cioffi FA, Cosmo AD, Palumbo A, d'Ischia M. 2006. Tubulin nitration in human gliomas. *Neurosci Lett* 394: 57–62.

74. Nakagawa H, Komai N, Takusagawa M, Miura Y, Toda T, Miyata N, Ozawa T, Ikota N. 2007. Nitration of specific tyrosine residues of cytochrome C is associated with caspase-cascade inactivation. *Biol Pharm Bull* 30: 15–20.

75. Casoni F, Basso M, Massignan T, Gianazza E, Cheroni C, Salmona M, Bendotti C, Bonetto V. 2005. Protein nitration in a mouse model of familial amyotrophic lateral sclerosis: possible multifunctional role in the pathogenesis. *J Biol Chem* 280: 16295–16304.

76. Sacksteder CA, Qian WJ, Knyushko TV, Wang HW, Chin MH, Lacan G, Melega WP, Camp DG 2nd, Smith RD, Smith DJ, Squier TC, Bigelow DJ. 2006. Endogenously nitrated proteins in mouse brain: links to neurodegenerative disease. *Biochemistry* 45: 8009–8022.

77. Danielson SR, Held JM, Schilling B, Oo M, Gibson BW, Andersen JK. 2009. Preferentially increased nitration of alpha-synuclein at tyrosine-39 in a cellular oxidative model of Parkinson's disease. *Anal Chem* 81: 7823–7828.

78. Yoon SW, Kang S, Ryu SE, Poo H. 2010. Identification of tyrosine-nitrated proteins in HT22 hippocampal cells during glutamate-induced oxidative stress. *Cell Prolif* 43: 584–593.

79. Zhang X, Monroe ME, Chen B, Chin MH, Heibeck TH, Schepmoes AA, Yang F, Petritis BO, Camp DG 2nd, Pounds JG, Jacobs JM, Smith DJ, Bigelow DJ, Smith RD, Qian WJ. 2010. Endogenous 3,4-dihydroxyphenylalanine and dopaquinone modifications on protein tyrosine: links to mitochondrially derived oxidative stress via hydroxyl radical. *Mol Cell Proteomics* 9: 1199–1208.

80. Chen CL, Chen J, Rawale S, Varadharaj S, Kaumaya PPT, Zweier JL, Chen YR. 2008. Protein tyrosine nitration of the Flavin subunit is associated with oxidative modification of mitochondrial complex II in the post-ischemic myocardium. *J Biol Chem* 283: 27991–28003.

81. Ai L, Rouhanizadeh M, Wu JC, Takabe W, Yu H, Alavi M, Chu Y, Miller J, Heistad DD, Hsiai TK. 2008. Shear stress influences spatial variations in vascular Mn-SOD expression. *Am J Physiol Cell Physiol* 294: C1576–C1585.

82. Liu B, Tewari AK, Zhang L, Green-Church KB, Zweier JL, Chen YR, He G. 2009. Proteomic analysis of protein tyrosine nitration after ischemia reperfusion injury: mitochondria as the major target. *Biochem Biophys Acta* 1974: 476–485.

83. Palamalai V, Darrow RM, Organisciak DT, Miyagi M. 2006. Light-induced changes in protein nitration in photoreceptor rod outer segments. *Mol Vis* 12: 1543–1551.

84. Murdaugh LS, Wang Z, Del Priore LV, Dillon J, Gaillard ER. 2010. Age-related accumulation of 3-nitrotyrosine and nitro-A2E in human Bruch's membrane. *Exp Eye Res* 90: 564–571.

85. Kato Y, Dozaki N, Nakamura T, Kitamoto N, Yoshida A, Naito M, Kitamura M, Osawa T. 2009. Quantification of modified tyrosines in healthy and diabetic human urine using liquid chromatography/tandem mass spectrometry. *J Clin Biochem Nutr* 44: 67–78.

86. Piroddi M, Palmese A, Pilolli F, Amoresano A, Pucci P, Ronco C, Galli F. 2011. Plasma nitroproteome of kidney disease patients. *Amino Acids* 40: 653–667.

87. Chaki M, Valderrama R, Fernández-Ocaña AM, Carreras A, López-Jaramillo J, Luque F, Palma JM, Pedrajas JR, Begara-Morales JC, Sánchez-Calvo B, Gómez-Rodríguez MV, Corpas FJ, Barroso JB. 2009. Protein targets of tyrosine nitration in sunflower (Helianthus annuus L.) hypocotyls. *J Exp Bot* 60: 4221–4234.

88. Aslan M, Ryan TM, Townes TM, Coward L, Kirk MC, Barnes S, Alexander CB, Rosenfeld SS, Freeman BA. 2003. Nitric oxide-dependent generation of reactive species in sickle cell disease. Actin tyrosine induces defective cytoskeletal polymerization. *J Biol Chem* 278: 4194–4204.

89. Webster RP, Brockman D, Myatt L. 2006. Nitration of p38 MAPK in the placenta: association of nitration with reduced catalytic activity of p38 MAPK in pre-eclampsia. *Mol Hum Reprod* 12: 677–685.

90. Ulrich M, Petre A, Youhnovski N, Prömm F, Schirle M, Schumm M, Pero RS, Doyle A, Checkel J, Kita H, Thiyagarajan N, Acharya KR, Schmid-Grendelmeier P, Simon HU, Schwarz H, Tsutsui M, Shimokawa H, Bellon G, Lee JJ, Przybylski M, Döring G. 2008. Post-translational tyrosine nitration of eosinophil granule toxins mediated by eosinophil peroxidase. *J Biol Chem* 283: 28629–28640.

91. Hamilton RT, Asatryan L, Nilsen JT, Isas JM, Gallaher TK, Sawamura T, Hsiai TK. 2008. LDL protein nitration: implication for LDL protein unfolding. *Arch Biochem Biophys* 479: 1–14.

92. Zhu JH, Zhang X, Roneker CA, McClung JP, Zhang S, Thannhauser TW, Ripoll DR, Sun Q, Lei XG. 2008. Role of copper, zinc-superoxide dismutase in catalyzing nitrotyrosine formation in murine liver. *Free Radic Biol Med* 45: 611–618.

93. Reed TT, Owen J, Pierce WM, Sebastian A, Sullivan PG, Butterfield DA. 2009. Proteomic identification of nitrated brain proteins in traumatic brain-injured rats treated postinjury with gamma-glutamylcysteine ethylester: insights into the role of elevation of glutathione as a potential therapeutic strategy for traumatic brain injury. *J Neurosci Res* 87: 408–417.

94. Casanovas A, Carrascal M, Abián J, López-Tejero MD, Llobera M. 2009. Lipoprotein lipase is nitrated in vivo after lipopolysaccharide challenge. *Free Radic Biol Med* 47: 1553–1560.

95. Sharov VS, Galeva NA, Dremina ES, Williams TD, Schöneich C. 2009. Inactivation of rabbit muscle glycogen phosphorylase b by peroxynitrite revisited: does the nitration of Tyr613 in the allosteric inhibition site control enzymatic function? *Arch Biochem Biophys* 484: 155–166.

96. Ohama T, Brautigan DL. 2010. Endotoxin conditioning induces VCP/p97-mediated and inducible nitric-oxide synthase-dependent Tyr284 nitration in protein phosphatase 2A. *J Biol Chem* 285: 8711–8718.

97. Sekar Y, Moon TC, Slupsky CM, Befus AD. 2010. Protein tyrosine nitration of aldolase in mast cells: a plausible pathway in nitric oxide-mediated regulation of mast cell function. *J Immunol* 185: 578–587.

98. Redondo-Horcajo M, Romero N, Martínez-Acedo P, Martínez-Ruiz A, Quijano C, Lourenço CF, Movilla N, Enríquez JA, Rodríguez-Pascual F, Rial E, Radi R, Vázquez J, Lamas S. 2010. Cyclosporine A-induced nitration of tyrosine 34 MnSOD in endothelial cells: role of mitochondrial superoxide. *Cardiovasc Res* 87: 356–365.

99. Chen HJ, Chen YC. 2012. Reactive nitrogen oxide species-induced post-translational modifications in human hemoglobin and the association with cigarette smoking. *Anal Chem* 84: 7881–7890.

100. Bigelow DJ, Qian WJ. 2008. Quantitative proteome mapping of nitrotyrosines. *Methods Enzymol* 440: 191–205.

101. Tao RR, Huang JY, Shao XJ, Ye WF, Tian Y, Liao MH, Fukunaga K, Lou YJ, Han F, Lu YM. 2013. Ischemic injury promotes Keap1 nitration and disturbance of antioxidative responses in endothelial cells: a potential vasoprotective effect of melatonin. *J Pineal Res* 54: 271–281.

102. Safinowski M, Wilhelm B, Reimer T, Weise A, Thomé N, Hänel H, Forst T, Pfützner A. 2009. Determination of nitrotyrosine concentrations in plasma samples of diabetes mellitus patients by four different immunoassays leads to contradictive results and disqualifies the majority of the tests. *Clin Chem Lab Med* 47: 483–488.

103. Lu N, Zhang Y, Li H, Gao Z. 2010. Oxidative and nitrative modifications of alpha-enolase in cardiac proteins from diabetic rats. *Free Radic Biol Med* 48: 873–881.

104. Cecconi D, Orzetti S, Vandelle E, Rinalducci S, Zolla L, Delledonne M. 2009. Protein nitration during defense response in Arabidopsis thaliana. *Electrophoresis* 30: 2460–2468.

105. Hui Y, Wong M, Zhao SS, Love JA, Ansley DM, Chen DD. 2012. A simple and robust LC-MS/MS method for quantification of free 3-nitrotyrosine in human plasma from patients receiving on-pump CABG surgery. *Electrophoresis* 33: 697–704.

106. Smallwood HS, Lourette NM, Boschek CB, Bigelow DJ, Smith RD, Pasa-Tolic L, Squier TC. 2007. Identification of a denitrase activity against calmodulin in activated macrophages using high-field liquid chromatography—FTICR mass spectrometry. *Biochemistry* 46: 10498–10505.

107. Mani AR, Moore KP. 2005. Dynamic assessment of nitration reactions in vivo. *Methods Enzymol* 396: 151–159.

108. Lin HL, Kenaan C, Zhang H, Hollenberg PF. 2012. Reaction of human cytochrome P450 3A4 with peroxynitrite: nitrotyrosine formation on the proximal side impairs its interaction with NADPH-cytochrome P450 reductase. *Chem Res Toxicol* 25: 2642–2653.

109. Lin HL, Myshkin E, Waskell L, Hollenberg PF. 2007. Peroxynitrite inactivation of human cytochrome P450 2B6 and 2E1: heme modification and site-specific nitrotyrosine formation. *Chem Res Toxicol* 20: 1612–1622.

110. Yamakura F, Kawasaki H. 2010. Post-translational modifications of superoxide dismutase. *Biochim Biophys Acta Proteins Proteomics* 1804: 318–325.

111. Kanski J, Schöneich C. 2005. Protein nitration in biological aging: proteomic and tandem mass spectrometric characterization of nitrated sites. *Methods Enzymol* 396: 160–171.

112. Spickett CM, Pitt AR. 2012. Protein oxidation: role in signaling and detection by mass spectrometry. *Amino Acids* 42: 5–21.

113. Tsikas D. 2012. Analytical methods for 3-nitrotyrosine quantification in biological samples: the unique role of tandem mass spectrometry. *Amino Acids* 42: 45–63.

68

FLUORESCENCE SPECTROSCOPY

Yevgen Povrozin and Beniamino Barbieri
ISS, Champaign, IL, USA

The number of fluorescence technique applications has been continuously growing over the last 20 years. While initially intended as an analytical tool for the determination of the presence of specific molecules in solutions, fluorescence is now routinely used in biochemistry and biophysics for studying molecular interactions and dynamics, both in solutions and in cells; in clinical immunoassays for the determination of the presence of specific antibodies and antigens; in drug discovery; in life sciences for DNA sequencing; and in nanotechnology and materials sciences for identification and characterization of new materials.

The reasons of the continuing increase in popularity are multiple: on one hand, it is due to the improvements in the sensitivity of the instrumentation that allows now for the observation of single-molecule events on a routine basis; on the other hand, the interface of the instrumentation with the personal computer has increased the automation of the data collection and the sophistication of the data analysis. A third reason for its increased success is due to the introduction in the past 30 years of innumerable and specific chemical probes used as markers for compounds that either do not display fluorescence or only emit a low level of it. The extent of the applications has benefited from the development of the green fluorescent protein (GFP) family that allows for the expression of fluorescent proteins in cells and tissues, a feature that allows the experimenter to follow the whereabouts of proteins in live cells and even tissues in live animals.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

Paradoxically, the capabilities of the instrumentation coupled to the computation power of the computer brings new challenges to the field, as novel practitioners are not always aware of the potential pitfalls that lie behind an experiment. In the past few years several articles and books have been published on the subject describing in detail the applications of the fluorescence techniques to the chemical and life sciences. A brief article cannot cover such details; our goal is rather to reiterate the fundamental principles of the technique and to mention some of the common pitfall that a user of the technique may encounter.

68.1 OBSERVABLES MEASURED IN FLUORESCENCE

Fluorescence is generally referred to as the emission of photons from a sample following the absorption of photons. There are other means for producing fluorescence in a sample (bioluminescence, sonoluminescence, and electroluminescence), but in the following we will refer exclusively to the phenomenon originated by the absorption of light.

Fluorescence is part of a general class of phenomena named *luminescence*; it is distinguished by the *phosphorescence* as the latter takes, typically, a time of the order of 1 µs (10^{-6} s) or longer, while the former takes a time of the order of 1 ns (10^{-9} s). As we will see in the following, the distinction between the two is described using the more precise terminology of quantum mechanics.

The main five parameters measured in fluorescence spectroscopy are:

1. Excitation spectrum
2. Emission spectrum
3. Decay times (fluorescence lifetimes)
4. Quantum yield
5. Anisotropy (or polarization)

Recent advancements in fluorescence microscopy have introduced the measurement of additional parameters (diffusion correlation times, brightness), but we will limit our discussion in this chapter to the five parameters listed earlier, which are measurable using a spectrofluorometer.

The description of the fluorescence measurable parameters is best understood with the introduction of the Perrin–Jabłoński diagram, which is a quantum mechanics representation of the energy levels of a molecular structure.

68.2 THE PERRIN–JABŁOŃSKI DIAGRAM

Figure 68.1 shows a classic representation of the electronic levels of a molecule in solution or in the gas phase (in solid phase the energy levels collapse into "bands" although the basic concepts are still valid).

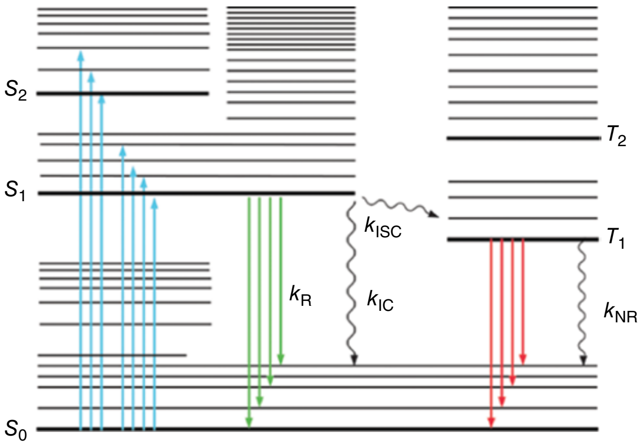


FIGURE 68.1 Perrin–Jabłoński energy diagram for a molecular structure. Singlet states are indicated by S_0, S_1, …, and triplet states by T_1, T_2, …. Internal conversion rate is k_{IC}; intercrossing conversion rate between singlet and triplet states is k_{ISC}; the fluorescence decay rate is k_R, while the nonfluorescence rate is k_{NR}.

The energy levels occupied by an electron are named "singlet states," and the letters S_0, S_1, S_2, …, indicate the ground state, the first excited state, etc.; upon absorption of a photon, an electron moves from the ground state S_0 to the excited states. Associated with each electronic level, there are several vibrational and rotational levels, which differ in energy by a smaller amount than the corresponding electronic levels.

Moreover, there are energy transitions that are not directly allowed (forbidden transitions). They are identified as "triplet states" and indicated by T_1, T_2, …, etc.; they also feature associated vibrational and rotational levels.

The absorption probability of a photon in each electronic level is described within the framework of quantum mechanics (energy separation between the levels and momentum and spin of the various levels). The molecules interact when in the presence of photons of the appropriate photon energy E, where

$$E = h\nu = h\frac{c}{\lambda} \tag{68.1}$$

In the relation, h is the Planck constant ($6.626 \times 10^{-34}\,\mathrm{J\,s}$), c is the speed of light ($2.9979 \times 10^{8}\,\mathrm{m\,s^{-1}}$), while v and λ are the frequency and wavelength of the electromagnetic wave describing the photon.

For absorption to occur, E has to be of the order of magnitude of the separation between the excited level and the ground state; that is,

$$E \approx E_{S_1} - E_{S_0} \tag{68.2}$$

Let us consider a population of N molecules in a solution. Upon absorption of photons (group of line to the left in the figure), a fraction of the molecules undergo a transition from the ground state S_0 to the upper electronic states, S_1, S_2, the final state depending ultimately by the energy of the absorbed photon. The absorption process takes an amount of time of the order of the femtosecond (10^{-15} s) or shorter.

Once in the excited electronic level, the molecules relax fairly rapidly (about 10^{-12} s) to the lowest level of the first excited state S_1; hence, they decay with rate k_R to emit fluorescence (group of lines in the middle in the figure). The characteristic time of the fluorescence is of the order of 1 ns (10^{-9} s).

There are additional decay routes that are not necessarily associated with the emission of photons; they are indicated by k_IC (internal conversion between two electronic states of the same spin multiplicity) and k_ISC (intersystem crossing conversion between the S levels and T levels). It is noteworthy that the excited level T_1 (triplet state) emits photons; this process is usually termed “phosphorescence,” and its characteristic time, as mentioned previously, is of the order of 1 μs (10^{-6} s) and longer.

The Perrin–Jabłoński diagram (Fig. 68.1) is instrumental to determine the law describing the decay time of fluorescence. If N_1 is the population of the excited level S_1, upon absorption of photons the population of the level changes is described by the relation

$$\frac{dN_1}{dt} = -\left(k_\mathrm{R} + k_\mathrm{NR}\right)N_1 + f_1 \tag{68.3}$$

where f_1 is a function that describes the process of the excitation photons (pulsed source, continuous-wave (CW) source, etc.). By solving the equation (and disregarding f_1), we find that

$$N_1 = N_1\left(0\right)e^{-\left(t/\tau_\mathrm{S}\right)} \tag{68.4}$$

where τ_S the decay time of the excited state S_1 is defined as

$$\tau_\mathrm{S} = \frac{1}{k_\mathrm{R} + k_\mathrm{NR}} \tag{68.5}$$

The fluorescence quantum yield is the fraction of excited molecules that return to the ground state with the emission of fluorescence. From direct examination of the Perrin–Jabłoński diagram, one simply divides the rate of radiative emission k_R by the total rates of deactivation, which includes both the radiative and nonradiative contributions:

$$\Phi = \frac{k_\mathrm{R}}{k_\mathrm{R} + k_\mathrm{NR}} \tag{68.6}$$

By using the definition of decay times, the quantum yield can also be expressed in terms of lifetimes:

$$\Phi = \frac{\tau_S}{\tau_R} \tag{68.7}$$

One can say that the quantum yield is the ratio of the number of emitted photons over the total number of absorbed photons.

The five measurable parameters of fluorescence are usually used to describe these processes, namely, the range in wavelengths of the absorption and emission of photons (excitation and emission spectra), the orientation changes during the time the molecules are in the excited states between absorption and emission of the photons (anisotropy or polarization), the fraction of photons emitted over the number of photons absorbed (quantum yield), and the emission rate (decay times). After a brief overview of the instrumentation, we will examine in detail the measurement of the five parameters.

68.3 INSTRUMENTATION

The peculiar parameters that characterize fluorescence are measured using "spectrofluorometers"; sometimes, instruments for the measurement of excitation and emission spectra are termed "spectrofluorimeters," while the ones for the measurements of the decay times are termed "spectrofluorometers." Yet, the distinction is not anymore as clearly demarked as several instruments allow, in the same unit, to measure both the steady-state (excitation and emission spectra) and the dynamic (decay times and rotational correlation times) properties of the fluorescence.

Usually, in all of the instruments, the fluorescence is collected at an angle of 90° with respect to the optical axis set by the excitation light beam. This geometry maximizes the efficiency of the emission collection and reduces the background due to the excitation light.

It is worthy to mention that absorption spectra can be measured using a spectrophotometer. In this type of instrument, the light detector is placed on the same optical axis of the excitation light beam, and the instrument detects the amount of light that is being transmitted (i.e., not absorbed) through the sample. A spectrophotometer measures the difference in the intensity of two signals (typically, sample transmittance is compared to 100% transmittance); instead, a spectrofluorometer measures a signal (the fluorescence) over a zero background.

The key elements of a spectrofluorometer are the light source, the monochromator, and the light detector.

68.3.1 Light Source

The typical light source utilized in a spectrofluorometer is a high-pressure xenon arc lamp. The bulb of this lamp includes xenon at high pressure that is excited to higher level by the electrical arc established by the current running through the electrodes. The emitted light is a continuous spectrum from (depending upon the models and geometries) about 250 up to 1100 nm. Figure 68.2 displays the spectrum of the lamp utilized by ISS. Although the spectrum is relatively flat up to about 800 nm, several sharp resonances are present above that wavelength.

It is worth noting that a variation of this lamp is the Hg–Xe lamp, which contains traces of mercury; this element displays resonances at around 295 nm, and this feature allowed for its use as an excitation source for the proteins containing tryptophan.

In the past several years, lasers have replaced the xenon arc lamp, specifically for time-resolved applications. Although they emit radiation only at specific wavelengths, their brightness is order of magnitude higher than that of the lamp. In addition, they can be pulsed with fairly narrow pulse widths (about 50 ps for the laser diodes). A recent advancement is the supercontinuum laser (or white laser) that delivers any wavelength in the range from 390 up to 2000 nm, featuring 5 ps pulse width and (in the model made by Fianium Ltd) the option of selecting the repetition rate up to 40 MHz.

Light-emitting diodes (LEDs) are also utilized as light sources especially in the region from 240 to 350 nm, where lasers are not available (with exceptions at 266, 315, 325 nm). They are compact, relatively inexpensive, and the source of choice when building an instrument dedicated to a specific application.

68.3.2 Monochromator

Monochromators are utilized to select the wavelength used for irradiating the sample when using a xenon arc lamp; in the collection channel of a spectrofluorometer, they are utilized to record the range of wavelengths emitted by a fluorophore (emission

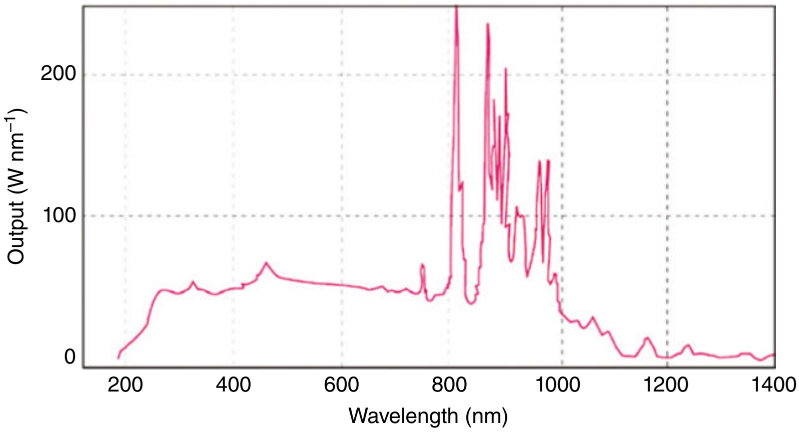


FIGURE 68.2 Spectral distribution for the 300 W xenon arc lamp. Source: Reproduced with permission of ISS.

spectrum, see Section 68.5.2). The simplest monochromator includes a diffraction grating and slits at the entrance and at the output. Light impinging at an angle on the grating is diffracted at a series of angles; usually, the first angle (or first order) is selected for the measurement.

It is important to realize that the transmission of the light traversing a monochromator is affected by two parameters:

1. The wavelength—the grating has a specific transmission curve, and some wavelengths are transmitted with a higher efficiency than other wavelengths, a feature to remember when collecting excitation and emission spectra.
2. The polarization status of the radiation—the grating of the monochromator transmits differently radiation with different planes of polarization.

Moreover, it is important to remember that when a monochromator is set to deliver radiation at wavelength λ, it also delivers radiation at 2λ (second order); as an example, if the excitation monochromator is set at 300 nm, it delivers radiation at 600 nm too. Typically the intensity of the second order is about 1/10 the intensity of the first order; still this amount is sufficient to contaminate the emission spectrum. The second order can be eliminated with a judicial selection of filters.

A characterization of every monochromator is the amount of stray light, that is, radiation present at any wavelength other than the specific wavelength the monochromator is set at. The stray light is usually measured as the amount of light that is transmitted outside the band pass of the 632.8 nm HeNe laser line. For typical holographic gratings it is 10^{-5} the intensity of the line. While this amount is not typically important for the study of fluorophores in thin solutions, it becomes important when the sample is in a turbid solution or even a solid state. Different strategies are available for the minimization of the stray light, the first being a judicial selection of the grating. Gratings are classified depending upon their fabrication process: the ruled gratings and holographic gratings, with the latter displaying less stray light inhomogeneity as the grooves are formed through the interference process of two laser beams in a photosensitive material, while in the former the grooves are formed mechanically.

Gratings can be arranged in different designs to build a monochromator, the two more popular being the Czerny–Turner and the Seya–Namioka.

68.3.3 Light Detectors

In all the instruments the fluorescence signal is converted into current by a photomultiplier tube (PMT) or photodiode (instruments for lifetime measurements may utilize other types of detectors too, such as hybrid PMTs, microchannel plate detectors, or streak cameras). PMTs are sensitive within a set wavelength range that is determined by the material used in the photocathode. Figure 68.3 displays the region of sensitivity for the PMT Model R928 by Hamamatsu. The PMT can be utilized in the region from

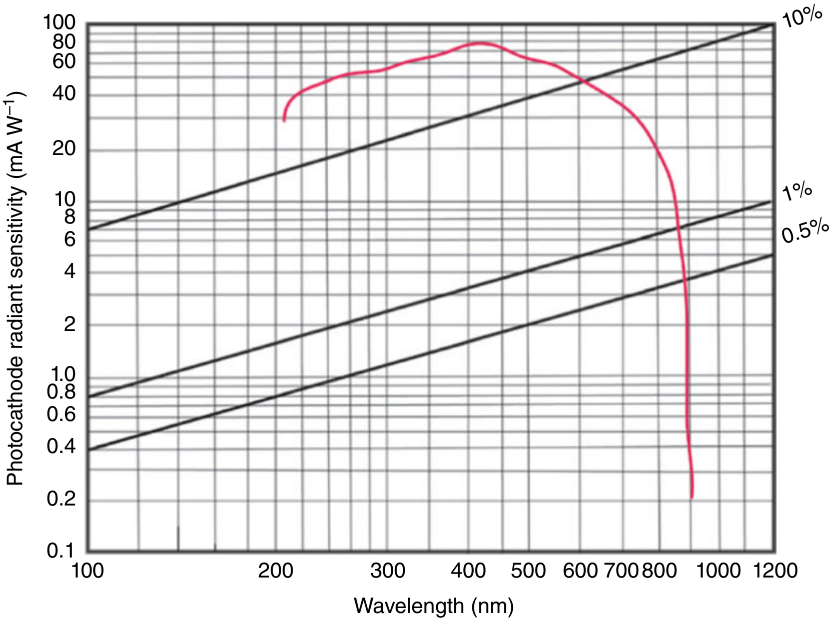


FIGURE 68.3 Wavelength range for a photomultiplier tube model R928 by Hamamatsu.

about 230 nm to about 830 nm. It is apparent that even within the operational wavelength region, the sensitivity is not the same; the nonlinearity of the sensitivity introduces an artifact in the data such that a correction to the data has to be introduced.

A spectrofluorometer includes other optical elements such as lenses and mirrors; moreover polarizers are utilized for anisotropy measurements. The operational region of the instrument is given by the superposition of the wavelength transmission of the various elements of the instruments. Even within this region, the variation in transmission has to be taken into account when measuring the fluorescence parameters. The procedures will be outlined in the measurement sections later.

Figure 68.4 displays the technical diagram of the K2 multifrequency phase fluorometer (MPF) made by ISS, an instrument capable of measuring all of the relevant fluorescence parameters.

The standard light source is a 300 W xenon arc lamp. CW lasers, pulsed lasers (including the multiphoton laser), and LEDs can be coupled to the K2 as well; typically these sources are utilized for the measurement of the decay times of fluorescence.

The light emitted by the source travels through the excitation channel that comprises the monochromator, a filter holder, and the polarizer holder; the monochromator selects the wavelength of the light that excites the sample. The fluorescence emitted by the sample is collected through the left or the right channels; the right channel includes the emission monochromator.

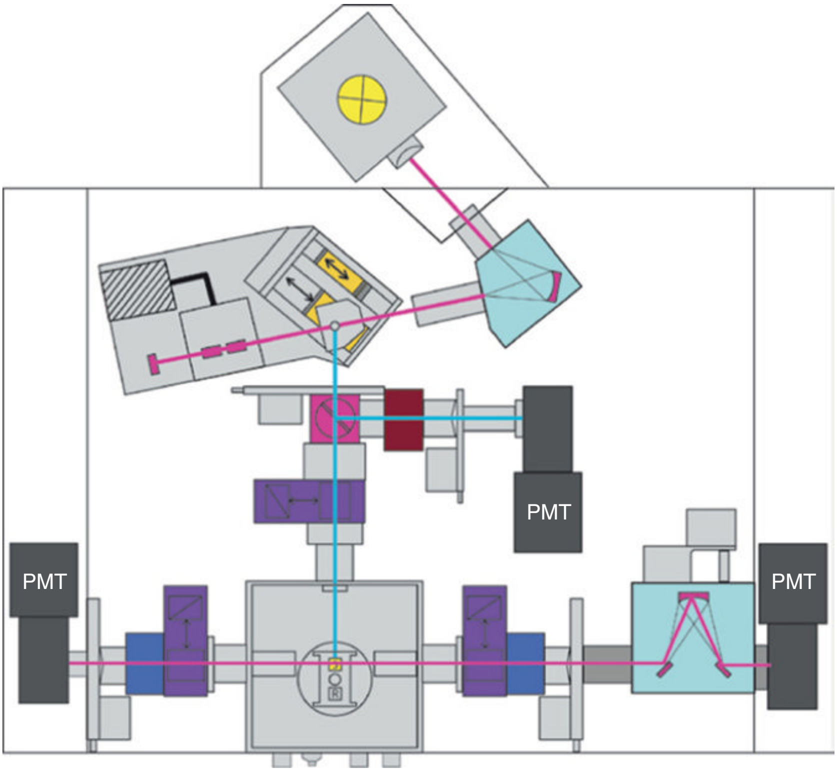


FIGURE 68.4 K2 multifrequency phase and modulation spectrofluorometer. Source: Reproduced with permission of ISS.

The instrument includes polarizer holders, filter holders, shutters for blocking the light from reaching the sample, and the detectors. All of these components are required for automated measurement acquisition.

68.3.4 Instrumentation for Steady-State Fluorescence: Analog and Photon Counting

Two general schemes are utilized to process the signal collected by the PMT: in one scheme, named *analog detection*, the signal from the PMT goes through a current-to-voltage converter, an amplifier, and finally, it is digitized by an analog-to-digital converter (ADC). The signal is then displayed on, and/or stored in, the computer.

In another scheme, named *photon counting detection*, the signal from the PMT goes through an amplifier discriminator that allows for the selection of pulses over a set threshold. A counter in the processing unit counts the number of photons collected per seconds by the detector. This parameter is then displayed by the software on, or stored in, the computer.

Although the advantage of analog detection is in the capability of processing signals within a high dynamic range and fast response, its overall sensitivity is lower than the sensitivity of photon counting detection. Ultimately the choice of one scheme over the other depends upon the specific application.

68.3.5 The Measurement of Decay Times: Frequency-Domain and Time-Domain Techniques

The instrumentation for the measurement of fluorescence decay times is broadly classified as belonging to one of two groups, time-domain and frequency-domain techniques.

The time-domain technique includes the single photon counting, the multiscaler, and the time-correlated single photon counting (TCSPC); the TCSPC is usually the technique utilized more often. The frequency-domain technique comes in an *analog* version (AFD) and a *digital* version (DFD), which has just been introduced.

In TCSPC, a photon is counted within a set time period with a high precision (Fig. 68.5). The time period is defined by the intervals between the pulses of the

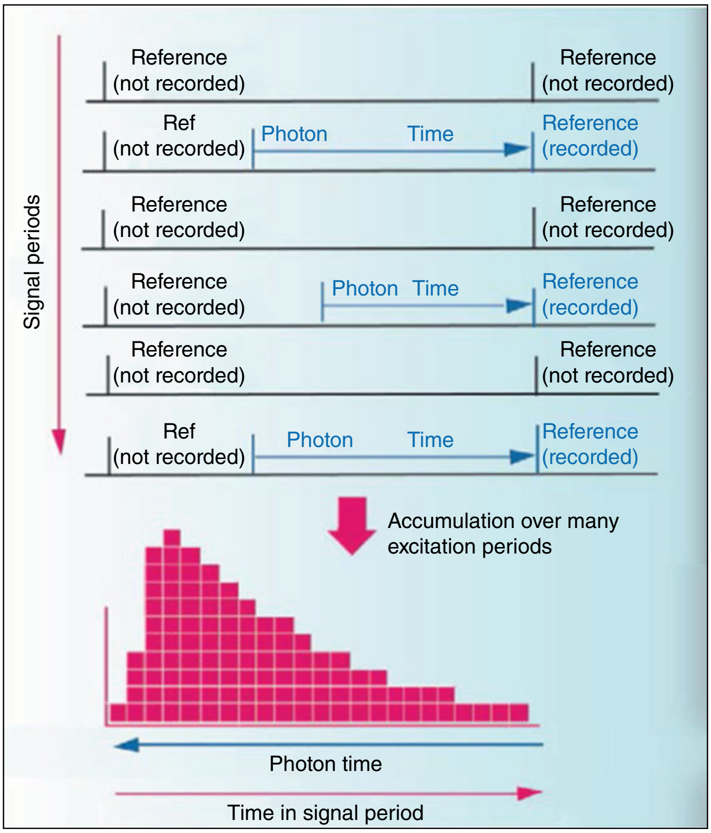


FIGURE 68.5 Principle of start–stop mechanism utilized in TCSPC data acquisition.

excitation light (repetition rate of the light source), and the precision is given by the acquisition electronics (mainly the time-to-amplitude converter (TAC) and the ADC components). For instance, when using an excitation light, emitting pulses at 80 MHz, the time period is the distance between two such pulses (12.5 ns). Typically, the repetition rate of some light sources can be set by the user.

At the arrival of each pulse on the light detector, a high precision timer is triggered, which records how much time has passed between the arrival of the excitation pulse and the emitted photon.

The TAC unit produces a signal proportional to the arrival time of the photon; different arrival time records are grouped in different memory locations (bins) of computer memory.

To interpret the lifetime time information obtained by a TCSPC instrument, a histogram of the arrival time records is built. For a single exponential decay, a curve similar to the one of Equation 68.4 is collected, and the decay time τ is determined using a minimization technique to fit the experimental data to the theoretical decay model.

The frequency-domain technique is more versatile as it can perform either with pulsed sources used for TCSPC or with the modulation of the excitation light source: the modulated excitation results in a modulated fluorescence with a phase and modulation, which is dependent on the lifetime of the excited fluorophores.

The instruments utilized in frequency-domain technique are called MPF or, simply, frequency-domain fluorometers. The underlying operational principle of an MPF is illustrated by Figure 68.6 for a CW source. The excitation light $E(t)$ is modulated at a frequency ω; its modulation is characterized by an alternating component AC_{EX} and an

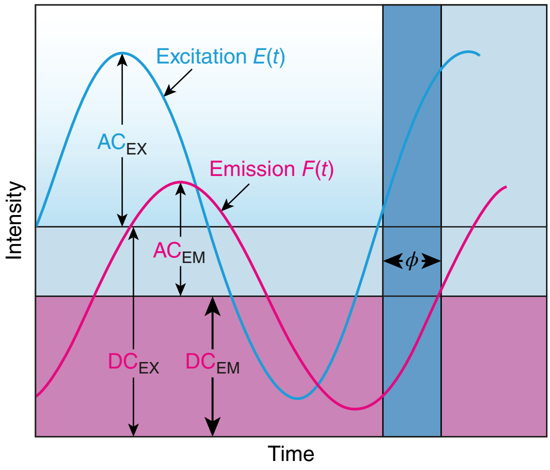


FIGURE 68.6 Schematics of the excitation and emission light in frequency-domain spectroscopy; the emission light is phase shifted and demodulated with respect to the excitation light.

average component DC_{EX}. The fluorescence light is modulated at the same frequency ω, but its phase is delayed by the quantity ϕ, and the overall modulation $(AC/DC)_{EM}$ is less than the original modulation of the excitation light. A frequency-domain instrument measures the phase shift ϕ and the demodulation m of the fluorescence; both quantities are related to the decay time (see Equations 68.8 and 68.9). For a single exponential decay, the decay time is related to the phase angle and to the modulation by the following relations:

$$\tau_P = \frac{1}{\omega}\tan\phi \tag{68.8}$$

$$\tau_M = \frac{1}{\omega}\sqrt{\frac{1}{m^2} - 1} \tag{68.9}$$

Such measurements are repeated at several different values of the modulation frequency ω ranging typically from two or three for a single exponential decay to up to 20–25 for multiple exponential decays. The decay times τ_i are determined using a minimization technique to fit the experimental data.

The first modern frequency-domain instrument has been introduced by Spencer and Weber in 1969 [1]. In this instrument the light source is modulated at a frequency ω, and the light detector is modulated at a frequency $(\omega + \Delta\omega)$; the two frequencies are being provided by phase-locked frequency synthesizers. The approach is also known as “heterodyning.” The output signal includes components at the sum (2ω) and the difference $(\Delta\omega)$ frequency; the low signal component $\Delta\omega$, called the “cross-correlation frequency,” which is typically in the range from 1 Hz to 20 kHz, is utilized to determine the phase shift and the demodulation of the fluorescence. From the phase and modulation of the $\Delta\omega$ frequency, the phase and the modulation of the fluorescence can be determined relative to that of a reference lifetime.

68.4 FLUOROPHORES

Generally fluorophores are divided into *intrinsic* and *extrinsic*. Intrinsic fluorophores are the natural components of a system (typically biological macromolecule) that exhibit fluorescence that can be measured, for instance, the aromatic amino acids tyrosine, tryptophan, and phenylalanine of the proteins, NADH, the flavins, and the porphyrins-based compounds such as chlorophylls. Extrinsic probes include all those molecules that are foreign to the system or were added to it artificially (fluorescent probes and labels—organic dyes, quantum dots, or biological fluorophores), such as fluorescein and 1,8-anilinonaphthalene sulfonic acid (ANS), which are introduced by the experimenter. Such molecules can be covalently linked to the molecule under study or noncovalently as in the case for diphenylhexatriene (DPH) used to study membranes.

68.5 MEASUREMENTS

68.5.1 Excitation Spectrum

The excitation spectrum displays the emission intensity distribution at one wavelength while scanning the excitation wavelength over a range of wavelengths (Fig. 68.7). Practically, for the acquisition of the excitation spectrum, the emission monochromator of the spectrofluorometer is set at a fixed wavelength (in the sample emission range), and the excitation monochromator is scanned over a range of wavelengths (the range that corresponds to the sample absorption range). Referring to the Perrin–Jabłoński diagram of Figure 68.1, when acquiring the excitation spectrum, one detects photons emitted by the molecules at a set wavelength (represented by one of the lines in the group at the middle of the figure) while scanning the wavelength of the radiation (energy of photons) sent to the sample from high energy to low energy (the group of lines at the left of the figure).

If there are no changes that occur to the molecule in the excited state, then the excitation spectrum closely resembles the absorption spectrum acquired with a spectrophotometer; yet, in most instances, it does not: in order for the two to match, a suitable correction of the instrumental factors has to be applied. The main culprit of the differences is due to the lamp; it features a peculiar emission spectrum of its own, that is, the intensity of the emitted radiation is not constant at all the wavelengths. In order to

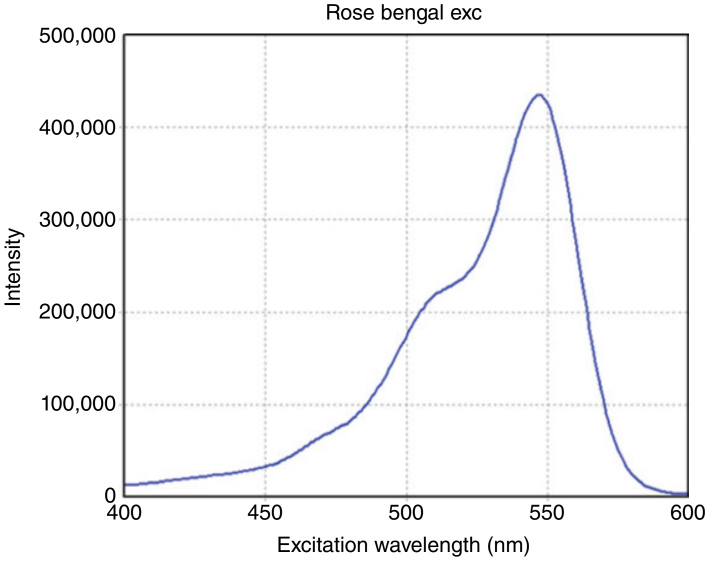


FIGURE 68.7 Excitation spectrum of rose bengal in a water solution, acquired using the K2 spectrofluorometer. The spectrum was acquired by scanning the excitation monochromator from 400 to 600 nm in steps of 1 nm; at each position data were acquired for 1 s. The fluorescence was observed at 610 nm. Source: Reproduced with permission of ISS.

correct for this effect, a small fraction of the excitation light is diverted in the reference channel of the spectrofluorometer (Fig. 68.4) where it passes through the quantum counter and it is collected by the reference detector. The quantum counter, usually a stable fluorophore at a high concentration in solution, delivers a number of photons proportional to the absorbed signal; therefore, at each wavelength, we have a signal proportional to the signal emitted by the lamp; this signal is utilized to correct the fluorescence signal collected in the emission channel. Although this correction addresses most of the concerns, it does not completely correct the excitation spectrum as the beam splitter utilized to divert part of the excitation light into the reference channel reflects differently the two planes of polarization. For a full correction to be implemented, one should place a cuvette with a scattering solution in the sample compartment and acquire an emission spectrum over the wavelength range of interest and then acquire the emission spectrum of the fluorophore and divide it by the spectrum of the scatterer. In this way, the excitation spectrum is fully corrected. Practically, the correction introduced by using the quantum counter and the reference channel is sufficient; one should nonetheless specify the experimental conditions when publishing the spectrum.

68.5.2 Emission Spectrum

The emission spectrum of a fluorophore is most likely the most popular experimental measurement carried out in fluorescence (Fig. 68.8). The spectrum is acquired by setting the excitation wavelength at a fixed value (one of the lines belonging to the group to the left of the Fig. 68.1) and then by scanning the emission monochromator over a range of emission wavelengths (group of lines in the middle of the Fig. 68.1).

There are a few general rules that apply to emission spectra:

1. The emission of fluorescence occurs at wavelengths longer than the excitation wavelength (Stokes shift).
2. The shape of the emission spectrum does not change by changing the excitation wavelength.
3. The emission spectrum is a mirror image of the excitation spectrum of lower energy.

An examination of Figure 68.1 explains as to why the first rule holds. When the molecules are excited, they relax to the lowest vibrational level of the excited states, and from there, they emit fluorescence. Fluorescence photons have a lower energy than excitation photons (i.e., the fluorescence occurs at longer wavelengths than the excitation). Hence, we also gather that the shape of the emission spectrum does not change by changing the excitation wavelength. Finally, rule 3 establishes that the emission spectrum ($S_1 \rightarrow S_0$ transition) is a mirror image of the absorption transition involving the same levels ($S_0 \rightarrow S_1$ transition). If the excitation spectrum includes transitions to higher levels, the emission spectrum will not be a mirror image of the excitation. There are exceptions to the mirror image rule: for instance, when *p*-terphenyl is excited the

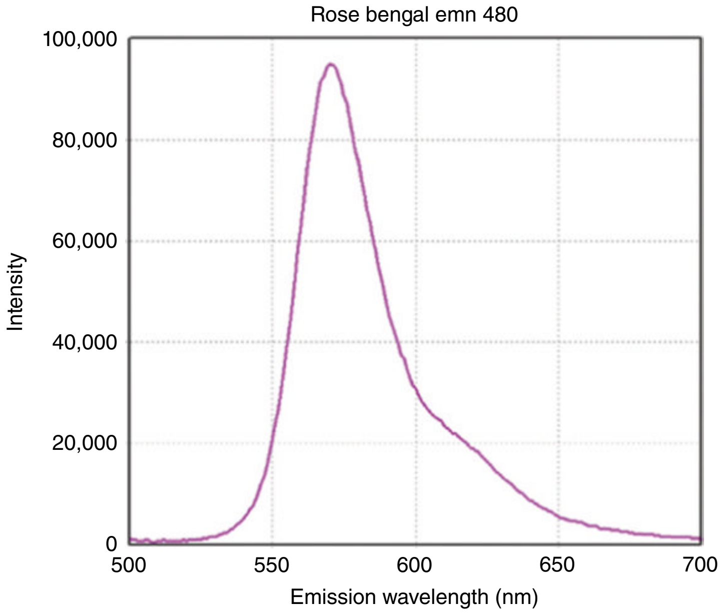


FIGURE 68.8 Emission spectrum of rose bengal in a water solution, acquired using the K2 spectrofluorometer. The excitation monochromator was set at 490 nm. The emission spectrum was acquired by scanning the emission monochromator from 500 to 700 nm in steps of 1 nm; at each position data were acquired for 1 s. Source: Reproduced with permission of ISS.

nuclei undergo a geometric rearrangement upon absorption, and the emission spectrum shows the additional vibrational structure. Excited-state reactions can also result in emission spectra that mark a departure from the mirror rule and so the formation of complexes (for instance, pyrene).

As for the excitation spectrum, the emission spectrum is affected by experimental artifacts, namely, the transmission of the emission monochromator and the sensitivity of the light detector: The transmission of the monochromator varies with the wavelengths, and moreover, it features different transmissions for the two planes of polarization of the light (see later for the definition of light polarization); the sensitivity of the light detector varies with the wavelength. All these variations have to be accounted for in order to acquire a "true" emission spectrum. To this respect, one distinguishes between technical spectrum (the spectrum acquired by an instrument) and the corrected spectrum (the technical spectrum that has been corrected for the experimental artifacts). Manufacturers typically provide correction files for an instrument; these factors are embedded in the software, and corrected spectra can be acquired online; or spectra can be corrected afterward. Practically, one does not need to correct a spectrum unless it is meant for publications; even in that event, it is completely acceptable to specify that the spectrum is a *technical* spectrum rather than a *corrected* one. There are some instances when corrected spectra are required; when calculating the quantum yield of a fluorophore, one has to calculate the area under the spectrum; the spectrum has to be corrected for providing the proper value. Another instance occurs when using the

Förster resonance energy transfer (FRET), a useful tool for estimating the distances between two interacting and close fluorophores.

Besides the instrumental artifacts, the emission spectra are sometimes distorted by experimental artifacts that a practitioner of the field needs to be aware of, namely:

1. Background fluorescence
2. The second order of the monochromator
3. The Raman spectrum of water

Background fluorescence occurs when the fluorophore is diluted in a solution, and the solvent (e.g., buffer) emits some fluorescence of its own at the emission wavelength utilized in the experiment; the resulting emission spectrum is the superposition of the individual spectra of the solvent and the fluorophore. In this case, one can acquire the emission spectrum of the solvent alone and subtract it from the emission spectrum of the solution in order to obtain the emission spectrum of the fluorophore.

We mentioned about the second order in the paragraph covering the monochromators: when a monochromator is set to deliver radiation at wavelength λ, it also delivers radiation at 2λ (second order); although the intensity is about 1/10 of the intensity of the first order, it is sufficient to introduce distortions when measuring turbid solutions and solid samples. The second order can be eliminated with a judicial selection of filters.

Finally, when working with water as a solution, the Raman peaks are present at a wavelength that is 3400 cm^{-1} longer than the excitation wavelength:

$$\lambda_{\mathrm{Ex}}^{-1} - \lambda_{\mathrm{R}}^{-1} = 3400\,\mathrm{cm}^{-1} \tag{68.10}$$

As an example, when exciting at 300 nm an emission peak appears at 334 nm; when exciting at 350 nm, an emission peak appears at 397 nm. Note that, while the position of the peak is fixed in unit of wavenumbers ($1/\lambda$), the position varies when dealing in wavelengths (λ); the change in the peak position with the change of the excitation wavelength allows for the user to discern the peak from other peaks or artifacts. The intensity of the Raman peak provides a simple tool to verify the status of the light source of the spectrofluorometer; measured periodically, one can have a pretty good idea of the derating of the xenon arc lamp and make a decision as to when to replace the lamp.

68.5.3 Decay Times of Fluorescence

The fact that the decay times of many fluorophores are in the range of 1–30 ns is truly amazing as this time scale is typical of molecular interactions in biological systems (enzyme conformational shifts, rotational motions in proteins, photosynthetic reactions, etc.) in physiologically active systems.

The decay time is affected by many parameters of the microenvironment (temperature, ions, polarity, viscosity, electric fields), and this is the reason it is widely utilized

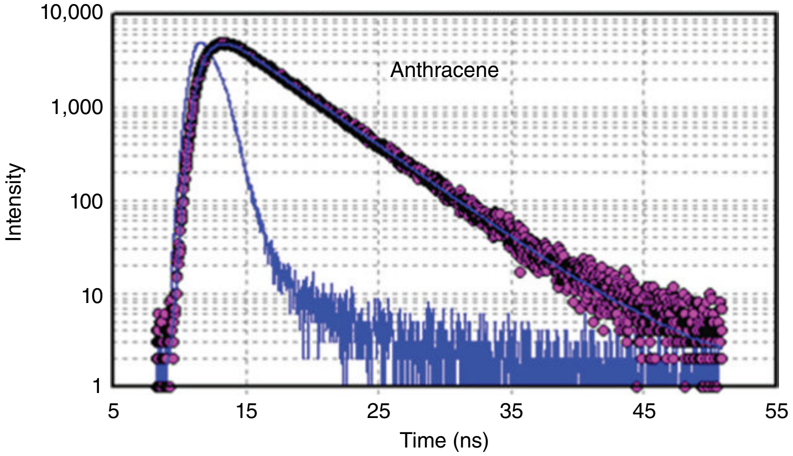


FIGURE 68.9 Decay curve of anthracene in ETOH using a TCSPC instrument (ChronosBH, by ISS). Source: Reproduced with permission of ISS.

for studying molecular interactions. For instance, the decay time of ANS in water is about 100 ps; when ANS is bound to a protein, the lifetime is 8–10 ns. The lifetime of ethidium bromide is 1.8 ns in water; it is 22 ns when bound to DNA and 37 ns when bound to tRNA.

Finally, the lifetimes can be used an analytical tool as well for the characterization of the presence of specific dyes or simply for the quantitation of complex fluorescent mixtures (the type of crude oil provided by a well, the dye in a hair spray or a soap, the production process of paper, the counterfeiting of banknotes and of drugs, etc.).

Back in 1962, Strickler and Berg [2] published a relation to estimate *a priori* the excited-state lifetime of a fluorescent molecule. Yet, its usefulness is limited because of the variation of lifetimes due to the experimental conditions. That is, the best way to know the lifetime of a fluorophores if to measure it directly.

Figure 68.9 displays the decay time of anthracene in ETOH using the ChronosBH, a TCSPC instrument, by ISS. The light source is a pulsed LED emitting at 335 nm. A high-pass filter (WG 385, 50% transmission at 385 nm) was used to separate the fluorescence. A single lifetime of 4.2 ns was determined using the fitting routine of the software.

Figure 68.10 displays the decay time of anthracene in ETOH using the ChronosFD, a frequency-domain instrument. Phase and modulation data were acquired at 14 different modulation frequencies ranging from 2 MHz to about 250 MHz. The light source is a pulsed LED emitting at 370 nm. A high-pass filter (WG 389, 50% transmission at 385 nm) was used to separate the fluorescence. A single lifetime of 4.2 ns was determined using the fitting routine of the software. In both techniques the decay times are recovered by using a fitting algorithm (least square analysis), the algorithm of the theoretical functions that best minimize the differences with the experimental points. Other approaches are available for the data analysis, such as the maximum entropy method (MEM) and the phasor analysis.

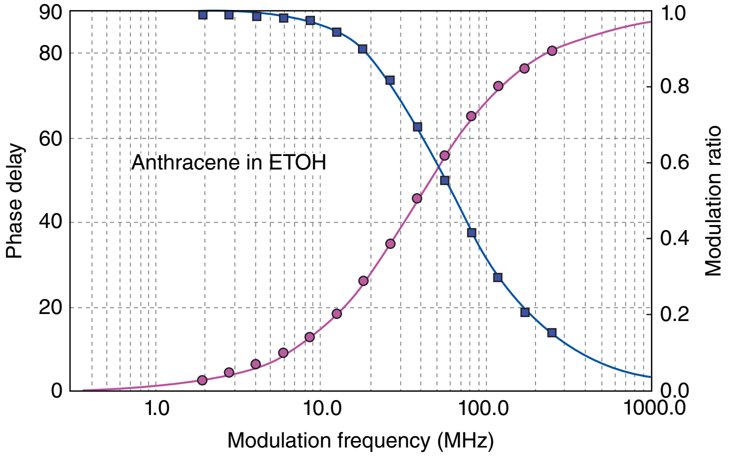


FIGURE 68.10 Decay curve of anthracene in ETOH using a frequency-domain instrument (ChronosFD, by ISS). Source: Reproduced with permission of ISS.

TABLE 68.1 Quantum Yield Values of Selected Molecules

Molecule	Wavelength Range (nm)	Temperature (°C)	Solvent	Quantum Yield
Benzene	270–300	20	Ethanol	0.04
Anthracene	360–480	20	Ethanol	0.27
Tryptophan	300–380	25	H_2O	0.14
Rhodamine 101	600–650	20	Ethanol	1.0

68.5.4 Quantum Yield

The quantum yield is a parameter that varies widely from molecule to molecule. A few examples are reported in Table 68.1. Clearly, when looking for a fluorescent probes, there are advantages in selecting one featuring a high quantum yield!

We refer the reader to the literature listed in Further Reading for the measurement of the quantum yield. We only recollect that there is a direct mode and a relative mode. The direct mode encompasses the use of the integrating sphere, an accessory of the spectrofluorometer that allows for the determination of the number of photons emitted by a sample. The relative mode allows for the determination of the quantum yield of a sample by comparison to a reference of known quantum yield. Both measurements require particular attention to the details.

68.5.5 Anisotropy and Polarization

Anisotropy (or polarization) is a popular application of fluorescence spectroscopy as it allows for the measurement of the rotation of molecules as well as of their shape and size and the rigidity of molecular structures.

A light beam is described as an electromagnetic wave with an electric vector $\vec{E}$ and a magnetic vector $\vec{B}$ perpendicular between them; both are also perpendicular to the

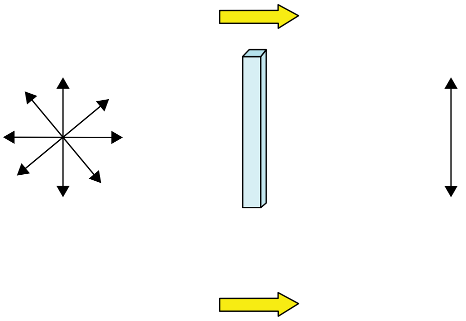


FIGURE 68.11 An unpolarized light beam traverses a polarizer; a plane of polarization is selected.

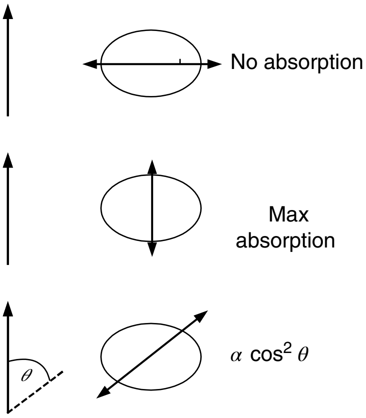


FIGURE 68.12 Molecules with the electric dipole featuring a component parallel to the direction of the electric field of the excitation light have a probability for absorption of a photon.

direction of propagation of the light beam $\vec{k}$. Natural light can be described as the superposition of innumerable single wave representations. When working with natural light a particular direction of the electric vector $\vec{E}$ can be selected by using a polarizer; such wave is said to be "polarized" (Fig. 68.11).

Polarized light can be utilized for interesting experiments and applications. When polarized light with the proper energy illuminates an ensemble of molecules (Fig. 68.12), only molecules with the excited-state dipole moment $\vec{M}_A$ (or transition moment) oriented in the same direction of the electrical field (polarization) can absorb the photons. If the direction of polarization of the excited beam and the direction of the dipole moment of the molecule are perpendicular to each other, no absorption takes place. In intermediate cases, the probability of the absorption is proportional to $\cos^2\theta$, where θ is the angle between the vector $\vec{E}$ of the exciting light and the vector $\vec{M}$ of the transition moment dipole (Fig. 68.12).

Because of the preferential absorption rules of the molecules, a polarized light introduces a *photoselection* of the molecules. As the distribution of the excited fluorophores is anisotropic, the fluorescence is anisotropic too. Any change in the direction of the

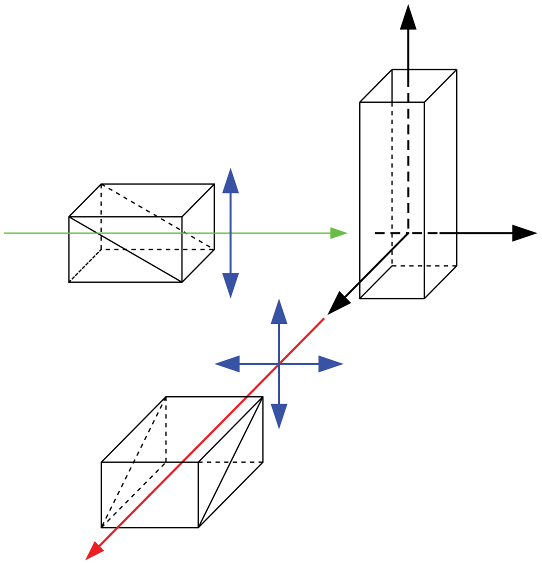


FIGURE 68.13 Experimental setup for anisotropy measurements. The spectrofluorometer has a polarizer in the excitation channel, and a second polarizer in the emission channel. The intensity of the fluorescence reaching the light detector is measured for the different orientation of the polarizers (see Equation 68.11).

transition moment $\vec{M}_A$ during the time the molecule spends in the excited level will result in a decrease of the anisotropy, that is, the overall polarization of the fluorophore solution will decrease. The decrease in the anisotropy can be due to several reasons:

- Difference in direction between the absorption and emission transition moments. This happens as the transition moments of the excited states S_1 and S_2 may not be the same; yet, molecules emit from the lowest vibrational level of S_1.
- Brownian motion. Molecules in the excited state enter into collisions with the molecules of the solvent or with the molecules of the same species, and as a result, the direction of the emission transition moment changes.
- Energy transfer to another molecule featuring a different orientation.

Anisotropy is measured using a spectrofluorometer equipped with polarizers; one polarizer is mounted in the excitation beam (Fig. 68.13), and a second polarizer is inserted in the emission channel. The anisotropy is defined as

$$r = \frac{I_{\mathrm{VV}} - gI_{\mathrm{VH}}}{I_{\mathrm{VV}} + 2gI_{\mathrm{VH}}} \tag{68.11}$$

And the polarization is

$$P = \frac{I_{\mathrm{VV}} - gI_{\mathrm{VH}}}{I_{\mathrm{VV}} + gI_{\mathrm{VH}}} \tag{68.12}$$

The two parameters, anisotropy and polarization, describe the same phenomenon; they are related to each other by

$$P = \frac{3r}{r+2} \tag{68.13}$$

(In the following description we will refer to anisotropy only.) In the relations earlier, I_{VV} is the measured fluorescence intensity with the polarizer in the excitation channel in the (V)ertical position and the polarizer in the emission channel in the (V)ertical position; I_{VH} is the measured fluorescence intensity with the polarizer in the excitation channel in the (V)ertical position and the polarizer in the emission channel in the (H)orizontal position.

The number g, called the *g-factor*, is given by $g = I_{\mathrm{HV}}/I_{\mathrm{HH}}$, where the letters V and H refer to the positions of the polarizers in the excitation and emission channel, respectively. The *g-factor* corrects the anisotropy values for the artifact introduced by the instrument; as is the case for emission spectra, the instrument has different transmission properties for the two planes of polarization.

Figure 68.14 displays the excitation polarization spectrum for erythrosine (top line until 500 nm); the other line represents the excitation spectrum in the range from 300 to 530 nm. The fluorescence is collected at 550 nm. The polarization is negative for wavelengths below 360 nm and then rises sharply up to 400 nm and stays almost constant above 400 nm. The reason for this behavior is due to the fact that the excitation at the short wavelengths favors the transition $S_0 \to S_2$, while at the longer wavelengths the transition $S_0 \to S_1$ is the one excited: as the fluorescence is always emitted by the lowest vibrational level of S_1, it is an indication of the different orientation of the transition moments of the excited levels S_1 and S_2. Practically, when using anisotropy measurements one has to select and specify the excitation wavelength (and choose a wavelength displaying a high value of polarization).

What are the values that the anisotropy can assume? In order to answer this question, one has to introduce the emission transition moment $\vec{M}_{\mathrm{E}}$ and distinguish the two cases:

1. $\vec{M}_{\mathrm{E}}$ and $\vec{M}_{\mathrm{A}}$ are parallel.
2. $\vec{M}_{\mathrm{E}}$ and $\vec{M}_{\mathrm{A}}$ are not parallel.

Without going into the details of the calculations (the interested reader can consult the book by Valeur cited in References), we note that for the case of the two moments being parallel and in absence of any motion, it is $r_0 = 0.4$; this value is called the *fundamental anisotropy*. When the two moments are not parallel, the values are confined in the range

$$-0.2 \le r_0 \le 0.4 \tag{68.14}$$

The case of the decrease of anisotropy due to Brownian motion collisions is a very interesting one for its practical applications. This is the case when molecules in the

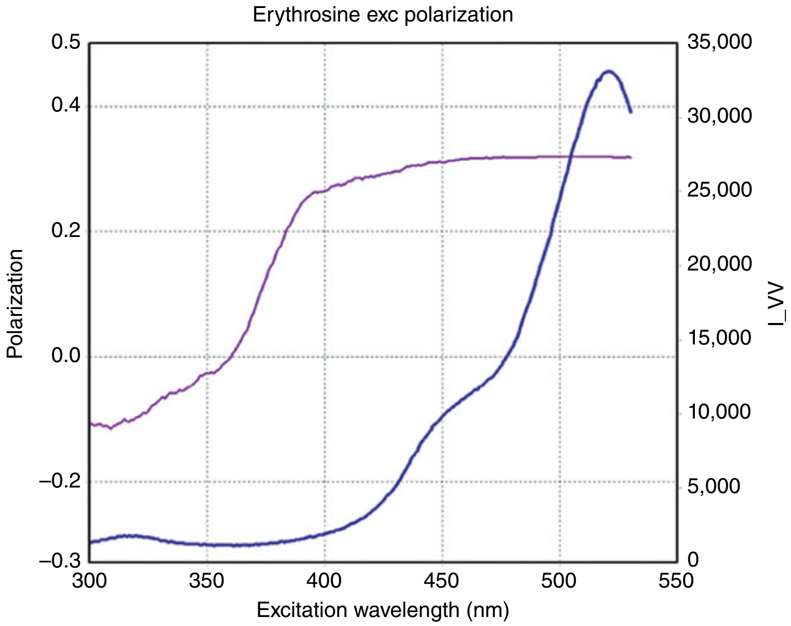


FIGURE 68.14 Excitation polarization spectrum for erythrosine (top line until 500 nm); the other line represents the excitation spectrum in the range from 300 to 530 nm. The fluorescence is collected at 550 nm.

excited state rotate due to collisions with the solvent. The amount of the depolarization depends upon the value of the decay time of the molecule, the size of the molecule, and the viscosity and temperature of the solvent. In fact, let us suppose that the decay time is of the same order of the rotational time; it is found that the anisotropy decays, for a spherical molecule, according to the following relation:

$$r(t) = r_0 \exp(-6D_r t) \tag{68.15}$$

where D_r is the rotational diffusion coefficient. From the Stokes–Einstein relation $D_r = RT/6V\eta$, V is the hydrodynamic volume of the molecule, η is the solvent viscosity, R is the gas constant, and T is the absolute temperature. D_r can be determined by resolving Equation 68.15 using time-resolved fluorescence techniques. Alternatively, if the decay is a single exponential decay, it can be solved using steady-state technique. As

$$\bar{r} = \frac{1}{\tau}\int_0^\infty r(t)\exp\left(\frac{-t}{\tau}\right)dt \tag{68.16}$$

By direct substitution one finds

$$\frac{1}{\bar{r}} = \frac{1}{r_0}(1 + 6D_r\tau) \tag{68.17}$$

TABLE 68.2 Selected Applications of Anisotropy Measurements

Spectroscopy	Separation of Excited States
Polymers	Local viscosity Molecular orientation Chain dynamics
Immunology	Antigen–antibody reactions Immunoassays
Molecular biology	Proteins interactions Nucleic acid–protein interactions Biological membranes Micellar systems

This is the Perrin equation; it allows for the evaluation of the decay times by measurements of the steady-state polarization! In some literature, the quantity $\tau_C = 1/6D_r$, called the rotational correlation time, is used. This case is strictly valid for a spherical molecule. When the more complex shape of a general ellipsoid is considered, the motion is described by three rotational diffusion coefficients associated with each of the rotational axis. The relation between the rotational correlation times and the rotational diffusion coefficients is no longer simple. The anisotropy decay is described by

$$r(t) = \beta_1 e^{-t(4D_1+2D_2)} + \beta_2 e^{-t(D_1+5D_2)} + \beta_3 e^{-t(6D_2)} \tag{68.18}$$

where

$$\begin{aligned} \tau_1 &= \frac{1}{(4D_1 + 2D_2)} \\ \tau_2 &= \frac{1}{(D_1 + 5D_2)} \\ \tau_3 &= \frac{1}{(6D_2)} \end{aligned} \tag{68.19}$$

In this expression the quantities β_1, β_2, β_3 represent expressions for the angles between the absorption and emission dipoles and the axes of the ellipsoid; D_1 and D_2 are the diffusion coefficients around the axis of symmetry and equatorial axes, respectively.

There are physical conditions where a probe is restricted to motion within an angle, for instance, the case of a probe in a membrane. In these cases, the anisotropy does not decay to zero. A hindered rotator is described by the following expression:

$$r(t) = (r_0 - r_\infty)\exp\left(\frac{-t}{\tau_c}\right) + r_\infty \tag{68.20}$$

Table 68.2 lists a few applications of the technique that spans from the physical chemistry research all the way to clinical applications.

68.6 CONCLUSIONS

Fluorescence is a sensitive technique that, although started as an analytical tool, is used more and more for the study of molecular interactions *in vitro* and in cells; in fact, it is nowadays capable of detection of single molecules on a routine basis. The fluorescence decay time of typical fluorophores falls in a window (1–20 ns) suitable for the observation of several molecular processes of biological relevance. The spectral properties of fluorophores are changed by several processes including collisions with other molecules, rotational diffusion, and formation of complexes; moreover, the fluorescence properties are sensitive to changes of the environment such as pH, electrical fields, concentration, temperature, and polarity. These features have expanded the applications of fluorescence to fields as diverse as the development of sensors for monitoring the presence of specific analytes (O_2, ions) *in vitro* and *in situ* to the development of sensors for the measure of physical parameters (materials under high pressure, mechanical properties of materials). A variety of research instruments is available for the measurement of the general and specific parameters of the fluorescence. Dedicated instruments are utilized for the measurements in specific immunoassays (polarimeters), in drug discovery (microwell plates and microarrays), cell sorting (cytofluorometers), and genome sequencing.

REFERENCES

1. R.D. Spencer and G. Weber, 1969. Measurements of subnanosecond fluorescence lifetimes with crosscorrelation phase fluorometer. Ann. N. Y. Acad. Sci. 158, 361–376.
2. J.S. Strickler and R.A. Berg, 1962. Relationship between absorption intensity and fluorescence lifetime of molecules. J. Chem. Phys. 37, 814–822.

FURTHER READING

W. Becker, 2005. Advanced Time-Correlated Single Photon Counting Techniques; Springer-Verlag, Berlin/Heidelberg.

D.M. Jameson, 2014. Introduction to Fluorescence; CRC Press/Taylor & Francis Group, Boca Raton.

J.R. Lakowicz, 2006. Principles of Fluorescence Spectroscopy, 3rd Edition; Springer–Verlag, New York.

B. Valeur, 2005. Molecular Fluorescence; Wiley-VCH Verlag Gmbh, Weindheim.

69

X-RAY ABSORPTION SPECTROSCOPY

Grant Bunker
Department of Physics, Illinois Institute of Technology, Chicago, IL, USA

69.1 INTRODUCTION

X-ray photons are products of natural (e.g., astrophysical) processes that also find many uses as the basis of powerful research tools in science, medicine, and engineering. In addition to the familiar use of X-rays for industrial and medical radiography, X-ray absorption spectroscopy is of great utility for providing information on molecular and electronic structure in chemistry and in condensed matter physics, materials science, biology, geology, and other fields. This chapter describes what is measured, how the measurement is accomplished—sources, X-ray optics, detectors, measurement modes, and samples—and how to minimize errors. The quantum physics of X-ray absorption processes is beyond the scope of this chapter on measurement, as is the detailed analysis of the spectra. For further information on fundamentals, see, for example, [1–3].

69.2 BASIC PHYSICS OF X-RAYS

In this section we briefly describe the basic physics describing the interaction of X-rays with matter. These processes (scattering and absorption) are central to understanding the mode of operation and characteristics of the X-ray optics that are used to manipulate X-ray beams, as well as the interactions of X-rays with materials, and methods of detection of X-rays.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

69.2.1 Units

Although SI units are commonly used in X-ray science, angstrom units (1 Å = 10^{-10} m = 1 nm) also are used, because this unit corresponds well to atomic dimensions, a property that is important for both diffraction and spectroscopic experiments. The wavelength range of X-rays is much shorter than that of visible light, which is approximately 0.4–0.7 micrometer (μm), about 10^4 times longer wavelength than typical X-ray wavelengths.

Electron volt (eV) units (1 eV ≈ 1.60×10^{-19} J) often are used to describe energies, because the spacings between energy levels in atoms are on the order of eV or greater—typically 10^4 times greater in the case of X-rays, in which case keV (10^3 eV) units are often used. 1 eV is the magnitude of energy change when one electron charge is moved through an electric potential difference of 1 V.

69.2.2 X-Ray Photons and Their Properties

X-radiation is the portion of the electromagnetic radiation spectrum with a wavelength that is shorter than ultraviolet and "vacuum ultraviolet" radiation. X-ray wavelengths are on the order of 1 nanometer (nm) or shorter, typically about 1 Å, or 0.1 nm.

In classical physics an electromagnetic wave in free space consists of oscillating electric ($\vec{E}$) and magnetic ($\vec{B}$) fields that are mutually perpendicular and also perpendicular to the direction of motion. The fields oscillate in phase at a frequency f (with angular frequency $\omega = 2\pi f$) and wavelength λ, where the speed of light is $c = f\lambda$, and $|\vec{E}| = |\vec{B}|c$. The wave vector $\vec{k}$ (where $|\vec{k}| = 2\pi/\lambda$) points along the direction of propagation ($\vec{E} \times \vec{B}$). The X-ray polarization vector $\hat{\epsilon}$ is a unit vector in the direction of $\vec{E}$.

When considering X-ray absorption by atoms, it must be recognized that light, including X-rays, comes in photons. It is generally necessary to use a quantum mechanical and special relativistic description to understand the interaction of light with matter, but when a large number of photons are present in a beam, it may be sufficient for some purposes to treat the beam as a classical electromagnetic field. This approximation is commonly made in X-ray diffraction experiments.

The quantum of electromagnetic field is a photon, a massless particle that carries energy $E = hc/\lambda$, where h is Planck's constant and c is the speed of light; photon momentum $p = E/c = h/\lambda$; and the photon's angular momentum along the direction of motion has the quantized values $\pm\hbar$, with $\hbar = h/2\pi$. The product $hc \approx 12398.5$ eV Å, so that a 12.4 keV photon has a wavelength of 1 Å = 0.1 nm and a 6.2 keV photon has a wavelength of 2 Å. In this chapter we shall largely restrict our attention to energies from around 1 to 100 keV. The rest energy of an electron is approximately $mc^2 = 511$ keV.

Physically X-rays are fundamentally no different than gamma rays, which also are high-energy photons; gamma rays generally have their origins in nuclear processes, whereas X-ray photons generally are produced by transitions of electrons between atomic energy levels and by accelerating charges. Both are produced in various energetic astrophysical processes.

69.2.3 X-Ray Scattering and Diffraction

The simplest interaction between a photon and an atom is the scattering of a photon from electrons in an atom. Although they are charged particles, protons do not scatter significantly compared to electrons because their mass is much greater. For this reason X-ray scattering from atoms depends only on the electron distribution and not (directly) on the nuclear charge.

If a photon scatters from a single weakly bound (quasi-free) electron, the momentum change of the photon causes the electron to recoil. The kinetic energy of the electron carries off some energy so that the scattered (outgoing) photon has a lower energy than the incoming photon. The change in wavelength between the scattered and incoming photons is given by

$$\lambda' - \lambda = \frac{h}{mc}\left(1 - \cos\theta\right),$$

where θ is the scattering angle, the angle between the directions of incoming and scattered photon directions. This is called "Compton scattering" and it is a type of inelastic scattering: the scattered photon has a lower energy than the incoming photon. The constant $h/mc = hc/mc^2 \approx 12.4\,\text{keV}\,\text{Å}/511\,\text{keV} \approx 0.0243\,\text{Å}$ is called the Compton wavelength.

The incident photon can also transfer part of its momentum to the whole atom, in which case its recoil is minimal, and the energy change between incoming and outgoing wavelengths in that case is negligible. This is elastic scattering: the incident and scattered photons have the same energy. Both elastic and inelastic scattering depend on the number of electrons in an atom and their spatial distribution within it. Multiple (say, N) electrons may coherently participate in the scattering process, generating a scattered wave of amplitude N times as strong as that from a single electron. The intensity of the wave varies as the square of the amplitude, giving N^2 times the beam intensity as that from a single electron. The number of electrons that coherently scatter depends on the wavelength and spatial distribution of electrons in the sample.

If a photon (or beam of photons) scatters from a material with spatial periodicities, such as a crystal, the coherent scattering will be greatly enhanced at certain angles of incidence: this is called X-ray diffraction. It is an X-ray analog of using diffraction grating with visible light. The atoms within a crystal are arranged in planes with various separations d_{hkl}, where the integer ("Miller") indices h, k, l serve to identify which sets of planes one is referring to. The geometric condition for "Bragg diffraction" is $n\lambda = 2d_{hkl}\sin\theta_{\text{B}}$, where λ is the X-ray wavelength and θ_{B} is the angle ("Bragg angle") between the incoming wave direction and the diffracting planes and n is a positive integer. The geometry of Bragg diffraction is shown in Figure 69.1.

Bragg diffraction from crystals is an important means of studying structures of materials such as semiconductors, alloys, minerals, and biological molecules such as proteins, lipids, and nucleic acids; it also provides a means to select specific wavelengths (energies) for X-ray absorption experiments using X-ray monochromators. X-ray

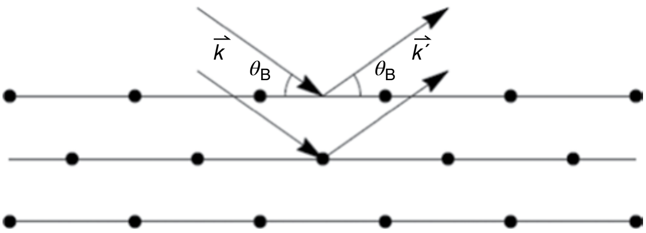


FIGURE 69.1 Bragg diffraction from atomic planes of crystal.

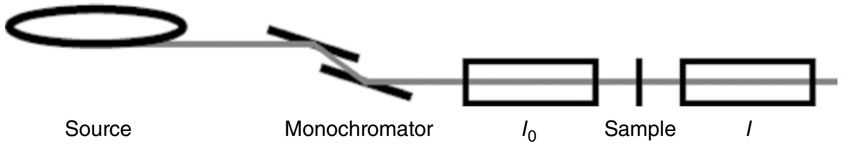


FIGURE 69.2 Schematic transmission mode X-ray absorption measuring apparatus (not to scale).

absorption fine structure (XAFS) is a distinct and complementary technique because it uses photoelectron (not X-ray) interference to probe the sample, and it does not require translational order (e.g., crystalline structure) to determine molecular structure.

69.2.4 X-Ray Absorption

Light can be absorbed by a material when the energy of the photons $E = h\nu = \hbar\omega$ is sufficient to create excitations in it. Microwaves excite transitions between rotational states of molecules; infrared light excites transitions between vibrational states of molecules. Visible, ultraviolet, and X-ray photons excite transitions between electronic states of molecules and atoms. The energies at which specific transitions of materials can be excited are characteristic of the structure and bonding of the atoms, molecules, or crystalline domains of which it's composed.

An X-ray photon may be absorbed when it has an energy that corresponds to a difference between two electronic quantum states in an atom or molecule. For X-rays the transitions are between a deeply bound core electronic state in an atom (such as the lowest energy level, the 1*S*) and an empty final state of suitable symmetry (e.g., a *P* state). The energies of these transitions are characteristic of the atoms in the material.

A typical setup for measuring X-ray absorption spectra is shown in Figure 69.2. The X-ray source (e.g., synchrotron radiation source) generally produces a range of energies from which a specific narrow band of energy is selected by a monochromator, typically one employing Bragg diffraction from set of crystals. Following the monochromator the beam flux (photons/s) is measured with a partially transparent detector I_0 (labeled by the current output signal it produces), and the beam is then transmitted through the sample, and the transmitted flux I is then measured. Often focusing optics, shutters, and beam defining slits are placed between the monochromator and the I_0 detector.

X-rays passing through a homogeneous uniform sample are attenuated exponentially, so that $I/I_0 = G(E)\exp(-\mu(E)x)$, where $\mu(E)$ is the X-ray linear attenuation coefficient of the material, E is the X-ray beam (photon) energy, x is the sample thickness (path length through the sample), and $G(E)$ accounts for detector sensitivities and absorption by air paths and detector windows. This equation is equivalent to $\mu(E)x = \ln(I_0/I) + \ln G(E)$; the second term ($\ln G(E)$) ordinarily gives a slowly varying additive background that can be readily subtracted out numerically or better measured independently by removing the sample from the beam path.

The quantity $\mu(E)$ measures the attenuation of X-rays as they pass through the sample because of absorption and scattering. In most cases of practical interest absorption is the dominant process, but scattering is not generally negligible.

69.2.5 Cross Sections and Absorption Edges

The ability of an atom to absorb or scatter is characterized by atomic *cross sections*, which are energy-dependent quantities that have dimensions of area. A beam of photons of intensity N[photons/s/m^2] hitting a thin target of total absorption cross section σ_{tot}[m^2] absorbs photons at a rate $N \times \sigma_{\mathrm{tot}}$[photons/s] where σ_{tot} is the sum of the absorption cross sections of the atoms that are illuminated by the beam.

The *barn* (10^{-28} m^2) is a non-SI unit to describe cross sections, commonly used in nuclear physics. It is also common to specify cross sections per mass or per atom.

If the cross sections σ for a pure element are specified on a per-mass basis, we can write $\mu(E) = \rho\sigma(E)$ where ρ is the mass density of the material in the sample. For a sample consisting of a mixture or compound of different elements, we have approximately $\mu(E) = \rho\sum_i f_i\sigma_i(E)$ and f_i and σ_i are, respectively, the mass fraction m_i/M_{total} and cross section of the ith atomic species in the sample. Approximate atomic cross sections are tabulated and readily available online. The ones shown here are from Chantler et al. [4].

A log-log plot of the specific cross section for gold is shown in Figure 69.3. It consists essentially of step increases in absorption at certain energies, with straight lines between them. The jumps are called *absorption edges*; they correspond to specific atomic excitations. The highest energy one at about 80 keV comes from exciting electrons out of the $n = 1$ atomic shell (K-edge); the next lower ones in energy between about 12 and 14 keV are due to excitations from the $n = 2$ atomic levels (L-edges LI, LII, LIII); the next lower ones in energy are the *M-edges*. Between the edges the curves are approximately straight lines on a log-log plot, which implies a power law dependence. A good rule of thumb is that, between the absorption edges, absorption varies as $1/E^3$, so the cross section (and absorption coefficient μ) decreases by a factor of 8 at twice the energy (if there are no absorption edges in that region). It will be noticed (Figs. 69.3, 69.4, and 69.5) that for heavy elements (gold, copper) the scattering cross sections are much less than absorption cross sections except at high energies; for lighter elements such as carbon, absorption and scattering are equal at about 20 keV.

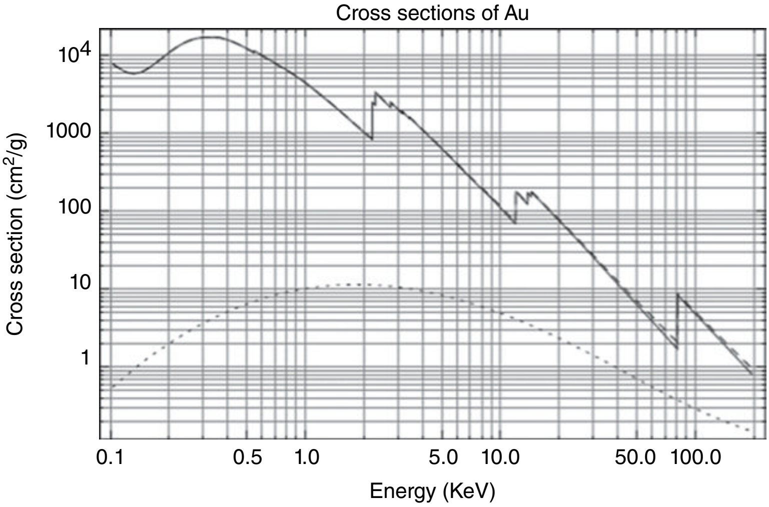


FIGURE 69.3 Log-log plot of the (semiempirical) X-ray absorption cross section of gold (Z=79) versus X-ray energy. The K, L_1, L_2, L_3, and M-edges are shown; fine structure is *not* shown. The solid line is the photoelectric absorption, and the dotted line is total elastic + inelastic scattering cross section; the dashed line is the sum of absorption and scattering.

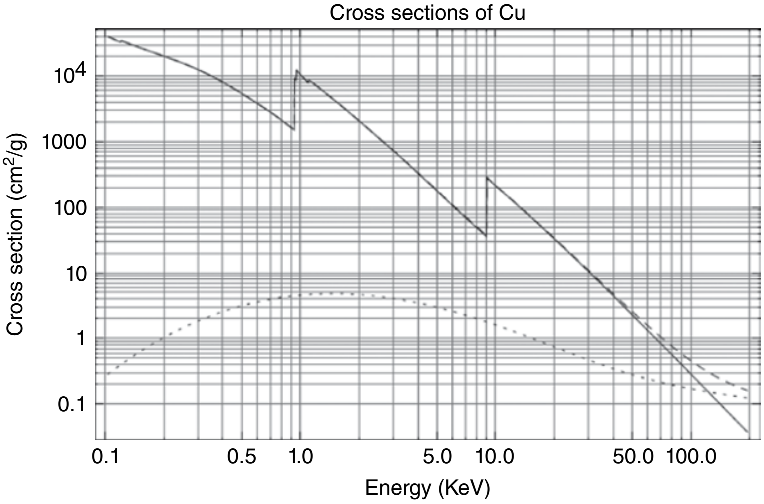


FIGURE 69.4 Log-log plot of the (semiempirical) X-ray absorption cross section of copper (Z=29) versus X-ray energy. The K, L_1, L_2, L_3, and M-edges are visible; fine structure is *not* shown. The solid line is the photoelectric absorption, and the dotted line is total elastic + inelastic scattering cross section; the dashed line is the sum of absorption and scattering.

The energies at which the absorption edges occur are characteristic of the atomic number of the element. For example, the calcium (Z=20) K-absorption edge is at 4.04 keV, iron (Z=26) is at 7.11 keV, zinc (Z=30) is at 9.66 keV, and molybdenum (Z=42) is at 20.0 keV. Empirically the K-edge energy depends on atomic number Z as

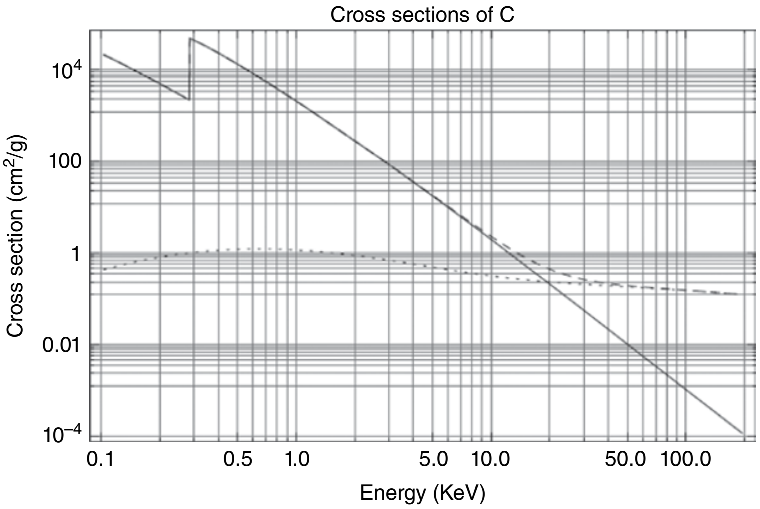


FIGURE 69.5 Log-log plot of the (semiempirical) X-ray absorption cross section of carbon ($Z=6$) versus X-ray energy. The K-edge is visible; fine structure is *not* shown. The solid line is the photoelectric absorption, and the dotted line is total elastic + inelastic scattering cross section; the dashed line is the sum of absorption and scattering.

$E_K \propto Z^{2.16}$. Similarly the L-edge energies are fairly smooth functions of the atomic number. Historically the simple dependence of core-level energies with atomic number was important in completing the periodic table. This predictability and the substantial separation between edges allow experimenters to easily "tune into" particular atomic species for study. Good sources of data on absorption edges, fluorescence energies, and absorption cross sections are the X-ray data booklet [5], the servers at NIST [4], and the online server at IIT [6].

The X-ray absorption cross section curves shown in Figures 69.3, 69.4, and 69.5 are composite averaged representations of various experiments and theoretical computations. They do not represent the XAFS, which consists of peaks and periodic oscillations that are observed in experimental absorption edge spectra. Once measured, experimental XAFS spectra can be analyzed to provide information on average distances to atoms around the element of interest (whose absorption edge it is). This is beyond the scope of this chapter but it is described in [3].

69.3 EXPERIMENTAL REQUIREMENTS

The accurate measurement of XAFS spectra places stringent requirements on the apparatus, because one must not only measure absorption spectra as a function of incident X-ray photon energy to a precision of 10^{-3} to 10^{-4} of the size of the edge step (the difference in absorption above and below the edge), but the incident photon beam

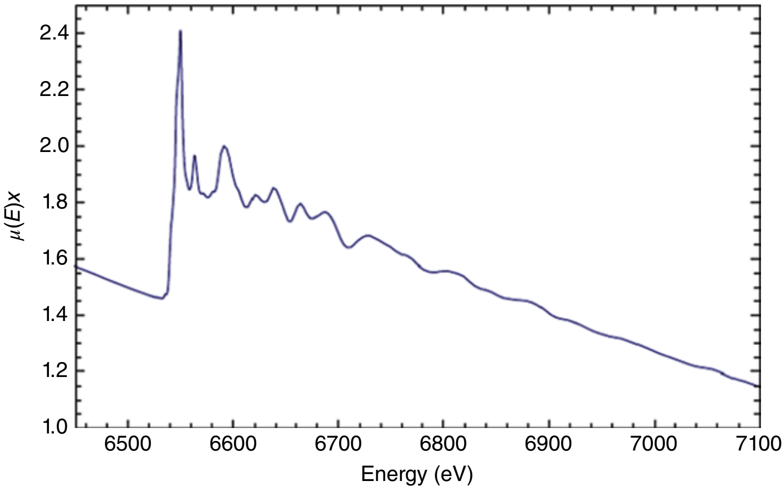


FIGURE 69.6 Plot of experimental transmission mode $\mu(E)x$ data for manganese oxide (MnO) measured at 80 K. The absorption edge occurs at approximately 6540 eV.

energy must be smoothly varied over an extended energy range of about 1 keV above the edge. Preserving positional stability of the beam and shape of the beam is also important. Furthermore certain systematic errors must be eliminated, as described below; otherwise conclusions regarding the structure will be incorrect. For this reason here we will concentrate on XAFS spectra.

XAFS spectra consist of absorption peaks within approximately 20 eV of the absorption edge, some on the order of 1 eV in width, plus much slower oscillations extending about 1 keV above the edge. The width of the near-edge peaks depends on the structure and also the core hole lifetime of the edge in question, because the finite lifetime of the excited state after the X-ray absorption event introduces an uncertainty principle-related energy broadening, which smooths the spectrum.

An example of an XAFS spectrum can be seen in Figure 69.6, which plots $\mu(E)x$ for a thin sample of manganese oxide (MnO). The oscillations become less rapid (in energy space) and smaller in amplitude as energy above the edge increases. Depending on the core hole level width, to adequately measure the structure near the edge, it is necessary to sample at approximately 1 eV or better energy spacing; above the edge the sampling can be several times larger than that. There are some advantages to sampling uniformly in k-space at intervals of 0.05–0.07 Å^{-1}, where $k^2 = 0.2625\ (E - E_0)$, E is the photon energy, and E_0 is the edge energy. Above the edge the key criterion is to sample the spectra at least twice per cycle of the highest-frequency oscillation. Oversampling is helpful to prevent aliasing of noise into the signal passband. For a more detailed discussion and rationale, see the section on “Instrument Control and Scanning Modes” and Bunker [3].

69.4 MEASUREMENT MODES

As described above, the basic mode for measuring XAFS is transmission mode, in which the fraction of X-rays that are absorbed by a sample is measured as a function of photon energy. Other modes are of considerable use however. For dilute samples, in which the element of interest is a small fraction of the total, it is often helpful to instead measure the X-ray fluorescence that is given off as a consequence of the initial absorption event. For example, exciting the K-edge of iron (at about 7112 eV) creates a vacancy ("core hole") in the $1S$ level, which is unstable; an electron from a higher level, most often $2P$, can fall into the core hole and give up a fluorescence photon in the process, with an energy equal to the difference between the $2P$ and $1S$ level, in this case about 6400 eV. This fluorescence photon can be detected and used as a proxy for the measuring the absorption directly. The advantage of doing this is that one only gets those photons if the absorption occurred in the first place, so it increases the sensitivity of the measurement.

Another method of detection is conversion electron detection or electron yield. In the case described above, fluorescence is not the only way a core hole can deexcite: it can also relax nonradiatively, by ejecting electrons, which propagate through the material and either escape the sample or create secondary excitations. Electrons within ≈1000 Å of the surface can escape the sample. If the sample is surrounded by a gas, for example, He, it will ionize the gas and create a number of secondary electrons, which can be collected and amplified as in an ionization chamber. This electron yield detection confers surface sensitivity that can be very helpful experimentally. The sample must be slightly conductive and the beam intensities relatively low for this to work reliably.

Another indirect method of detection (which is very infrequently used) is optical detection of XAFS: measuring low levels of light that are produced as a consequence of the absorption event.

It is practical to measure XAFS-like data by scattering electrons or high-energy photons off a sample and analyzing the energy loss spectrum of these scattered particles. These techniques (which are beyond the scope of this chapter) are, respectively, called electron energy loss spectroscopy (EELS) and inelastic X-ray scattering (IXS).

69.5 SOURCES

69.5.1 Laboratory Sources

XAFS makes particularly stringent demands on the measuring apparatus, but it is still possible to use laboratory sources for measuring X-ray absorption spectra; this was done for many years before synchrotron radiation sources became available. Conventional X-ray tubes work by allowing electrons to accelerate under the influence of a high DC voltage and collide with a metallic anode. The decelerating electrons radiate a broad bremsstrahlung ("braking radiation") X-ray spectrum, and they also excite the atoms of the anode, which give off fluorescence X-rays that are characteristic of the anode material. For most experiments the narrow bandwidth fluorescence is used, but it is also possible to use the broad spectrum and pass it through a monochromator to select specific energies. This is not particularly effective for measuring XAFS because of the low throughput, but it can be, and has been, done; see, for example, [2].

Another type of lab source is a laser-induced plasma X-ray source, which can be used for stroboscopic experiments. For extremely rapid time-resolved experiments, free electron lasers are coming into use [7].

69.5.2 Synchrotron Radiation Sources

The most common X-ray sources for the measurement of X-ray absorption spectra are synchrotron radiation sources [8]. Free electron laser sources also have recently come online to facilitate experiments that require ultrafast measurements or extreme brightness. These large multiuser facilities are the technological by-product of high-energy physics facilities like Fermilab and the Large Hadron Collider, with an important difference that they accelerate electrons (rather than protons) to speeds extremely close to the speed of light with a booster synchrotron and then store them for extended periods within a ring-shaped pipe containing ultrahigh vacuum in order to minimize scattering of the electrons by residual air molecules. At specific locations around the ring strong magnets ("bend magnets") are placed so as to bend the trajectory of the electrons, so they can circulate around the storage ring continuously for hours or days. The total current in the ring slowly degrades because of scattering from residual gas molecules in the evacuated ring, and other processes, so the current decreases exponentially, after which it is dumped and electrons are reinjected to restore the current. In some cases it is possible to operate in "top-off mode" in which the current is kept approximately constant by adding small amounts of charge to the bunches while the ring undergoes normal operations. This has significant benefits in keeping heat load on optics and ring components constant.

The electrons are not distributed uniformly around the ring; rather they are concentrated in "bunches," the maximum number of which depends on the storage ring lattice design. The ring can be populated with a single bunch, or many bunches, depending on experimental needs. The speed of light divided by the circumference of the ring gives the single bunch repetition frequency. For example, the repetition rate for a 600 m circumference ring would be 500 kHz, and if 20 equally space bunches were populating the ring, the bunch repetition rate would be 10 MHz. Typical light pulse lengths are ≈100 ps long.

At such high speeds (close to the speed of light c) Newtonian mechanics is inadequate to describe the physics, and Einstein's special theory of relativity is required to understand it. Synchrotron radiation is inherently a relativistic effect. The total energy of E of each electron (rest energy plus kinetic energy) amounts to billions of electron volts. For example, the electron energy is 7–8 GeV for the APS, SPRing8, and ESRF and around 1–3 GeV for most other sources. The rest energy of an electron $mc^2 \approx 0.511$ MeV; the relativistic parameter $\gamma = E/(mc^2)$ is typically of order 10^3 to 10^4.

A principle of electrodynamics is that accelerating charged particles radiate electromagnetic waves; those moving at constant velocity do not radiate. Although the force produced by an electron moving in the field of the bend magnet does not change the electron's speed, it does change the direction of the electron's velocity, and so the electrons radiate. It is this radiation that produces a broad spectrum extending into X-ray energies. At nonrelativistic speeds the radiation pattern and spectrum would resemble that of a radio-frequency antenna. However relativistic effects radically transform the radiation spectrum and angular radiation pattern, as a consequence of the relativistic length contraction and time dilatation, shifting the spectrum up into the X-ray region, and causing the radiation pattern to be limited to angles of order $1/\gamma \approx 10^{-4}$ (radians) within the orbital plane. This concentrates the X-rays in angle so that a fan of radiation that is highly collimated in the vertical direction (i.e., localized in the plane of the ring) is emitted from the bend magnets. This is ideal for use by silicon monochromators based on Bragg diffraction.

The total radiated power is given by Lienard's generalization of the Larmor formula

$$P = \frac{\mu_0 q^2 \gamma^6}{6\pi c}\left(a^2 - \left|\frac{\vec{v}\times\vec{a}}{c}\right|^2\right),$$

where q is the electronic charge, $\vec{a}$ and $\vec{v}$ are the acceleration and velocity of the electron, c is the speed of light, and μ_0 is the magnetic permeability of free space. Note the very strong dependence of the radiated power on γ.

69.5.3 Bend Magnet Radiation

The spectrum of light that is radiated from an electron in circular motion can be expressed in a general manner through a function $g_1(x) = x\int_{t=x}^{\infty} K_{5/3}(t)dt$, where K_n is the modified Bessel function of order n; $g_1(x)$ is plotted in Figure 69.7. For many purposes a useful approximation is $g_1(x) \approx 1.72x^{0.282}e^{-0.969}$. This function allows us to calculate the spectral flux that can be expected from a bend magnet source (or wiggler—see below) for a given beam current and geometry. The number of photons/second/milliampere/milliradian of horizontal angular acceptance and bandwidth $\Delta\epsilon/\epsilon = 10^{-4}$ is $1.256 \times 10^6\ \gamma\ g_1(x)$, where $x = \epsilon/\epsilon_c$, ϵ is the photon energy, and ϵ_c is a parameter called the critical energy; it depends on electron beam energy E and magnetic field B of the bend magnet (or alternatively the bend radius ρ). In practical units $\epsilon_c \approx 0.665E^2B$, $\rho \approx 3.3E/B$, and $\gamma \approx 1957E$ with ϵ_c in keV, E in GeV, B in tesla, and ρ in meters.

69.5.4 Insertion Devices: Wigglers and Undulators

Bend magnets are required to get the electrons to circulate around the ring, and they make fine sources for many purposes. After a time accelerator physicists realized that there are strong benefits to inserting specifically designed magnetic structures into the beam path in the straight sections between the bend magnets. These insertion devices

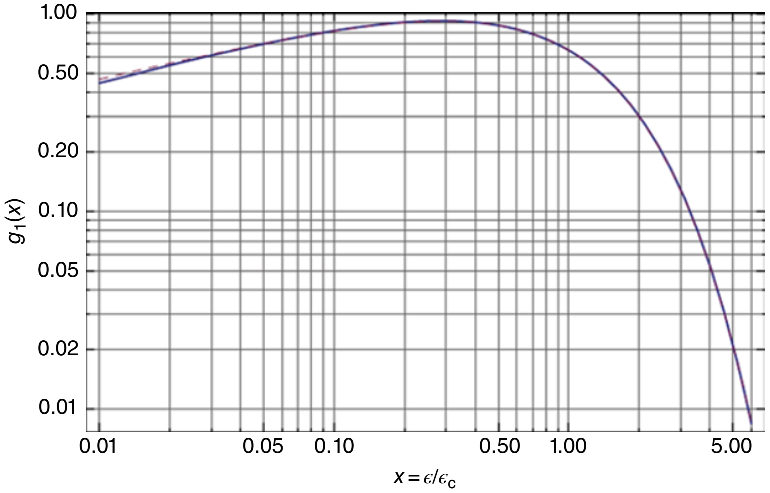


FIGURE 69.7 Energy spectrum $g_1(x)$ for bend magnets and wigglers. The solid curve is the exact function and the dashed curve is the approximation $g_1(x) \approx 1.72x^{0.282}e^{-0.969}$.

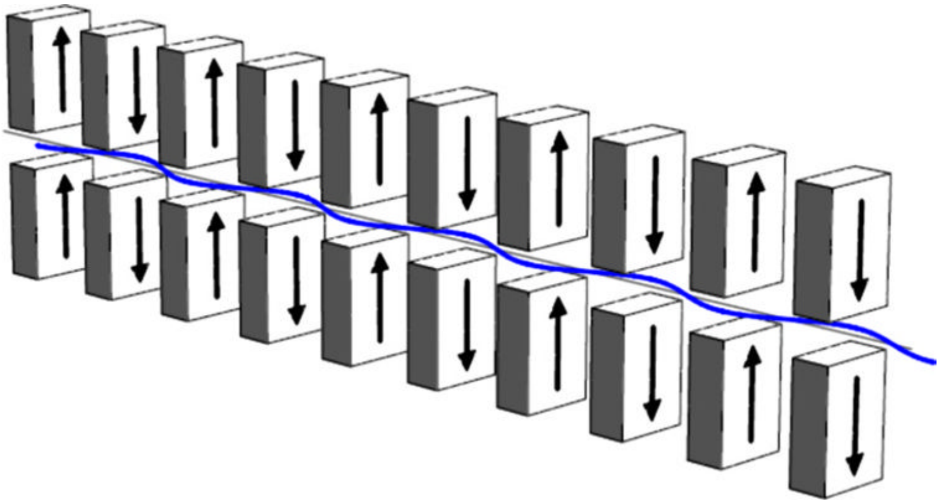


FIGURE 69.8 Schematic of an insertion device (wiggler or undulator). These devices use permanent magnets or electromagnets, either conventional or superconducting. The alternating vertical magnetic field (indicated by arrows) causes the path of the electron to undulate.

consist of periodic arrays of typically $N = 50$–100 magnetic poles (of spatial period λ_0, which is typically a few centimeter) as shown in Figure 69.8. If the angular deflection of the path is large compared to the angular width of the electron radiation pattern ($\approx 1/\gamma$), the X-rays emitted at each pole add up incoherently with those emitted at the other poles, so the spectrum is described by the $g_1(x)$ function with appropriate critical energy and multiplied by the number of poles. This is a wiggler.

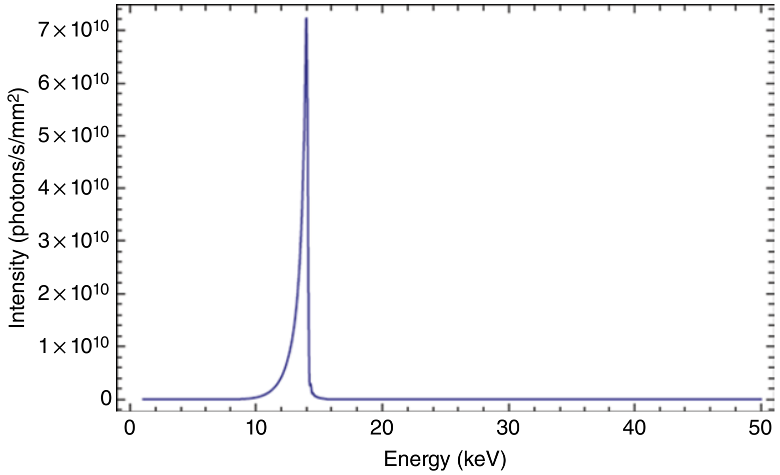


FIGURE 69.9 Computed APS type A undulator spectrum, $K=0.01$. The undulator period is 3.3 cm and the electron beam energy is 7 GeV. In this case a single peak is produced.

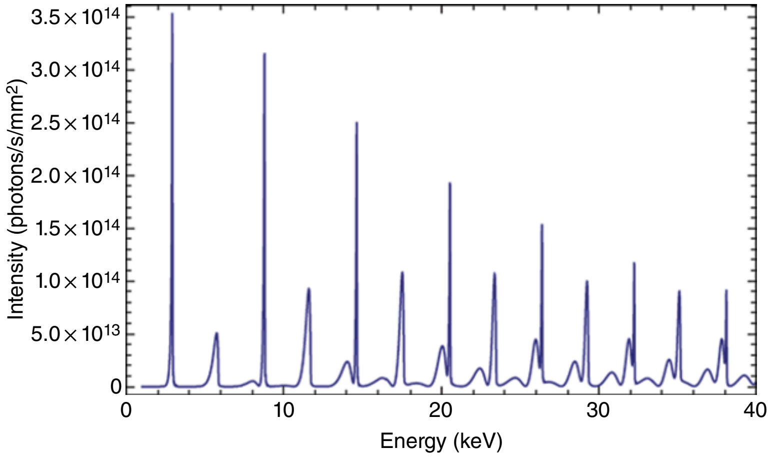


FIGURE 69.10 Computed APS type A undulator spectrum, $K=2.76$. The undulator period is 3.3 cm and the electron beam energy is 7 GeV. In this case many peaks are produced. The odd order harmonics are much stronger than the even order ones.

On the other hand, if the angular deflection $\delta_{\mathrm{w}}=\lambda_0/(2\pi\rho)$ in the insertion device (where ρ is the bend radius of the trajectory) is less than the intrinsic divergence $1/\gamma$, the light emitted at each pole interferes with light emitted at the other poles. The interference effects cause the spectral peaks to be concentrated at specific energies, as shown in Figures 69.9 and 69.10, and the radiation pattern to be collimated horizontally and vertically along the axis of the insertion device within an angle $\Delta\theta \approx 1/(\gamma\sqrt{N})$. This is an undulator. The deflection is quantified using the "deflection parameter" $K=\gamma\delta_{\mathrm{w}}$. The energy spectrum depends on observation angle because of the relativistic Doppler shift, with lower energies detected off-axis. The fractional width of the peaks in energy $\Delta\epsilon/\epsilon$ is inversely proportional to the number of poles N. The wavelength of the peak radiation is $\lambda = \lambda_0/(2\gamma^2)(1 + K^2/2 + \gamma^2\theta^2)$, where θ is the off-axis observation angle; the photon energy is then $\epsilon = hc/\lambda$. By adjusting the vertical separation between the magnets, the magnetic field strength can be systematically varied, which changes the deflection parameter, and shifts the energy of the peaks, which is used to maximize the flux that is passed subsequent X-ray optics.

69.6 BEAMLINES

Once generated, the X-rays that are produced by bend magnets or insertion devices are allowed to pass through a thin beryllium window (beryllium has very low atomic number, so it does not attenuate the beam much) and pass from the machine vacuum down into the beamline. Beamlines are complex instruments, usually tens of meters long, which through use of shielded enclosures safely convey the intense polychromatic (X-ray + UV–Vis + infrared) beam to the various X-ray optical systems, experimental enclosures, detectors, and data acquisition systems used for the measurements. The optics normally reside in shielded high-vacuum or ultrahigh vacuum environments to reduce air scattering and reduce damage by preventing ionization of gases. Cooled apertures or slits are used to define the central portion of the beam and absorb the undesired low-energy off-axis light. Many beamline components are now commercially available.

69.6.1 Instrument Control and Scanning Modes

Computer control is needed to orchestrate the various motions of the monochromator and other optics (shutters, slits, mirrors), detector readout, and sample position, laser firing, and other experimental variables. The incident flux cannot be assumed to be constant to 10^{-3} so it is essential to measure the detector signals precisely over identical fixed time intervals, so that the fluctuations in the incident intensity divide out properly. Typically motion control and data acquisition is done using dedicated VME- or VXI-based (or CAMAC-based) scalers, or GNU-/Linux-based computers, with fast real-time instrument control software such as that available in EPICS [9], Tango [10], MX [11], and others.

In step-scan mode the monochromator and other optics positions are moved (under stepper or DC servo motor control) to obtain a harmonic-free beam of the desired energy; a brief waiting period of a few hundred milliseconds is inserted to allow the vibrations of the optics to damp out; the detectors are then read out for a fixed time, or a fixed count; and the counts are recorded in a computer file. A different energy is then selected and the process is repeated to build up the sampled spectrum. In this step-scan mode, the energies are selected to provide adequate sampling of the spectrum. A pre-edge region of several hundred eV is normally measured with a coarse ($\approx$10 eV) grid to measure the background trend. Over the absorption edge, where the spectrum may be changing rapidly, energy intervals on the order of 1 eV or less are commonly used.

Above the edge, the oscillations are approximately periodic in $k = \sqrt{0.263(E - E_0)}$ (in eVÅ units), where E is the photon energy and E_0 is the absorption edge energy. By the Nyquist theorem, to obtain adequate sampling of the oscillations, a minimum sampling rate in k-space is needed; $\delta k \approx 0.05\,\text{Å}^{-1}$ is sufficient, which implies a maximum spacing between energy points of $\delta E \approx 0.2\sqrt{E - E_0}$ at energy E above the edge. Although a uniform k-space grid above the edge is optimal, dividing the scan into ranges of different (but adequately sampled) energy subgrids is satisfactory. This is described further in [3].

An alternative approach is to use continuous scanning, by slewing the monochromator in a continuous motion, and averaging the acquired data points over known time intervals. This averages the spectrum over nearby energy values; corresponding averages must be made of the energy values themselves. This approach has the advantage of offering potentially shorter scan times, which can be used among other things to record transient phenomena. Since it is not necessary to wait for motors to move, and settling times to elapse, the duty cycle is generally better in continuous-scan mode than in step-scan mode. The time intervals for each data point must be carefully matched to the slew rate so as to get adequate spectral sampling.

"Dispersive XAFS" is done by dispersing photons of different energies at different angles using a bent crystal. The angle of incidence onto the crystal varies with position, so different wavelengths are diffracted from different parts of the bent crystal. The diffracted X-rays are arranged to pass through the sample and are detected with a spatially sensitive detector such as a photodiode array, area detector such as a CCD or pixel array detector, or (many decades ago) film. Since each small area of the bent crystal diffracts only a small range of wavelengths, this approach does not increase the overall intensity, but it does make it unnecessary to use any mechanical motions to do a scan, so it can be useful for fast phenomena.

Time-resolved experiments have become more common in recent years. These are usually pump-probe experiments in which a perturbation (e.g., laser pulse) is made to the sample, and at a fixed delay time relative to it, a fast measurement is made at fixed energy, typically with avalanche photodiode (APD) detectors. By altering the delay time, a record of the signal versus time can be built up. By altering the energy, an energy scan can be built up. The signal from a single pulse (emitted by a single electron bunch in the ring) normally is insufficient for adequate signal-to-noise (S/N) ratio, so signal averaging over many repetitions is required.

69.6.2 Double-Crystal Monochromators

A double-crystal monochromator (Fig. 69.11) selects a specific (but variable) energy band from the polychromatic beam for use in the experiment. Normally the monochromator uses Bragg diffraction from a pair of crystals (or diffraction gratings, for soft X-ray experiments). The first crystal defines the energy by setting the angle of incidence onto the crystal planes; the second crystal, parallel to the first, essentially acts like a mirror to direct the monochromatic diffracted beam parallel to its original

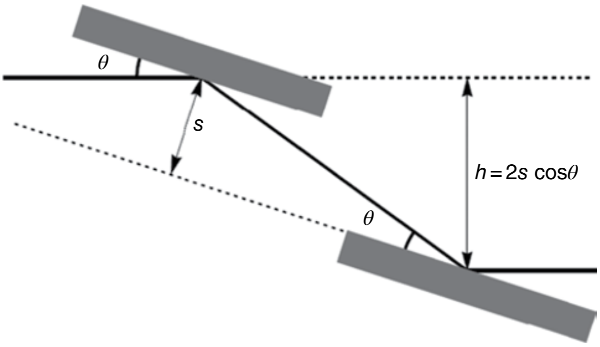


FIGURE 69.11 Schematic double-crystal monochromator. The rectangles represent crystals (typically silicon) and the solid lines show the path of the X-ray beam. The beam is displaced by a distance *h*, which depends on *θ* unless *s* is varied so as to compensate.

direction, down the beamline. Because the incidence angle and exit angle are equal, just as light reflects from a mirror, conventionally (but incorrectly) the diffracted beams are often referred to as "reflections."

High-quality silicon crystals of large size are readily available from the semiconductor industry, and they have suitable properties for many purposes at hard X-ray energies. Germanium crystals are useful at high energies, and diamond crystals are of use because of their very high thermal conductivity. Silicon, germanium, and diamond all have the same crystallographic structure. Diffraction does not occur significantly for all values of *hkl* in the Bragg diffraction equation $n\lambda = 2d_{hkl}\sin\theta_B$, but only when *hkl* are all odd, or when *hkl* are all even and their sum is an integer multiple of 4: for example, *hkl* = 111, 220, 311, 331, 400 are all "allowed reflections." Other crystal types will have different selection rules.

Synthetic multilayers are artificial structures consisting of alternating layers of high and low atomic number materials that act as X-ray interference coatings. Diffraction from these structures can be used for monochromators, but they do not provide sufficient energy resolution to define the beam energy in XAFS experiments. They can be used for collecting fluorescence however [12].

At a third-generation synchrotron source, a large amount of power (hundreds to thousands of watts) is typically deposited in the first crystal, and thermal management is a key aspect of the design. The angular tolerances for alignment of the monochromator crystals are quite demanding, typically on the order of seconds of arc, that is, tens of microradians. The second crystal may be sagittally bent to provide horizontal focusing.

69.6.3 Focusing Conditions

X-ray beams can be focused using reflective optics, that is, X-ray mirrors; diffractive optics such as bent crystals, asymmetrically cut crystals, and Fresnel zone plates; and refractive optics such as beryllium lenses. In comparison to properties at visible wavelengths, at X-ray wavelengths the performance of these devices is strongly constrained because of the index of refraction of materials at X-ray energies, as described below.

Over a small region of a mirror, it will be locally flat, and the specularly reflected beam will have equal angles of incidence and reflection. The incident and reflected beams define a plane of reflection, and the mirror can be bent along an axis that is either perpendicular to that plane (meridional focusing) or parallel to the plane (sagittal focusing). For example, a horizontal mirror can be meridionally bent to provide vertical focusing or sagittally bent to provide horizontal focusing. The focusing equation for the sagittal case is $2\sin(\theta)/R_s = (1/u + 1/v)$; the focusing equation for the meridional case is $2/(R_m \sin(\theta)) = (1/u + 1/v)$. Evidently $R_s = R_m \sin^2(\theta)$, so $R_m \gg R_s$ for small angles of incidence. Here R_s and R_m are the radii of curvature for the sagittal and meridional cases, θ is the local angle of incidence onto the optic (mirror or bent crystal), and u and v are the source to optic distance and the optic to focus distance, respectively. By using a large source to optic distance u and small optic to focus distance v, the apparent size of the X-ray source (e.g., undulator) can be demagnified by the v/u, significantly reducing the size of the beam, at the cost of increased angular spread.

69.6.4 X-Ray Lenses and Mirrors

In the visible energy range, a light beam propagating in a medium with high index of refraction that is incident on an interface with a material of lower index will undergo total internal reflection, provided the angle of incidence is shallow enough. In the X-ray region the index of refraction is a complex number with a magnitude very slightly less than 1.0. In this case X-ray beams will undergo total external reflection at sufficiently shallow angles so that all materials act as X-ray mirror at shallow angles on the order of milliradians. This effect is used to make X-ray mirrors. Synchrotron X-ray mirrors are typically polished to a mean square roughness on the order of several Ångström.

The complex index of refraction $\tilde{n}$ of materials in the X-ray regime can be written as $\tilde{n} = 1 - \delta - i\beta$, where $\delta = ne^2\lambda^2/2\pi mc^2$, $\beta = \mu\lambda/4\pi$, and μ is the X-ray absorption coefficient; λ is the X-ray wavelength, n is the number density of mobile electrons in the material, $e^2 \equiv q^2/4\pi\epsilon_0$, and q and m are the charge and mass of the electron. In pure elements this can be expressed as $\delta = N(Z/A)e^2\lambda^2/2\pi mc^2$ where ρ is the mass density, Z and A are, respectively, the atomic number and the atomic weight of the element, and N is Avogadro's number. The real and imaginary parts of $\tilde{n} = 1$, respectively, describe dispersion and absorption of the X-rays.

The closeness of $\tilde{n}$ to 1.0 shows that the ability of materials to refract is quite limited. In recent years X-ray lenses have come into more common use at synchrotron radiation sources, particularly at high energies. In the X-ray energy region, a lens typically consists of a series of lens-shaped voids (or a more easily fabricated series of cylindrical holes) in a low atomic number material such as beryllium, which is used to minimize absorption.

Total external reflection occurs at angles $\theta<\theta_{\mathrm{c}}$, where the critical angle $\theta_{\mathrm{c}}=\sqrt{2\delta}$. θ_{c} varies inversely with X-ray energy, so that reflection of high-energy X-rays requires shallower angles of incidence onto the mirror. By choosing the angle appropriately, the mirror can be used as a low-pass filter, so that low energies are reflected and higher energies are absorbed by the mirror.

The product of energy E and θ_{c} is a property of the mirror surface and depends on its composition. Representative values of $E\cdot\theta_{\mathrm{c}}$ (in keV-mrad) are $\approx$ 31(Si), 59(Ni), 62(Pd), 67(Rh), 80(Au), and 82(Pt). θ_{c} depends on the nature of the coating, so it is often useful to deposit selected metals in stripes onto the X-ray mirrors. For example, it has proven useful to apply Pt and Rh coatings in stripes to an uncoated substrate made of a low expansion ceramic. By shifting the mirror sideways, different mirror coatings can be chosen without breaking vacuum. A high atomic number coating extends the range of angles that are reflected and allow use of a shorter mirror by reflecting at larger incidence angle. Absorption edges from the mirror coating over the energy range of interest generally are undesirable.

The reflectivity of mirrors (and multilayers) can be calculated using Fresnel's equations [13, 14]. For a sufficiently thick (at least tens of nanometers) mirror material or coating, the reflectivity (assuming zero roughness) may be written [15] in terms of the reduced angle $\phi=\theta/\theta_{\mathrm{c}}$ as

$$R(\phi)=\frac{(\phi-A)^2+B^2}{(\phi+A)^2+B^2},$$

where

$$2A^2=\left[\left(\phi^2-1\right)^2+\left(\beta/\delta\right)^2\right]^{1/2}+\left(\phi^2-1\right),$$

and

$$2B^2=\left[\left(\phi^2-1\right)^2+\left(\beta/\delta\right)^2\right]^{1/2}-\left(\phi^2-1\right).$$

A contour plot of this theoretical reflectivity versus angle and energy is shown in Figure 69.12.

69.6.5 Harmonics

Inspection of the formula for the Bragg condition indicates that photons of an energy that is an integer multiple of the desired fundamental energy may pass through the monochromator. Some of these may be "forbidden reflections," which are very weak, while others are "allowed reflections." Whether a reflection is allowed or forbidden depends on crystal symmetry.

These higher harmonics must be eliminated or otherwise reckoned with, or serious systematic errors will occur in the absorption spectra. For diamond-structure

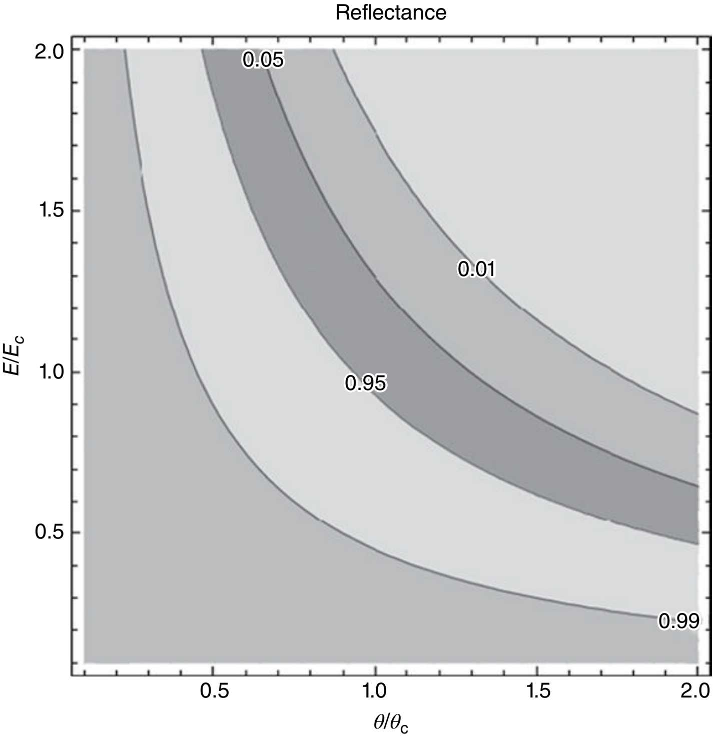


FIGURE 69.12 Contour plot of calculated mirror reflectivity versus angle for $\beta/\delta = 0.1$. Contour levels shown are 1, 5, 95, and 99% of maximum reflectivity.

monochromator crystals such as silicon and germanium (or diamond), reflections are only allowed if the indices $\langle h, k, l \rangle$ satisfy either of the following conditions: $\langle h, k, l \rangle$ are all odd integers, or $h+k+l=4m$, where h, k, l, m are integers, because otherwise the structure factor [3] vanishes. Examples of allowed reflections are ⟨1, 1, 1⟩, ⟨2, 2, 0⟩, ⟨3, 1, 1⟩, ⟨3, 3, 1⟩, and ⟨3, 3, 3⟩. The allowed ⟨3, 3, 3⟩ reflection is three times the energy ("third harmonic") of the ⟨1, 1, 1⟩ and must be eliminated if the ⟨1, 1, 1⟩ reflection is used in the experiments. The ⟨2, 2, 2⟩ reflection is forbidden.

Harmonics can be measured in several ways: by attenuating the beam with a series of attenuators of different thickness and known composition (which increases the proportion of harmonics, "beam hardening") and solving a simple set of equations or using a scatterer and an energy dispersive detector (and some calibration) to measure the proportion of harmonic and fundamental.

There are several standard ways to eliminate harmonics. One method ("detuning the monochromator") is to intentionally misalign the second crystal of the monochromator a small amount with respect to the first crystal, using a piezoelectric actuator. This reduces the amount of the fundamental energy photons that get through, typically by about half, but the higher-energy harmonics are attenuated to a much greater degree than is the fundamental.

Another common method of reducing harmonics is to use a harmonic rejection mirror, set to an angle that reflects the fundamental but does not reflect the harmonics. See "X-Ray Lenses and Mirrors."

A third method that is not widely employed at synchrotron sources (because of their high intensities) is to use energy-sensitive detectors to selectively record photons at the fundamental energies and reject the harmonics by electronic means.

A type of device called a "beam cleaner" [16] is a medium-resolution (≈100 eV) bent Laue optic that can be used to select a particular harmonic (or fundamental) energy that is transmitted by the beamline's primary high-resolution monochromator. This allows the experimenter to use higher-order harmonics for experiments, effectively extending the useful energy range of the beamline. Conversely it can be used to eliminate harmonics from the beam. Use of the medium-energy-resolution optic simplifies electromechanical tracking of the secondary monochromator with the primary monochromator, significantly reducing the cost and complexity.

69.7 DETECTORS

XAFS measurements require precise monitoring of the incident and transmitted fluxes I_0 and I. The incident intensity from a synchrotron source can fluctuate for a variety of reasons, but if the detectors are sufficiently linear and other precautions are taken, the fluctuations will divide out between the two detectors. Good linearity $\approx 10^{-4}$ is required separated for each detector. Even if the incident beam intensity I_0 were kept perfectly constant (which can be approximately arranged through use of beam intensity leveling servo controls), the I beam transmitted by the sample can vary by more than an order of magnitude, because of the energy-dependent absorption by the sample. Therefore it is not sufficient to simply make use of identical nonlinear I_0 and I detectors: it is necessary to use good quality detectors and be sure to operate them in the linear range or to otherwise compensate for their nonlinearities on a detector by detector basis. A good reference on detectors is the book by Knoll [17].

In fluorescence detection a large area detector is placed near the sample, typically at 90° to the beam in the horizontal plane, to collect the fluorescence. The X-rays produced by bend magnets, and planar wigglers and undulators, are normally polarized in the horizontal (orbital) plane of the ring. The elastic scattering, which causes an undesirable background, is minimum along the direction of the X-ray polarization vector.

Background is caused by scattered X-rays from the sample or undesired fluorescence from elements other than the one of interest. If the number of detected background photons were strictly constant, that background could simply be subtracted out, but because the signal; consists of photons, their number fluctuates, contributing noise. Let signal S be the number of the desired fluorescence photons that are detected in a counting interval, and let B be the number of background photons. The noise fluctuations vary as $\sqrt{S+B}$, so the S/N ratio is $S/\sqrt{(S+B)}$. It is convenient to define

$N_{\text{eff}} = S^2/(S+B) = S/(1+B/S)$, the effective number of counts, as the number of signal photons that would give the same S/N ratio in the absence of background. This shows that the presence of background reduces the effective counts by the ratio $1/(1+B/S) \approx S/B$ for large background-to-signal ratio. A background-to-signal ratio of 100 effectively increases the necessary counting time per point to reach a specified S/N ratio by a factor of 100. Therefore it is essential to reduce the amount of background when measuring dilute species.

The detectors for fluorescence are selected by the experimenter for a (sample-dependent) compromise between large area (or solid angle), good energy resolution/background rejection, and high maximum count rate. The most common detectors for fluorescence detection are multielement solid-state detectors (silicon or germanium), including silicon drift detectors (SDDs); large area fluorescence ionization chambers with Z-1 filters and slits; positive-intrinsic-negative (PIN) diodes/PIPS detectors; and scintillators with photomultipliers. Each of these can be supplemented with diffractive analyzers (synthetic multilayers, or Bragg or Laue geometry diffractive optics) in order to reduce background.

69.7.1 Ionization Chambers and PIN Diodes

The most common detectors to use in transmission mode XAFS experiments are ionization chambers, which consist of parallel conducting plates, internally supported by insulators, in a conducting box with X-ray transparent windows on each end, and containing a suitably chosen “fill gas,” for example, N_2, Ar. Ionization chambers do not provide energy resolution—basically they measure the total amount of energy that is deposited in the chamber by absorption of photons. It takes approximately 30 eV on average to create an electron–hole pair in fill gases.

A voltage (typically several hundred volts) is applied between the plates so that when X-rays pass through the windows and between the conducting plates, partially ionizing the gas, the electric field pushes the negative electrons toward the anode and the positive ions toward the cathode. The resulting currents (typically 10–100 nA) are converted to voltage signals using a current amplifier (transimpedance or transconductance amplifier) and are measured by the data acquisition system. This often is done using a voltage to frequency converter (usually 100 kHz/V), which generates pulses at a frequency that is proportional to the input voltage. The pulses are then counted with a multichannel scaler for a fixed time on the order of 1 s. Alternatively, A/D converters can be used to digitize the signal. The time constants over which the input currents are smoothed/integrated by the amplifier should be the same for the I_0 and I channels; otherwise variations in flux of the incident beam will not accurately divide out.

It is essential to operate ionization chambers in what is called the “plateau region.” If the voltage applied to the ionization chamber is insufficient to separate the electron–ion pairs from each other, they will recombine, and the ionization event not be recorded, giving nonlinear response. For this reason the experimenter should verify that the voltage is sufficient to obtain linear response. However if the voltage is too large, liberated electrons will be accelerated to sufficient energy before colliding with another gas atom that they can ionize that atom and create secondaries, so that one obtains a pulse of current whose total charge is proportional to the energy of the absorbed photon. This is a proportional counter. Increasing the voltage further yields a Geiger counter. These have their uses but have limited count rates, so normally for XAFS measurements a linear ionization chamber response is required.

PIN diodes and PIPS detectors (Canberra Inc.), when operated in current mode, essentially behave like solid-state ionization chambers, except they normally absorb nearly all of the X-rays incident upon them, and are therefore not used as semitransparent intensity monitors in the way ionization chambers are. PIN diodes can be used as intensity monitors by detecting scatter or fluorescence from thin foils inserted in the beam. They do not require high-voltage supplies to create sufficient electric field to separate the electrons from their corresponding positively charged "holes." Additionally they can be operated in pulse-counting mode as described in the next section. PIN diodes are convenient for monitoring X-ray intensities when space is limited. They are sensitive to near-infrared and visible light so they must be shielded from it when using them for X-ray detection.

69.7.2 Solid-State Detectors, SDDs, and APDs

Solid-state (semiconductor) detectors, such as silicon or germanium detectors, and SDDs are "energy dispersive," that is, they are normally used to provide a measure of energy resolution of the absorbed photons. A photon that is absorbed in the semiconductor material (Si or Ge) creates electron–hole pairs, which, as in an ionization chamber, are separated by application of an electric field, often associated with a high applied voltage, and collected by electrodes. This electric field would result in a current sufficient to damage the device, if the semiconductor were kept at room temperature, so the detector elements normally are cooled with liquid nitrogen or Peltier cooler, which also reduces electronic noise.

The charge pulses are preamplified and then passed through a shaping amplifier, which returns voltage pulses with heights that are proportional to the total charge of each pulse. These pulses have energies that are proportional to the energy of the absorbed photon, and they can then be selected with discriminators or sorted into bins with a multichannel analyzer (MCA) to provide an energy spectrum.

Alternatively, a more modern approach (e.g., X-ray Instrumentation Associates Digital X-ray Processor) uses a very high-speed analog-to-digital converter (flash ADC) to sample and digitize the pulses and proprietary algorithms running on field-programmable gate arrays (FPGAs) to perform the quantitative pulse analysis. These devices effectively combine the functionalities of a high-voltage supply, preamplifier, shaping amplifier, and MCA into a single compact unit per detector channel. Devices similar to this are available from several vendors.

APDs are fast detectors that operate effectively like solid-state Geiger counters. They offer little to no energy discrimination but are useful for ultrafast experiments.

69.8 SAMPLE PREPARATION AND DETECTION MODES

69.8.1 Transmission Mode

As described above, in transmission mode the quantity that is measured is the fraction of X-rays that are transmitted through a sample at a given energy. To do this the sample must be thin enough to allow sufficient X-rays to penetrate the sample so that adequate photon statistics are obtained, while it must be thick enough to provide adequate absorption contrast as a function of energy. Although theoretically the optimal S/N ratio is obtained when the edge step $\Delta\mu x \approx 2.5$, where $\Delta\mu$ is the difference in absorption coefficient above and below the edge and x is the sample thickness, in practice systematic errors such as "thickness effects" become worse for thicker samples, and experience has shown that in most cases restricting $\Delta\mu x \approx 1$ gives more reliable results. If the edge step is less than about 0.1, it may be better to consider fluorescence measurements. Thickness effects generally reduce the apparent amplitude of structure in the spectra. These effects are described in more detail in Bunker [3].

Samples for transmission mode can be polycrystalline, crystalline, or amorphous films, coatings, and powders on an X-ray thin support or bound in a compressed pellet. It is important that the sample is homogeneous and of uniform thickness (meaning the integrated absorption through the sample is constant from point to point over the illuminated area of the sample). A variation in thickness introduces a nonlinear dependence of the apparent absorption $(\mu x)_{\text{eff}}$ on the absorption coefficient μ:

$$(\mu x)_{\text{eff}} = \mu\bar{x} - \mu^2\sigma^2/2 + \cdots$$

where σ^2 is the variance of the thickness distribution [3].

69.8.2 Fluorescence Mode

A schematic fluorescence mode experiment is shown in Figure 69.13. In this case the sample typically is placed at 45° to the incident beam, and the fluorescence detector is placed at 90° in the horizontal plane, to minimize elastic scatter. If there are effective ways to eliminate scatter, it can be beneficial to collect a greater solid angle by using multiple, or larger, detectors.

The measured fluorescence signal is proportional to

$$\mu_i(E)\csc(\theta_{\text{in}})\big(1-\exp(-\alpha\tau)/\alpha\big),$$

where $\mu_{\text{T}}(E)$ are, respectively, the absorption coefficients of the element of interest and the total sample absorption coefficient at energy E, τ is the sample thickness, $\alpha = \mu_{\text{T}}(E_{\text{f}})$

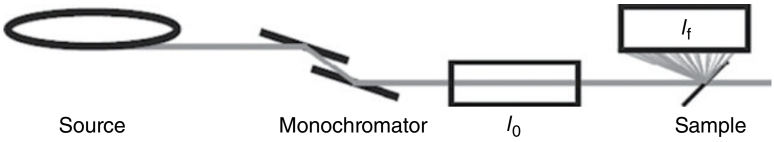


FIGURE 69.13 Schematic fluorescence mode X-ray absorption measuring apparatus (not to scale).

$\csc(\theta_{\text{in}}) + \mu_{\text{T}}(E)\csc(\theta_{\text{out}})$, $\mu_i(E)$, and θ_{in} and θ_{out} are the incidence and exit angles measured relative to the surface. This expression in principle should be integrated over the various exit angles subtended by the detector, but this simple form is sufficient for our purposes.

For thin samples, that is, $\alpha\tau \ll 1$, this becomes $\mu_i(E)\csc(\theta_{\text{in}})\tau$ as expected: the measured signal is proportional to the absorption coefficient times the projected sample thickness. In this case the measured fluorescence is proportional to the absorption coefficient.

For samples that are thick, if the element of interest does not contribute significantly to the total absorption coefficient μ_{T}, then the measured fluorescence depends linearly on the absorption coefficient μ_i but with an energy-dependent prefactor that depends on the composition of the sample matrix. Generally this can be calculated and compensated for.

For samples that are thick, if the element of interest *does* contribute significantly to the total absorption coefficient μ_{T}, the measured fluorescence depends nonlinearly on the absorption coefficient μ_i. In this case measured spectra can be seriously distorted if no other measures are taken. The nature of these distortions typically is that fine structure in the measured spectra is suppressed. This effect is called self-absorption or overabsorption, because the origin of the effect is a variation in effective penetration depth that depends on the absorption coefficient of the element of interest. As the true absorption $\mu_i(E)$ increases, the measured fluorescence does not increase in proportion, and features may be suppressed.

Algorithms have been developed [3] to computationally correct for these effects if the composition of the sample is known. Alternatively, by choosing $\theta_{\text{in}} \approx 90°$ and $\theta_{\text{out}} \approx 0°$, these effects can be significantly reduced, at the cost of reduced solid angle subtended by the detector.

69.8.3 HALO

A mnemonic, "HALO," has proved helpful in remembering to take measures that will reduce errors in XAFS measurements. It stands for harmonics, alignment, linearity, and offsets.

Harmonics must be eliminated from the beam; sample alignment must be checked, so that nothing intercepts the beam between the I_0 detector and I (or I_f) detector except for the homogeneous sample and homogeneous windows; linearity of detectors must be maintained (e.g., by checking ion chambers are used in their plateau region) or corrected, if nonlinear (e.g., by performing dead time corrections in pulse-counting detectors); and offsets (dark currents/amplifier offsets) must be measured and subtracted out before dividing by reference signals.

It should be mentioned that errors will be introduced if the I_0 detector (and in transmission the I detector) is placed too close to the sample, allowing fluorescence from the sample to get into the detector(s). This additional signal slightly corrupts the measured signal(s) and should be avoided.

69.8.4 Sample Geometry and Background Rejection

Several strategies are used to reject scattered background and unwanted extraneous fluorescence before they reach the detector. These include use of beam polarization; grazing incidence; attenuators, filters, and slits; and diffractive analyzers.

69.8.4.1 Use of Polarization Normally at synchrotron radiation sources (bend magnets, wigglers, planar undulators), the X-rays from the source are linearly polarized with the electric field vector in the horizontal plane; also the polarization is perpendicular to the direction of propagation. The scattering is minimum along the polarization direction, so fluorescence detectors are often placed in that location relative to the sample, that is, centered in the horizontal plane and perpendicular to the direction of travel of the X-ray beam. This significantly reduces background from elastic scatter.

69.8.4.2 Grazing Incidence At sufficiently small angles of incidence (typically milliradians) onto a flat surface, total external reflection from the sample will occur (see "X-Ray Lenses and Mirrors"). This has the effect of enhancing the X-ray field and confining it to a region very close to the surface of the sample. This confers a surface sensitivity to fluorescence experiments. It is also possible to determine the absorption coefficient and provide depth-dependent information by carefully measuring the reflected beam intensity as a function of angle and energy.

Even when the angle of incidence is too large for total external reflection to occur, a shallow ("grazing" or "glancing") angle of incidence enhances the surface sensitivity for purely geometrical reasons, which can be beneficial, particularly if elastic scatter or fluorescence from the substrate produces undesired background.

69.8.4.3 Electron Yield for Reducing Background Use of electron yield detection can eliminate problems associated with self-absorption effects in fluorescence and can reduce the effects of fluorescence from the sample substrate. As described above, most XAFS experiments are carried out in transmission mode or fluorescence mode. A third mode is to measure (directly or indirectly) the electrons that are ejected from the surface of the sample from nonradiative deexcitation of the core hole that is produced by the initial absorption event. If the samples can be placed in a vacuum (as is typically done for soft X-rays), the electrons can be collected and integrated to produce a signal current. Alternatively, the sample can be surrounded by helium, and the "conversion electrons" that are produced by collisions of the electrons that ejected from the sample with the helium atoms can be collected by applying an electric field. The advantage of this method is that it confers strong surface sensitivity, because the electrons within the sample have path lengths only on the order of 1000 Å, so only those atoms that are near the surface are visible by this method. It should be mentioned that this method can be difficult when using high-intensity undulator beams, because of sample charging and other effects.

69.8.4.4 X-Ray (Z-1) Filters and Attenuators In fluorescence detection of dilute species, if the fluorescence energy is significantly below the absorption edge of the element of interest, it may be possible to make a foil of containing a different element (typically one atomic number lower in the periodic table) that has its own absorption edge between the fluorescence energy of interest and the energy of excitation. When this is possible, it may be possible to selectively attenuate the elastically scattered background while only moderately absorbing the desired fluorescence. Of course this "Z-1 filter" itself will fluorescence, but through careful use of slits much of this refluorescence can be blocked from entering the detector. Significant improvements in S/N ratio can be obtained with an optimized system.

If there is a large amount of unwanted fluorescence at much lower energies than the desired fluorescence, for example, fluorescence from an element of lower atomic number, it may be beneficial to place an attenuator such as aluminum foils between the sample and the detector. The absorption coefficient of most elements varies as $1/E^3$ over energy regions in which there are no absorption edges, so the absorption coefficient is approximately eight times larger at half the energy. Choosing an attenuator of optimal thickness can significantly improve the S/N ratio. The optimal thickness can be calculated or measured experimentally by optimizing the effective counts as a function of attenuator thickness.

69.8.4.5 Diffractive Analyzers Since the desired fluorescence is found at a well-defined energy, it may be possible to accept a useful fraction of the fluorescence with a diffractive analyzer, provided the spot size on the sample is sufficiently small. The difficulty is that the fluorescence radiation is emitted in all directions into 4π solid angle, and it is difficult for an optic to collect a large fraction of that. In order to meet the Bragg condition across a curved crystal, the correct shape is a logarithmic spiral that is specific to the fluorescence energy and the crystals and crystal cut that are used. To approximate that precise shape, the crystals must either be bent or a series of crystals may be used to approximate a bent surface. The incident and diffracted X-rays may be on the same side of the crystal (Bragg geometry) or the diffracted X-rays may pass through the crystal and emerge from the other side (Laue geometry). Laue geometry can be advantageous for improving the collection efficiency because the X-rays can be incident at a small angle with respect to the local surface normal. The angle tolerances for perfect crystals in Bragg geometry are very small, requiring high precision. In Laue geometry, under strongly bent conditions and suitable asymmetric cut crystals (in which the diffracting planes are parallel to the crystal surface), the perfect crystals have a much broader angular acceptance, which improves the throughput. Diffractive analyzers (some of which are commercially available) have been made with silicon (Bragg and Laue) [18–20], LiF (Bragg), pyrolytic graphite (Bragg) [21], and graded synthetic multilayers (Bragg) [12].

69.8.5 Oriented Samples

In single crystals or otherwise oriented samples, the absorption coefficient depends on the relative orientation of the X-ray polarization vector and the crystal axes. The absorption coefficient is well approximated as a second rank tensor (in the dipole approximation). It is usually assumed that these complications average out to near zero when the sample is in polycrystalline or solution form, and they automatically vanish for crystals of cubic symmetry (the absorption is isotropic). However if the sample is a powder and the grains have a nonrandom orientation ("preferred orientation" or "texture"), systematic errors can result. Magic angle spinning can be used to average out such undesired effects for crystalline samples [3].

In this case the sample is spun around the sample normal direction $\hat{n}$, which makes an angle ($\arccos(1/\sqrt{3}) \approx 54.74°$) ("magic angle") relative to the X-ray polarization direction (Fig. 69.14). This averages out the off-diagonal terms of the tensor and makes the resulting averaged absorption equivalent to the isotropic case. There should be an integer number of full revolutions of the sample rotation (or at least a large number of rotations) within a signal integration period.

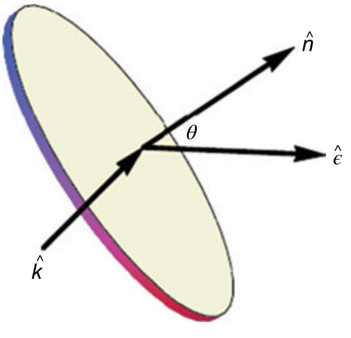


FIGURE 69.14 Magic angle spinning geometry. $\hat{n}$ is the sample normal direction, $\hat{\epsilon}$ is the electric polarization direction, and $\hat{k}$ is the direction of the incident beam. θ is set to the magic angle 54.7°.

69.9 ABSOLUTE MEASUREMENTS

In routine XAFS measurements it is normal to measure the absorption of windows and air path integrated with the data and then to remove it through numerical background subtraction using cubic splines or other means. Similarly the energy dependence of detector efficiency contributes an additive background to the transmission mode XAFS spectra, and an estimate of that effect can be computationally generated and removed. In fluorescence and electron yield mode, this same effect multiplies the data by a slowly varying energy-dependent function, which will introduce a small systematic error in the amplitudes if uncorrected. Of course these background effects can be measured directly and corrected for but at the cost of valuable additional beam time. In routine experiments the numerical approach is conventionally used.

However for more accurate or absolute measurements, these effects and others must be accounted for. An example is measuring or correcting for the signals from scattered and fluorescence X-rays that are emitted by the sample and that may be absorbed in the detectors. If the detectors are too close to the sample, these effects can be significant; they can be reduced through use of suitable masks and evacuated pipes ("flight tubes") to increase the distance between the sample and the detectors.

Other confounding effects can stem from temperature and pressure variations within ionization chambers, which can affect the gas density in a time-dependent manner. A 3°C variation in the ionization chamber temperature can cause a 1% variation in detector output. These environmental influences can be controlled in obvious ways.

Examples of high-accuracy absolute measurements of X-ray absorption are given in Glover and Chantler [22] and subsequent work. Absolute measurements of the X-ray absorption coefficient also require precise characterization of the sample uniformity and thickness.

REFERENCES

1. E.A. Stern and S.M. Heald "Basic Principles and Applications of EXAFS", in E.E. Koch (ed), *Handbook on Synchrotron Radiation*, North-Holland Publishing Company, Amsterdam/New York/Oxford, 955–1014, (1983).
2. D.C. Koningsberger and R. Prins *X-Ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS, and XANES*, John Wiley & Sons, New York, (1988).
3. G.B. Bunker *Introduction to XAFS: A Practical Guide to X-Ray Absorption Fine Structure Spectroscopy*, Cambridge University Press, Cambridge, UK, (2010).
4. C.T. Chantler, K. Olsen, R.A. Dragoset, J. Chang, A.R. Kishore, S.A. Kotochigova, and D.S. Zucker, X-Ray Form Factor, Attenuation and Scattering Tables (version 2.1) (2005). [online] http://physics.nist.gov/ffast (accessed on December 2, 2015).
5. Lawrence Berkeley Labs, Center for X-Ray Optics, X-Ray Data Booklet. [online] http://xdb.lbl.gov/ (accessed on December 2, 2015).

6. [online] http://www.csrri.iit.edu/periodic-table.html (accessed on November 4, 2015).
7. [online] http://www.lightsources.org/fels (accessed on November 4, 2015).
8. [online] http://www.lightsources.org (accessed on November 4, 2015).
9. [online] http://www.aps.anl.gov/epics/ (accessed on November 4, 2015).
10. [online] http://www.tango-controls.org/ (accessed on November 4, 2015).
11. [online] http://mx.iit.edu/ (accessed on November 4, 2015).
12. K. Zhang, G. Rosenbaum, R. Liu, C. Liu, C. Carmeli, G. Bunker, and D. Fischer "Development of multilayer analyzer array detectors for X-ray fluorescence at the third generation synchrotron source eighth international conference on synchrotron radiation instrumentation" *AIP Conf. Proc.*, 705, 957–960, (2004).
13. D.J. Griffiths *Introduction to Electrodynamics*, Third Edition, Prentice Hall, Saddle River, NJ, (1999).
14. J. Als-Nielsen and D. McMorrow *Elements of Modern X-Ray Physics*, John Wiley & Sons, Ltd, Chichester, UK, (2011).
15. T. Matsushita and H. Hashizume "X-Ray Monochromators", in E.E. Koch (ed), *Handbook on Synchrotron Radiation*, Vol. 1, North-Holland Publishing Company, Amsterdam/New York/Oxford, 261, (1983).
16. C. Karanfil, L.D. Chapman, G.B. Bunker, C.U. Segre, and N.E. Leyarovska "A 'beam cleaner' for harmonic selection/rejection" *Rev. Sci. Instrum.*, 73, 3, 1505, (2002).
17. G.F. Knoll *Radiation Detection and Measurement*, John Wiley & Sons, Inc., Hoboken, NJ, (2002).
18. B.W. Adams and K. Attenkofer "An active-optic X-ray fluorescence analyzer with high energy resolution, large solid angle coverage, and a large tuning range" *Rev. Sci. Instrum.*, 79, 023102-1–12, (2008).
19. Z. Zhong, L. Chapman, B. Bunker, G. Bunker, R. Fischetti, and C. Segre "A bent Laue analyzer for fluorescence EXAFS detection" *J. Synchrotron Radiat.*, 6, 212, (1999).
20. C. Karanfil, G. Bunker, M. Newville, C.U. Segre, and D. Chapman "Quantitative performance measurements of bent crystal Laue analyzers for X-ray fluorescence spectroscopy" *J. Synchrotron Radiat.*, 19, 375380, (2012).
21. D.M. Pease, M. Daniel, J.I. Budnick, T. Rhodes, M. Hammes, D.M. Potrepka, K. Sills, C. Nelson, S.M. Heald, D.I. Brewe, and A. Frenkel "Log spiral of revolution highly oriented pyrolytic graphite monochromator for fluorescence X-ray absorption edge fine structure" *Rev. Sci. Instrum.*, 71, 3267, (2000).
22. J.L. Glover and C.T. Chantler "The analysis of X-ray absorption fine structure: beam-line independent interpretation" *Meas. Sci. Technol.*, 18, 29162920, (2007).

70

NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY

Kenneth R. Metz
Chemistry Department, Merkert Chemistry Center, Boston College, Chestnut Hill, MA, USA

70.1 INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is arguably the single most powerful instrumental technique used in modern chemistry and biochemistry, and it also plays significant roles in physics, materials science, physiology, and medicine. The amount of information available from NMR studies is enormous and ranges from the structures of molecules to the details of molecular interactions and the dynamics of molecular motions. It is nondestructive and noninvasive, a crucial asset for many investigations. Virtually any sample type (solid/liquid/gas, or animal/vegetable/mineral) is amenable to study by this method. The primary limitation of NMR is its relatively low sensitivity compared to other forms of spectroscopy, which usually mandates the use of large samples or long data acquisition times, if not both.

Although it is fair to characterize NMR as a mature technique, the field continues to evolve dynamically with frequent, regular advances in both hardware and methodology. Nearly 70 years of continuous, aggressive development has provided investigators with a staggering range of NMR techniques and experiment types. At first glance, so much technology can seem bewildering and intimidating. However, the great majority of NMR studies employ a relatively small set of pulse experiments. This chapter presents the basic principles and some experimental aspects of NMR.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

70.2 HISTORICAL REVIEW

At the heart of NMR is nuclear spin, the existence of which was confirmed by the 1924 atomic beam experiments of Stern and Gerlach [1]. Magnetic resonance with atomic beams was actually achieved by Rabi in 1938 [2]. (Nobel Prizes in Physics for that work were awarded to Stern in 1943 and to Rabi in 1944.) The next logical step was to observe NMR in condensed matter. Initial attempts at ^{7}Li NMR by Gorter and Broer [3] in 1942 were unsuccessful, mainly due to bad luck, before World War II interrupted research in the field. Aided by wartime developments in radio technology and knowing that NMR represented a still-unplucked "plum," several major physics groups were poised to tackle the problem when the war ended. As a result, in 1945, two American groups succeeded in observing ^{1}H NMR signals in water and paraffin for the first time [4, 5]. The importance of that work resulted in physics Nobel Prizes in 1952 for the lead investigators of both groups (Felix Bloch at Stanford and Edward Purcell at Harvard). Physicists initially hoped that NMR methods would allow the measurement of exquisitely accurate values for magnetic moments and fundamental physical constants like the gyromagnetic ratio (γ), but Knight [6], followed by Proctor and Yu [7], soon discovered that the experimental values depended on which chemical compound was studied. This was termed the "chemical shift," and it suddenly made NMR highly interesting to chemists for studies of molecular structures. The 1950s and the early 1960s witnessed enormous development in NMR technology, including the introduction of the first commercial spectrometers, routine line narrowing by sample spinning, magnetic field stabilization (which revealed spin–spin coupling), decoupling methods, and "magic angle" sample spinning for studies of solid samples. In 1966, Richard Ernst and Weston Anderson, working at Varian Associates, took advantage of the growing availability of digital computers to introduce pulsed Fourier transform NMR (FTNMR) [8]. By the mid-1970s, spectrometers could be purchased with dedicated minicomputers to perform the data acquisitions and calculations required for FTNMR. During that same decade, high-field superconducting magnets became available, which displaced the large electromagnets that dominated early NMR. The signal/noise improvements that high-field spectrometers provided stimulated the extension of NMR to nuclides beyond ^{1}H, ^{13}C, ^{19}F, and ^{31}P. In fact, virtually every element in the periodic table became susceptible to study by NMR, opening the door to important new investigations in physical, inorganic, biological, and organometallic chemistry. The dynamic decade of the 1970s witnessed the beginnings of another pivotal development in NMR: multidimensional experiments. Much of the work in two-dimensional (2D) NMR occurred in the research groups of Richard Ernst (Zürich) and Ray Freeman (Oxford), with Ernst receiving the 1991 Nobel Prize in Chemistry for that and his many other contributions to NMR. Multidimensional methods proved to be powerful tools for structure assignments in small molecules, and they revolutionized the determination of three-dimensional (3D) solution-state structures of macromolecules such as proteins. The 2002 Nobel Prize in Chemistry was shared by Kurt

Wüthrich for the work of his group in that area. Biomedical applications of NMR have driven a great deal of development since the mid-1970s. That began with techniques for noninvasive ^{1}H, ^{13}C, and ^{31}P measurements of metabolite concentrations in perfused tissues and live animals. It soon came to include mapping the spatial distributions of water and fat ^{1}H NMR signals in humans and animals, which, under the name "magnetic resonance imaging" (MRI), plays a crucial role in modern medicine. In 2003, Paul Lauterbur and Peter Mansfield shared the Nobel Prize in Physiology or Medicine for their early work in that area. Clearly, NMR has been a mainstream technique for over half a century and has contributed to major progress in a very wide range of applications, yielding Nobel Prizes for eight investigators!

70.3 BASIC PRINCIPLES OF SPIN MAGNETIZATION

NMR spectroscopy operates by manipulating nuclear spins. Nuclear spin is quantized, with a nuclear spin quantum number generally represented by the symbol I. A great many different isotopes (or "nuclides") are endowed by nature with nuclear spins and are thus amenable to study by NMR. That includes at least one nuclide for nearly every element in the periodic table. However, NMR observability is by no means universal. Specifically, nuclides having even atomic numbers *and* even mass numbers possess no nuclear spin (i.e., $I=0$) and cannot be studied by NMR. For example, ^{12}C with its atomic number of 6 and atomic mass of 12 yields no NMR signal. Nuclides that combine an odd atomic number with even atomic mass possess integer spin values, such as $I=1$ for ^{2}H and ^{14}N and $I=3$ for ^{10}B. Finally, an odd atomic mass (regardless of atomic number) produces half-integer spins, such as ^{1}H, ^{13}C, and ^{31}P with $I=1/2$, ^{23}Na and ^{11}B with $I=3/2$, and ^{17}O with $I=5/2$. Table 70.1 lists some of the more important of the available NMR nuclides, along with their I values and other NMR parameters.

Fortunately, some very common nuclides like ^{1}H and ^{31}P are NMR active, and that has certainly contributed to the tremendous utility and explosive growth of NMR over the years. At the same time, it is tempting to lament the NMR inactivity of other nuclides such as ^{12}C and ^{16}O, the most abundant isotopes of those important elements. Although it is still possible to perform NMR studies with carbon and oxygen, the rare nuclides ^{13}C and ^{17}O must be used, resulting in relatively noisy spectra and long data acquisition times. If ^{12}C and ^{16}O did possess nuclear spin, however, spin coupling would greatly complicate routine ^{1}H NMR spectroscopy for most samples. So, in that respect, one can view the absence of nuclear spin in ^{12}C and ^{16}O as a "glass half full."

The nonzero spin of a nucleus endows it with a magnetic moment μ that behaves like a tiny, spinning bar magnet (Fig. 70.1). The magnitude of μ for a particular nucleus is given by

$$\mu=\gamma\left(\frac{h}{2\pi}\right)\left[I\left(I+1\right)\right]^{1/2}, \tag{70.1}$$

TABLE 70.1 NMR Properties of Selected Nuclides

Nuclide	% Abundance	Spin I	Gyromagnetic Ratio γ (MHz T^{-1})	Larmor Freq. ν (MHz) at 11.744 T[a]	Relative Sensitivity[b]	Relative Receptivity[c]
^{1}H	99.985	1/2	42.576	500.01	1.00	1.00
^{2}H	0.0156	1	6.5355	76.753	9.65×10^{-3}	1.50×10^{-6}
^{7}Li	92.58	3/2	16.546	194.32	2.93×10^{-1}	2.72×10^{-1}
^{10}B	19.58	3	4.5754	53.733	1.99×10^{-2}	3.89×10^{-3}
^{11}B	80.42	3/2	13.660	160.42	1.65×10^{-1}	1.33×10^{-1}
^{13}C	1.108	1/2	10.705	125.72	1.59×10^{-2}	1.76×10^{-4}
^{14}N	99.63	1	3.0755	36.119	1.01×10^{-3}	1.00×10^{-3}
^{15}N	0.37	1/2	−4.3142	50.666	1.04×10^{-3}	3.85×10^{-6}
^{17}O	0.037	5/2	−5.7719	67.785	2.91×10^{-2}	1.08×10^{-5}
^{19}F	100	1/2	40.054	470.39	8.33×10^{-1}	8.33×10^{-1}
^{23}Na	100	3/2	11.262	132.26	9.25×10^{-2}	9.25×10^{-2}
^{27}Al	100	5/2	11.103	130.39	2.07×10^{-1}	2.07×10^{-1}
^{29}Si	4.70	1/2	−8.4577	99.327	7.84×10^{-2}	3.68×10^{-4}
^{31}P	100	1/2	17.235	202.41	6.63×10^{-2}	6.63×10^{-2}
^{39}K	93.08	3/2	1.9869	23.334	5.08×10^{-4}	4.73×10^{-4}
^{59}Co	100	7/2	10.054	118.07	2.77×10^{-1}	2.77×10^{-1}
^{109}Ag	48.48	1/2	−1.9813	23.268	1.01×10^{-4}	4.89×10^{-5}
^{113}Cd	12.26	1/2	−9.4427	110.90	1.09×10^{-2}	1.34×10^{-3}
^{133}Cs	100	7/2	5.5846	65.586	4.74×10^{-2}	4.74×10^{-2}
^{199}Hg	16.84	1/2	7.5902	89.139	5.67×10^{-3}	9.54×10^{-4}
^{203}Tl	29.52	1/2	24.332	285.76	1.87×10^{-1}	5.51×10^{-2}
^{205}Tl	70.48	1/2	24.570	288.55	1.92×10^{-1}	1.35×10^{-1}

[a] Larmor (NMR resonance) frequency calculated from Equation 70.2. Signs are irrelevant in this context and have been ignored.

[b] NMR sensitivity at constant magnetic field relative to an equal number of ^{1}H nuclei. Calculated from $\gamma^{3}\cdot I(I+1)$, ignoring signs.

[c] NMR receptivity, including effects of natural abundance, calculated from Equation 70.6.

FIGURE 70.1 Magnetic moment μ of a nucleus with nonzero spin.

where h is Planck's constant and γ is the gyromagnetic (or "magnetogyric") ratio, a fundamental constant for each nuclide. Table 70.1 lists the γ value for each nuclide. The standard units are rad s^{-1} T^{-1} (where T represents teslas), but for the table, radians have been converted to Hertz to permit the use of somewhat more convenient units of MHz T^{-1}.

In an NMR experiment, the nucleus is placed in an external magnetic field, causing the μ vector to precess about the field direction in analogy to the precession of a gyroscope in a gravitational field. The precession frequency is given by the so-called Larmor equation,

$$\begin{aligned} \omega_0 &= \gamma B_0 \left(\text{in rad s}^{-1}\right), \text{or} \\ v_0 &= \frac{\gamma B_0}{2\pi} \left(\text{in Hz}\right), \end{aligned} \tag{70.2}$$

where B_0 represents the magnetic field strength or, more correctly, its flux density (in teslas). Typical laboratory magnets produce fields that result in Larmor frequencies in the radiofrequency (RF) range for most nuclides, as shown in Table 70.1. Thus, for a B_0 of 11.744 T, the Larmor frequency of ^{1}H is 500.0 MHz, as required by the ^{1}H gyromagnetic ratio of +42.5770 MHz T^{-1}.

In addition to causing nuclear precession, the external magnetic field forces the nucleus to adopt a specific orientation. A spin-1/2 nucleus orients either parallel or antiparallel to the direction of the field, as illustrated in Figure 70.2. The two states differ in the sign of μ and have slightly different energies U:

$$U = -\mu B_0. \tag{70.3}$$

A small majority of spins adopt the lower-energy parallel orientation, in turn causing the entire sample to develop a small bulk magnetization vector μ_{bulk} in the direction of the external magnetic field. Because the energy difference between the two states is tiny, only a slight excess of spins (on the order of 0.001–0.0001% for ^{1}H) adopt the parallel orientation to the field, and the bulk magnetization vector of the sample is miniscule. It is this vector that produces the NMR signal, which explains why NMR signals are generally so weak.

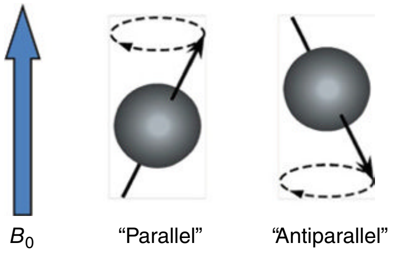


FIGURE 70.2 Possible orientations of spin-1/2 nuclei in an external magnetic field.

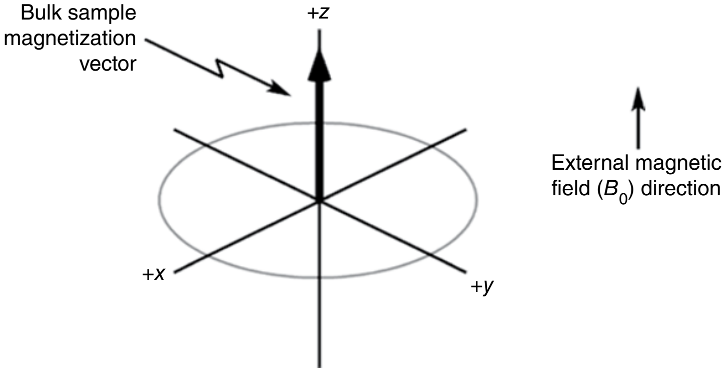

Bulk sample magnetization vector
+z
External magnetic field (B_0) direction
+x
+y

FIGURE 70.3 Representation of an NMR sample's bulk magnetization vector in the "laboratory" frame of reference.

The sample's bulk magnetization vector points in the same direction as the external magnetic field, as shown in Figure 70.3. Two important conventions apply to the Cartesian axis system in the figure. First, the main, external magnetic field vector is always assumed to point in the direction of the +*z*-axis (which, of course, is why the bulk sample magnetization also develops there). Its orientation along the *z*-axis is not meant to imply that the magnetic field itself must actually be vertical in the laboratory. Although the superconducting magnets used in commercial high-resolution NMR spectrometers happen to produce vertical B_0 fields, it is perfectly possible to perform NMR with the main field oriented in any physical direction. There is a second important convention: if the Cartesian axis system is time invariant (i.e., stationary), it is defined as the "laboratory reference frame." A useful alternative is discussed later in this chapter.

70.4 EXCITING THE NMR SIGNAL

Once the sample inside the magnet has developed its bulk magnetization vector, the NMR excitation can be performed. Usually it is done by applying a pulse of RF current to a wire coil that surrounds the NMR sample centered in the external B_0 magnetic field. That creates a second external magnetic field, B_1, along the laboratory reference frame's +*x*-axis to tip the sample magnetization vector away from the +*z*-axis. Once it departs the axis, the magnetization vector precesses about the *z* direction at the Larmor frequency v_0 just like the individual nuclear spin vectors from which it arises. If the sign of the gyromagnetic ratio γ is positive, the most common case, the precession occurs as shown in Figure 70.4. Negative gyromagnetic ratios cause precession in the opposite direction.

While the bulk sample magnetization vector is tipping away from the +*z*-axis, it simultaneously precesses as discussed earlier. In other words, the tip of the vector undergoes "nutation," tracing an oscillatory spiral path in the laboratory reference

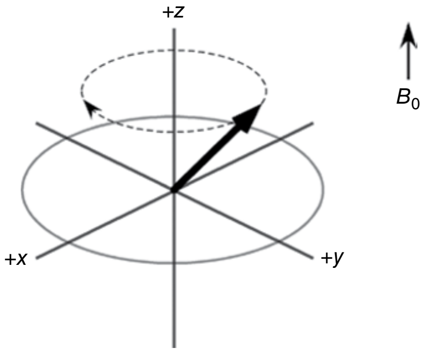


FIGURE 70.4 Precession of the bulk sample magnetization vector about the *z*-axis in the laboratory reference frame. The direction of precession shown assumes the gyromagnetic ratio has a positive value.

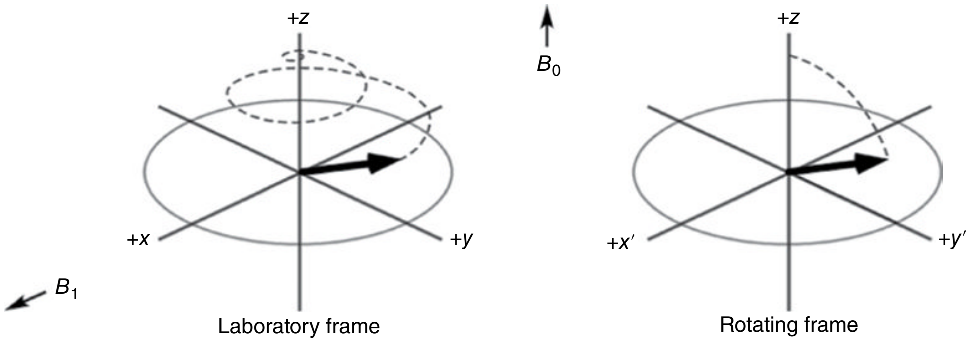


FIGURE 70.5 The motion of a sample's bulk magnetization vector in laboratory and rotating reference frames. If the rotating frame rotates at the precession frequency of the vector, the oscillatory motion disappears. (That condition is called "on resonance.") Using the rotating frame greatly simplifies discussions of NMR experiments.

frame. This complex motion makes it difficult to visualize many NMR experiments, so it is common to abandon the laboratory reference frame and use a frame that rotates about the *z*-axis at frequency v_0. Conceptually, this is a bit like jumping on a carousel: a motion that appears spiral to an observer standing next to the carousel will lack its rotational component when observed by someone who stands on the carousel. This new frame of reference is called (not surprisingly) the rotating reference frame. The difference between vector motions in the two frames is illustrated in Figure 70.5. The precessional motion stops only if the frame's rotation frequency matches the precession frequency of the vector exactly, a condition known as being "on resonance." If the sample produces multiple vectors with different frequencies, the rotating frame cannot match all of them simultaneously and, at most, only one will show no oscillation. Note that a prime is often added to *x* and *y* to denote the rotating frame. (It can be added to *z* as well, but there is little point since the *z*-axes coincide in both frames.)

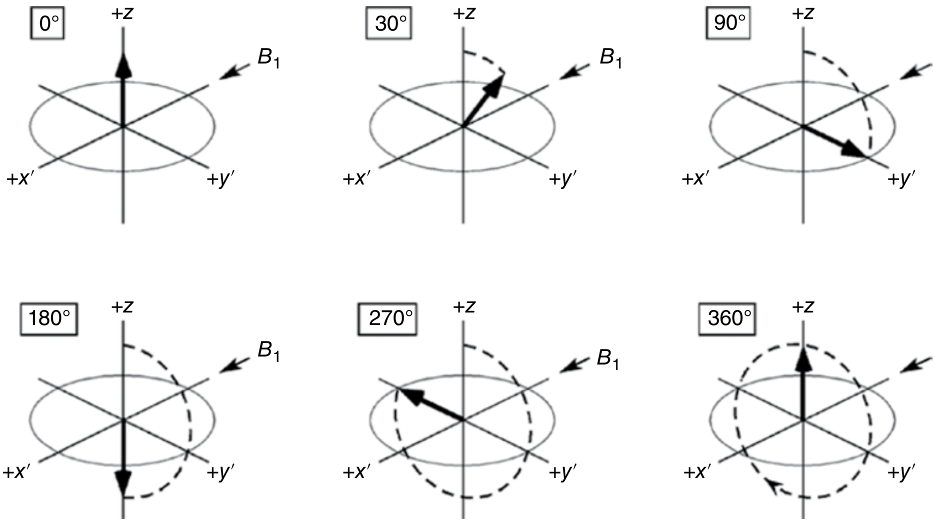


FIGURE 70.6 Nutation (or "tip") angles produced by various B_1 fields applied along the $+x'$-axis in the rotating frame.

Since the bulk sample magnetization precesses at frequency ν_0, an *oscillating* B_1 field at that frequency must be applied along $+x$ in the laboratory frame to tip the vector. In the rotating frame, the field is still applied along $+x'$ but it appears to be *static* and not oscillating. The vector can then be considered to precess about the new $+x'$ magnetic field for the length of time that field remains on. Once the field is turned off, the vector is left somewhere in the $y'z$ plane at an angle θ relative to $+z$. This is referred to as the "nutation angle" or often just the "tip angle." By analogy with Equation 70.2, the vector's precessional frequency ν_1 about the B_1 magnetic field is

$$\nu_1 = \frac{\gamma B_1}{2\pi} \ (\text{in Hz}). \tag{70.4}$$

In practice, the B_1 field is normally applied as a pulse of RF energy of some finite duration t_{pulse}, making it easy to calculate the resulting nutation angle:

$$\theta = \nu_1 \ (t_{\text{pulse}}) \ (360°) = \gamma \ B_1 \ (t_{\text{pulse}}) \ (360°). \tag{70.5}$$

Longer pulses and greater RF field strengths clearly produce greater tip angles. Examples of nutation are shown in Figure 70.6, from which it is clear that the pulse width or B_1 can be chosen to leave the sample magnetization vector at any desired nutation angle. Commercial NMR spectrometers use B_1 fields that are orders of magnitude smaller than their B_0 fields, so pulse widths of 1–100 μs are usually needed to produce a 90° tip.

NMR spectra usually contain multiple signals distributed over a range of precession frequencies. So, it is vital to be able to tip vectors over that, or preferably a greater,

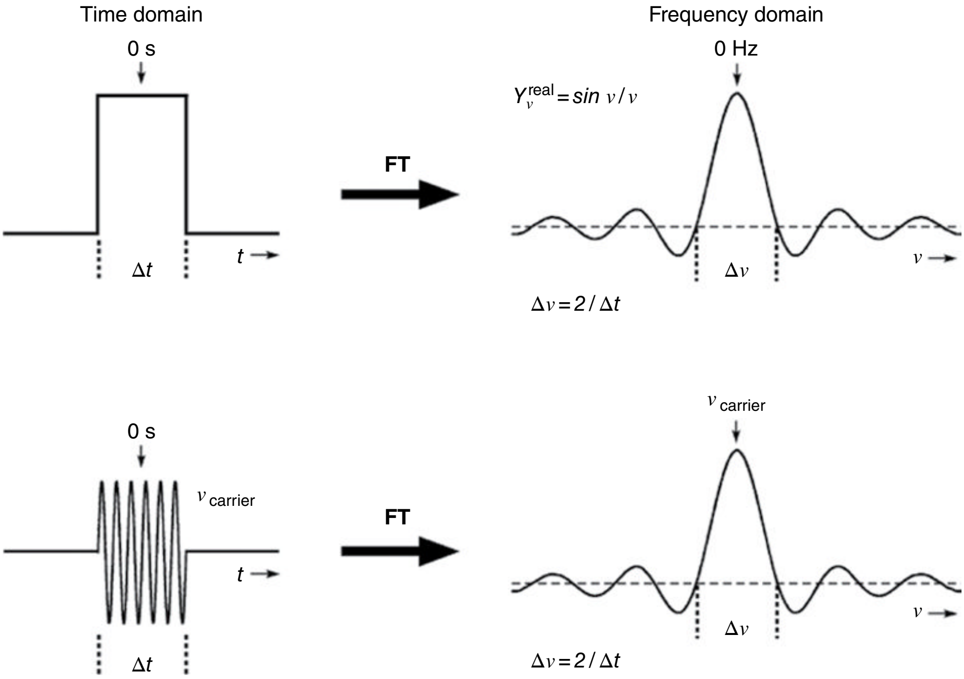


FIGURE 70.7 Fourier pairs showing the relationship between time-domain excitation pulses and the frequencies they excite. The greatest RF power is produced at the "carrier" frequency within the time-domain pulse, but substantial (though not uniform) power is applied at a wide range of other frequencies.

range of frequencies. That is why short RF *pulses* are used for excitation. According to Fourier theory, applying a time-domain pulse excites a range of frequencies simultaneously. As illustrated in Figure 70.7, the Fourier transform of a simple rectangular time-domain pulse waveform produces a frequency-domain (sin *v*)/*v* amplitude function centered at 0 Hz. When the pulse is a rectangular envelope containing a signal oscillating at frequency v_{carrier}, then the frequency-domain excitation profile is still a (sin *v*)/*v* function, but it is centered at v_{carrier} instead of at 0 Hz. An inverse relationship exists between the pulse width Δt and the frequency range Δv that is excited (i.e., $\Delta v = 2/\Delta t$). It is generally best to use the shortest practical pulse since that will excite the widest range of frequencies. As an example, suppose a 500 MHz NMR spectrum is expected to exhibit peaks over a frequency range Δv of 2000 Hz. In that case, the maximum practical RF pulse width is $\Delta t = 2/2000\,\text{Hz} = 1\,\text{ms}$. Provided the 500 MHz carrier frequency is exactly centered in the NMR spectrum, the spectrum will just fit into this 2000 Hz window and all the peaks will be excited. However, from the shape of the (sin *v*)/*v* function in the figure, it is clear that they will not be excited *equally*! If reasonably uniform excitation is required, the entire spectrum should occupy only the middle 5–10% of the excitation bandwidth Δv. In the case of the 2000 Hz spectrum, Δv

must then be ≥20,000 Hz, necessitating an RF pulse width of ≤100 μs (or preferably ≤50 μs). Lest you think that the best way to run all NMR experiments is simply to use extremely short RF pulses all the time, recall that the pulse must also cause a substantial nutation and that large nutation angles result from *long* RF pulses (see Eq. 70.5). Consequently, a compromise pulse width must be employed that (1) is long enough to produce the necessary nutation and (2) short enough to do so more or less uniformly over a suitably wide range of frequencies. To reconcile these competing needs, commercial NMR spectrometers incorporate 50–1000 W RF power amplifiers so that quite short pulses can still produce significant nutations.

In special cases, the inverse relationship between Δt and $\Delta \nu$ can be exploited to excite only a narrow range of frequencies. The most common application of that in NMR is for suppressing intense solvent peaks that would otherwise obscure small analyte signals. A long RF pulse centered at the solvent peak frequency will tend to saturate the peak without perturbing the rest of the spectrum too much. (Saturation results in a weak peak, as discussed in the succeeding text.) An alternative approach is to use long, shaped pulses to selectively excite only the parts of the spectrum one wishes to observe while leaving the interfering solvent peaks mostly unexcited.

70.5 DETECTING THE NMR SIGNAL

Once the sample spins have been excited, the resulting NMR signal must be detected. The same physical coil of wire that is used as the RF transmitter during excitation is also used as the RF receiver during signal detection. Conceptually, however, it is easiest to consider the transmitter and receiver coils to be separate and aligned along the laboratory-frame x and y axes, respectively, as illustrated in Figure 70.8. When the transmitter

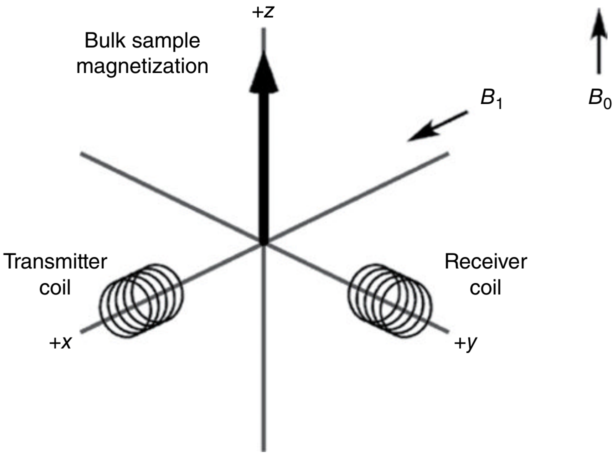


FIGURE 70.8 Conceptual positions of the NMR RF transmitter and receiver coils in the laboratory reference frame.

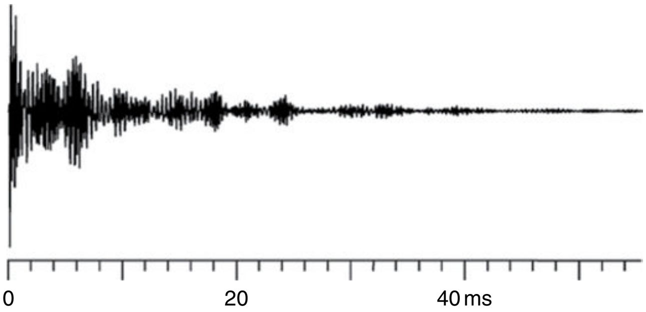


FIGURE 70.9 An experimental ^{13}C NMR free induction decay obtained at 75 MHz from a sample of rat liver in a 7.05 T magnetic field. Nearly all the signals arise from tissue lipids. The oscillatory nature of the FID is evident, as is constructive and destructive interference among the many different frequencies it contains. The progressive decrease in signal intensity is a result of relaxation processes that destroy the *xy* magnetization.

coil's B_1 field is turned on, the sample magnetization vector nutates continuously until the B_1 field is turned off, at which time the vector is left θ° away from +*z* (Eq. 70.5). In the laboratory frame, the vector then precesses about the *z* direction, creating a small magnetic component in the *xy* plane and a weak oscillating magnetic field in the receiver coil. That oscillating field induces an oscillating current in the coil that is converted to voltage, amplified, and ultimately digitized as a time-domain signal called the "free induction decay" (FID). An example is shown in Figure 70.9. The FID constitutes the fundamental data set in an FTNMR measurement. As shown in the figure, its overall amplitude decreases with time. That results partly from destructive interference among its various component frequencies but is mostly caused by relaxation as the signal's *xy* components disappear and magnetization is restored along the *z*-axis.

Since the NMR receiver coil is considered to lie on the +*y*-axis, it is clear that the most intense FID (and spectrum) will be obtained by applying a 90° excitation pulse to tip the initial *z* magnetization fully into the *xy* plane. If multiple FIDs are averaged to form the spectrum, then relaxation effects usually become important and smaller tip angles work best, as discussed later. Figure 70.6 also predicts that a 180° excitation pulse should produce no signal, since the magnetization vector is left aligned along −*z* with no component in the *xy* plane. In ideal cases, that is quite correct, and it is often nearly true in practice.

The intensity of the received NMR signal is governed by more than the excitation tip angle. Some nuclides inherently produce stronger signals than others. For equal numbers of nuclei at a constant magnetic field, a nuclide's NMR sensitivity is proportional to $\gamma^3 \cdot I(I+1)$ in the absence of complicating relaxation effects. The signal intensity obtained from a normal, unenriched sample also depends on the natural abundance *A* of the nuclide. For example, a sample of benzene (C_6H_6) would be expected to produce a stronger NMR signal for ^1H than for ^{13}C simply because nearly all the hydrogen atoms are NMR-active ^1H ($A = 99.985\%$) while few of the carbon atoms are ^{13}C

($A = 1.108\%$). Scaling the sensitivity by the nuclide's natural abundance yields the NMR "receptivity":

$$\text{receptivity} \propto A\ \gamma^3\ I\left(I+1\right), \tag{70.6}$$

an informative parameter that expresses the overall ease of observing the nuclide in an NMR experiment. Since ^{1}H has the highest receptivity of all the common nuclides, values are usually tabulated relative to it, as in Table 70.1. Owing to relaxation effects, the relative receptivity parameter is useful mainly for comparing nuclides having the same spin quantum number. Among spin-1/2 nuclides, ^{1}H and ^{19}F are the most observable. It may be surprising to learn that ^{205}Tl is even more receptive than widely studied ^{31}P, resulting in a substantial literature on ^{205}Tl (and ^{203}Tl) NMR despite the rather high toxicity of thallium compounds. With a relative receptivity of only 0.00018, ^{13}C is not an especially attractive NMR nuclide. However, the prevalence of carbon in chemical compounds makes ^{13}C NMR disproportionately important.

70.6 COMPUTING THE NMR SPECTRUM

Once the FID has been digitized and saved in the computer, it is subjected to discrete Fourier transformation (FT) to form the final frequency-domain NMR spectrum. If the spectrometer is "on resonance" (i.e., if its rotating-frame frequency exactly equals the Larmor frequency), then the FID does not oscillate but simply decays with time. For most samples, the on-resonance time-domain decay is monoexponential, and the FID is just the simple function $y = M_0 \cdot \exp(-t/\tau)$, where M_0 is the initial amplitude of the signal at time $t = 0$. The parameter τ is called the "relaxation time" and $1/\tau$ is the relaxation rate for the process. When such a function is Fourier transformed, a "Lorentzian" frequency-domain line shape is produced, as shown in Figure 70.10. Samples dissolved in liquid solvents normally produce NMR signals that decay monoexponentially, so they produce Lorentzian peaks. FT yields both a real spectrum and a corresponding imaginary spectrum. As case (a) illustrates, the real FT of an on-resonance exponential decay forms a well-behaved, symmetrical "absorption mode" Lorentzian line, while the imaginary FT produces a complicated "dispersion mode" line shape. The latter is difficult to interpret, so only the absorption mode spectrum is usually observed. In practice, the frequency-domain spectrum obtained immediately following FT is rarely either pure absorption or pure dispersion mode. However, it is simple mathematically to process the spectrum to produce the pure modes. That process, called "phasing," can be performed automatically by the computer or interactively under operator control.

What if the FID is acquired "off resonance" (i.e., at a rotating frame frequency that differs from the precession frequency)? That is extremely common and causes the FID to oscillate at the frequency difference while simultaneously decaying exponentially. Figure 70.10 illustrates the effect (case b). Luckily, FT still produces a Lorentzian line,

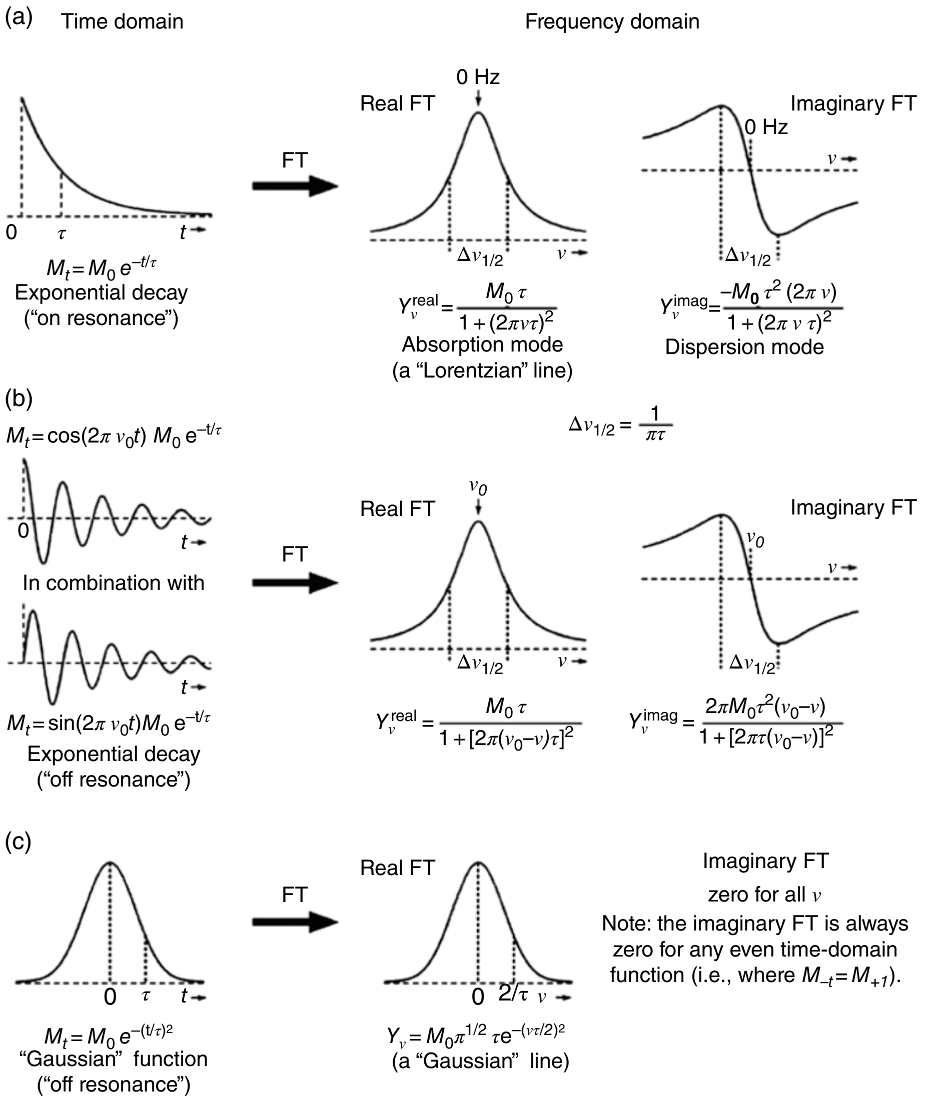


FIGURE 70.10 Fourier pairs relating to NMR line shapes. (a) Fourier transforming a simple monoexponential time-domain decay produces a Lorentzian line having real and imaginary parts as shown. Usually, only the real (absorption mode) part is displayed in spectra. (b) An oscillating monoexponential decay function also produces a Lorentzian line, but one centered at the frequency of the oscillation. (c) Fourier transforming a Gaussian time-domain decay function yields a Gaussian NMR line.

but its frequency is shifted away from zero. The horizontal scale in an NMR spectrum is usually labeled "chemical shift," but in fact it is a frequency scale and the various peaks appear at their frequency differences relative to the spectrometer's rotating frame. Since instrumental factors can shift all the frequencies simultaneously, an internal standard compound such as tetramethylsilane (TMS) is usually added to the sample, and its peak position is arbitrarily defined as zero on the scale. Other NMR peak positions are defined by their differences relative to that reference frequency.

Unlike solutions, solid NMR samples generally produce FIDs that have maximum amplitude at time 0 and then exhibit Gaussian decay (i.e., as a function of $\exp(-t/\tau)^2$ instead of $\exp(-t/\tau)$). Interestingly, the Fourier transform of a Gaussian function yields another Gaussian function (case (c) in the figure). So, NMR spectra of solid samples typically contain Gaussian, and not Lorentzian, peak shapes.

So far we have mostly assumed the presence of only a single resonance line in the spectrum. NMR would not be very useful if its applications were restricted to samples with single lines! Exciting a multiple-spin system produces an FID that is a sum (i.e., interferogram) of individual oscillatory functions. As shown in Figure 70.9, the appearance can be quite complicated. Fortunately, the FT of a sum of functions is equal to the sum of the FTs of each separate function:

$$\mathrm{FT}\left\{f_1(t)+f_2(t)+f_3(t)+\cdots\right\}=\mathrm{FT}\left\{f_1(t)\right\}+\mathrm{FT}\left\{f_2(t)\right\}+\mathrm{FT}\left\{f_3(t)\right\}+\cdots. \tag{70.7}$$

This is called the Fourier addition theorem [9]. Each component of the intimidating FID transforms independently and contributes a separate, easily discernible peak to the frequency-domain spectrum. For ideal liquid samples, all the peaks will have Lorentzian shapes, but that does not mean that all the line *widths* are necessarily the same. As shown by the Fourier pairs in Figure 70.10, the width of a Lorentzian peak is an inverse function of its decay time constant τ. So, if the various spins in the sample produce time-domain signals that decay at different rates, their spectral line widths will vary. That can easily be caused by differing relaxation rates, as discussed later.

70.7 NMR INSTRUMENTATION

Despite the impressive appearance and expense of a typical NMR spectrometer, the hardware design concept is relatively straightforward (Fig. 70.11). It begins with a high-quality RF transmitter (usually a computer-controlled digital synthesizer). The transmitter is "on" continuously. Its signal of, typically, a few tenths of a volt is applied to the input of a computer-controlled RF switch. That device creates the RF pulse by turning on, passing RF energy for the desired number of microseconds, and then turning off again. The resulting small RF pulse then enters an RF power amplifier and emerges with an amplitude of, typically, tens of volts. (Recall that high RF power is needed so the sample's bulk magnetization can be nutated even when using a very short pulse.) The amplified RF pulse then passes through series crossed diodes that have essentially no effect on it. The diodes are needed because RF power amplifiers tend to produce unwanted noise when no pulse is present, and since the diodes block signals of less than about 0.5V, they prevent the amplifier noise from contaminating the tiny NMR signal when it is detected later. After passing through the diodes, the RF pulse is applied to the

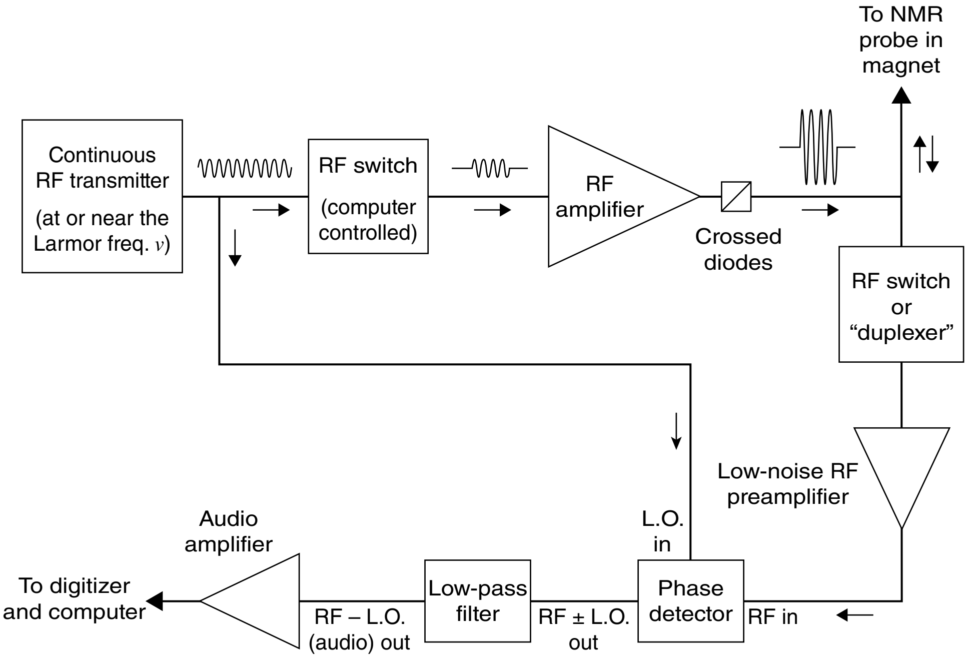


FIGURE 70.11 Block diagram of a pulsed Fourier transform NMR spectrometer operating at the Larmor frequency *ν*. This conceptual diagram shows the major features but excludes some practical details, such as additional tuned filters at various stages that help to suppress noise.

NMR probe, where it excites the sample. At that point, the excitation part of the experiment has been accomplished and it only remains to detect the sample's resulting NMR signal. That tiny (microvolts) signal returns from the probe and proceeds downward in the diagram to an active RF switch or, equivalently, a passive "duplexer." That device is not fundamental to the principle of the measurement, but it plays a crucial practical role: it is turned "off" during the high-power RF pulse to prevent the pulse from reaching the extremely sensitive RF preamplifier and destroying it. Once the pulse ends, this switch is closed again, providing a path for the sample's NMR signal to reach the low-noise preamplifier. The role of the preamplifier is to amplify the NMR signal voltage, typically by 1000-fold or more, while contributing little additional noise itself. The boosted NMR signal is then applied to the input of a "phase detector." This device is "borrowed" from superheterodyne radio technology. It combines the NMR RF signal (A_1 cos(2 π ν_1 t)) with a local oscillator or "L.O." signal (A_2 cos(2 π ν_2 t)) that comes directly from the original transmitter. The output of the phase detector can be considered to be the product of the two input signals (A_1 A_2 cos(2 π ν_1 t) · cos(2 π ν_2 t)), which, due to a mathematical identity, is equivalent to the sum and difference of their frequencies:

$$\text{output signal} = \left\{\frac{1}{2}\ A_1A_2\cos\left[2\ \pi\left(\nu_1+\nu_2\right)t\right]\right\}+\left\{\frac{1}{2}\ A_1A_2\cos\left[2\ \pi\left(\nu_1+\nu_2\right)t\right]\right\}.$$

For example, if the transmitter produces 500.000 MHz and the actual NMR signal has a Larmor frequency of 500.003 MHz, the output of the phase detector will be an interferogram containing signals of 1000.003 MHz (the sum) and 0.003 MHz (the difference). By then passing this through a low-pass audio filter, the high-frequency component is blocked and only the 0.003 MHz (or 3 kHz) signal reaches the remaining amplifier, digitizer, and computer. Why do this instead of digitizing the NMR signal directly? The reason is a practical one. NMR spectra nearly always occupy only a tiny fraction of the whole frequency range. Normal ^{1}H NMR spectra, for example, are fully contained within a 20 ppm range of frequencies, so in our example, all the spectral information of interest would occupy about a 0.010 MHz range of frequencies near 500 MHz (e.g., between 500.000 and 500.010 MHz). By removing the frequencies below 500 MHz, the phase detector/low-pass filter effectively expands just the small range that contains the NMR information. This low-frequency (audio) signal is easy and cheap to digitize. A possible alternative would be to eliminate the phase detector/low-pass filter and just use an extremely high-speed digitizer operating at over 1000 megasamples per second to characterize all frequencies from 0 to 500 MHz or more. However, since the experiment excites no observable NMR signals below the ^{1}H range near 500 MHz, no interesting or useful data exist below that frequency and the capabilities of the expensive fast digitizer would be wasted. The phase detector approach is far more practical.

NMR probes contain tuned circuits to sense the precession of the sample's tiny bulk magnetization vector. There is a common misconception that the NMR signal consists of RF photons emitted or "released" by relaxing molecules after they have been excited, in analogy with spectroscopies at higher frequencies such as in the ultraviolet, visible, and infrared regions of the electromagnetic spectrum. By contrast, the NMR signal is detected as an oscillating voltage that the bulk magnetization vector induces in the receiver coil that surrounds the sample. The coil essentially acts as an RF antenna to detect the sample's oscillating magnetic field. Since that field is exceedingly small, the coil is incorporated into a resonance circuit tuned to oscillate at or near the Larmor frequency of the sample (Fig. 70.12). The probe circuit also contains a matching capacitor that matches the extremely high input impedance of the parallel-tuned resonance circuit to the relatively low (e.g., 50 Ω) impedance of the rest of the spectrometer electronics. Impedance mismatches cause RF signals to reflect, much as water waves in a swimming pool reflect from the walls of the pool. Consequently, without the matching capacitor, little RF energy from the power amplifier would enter the resonance circuit and the sample would not be excited. A mismatch would also prevent the small NMR signal from escaping the circuit to reach the preamplifier. The physical shape of the transmitter/receiver coil varies depending on the configuration of the NMR probe. The key is that the B_1 field produced by the coil must be orthogonal to the B_0 field produced by the main magnetic field. For magnets that produce horizontal fields, such as low-field permanent magnets, a vertical solenoid makes an excellent RF coil. For superconducting magnets, which normally produce vertical B_0 fields, a "saddle" RF coil design is normally used since it produces a horizontal B_1 field.

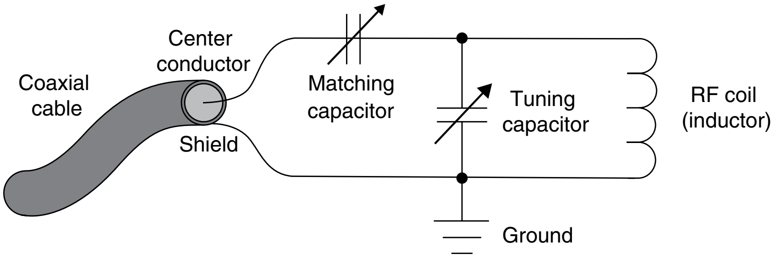


FIGURE 70.12 A parallel-tuned resonance circuit for NMR. The RF coil surrounds the sample and detects the oscillating magnetic field produced by the sample's net magnetization vector. A tuning capacitor connected in parallel with the coil forms a resonance circuit with an oscillation frequency at or near the Larmor frequency of the sample. A separate matching capacitor connected in series matches the input impedance of the parallel-tuned circuit to the, typically, 50 Ω impedance of other spectrometer components such as the coaxial cable. Without this matching capacitor, the impedance mismatch would prevent RF energy from entering the resonance circuit to excite the sample, and the tiny NMR signal would not escape the circuit to be amplified.

The magnetic field B_0 is crucial to the NMR experiment, and it is important to ensure that field is both as intense and as uniform as practical. It is these requirements that make the magnet the most expensive part of a typical NMR spectrometer. Modern superconducting magnets are designed as shown in Figure 70.13. The actual magnet is only a small part of the overall assembly, being enclosed within a Dewar of liquid helium at 4.55 K, which, in turn, is inside a large outer Dewar of liquid nitrogen at 77.4 K. The magnet must be cooled to liquid helium temperatures since a superconducting wire alloy is used to make the magnet. Unfortunately, liquid helium is expensive and sometimes difficult to get at all, so the helium Dewar is immersed in the liquid nitrogen so that small heat leaks will tend to boil away the relatively inexpensive nitrogen instead of the helium. During initial installation, the magnet is energized but then the power supply is removed and the current continues indefinitely inside the wire. The only time reenergization should be required is in the case of a magnet "quench," where the wire suddenly becomes resistive and the current drops precipitously, usually boiling off the cryogens in a spectacular, and occasionally dangerous, fashion. A quench can damage a magnet permanently, but designs have improved over the years and they now usually survive quenches successfully.

All magnets are prone to drift slightly. Consequently, NMR samples are usually prepared in deuterated solvents, and a resonance circuit built into the probe continuously monitors the ^{2}H NMR signal to detect any change in its resonance frequency. Circuitry then compensates for the drift to avoid smearing the NMR signals. This is referred to as "field-frequency" locking.

While locking is generally a robust technique, the deuterated solvent must be chosen based on the analyte's chemical compatibility and solubility. Expense is another consideration, as fully deuterated solvents can easily cost $1 per gram, especially if

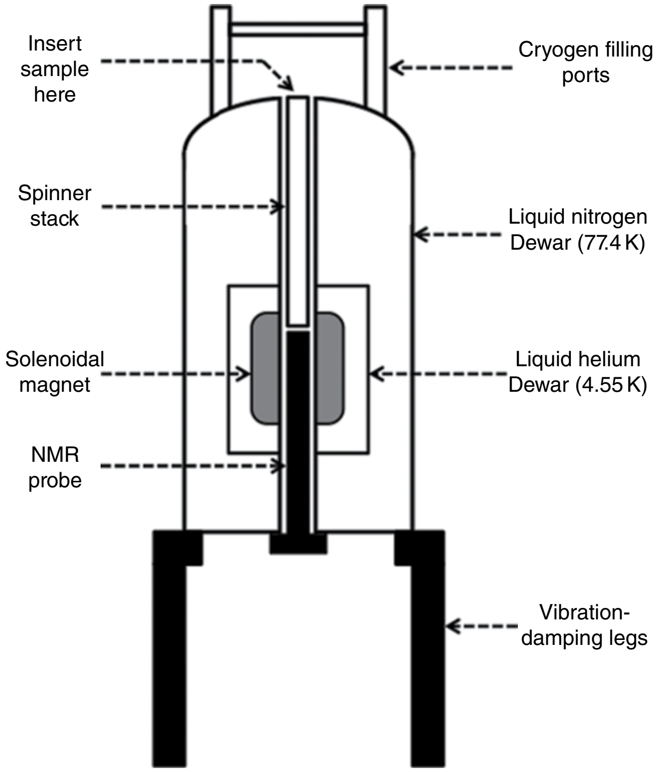


FIGURE 70.13 Cross-sectional diagram of an NMR magnet. Most of the volume is consumed by the liquid nitrogen Dewar. The liquid helium Dewar is contained inside, with the actual superconducting magnet located inside that. Filling ports for the cryogens are located on top of the magnet. The NMR probe inserts from the bottom, and the sample is placed in the top. The entire assembly is supported by vibration-damping legs to prevent building vibrations from causing side bands in the spectra.

high deuterium enrichment (e.g., 99.9%) is needed. The most common solvents are deuterium oxide (D_2O) for polar analytes, chloroform-d (i.e., $CDCl_3$) for nonpolar analytes, and perdeuterated dimethyl sulfoxide (DMSO-d_6), which is a surprisingly good solvent both polar and nonpolar analytes. Since absolutely complete deuteration is not possible, even the best available solvents exhibit small peaks due to residual protonated molecules. Table 70.2 lists several common deuterated solvents, along with the positions and multiplicities of their ^{1}H peaks due to incomplete deuteration. In D_2O, it is unlikely that any given water molecule contains more than one ^{1}H, so its residual protonated water peak is referred to as an "HOD" peak. Table 70.2 also lists the solvents' ^{13}C chemical shifts and multiplicities. Unlike the typically small residual ^{1}H peaks, the ^{13}C peaks are relatively intense since samples contain mostly solvent and the solvent contains full natural-abundance ^{13}C, just like the dissolved analytes. ^{13}C-depleted solvents can be purchased for special work, although they are quite

TABLE 70.2 NMR Properties of Common Deuterated Solvents[a]

Solvent	Formula	Cost	^{1}H Chem. Shift	^{1}H Multiplicity	^{13}C Chem. Shift	^{13}C Multiplicity
Acetone-d_6	$(CD_3)_2CO$	$$$	2.06	5	29.9 206.7	7 1
Benzene-d_6	C_6D_6	$$$	7.16	1	128.4	3
Chloroform-d	$CDCl_3$	$	7.24	1	77.2	3
Deuterium oxide	D_2O	$$	4.80	1	—	—
Dimethyl sulfoxide-d_6	$(CD_3)_2SO$	$$$	2.50	5	39.5	7
Methylene chloride-d_2	CD_2Cl_2	$$	5.32	3	54.0	5

[a] Chemical shifts are in ppm relative to TMS or TSP at 0 ppm. Multiplicities indicate the number of peaks present in the solvent's resonance.

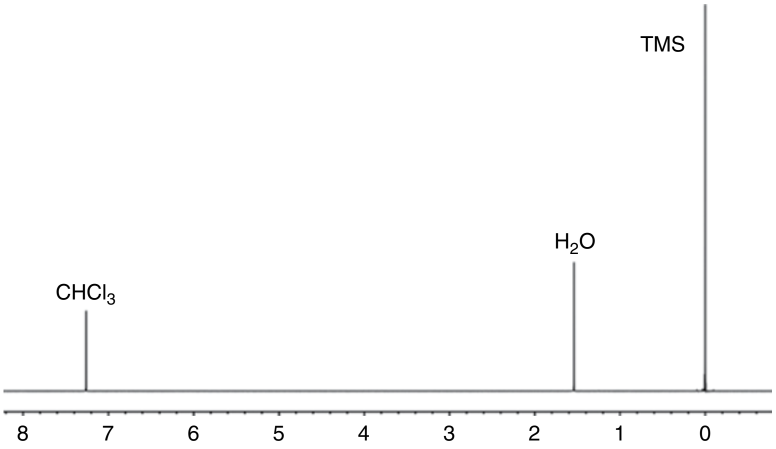


FIGURE 70.14 Proton NMR spectrum of deuterochloroform ($CDCl_3$) solvent containing 0.1% TMS. Parameters: 599.7 MHz; 25°C; 45° tip; 1 average; 20 Hz spin rate. Residual protonated chloroform is visible near 7.24 ppm, and a small quantity of contaminating moisture produces the peak near 1.5 ppm.

expensive. One additional problem associated with NMR solvents is the presence of moisture. Even relatively nonpolar solvents like $CDCl_3$ can contain enough H_2O to produce a significant peak, as illustrated in Figure 70.14. To make matters worse, the chemical shift of the water peak can vary enormously depending on the amount of hydrogen bonding present. As shown in Table 70.2, the residual HOD peak in D_2O appears near 4.8 ppm, but the peak shifts far upfield (1–3 ppm) in solvents that interfere with hydrogen bonding. That can confound structure assignments and may require running a solvent blank to determine exactly which peaks belong to the analyte. Even that can fail, however, since the analyte itself may be the main source of the moisture.

Extraordinarily high-quality NMR magnets are now available commercially. Not only are their magnetic fields intense but, nearly as important, their fields are extremely homogeneous (i.e., uniform). That homogeneity is crucial to achieving the best spectral line shapes and resolution. The reason is easy to understand by making a small modification to the Larmor equation:

$$\Delta \nu_0 = \gamma \, \Delta B_0, \tag{70.8}$$

where ΔB_0 represents a range of magnetic fields over the sample volume and $\Delta \nu_0$ is the resulting range of resonance frequencies. So, variations in the magnetic field tend to smear the NMR lines over a range of frequencies, producing deleterious broadening. For optimum results, the field should vary by no more than about one part per billion over the sample volume. That would correspond, for example, to field-induced line broadening of 0.6 Hz on a 600 MHz spectrometer. Such superb homogeneity cannot be obtained from the basic magnet alone, but the residual field nonuniformities can be corrected by using "shim coils" built into the inner bore of the magnet. Current is passed through those coils to create additional small fields that "shape" B_0 and make it as uniform as possible. Every sample should be "shimmed" individually before its spectrum is acquired, in part because the sample itself can induce field inhomogeneities, particularly at interfaces such as the upper surface where the sample liquid meets air above it. Older spectrometers required the investigator to adjust the currents through the shim coils manually while monitoring the FID to produce the slowest possible signal decay and, therefore, narrowest lines (see Fig. 70.10). Experienced operators usually achieved a good shim in several minutes. With modern instruments, the shimming process is fully automated, and the software usually produces an excellent shim in only a minute or two without operator intervention. In addition to shimming, sample spinning is used to obtain narrow peaks. Spinning the tube about its long (z) axis tends to average field inhomogeneity in the transverse (x and y) directions. A spinning rate of about 20 Hz is usually sufficient, and the improvement obtained can be dramatic (Fig. 70.15). Small spinning side bands are often be observable at the spinning rate or integer multiples of it, as shown in the figure. These can easily be distinguished from other, more meaningful small peaks because (1) they appear on every peak in the spectrum, (2) their positions change if the spinning rate is varied, and (3) they disappear altogether if the spinning is stopped. To keep the spinning side band amplitudes low, it is important to use both the best practical shim and the highest quality NMR sample tubes. Tubes are marketed by several manufacturers and must conform to rigid specifications for roundness, concentricity, and camber (Fig. 70.16). Low-quality tubes may suffice for undemanding applications at low magnetic fields, for example, in nonspinning studies at 60 MHz, but high-field magnets require quality sample tubes that can be expensive. For most spectrometers, 5 mm outer diameter sample tubes are used, but "wide-bore" instruments allow 10 and even 20 mm tubes to fit. Achieving an excellent shim is often more challenging with large-diameter tubes since the homogeneity must be optimized over a

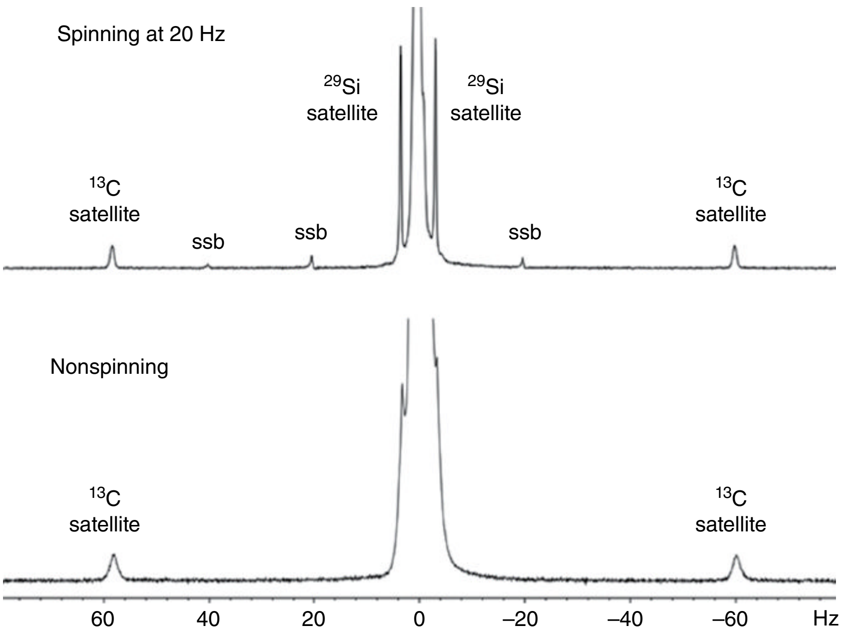


FIGURE 70.15 Expanded proton spectra of 0.1% (v/v) tetramethylsilane (TMS) in $CDCl_3$. Acquisition parameters: 599.7 MHz; 25°C; 45° tip; 27.2 s total cycle time; 32 averages. The displayed spectral width is about 0.27 ppm. In addition to satellite peaks due to ^{29}Si- and ^{13}C-containing molecules, the spinning spectrum shows multiple spinning side bands (ssb) that are separated from the central peak by the spinning rate (20 Hz) or integer multiples of it. Turning off the spinning (bottom spectrum) eliminates the spinning side bands but degrades the NMR line widths.

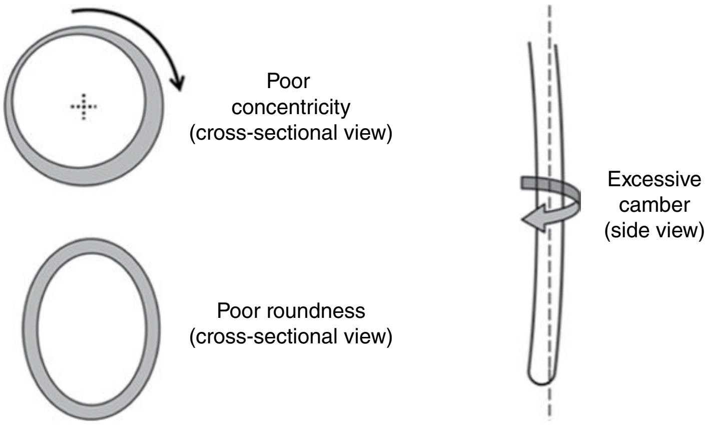

Poor concentricity (cross-sectional view)
Excessive camber (side view)
Poor roundness (cross-sectional view)

FIGURE 70.16 NMR sample tube characteristics. Concentricity indicates how well the inner and outer surfaces of the tube wall are centered relative to each other. Roundness is a measure of how closely the tube's cross section conforms to a perfect circle. Camber reflects the tube's deviation from straightness. The concentricity, roundness, and camber shown in this figure are all dismal. Spinning a sample inside such a tube would cause a periodic modulation of the NMR signal, resulting in intense spinning side bands that flank every peak in the spectrum. Excellent sample tube quality is important for all NMR applications but is especially crucial for studies in high magnetic fields. The best tubes marketed for use at or above 600 MHz can cost over $30 each.

greater volume. Although sample spinning and a quality shim are the best ways to obtain good NMR line shapes and resolution, major improvements are also possible after the data have already been collected by using the "reference deconvolution" technique [10, 11]. The essence of the technique is to isolate an intense, well-resolved single peak in the spectrum (often the solvent) and then to deconvolve its shape from the rest of the spectrum. Since B_0 inhomogeneity affects the shapes of all the peaks in the same manner, deconvolving should ideally remove the effects and leave perfectly Lorentzian and narrow lines throughout the spectrum. The method can work remarkably well and, in many cases, without degrading the signal/noise ratio very greatly (or even at all!).

70.8 THE BASIC PULSED FTNMR EXPERIMENT

The great majority of NMR studies are performed on liquids or dissolved samples using a simple "1-pulse" sequence (also known as "pulse and collect"). The sequence is illustrated in Figure 70.17. It begins with the application of a short (microseconds) rectangular pulse of RF energy to the sample, which tips the bulk magnetization vector away from the *z*-axis. Unfortunately, the RF pulse also saturates the sensitive preamplifier in the spectrometer's receiver, which responds by emitting wild signal excursions for, typically, 1–20 μs. To avoid acquiring that meaningless signal, a short receiver recovery delay is inserted immediately following the pulse. Then the computer's digitizer is turned on and the FID is acquired as a time series of points. Finally, a relaxation delay is used to give the sample time to relax before the next pulse is applied.

The sequence is usually repeated multiple (*n*) times, and the resulting *n* FIDs are averaged to improve the signal/noise ratio. Signal averaging is a routine operation in

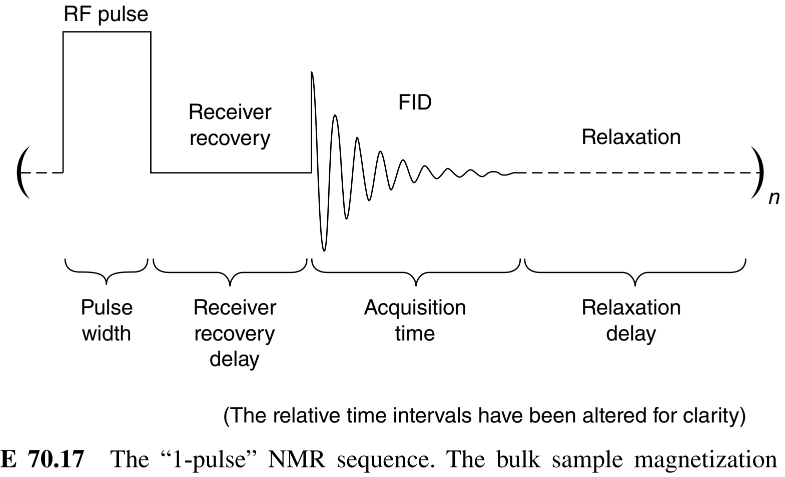


FIGURE 70.17 The "1-pulse" NMR sequence. The bulk sample magnetization vector is tipped away from the *z*-axis by a short (microseconds) rectangular pulse of RF energy. Following a brief recovery delay, the FID is acquired. The spin system is then allowed to relax during the relaxation delay. The sequence is repeated *n* times, recording the *n* FIDs and averaging them.

NMR since the basic signal is nearly always weak and noisy. When n independent FIDs are summed, the signal adds linearly as $n \cdot S_1$, where S_1 is the amplitude of the signal obtained in a single scan. The noise also adds, but only in proportion to $\sqrt{n} \cdot N_1$, where N_1 is the time-domain noise amplitude in a single scan. Consequently, the ratio of signal/noise improves in proportion to the square root of the number of averaged scans:

$$\left(\frac{S}{N}\right)_n = \frac{n \cdot S_1}{\sqrt{n} \cdot N_1} = \sqrt{n} \cdot \left(\frac{S}{N}\right)_1 . \tag{70.9}$$

Averaging 4 FIDs doubles the *S*/*N*. Clearly, time averaging can easily reach a point of diminishing returns. For example, if 5 s are needed to acquire a single FID, doubling the *S*/*N* requires 20 s, which is usually very acceptable. However, to improve the *S*/*N* by 10-fold, 100 scans and 500 s (8.3 min) are needed. It will be necessary to average 10,000 scans to improve the *S*/*N* by 100-fold, requiring 50,000 s (833 min)! That would be prohibitive in all but the most dire cases. Alternatives to signal averaging could include isotopic enrichment, increasing the sample concentration, or using a larger-diameter sample tube to place more sample in the NMR coil (if the probe allows). A larger tube increases the *S*/*N*, but generally not in direct proportion to the increased sample volume since (1) NMR coils lose efficiency as their diameters increase and (2) the sample itself becomes a more significant source of noise.

Once the averaged FID is obtained, it is Fourier transformed to produce the one-dimensional (1D) NMR spectrum. That spectrum must normally be "phased" to produce pure absorption-mode peaks (the real FT of Fig. 70.10) instead of dispersive peaks (the imaginary FT of Fig. 70.10) or some mixture of the two modes. Phasing is largely automatic on modern spectrometers, although a little manual touch-up is sometimes needed.

Prior to acquiring the NMR spectrum as in the earlier text, it is often necessary to determine the RF pulse width that corresponds to a 90° (or other) tip angle. In that case, a set of spectra is acquired by using a different, incremented pulse width for each. Following FT and phasing, it is possible to determine the amount of nutation caused by each of the pulse widths (Fig. 70.18). It is common practice to search for the pulse width that yields a 360° nutation, since its "null" signal is superior to that produced by a 180° pulse if the sample extends outside the RF coil where the B_1 field is very inhomogeneous. Once the 360° pulse width has been found, dividing it by four yields the 90° pulse width.

70.9 CHARACTERISTICS OF NMR SPECTRA

NMR spectra are often discussed by using special terms, as illustrated in Figure 70.19. The rationale for some of these is purely historical and not rooted in modern practice, but the terms continue to be used anyway. One important modern convention is that the frequency increases from right to left, just as in other types of spectroscopy.

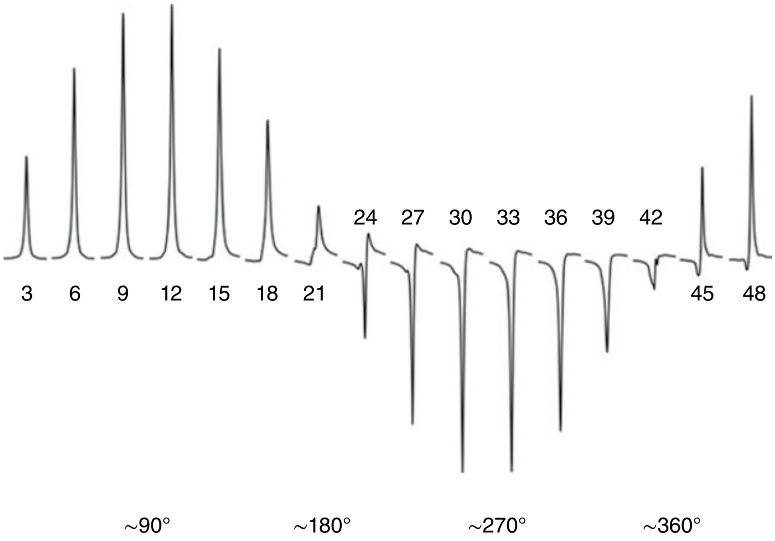


FIGURE 70.18 Proton RF pulse width calibration data for the H_2O peak in a $CDCl_3$ solvent. Acquisition parameters: 599.7 MHz; 25°C; 42 s total cycle time; 2 averages; 20 Hz spin rate. Each peak is labeled with the pulse width (in μs) used to produce it. Approximate tip angles are indicated along the bottom of the figure. A 43 μs pulse produces close to a 360° tip, meaning the 90° pulse width is about 43/4 ≈ 10.8 μs.

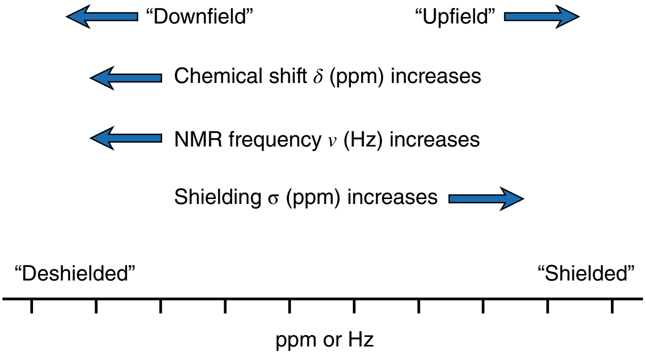

"Downfield"
"Upfield"
Chemical shift δ (ppm) increases
NMR frequency ν (Hz) increases
Shielding σ (ppm) increases
"Deshielded"
"Shielded"
ppm or Hz

FIGURE 70.19 Terminology conventions in NMR spectroscopy. All common chemical shift references (TMS, TSP, and DSS) produce signals at 0 ppm, significantly upfield from nearly all other ^{1}H and ^{13}C NMR lines.

70.9.1 The Chemical Shift

The most important parameter in NMR spectroscopy is the chemical shift. The horizontal scale of the spectrum actually represents frequency but is almost universally labeled chemical shift with units of parts per million (ppm). Zero on the scale is assigned based, ideally, on a standard compound actually dissolved in the sample. For ^{1}H and ^{13}C NMR, three different compounds are widely used as standards (Fig. 70.20).

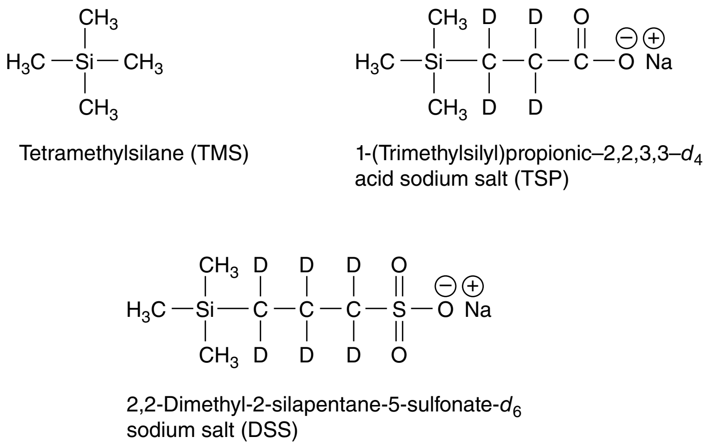

Tetramethylsilane (TMS)
1-(Trimethylsilyl)propionic–2,2,3,3–d4 acid sodium salt (TSP)
2,2-Dimethyl-2-silapentane-5-sulfonate-d6 sodium salt (DSS)

FIGURE 70.20 Chemical structures of the three most common chemical shift reference compounds for ^{1}H and ^{13}C NMR. TMS is the ultimate reference standard, but its poor solubility in polar solvents requires the use of the ionic salts TSP or DSS instead. All three compounds exhibit a ^{1}H reference peak at 0.00 ppm.

TMS is the ultimate reference. It is quite nonpolar and dissolves readily in relatively nonpolar NMR solvents such as $CDCl_3$ and DMSO-d_6. It also has the advantage of a low boiling point (27°C), which facilitates recovering the analyte from the sample without contamination. For polar solvents, low solubility precludes the use of TMS, and ionic salts are employed instead. TSP is the most common, though DSS is also widely used. All three compounds produce ^{1}H peaks at virtually the same chemical shift and may be assigned values of 0.00 ppm for practical purposes. The pure compounds can be purchased commercially, but deuterated solvents can easily be purchased that already contain 0.1 or 1% of them. When an internal standard cannot be used, due perhaps to a fear of contaminating an important sample, it is possible to fall back on the solvent as a secondary chemical shift reference. Table 70.2 gives the ^{1}H and ^{13}C chemical shifts of typical solvent peaks relative to TMS, TSP, or DSS. Another alternative is to place the standard inside a glass capillary inserted into the sample, but that is less desirable since (1) the shim is usually degraded and (2) magnetic susceptibility differences between the sample and capillary can easily shift the reference peak by 0.05 ppm or more.

The chemical shift of a nuclide in a molecule can be influenced by several factors (Fig. 70.21). Electrons shield the nuclei from the main B_0 field according to a modified Larmor equation:

$$\nu_0 = \gamma\left(1-\sigma\right)B_0, \tag{70.10}$$

where σ is called the shielding constant. High electron density about a nucleus makes its shielding constant large and that, in turn, produces an upfield shift to low frequency and a small chemical shift value. Any electronegative atom X present in the molecule

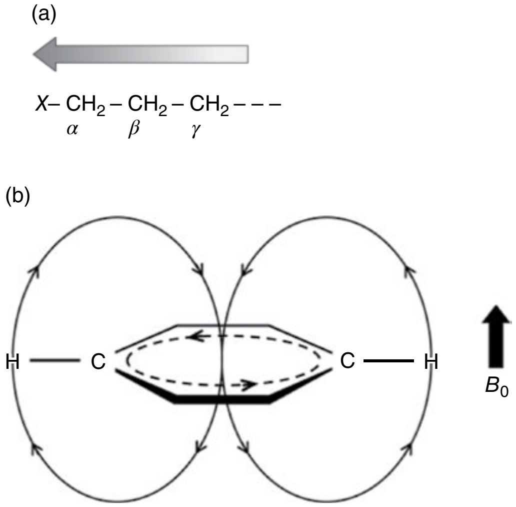


FIGURE 70.21 Some causes of chemical shifts. (a) When relatively electronegative elements *X*, such as fluorine or oxygen, are covalently bonded in the molecule, they withdraw electron density from nearby atoms. These deshielded atoms produce resonance lines that are shifted downfield (to high frequencies) in the NMR spectrum. (b) Side view of the planar ring of an aromatic compound, such as benzene (C_6H_6), that contains delocalized electrons in π molecular orbitals. In the presence of an external magnetic field B_0, these electrons circulate (dotted line), creating a small opposing magnetic field in the center of the ring. By tracing the solid lines of flux, it can be seen that this field actually reinforces B_0 at the positions of the molecule's hydrogen atoms in the periphery. To satisfy the Larmor equation, the resonance frequency must increase in response to this slightly higher field, producing a downfield shift for the attached ^{1}H.

tends to withdraw electron density via the inductive effect, leading to downfield (high-frequency) shifts for atoms one or two bonds away. The effect is evident in the chemical shifts of, for example, alcohols *versus* aliphatic compounds shown in Table 70.3. In special cases, more subtle chemical shift effects can occur. For example, an iodine atom, which might be expected to withdraw a little electron density from a molecule, often does just the opposite due to its high polarizability and causes an upfield, instead of downfield, shift. Anionic substances in solution often show relatively large upfield chemical shifts due to shielding by their extra electron(s). For example, the ^{1}H chemical shift of the four equivalent protons of borohydride ion (BH_4^-) coincides almost exactly with that of TMS. Cations can show just the opposite effect. Elegant studies of the distribution of electric charge along hydrocarbon chains have been performed by measuring changes in NMR shielding [12].

Chemical shifts are normally expressed in ppm from the reference line of TMS, TSP, etc. An important ramification is that the separation *in ppm* between two peaks with different chemical shifts should be constant even when measured with spectrometers operating at different magnetic fields. However, the separation *in Hz* will differ. For example, suppose the chemical shift difference between two proton peaks is

TABLE 70.3 Typical Chemical Shift Ranges for ^{1}H and ^{13}C Relative to TMS

	Approximate Chemical Shift (ppm)	
Functional Group	^{1}H	^{13}C
CH (aliphatic, alicyclic)	0.0–2.0	0–55
CH (alkyne)	2.0–3.0	60–90
XCCH (α-monosubstituted aliphatic)	2.0–5.0	25–90
X_2CCH (α,α-disubstituted aliphatic)	2.4–7.0	25–90
COH (alcohols, water)	0.5–6.3 (4–6 typ.)	60–80
CNH_2 (amine)	1.7–5.0	25–55
$-CONH_2$ (amide)	5.0–8.7	150–185
CH (alkene)	4.3–7.5	110–150
CH (aromatic, heteroaromatic)	6.0–9.1	90–160
CH (aldehyde)	8.9–10.2	175–220
COOH (carboxylic acid)	10.0–13.2 (when dimerized)	150–185

2.0 ppm. In a B_0 field of 9.4 T, corresponding to 400 MHz for ^{1}H, there is 400 Hz per ppm so the separation between the peaks is 2 × 400 Hz or 800 Hz. In a field of 14.1 T, the proton resonance frequency is 600 MHz and 2 ppm corresponds to 2 × 600 Hz or 1200 Hz. So, a higher magnetic field puts more "distance" between the peaks, in Hz, even though the separation is still 2 ppm. At higher fields, the spectrum will contain more space for peaks, decreasing the likelihood of peak crowding and overlap. This is one major advantage of using strong magnets in NMR.

It is clear from Table 70.3 that aromatic protons are shifted rather far downfield, even in the absence of any significantly electronegative atoms. For example, the ^{1}H chemical shift of the six equivalent protons in benzene is about 7.2 ppm, compared to simple aliphatic protons that normally appear in the 0–2 ppm region. A different mechanism is responsible for this effect: ring currents. Aromatic compounds contain loosely bound electrons in π molecular orbitals that are free to circulate in response to the B_0 field. This electron circulation sets up an extra magnetic field that detracts slightly from the main B_0 field in the center of the aromatic ring but reinforces it in peripheral regions outside the ring. Thus, protons attached around the ring experience a slightly increased B_0 field and resonate at a higher frequency. Elegant confirmation of this mechanism is found in the chemical shifts for the aromatic compound [18]-annulene, which contains protons both on the inside and outside of the ring (Fig. 70.22). At −60°C, distinct [18]-annulene proton peaks are observed at −3.0 and +9.3 ppm [13]. The −3.0 ppm peak is due to the interior protons that experience the reduced magnetic field, and the peak at +9.3 ppm arises from outer protons that are exposed to a stronger field and resonate at a higher frequency. Other large, cyclic aromatic compounds such as porphyrins often exhibit similar effects. The presence or absence of substantial ring current is generally regarded as a reliable test for aromaticity in newly synthesized compounds [14].

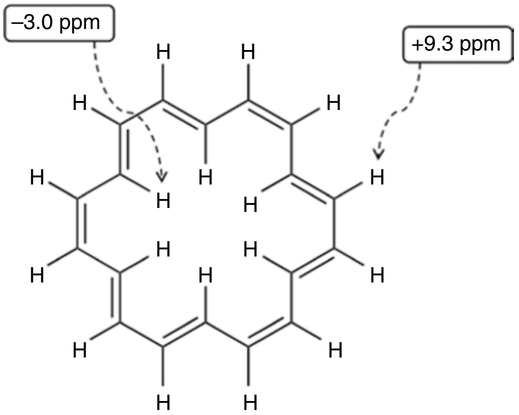


FIGURE 70.22 The chemical structure of [18]-annulene. The six protons in the center of the aromatic ring inhabit a region of decreased magnetic field due to electron ring currents and produce a ^{1}H peak at −3.0 ppm relative to TMS. The 12 exterior protons experience an enhanced field and produce a peak at +9.3 ppm near the traditional aromatic chemical shift range of the proton NMR spectrum.

In many alcohols and carboxylic acids, hydrogen bonding can profoundly affect the chemical shifts of —OH peaks. Strong hydrogen bonding shifts the peaks downfield, while reduced hydrogen bonding results in an upfield shift. This is reflected in the wide range of values shown in Table 70.3 for —OH species and in the upfield shift for water in $CDCl_3$ (Fig. 70.14). Compounds that form especially strong hydrogen bonds, such as carboxylic acid intermolecular dimers, exhibit large shifts to low field, as shown by their roughly 10–13 ppm chemical shift. An even more impressive case occurs in compounds like enols that can form especially stable, intramolecular hydrogen bonds in six-membered rings. One such compound, the enol tautomer of 2,4-pentanedione (Fig. 70.23) exhibits an —OH peak shifted nearly to +15.3 ppm, beyond the range for any "normal" proton. NMR peaks caused by —OH protons must be interpreted with care since, depending on the solvent and molecular structure, they can appear at nearly any position in the spectrum.

Table 70.3 includes chemical shift information for ^{13}C along with ^{1}H. The shift trends for ^{13}C roughly parallel those of ^{1}H, but the range is over ten times bigger. Since ^{13}C is surrounded by more electrons than is ^{1}H, there is greater potential for changing the electron density near ^{13}C and that leads to a wider chemical shift range. The effect is even bigger for heavier nuclides. For example, ^{59}Co (atomic number 27) has a known chemical shift range of over 18,000 ppm! Its shift is so sensitive that separate peaks appear for the (+) and (−) enantiomers of tris(ethylenediamine)cobalt(III) ion in solution when they are merely ion paired with optically active tartaric acid [15]. Nearly every element in the periodic table has at least one NMR-active nuclide, and virtually all of those have been studied at one time or another. Extremely useful chemical shift compilations have been published for the less common NMR nuclides [16, 17].

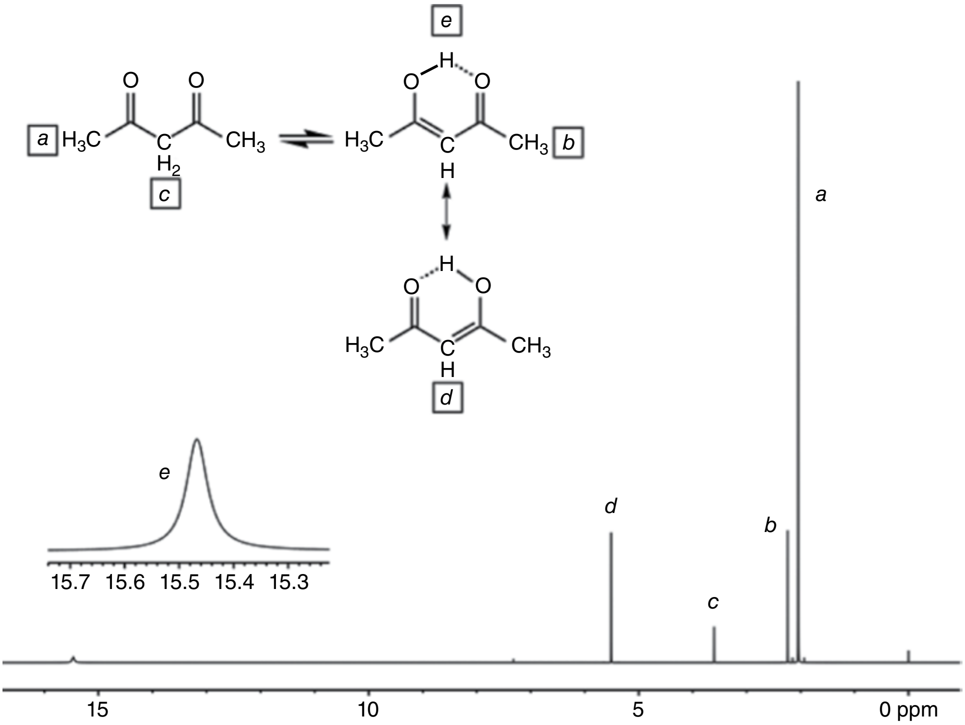


FIGURE 70.23 Proton NMR spectrum of 10% (v/v) 2,4-pentanedione in $CDCl_3$ containing 0.1% TMS. Acquisition parameters: 599.7 MHz; 25°C; 30° tip; 30.0 s total cycle time; 8 averages; nonspinning. Tautomerization yields an equilibrium mixture of the keto form with a lesser quantity of the enol. Exchange between the two forms is slow on the "NMR time scale," allowing distinct resonance peaks to be observed for each (assignments shown). Resonance stabilization of the 6-member ring in the enol form produces an unusually strong hydrogen bond and a corresponding large downfield shift for the OH proton resonance. The tiny peak near 7.24 ppm arises from $CHCl_3$ contamination in the solvent.

Many references provide chemical shift data for standard nuclides like ^1H and ^{13}C. Early compilations [18, 19] remain extremely useful, but more recent monographs contain updated information [20–22]. In addition, the extensive $^1H/^{13}C$ spectral libraries published by Aldrich [23] can be quite helpful. Even if the specific compound of interest is not included, the library will often provide spectra for related compounds that can assist in making the assignments.

70.9.2 Spin–Spin Coupling

Spin coupling between various protons, carbons, and other NMR-active nuclides can be exploited for assigning spectral lines to individual nuclei and for determining molecular structures by NMR. Coupling in molecules occurs through chemical bonds via the Fermi contact mechanism. The concept is easy to grasp. When a nuclear spin 1 in a molecule adopts its quantized orientation relative to B_0, its tiny magnetic moment polarizes the adjacent bonding electrons. Those, in turn, set up a weak magnetic field in the vicinity of another nearby nucleus 2, causing its resonance frequency to shift slightly. In a different molecule, spin 1 may adopt a different quantized orientation, which changes the shielding on nucleus 2. The overall effect in large ensembles of spins is to split the single resonance line of nucleus 2 into multiple peaks that reflect the possible orientations of nuclear spin 1. That information is conveyed by the electrons in the chemical bonds and is called scalar coupling. The effect is most likely to be strong when the two nuclei are in close proximity and tends to grow weaker with the number of intervening chemical bonds.

The magnitude of the scalar coupling is indicated by the spacing, in Hz, between the peaks in the coupled multiplets, and is designated by the letter J. The strength of scalar coupling is controlled by the bonding details of molecules and *has no dependence on the external* B_0 *field*. Consequently, J values are always expressed in Hz and never in ppm. If measured using a high-field NMR system, a J value will be identical to that found by using lower-field instruments.

In the absence of complicating effects, if two nuclei are coupled such that one exhibits a doublet of peaks, the other will also be a doublet. The two peaks in both doublets will be separated by J Hz, so it is often easy to determine which nuclei are coupled simply by inspecting the spectrum. Coupling between two directly bonded nuclides (such as C—H) is called 1-bond coupling and is symbolized by ${}^1J_{\mathrm{CH}}$, where the superscript indicates the number of chemical bonds that separate the coupled nuclei. Coupling between two vicinal protons bonded as H—C—C—H would be ${}^3J_{\mathrm{HH}}$ since there are three intervening bonds between the two protons. As a general trend, ${}^1J > {}^2J > {}^3J$, and so on since the Fermi contact mechanism loses efficiency with the successive addition of more bonds.

When spin-1/2 nuclei couple, the appearance of the resulting coupling patterns can be predicted nicely by Pascal's triangle in most cases (Fig. 70.24.) For example, the ^{1}H NMR spectrum of TMS (Fig. 70.15) displays doublets due to spin coupling of the observed methyl protons with directly bonded ^{13}C. Since only 1.1% of all carbon is ^{13}C, the doublet appears as the weak "^{13}C satellites" indicated in the figure. ${}^1J_{\mathrm{HC}}$ in this case is 118.0 Hz. Nearer to the center of the TMS ^{1}H line, two clear peaks are visible due to ^{1}H–^{29}Si coupling. The coupling constant is comparatively small (${}^2J_{\mathrm{HSi}} = 6.57$ Hz) since it results from 2-bond coupling. A careful examination of the spectrum also reveals that neither the ^{29}Si nor the ^{13}C doublet is centered exactly on the main peak. That is because the ^{1}H chemical shifts change slightly when ^{13}C is substituted for ^{12}C and ^{29}Si is substituted for ^{28}Si.

The simple coupling patterns predicted in Figure 70.24 are ideal approximations. In practice, spectra conform well to them provided $\Delta\nu \gg J$, where $\Delta\nu$ is the difference in Hz between the chemical shifts of the coupled partners. When the chemical shift difference between two coupled nuclei begins to approach the coupling constant, so-called second-order effects usually occur, distorting the expected peak intensities and spacings and even

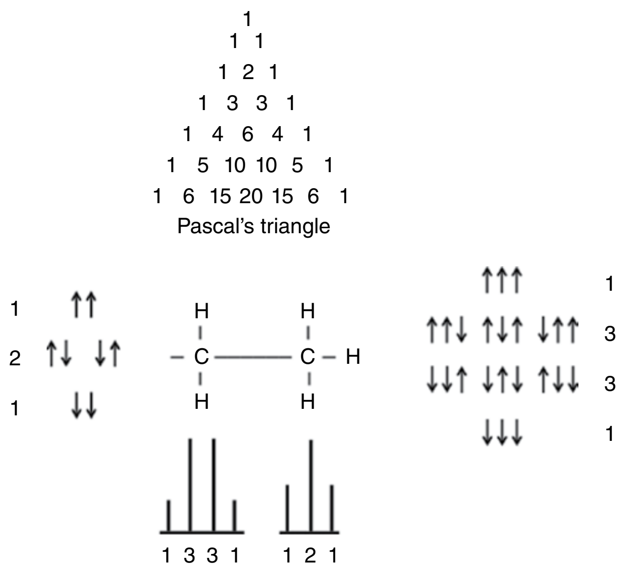


FIGURE 70.24 Predicting the number and intensities of lines in NMR multiplets due to spin–spin coupling between spin-1/2 nuclides. The CH_2 group is split by the three methyl H's, which can take on four different overall spin energies. The middle two energies are three times more likely. So, the CH_2 protons are split into four peaks with intensity ratios of 1 : 3 : 3 : 1. This corresponds to the fourth row of Pascal's triangle. A similar argument accounts for the triplet of 1 : 2 : 1 intensity ratios for the CH_3 protons.

forming extra peaks that are not predicted by first-order theory. An example of small second-order coupling effects is in Figure 70.25, which shows spectra of ethyl *trans*-crotonate acquired at 60 and 600 MHz. The inset in (a) expands the spectrum of vinyl proton e, which is a beautiful, though slightly overlapped, doublet of quartets due to its spin coupling with the three protons of the nearby methyl group and the one other vinyl proton. The spectrum exhibits good symmetry and nearly ideal relative peak amplitudes. By contrast, the downfield quartet in the corresponding 60 MHz spectrum (b) shows markedly reduced intensity due to second-order coupling effects since the chemical shift difference, in Hz, is more similar to J at 60 MHz than it is at 600 MHz. The doublet of quartets at 5.8 ppm that corresponds to proton d also shows intensity distortions for the same reason. The spectra of Figure 70.25 remind us of one more lesson: using a high-field NMR spectrometer provides more "empty" space between the multiplets in a spectrum, making it possible to study more complicated molecules without risking significant peak overlap. Obviously that could be a bigger problem in the 60 MHz spectrum of spectrum (b), despite the fact that all the coupling constants in Hz are the same in both spectra.

Spin coupling patterns are a bit more complicated when the coupled nuclides are not spin 1/2. The general rule for predicting the number of lines (assuming first-order coupling) is

$$\text{number of lines} = 2nI + 1, \qquad (70.11)$$

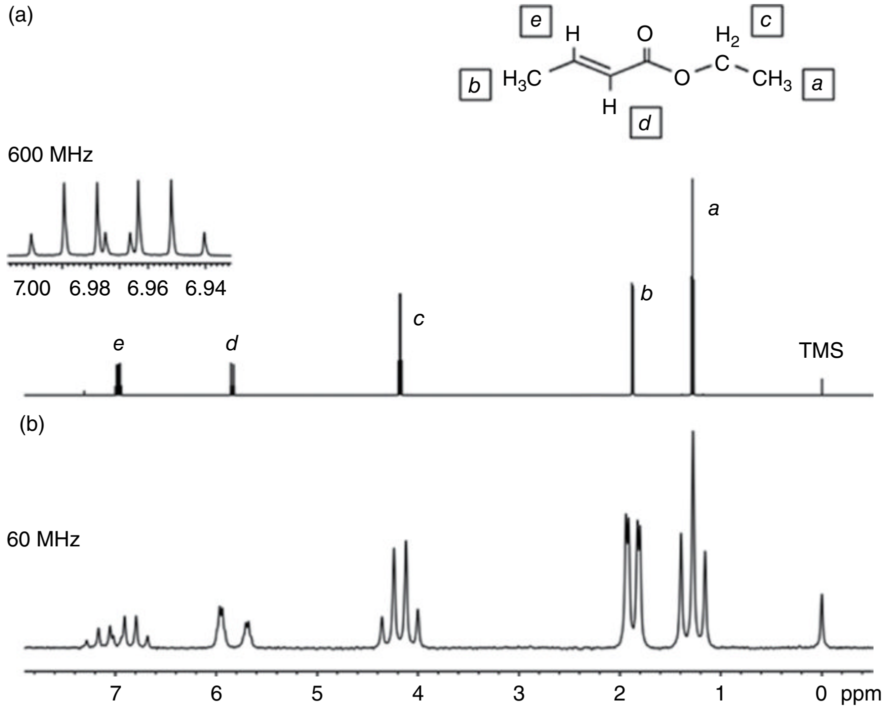


FIGURE 70.25 Proton NMR spectra of 10% (v/v) ethyl *trans*-crotonate in $CDCl_3$ with TMS. Similar acquisition parameters were used for both spectra except that spectrum (a) was measured at 600 MHz while spectrum (b) was acquired at 60 MHz. The inset shows an expansion of the 600 MHz spectrum for the doublet of quartets assigned to proton *e* in the structure. Note the spacing between peaks within each spin-coupled multiplet is the same *in Hz*, regardless of magnetic field strength, but it differs greatly *in ppm*.

where *n* is the number of nuclei to which the observed nuclide is coupled and *I* is the spin quantum number of those nuclei. As a simple example, consider spin-1 nuclides like deuterium. They can take on three different quantized spin states: +1, 0, and −1, all with equal probabilities. Consequently, the ^{13}C NMR peak for $CDCl_3$ is a 1 : 1 : 1 triplet due to coupling of the ^{13}C with the attached deuterium atom. According to the equation, $(2{\cdot}1{\cdot}1)+1=3$ peaks are expected for the ^{13}C. More complicated coupling can give rise to beautiful, elegant spectra in many cases. As illustrated in Figure 70.26, the ^{1}H spectrum of borohydride ion (BH_4^-) exhibits two distinct sets of peaks, a prominent set of four accompanied by a less intense set of seven. Boron exists as approximately 80% ^{11}B and 20% ^{10}B. ^{11}B is spin 3/2, so according to Equation 70.11, $(2{\cdot}1{\cdot}3/2)+1=4$ proton lines should result from the spin coupling. ^{10}B is spin 3, so the equation predicts $(2{\cdot}1{\cdot}3)+1=7$ lines, precisely as found. Careful measurements of the peak areas confirm the four larger peaks combine to produce the expected 80/20 ratio relative to the sum of the seven smaller peaks. The inset at the top of Figure 70.26 shows the ^{11}B NMR spectrum of the same sample, and it contains the 1 : 4 : 6 : 4 : 1 quintet predicted

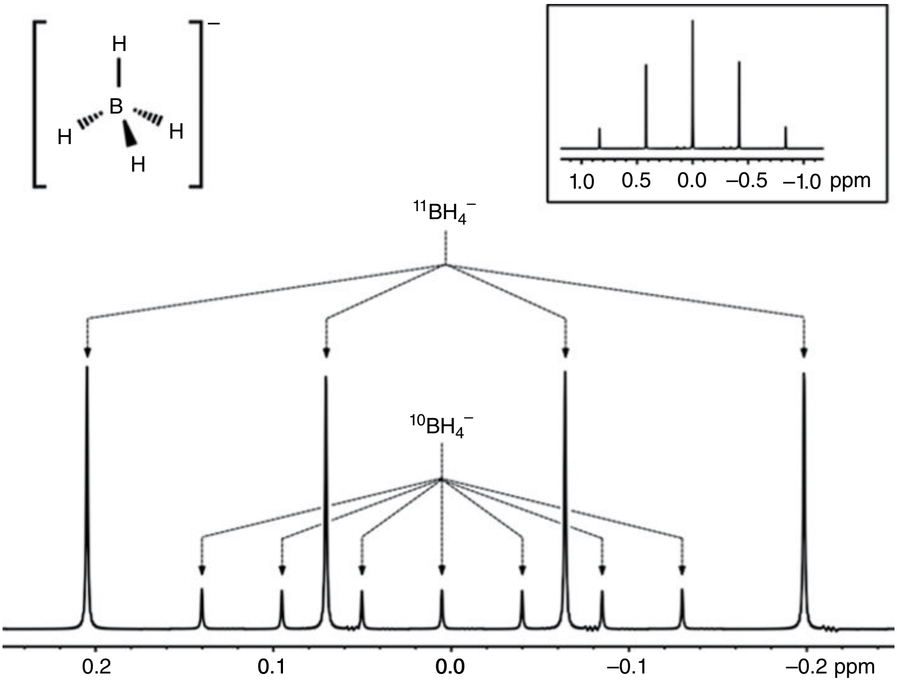


FIGURE 70.26 Proton NMR spectrum of 10% (w/v) sodium borohydride ($NaBH_4$) dissolved in D_2O. ^1H acquisition parameters: 599.7 MHz; 25°C; 30° tip; 26.7 s total cycle time; 4 averages. Chemical shifts are relative to the HOD peak at 4.80 ppm. The borohydride ion is tetrahedral, as shown. Boron exists as 19.58% ^{10}B ($I=3$) and 80.42% ^{11}B ($I=3/2$). Consequently, ^{10}B-containing borohydride splits the proton spectrum into seven lines, and ^{11}B-borohydride exhibits four stronger lines. The boxed inset shows the ^{11}B NMR spectrum of the same sample, with its central line set arbitrarily to 0.0 ppm. The spectrum is a 1 : 4 : 6 : 4 : 1 quintet due to boron coupling with the four equivalent protons, with $^1J_{^1H-^{11}B} = 80.62$ Hz. ^{11}B acquisition parameters: 192.4 MHz; 25°C; 30° tip; 12.4 s total cycle time; 4 averages.

by Pascal's triangle for coupling to the four spin-1/2 protons in the ion. The coupling is completely first order since the chemical shift difference between ^1H and ^{11}B is gigantic (407.3 MHz) compared to the $^1J_{HB}$ value of 80.62 Hz. The ^{10}B NMR spectrum is not shown, but it resembles the ^{11}B spectrum except that the signal/noise ratio is reduced (due to the lower natural abundance and gyromagnetic ratio of ^{10}B), and the splitting between the five lines is only 27.00 Hz, equivalent to the splitting for the ^{10}B species in the ^1H spectrum. Careful measurements reveal that the center of the $^{11}BH_4^-$ quartet is at +0.00325 ppm, while it is at +0.00500 ppm for the $^{10}BH_4^-$ septet. This same phenomenon was encountered before in Figure 70.15 and results because the chemical shift changes slightly when one isotope is substituted for another.

Spin–spin coupling can be a useful tool for understanding spectra and determining chemical structures. However, it can also be an impediment in some cases. Early in the history of NMR, spin "decoupling" techniques were developed to eliminate spin

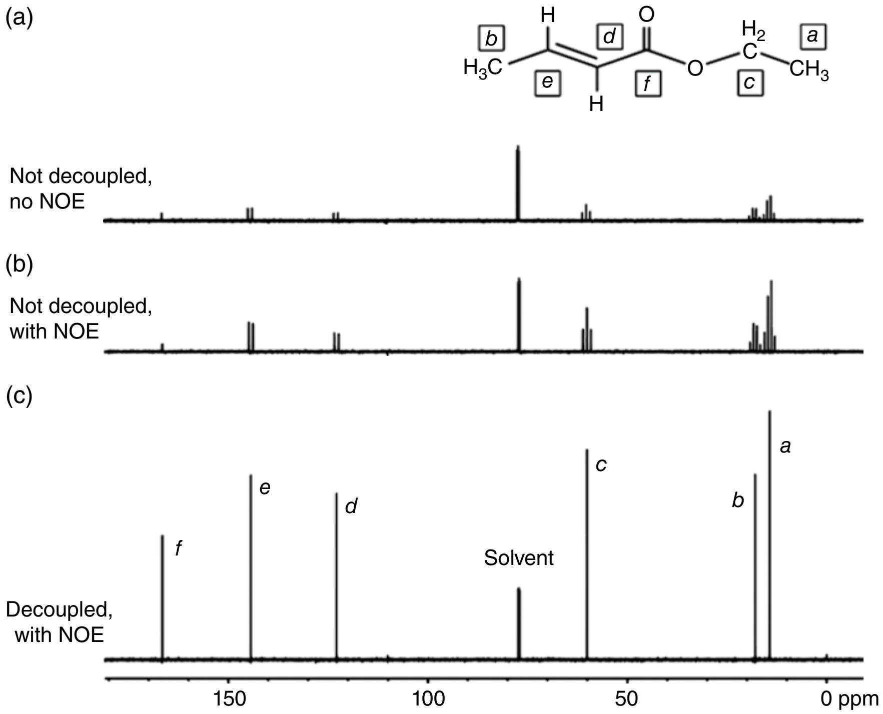


FIGURE 70.27 ^{13}C NMR spectra of 10% (v/v) ethyl *trans*-crotonate in $CDCl_3$ with 0.1% TMS. Acquisition parameters for all spectra: 150.8 MHz; 25°C; 45° tip; 0.87 s acquisition time and 2.0 s relaxation delay (giving 2.87 s total cycle time); 32 averages; 20 Hz spin rate. (a) With the proton decoupler turned off; (b) with the proton decoupler turned off during FID acquisition but on during the relaxation delay; (c) with the proton decoupler turned on continuously throughout all periods in the experiment.

coupling, either selectively for certain multiplets or throughout the entire spectrum in the case of broadband decoupling. A nice, routine application for broadband decoupling is in the acquisition of ^{13}C NMR spectra. In carbon-containing compounds, the carbon atoms are usually bonded to hydrogens. So, ^{13}C NMR spectra will exhibit spin coupling with ^{1}H. That can be a very useful peak assignment tool since it indicates whether the carbon is part of a methyl, methylene, etc., group. However, the signal/noise ratio in ^{13}C spectra is often poor and can be improved by decoupling the protons, thus collapsing the multiplets and concentrating all the available signals into a single peak. In addition, if the decoupling ^{1}H RF field is left on during a substantial part of the experiment, an extra ^{13}C signal/noise improvement of up to almost threefold can be gained due to the so-called nuclear Overhauser enhancement (NOE). Decoupling and NOE effects are illustrated in Figure 70.27. The ^{13}C spectrum (a) was acquired with the proton decoupler turned completely off, and it exhibits full coupling with the ^{1}H. From the multiplicities, each carbon type can be determined clearly. For example, *a* and *b* are both quartets and must be methyl carbons, while *f* appears to be a singlet and must not contain an attached proton. In spectrum (b), the decoupler was turned on during the relaxation delay of the experiment but was turned off just before the FID was acquired so as not to interfere with the coupling. The resulting ^{13}C spectrum still exhibits multiplets, but, due to the NOE, the peaks from carbons bearing attached protons are substantially more intense than in spectrum (a). (The solvent peak did not grow since it contains ^{2}H instead of ^{1}H.) In spectrum (c), the proton decoupler was turned on continuously throughout the entire experiment to gain the NOE and to collapse the multiplets into singlets. The improvement in signal/noise is obvious, and it is now trivial to count the total number of carbon atoms in each molecule simply by counting the number of peaks in the ^{13}C spectrum. Because the NOE preferentially enhances the signals of carbons bearing attached protons, it distorts the relative peak areas in the ^{13}C spectrum. Consequently, there is usually little to be gained by integrating the ^{13}C peaks in the (forlorn) hope of determining the relative numbers of them in the molecule.

Proton decoupling in a ^{13}C spectrum is an example of *heteronuclear* decoupling. RF pulses are applied at the proton resonance frequency at the same time the spectrometer's receiver is acquiring the ^{13}C FID. Since the resonance frequencies for the two nuclides typically differ by hundreds of MHz, it is easy to build electronic circuitry that blocks even high-power ^{1}H RF pulses and prevents them from "bleeding" into the ^{13}C receiver channel. Now suppose one wishes to acquire a proton NMR spectrum where all the protons are fully decoupled from each other. That would be an example of *homonuclear* broadband decoupling. Such a spectrum would contain only a single peak at each chemical shift without the complexities and peak overlap that the spin-coupled multiplets often produce. The practical problem is that RF pulses applied in the same frequency range as the one being detected will interfere grossly with the detector electronics and produce a useless spectrum. So, broadband homonuclear decoupling has been a long-standing technological challenge in NMR, and attempts to solve the problem have only met with limited success. Fortunately, a very recent approach now promises much better results [24].

70.10 NMR RELAXATION EFFECTS

Discussions so far have emphasized exciting NMR spin systems and acquiring and analyzing the resulting spectra. Little has been said about how the excited sample relaxes back to its original state, yet that can profoundly affect the spectrum and NMR spectroscopists ignore it at their peril.

70.10.1 Spin–Lattice Relaxation

First consider the microscopic details of how bulk $+z$ magnetization first develops in a spin-1/2 sample. At the instant the sample is inserted into the magnet, the individual nuclear spin vectors align either parallel or antiparallel to the direction of B_0. A given nucleus is equally likely to adopt either orientation, so equal numbers of nuclei end up in each state and their individual vectors cancel. Consequently, the bulk sample is initially unmagnetized. As the sample "soaks" for a time inside the B_0 field, a small population of antiparallel nuclei flip to the lower-energy, parallel orientation, producing the tiny bulk sample magnetization along +*z*. These flips to the low-energy state release heat, an exothermic process that increases (very slightly) the kinetic energy of molecules inside the sample. Magnetizing an NMR sample is therefore an *enthalpic* process.

The growth of sample magnetization along the *z*-axis is characterized by the sample's "spin–lattice" (or "longitudinal") relaxation time, universally designated by the symbol T_1. When that magnetization is perturbed in an NMR experiment, the T_1 value determines how long it takes the sample to reestablish its magnetization along the +*z*-axis. Following an RF pulse, the *z* component of the sample magnetization increases exponentially according to the equation:

$$M_\tau = M_\infty \cdot \left(1 - A \cdot e^{-\tau/T_1}\right), \tag{70.12}$$

where M_τ is the amplitude of the *z* magnetization at time τ after the pulse, M_∞ is the fully relaxed *z* magnetization amplitude (i.e., at $\tau=\infty$), T_1 is the spin–lattice relaxation time, and A is a constant that depends on the initial nutation angle. If a 90° RF pulse is applied, $A=1$ so that M_τ varies from 0 to $+M_\infty$. Applying an initial 180° pulse causes M_0 to be $-M_\infty$, making $A=2$ in the equation. The time course for the recovery of magnetization along the *z*-axis in these two cases is illustrated in Figure 70.28.

T_1 turns out to be a very important practical parameter. Most NMR spectra are measured by repeating an RF pulse sequence many times and averaging the resulting signals.

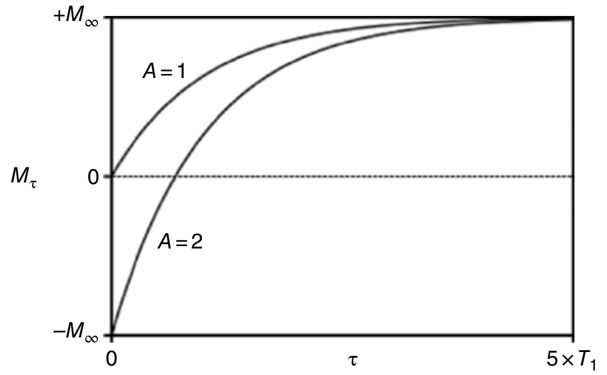


FIGURE 70.28 Recovery of *z* magnetization following an RF pulse. A perfect 90° pulse makes the *z* magnetization zero at $\tau=0$, and it returns to its initial value according to the $A=1$ curve. (That curve also illustrates how bulk sample magnetization forms when the sample is first inserted into the NMR magnet.) A 180° pulse inverts the initial magnetization, which recovers as shown by the $A=2$ curve. In both cases, *z* magnetization is almost completely reestablished after a time period of $5\times T_1$, where T_1 is the spin–lattice relaxation time.

If the sample's +*z* magnetization does not recover fully between pulses, then the magnetization becomes partly "saturated" and only a fraction of the possible NMR signal will be obtained. Thus, T_1 controls how long one must wait before repeating pulses. Samples having long T_1 values can force the investigator to use long delays, leading to inefficient data collection and poor signal-to-noise ratios in the spectra. To make matters worse, the various types of nuclei in the sample may relax at different rates, causing some of them to saturate more than others. That distorts the relative peak areas in the NMR spectrum, leading to errors in determining the relative numbers of, say, methyl *versus* methylene protons and in quantifying analyte concentrations. Quantitative applications of NMR require an acute appreciation of the sample's T_1 values!

70.10.2 Spin–Spin Relaxation

After an RF pulse nutates a sample's bulk magnetization vector away from the +*z*-axis, the vector has a magnetic component in the *xy* plane that is detected as the FID. The amplitude of the FID is known to decrease with time (Fig. 70.9), corresponding to a decrease in the *xy* component of the magnetization vector. This time-dependent loss of *xy* magnetization is caused by "spin–spin" (or "transverse") relaxation and, in most samples, is represented well by the equation:

$$M_\tau = M_0 \cdot e^{-\tau/T_2}, \tag{70.13}$$

where M_0 is the magnitude of the *xy* magnetization component immediately following the pulse (i.e., at $\tau = 0$). T_2 is the time constant for the exponential decay and is called the spin–spin or transverse relaxation time. The time course for the loss of *xy* magnetization is illustrated in Figure 70.29. Since magnetization in the *xy* plane is the result of partial phasing of individual precessing nuclei in the sample, spin–spin relaxation occurs when those phases become more random. It is complete when full

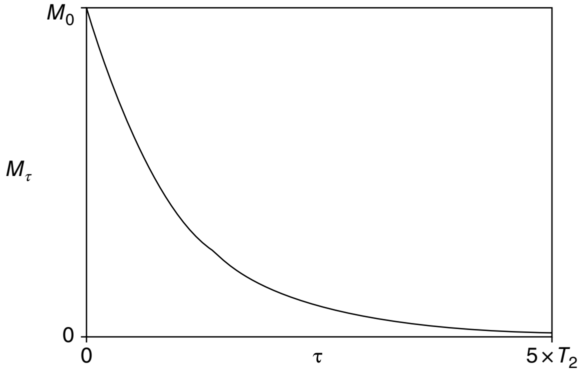


FIGURE 70.29 Loss of *xy* magnetization following an RF pulse. Virtually no transverse magnetization remains after a time period of $5 \times T_2$, where T_2 is the spin–spin relaxation time.

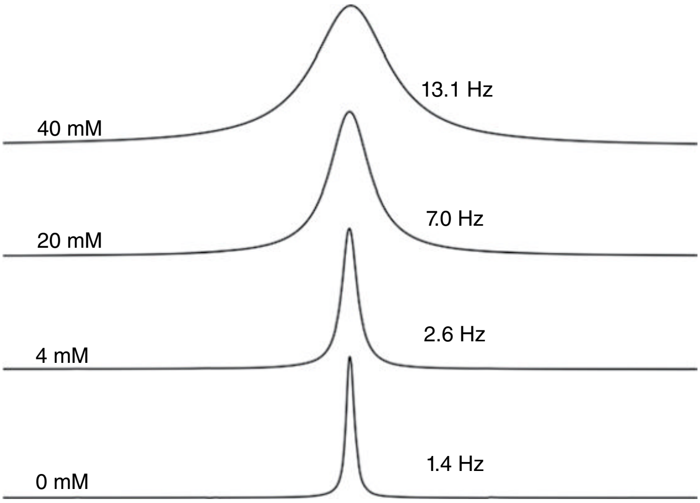


FIGURE 70.30 The effect of paramagnetic ions on proton NMR line widths of HOD. Acquisition parameters: 599.7 MHz; 25°C; 30° tip; 10.5 s total cycle time; 4 averages; 20 Hz spin rate. The samples contained approximately 20% (v/v) HOD in D_2O, along with 0–40 mM $CuSO_4$. A spectral region 100 Hz wide is shown in each case. Full widths at half maximum (FWHM) for the lines varied from 1.4 to 13.1 Hz. The $T_2^* = 1/(\pi * \text{FWHM})$ so, for the 4 mM solution, T_2^* was 0.12 s. Its actual T_2 was measured to be 0.30 s, so most of the width was due to magnetic field inhomogeneity.

phase randomization is achieved. This process involves no energy change, but it does increase the sample's spin entropy. Therefore, spin–spin relaxation is an *entropic* process, in contrast to *enthalpic* spin–lattice relaxation.

The mathematical form of Equation 70.13 determines the line shapes in NMR spectra. As illustrated in Figure 70.10, exponential decay in the time domain produces Lorentzian lines. Moreover, the T_2 value for a time-domain NMR signal determines the minimum width of its Lorentzian peak: $\Delta v_{1/2} \geq 1/(\pi \cdot T_2)$, where $\Delta v_{1/2}$ is the full width of the peak at half-maximum (FWHM) height, in Hz. That has practical importance: fast relaxation (small T_2) produces broad NMR lines (large $\Delta v_{1/2}$), resulting in peak overlap, poor resolution, and perhaps reduced signal/noise ratio. Figure 70.30 shows proton NMR lines for aqueous solutions of copper(II) ion, a paramagnetic species that hastens relaxation. The line widths clearly increase with the ion concentration. The reason the sample's T_2 only sets the minimum width for the line is because static magnetic field inhomogeneity also broadens all the spectral lines, so only part of the width of any given experimental peak is due to its T_2. The width that results from the combination of T_2 and magnetic field inhomogeneity is so widely encountered that it is given a special symbol: T_2^* (pronounced "tee too star"). This is an "effective" spin–spin relaxation time. In other words, T_2^* is the T_2 value that *would be* required to produce the experimentally observed line width *if* the B_0 field were perfectly homogeneous. When a raw NMR line width is used to calculate the relaxation time, it is always T_2^* and not T_2 that is being measured. By using a well-shimmed magnet, the true T_2 value can often be estimated with rough accuracy from the line width alone, as shown in Figure 70.30. Mathematically, the relationships between T_2, T_2^*, and their corresponding rates R ($=1/T$) are expressed by

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{\text{field inhomog}}} = R_2^* = R_2 + R_{\text{field inhomog}}.$$

Spin–lattice and spin–spin relaxation times can be measured accurately by using pulse experiments. The most common sequence for measuring T_1 is inversion recovery, but saturation recovery is also used. For T_2, the Carr–Purcell–Meiboom–Gill (CPMG) sequence is usually employed. Such measurements are beyond the scope of this work but may easily be found in standard NMR references. All commercial NMR spectrometers come equipped with the required pulse sequences.

NMR spectra are almost always measured by acquiring multiple FIDs, which are then averaged to produce the final result. That requires repeating the pulse sequence multiple (n) times, in cycles of (RF pulse—acquisition—relaxation delay)$_n$. The total time between RF pulses is called the “cycle time.” It is tempting to acquire NMR spectra by using the smallest possible cycle time so the greatest number of FIDs can be acquired in a given period. By averaging more FIDs, the signal/noise ratio should increase. That line of reasoning is seductive, but it carries a hidden flaw: if the sample is not allowed to relax completely between pulses, then each pulse produces only a fraction of the full signal. A 90° RF pulse might produce a big NMR signal the first time it is applied, but if the pulse is repeated before the spin system has time to relax completely, the second FID will be less intense than the first. The two approaches, using a short cycle time so many FIDs can be acquired *versus* using a long cycle time that permits full relaxation, are mutually exclusive. There must be some intermediate cycle time that produces the best overall result. That optimum compromise is expressed by the equivalent equations:

$$\cos\alpha = e^{-\left(T_{\text{cycle}}/T_1\right)} \qquad \alpha = \cos^{-1}\left\{e^{-\left(T_{\text{cycle}}/T_1\right)}\right\} \qquad \frac{T_{\text{cycle}}}{T_1} = -\ln\left(\cos\alpha\right).$$

In the equations, α is the so-called Ernst angle, T_{cycle} is the total time between RF pulses, and T_1 is the sample’s spin–lattice relaxation time. The Ernst angle is the tip angle that maximizes the signal/noise ratio in the spectrum for any given ratio of cycle time to T_1. In practice, tip angles of 30°, 45°, and 60° are most widely used, and the equations show the corresponding optimum cycle times to be $0.14\times T_1$ (for 30°), $0.35\times T_1$ (for 45°), and $0.69\times T_1$ (for 60°). Using a bigger tip angle produces more signal per FID, but it tends to saturate the spin system more and therefore requires a longer delay to permit relaxation. Note that these combinations maximize the signal/noise ratio but they also saturate the signals to some extent and will therefore distort the relative peak areas if the various spins have different T_1 values.

70.10.3 Quantitative Analysis by NMR

Providing spin saturation is avoided, the area under an NMR peak is directly proportional to the number of nuclei that produce the peak. There is no need to scale the area to account for differences in detector sensitivity, making NMR an obvious method for quantitative analysis. There are two problems, however. First, compared to other instrumental methods, NMR is not very sensitive so large samples and/or high analyte concentrations must be used. The second problem is relaxation. Measurements must be performed in a way that does not distort the relative peak areas in the spectrum. That automatically precludes running under conditions that yield the optimum signal/noise ratio per unit time. The only practical method is to choose a cycle time much longer than T_1, usually at least $3\times T_1$ but often as much as $5\times T_1$. From Equation 70.12 with $A=1$, it is easy to show that using a cycle time of $3\times T_1$ causes a 4.98% error in the peak area while a cycle time of $5\times T_1$ produces only a 0.67% error. The latter is generally acceptable and is widely used for the best work. Having to wait $5\times T_1$ between pulses implies a slow experiment, and that can certainly be the case. However, it is often possible to speed relaxation by adding a paramagnetic substance. Paramagnetics not only can reduce the T_1 values, but, in samples with nuclei that relax at very different rates, the paramagnetic can help equalize the T_1's as it simultaneously reduces them. Although the advantages are clear, paramagnetic compounds also tend to broaden the NMR peaks (Fig. 70.30) so the temptation to use too much must be resisted.

An example of tissue phospholipid quantitation by ^{31}P NMR [25, 26] is shown in Figure 70.31. The T_1's for the various compounds were long (seconds), especially for the tributylphosphate internal standard, so a paramagnetic was added to decrease them and the differences between them. Since lipid analytes are generally nonpolar, the solvent (chloroform) was also necessarily nonpolar, making it impossible to dissolve simple ions like copper(II). Fortunately, highly paramagnetic chromium(III) can be purchased as the relatively nonpolar acetylacetonate complex (Cr(acac)$_3$). By dissolving that complex in the sample, the ^{31}P T_1 values were reduced enough to allow accurate quantitation of phospholipids from 0.5 to 1 g of tissue in a 2 h measurement. With the 7.06 T magnetic field used, considerable peak overlap was unavoidable. However, accurate peak areas were still obtained by fitting theoretical Lorentzian peaks to the experimental spectrum.

70.11 DYNAMIC PHENOMENA IN NMR

NMR is a powerful tool for analyzing molecular structures, but it has another important application: it can reveal crucial information about molecular dynamics. Molecules in solution undergo rapid, continuous motion, and those motions greatly simplify NMR spectra by averaging the chemical shifts, which depend on the molecule's orientation relative to the B_0 direction. Most solution spectra contain narrow, discrete lines

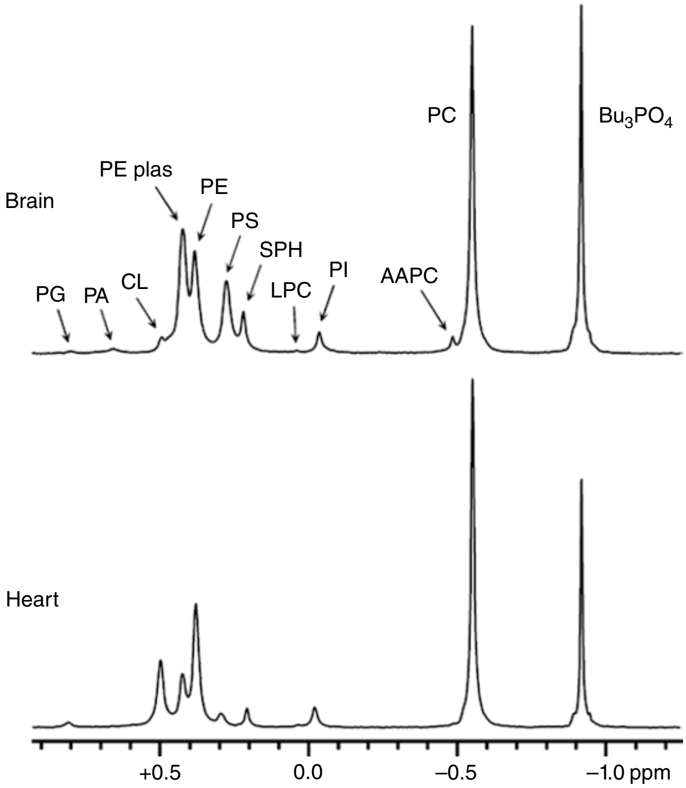


FIGURE 70.31 ^{31}P NMR spectra used to quantify the phospholipids in extracts of rat brain and heart. Tributylphosphate was added as an internal standard, and the acetylacetonate complex of chromium(III) was included to reduce the T_1 values of the compounds. A 7.06T magnet was used, giving a ^{31}P resonance frequency of 121.6MHz. The peak labels designate various classes of phospholipids, which are dominated by phosphatidylcholine (PC).

due to that motional averaging and are said to be in the "fast exchange limit." In some cases, the motion is slow enough or can intentionally be made slow enough to prevent complete averaging, thus allowing NMR to be used to study the dynamics of processes. Whether such effects can be observed depends on their rates relative to the "NMR time scale." Suppose a molecule can adopt either of two conformations that exhibit two ^1H NMR lines separated by, say, 1 ppm. If we acquire the spectrum of this compound with a 300MHz spectrometer, the two lines will be separated by 300Hz. If the compound is able to switch conformations faster than roughly 300 times per second, the two distinct lines will tend to average and merge into one. On a 600MHz spectrometer, the exchange rate between the two conformations must occur twice as fast to average the same lines, which are now separated by 600Hz. So, it is clear that the "NMR time scale" is not fixed but depends on the conditions of the experiment. Despite that, NMR can provide insights into molecular dynamics that are challenging to obtain in other ways.

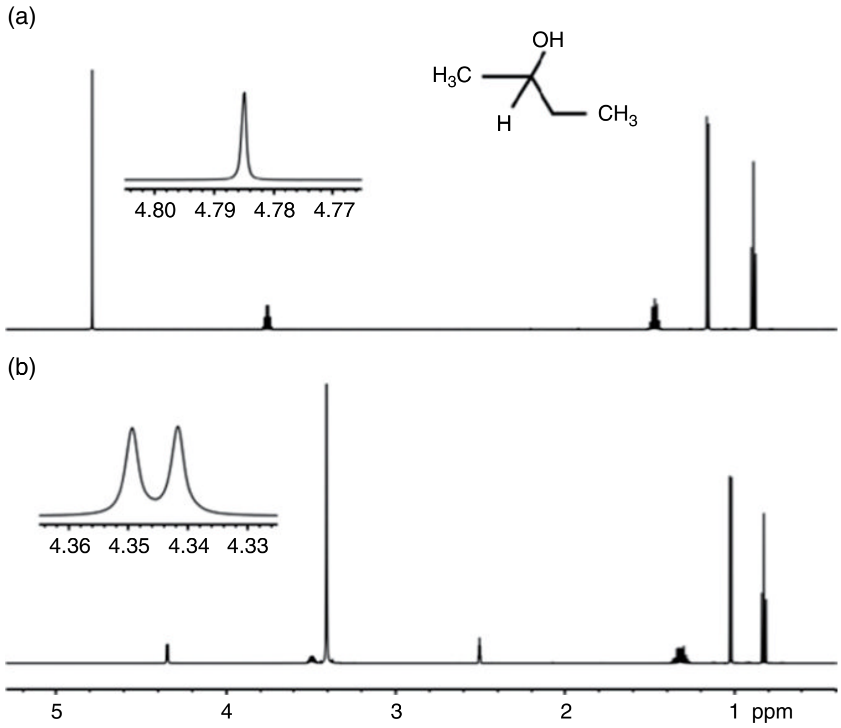


FIGURE 70.32 Proton NMR spectrum of approximately 5% (v/v) 2-butanol dissolved in (a) D_2O and (b) DMSO-d_6. Acquisition parameters: 599.7 MHz; 25°C; 30° tip; 10.5 s total cycle time; 8 averages; 20 Hz spin rate. The chemical shift reference was internal TSP (for D_2O) or TMS (for $CDCl_3$). In spectrum (a), the butanol —OH proton exchanged rapidly with the solvent and merged with the HOD peak to form a singlet near 4.785 ppm. In DMSO-d_6 (b), the exchange was slowed sufficiently to produce a separate butanol —OH peak near 4.35 ppm. The water peak appeared near 3.4 ppm, and residual protonated DMSO produced the multiplet at 2.5 ppm.

If the molecular exchange rate is slow on the NMR time scale, distinct peaks will be observed for each form of the molecule. This is the "slow exchange limit." An example of that has already been encountered in Figure 70.23, which contains the spectrum of 2,4-pentanedione, a keto-molecule that can tautomerize to form an enol. In solution both tautomers exist in equilibrium simultaneously and, at room temperature, the exchange rate from one form to the other is slow enough to produce very distinct NMR peaks. By integrating peaks associated with each form, the equilibrium constant can be calculated.

The exchange rate between chemical forms can sometimes be manipulated to either produce or eliminate multiple peaks in the spectrum. An example of using the solvent to do that is shown in Figure 70.32. The ^{1}H spectrum (a) is that of 2-butanol dissolved in D_2O. Hydrogen bonding between the alcohol and solvent promotes relatively fast exchange of the alcohol —OH proton with the D_2O to produce a single, averaged singlet near 4.8 ppm. Dissolving the alcohol in DMSO-d_6 slows the exchange rate

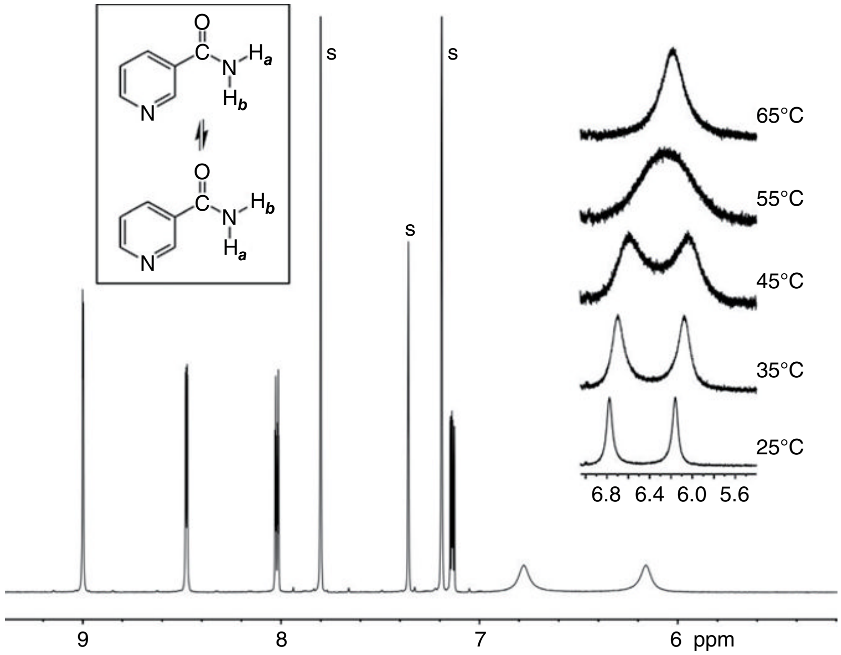


FIGURE 70.33 Proton NMR spectra of nicotinamide in nitrobenzene-d_5. Acquisition parameters: 599.7 MHz; 30° tip; 5.5 s total cycle time; 8 averages; 20 Hz spin rate. The temperature was 25°C except where noted. Peaks labeled "*s*" arise from residual ^{1}H in the solvent, which was used due to its high boiling point. Delocalization of lone-pair electrons from the nitrogen atom into the carbonyl group endows the C—N bond with partial double-bond character, inhibiting free rotation. Consequently, the two amido protons are not equivalent and produce separate peaks. Raising the temperature increases the rotation rate about the bond and averages the amido proton chemical shifts. The two peaks just merge at 54°C.

enormously, allowing a separate, distinct alcohol —OH peak to be observed (b). Diminished hydrogen bonding also shifts the peak upfield from its position in spectrum (a). With slow ^{1}H exchange, the alcohol —OH proton remains attached to the molecule long enough to manifest spin coupling with the lone proton on the adjacent carbon atom, yielding a doublet that proves it to be a secondary alcohol. This technique can usually be used to determine unambiguously whether an unknown alcohol is primary, secondary, or tertiary.

The choice of solvent is not the only approach for manipulating the dynamics of molecular exchange processes. Figure 70.33 shows how temperature can be used to the same end. In nicotinamide, as in most amide compounds, partial double-bond character in the C—N bond slows its rotation and makes the two amido hydrogen atoms inequivalent. Each hydrogen produces a separate broad resonance peak at a different chemical shift. (Broadness in NMR peaks usually suggests the presence of some dynamic process.) As shown in the figure, the molecule can be considered to exist in two distinct states where the positions of those hydrogens are interchanged. By raising the sample temperature, the rotational rate about the C—N bond increases and leads to averaging on this time scale, causing the two spectral peaks to converge and broaden. The temperature at which they just merge into a single, flat-topped peak is called the coalescence temperature (54°C, in this case). Continued increases in temperature narrow the peak. At the coalescence temperature, the lifetime $\tau_{\text{coalescence}}$ for each of the two states can be calculated from

$$\tau_{\text{coalescence}} = \frac{1}{k_{\text{coalescence}}} = \frac{\sqrt{2}}{\pi \cdot \Delta v}, \tag{70.14}$$

where $k_{\text{coalescence}}$ is the pseudo-first-order rate constant (in Hz) for exchange between the two states and Δv is the chemical shift difference (in Hz) between the two peaks in the slow exchange limit (i.e., at very low temperatures). By using the chemical shift difference (369.5 Hz) at 25°C as an estimate of Δv for nicotinamide, Equation 70.14 shows the lifetime to be 1.22 ms at 54°C, with a corresponding exchange rate of 821 s^{-1}. Since the true Δv is undoubtedly larger than the estimate, perhaps by severalfold, these values should be considered as the upper limit for the lifetime and lower limit for the exchange rate. Even so, they are informative. Although the exchange phenomenon in this particular example involves peaks with different chemical shifts, exactly the same kind of averaging can occur when the peaks arise from scalar coupling instead. In other words, molecular motions can average both chemical shift and scalar coupling information.

It is possible to calculate the shapes of NMR spectral lines in exchanging systems for any combination of Δv and lifetime τ. (See Ref. 27 for a particularly clear discussion.) Fitting the theoretical shapes to experimental spectra allows the exchange rate constant to be found at temperatures that differ from the coalescence temperature. Variable-temperature NMR is widely used to find thermodynamic parameters for molecular inversion and other dynamic processes [28]. For intramolecular exchange processes such as rotation, the thermodynamic activation parameters can be calculated from the relationship

$$k_{\text{r}} = \kappa \frac{kT}{h} \cdot e^{\left(-\Delta G^{\ddagger}/RT\right)} = \kappa \frac{kT}{h} \cdot e^{\left(-\Delta H^{\ddagger}/RT\right)} \cdot e^{\left(-\Delta S^{\ddagger}/RT\right)}, \tag{70.15}$$

where k_{r} is the exchange rate constant (=$1/\tau$), κ is a constant that depends on the nature of the process, k is Boltzmann's constant, T is absolute temperature, h is Planck's constant, R is the gas-law constant, $\Delta G^{\ddagger}$ is the free energy of activation, $\Delta H^{\ddagger}$ is the enthalpy of activation, and $\Delta S^{\ddagger}$ is the entropy of activation. The value of κ depends on the particular system, but in practice, it is often approximated as 1.

A practical complication in variable-temperature NMR studies is that, while commercial spectrometers are reasonably good at maintaining a constant, uniform sample temperature, they often do not report it very accurately. The actual and displayed temperature can easily differ by several degrees. That is not necessarily the fault of the manufacturer. The actual sample temperature depends on several variables such as the gas flow rate through the NMR probe, the volume of sample in the NMR tube, and whether broadband decoupling is being used. In the latter case, the RF pulses from the decoupler can actually heat the sample by ≥10°C, especially when using a solvent with a high dielectric constant. So, it is difficult to produce a temperature calibration that is reliable under all conditions. Consequently, careful variable-temperature work usually requires one or more extra NMR experiments to measure the actual sample temperature. These are not difficult to perform and rely on known effects of temperature on the chemical shifts for various compounds. Among the most widely used compounds for ^{1}H NMR are methanol (MP = −97.8°C; BP = 64.7°C) for work at low temperatures and ethylene glycol (MP = −13°C; BP = 197.6°C) for high-temperature studies. The ^{1}H NMR spectra of both of those compounds contain only two peaks, and the chemical shift difference between them depends almost linearly on temperature [29]. The compound tris(trimethylsilyl)methane has been proposed as an "NMR thermometer" for measurements in ^{13}C NMR [30]. It is normally unwise, if not impossible, to add the temperature calibrant directly to the sample. Instead, the usual procedure is to place the compound in a glass capillary, insert the capillary into the NMR sample, allow the temperature to equilibrate, and then run a quick spectrum to obtain the chemical shift data. The capillary is likely to degrade the shim, but it will still be good enough to permit accurate temperature measurements. Another option is to fill a separate NMR tube with the calibrant, insert it into the spectrometer in place of the original sample, allow its temperature to equilibrate, and then measure the necessary chemical shift(s). A final possibility is to insert a thermistor into the center of the sample. That can work well, but the wire leads from the thermistor will prevent the tube from spinning, which might affect the sample temperature slightly. RF pulses should never be applied while the thermistor is in place since its leads can act as antennas, concentrating the energy and causing the sample to heat.

70.12 MULTIDIMENSIONAL NMR

The NMR methods described earlier demonstrate the enormous power of the technique. However, with increasing applications to larger and more complex molecules, and especially to biopolymers, the spectra tend to become extremely crowded and exhibit extensive peak overlap even when the highest available magnetic fields are employed. One solution to the crowding problem is to spread the peaks into more than one dimension, and that is exactly what multidimensional NMR techniques do. However, merely dispersing the peaks better, as important as that may be, is only one of the many advantages of multidimensional NMR. Indeed, this approach can provide a tremendous range of information quickly that was previously impossible, or at least difficult and slow, to obtain by 1D methods. As a result, multidimensional methods have revolutionized the practice of NMR since their introduction in the 1970s.

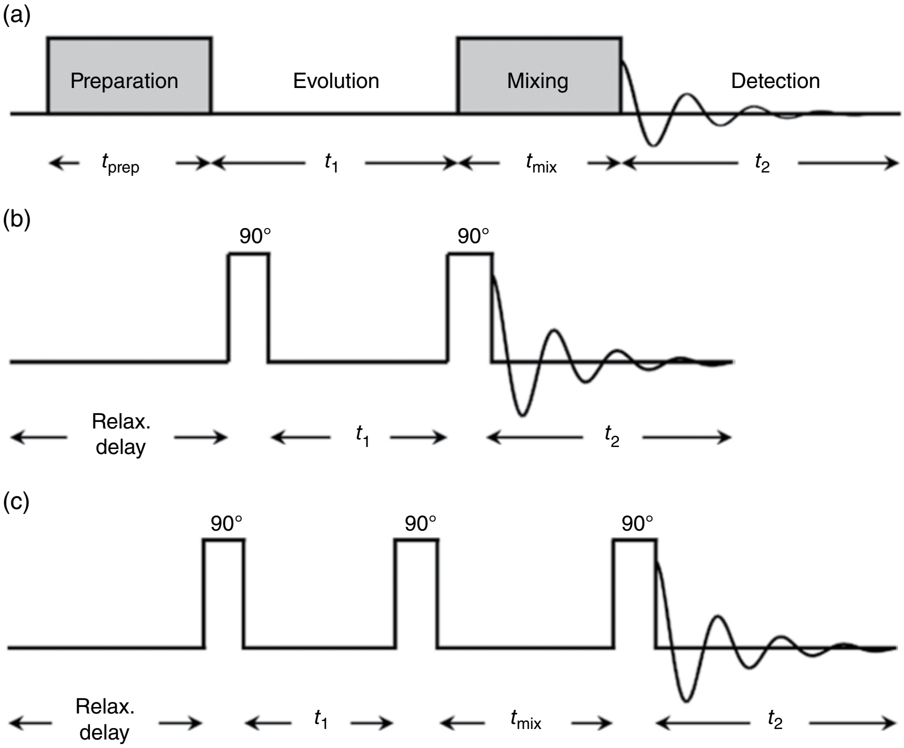


FIGURE 70.34 Two-dimensional NMR pulse sequences. (a) General characteristics. (b) The basic pulse sequence for correlation spectroscopy (COSY). (c) The basic pulse sequence for nuclear Overhauser effect spectroscopy (NOESY).

This brief discussion will be limited to 2D techniques (2D NMR), although extension to additional dimensions is straightforward. In addition to the relaxation delay, short receiver recovery delay, and RF pulse width, a 1D NMR spectrum is normally acquired by using only one other time period: the acquisition time during which the FID is recorded. We define that time period as t_2 (not to be confused with the spin–spin relaxation time!). The principle of 2D NMR, first expressed by Jean Jeener at an AMPERE Conference in 1971, involves using a pulse sequence that contains *two* independent time periods t_1 and t_2. These are called the evolution and detection periods and are illustrated in Figure 70.34. The actual FID is acquired during period t_2, just as in 1D NMR, but any evolution the spin system undergoes during period t_1 will also be reflected in the FID. In addition to these two periods, the 2D pulse sequence also includes an initial preparation period, which is usually a relaxation delay of fixed length followed by a 90° RF pulse to excite the spin system. A few sequences include a mixing time, also of fixed length, to allow the exchange of information between different spins.

Figure 70.34 shows the basic pulse sequence for the "correlation spectroscopy" (COSY) experiment. It begins with a relaxation delay and a 90° RF pulse. That is followed by the time period t_1. On the first pass, a t_1 value of essentially zero is used. Then

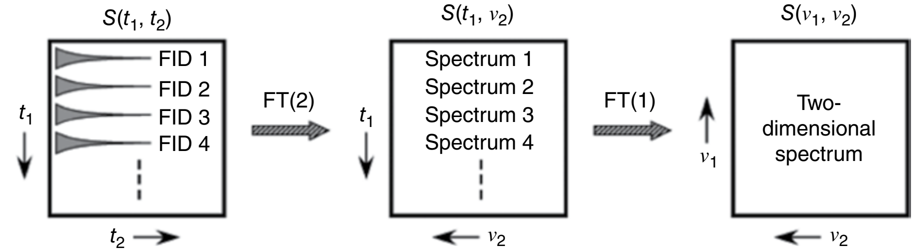


FIGURE 70.35 Steps in processing a 2D NMR data set. Each FID is first Fourier transformed in the t_2 direction, as usual. The first points in each of resulting spectra can be assembled into a set and considered to be another FID but this time along the t_1 direction. Likewise, the set of second points in all the spectra form another FID in the t_1 direction and so on. These FIDs are Fourier transformed in the t_1 direction. Ultimately, the data set that began as a function of (t_1, t_2) has been transformed into the (ν_1, ν_2) frequency domain: a 2D NMR spectrum.

the second RF pulse is applied, and the FID is acquired and stored in the computer. We will define it as FID 1. The sequence is then repeated by starting again with the relaxation delay and 90° pulse. This time, we use a t_1 delay that is increased by a "dwell" time (DW) of, often, several hundred microseconds to give the spin system time to evolve. Then the second pulse is applied and FID 2 is acquired and stored separately from FID 1. The whole process is repeated again, this time using $t_1 = 2 \times \text{DW}$ and saving FID 3. It is common to repeat the sequence $n = 256$–1024 times to acquire an array of n FIDs, with each one representing an evolution time of $t_1 = (n - 1) \times \text{DW}$.

Now, what do we do with the set of n FIDs? As illustrated in Figure 70.35, we first Fourier transform each FID in the usual manner. Since they were all acquired during the period t_2, the frequency-domain axis is designated as ν_2. Inspecting the resulting n frequency domain spectra reveals that their peak amplitudes change (oscillate) from one spectrum to the next. So, we Fourier transform in the t_1 direction to create frequency-domain spectra along a ν_1 axis in that direction. The data set has now been transformed from a 2D time-domain function $s(t_1,t_2)$ to a frequency-domain function $S(\nu_1,\nu_2)$. This is a 2D NMR spectrum. If nothing "interesting" occurred during the evolution period t_1, then the 2D NMR spectrum will simply contain the conventional 1D spectrum spread along its diagonal axis. However, "interesting" things normally do happen during t_1, resulting in the appearance *extra peaks* in the 2D spectral plane to accompany the 1D spectrum that lies along the diagonal.

When extra peaks appear in a 2D NMR spectrum, their meaning depends on exactly which pulse sequence was used to acquire the data. Figure 70.36 contains a schematic illustration of a COSY spectrum, where cross peaks arise only when the spins are J coupled. In the diagram, the cross peaks labeled "A" appear at the intersection of the chemical shifts of peaks 1 and 4 along the diagonal, meaning that the spins that produce peaks 1 and 4 are spin coupled. The cross peaks labeled "B" show that the nuclei of peak 4 are also coupled to the nuclei that produce peak 2. The cross peaks "C" also indicate coupling between the nuclei that give rise to peaks 4 and 5.

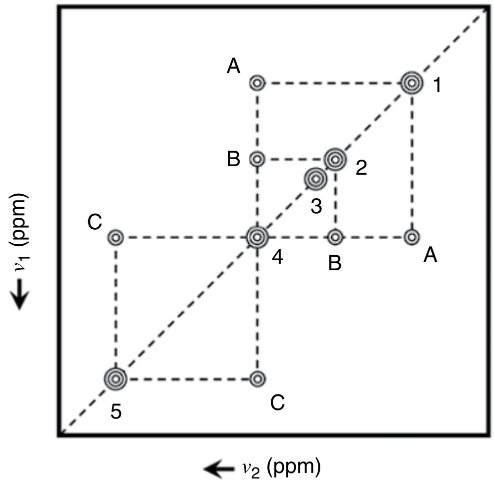


FIGURE 70.36 Schematic diagram of a COSY spectrum shown as a contour plot. The 1D NMR spectrum lies along the diagonal. Cross peaks correlate various peaks in the 1D NMR spectrum, showing that they are spin coupled. The cross peak positions shown lead to the following conclusions. The nuclei that produce peak 1 are coupled to those that produce peak 4 (cross peaks A). The peak 1 nuclei are not coupled to those of peak 2 (no cross peaks). Peak 2 is coupled to peak 4 (cross peaks B). Peak 3 is not coupled to any other peaks. Peak 4 is coupled to peaks 1, 2, and 5. Peak 5 is coupled only to peak 4 (cross peaks C).

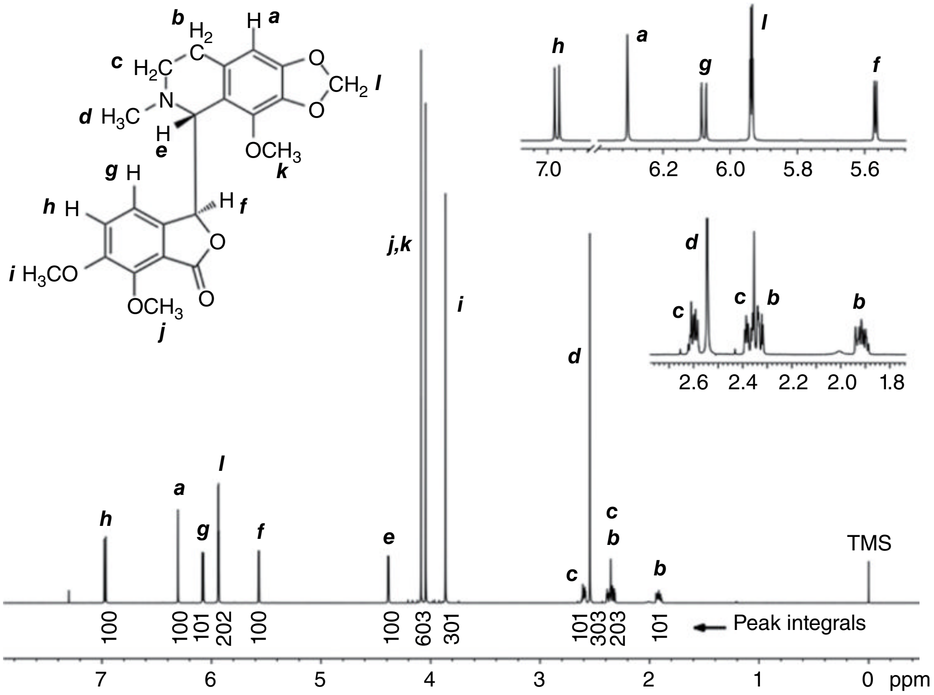


FIGURE 70.37 ^1H 1D NMR spectrum of 150 mg noscapine in 1 mL $CDCl_3$. Acquisition parameters: 599.7 MHz; 25°C; 30° tip; 11.9 s total cycle time; 8 averages; 20 Hz spin rate; TMS chemical shift reference. Some of the peaks can be assigned by using the 1D spectrum alone, such as the lone aromatic proton ***a*** singlet, the singlet for the two ***l*** protons that are highly deshielded by adjacent oxygen atoms, and the singlet for the three methyl protons ***d***. Some of other assignments are more challenging and are aided by 2D NMR methods.

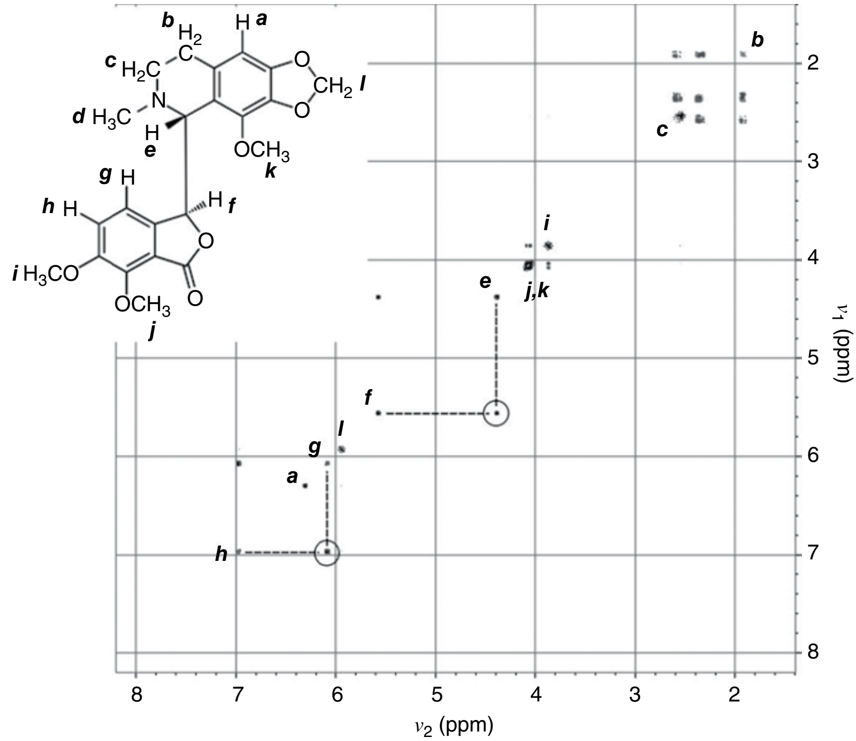


FIGURE 70.38 Experimental ^{1}H COSY spectrum of 150 mg noscapine in 1 mL $CDCl_3$. Acquisition parameters: 599.7 MHz; 25°C; 90° tip; 1.0 s relaxation delay; 256 t_1 values; 8 averages per t_1 value; 20 Hz spin rate; TMS chemical shift reference. Total data acquisition time was 41.5 min. Cross peaks confirm J coupling between the three upfield multiplets due to protons ***b*** and ***c***. Cross peaks also show coupling between protons ***e*** and ***f***, as well as between ***g*** and ***h*** (dashed lines). The apparent cross peaks between proton peaks ***i*** and ***j; k*** are artifacts due to overlap at the bases of these intense peaks.

An experimental ^{1}H 1D NMR spectrum of noscapine is shown in Figure 70.37. Some of the peaks can be assigned unambiguously, such as the singlets. However, the assignment of other peaks is more difficult to perform and can benefit from 2D NMR data. Figure 70.38 shows the COSY spectrum, which not only confirms J coupling between peaks ***e*** and ***f***, as well as ***g*** and ***h***, but also shows that all three multiplets in the 1.8–2.6 ppm region are mutually coupled.

Even more informative is the NOESY spectrum (Fig. 70.39). In NOESY, the pulse sequence includes a mixing time that allows nuclear Overhauser interactions to occur between excited spins. The NOE is a *through-space* interaction, so the appearance of cross peaks in a NOESY spectrum is evidence that the corresponding nuclei are near each other in space. If the mixing time is short, cross peaks appear only if the spins are very near each other. A longer mixing time allows spins that are more widely separated to produce cross peaks. In the case of noscapine, the NOESY spectrum confirms that

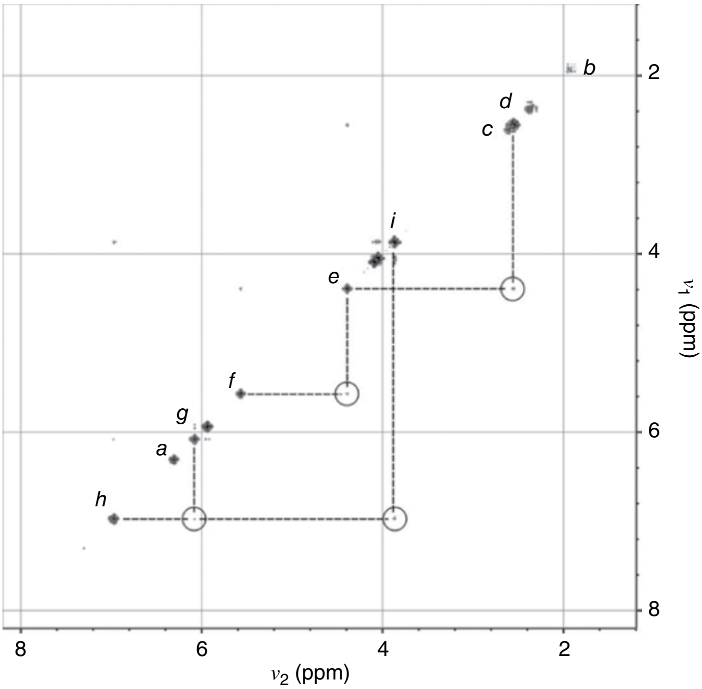


FIGURE 70.39 Experimental ^{1}H NOESY spectrum of 150 mg noscapine in 1 mL $CDCl_3$. Acquisition parameters: 599.7 MHz; 25°C; 90° tip; 2.0 s relaxation delay; 256 t_1 values; 4 averages per t_1 value; 500 ms mixing time; 20 Hz spin rate; TMS chemical shift reference. Total data acquisition time was 1.5 h. All cross peaks appear symmetrically relative to the diagonal, but for clarity, only those in the lower half of the spectrum are circled. The cross peaks show that proton ***e*** is spatially near protons ***d*** and ***f*** but that ***d*** and ***f*** are not near each other (no cross peaks link the two). Likewise, proton ***h*** is near protons ***g*** and ***i***, but ***g*** and ***i*** are not near each other (again, no cross peaks link them directly).

proton ***e*** is spatially near protons ***d*** and ***f***, although protons ***d*** and ***f*** are not near each other. It is therefore logical that proton ***e*** should occupy a position between ***d*** and ***f***, as shown in the chemical structure. In addition, the peak for proton ***h*** exhibits cross peaks to the peaks of both protons ***g*** and ***i***, again consistent with the proposed structure. That also allows the singlet for the methoxy protons ***i*** to be assigned to the peak at 3.87 ppm, narrowing the possible assignments of the other two methoxy groups ***j*** and ***k*** to the intense singlets at 4.05 and 4.08 ppm (though the order is not clear). 2D NMR methods provide more information than can be obtained from 1D NMR spectra alone, and the figures demonstrate how that additional information can sometimes prove crucial in making NMR peak assignments.

A great many different 2D NMR experiments have been developed, but the great majority of studies only employ a few of them. Table 70.4 lists the most important ones and includes brief descriptions and other information.

TABLE 70.4 Some Common Two-Dimensional NMR Pulse Experiments[a]

Experiment Name	Typical Speed	Features
COSY: Correlation Spectroscopy	Relatively fast (e.g., 10–40 min)	This homonuclear experiment is widely used for proton spectra. Cross peaks reveal which spins are scalar coupled. It works best if $J = 3–15$ Hz (values typical of geminal or vicinal protons). Excellent for molecular structure analysis and assigning spectra.
TOCSY: Total Correlation Spectroscopy	Relatively fast (e.g., 10–40 min)	Like COSY, cross peaks show which pairs of spins have scalar coupling. Compared to COSY, this experiment is more sensitive to weak homonuclear coupling and yields cross peaks even from widely separated spins with small J values.
HSQC: Heteronuclear Single-Quantum Correlation	Moderately fast (e.g., 20–60 min)	A heteronuclear 2D method with the spectrum of one nuclide I (e.g., ^{1}H) along axis v_2 and that of the other nuclide S (e.g., ^{13}C) along the other axis v_1. A cross peak appears if an I spin is directly J coupled to an S spin. Very useful for assigning spectra. In the v_2 direction, the cross peaks are multiplets split by $^1J_{\text{IS}}$. The HSQC signal is detected from the abundant I spins, so the experiment is relatively fast.
HMQC: Heteronuclear Multiple-Quantum Correlation	Moderately fast (e.g., 20–60 min)	Like HSQC, this heteronuclear experiment yields a 2D spectrum with the I nuclide (e.g., ^{1}H) along one axis and the S nuclide (e.g., ^{13}C) along the other. A cross peak indicates direct J coupling of an I spin with an S spin. The S peaks show splitting due to coupling between I spins and, so, are a little broader than those in HSQC.
HETCOR: Heteronuclear Correlation	Slow (usually hours)	This is also a heteronuclear experiment. It is analogous to HSQC, but the signal is detected from the low-abundance S spins (e.g., ^{13}C), resulting in relatively poor signal/noise ratio per unit time. The HETCOR experiment has been displaced by HSQC or HMQC for most purposes.
NOESY: Nuclear Overhauser Enhancement Spectroscopy	Slow (usually hours)	A homonuclear experiment, usually performed with ^{1}H. Cross peaks occur between spins that are near each other in space. (The cross peak intensities and growth rates depend on r^6.) Especially useful for small molecules. Molecules with formula weights of 1000–3000 often yield weak or no cross peaks due to their unfavorable molecular correlation times.
ROESY: Rotating Frame Overhauser Enhancement Spectroscopy	Slow (usually hours)	Similar to NOESY in that cross peaks indicate spatial proximity between different spins. Unlike NOESY, molecules with formula weights of 1000–3000 produce strong cross peaks. ROESY is often a bit faster than NOESY but can still require hours.
HOESY: Heteronuclear Overhauser Effect Spectroscopy	Slow (usually hours)	This is a heteronuclear analogue of NOESY. A cross peak appears in the spectrum when an I spin is spatially near an S spin. It can be used, for example, to investigate ^{1}H-^{13}C and ^{1}H-^{31}P distances within molecules.

[a] Listed are some 2D NMR experiments and their general features. The ones given here constitute a small fraction of all such experiments that have been developed, but they are among the most widely used. (*The* most popular is probably COSY.) Some of the experiments are *homonuclear* (i.e., the spectra along the v_1 and v_2 axes are from the same nuclide, such as ^{1}H). Other experiments are *heteronuclear*, having the spectrum of nuclide I (such as ^{1}H) along one axis and the spectrum of a different nuclide S (like ^{13}C) along the other axis. The experiments are divided into (i) homonuclear correlation methods, (ii) heteronuclear correlation methods, and (iii) methods that detect the spatial proximities of different spins.

70.13 CONCLUSION

From the information presented in this chapter, the power and enormous range of applications of NMR should be apparent. While it is not possible to cover 70 years of development in a single chapter, the goal has been to provide a clear introduction to both the principles of NMR and to many of its modern techniques. It is hoped that this information will serve as a useful overview and a springboard for those who seek additional information in the field.

REFERENCES

1. W. Gerlach and O. Stern, *Ann. Phys.* 1924, 74, 673–699.
2. I. I. Rabi, J. R. Zacharias, S. Millman, and P. Kusch, *Phys. Rev.* 1938, 53, 318.
3. G. J. Gorter and L. F. J. Broer, *Physica* 1942, 9, 591–596.
4. F. Bloch, W. W. Hansen, and M. Packard, *Phys. Rev.* 1946, 69, 127.
5. E. M. Purcell, H. C. Torrey, and R. V. Pound, *Phys. Rev.* 1946, 69, 37–38.
6. W. D. Knight, *Phys. Rev.* 1949, 76, 1259–1260.
7. W. G. Proctor and F. C. Yu, *Phys. Rev.* 1950, 77, 717.
8. R. R. Ernst and W. A. Anderson, *Rev. Sci. Instrum.* 1966, 37, 93–102.
9. R. N. Bracewell, "*The Fourier Transform and its Applications*," 2nd edition, McGraw-Hill Book Company, New York, 1978, p. 104.
10. G. A. Morris, *J. Magn. Reson.* 1988, 80, 547–552.
11. K. R. Metz, M. M. Lam, and A. G. Webb, *Concepts Magn. Reson.* 2000, 12, 21–42.
12. L. M. Tolbert, *Acc. Chem. Res.* 1992, 25, 561–568.
13. Y. Gaoni, A. Melera, F. Sondheimer, and R. Wolovsky, *Proc. Chem. Soc.* 1964, 397–398.
14. A. Kuhn, P. Sreeraj, R. Pöttgen, H.-D. Weimhöfer, M. Wilkening, and P. Heitjans, *Angew. Chem. Int. Ed.* 2011, 50, 12099–12102.
15. L. L. Borer, J. G. Russell, R. E. Settlage, and R. G. Bryant, *J. Chem. Educ.* 2002, 79, 494–497.
16. R. K. Harris and B. E. Mann (eds.), "*NMR and the Periodic Table*," Academic Press, New York, 1978.
17. P. Laszlo (ed.), "*NMR of Newly Accessible Nuclei*," Vol. 1 and 2, Academic Press, New York, 1983.
18. L. M. Jackman and S. Sternhell, "*Applications of Nuclear Magnetic Spectroscopy in Organic Chemistry*," 2nd edition, Vol. 5 (International Series of Monographs in Organic Chemistry), Pergamon Press, Oxford, 1969.
19. J. B. Stothers, "*Carbon-13 NMR Spectroscopy*," Academic Press, New York, 1972.
20. R. M. Silverstein, F. X. Webster, and D. J. Kiemle, "*Spectrometric Identification of Organic Compounds*," 7th edition, John Wiley & Sons, Inc., Hoboken, NJ, 2005.
21. J. B. Lambert, H. F. Shurvell, D. A. Lightner, and R. G. Cooks, "*Organic Structural Spectroscopy*," Prentice Hall, Upper Saddle River, NJ, 1998.

22. P. Crews, J. Rodriguez, and M. Jaspars, "*Organic Structure Analysis*," 2nd edition, Oxford University Press, New York, 2010.
23. C. J. Pouchert and J. Behnke, "*The Aldrich Library of* ^{13}*C and* ^{1}*H FT NMR Spectra*," 1st edition, Vol. 1–3, Aldrich Chemical Company, Inc., Milwaukee, WI, 1993.
24. M. Foroozandeh, R. W. Adams, N. J. Meharry, D. Jeannerat, M. Nilsson, and G. Morris, *Angew. Chem. Int. Ed.* 2014, 53, 6990–6992.
25. K. R. Metz and L. K. Dunphy, *J. Lipid Res.* 1996, 37, 2251–2265.
26. K. R. Metz and L. K. Dunphy, *J. Lipid Res.* 1996, 38, 1275.
27. R. K. Harris, "*Nuclear Magnetic Resonance Spectroscopy*," Pitman Books Limited, London, 1983, pp. 121–126.
28. L. T. Scott, M. M. Hashemi, and M. S. Bratcher, *J. Am. Chem. Soc.* 1992, 114, 1920–1921.
29. A. L. Van Geet, *Anal. Chem.* 1968, 40, 2227–2229.
30. W. Sikorski, A. W. Sanders, and H. J. Reich, *Magn. Reson. Chem.* 1998, 36, S118–S124.

71

NEAR-INFRARED SPECTROSCOPY AND ITS ROLE IN SCIENTIFIC AND ENGINEERING APPLICATIONS

Brad Swarbrick
Quality by Design Consultancy, Sydney, New South Wales, Australia

71.1 INTRODUCTION TO NEAR-INFRARED SPECTROSCOPY AND HISTORICAL PERSPECTIVES

71.1.1 A Brief Overview of Near-Infrared Spectroscopy and Its Usage

Until recently, the near-infrared (NIR) region of the electromagnetic spectrum was not discussed in too great detail in undergraduate university courses. This was primarily due to NIR spectra not containing the sharp, well-defined absorbance bands, typical of methods such as the mid-infrared (MIR) or nuclear magnetic resonance (NMR) spectroscopy, therefore making identification of functional groups difficult. However, during the past 20-year period, the availability of high power computers and the development of mathematical methods such as chemometrics (discussed in detail in Chapter 65) has helped scientists and engineers utilize this most informative region of the spectrum.

The NIR region lies between the visible and MIR regions (i.e., between 700 and 2500 nm) and is associated with bond stretching phenomena between molecules with large dipole moments, in particular, the stretching frequencies associated with molecules containing carbon—hydrogen (C—H), oxygen—hydrogen (O—H), and nitrogen—hydrogen (N—H). The specificity for these large dipole moment bonds has allowed the NIR method to be used in a number of industrial applications for the

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

prediction of quality parameters at the point of manufacture, including sectors such as agricultural, pharmaceutical, and food and beverage, to name just a few.

The main benefits of NIR spectroscopy lie in the high signal-to-noise (S/N) of the instrumentation and the minimization of sample preparation required to collect reliable spectra. Industries, such as the agricultural sector, have used the technique for many years to measure constituents such as protein, moisture, starch, and oil in grains and oilseeds when they are delivered to grain handlers during harvest (refer to Section 71.6.1 for more details). In these applications whole grains are loaded into the spectrometer and scanned "as is," yielding in one measurement as many as 10 or more constituents, using chemometric models. Of even greater value to the end user has been the implementation of NIR to pharmaceutical applications (Section 71.6.2). Whole pharmaceutical tablets can be analyzed nondestructively for the quantification of active ingredient content, again at the point of manufacture, thus bringing the laboratory to the process operator and providing greater assurance of quality through an entire manufacturing run. A much greater discussion of the many applications of the NIR method is presented in Section 71.6 of this chapter.

NIR spectrometers are also highly versatile. They can be installed in environments such as clean laboratories, pharmaceutical manufacturing plants, right through to harsh conditions found in feed mills, and petrochemical plants. The instrumentation can typically be used to standalone besides the process (at line) or can be integrated directly to the process through the use of specialized interfaces and fiber optic cables (in line). Recently, a number of manufacturers have developed so-called microspectrometers, which can be adapted to the most difficult of sampling situations. The various technologies used for developing NIR spectrometers are discussed in more detail in Section 71.3 of this chapter.

As the challenges of manufacturing high-quality products at competitive pricing continue into the future for all industrial sectors, the NIR technique has found and will continue to find many more applications for meeting such challenges. NIR can now be considered to be a mature technology and has widespread acceptance as an alternative testing method to establish reference methods, for example, the British, European, and US Pharmacopoeias have dedicated chapters to the NIR method for use in the pharmaceutical and biopharmaceutical sectors. Many groups such as the American Society for Testing and Materials (ASTM) and the US Food and Drug Administration (US FDA) have provided guidance on how to implement NIR into a number of industrial situations [1, 2].

NIR is also highly suited to research and development applications. The method can be highly sensitive to both constituents of interest and the entire sample matrix. Many research groups have used the NIR technique for classifying existing samples into known classes and also isolating new classes, particularly in biological and ecological research [3]. This is again attributed to the ability of NIR to measure the sample, as it exists in nature, without the need for sample preparation. Heterogeneity of natural samples is a typical challenge for the NIR scientist to address; however, with the portability of some instrumentation, the use of fiber optic probes, or using some smart sampling accessories, multiple spectra from a single sample can be collected rapidly, averaged, and used as a composite spectrum that is representative of the entire sample, even in the presence of heterogeneity.

One of the major hurdles for a new practitioner to the NIR technique is the development of robust calibration models. Calibration models are typically developed using multivariate analysis (MVA), also known as chemometric techniques. There are no real shortcuts to model development, and this requires excellent subject matter knowledge of the application at hand and a good working knowledge of the methods used for sampling and model development. This is covered in more detail in Section 71.5 of this chapter and a dedicated chapter on chemometrics (see Chapter 65). The good news is that once a calibration model has been developed, the end user of the method does not need to have a working knowledge of chemometrics. As long as they follow standard procedures of sampling and instrument usage, nonskilled workers can use the technology in their day-to-day tasks without the supervision of an expert.

The NIR method is highly sensitive to moisture content in samples, and in some cases, when the moisture content is too high, the detector can saturate quickly. NIR is therefore not a suitable method for analyzing highly aqueous solutions, although there are some applications where this is possible [4]. Methods such as Raman spectroscopy may be better suited for these applications, but for such a small limitation, the NIR technique is highly versatile for the many other applications that exist.

This chapter provides a concise overview of the NIR methods from basic theory right through to end user applications and their development. It should be used as a first reference point, and the literature cited in this chapter should be used if more detailed explanations are required.

71.1.2 A Short History of NIR

The discovery of the NIR region of the electromagnetic spectrum has been attributed to Frederick William Herschel [5]. Using a simple apparatus consisting of a prism and a thermometer, Herschel was performing experiments to disperse sunlight into its component colors in the visible region of the spectrum. He observed that toward the red part of the visible region, the temperature of the dispersed radiation increased, and just beyond the visible region, where the radiation became invisible, the temperature reached its maximum. Herschel attributed his finding to be a thermal band beyond red, which is termed infrared (IR) from the Latin word “infra,” and considered this region to be different from light [6]. This principle is shown diagrammatically in Figure 71.1.

It was not until Ampere’s experiment using a newly developed thermocouple that NIR was determined to have similar optical characteristics to that of visible radiation [7].

Sometime later, Maxwell, Planck, and some of the great scientists of the nineteenth and twentieth centuries developed theories that better understood the nature of the electromagnetic spectrum, and today, we understand the NIR region to lie between

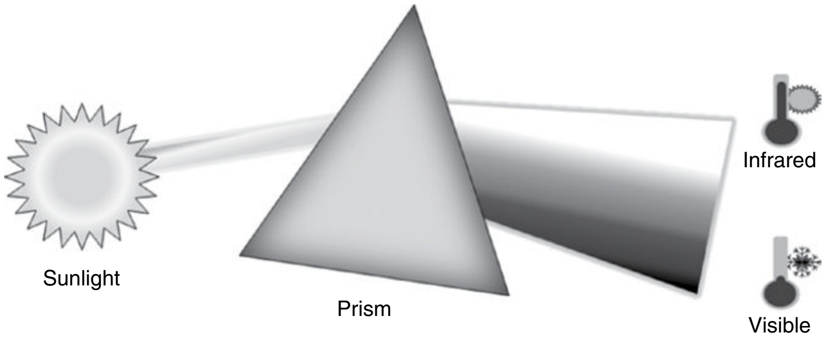


FIGURE 71.1 Dispersion of polychromatic light into its components using a prism. Toward the higher wavelength region of the visible spectrum and into the NIR region, the temperature of the radiation increases.

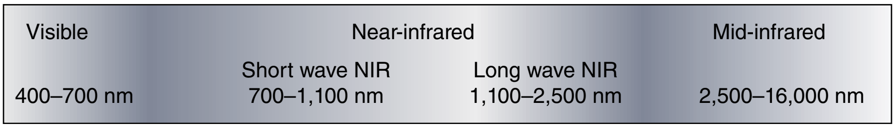


FIGURE 71.2 The NIR region of the electromagnetic spectrum with respect to the visible and MIR regions.

the visible and MIR regions [8]. Figure 71.2 provides a diagram of where the NIR region of the spectrum lies with respect to the visible and the MIR region.

Coblentz [9] developed a primitive spectrometer for measuring the NIR spectrum of a number of materials and thus provided a means for chemists to elucidate the structure of compounds. For a more detailed discussion of the early history of NIR and the first instrumentation used to collect IR spectra, the interested reader is directed to the work of Burns and Ciurczak [6].

For a long period of time, the use of NIR remained relatively limited with few applications being published. It was not until the 1960s that the pioneering works of Ben-Gera and Norris of the US Department of Agriculture (USDA) paved the way for the modern success of NIR [10]. This initial work focused on the analysis of fat and moisture content in agricultural products and is still cited as one of the most influential references in NIR applications [11].

Agricultural and food applications dominated the NIR literature throughout the 1970s and 1980s with workers such as Williams [12], Osborne [13], Shenk [14], Flinn [15], Blakeney [16], and Batten [17] all making significant contributions. Agricultural applications of NIR are discussed in more detail in Section 71.6.1.

Up until the late 1990s, instrumentation remained the key limitation for the widespread use of NIR in the process industries, particularly for real-time monitoring. Swarbrick [18] reviews the evolution of instrumentation since 2000 and discusses how these advances in instrument speed and portability have allowed NIR to be utilized in process applications, particularly in the pharmaceutical sector. Section 71.3 discusses the various instrumentation types available for generating NIR spectra.

In the pharmaceutical sector, the work by Ciurczak and Drennen [6], Mark [19], and Ritchie [20, 21] has paved the way for the modern usage of NIR for applications such as raw material identification, intact tablet analysis, monitoring of drying operations, and recently the monitoring of bioprocesses. A more in-depth discussion of NIR applied to pharmaceutical and biotechnology applications is provided in Section 71.6.2.

Other major applications of the NIR method can be found in industries such as petrochemical, wool, wood, dairy, and textile industries. As instrumentation and education in chemometrics improve over the coming years, there will be many more applications of NIR published due to its versatility and ruggedness. This chapter can only provide a brief outline of the myriad of applications where NIR is currently used, and the interested reader is encouraged to investigate the many forums that exist in the NIR community and to attend the two major conferences dedicated to the use and promotion of the NIR method.

In particular, the International Conference on NIR Spectroscopy (ICNIRS) [22] is a biannual meeting held globally that brings many of the leading practitioners together to present their work and discuss their ideas. During this meeting, the leading researcher in NIR spectroscopy (as nominated by their peers) is presented with the Tomas Hirschfeld award. Tomas Hirschfeld was an internationally recognized analytical scientist and inventor with over 100 patents to his name [23]. Tomas envisioned that research into microsensors and microinstruments would be the future and very rewarding to those who invested into this area [24]. After his unprecedented death in 1986, the ICNIRS inaugurated the Tomas Hirschfeld award for recognition of a significant contribution to the science of NIR spectroscopy, including research and development of new technology. Tomas' legacy is still recognized to this day as a key contributor to this important area of science and technology.

Every second year, in between the ICNIRS meetings, the Australian Near Infrared Spectroscopy Group (ANISG) [25] is held in Australia and New Zealand. A smaller conference with respect to the ICNIRS, the ANISG attracts a quality audience that come together to share ideas and stories regarding the development and application of NIR spectroscopy in a wide range of applications.

Overall, it is again stated that NIR is a mature technology, utilized in many research and industrial applications. The advancements in instrumentation, particularly since 2000, have enabled this technology to be implemented into many business critical and novel applications. The next section provides details on the theory behind the NIR method and how its characteristics allow it to have the versatility it has as an analytical, classification, and process monitoring tool.

71.2 THE THEORY BEHIND NIR SPECTROSCOPY

71.2.1 IR Radiation

The IR region of the electromagnetic spectrum is located between the visible and microwave regions (700–111,000 nm; refer to Fig. 71.2). It is associated with stretching and bending modes that occur prevalently within covalent bonds of organic molecules; however this definition extends to other chemical and physical bond types, including hydrogen bonding and other molecular interactions.

The IR region of the spectrum can be further broken down into three subregions:

1. The far infrared (FIR): This region lies between 16,000 and 111,000 nm primarily used for rotational spectroscopy and is associated with the measurement of inorganic materials and applications in astrophysics.
2. The MIR: This region traditionally has been used due to the large amount of chemical information related to molecular structure present in the sharp spectral features produced. This region lies between 2,500 and 16,000 nm.
3. The NIR: This region lies between the visible and MIR regions (700–2500 nm) and is associated with combination bands and overtones generated in the MIR region. Unlike the MIR, the spectral features of a typical NIR spectrum are broad and show much band overlap and require methods such as chemometrics to extract important information.

This chapter is concerned with the NIR region of the spectrum, and a more detailed description of the theory behind its interaction with matter is provided in the following sections.

71.2.2 The Mechanism of Interaction of NIR Radiation with Matter

When MIR radiation interacts with matter, depending on the frequencies, various molecular stretching and bending motions are induced in a molecule. The intensity of the spectral bands generated is a function of both the intensity of the incident radiation and the molar absorptivity of the chemical bond being excited. These excitations are known as the fundamental frequencies.

71.2.2.1 Overtone Frequencies Using the analogy of ringing a bell, when the bell is initially struck with a hammer (or other devices), the first sounds heard are loud and highly distinct in their audible frequencies. These are the fundamental notes of the bell sound. As the bell is left to vibrate, the sound intensity decreases rapidly, and when listened to closely, oscillating sounds can be heard after the fundamentals have decayed. These are known as the overtones of the fundamental frequencies and typically occur at integer values of the fundamental, that is, if the fundamental frequency occurs at a

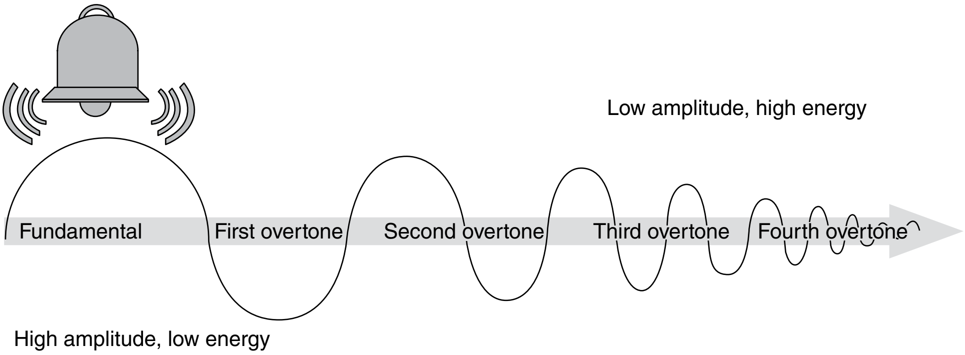


FIGURE 71.3 Analogy between ringing a bell and the overtones generated in the NIR region of the electromagnetic spectrum.

value *f*, theoretically, the first overtone will occur at 2*f*, the second overtone at 3*f*, etc. Figure 71.3 provides a diagrammatic representation of the fundamental and overtone frequencies using the bell analogy related to the NIR spectrum.

The earlier analogy of ringing a bell translates across to the MIR–NIR spectrum. When a molecular bond absorbs energy at a particular frequency and a bond stretching mode is induced in the molecule at fundamental frequency *f*, then at approximately 2*f*, the first overtone band of the fundamental will occur in the NIR region of the spectrum. As with the case of the bell ringing, the overtone frequency has an intensity of approximately an order of magnitude less than the fundamental. Successive overtones diminish in intensity by an order of magnitude from the last overtone.

In general, these overtone frequencies have a low molar absorptivity coefficient (ε) with respect to the fundamental frequency, and this property lends itself extremely useful in practical implementations of NIR spectroscopy (see Section 71.6).

71.2.2.2 Combination Frequencies In general, stretching frequencies contribute mainly to overtone frequencies in NIR; however, there are some strong bending vibrations that also contribute [8]. The region of the NIR spectrum between 2000 and 2500 nm is commonly known as the first overtone and combination band region. Combinations occur when the overtones generated in the MIR combine to form bands of higher intensity than would occur from the overtone alone, that is, they arise from the sharing of NIR energy between two and more fundamental absorptions.

The region of the NIR spectrum between 1100 and 2000 nm is typically where the second overtones are located. There are also some combination bands that are strong enough in intensity to generate bands in this region. In the region between 700 and 1100 nm, this is where third and fourth overtones occur. Their intensity is typically too small to reveal any combination bands of any practical use. Figure 71.8 in Section 71.2.3.2 provides an overview of the NIR region of the electromagnetic spectrum showing where combinations and overtones occur and also how the molar absorptivity varies across this region of the spectrum.

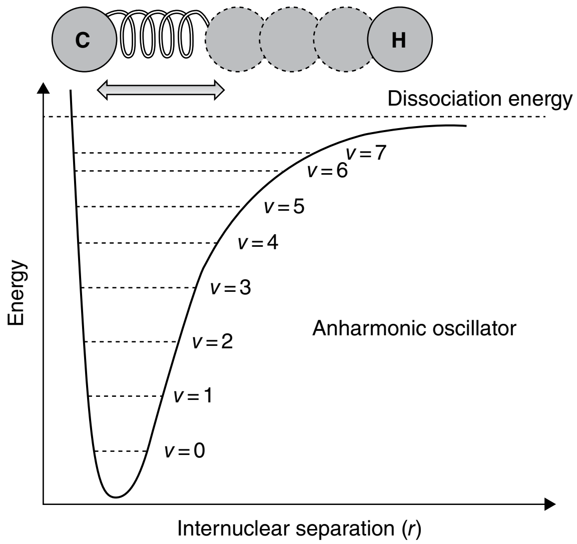


FIGURE 71.4 Description of molecular vibrations in terms of the anharmonic oscillator model.

71.2.2.3 Dipole Moments The atoms in a molecule can be considered as being held together by weak springs [8]. At their ground state, the molecules will naturally vibrate, and as more energy is applied (i.e., in the form of NIR energy in this case), the molecules will vibrate with a greater frequency. The energy levels of the vibrational states allowed are governed by the rules of quantum mechanics and in particular can be described as an anharmonic oscillator. Figure 71.4 shows this concept diagrammatically.

In the case of a two-atom molecule, only stretching of the bond between them can occur; for three or more atoms, bending of the bonds can also occur. Since the NIR region measures the combinations and fundamentals arising from the MIR region of the spectrum, the magnitude of the bands observed are 1–3 orders of magnitude lower than the fundamental bands. For molecular bonds containing C—H, O—H, and N—H bonds, there is a “large” difference in the atomic masses of carbon, oxygen, and nitrogen with respect to hydrogen. When NIR radiation is applied to molecules containing these functional groups, this sets up large dipole moment changes in the molecules, and the energy absorbed by them results in the bands observed in an NIR spectrum. Figure 71.5 provides a description of a dipole moment for the C—H bond.

Since the functional groups have different affinities (attraction) to each other, the dipole moments vary in magnitude. This fact allows for a distinction of bond type in the IR region of the spectrum, and this observation was first made by Coblentz [9] in his pioneering work on structural elucidation of molecules. For a more theoretical discussion on the theory behind NIR, the interested reader is referred to the literature [26].

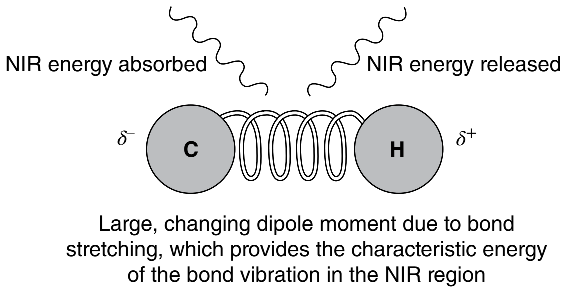

NIR energy absorbed
NIR energy released
δ−
C
H
δ+
Large, changing dipole moment due to bond stretching, which provides the characteristic energy of the bond vibration in the NIR region

FIGURE 71.5 Changes in dipole moments between two atoms of largely differing atomic mass result in the bands typically observed in the NIR region of the electromagnetic spectrum.

71.2.3 Absorbance Spectra

71.2.3.1 Calculating an Absorbance Spectrum

A brief overview of the theory behind NIR was presented in the previous sections. This section puts the theory into practice by describing how an NIR spectrum is collected. Section 71.3 describes the instrumentation used to collect NIR spectra, and Section 71.4 discusses the sampling methods commonly employed. An absorbance spectrum is calculated as the negative logarithm of the reflectance or transmittance of NIR light from a sample. The concepts of reflectance and transmission are discussed in more detail in Sections 71.4.1 and 71.4.2. The formal calculation of absorbance is provided as follows:

$$A_i = -\log\left(\frac{1}{X_i}\right)$$

where

A_i = absorbance calculated for the i^{th} wavelength of the spectrum.
X_i = the reflectance or transmittance of a sample at the i^{th} wavelength.

Absorbance is measured on a unitless scale and is represented by the symbol AU (absorbance units). Since the scale is logarithmic, each AU represents an order of magnitude less light intensity than the incident light source. Figure 71.6 shows this principle diagrammatically.

For example, an absorbance value of 1 represents 10 times less light incident on the detector compared to the original light source, and an absorbance value of 5 represents 100,000 times less light detected compared to the original light source. A major advantage of NIR spectrometers is the detection systems used have high S/N ratios and in some cases can generate reliable spectrum up to 6 AU. For more details of the detectors used in NIR spectroscopy, refer to Section 71.3.9.

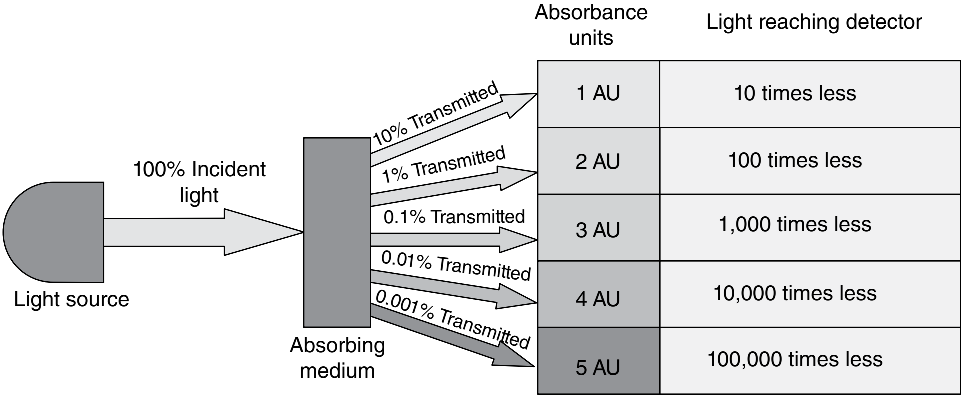


FIGURE 71.6 The relationship between light absorbed by a sample and the light intensity that reaches a detector.

A general process for collecting a spectrum on an NIR instrument typically proceeds as follows:

1. Collect a dark current (DC) spectrum. This is usually collected where the light source of the instrument is turned off and the electronic noise of the detector is measured for all wavelengths.
2. Collect a reference (Ref) spectrum. In transmission mode (Section 71.4.1) this is the signal from the light source unhindered by any sample, and for reflectance mode, the light source is reflected of a standard highly reflective material measured for all wavelengths (typically a material called Spectralon®).
3. Collect a sample (Sam) spectrum by placing the sample in front of the light source and collecting the light either transmitted through or reflected off the sample.

The raw reflectance or transmittance scan is calculated as follows:

$$X_i = \frac{(\mathrm{Sam}_i - \mathrm{DC}_i)}{(\mathrm{Ref}_i - \mathrm{DC}_i)}$$

In the case of transmission, $X_i = T_i$ where T_i is the transmittance at the ith wavelength, and for the case of reflectance, $X_i = R_i$ where R_i is the reflectance at the ith wavelength. Figure 71.7 provides examples of spectra collected in transmission and diffuse reflectance (Section 71.4.2) mode.

71.2.3.2 Characteristics of NIR Spectra The NIR region of the electromagnetic spectrum is split into two general subregions:

1. The short wave (SW) region (also known as the Herschel region) spans the wavelength range of 700–1000 nm [27]. This region is predominantly used for diffuse transmission measurements.

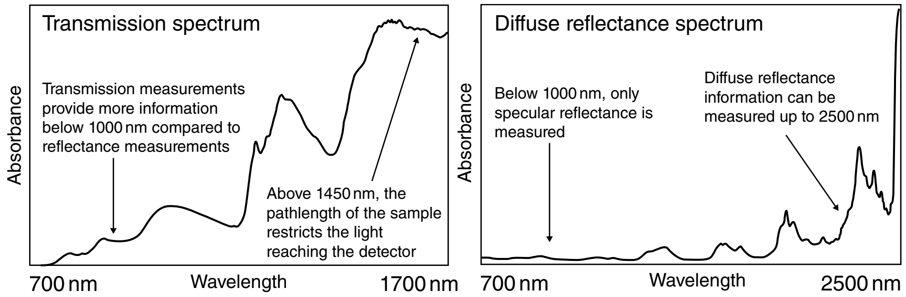


FIGURE 71.7 Examples of NIR spectra collected in transmission and diffuse reflectance modes.

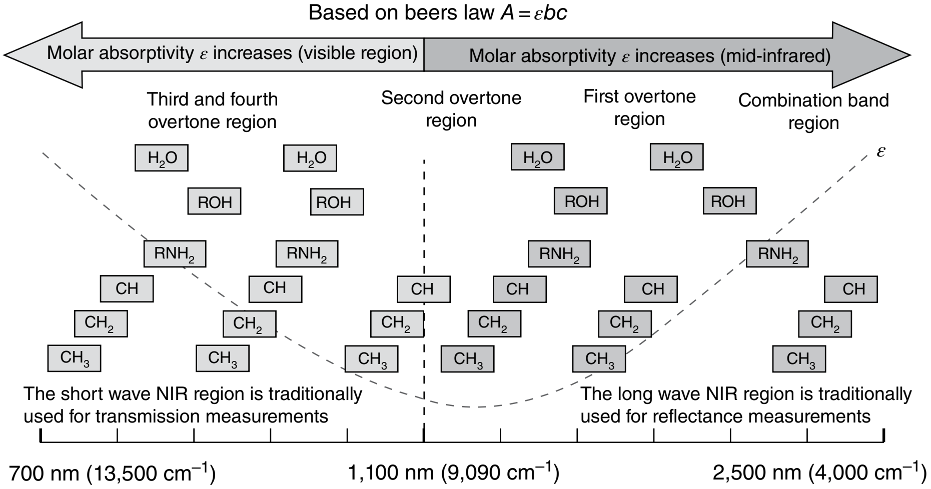

Based on beers law $A = \varepsilon bc$
Molar absorptivity ε increases (visible region)
Molar absorptivity ε increases (mid-infrared)
Third and fourth overtone region
Second overtone region
First overtone region
Combination band region
ε
H_2O
ROH
RNH_2
CH
CH_2
CH_3
The short wave NIR region is traditionally used for transmission measurements
The long wave NIR region is traditionally used for reflectance measurements
700 nm (13,500 cm^{-1})
1,100 nm (9,090 cm^{-1})
2,500 nm (4,000 cm^{-1})

FIGURE 71.8 Characteristics and chemical information available in the NIR region of the electromagnetic spectrum.

2. The long wave (LW) region spans the region from 1000 to 2500 nm. This region is predominantly used for diffuse reflectance measurements.

Figure 71.8 shows the various regions of the NIR spectrum.

In particular, the region between 850 and 1100 nm has a very low molar absorptivity coefficient (ε) associated with it. Based on the Beer–Lambert law, this means that more light can be passed through samples in this region, and this is why this region is ideal for transmission measurements. This region contains the third and fourth overtones, and although the signals generated in this region are 2–3 orders of magnitude smaller than the fundamental that they arise from, the combination of high pathlength and low molar absorptivity means that light can be passed through pathlengths up to 30 mm thick to generate high-quality spectra.

Between 1000 and 2500 nm this is where the LW region of the spectrum occurs. Up to 1500 nm transmission measurements are still possible; however, beyond 1500 nm, two mechanisms combine that make transmission measurements difficult:

1. The wavelength of the radiation becomes similar in size to the particles of the sample being measured.
2. The molar absorptivity coefficient increases rapidly as the MIR region is approached.

According to the Beer–Lambert law [28], absorbance is related to the concentration of active sample constituent via the relationship

$$A = \varepsilon bc$$

where

A = absorbance

ε = molar absorptivity coefficient

b = pathlength that the radiation has to pass through

c = concentration of the constituent in the sample

In clear, nonscattering liquid samples, the pathlength b of the sample can be controlled by fixing it with a cuvette. According to the Beer–Lambert law, if the pathlength is fixed and the measurement is performed at a single wavelength, absorbance is proportional to concentration; therefore a linear calibration model can be developed where concentration can be predicted from the sample absorbance (provided the absorbances lie in the range 0–1).

When this principle is extended to the LW NIR region, the fact that the wavelength of the radiation is similar in size to the particles being measured means that its interaction with the particles is more elastic than specular in nature. This phenomenon is known as diffuse reflectance. In both diffuse reflectance and diffuse transmission measurements, the pathlength that the radiation has to pass through is not constant and cannot be made constant due to the different ways solid samples pack each time they are prepared. As a result, typical diffuse reflectance NIR spectra have a quadratic baseline beyond 1500 nm (refer to Fig. 71.7).

In the section on preprocessing (Section 71.5), it is explained how additive and scatter correction algorithms can be used to correct for these spectral features caused by physical effects in the sample.

Overall, NIR spectra contain broad spectral features arising from overtone and combination bands that are highly overlapping. Until the rise of the personal computer and chemometric methods, the NIR region of the spectrum was thought (and taught) to contain little information useful for structural elucidation or any other application. Today, the number of applications is growing because of the development of fast and reliable instrumentation. The next section provides an overview of the instrumentation available for collecting NIR spectra and their modes of operation.

71.3 INSTRUMENTATION FOR NIR SPECTROSCOPY

This section provides a brief overview of the NIR instrumentation currently available at the time of writing of this chapter and is not meant to be an exhaustive description of each instrument type. The interested reader is referred to the referenced literature for more information regarding the theory and construction of each instrument type discussed.

71.3.1 General Configuration of Instrumentation

The development of instrumentation for NIR spectroscopic applications has seen a massive growth and improvement since the mid 1990s. The first instruments were based on specific filters (typically 10–20 individual wavelengths), which were chosen for a particular application; however, today's modern spectrometers can generate thousands of data points per spectrum. The following sections provide a discussion of the instruments currently available and their applicability to various applications.

In general, all spectrometers consist of the following components:

1. A suitable light source (typically a tungsten halogen lamp optimized for the IR region)
2. A suitable sampling device (see Section 71.4)
3. A light dispersion device (monochromator)
4. A detector

Figure 71.9 shows a generic spectrometer setup used in most instruments.

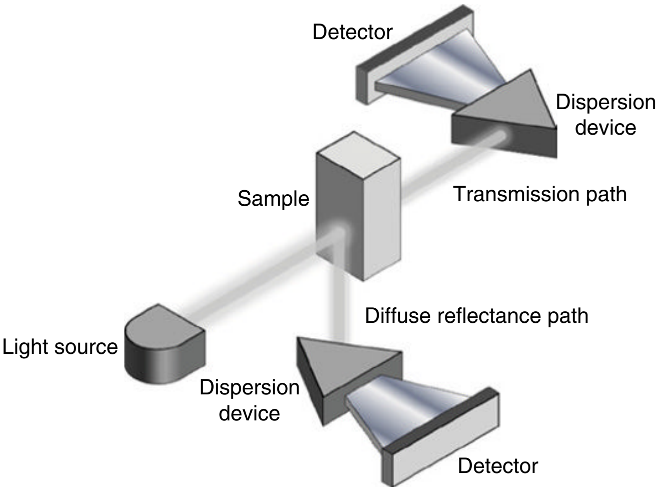


FIGURE 71.9 Generic configuration of a spectrometer used to collect NIR spectra.

71.3.1.1 Diffuse Reflectance Configuration Many of the modern NIR instruments today are configured to collect spectra in diffuse reflectance mode. Figure 71.10 provides a general layout of a typical diffuse reflectance setup.

In any spectrometer configuration, there are two main ways in which the instrument works:

1. Predispersive instruments place the monochromator just after the light source and pass the dispersed light through the sample. These systems have a single, larger surface area detector that synchronizes with the dispersion system such that a spectrum can be generated over many wavelengths.
2. Postdispersive instruments pass the light from the source through the sample and then send the reflected/transmitted light typically to a stationary monochromator, which disperses the light over a diode array detector, such that the entire spectrum is generated in one operation.

71.3.1.2 Transmission Configuration Transmission instruments are typically designed for two main applications in NIR:

1. Diffuse transmission, where an intense light source is passed through a solid/semisolid medium, and what light emerges from the sample is used to generate a spectrum.
2. Normal transmission, where the spectrometer acts as a typical instrument for collecting spectra through clear, nonturbid liquids (typically meeting the requirements of the Beer–Lambert law).

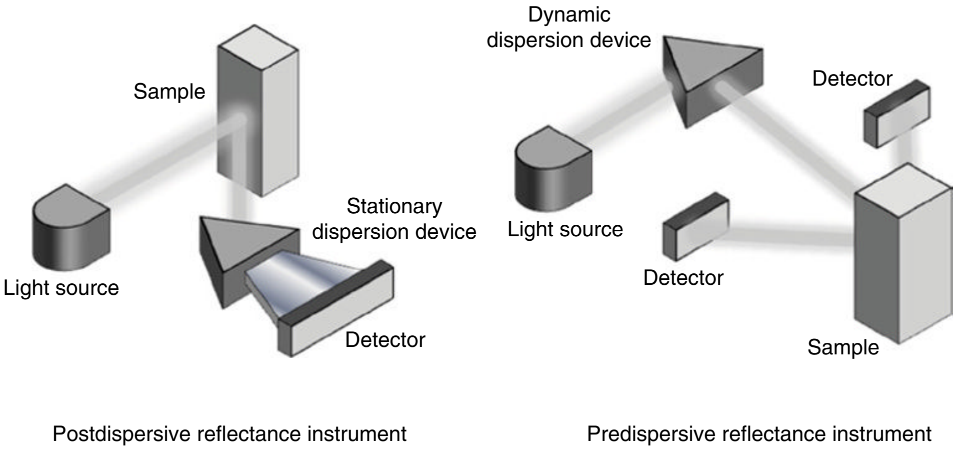


FIGURE 71.10 General NIR instrument configuration for collecting spectra in diffuse reflectance mode.

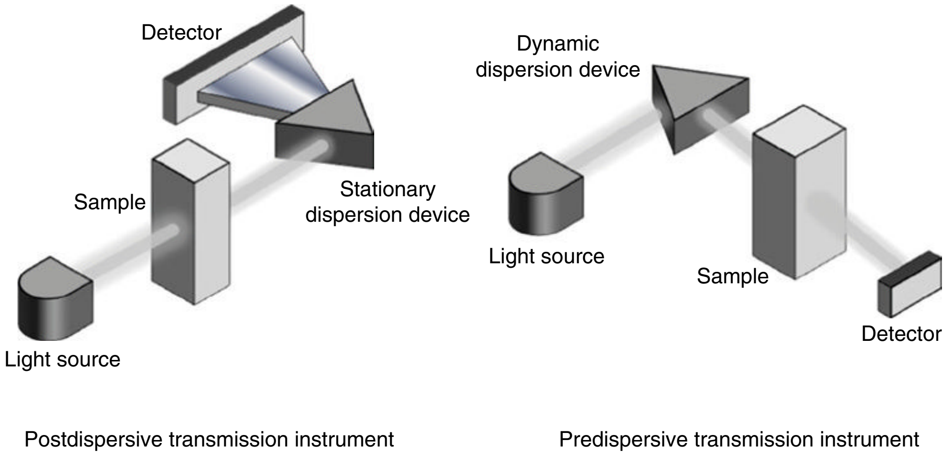

Detector
Sample
Stationary dispersion device
Light source
Postdispersive transmission instrument
Dynamic dispersion device
Light source
Sample
Detector
Predispersive transmission instrument

FIGURE 71.11 General NIR instrument configuration for collecting spectra in transmission mode.

Figure 71.11 provides a general layout of a typical transmission setup.

Although there are other sampling techniques available (such as interactance probes; Section 71.4.4.1), in general, these techniques are a variant of the two modes of collection described previously. These configurations are typically used with the instrument types described in the following sections.

71.3.2 Filter-Based Instruments

Figure 71.12 provides a schematic view of how a typical filter-based instrument works.

Light from the source is incident on a spinning filter wheel, which dispersed the light into the component wavelengths defined by the individual filters. A chopper was typically implemented to separate the wavelengths detected after the NIR radiation was passed through (or reflected off) a sample. In a filter instrument, a continuous spectrum cannot be generated; only a set of absorbance values are generated when the sample signal is divided by the reference material and the logarithm of the transmission/reflectance value is taken (see Section 71.2.3.1 for details on the calculation of an absorbance spectrum).

Filter instruments typically consist of a filter wheel with anywhere between 5 and 30 individual filters that represent different wavelengths in the NIR region. The main reason why filters were used in earlier instrumentation was the limitation that mathematical tools and fast computers are able to handle the data generated by modern spectrometers. The main method of analysis of the data generated by a filter instrument was multiple linear regression (MLR), discussed in Chapter 65.

The analyst would take the data collected on a predefined set of samples and use either a step-forward analysis, where wavelengths (filters) were added to a model to see if it improved the fit, or a step-backward approach, where all wavelengths (filters) are added and successively removed from the analysis until an optimal model could be generated.

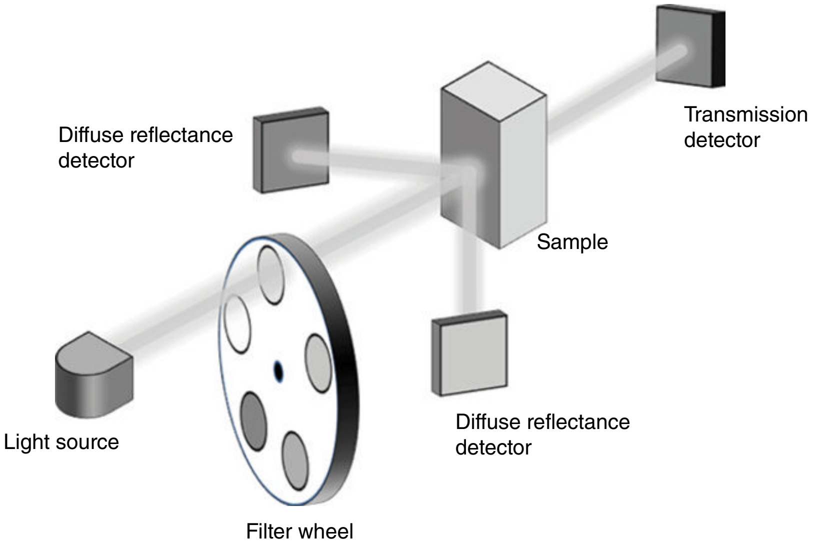


FIGURE 71.12 Filter-based NIR spectrometer.

Filter instruments still find practical use today as moisture analyzers in agricultural and pharmaceutical applications; however, their use is limited as newer and faster (and cheaper) instruments provide more versatility.

71.3.3 Holographic Grating-Based Instruments

A holographic grating for NIR applications is typically a highly polished concave glass surface with a thin coating of gold and blazed with a series of parallel lines that form a diffraction grating. Light incident onto the grating may be used in two ways:

1. Moving grating system, where the holographic grating is mounted onto a precise stepper motor (known as an encoder), and as the grating sweeps through an arc, defined by the encoder, the incident light is dispersed into its component wavelengths.
2. Stationary gratings are fixed in a predefined position such that the light is dispersed over an array detector for a specified wavelength region. This type of instrument is discussed in more detail in Section 71.3.4.

The moving grating system will be discussed in more detail in this section. These instruments can be configured to be either pre- or postdispersive, and as the grating moves through a single arc, the entire wavelength range of the spectrum is generated. Each component wavelength is passed onto the sample, and the detector is used to measure the signal. In order to generate a spectrum, the encoder is used to synchronize the wavelength measured to the detector signal.

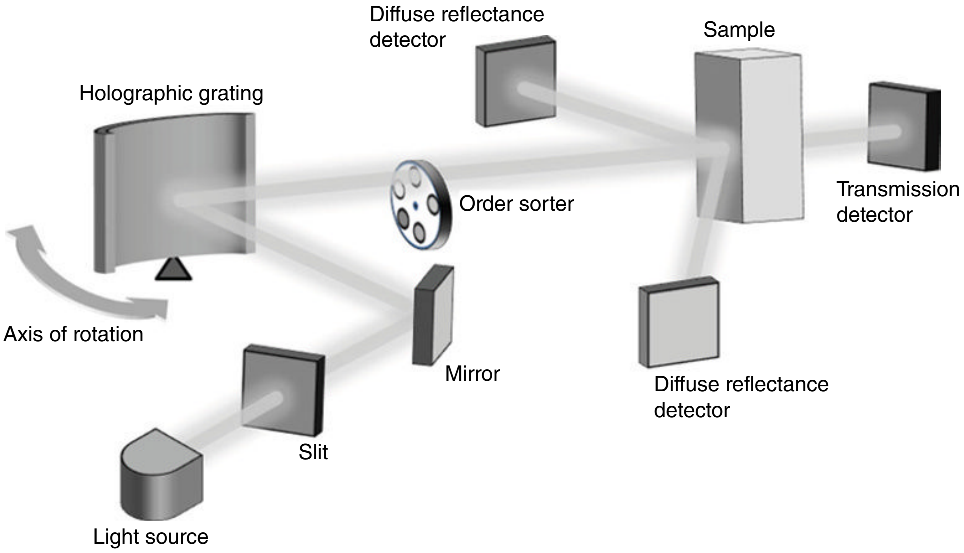


FIGURE 71.13 General instrument configuration of a predispersive holographic grating-based NIR spectrometer.

Holographic gratings also generate higher- and lower-order diffraction patterns [29] that must be eliminated in order to generate as pure a spectrum as possible. These instruments are also equipped with mechanical order sorters, which work in synchronization with the encoder and the detector. Thus, the holographic grating instruments can result in a highly mechanical system, which, when any service or repairs need to be performed, may lead to intensive recalibration of the instrument.

Figure 71.13 provides a schematic view of a typical predispersive grating-based spectrometer.

The spectral resolution of this spectrometer type is typically determined by the slit width of the instrument. There is a trade-off to be made here; if the slit width is too large, the resolution of the instrument decreases; however, more light can be passed onto the detector, thus increasing the S/N ratio. Conversely, if the slit width is made too small, very little light can reach the detector (particularly when diffuse transmission is used); therefore the S/N ratio is decreased; however, the resolution is increased.

Typical grating-based instruments use a 10 nm bandpass for collecting a spectrum. This means that the true resolution of the instrument can discriminate between spectral features separated by 10 nm distances. Spectral resolution must be distinguished from point to point resolution. Most spectrometers will generate spectra with anywhere from 5 to 0.5 nm point to point spacing. This is because the instrument internal software will interpolate in between the spectral resolutions of the instrument to generate a smoother spectrum.

This principle of optical efficiency is also applicable to diode array-based systems using a stationary grating as the slit width also determines the resolution. Also in this case, the slit width is typically chosen to match the pitch of the detector spacing in the diode array detector.

In general, moving grating instruments are slow with respect to modern spectrometer setups that utilize no moving parts, and reliable spectrum generation usually requires the moving grating to complete 32 passes (coadds) to generate a single spectrum. Their mechanical nature requires them to be housed in rugged and environmentally controlled housings to minimize mechanical shock and temperature fluctuations. For this reason, this type of instrument finds limited use in real-time monitoring applications; however, for general laboratory and research applications, these instruments find many applications.

71.3.4 Stationary Spectrographic Instruments

In order to overcome the mechanical limitations of the moving grating instruments discussed in Section 71.3.3, a postdispersive configuration of the holographic grating monochromator can be used. Figure 71.14 shows a schematic of the general configuration of a fixed spectrograph instrument.

In this instrument type, incident light is passed through or reflected off a sample and passed onto the fixed monochromator. The position of the monochromator is fixed such that is disperses the incident light onto a diode array detector over the desired wavelength region of the spectrum. As with any grating-based instrument, spectral resolution is determined by the slit width. Since the detector is based on a diode array, the

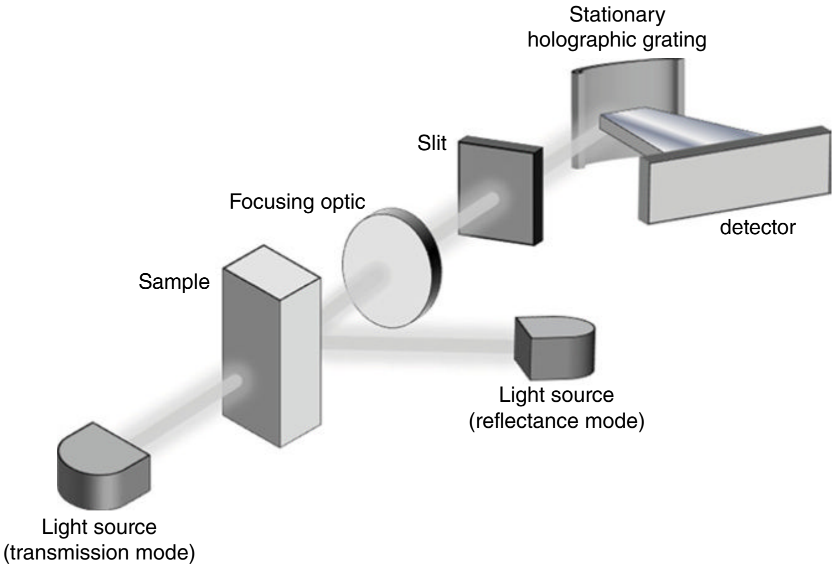

Stationary holographic grating
Slit
Focusing optic
detector
Sample
Light source (reflectance mode)
Light source (transmission mode)

FIGURE 71.14 General instrument configuration of a spectrograph utilizing a diode array detection system.

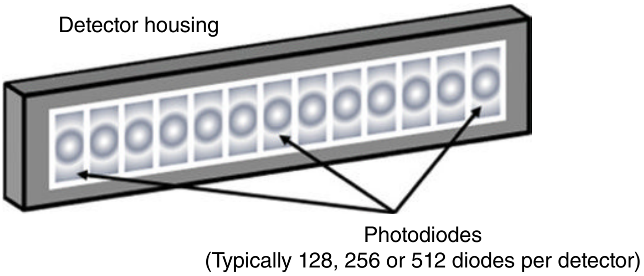


FIGURE 71.15 Diagrammatic representation of a diode array detector.

entire spectrum is collected in one step. This provides a major speed advantage over the moving grating systems. The same S/N issues exist, that is, if the slit width is too small, very little light will reach the monochromator; however, this can be improved by increasing the integration time, that is, the exposure time the light is incident on the detector for collecting a single scan. Therefore, like the moving grating instruments, spectra can be coadded to generate a spectrum with a higher S/N ratio, and since the typical integration time of a diode array detector is in the range of 10–100 ms, a high-quality spectrum can be generated in 3 s or less.

A generic diode array detector is shown diagrammatically in Figure 71.15. There are many different detector types that can be used, and these are discussed in more detail in Section 71.3.9.

Diode array detectors come in sizes as orders of the power of 2, that is, 2^k with the most common pixel numbers being 128, 256, and 512 arrays (although 1024 and 2048 systems do exist). It is a general rule that the pitch between the detectors is matched to the slit width; thus to take advantage of any resolution gains by using larger arrays, the slit width may be too restrictive for collecting high-quality spectra. If the slit and array are not matched, then cross talk between diode array pixels may result, meaning the same information is collected on consecutive pixels. Instrument vendors spend much time and resource into the development of instrumentation in order to minimize such inefficiencies [30].

Overall, stationary gratings coupled with diode array detectors offer a speed advantage over moving grating systems. Since they also contain no moving parts, they are essentially ruggedized for use in process applications. Early instruments of this type suffered from issues arising from overheating and drift; however, today's manufacturing processes and the introduction of more stable detector systems have seen a rise in the use of such systems in real-time applications.

71.3.5 Fourier Transform Instruments

For many years, the moving grating-based instruments dominated the NIR market particularly in agricultural applications and were seen as the gold standard of instrumentation. In the late 1990s–early 2000s the Fourier transform near infrared (FT-NIR)

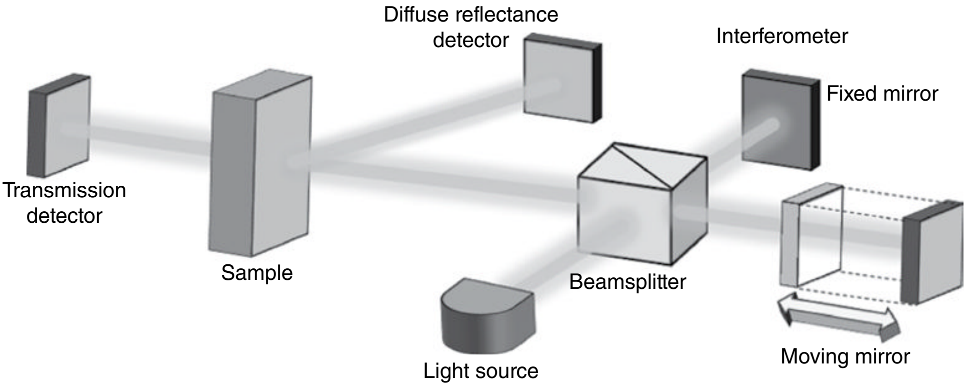


FIGURE 71.16 General instrument configuration of a Fourier transform NIR spectrometer utilizing a Michelson interferometer.

spectrometer was introduced as an alternative to the currently available grating-based instruments.

Fourier transform (FT) instruments are not slit width limited (Section 71.3.3) and therefore provide the advantage of allowing more light to be incident onto a sample or detector (known as the Jacquinot advantage [31]). FT-NIR instruments are typically predispersive, passing the incident light onto an interferometer (in particular, Michelson interferometers [32] are typically used). Figure 71.16 provides a schematic representation of an FT-NIR instrument that utilizes a Michelson interferometer.

The interferometer uses a combination of a fixed and moving mirror system to generate interference patterns. This packet of light patterns is passed onto a sample in either diffuse reflectance or transmission mode. Since the signal is sent to the sample in the time domain, by applying a fast Fourier transform (FFT) [33], the time-domain signal is transformed into a frequency-domain signal, and therefore an NIR spectrum can be generated. This is known as the Fellgett advantage [34]. The theory behind the FT is outside of the scope of this book, and the interested reader is referred to the literature for more details [33].

The accuracy of the wavelength scale is determined by an onboard laser source in the instrument (also known as the Connes advantage [35]). The position and frequency of the laser pulse allow FT-NIR instruments to be set to any spectral resolution desired by the developer. Since the wavelength accuracy is determined by a laser, the resolution of an FT-NIR spectrometer is a true spectral resolution. Typical resolutions that are used in FT-NIR applications range from 8 to 64 cm^{-1} depending on the application. It is noted here that resolution in an FT-NIR spectrometer is linear in the cm^{-1} scale and nonlinear in the nm scale. It is therefore very important to be aware of this fact when comparing results from an FT instrument to a grating-based instrument (which is linear in the nm scale).

In practical terms, the major benefit of the FT systems over the moving grating systems is the minimization of moving parts in the FT system. The so-called resolution advantage is not really an advantage unless a user is interested in water vapor analysis (or similar applications). For most practical industrial applications, using a resolution of 16 cm^{-1} typically results in good calibration models. When the resolution is improved to below 16 cm^{-1}, the size of the spectral files increases and the inherent noise (shot noise) [30] of the spectrum increases since the instrument is using true spectral resolution and the final signal is not interpolated and smoothed. In terms of time benefits, both FT and grating instruments collect spectra at a similar frequency, and FT instruments also require considerable coadding to generate a high-quality spectrum.

In terms of real-time process applications, FT-NIR instruments offer the following major advantages compared to moving grating-based instruments:

1. They are more rugged by design and require less environmental protection.
2. FT instruments do not suffer from Wood's anomaly [36], a spectral inconsistency related to diffraction gratings, and it is known as passing of orders. These result in spectral features that inconsistently arise (particularly when moving samples are measured) and can result in false predictions if the anomaly becomes too intense.
3. Since the FT instruments can be manufactured in a more consistent manner than grating-based instruments, calibration transfer is easier with FT instruments.

Other than the previously mentioned advantages, FT and moving grating instruments typically have similar performance characteristics. Although preferred over moving grating instruments for process applications, FT instruments are still relatively slow compared to more modern NIR spectrometer designs and are now finding many applications in the analysis of processes that slowly evolve over time or for research applications.

At the time of the writing of this chapter, a new generation of micro-FT-NIR spectrometers has been developed [37]. Based on a continuous manufacturing (CM) process, the entire spectrometer including lamps has been integrated onto a single board with reported interferometer frequencies of 400 Hz. It will be interesting to monitor the progress and performance of such instruments in the future.

71.3.6 Acoustooptical Tunable Filter Instruments

Acoustooptical tunable filter (AOTF) spectrometers were first proposed by Harris and Wallace in 1969 [38]. They utilize a crystal of tellurium dioxide (TeO_2) as the light dispersion device, and through the application of specific radio frequencies (RF), the crystal acts like a tunable bandpass filter, therefore the name acoustooptical. Figure 71.17 provides a schematic of a generic AOTF instrument.

When RF waves are passed through the crystal, the lattice successively compresses and relaxes, and the resulting effect is similar to a transmission or Bragg refractor [39], the main difference being that the AOTF device only emits one wavelength band at a

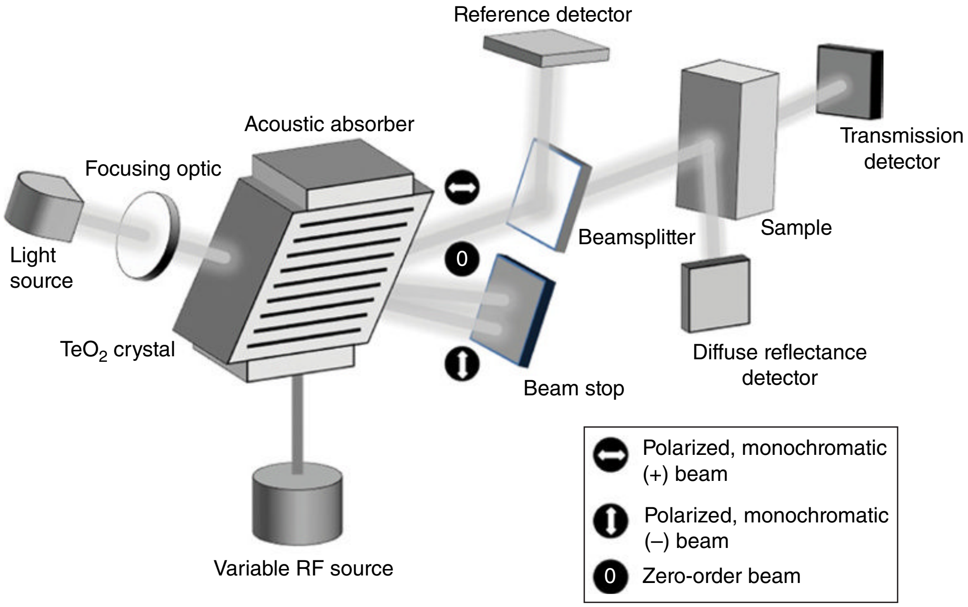


FIGURE 71.17 General instrument configuration of an acoustooptical tunable filter (AOTF) NIR spectrometer.

time. As is shown in Figure 71.17, the light emitted from the AOTF crystal is two orthogonally polarized beams. To use the AOTF as an NIR spectrometer, a beam stop is used to exclude all but the desired monochromatic light beam that is directed toward the sample.

A major advantage of the AOTF spectrometer over other spectrometers is the ability to precisely and rapidly tune the system to a specific wavelength. This is useful when a single absorbance band has been found to monitor a process in real time. Since the RF can be tuned in a matter of microseconds, the collection of an entire NIR spectrum is possible in a short time period.

AOTF instruments find widespread usage in high-speed monitoring applications in the pharmaceutical and agricultural sectors. They are most applicable for industrial applications due to their speed, ruggedness, and no moving parts.

71.3.7 Microelectromechanical Spectrometers

Microelectromechanical spectrometers (MEMS) are mass-produced silicon chip-based devices, which consist of a tunable Fabry–Perot filter [40]. The manufacturing process builds various substrate layers onto the surface of the chip, and the result is a microelectromechanical device consisting of two dielectric mirrors, facing parallel to each other and separated by a small distance from each other. The application of precise voltage levels to the device changes the distance between the two mirrors, thus allowing

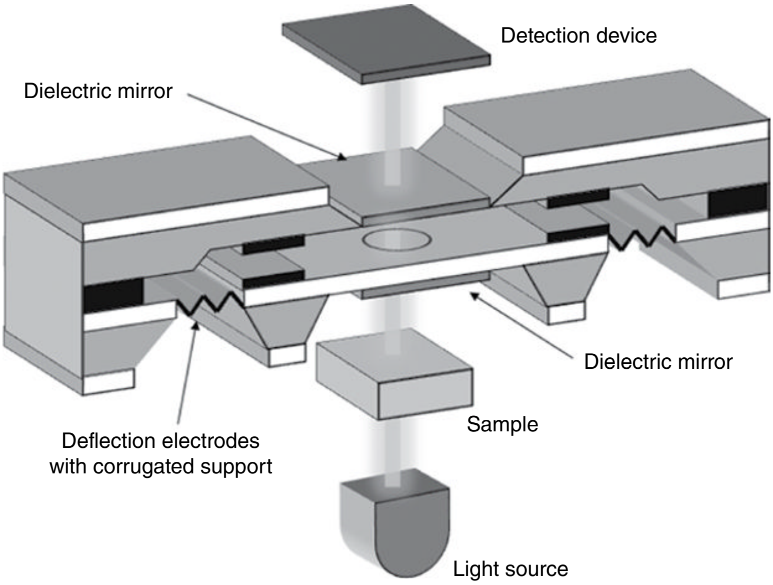


FIGURE 71.18 General configuration of a microelectromechanical spectrometer (MEMS).

the device to be tuned to specific wavelength bands. Figure 71.18 provides a schematic representation of the light dispersion devices for MEMS.

The result is a fast tunable bandpass filter capable of recording a complete spectrum in 50 μs. MEMS are typically used in small, handheld devices due to their size, and for more information on this interesting class of spectrometer, the interested reader is referred to the literature [30].

71.3.8 Linear Variable Filter Instruments

In recent times, miniaturization of instrumentation has taken a massive leap forward with the introduction of stable micro-NIR spectrometers, which utilize Linear Variable Filters (LVFs). It is interesting to note that the early work on NIR was performed by Professor Karl Norris using instruments based on LVFs [41], and now this technology is making its renaissance in NIR technology.

An LVF is a bandpass filter coating intentionally wedged in one direction. As a result of the varying film thickness, the wavelength transmitted through the filter varies linearly in the direction of the wedge. The LVF is coupled to a linear detector array using a semiconductor manufacturing process such that the entire LVF and detector are seamlessly joined together [42]. Figure 71.19 provides a schematic overview of a currently available LVF device.

As with the new FT instruments discussed in Section 71.3.5, the LVF instruments can be manufactured as a complete unit from highly repeatable manufacturing processes. Much time and investment have been made into the development of these

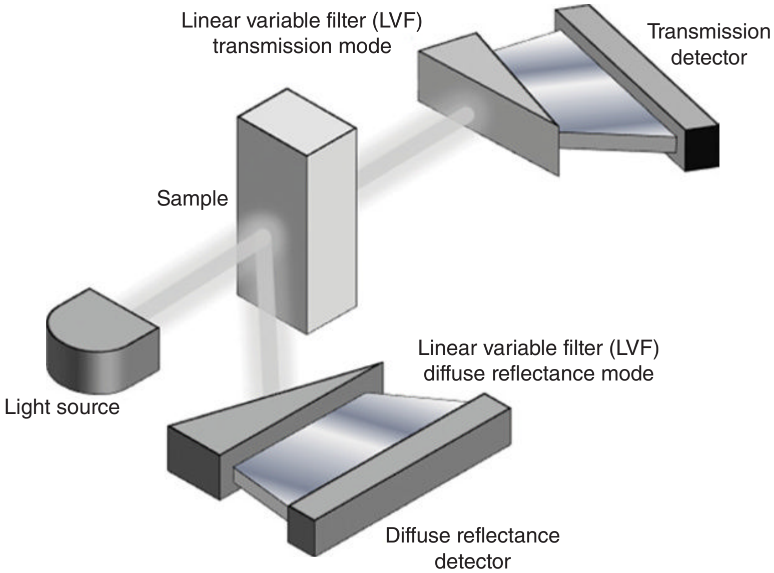


FIGURE 71.19 General configuration of a micro-NIR device utilizing a Linear Variable Filter (LVF).

systems in recent times, and they will form the cornerstone of many process applications, particularly in process analytical technology (PAT) applications moving into the future (refer to Section 71.6.2.3 for more details).

The resolution of these instruments is generally between 0.5 and 1% of the center wavelength of measurement, and they can be configured to cover a wide wavelength range (depending on the detector system fused to the LVF). Scanning times are in the order of 1–2 s, depending on the integration time and scan count used to collect a spectrum. They contain no moving parts at all and are highly ruggedized to be used in harsh conditions.

71.3.9 A Brief Overview of Detectors Used for NIR Spectroscopy

To this point, the discussion has been primarily focused on the dispersion devices that characterize an instrument type. It is, however, the combination of dispersion device and detector type that determines the wavelength region covered by the final instrument.

In Section 71.2.3.2, Figure 71.8 shows the regions of the NIR spectrum and the chemical phenomena measured. Correspondingly suitable detector types must be matched to the region being measured. In particular, only three detector types will be discussed in this section:

1. Silicon (Si)
2. Lead sulfide (PbS)
3. Indium gallium arsenide (InGaAs)

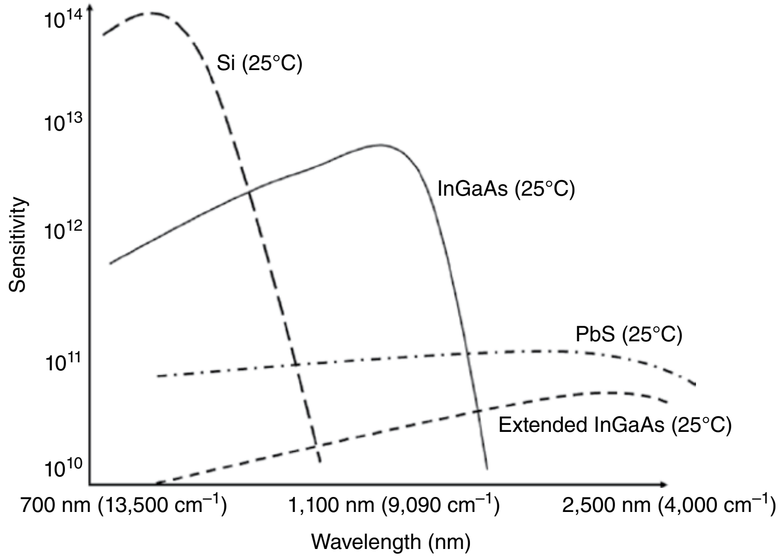


FIGURE 71.20 Comparison of detector sensitivity and wavelength coverage in the NIR region of the electromagnetic spectrum.

Figure 71.20 provides an overview of the detector sensitivities over the NIR region of the spectrum, and a brief discussion of the detector applicability is provided in the following subsections.

71.3.9.1 Silicon Detectors Silicon (Si) detectors are used primarily in the SW NIR region (refer to Section 71.2.3.1). In this region, the third and fourth overtones are measured, and this requires a detector with the high sensitivity characteristics of the Si detector (refer to Figure 71.20). In this region of the spectrum, specular reflectance phenomena (i.e., purely light reflection phenomena) dominate, and therefore, diffuse reflectance methods provide little chemical information of the sample being measured. Therefore, this region of the spectrum is mainly used for transmission measurements, and the Si detector is capable of measuring the small light levels (typically 3–5 AU), which typically emanate from passing light through densely packed solid materials such as grain or pharmaceutical tablets. The Si detector has a maximum sensitivity at 950 nm and has an effective working range between 350 and 1100 nm.

71.3.9.2 Lead Sulfide Detectors Lead sulfide (PbS) detectors are primarily used in the LW region of the NIR spectrum. Typically used for diffuse reflectance applications, the PbS detector has a maximum sensitivity at 2500 nm and an effective working range between 1100 and 2500 nm (actually, the PbS detector can also be used in parts of the MIR).

Some grating-based instruments have utilized a combination of Si and PbS detectors to create an instrument capable of measuring between 700 and 2500 nm. The issue with this type of instrument configuration is that at the junction between the two detectors at 1100 nm, an inconsistent join is visible in the generated spectrum. This junction has to be removed when developing predictive models.

71.3.9.3 Indium Gallium Arsenide Detectors The indium gallium arsenide (InGaAs) detector is a midpoint between the Si and PbS detectors. The normal InGaAs detector has a maximum sensitivity comparable to the Si detectors with a maximum sensitivity around 1400 nm and an effective working range between 900 and 1700 nm. An extended version of the detector is available by doping the InGaAs with phosphorous. This provides a detector with a wavelength range between 1200 and 2600 nm with a maximum sensitivity around 2000 nm; however, the doping process reduces the sensitivity of the detector with respect to the nondoped version.

InGaAs detectors are the most widely used detectors in NIR instruments since they can cover the 900–1700 nm range where a majority of NIR information can be found, and also it allows some limited applicability to both transmission and reflectance applications.

71.3.10 Summary

This section summarized the many types of NIR instrument available on the market. Each spectrometer type has its advantages and disadvantages with respect to performance. Instruments can be loosely characterized into two categories:

1. Research-grade instruments
2. Portable instruments

Research-grade instruments are typically slower in performance (requiring between 10 and 30 s to collect a spectrum); however, they are highly stable, provide multiple sampling options to be utilized by a single spectrometer, and are highly useful for method development purposes before deploying a smaller, portable system into a real-time process application.

Portable instruments utilize fast scanning dispersion devices, which are typically lower in spectral resolution compared to research-grade instruments; however their speed characteristics more than make up for the resolution deficiencies. The smaller systems are typically used in handheld devices or in process applications where space is limited and where the implementation of a larger system would be excessive in terms of system build and budgetary requirements.

Section 71.4 provides details on the various sampling modes used for the instrument types defined in this section, and Section 71.6 provides example applications where these instruments have found to be highly valuable for business critical operations.

71.4 MODES OF SPECTRAL COLLECTION AND SAMPLE PREPARATION IN NIR SPECTROSCOPY

The low molar absorptivity constant and the high S/N ratio associated with instrument detectors provide NIR with a number of key advantages over many other spectroscopic methods when used for on-line, in-line, and at-line applications. Before continuing with a discussion of spectral collection and sampling, a brief explanation of the general types of application are provided:

- At-line data collection: In this mode, samples are physically taken from the process/bulk material and presented to the instrument for analysis in a remote laboratory.
- On-line data collection: A side port or other sampling systems are used to divert a sample from the main process into a cell (or other devices) for analysis by the spectrometer, with no user intervention required. Typically the sample is not returned to the process after analysis.
- In-line data collection: The sampling system, for example, fiber optic probe, is inserted directly into the process, and samples are measured as they exist in the process.

Based on the earlier definitions, there are a number of ways a sample can be presented to an NIR spectrometer for analysis. As discussed in Sections 71.3.1.1 and 71.3.1.2, to generate an NIR spectrum, the major modes of spectral collection are either transmission or reflectance.

71.4.1 Transmission Mode

In transmission mode, NIR radiation is passed through a sample, usually in the state it exists in, and the light that is not absorbed is collected on the detector for each wavelength scanned and presented as either a transmission or absorbance spectrum (Section 71.2.3). Figure 71.9 provides a diagrammatical overview of the transmission process.

To collect a spectrum in transmission mode, incident light from an NIR source is passed onto the dispersion element of the spectrometer (refer to Section 71.3 for dispersion devices used in NIR spectroscopy), and this serves as the background (or reference) spectrum (I_0). The sample is then presented to the sampling device where light is passed through the sample. This generates a signal based on the light that is transmitted (i.e., not absorbed) by the sample for each wavelength (I). The ratio of I/I_0 for each wavelength generates the transmission spectrum.

Taking the negative logarithm of the transmission values at each wavelength generates the absorbance spectrum. In many cases, the absorbance spectrum is considered the first preprocessing of the data as the logarithm acts to linearize the data before analysis. Section 71.5 provides more details of the preprocessing methods commonly used for NIR spectroscopy. Depending on the physical nature of the sample, this dictates the wavelength region used to collect a transmission spectrum.

71.4.1.1 Normal Transmission By normal, it is meant that the sample being analyzed is a clear, homogeneous liquid sample contained in a cuvette or sampling cell where the effects of scatter have been minimized. The low molar absorptivity coefficient works into the favor of the method developer and allows the optimization of pathlengths to obtain the best spectral quality for the application.

For example, NIR is highly specific (see Fig. 71.8, band assignment table) to moisture, and if the pathlength is too large, detector saturation at higher wavelengths, that is, greater than 1500 nm, may result. In the region between 900 and 1000 nm, moisture absorbs relatively unhindered by other molecular groups. Therefore the combination of low absorptivity and high pathlength can be used to measure larger sample volumes, using a lower-cost spectrometer setup and providing an effective means to monitor a business critical operation using a simple, yet robust setup.

Using another example, the region of the NIR spectrum between 1200 and 2000 nm lends itself very well to the analysis of octane number and other properties in gasoline using transmission. In particular, NIR is sensitive to aromatic C—H absorbances and can be used to distinguish between aromatic components and straight chain hydrocarbons. In this application, the pathlength of the cuvette or cell is controlled to avoid detector saturation due to increasing molar absorptivity coefficients.

71.4.1.2 Diffuse Transmission To reiterate the importance of low molar absorptivity coefficient in the NIR region, the short wavelength region is most useful for diffuse transmission. The difference between normal transmission (discussed in Section 71.4.1.1) and diffuse transmission (discussed in Section 71.4.1.2) is that the sample is typically solid or a highly particulate suspension/emulsion that absorbs and scatters most of the NIR radiation before it reaches the detector. In sectors such as the agricultural industry (Section 71.6.1), grain traders use NIR to measure commodity prices of the grains either during harvest or after blending. This requires that the NIR radiation passes through up to 30 mm of solid material in order to generate a reliable spectrum.

The absorbance values associated with diffuse transmission are pathlength and material dependent but often are in the 3–5 AU range. Reliable spectra can be recorded in some cases above 5 AU owing to the high S/N ratios of the instrument detector system (see Section 71.3.9).

Another application of diffuse transmission is the on-line measurement of products that are in suspension or emulsion form. The nature of the sample measured is typically heterogeneous and is prone to scatter effects that may distort the spectral consistency between successive scans. There are a number of excellent preprocessing methods used to modify NIR spectra to minimize the effects of scattering due to particulate matter, oil bubbles, and air pockets that may exist in these samples, and Section 71.5 outlines some of the common methods applied before building NIR quantitative or qualitative models.

71.4.2 Diffuse Reflectance

In many applications, the use of transmission is not always practical, particularly when measuring solid samples in a high-speed environment. While transmission measurements may provide more robust quantitative models due the greater sampling cross sections achievable to generate a reliable spectrum, up to 120 coadded scans may be required. This is not a feasible solution for high-speed applications.

Diffuse reflectance measurements typically use a direct illumination, fiber optic device, or integrating sphere (Section 71.4.2.1) to collect sample spectra. The process of diffuse reflectance spectral collection is provided in Figure 71.21, and the general instrument setup was previously shown in Figure 71.10.

Since reflectance measurements are made on the surface of a sample, the general assumption that has to be made is that the sample is homogeneous throughout the depth of penetration of the radiation. In NIR spectroscopy, the depth of penetration of the radiation is in the order of 2–10 mm, depending on the nature of the samples and the particle size distribution. Figure 71.22 provides a diagrammatical representation of the depth of penetration and particle size and how it influences the absorbance values of the resulting spectrum.

A practical means of determining the depth of penetration of the radiation for a particular application is to fill a borosilicate glass vial with varying levels of the material

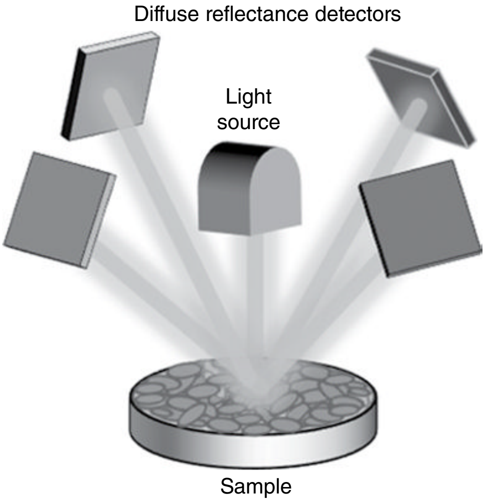


FIGURE 71.21 General instrument configuration for collecting NIR spectra in diffuse reflectance mode.

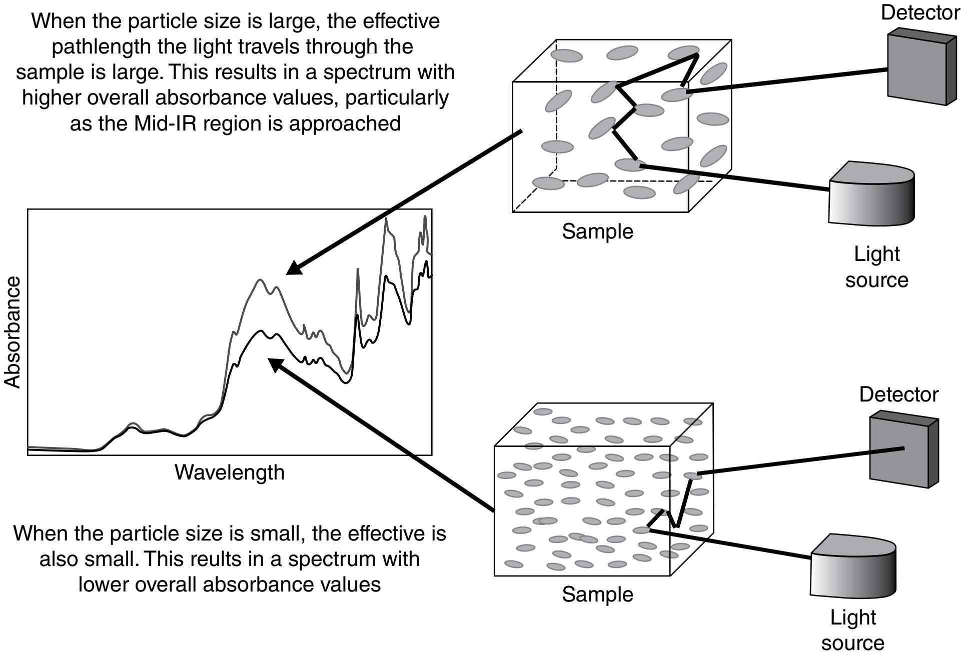

When the particle size is large, the effective pathlength the light travels through the sample is large. This results in a spectrum with higher overall absorbance values, particularly as the Mid-IR region is approached
Detector
Sample
Light source
Absorbance
Wavelength
When the particle size is small, the effective is also small. This reults in a spectrum with lower overall absorbance values
Detector
Sample
Light source

FIGURE 71.22 The relationship between a sample particle size and the absorbance measured by an NIR spectrometer.

to be analyzed and place a material of known spectral characteristics on top of the material. The depth at which only the spectral features of the material to be analyzed are observable in the spectrum can be used as an assessment of the depth of penetration. Another mode of reflectance in NIR spectroscopy is known as specular reflectance. Modern NIR sampling devices are designed to minimize this effect by placing the detectors at 45° to the incident light source. Specular reflectance results in the incident light being reflected back to the source at approximately 180° to the source. This radiation therefore has minimal interaction with the absorbing matter and therefore contains little to no chemical information. The ultimate goal of the NIR method is to collect the most relevant chemical and physical information from the sample, and this information can only be reliably found in a diffuse reflectance spectrum.

71.4.2.1 The Integrating Sphere An extremely useful sampling device in many types of spectroscopy is the integrating sphere. It is particularly well suited to diffuse reflectance measurements and is utilized by a number of instrument vendors. Figure 71.23 provides a diagrammatic representation of an integrating sphere used for diffuse reflectance measurements.

The device consists of a precisely constructed glass sphere coated with a highly efficient material that reflects light (typically gold for NIR applications). The light source is located at 180° to the sample. When light is incident on the sample surface, the

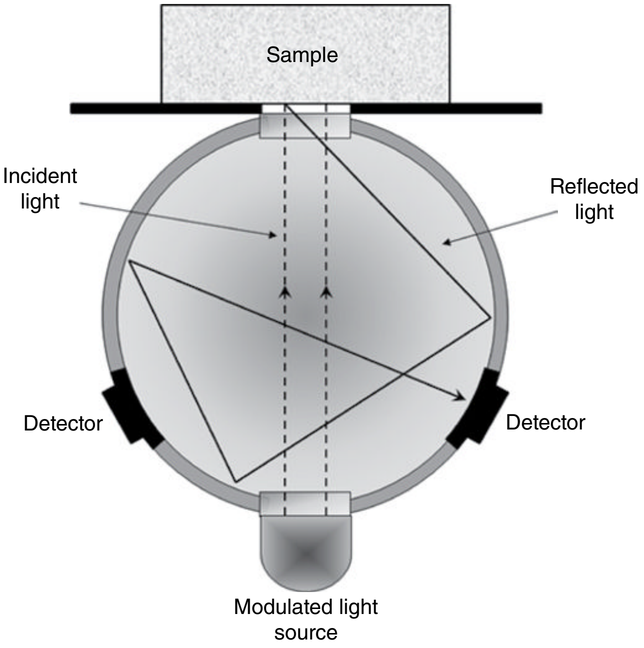


FIGURE 71.23 The integrating sphere.

specular reflectance is sent back to the light source and is therefore not detected. The light that is not specularly reflected can be considered to be diffusely reflected, and this light bounces inside the sphere until it finally reaches the detector.

71.4.2.2 Interactance Measurements Due to the large effects of specular reflectance in the SW region of the NIR spectrum, diffuse reflectance sampling devices provide very little information to the analyst. Figure 71.24 shows this region of the spectrum for a sample measured in diffuse reflectance mode.

In order to use reflectance devices in the SW NIR region, an interactance device is used. This is just a large area of light and detection surface designed to push as much light as possible onto a sample and collect as much of the reflected signal as possible. It is designed to have close contact with the sample and has found much use for the measurement of emulsions and gels, where scatter effects can dominate when measuring at longer wavelengths in the NIR region. Figure 71.28 shows an example of an interactance probe and its fiber configuration.

71.4.3 Sample Preparation

Samples present themselves in many forms, and these can be such items as pharmaceutical tablets, bottled wine, crushed sugarcane, and gasoline, just to name a few. This section provides a practical guideline on what type of sample preparation is required to obtain high-quality NIR spectra for the vast number of applications.

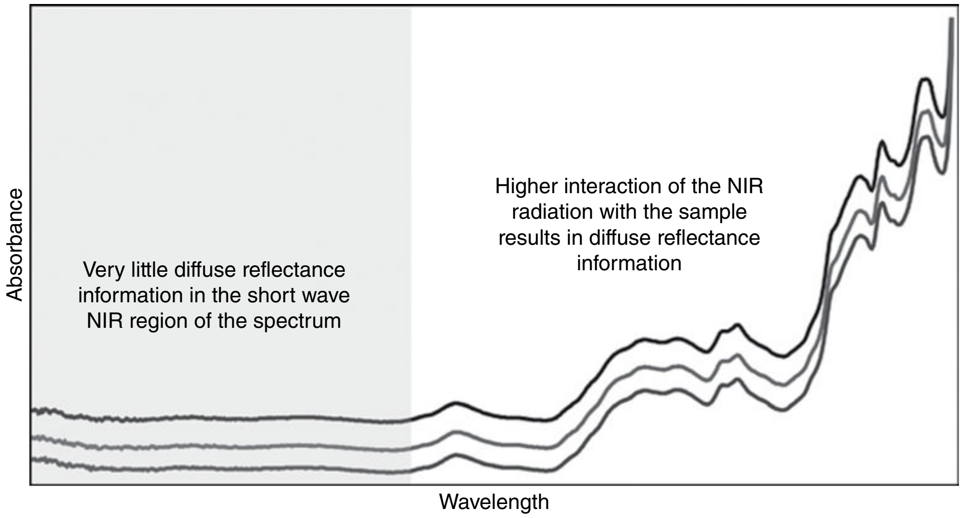


FIGURE 71.24 Sample spectrum measured in diffuse reflectance mode showing little chemical information in the short wave (SW) region.

71.4.3.1 Solid Powders Solid powders can be comprised of a single "homogeneous" material, such as a pure raw material, or may be a blend of two or more components. Powder samples should typically be loose and free of visible clumps. Clumping may be the result of moisture contamination, and if this is not the usual state of the material, it is important to assess the sample quality visually before analysis by NIR spectroscopy.

The method developer must ensure some level of homogeneity of powder blend samples before analysis to avoid nonrepresentative sampling. If this is a concern, multiple scan averages should be collected to avoid heterogeneity problems. Some NIR instruments come with a large sample area device that rotates the sample through a full revolution and takes multiple scans. The average of this scan is used as the final spectrum for analysis or method development purposes.

In the case of flaky materials (usually in the form of small plates), some sample preparation may be required to reduce the particle size and therefore reduce the high sampling variations associated with poorly packed materials. This is easily achieved through a light preparation in a mortar and pestle.

If the powder material is to be measured at line, sampling methods such as borosilicate glass vials can be used to measure the material in a manner that allows visual inspection of the material before scanning. This method also allows complete control over the sample preparation and due to the nondestructive manner of the analysis allows for retention of the same sample used for the analysis. This is highly important in cases where sample traceability is required (particularly in the pharmaceutical industry).

71.4.3.2 Grains and Seeds The agricultural sector utilizes NIR for many of its operations (refer to Section 71.6.1) and, in particular, for the analysis of grains and

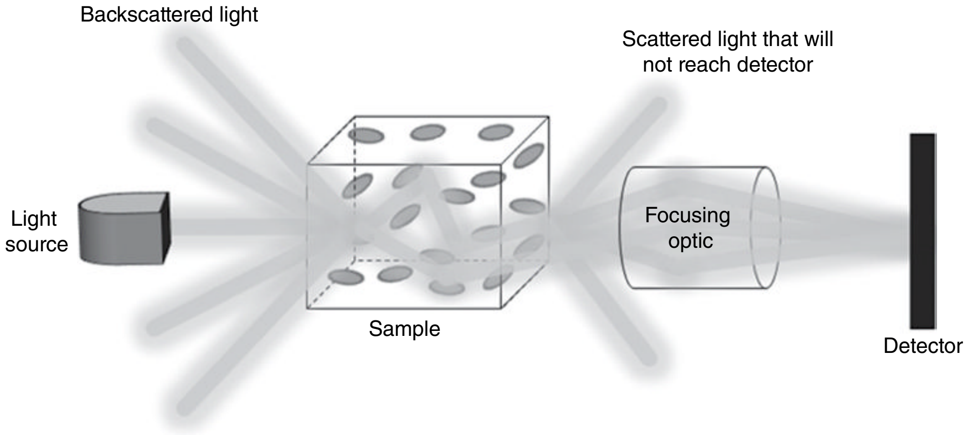


FIGURE 71.25 The process of diffuse transmission as applied to grains and oilseed analysis.

oilseeds. The main mode of analysis has been transmission mode using specially designed grain hoppers. The grain is fed into the top of the hopper of an at-line instrument where it is fed into a fixed pathlength cell. NIR radiation is passed through the sample, and after scanning, the old sample is ejected and a new sample is received. This process is performed multiple times in order to minimize packing differences in the sample and heterogeneity. Figure 71.25 shows the process of diffuse transmission as applied to grains and oilseeds.

71.4.3.3 Silage and Sugarcane During the harvest of grains, the remaining stalks (also known as silage) are also harvested. The material presented to the instrument is either high in moisture or as a dry straw. Sugarcane, after crushing in a mill, also has a similar consistency. To analyze this sample type, the use of overhead NIR instrumentation is required.

Silage is typically monitored in a harvester in real time using a device located on the exit chute to a hopper (Fig. 71.26a) or utilizing an innovative sampling method on a conveyor belt in the case of the sugarcane industry (Fig. 71.26b). Again, due to the heterogeneity of the samples, scans are taken over a specific sample area and averaged to result in a final scan, which minimizes density differences and heterogeneity. More details of these applications are provided in Section 71.6.1.2.

71.4.3.4 Pharmaceutical Tablets Many instrument configurations exist for the analysis of intact pharmaceutical tablets. The most common sampling approach is to develop custom-made holders for the tablet type and analyze the tablets in transmission mode. Figure 71.27 shows this setup diagrammatically.

By measuring tablets in transmission mode, a greater area of sample is measured through the interior of the tablet. In this case, all of the sampling considerations of diffuse transmission must be taken into account, that is, tablet thickness (pathlength

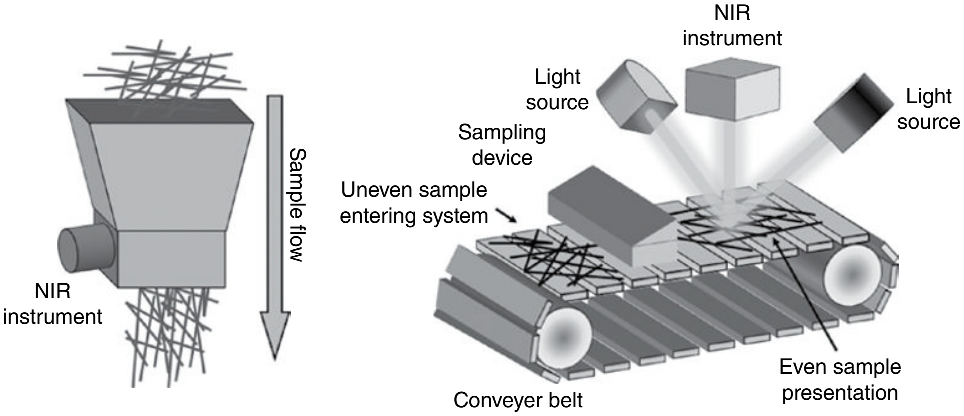


FIGURE 71.26 Analysis of silage (left image) or sugarcane (right image) requires sampling devices that best account of sample heterogeneity.

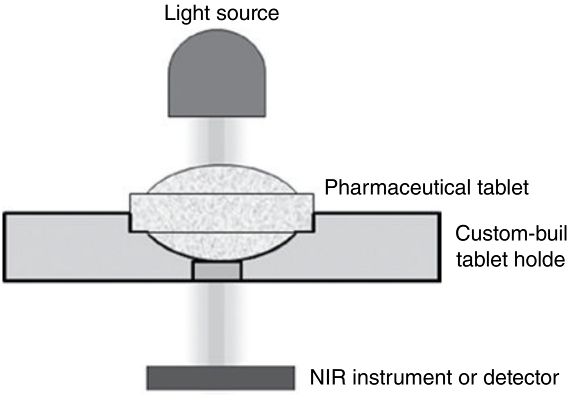


FIGURE 71.27 General sampling device used for measuring intact pharmaceutical tablets in diffuse transmission mode.

considerations), region of the spectrum used, etc. Due to detector response saturation above 1500 nm in transmission mode, most of the spectral information is useless for analysis (refer to Fig. 71.7). It is up to the method developer to exclude these noise parts of the spectrum before a calibration model can be developed. Diffuse reflectance measurements can also be used for measuring intact tablets, provided the following conditions hold:

1. The concentration of the active material in the tablet is high (typically >5% w/w).
2. It can be assumed that the distribution of the active material is homogenous over the surface of the tablet.

This will be discussed more in Section 71.6.2.2.

71.4.4 Fiber Optic Probes

This section provides a brief overview of some of the common types of fiber optic probe currently in use for NIR applications. Fiber optic probes allow for larger instruments or instruments that are sensitive to mechanical shock or environmental conditions to become “portable.”

Fiber optic probes bring the instrument to the process and can be configured in any of the sampling arrangements previously discussed in Section 71.4. Fiber optic probes are most commonly used in process applications, where measurements are performed on the sample as it exists in the process. The use of fiber optic probes is now becoming limited due to the emergence of smaller and more robust instruments, which can make direct contact with the process. While practically advantageous in some applications, fiber optic probes add another level of complexity, particularly when it comes to instrument matching for calibration transfer purposes. The following subsections discuss the various types of fiber optic probe available.

71.4.4.1 Reflectance Probes There are a number of configurations for a direct insertion reflectance probe, and these can be used to measure both solid and liquid interfaces. Reflectance probes are typically designed for the collection of diffuse reflectance spectra, but they also come in designs that accommodate attenuated total reflectance (ATR) and interactance. Figure 71.28 provides a schematic diagram of the probe types available.

These probes are designed such that they can be permanently mounted into a vessel or pipe (with a self-referencing and self-cleaning mechanism, as is a design feature of

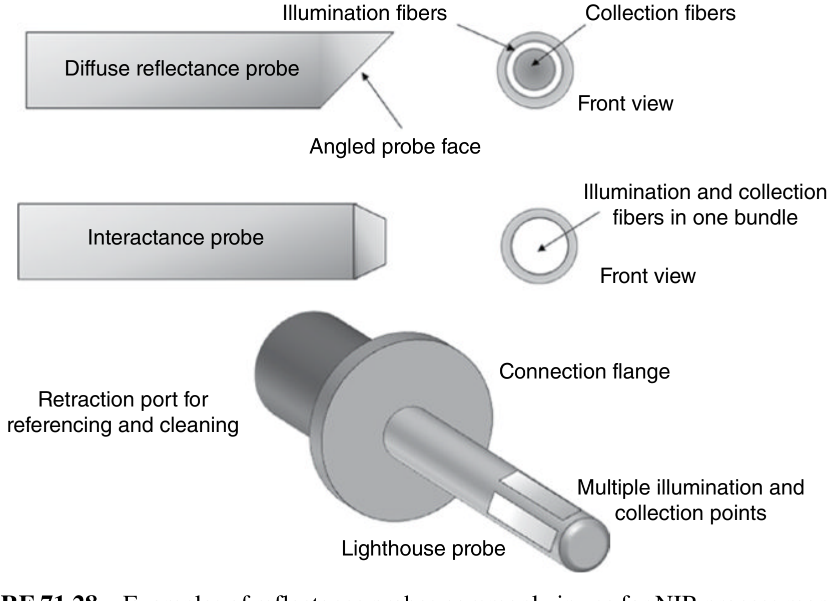

Illumination fibers
Collection fibers
Diffuse reflectance probe
Front view
Angled probe face
Illumination and collection fibers in one bundle
Interactance probe
Front view
Connection flange
Retraction port for referencing and cleaning
Multiple illumination and collection points
Lighthouse probe

FIGURE 71.28 Examples of reflectance probes commonly in use for NIR process monitoring applications.

the lighthouse probe) or designed to be inserted and withdrawn from a suitably designed housing that controls the depth of the probe into the sample.

Reflectance probes have found widespread use in applications from the pharmaceutical, biopharmaceutical/biotechnology, and petrochemical sectors where monitoring of critical processes can be performed in real time and changes can be made to the process before they become an issue.

71.4.4.2 Transmission Probes Transmission probes typically find use in on-line process applications where the sample can be sidestreamed (and sometimes conditioned) before analysis. Typically suited for clear to relatively turbid (high scattering) liquid applications, transmission probes can offer some real advantages over reflectance probes due to the higher sample volumes measured. Figure 71.29 shows two common types of transmission probe in use for industrial applications.

When the transmission probe is configured as a direct insert probe, the light path is directed from the source to the sample. The pathlength can be adjusted by the use of a calibrated scale on the probe, which can be precisely set; however, in most cases probes come in a fixed pathlength configuration. Once the light passes through the sample (defined by the pathlength), the nonabsorbed light is bounced off an internal mirror and can be sent back to the detector in one of two ways:

1. It can be reflected back through the sample. In this case, the pathlength is doubled since the light has to pass through the sample twice before it reaches the detector.
2. It can be reflected back. This time the light path is through a separate fiber optic that does not pass through the sample twice.

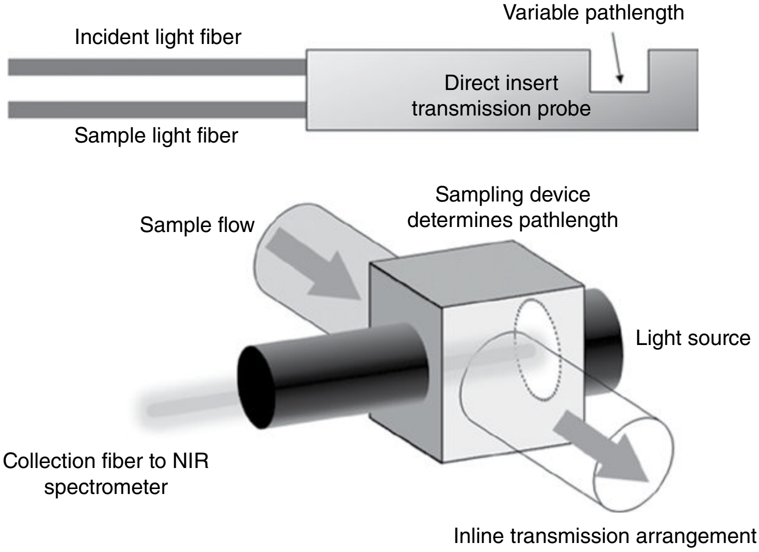

Variable pathlength
Incident light fiber
Direct insert transmission probe
Sample light fiber
Sampling device determines pathlength
Sample flow
Light source
Collection fiber to NIR spectrometer
Inline transmission arrangement

FIGURE 71.29 Examples of transmission probe systems typically used for monitoring processes using NIR spectroscopy.

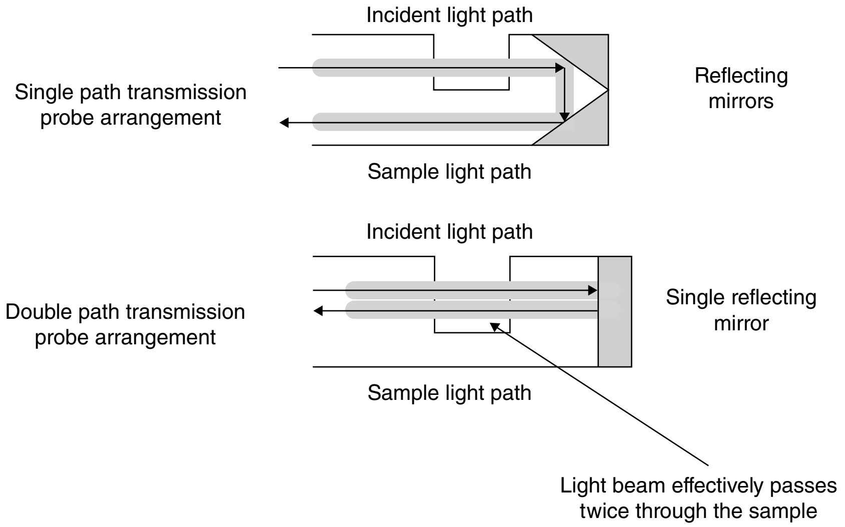

Incident light path
Single path transmission probe arrangement
Reflecting mirrors
Sample light path
Incident light path
Double path transmission probe arrangement
Single reflecting mirror
Sample light path
Light beam effectively passes twice through the sample

FIGURE 71.30 Configurations of direct insert probes used for measuring transmission NIR spectra.

These two situations are shown in Figure 71.30.

As was the case of reflectance probes, transmission probes can be permanently assembled into a process line or inserted and withdrawn from a fixed housing. The requirement of a transmission probe is that it needs a clear optical path to collect a reference spectrum.

71.4.5 Summary of Sampling Methods

This section provided a brief overview of the sampling methods that can be employed to capture reliable NIR spectra in a laboratory and a process monitoring situation. When approaching the design of a system capable of measuring meaningful NIR spectra for a particular application, the design engineer or scientist must use all of the information available to them in order to optimize the system for the application. This includes ensuring that the following main criteria have been taken into consideration:

1. Ensure that the NIR method is capable of performing its task by conducting feasibility studies.
2. Determine the required speed of analysis as this will determine the type of dispersion device used for the application (refer to Section 71.3).
3. Determine whether spectral resolution is a limiting factor in the analysis and choose an instrument that meets both the speed and resolution requirements.
4. If a moving grating or FT instrument is to be used, the sampling requirements must be optimized for the application. Determine whether the sample can be brought to the instrument (at line) or if the instrument needs to be interfaced with the process via fiber optic probes (refer to Section 71.4.4).

5. If a portable instrument is to be used, can it be interfaced with a minimum or no fiber optic cabling? If so, this is the ideal situation, and this requires that the instrument needs a permanent mounting to keep it in position for analysis. Also assess the environment for the possibility of explosion and ensure that the instrument configuration is rated to comply with the requirements of such environments.
6. Once the instrument configuration is built and the housings are all secured, a suitable graphical user interface (GUI) must be decided upon. This can be as simple as the vendor of the instrument native platform or can be customized for process operators using traffic light systems or simple pass/fail messages.
7. Determine whether the results generated by the NIR analysis are to be used for process control. In this case, the output of the instrument must be sent via a standard communication protocol to the control system such that action can be taken.

Once all of these factors have been taken into account, this should give the end application the best chance of success. It cannot be stressed more; representative sampling is the key to reliable NIR method development. The Theory of Sampling (TOS) is outside of the scope of this chapter, but it must be taken into account before designing a reliable system. The interested reader is referred to the excellent literature on TOS [43].

Representative sampling ensures that the data collected for model building is of the highest quality. There will always be random effects present in spectral data due to particle size of materials and varying sample density. This is where preprocessing of data can be used to minimize such effects, and this is discussed in the next section of this chapter.

71.5 PREPROCESSING OF NIR SPECTRA FOR CHEMOMETRIC ANALYSIS

NIR spectroscopy in general does not produce the sharp absorbance bands, which are typical of other spectroscopic methods such as MIR, Raman, and NMR spectroscopy. For this reason, early researchers did not investigate this region as it provided little structural information that could be interpreted in the traditional ways.

It is true to say that NIR and chemometrics (see Chapter 65) go hand in hand. This section does not go into the details of the algorithms used to generate predictive or classification models but focuses primarily on the preprocessing methods used to minimize unwanted effects before the modeling process is performed.

It is stressed here that preprocessing is not a substitute for good sampling! Ninety percent of the initial effort for implementing an NIR system should be given to optimal sampling. If the data that are generated by the instrument are nonrepresentative or substandard, the final model is either bad or requires more components to account for the bad sampling. This results in a complex model that, in the case of failure, may not be interpretable. This is a case of garbage in–garbage out.

71.5.1 Preprocessing of NIR Spectra

NIR spectroscopy is typically used to measure solid samples either in powdered form or as grains (agriculture). This introduces variability into the packing as the solid sample cannot be packed the same twice in a row. This results in packing density variations that affect both reflectance and transmission measurements, and these are commonly known as additive effects, since they primarily affect the baseline of the spectra.

Varying particle size distributions also contribute to spectral variations between samples. In the NIR region of the spectrum, the wavelength of the radiation greater than 1600 nm becomes of comparable size to the particles in the sample, and thus elastic rather than specular interactions occur between the incident radiation and the sample (refer to Section 71.4.2). The overall effect is to introduce nonlinear (or scatter effects) into the spectra, also known as multiplicative effects. The following sections briefly describe the most common preprocessing methods used in NIR spectroscopy for minimizing both additive and multiplicative effects.

71.5.2 Minimizing Additive Effects

An additive effect can be considered as a constant offset in the spectra. To remove such effects, a differencing factor must be introduced such that a common baseline is achieved. In NIR spectroscopy, only in certain cases (particularly the analysis of nonscattering liquids) is the baseline ever linear. The most common approach to minimizing purely additive effects in NIR spectra is by the use of derivatives. There are a number of commonly used approaches to derivatives, and most software packages either employ:

1. The Savitzky–Golay derivative [44] or the
2. Segment–Gap derivative [44]

These two derivative types are discussed in more detail as follows.

71.5.2.1 How Derivatives Work By definition, derivatives (also known as differentiation) are (generally speaking) calculated as the difference between the second point in the spectrum and the first (divided by a constant centering factor). This acts to center the entire spectrum around the zero line. It also measures the slope of the spectral features (since by definition, derivatives measure the rate of change of data).

Figure 71.31 provides some Gaussian curves of various intensities offset from each other in a linear manner along with their first derivatives.

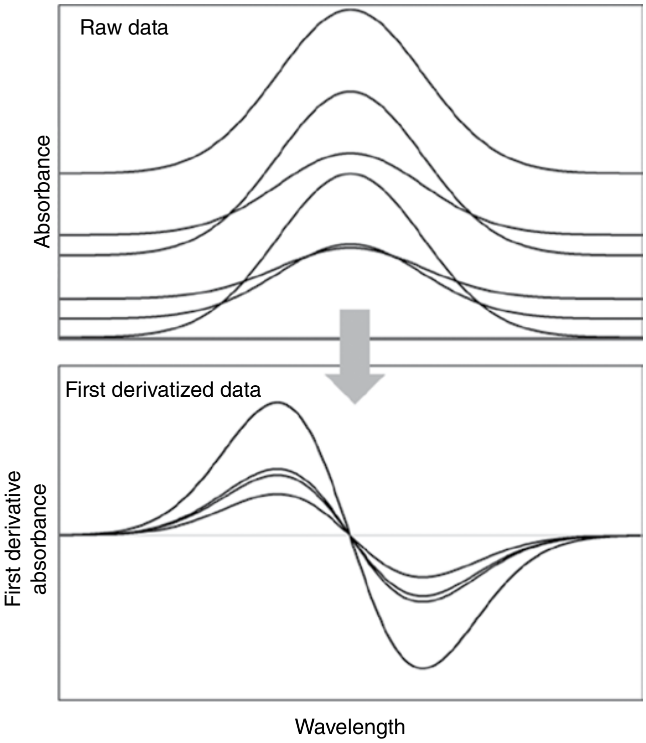


FIGURE 71.31 Gaussian curves of various intensities and offsets along with their first derivatives.

By taking the first derivative of these spectra, Figure 71.31 shows that the data are all centered around zero and the relative intensities of the curves can now be compared on a quantitative basis.

To further explain how the first derivative works, consider a single Gaussian curve and its corresponding derivative in Figure 71.31. Moving from left to right in the Gaussian curve, the slope of the curve is close to zero. From there it rapidly increases until the inflection point is reached. The slope then rapidly decreases to zero when the peak maximum is reached in the Gaussian peak (this is the zero point in the derivative curve). The slope then rapidly decreases as the curve moves to the right of the maxima until it reaches the inflection point. From there, the slope then rapidly approaches close to zero when the end of the curve is reached.

The second derivative is defined as the curvature of the original curve or the slope of the first derivative. Consider the points shown in Figure 71.32.

The first and second derivatives can be described mathematically as follows:

General first derivative:

$$\frac{\partial Y}{\partial \lambda} = \frac{(B - A)}{\Delta \lambda}$$

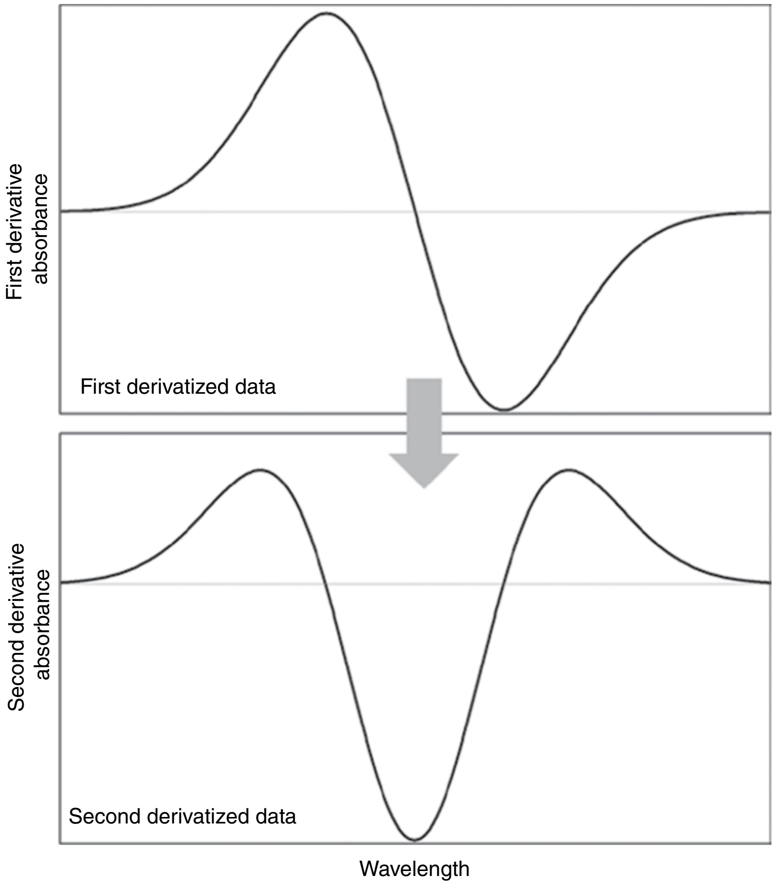


FIGURE 71.32 First and second derivatives of a Gaussian curve.

General second derivative:

$$\frac{\partial^2 Y}{\partial \lambda^2} = \frac{(C - B) - (B - A)}{\Delta \lambda}$$

$$\frac{\partial^2 Y}{\partial \lambda^2} = \frac{C - 2B + A}{\Delta \lambda}$$

where

Y is the *y*-axis absorbance scale

λ is the *x*-axis wavelength scale

A, B, and *C* are the first, second, and third points, respectively, in the spectrum to be derivatized

In the definition of the second derivative, the first equation represents the first derivative of the first derivative. Figure 71.33 shows the application of both first and

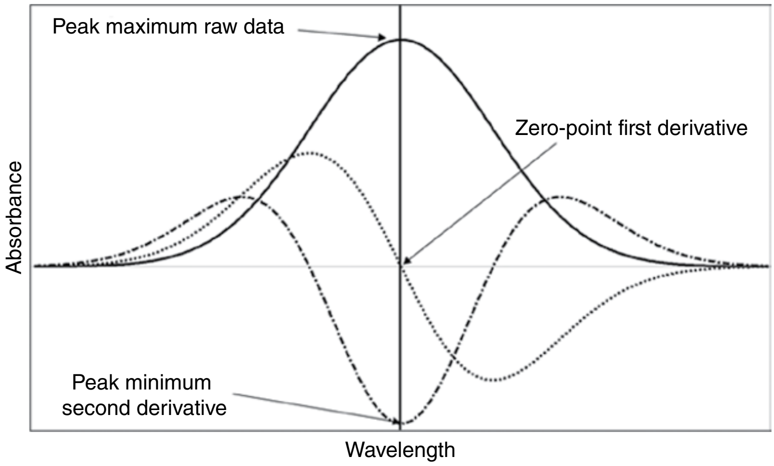


FIGURE 71.33 The relationship of curve maxima, zero point, and minima for Gaussian, first, and second derivative curves.

second derivatives to a Gaussian curve. In the case of the second derivative, the peak maximum in the original curve becomes the peak minimum in the second derivative. This is why the second derivative is preferred sometimes as peak minima are more interpretable than zero points in the first derivative.

Another important feature of the second derivative particularly for application to spectral data that are collected above 1600 nm is that they take curvature into account. Since the baseline of the spectra between 1600 and 2500 nm exhibits quasi-quadratic behavior, the second derivative is particularly useful in minimizing additive effects in this region.

71.5.2.2 The Savitzky–Golay Derivative The Savitzky–Golay algorithm was proposed in 1964 as a digital smoothing filter [44]. It was extended to be a derivative and has found widespread usage in analytical chemistry. The algorithm is based on performing a least squares linear regression fit of a polynomial around each point in the spectrum to smooth the data. The derivative is then the difference of the fitted polynomial at each point. The algorithm includes a smoothing factor that determines how many adjacent variables will be used to estimate the polynomial approximation of the curve segment. Figure 71.34 shows how the algorithm works.

In general, the derivative works by fitting a polynomial to the data defined by the smoothing window. The window size must be an odd number as the smoothed point in the window lies at the center, and this point becomes A in the equations defined in Section 71.5.2.1. The window is then moved along by a one-point increment, and a polynomial is fit to the data spanning the smoothing window and becomes the point B in the equations listed in Section 71.5.2.1. The point C can be calculated in a similar manner. The first and second derivatives are calculated as per the equations. The effect of the polynomial is to provide a smoothed point that is less noisy than the original

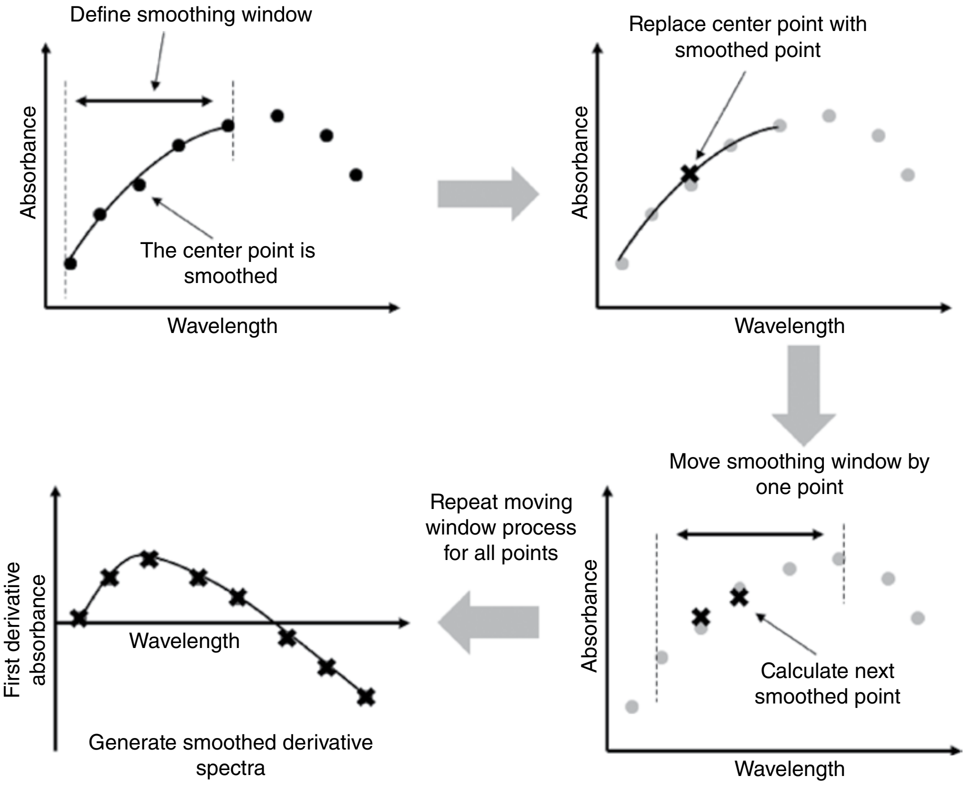

Define smoothing window
Absorbance
The center point is smoothed
Wavelength
Replace center point with smoothed point
Absorbance
Wavelength
Move smoothing window by one point
Absorbance
Calculate next smoothed point
Wavelength
Repeat moving window process for all points
First derivative absorbance
Wavelength
Generate smoothed derivative spectra

FIGURE 71.34 Diagrammatic representation of the Savitzky–Golay derivative.

data; therefore the final derivative is smooth with minimal noise characteristics. More details on how to set window sizes are discussed in Section 71.5.2.4.

The Savitzky–Golay derivative is used in most software packages that are supplied with NIR instruments. The algorithm was corrected for errors in the original tables by Steiner et al. [45] and is the preferred method by most chemometricians because it fits exact mathematical functions to the data to generate a continuous derivative.

71.5.2.3 Segment–Gap Derivatives The Segment–Gap derivative enables the computation of derivatives using an algorithm that allows selection of a gap factor and a smoothing factor. The principles of the Segment–Gap derivative are based on a modification of the moving average algorithm where a suitable smoothing window is used to calculate the average point in the center of the window. The gap size is set such that different sizes between the windows can be set; however, the most common value for the gap size is 1. For such functions, Norris suggested that derivative curves with less noise could be obtained by taking the difference of two averages, formed by points surrounding the selected x locations [46]. As a further simplification, the division of the difference in y-values or the y-averages, by the x-separation x, is omitted.

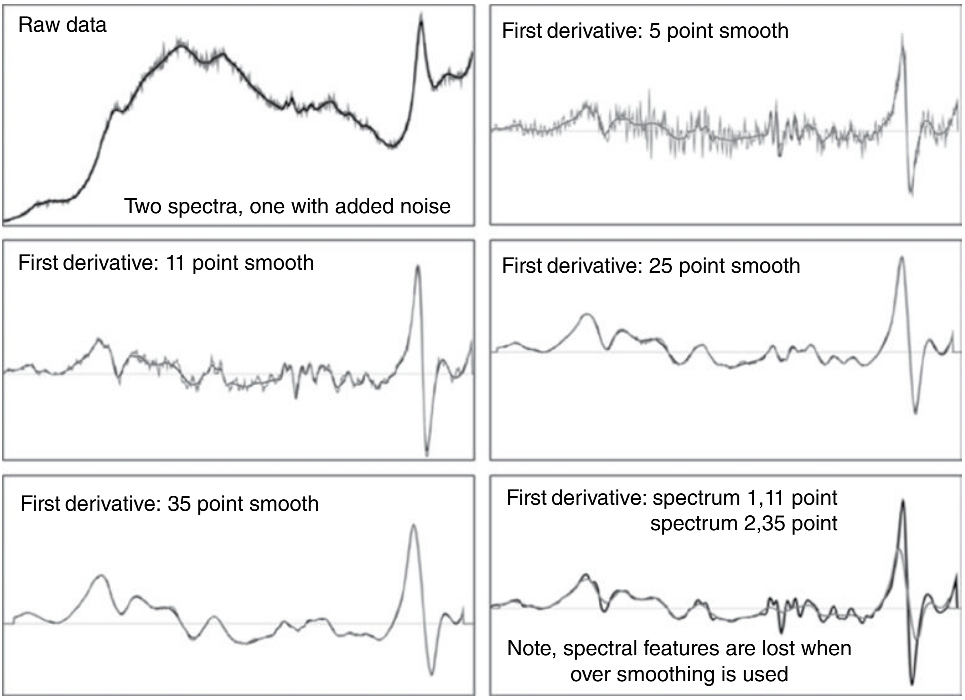


FIGURE 71.35 Comparison of derivative data. One curve is smooth, while the second is the original curve with added noise. It can be seen from this figure that in attempt to remove all of the noise from the spectrum, specific spectral features are smoothed out.

Norris [47] introduced the term segment to indicate the length of the x interval over which y-values are averaged, to obtain the two values that are subtracted to form the estimated derivative. If too large a segment is defined, the resolution of the peaks will decrease. Too narrow a segment (smaller than the half bandwidth of the peak) may introduce noise in the derivative data. This is illustrated in Figure 71.35.

71.5.2.4 Some Pitfalls of Derivatives The definition of suitable window segment sizes is the key challenge to optimizing a derivative. There are a number of factors that must be taken into account when applying a derivative, and if these are overlooked, valuable information may be lost in the spectral data. The following is a checklist that must be taken into account before applying a derivative:

1. Is the baseline shift linear or nonlinear? This aids in the determination of the derivative order to use.
2. Is the wavelength scale linear over the entire range? If a nonlinear scale is used, regions of higher resolution will be smoothed less than regions of lower resolution.

3. The larger the smoothing window, the more noise is removed; however, if too large a window is used, important spectral information may be reduced or even lost through "averaging out."
4. Derivatives enhance noise! Since the process of derivatization is successive differences between points, depending on the data point resolution of the spectrum, the first derivative scale can be 1–2 orders of magnitude lower in y-axis scale than the original scale of the data. The second derivative has an even smaller scale (sometimes close to the noise level of the instrument).

   This is why points 3 and 4 are so important and go hand in hand with each other. It is important to focus any preprocessing on the region of interest to the detriment of other regions. A simple rule of thumb with derivatives is to understand the Full Width at Half Maximum (FWHM) of the most important spectral features and set the smoothing window such that it has minimum effect on the peak shape and intensity.
5. When applying a smoothing window, by definition, the value of the window size is odd such that the center point in the window is the one that is smoothed. This means that the points to the left of the center point on the edge of the spectrum will be eliminated from the preprocessed data; the same also holds at the end of the spectrum for the points to the right of the center point. A method developer must be aware of this fact if they are preprocessing a region of data within the full spectrum range. It is important to leave a number of slack points before and after the region of interest to avoid removing important information.

Overall, derivatives are probably the most used preprocessing technique for NIR spectra. If the previous rules are followed, the preprocessing can be optimized efficiently and can be applied to new data before any predictions are made using chemometric models.

71.5.3 Minimizing Multiplicative Effects

71.5.3.1 Standard Normal Variate and Detrending The Standard Normal Variate (SNV) algorithm is a nonmodel-based scatter correction algorithm first proposed by Barnes et al. in 1989 [48]. The algorithm requires no external spectral information for correction; it uses only the spectrum itself for the correction information. The algorithm works by first calculating the mean value (m_i) and the standard deviation (s_i) over all wavelengths in the spectrum. Then for each wavelength in the spectral region chosen, the mean absorbance is subtracted from each wavelength, and it is then divided by the overall spectrum standard deviation. This can be expressed mathematically as

$$z_{i,j(\mathrm{SNV})} = \frac{z_j - m_i}{s_i}$$

where

$z_{i,j(\mathrm{SNV})}$ is the SNV corrected absorbance value at wavelength (j) for spectrum (i)
z_j is the absorbance value of the original spectrum at the j^{th} wavelength
m_i is the overall mean absorbance value for the i^{th} spectrum
s_i is the standard deviation of all absorbance values for the i^{th} spectrum

Figure 71.36 shows the effect of applying the SNV algorithm to a set of NIR spectra highly affected by scatter.

SNV centers the data around the center of gravity of the spectrum and then scales the spectrum to ±3 standard deviations around the zero line. The overall effect of this preprocessing method is to minimize the physical scatter and packing variations and reveal chemical information for chemometric modeling. It is a simple scatter correction technique and is implemented by many vendor software packages.

In their original paper [48], the authors also suggest that the detrend algorithm be applied to SNV corrected spectra. Detrending is a transformation that minimizes nonlinear trends, thus SNV and detrend in combination reduce multicollinearity, baseline shift, and curvature. The detrending calculates a baseline function as a least squares fit of a polynomial to the sample spectrum. These transformations are applied to individual spectra and are distinct from other transformations that operate at each wavelength in a given set of spectra. As the polynomial order of the detrend increases, additional baseline effects are removed (0 order: offset; first order: offset and slope; second order: offset, slope, and curvature).

Typically, detrending is performed by using a second-order (or higher-degree) polynomial in regression analysis where absorbance values (or y-variables) and the independent variable or x-variable (W) are given by the corresponding wavelengths:

$$z_{\mathrm{Detrend},i} = A_{\mathrm{SNV}} + B_{\mathrm{SNV}} + C_{\mathrm{SNV}}W^2 + \left[D_{\mathrm{SNV}}W^3 + E_{\mathrm{SNV}}W^4\right]$$

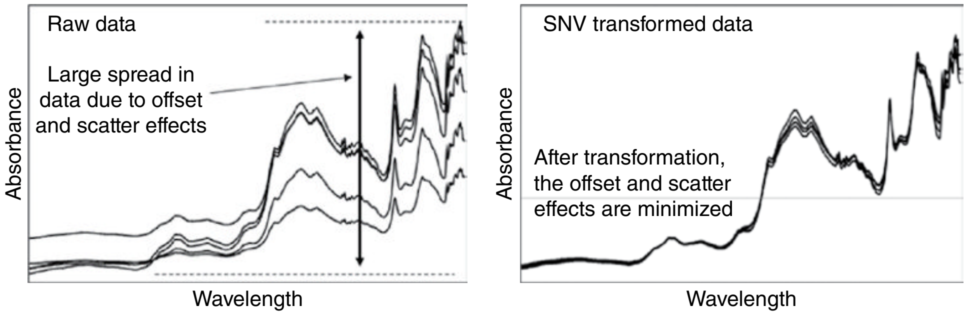


FIGURE 71.36 The Standard Normal Variate (SNV) preprocessing minimizes the effect of offset and scatter in NIR spectra while retaining the original spectral profile.

where *A*, *B*, *C* (and *D*, *E*) are the regression coefficients. The base curve in the earlier relationship is given by the fitted values $z_{\mathrm{Detrend},i}$, and thus derived spectral values subjected to SNV followed by detrend become

$$z_{\mathrm{SNV/Detrend}} = z_{\mathrm{Detrend},i} - z_{\mathrm{SNV},i}$$

This calculation removes baseline shift and curvature, which may be found in diffuse reflectance NIR data of powders, particularly if they are densely packed. The use of this transform does not change the shape of the data, as is the case with the application of derivatives (Section 71.5.2).

71.5.3.2 Multiplicative Scatter Correction Algorithms Multiplicative Scatter Correction (MSC) is a model based transformation method used to compensate for pure additive and/or multiplicative effects in spectral data. Extended Multiplicative Signal Correction (EMSC) is an extension of the original method that in addition allows modeling of different types of wavelength-dependent effects that can be found in the spectra.

Multiplicative Scatter Correction MSC (also known as multiplicative signal correction) was originally designed to deal with multiplicative scattering in reflectance spectroscopy [49]. However, a number of similar effects can be successfully treated with MSC such as:

- Pathlength variations
- Offset shifts
- Interference

The simplest form of light scattering can be described by a simple additive baseline shift (a_i). The measured spectra are then given as

$$z_i = a_i + z_{i,\mathrm{chem}} + e_i$$

where $z_{i,\mathrm{chem}}$ is the theoretical spectra without any noise or scattering effect. The error vector e_i describes the random measurement noise and other scattering effects not described by the equation. It should be emphasized that the baseline shift will be different for each individual spectrum and is modeled separately.

A multiplicative effect (b_i) is also included in MSC so that the measured spectra are given as

$$z_i = a_i + b_i z_{i,\mathrm{chem}} + e_i$$

Since the theoretical spectrum $z_{i,\mathrm{chem}}$ is not known, the individual spectra are instead approximated as

$$z_i = a_i + b_i m + e_i$$

where *m* is some reference spectrum, for example, a typical sample or the mean of a set of spectra. The unknown additive and multiplicative parameters a_i and b_i are then estimated by simple linear regression of the individual spectrum on the reference spectrum. Unlike normal least squares fitting, which aims to minimize the term e_i, MSC aims to model the scatter effects in the data (defined by $b_i m$) therefore removing these effects and maximizing the chemical information in the data. Thus e_i is maximized to retain chemical information.

The coefficients can then be used to calculate MSC corrected spectra:

$$z_{i,\text{corrected}} = \frac{(z_i - a_i)}{b_i}$$

The effect of the MSC algorithm is shown in Figure 71.37.

To investigate whether a set of measured spectra have additive or multiplicative effects, it is often instructive to plot the individual wavelengths against the reference (or average) spectrum. If the spectra have multiplicative scattering, they will have different slopes in such a plot. Additive baseline shifts are seen when the spectra intersect differently with the vertical axis.

As the MSC preprocessing uses the mean spectrum as reference for the data set, the assumption is that this calculated mean is representative for all current and future spectra collected using similar experimental conditions. It is also assumed that the observed scatter effects are not significantly associated with the chemical signal that is to be retained. To safeguard against removing the chemical signal along with the scatter effects, it is recommended to keep known important spectral regions or peaks out of the calculations. If such information about chemical signal is sparse, including the full spectra might still work under the assumption that the chemical signal cancels out when seen across the entire wavelength range.

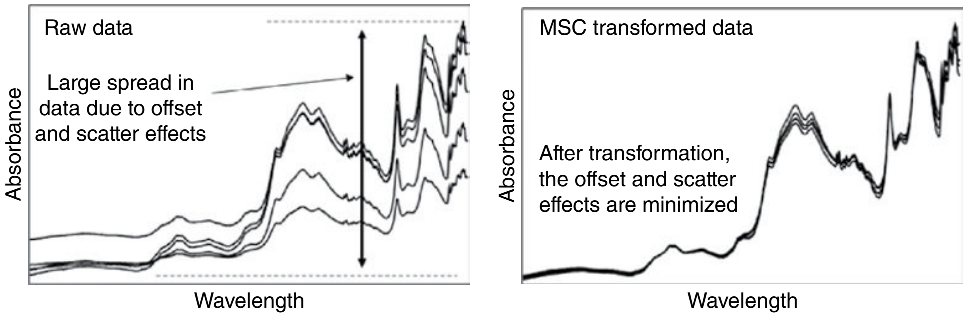


FIGURE 71.37 The application of the MSC algorithm to homogenous data allows the correction of additive and multiplicative effects caused by sample packing and light scatter.

Another way to avoid removing important variation from the data is to include known good spectra in an EMSC model (see section "Extended Multiplicative Signal Correction").

Extended Multiplicative Signal Correction EMSC is an extension to conventional MSC, which is not limited to only removing multiplicative and additive effects from spectra [50]. This extended version allows a better separation of physical light scattering effects from chemical light absorbance effects, by including wavelength-dependent effects or *a priori* information in the modeling.

The MSC model is then extended with additional terms according to the following equation:

$$z_i = a_i + b_i m + d_i \lambda + g_i \lambda^2 + h_i \mathrm{BS} + I_i \mathrm{GS} + m_i m^2 + e_i$$

where d_i, g_i, h_i, I_i, and m_i are additional effects as explained as follows.

Channel Effects Wavelength-dependent light scattering variations are modeled by the third and fourth terms in the equation ($d_i\lambda + g_i\lambda^2$). The assumption is that the wavelength dependency follows a smooth polynomial of first or second order.

Squared Spectrum The reference spectrum is given as a linear term in the original MSC model. A squared term can also be included to allow for nonlinearity ($m_i m^2$).

Good Spectra The MSC and EMSC models fit a regression equation with the included model parameters explaining the different scattering effects. The residual term from the models will therefore contain the chemical signal of interest as well as any unmodeled scattering effects and noise. This is somewhat different from traditional applications of least squares regression where the residuals should be as small as possible. One assumption behind MSC and EMSC is that the chemical signal of interest is poorly correlated with the model parameters that describe the scattering. If such a correlation exists, parts of the chemical signal will also be removed from the model.

One approach to improve the EMSC model is to include a term modeling the chemical signal of interest. This is done with the term I_iGS where the GS is a matrix of good spectra. The coefficient I_i is a vector since there will be one coefficient for each of the good spectra. Good spectra in this context might, for instance, be spectra of pure components present in the samples. For applications where the user has no such information, this term may simply be left out from the model equation.

When good spectra are included, the associated chemical signal will be explicitly modeled and hence not included in the residual term. After estimating the model parameters, the term I_iGS will therefore not be subtracted like the other model terms but retained in the corrected spectra.

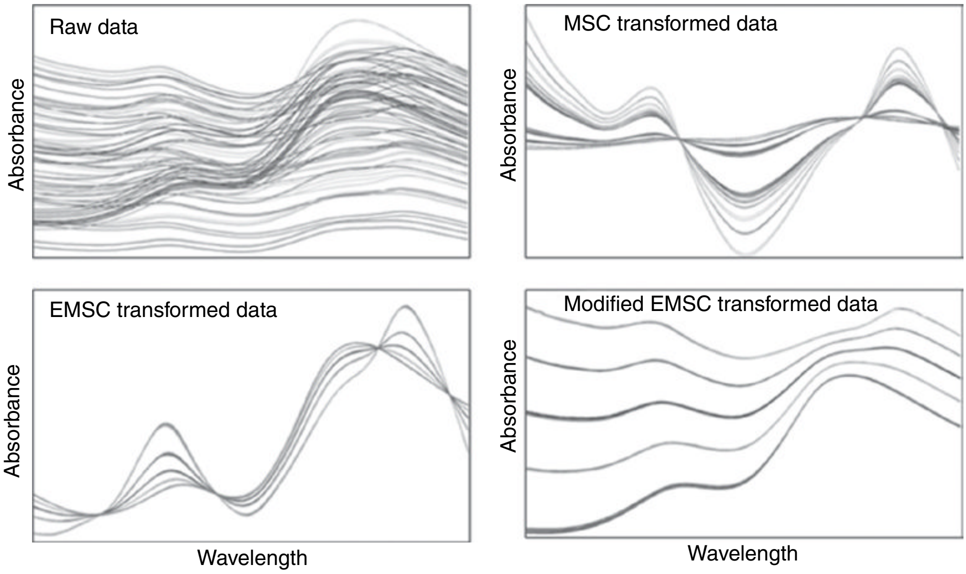


FIGURE 71.38 Effects on nonhomogeneous NIR spectra by the application of MSC, EMSC, and a modified version of EMSC, which takes into account external spectral information.

In a classical experiment conducted by Martens et al. [51] on gluten/starch mixtures, the use of good spectra is presented. The raw NIR data are shown in Figure 71.38 along with the EMSC corrected spectra using a good spectrum.

The good spectrum is a typical example of the underlying chemical constituents and in this case was the difference spectrum of gluten and starch. The overall effect of using a good spectrum is immediately seen in the EMSC corrected data. It is noted here that use of a simple MSC correction was detrimental to the data due to the chemical diversity present in the spectra. In this case, chemical information was confused with scatter information, and this highlights one of the major pitfalls associated with indiscriminate usage of the MSC method.

Bad Spectra Most of the terms in the EMSC equation described so far are used to model specific effects such additive, multiplicative, or linear/quadratic channel effects. However, the scattering may manifest itself in many different ways. To correct for any general scattering effects, bad spectra can be modeled by the term h_iBS. Any spectra that explain the effect that should be corrected for can be included in the algorithm and that has not yet been described by the other terms in the EMSC equation.

In contrast to any good spectra included in the EMSC model, the effects of the bad spectra are subtracted from the input data along with the other estimated parameters.

Reference Spectrum The default and most commonly used reference spectrum is simply the mean of a set of representative spectra. However, it is sometimes more

TABLE 71.1 The Usage and Effect of Preprocessing Techniques Used in NIR Spectroscopy

Transform	Additive	Multiplicative
Derivative	Yes	Yes (second derivative only)
Baseline	Yes	No
SNV	Yes	Yes
Detrending	Yes	No
MSC/EMSC	Yes	Yes

useful to choose a different reference spectrum. If it is known *a priori* that another spectrum is more representative than the mean, this can be used instead.

71.5.4 Preprocessing Summary

Due to the nature of reflectance and diffuse transmission sampling typically used for the collection of NIR spectra, additive and multiplicative effects are a common occurrence and must be minimized in order to develop robust predictive models. In the SW NIR region where transmission measurements are typically performed, additive effects are the major contributor to spectral differences of similar samples. In the longer wavelength region, where diffuse reflectance measurements are typically performed, a combination of additive and multiplicative effects can be observed and in most cases must be jointly corrected for using preprocessing techniques such as MSC or SNV.

Table 71.1 provides an overview of the preprocessing methods commonly used for the minimization of additive and multiplicative effects and briefly describes their intended use.

There also exist many other preprocessing techniques used in NIR spectroscopy; however, this would require an entire chapter in itself to assess, and the interested reader is referred to the excellent literature for more details [52].

71.6 A BRIEF OVERVIEW OF APPLICATIONS OF NIR SPECTROSCOPY

NIR spectroscopy has found application in many industrial and research settings. In its early years, between 1935 and 1980, it was reported that there were 205 publications available [11]. Today, there are thousands of papers and applications with periodicals dedicated to work performed by researchers in NIR.

This section provides a concise overview of some of the industries that have gained the full benefits of the NIR technique. This overview is by no means exhaustive, and the interested reader is referred to the literature references cited in each section for more details of a particular application.

71.6.1 Agricultural Applications

Until the early to mid 1990s, agricultural applications of NIR spectroscopy dominated the literature and practical use of this method. Much emphasis was devoted to this area because the primary methods of analysis available for measuring agricultural products are detailed and time consuming.

NIR is well suited to the analysis of agricultural products since the main constituents of these products are proteins (N—H absorbances), carbohydrates (O—H and C—H absorbances), oils (C—H absorbances), and moisture (O—H absorbances). A major challenge in the development of agricultural methods is sample heterogeneity. Section 71.4 discussed some of the sampling techniques for such materials, and the following sections provide some practical examples of how these have been implemented.

71.6.1.1 At-Line Applications During grain and oilseed harvest time, producers of agricultural products deliver these to receiving points located primarily in rural regions. Before the use of NIR, samples would be taken from each delivery and sent to external laboratories for wet chemical analysis. In peak seasons, this places a tremendous workload on these laboratories and thus delayed the payment of the producer and sometimes leading to great frustration.

The pioneering work of researchers such as Norris [10], Shenk [14], and Williams [12] resulted in the utilization of NIR for the rapid, nondestructive analysis of whole or ground grain at the point of receival.

A number of instrument vendors started to produce instruments based on diffuse transmission and diffuse reflectance that were capable of measuring multiple constituents in grains and oilseeds simultaneously. In the case of whole grain analyzers, samples collected from the delivery truck are placed into a grain hopper, and the instrument scans the entire lot by subsampling. The spectra collected are then averaged, and a quantitative model(s) is applied. The values of the properties of the grain are then used as the payment method, thus relieving the workload of external laboratories as the analysis time required is no longer than 30 s.

In the case of ground grain analyzers, these systems work in reflectance mode and require some sample preparation before analysis. This is typically performed using an industrial grade mill to produce samples of consistent particle size (thus avoiding the need to minimize such variability using preprocessing; see Section 71.5). Modern sampling techniques have also extended the usage of reflectance techniques to measure whole grains, and the use of ground grain analyzers is becoming limited.

A number of instrument vendors offer turnkey solutions for measuring agricultural products through the development of in-house calibration models. One of the challenges of developing models for the agricultural sector is season to season variability. This requires that calibration models must be developed over many years and is one of the major reasons why NIR is often overlooked as a viable alternative to laboratory methods by some as it requires a training phase. Some words of wisdom: day 1 starts when an organization makes a conscious effort to build a model. The longer the delay, the longer it takes to get to day 1!

Some instrument vendors have collected spectra for many years and sent the samples used to collect the spectra to certified laboratories for reference analysis. The calibration models developed are based either on local–global calibration approaches [14] or artificial neural networks (ANN) [53].

Other applications of NIR for agricultural products include the analysis of grains and feed pellets for animal consumption and the nutritional and energy outputs of the batches produced [54]. One group is now using NIR to detect high levels of aflatoxin (an extremely toxic material, which results in many human deaths) in peanuts in third world countries [55]. In the area of rice production, fertilization management is a key economic driver, and the work of Batten [56] involved the development of a leaf nitrogen test to understand the fertilization requirements on rice crops during the growing season.

NIR has also been utilized heavily on the wine and grape sector [57]. During harvest, grape quality parameters can be assessed to determine color (anthocyanin content) and taste (tannins, sugars, total acids) at the point of receival. Attempts to make more automated equipment for the analysis of grapes have met with some technical challenges, and in order to analyze a sample, some detailed preparation work is required.

The next section discusses extensions of the at-line method into real-time processes utilized by the agricultural sector.

71.6.1.2 In-Line Applications The scientific management of fertilization during the growing season is known as precision agriculture [58]. In order to assess the effects of fertilization management on a crop plot, a real-time analysis technology is required; fortunately, this is where NIR has excelled.

Numerous NIR instrument vendors have collaborated with harvest equipment manufacturers to produce systems that can analyze grain and silage during harvest. These NIR instruments are located close to the clean grain elevator or the exit chute of the harvester to make real-time predictions. Since the results of an analysis can be merged with GPS coordinates of the harvester, maps of protein and moisture, for example, can be generated for a particular field, and this information can be used to better understand how to manage fertilizer usage in the coming years and also to understand the water table profile associated with a field. In this way, if fertilizer is placed at the top of a water gradient, it will have more likelihood of dispersing throughout the entire field, rather than just blanket coating the field.

In the area of sugarcane analysis, one vendor [59] has patented a sampling system combined with an overhead NIR analyzer for monitoring parameters such as Brix (sugar), fiber, and other quality attributes. This type of analysis allows for better segregation of materials based on their quality. In the grain industry, traders can use NIR to blend lower-quality grains with premium grains in a way that still meets product specification but allows them to reduce the amount of low-quality material they store.

Figure 71.26 shows some of the in-line setups available for the analysis of agricultural products in real time.

71.6.2 Pharmaceutical/Biopharmaceutical Applications

It is true to say that the pharmaceutical industry (and for that matter any regulated industry) moves at a very slow pace. Quality control (QC) has traditionally been a laboratory-based activity where samples are sent for analysis, and based on the results, raw materials, intermediates, and final products are released.

In the early 2000s, the US FDA released a number of landmark publications to move the pharmaceutical industry into the twenty-first century. The initial work entitled Pharmaceutical cGMPs for the 21st century: A risk-based approach [60] defined a road map for pharmaceutical manufacturers to become innovative in the use of modern, state-of-the-art process analyzers. This publication was followed up soon after with the PAT framework guidance document [61], showing the agency willingness to support the concepts of the cGMPs for the twenty-first-century document. Since then, numerous industry groups and standards organizations have delivered specific guidance to industry, and a more detailed discussion can be found in the review by Swarbrick [18]. NIR in the early days was considered to be the driving force behind the PAT initiative and still forms a major part of it. Today, PAT extends to other spectroscopic methods such as Raman, MIR, and technologies such as in-line particle size analyzers.

71.6.2.1 Raw Material Identification and Conformity Based on Coblentz's early work, he found that molecules exhibit unique features or "fingerprints" in the IR region that can be used to distinguish between them [9]. This observation sets the basis for using MIR for raw material identification, and this method was (and still is) employed by pharmaceutical companies. The biggest issue with the use of MIR methods for raw material identification was that the method tended to be subjective and requires sample preparation. An analyst would typically take a sample (either from a single lot or a composite sample) and prepare either a paraffin-based mull or a potassium bromide disc of the material for MIR analysis. Once the spectrum was collected, it would be visually compared with a reference spectrum of the material kept on file, and a decision would be made on its identity.

One of the major advantages already discussed about NIR spectroscopy is based on it requiring minimal to no sample preparation. Ciurczak [62] in his early work described procedures for developing raw material identification approaches using NIR, and a brief description of the method is given here.

To develop a robust NIR raw material identification library, an analyst will collect between 6 and 15 samples of each of a number of raw materials, and using one of the sampling methods discussed in Section 71.4, samples are presented to the spectrometer in a consistent way such that sampling variations are minimized. Then, using a suitable preprocessing method to minimize the residual sampling effects (refer to

Section 71.5), raw material libraries for each material are developed, using chemometric methods, either based on correlation or Principal Component Analysis (PCA) (refer to Chapter 65).

To test the ability of the libraries for correctly classifying the raw materials, the model is used to predict the samples used to make the libraries. This tests for uniqueness of classification. When ambiguities arise between materials, a method developer must consider the use of additional tests to confirm the identity of the material (i.e., using a simple pharmacopoeial test) or through the use of a hierarchical model that identifies known ambiguities and applies second- or third-level models to resolve the ambiguities.

To test the developed model's ability to assess future materials, a set of data, known as a test set, must be used that contains samples not included in the library. The library must also be tested against negative samples, that is, materials not included in the library to test for their rejection. Once the library has been validated in this way, it can be used as an alternative to pharmacopoeial monographs, as stated in the general chapters on alternative testing methods.

The biggest advantage of using the NIR method for raw material analysis is its objectiveness. If the library is based on a set of known, high-quality materials and has been validated using a suitable test set, the spectra are compared via a chemometric model, and the assessment is based on a statistical, not a visual, basis. A secondary but equally important advantage of NIR for raw material analysis is that the sample analyzed is not destroyed and can be kept as a retention sample for further evidence of material quality in the event an audit or if a process investigation is performed.

NIR is used widely as a raw material identification method, and in some companies the instrument is located outside of the QC laboratory, particularly at the point of receival for instant lot rejection. Other organizations use NIR in the raw material dispensary as a 100% check of lots before they are used for release to production. Other applications of NIR include raw material conformity and process ability prediction. These methods determine whether the material falls within predefined limits (either based on standard deviation limits or multivariate limits) and are used in conjunction with identification results to partition raw material lots to specific products based on knowledge gained over the product history. This is a quality by design (QbD) approach to raw material and process understanding.

71.6.2.2 Intact Tablet Analysis The most common method used to analyze the active pharmaceutical ingredient (API) in tablets, capsules, and gels is by the use of high-performance liquid chromatography (HPLC). This method requires many preparation steps and instrument calibration to known standards before an analysis can be performed. In high-volume pharmaceutical plants, this can put significant pressure on a QC laboratory to release results to keep the manufacturing plant running in an efficient manner.

A number of instrument vendors provide sampling options for measuring pharmaceutical tablets in transmission mode. This can also be performed in diffuse reflectance mode when the API content is high (typically >10% w/w) and where it can be assumed that the API is distributed evenly over the tablet surface. This was previously discussed in Section 71.4.3.4.

In order to develop an intact tablet analysis method, a number of important steps must be followed. These are described as follows:

1. A feasibility study must be performed that shows that the NIR method is both specific and selective for the API over an extended concentration range (typically between 70 and 130% of product label claim).
2. If transmission mode is to be used for the analysis, the tablet thickness must be taken into consideration, and the spectral regions where the API absorbs must show a high-quality signal. Otherwise, if the API concentration is high enough, diffuse reflectance can be considered.
3. Once the feasibility and sampling methods have been performed, the sampling must be optimized, and a calibration set of samples must be chosen over the widest possible production time range of the product. These samples will be highly consistent due to the tight nature of pharmaceutical manufacturing, and they must be extended with samples prepared in the laboratory (using rational designed experiments) to achieve the 70–130% label claim range.
4. Laboratory samples must also be produced at the target label claim range, and all of the laboratory developed tablets should (wherever possible) be produced on the equipment used to make commercial batches.
5. A pool of samples is now developed with a suggested initial target of 180 individual samples. This set will be split 2 : 1 for calibration and test set samples, respectively. The calibration set will contain at least 90 production samples and 30 laboratory-made samples such that their distribution is boxcar in character. The test set will contain 45 production and 15 laboratory samples for the same reasons as for the calibration set.
6. Each sample is now scanned using the spectrometer, and the data are analyzed using PCA (or other techniques) to determine if there is a trend in the active ingredient in the region where the API absorbs and also to confirm that samples produced at the target API level have the same spectral characteristics as the production samples. If this is the case, the laboratory and production samples can be pooled together to make a robust calibration set. Otherwise, further work must be performed on the laboratory samples to make them equivalent to the production samples.
7. Each tablet is now analyzed using the validated reference method that has a known standard error of laboratory (SEL). This is the baseline standard that the NIR model must be compared to. According to PASG and EMA guidelines [1, 63], the standard errors of calibration and prediction (SEC and SEP, respectively) must be less than 1.4 × SEL.

8. Once the individual sample API concentrations have been matched to their NIR spectrum, a chemometric model (based on Partial Least Squares Regression (PLSR), Principal Component Regression (PCR), or MLR) is developed, and the Standard Errors for calibration are compared to the SEL. It is important to note here that the SEC/SEP values can be improved by adding more components to the model; however, the number of components must be commensurate with the complexity of the system. Typically, 1–5 components are the usual number. Since all loadings, loading weights, and regression coefficients of the models must be given a physical interpretation, the use of smaller component numbers is suggested; otherwise the model cannot be used.
9. Once the model has been developed and internally validated using an optimized preprocessing method and test set validation, the model must be applied to new samples to test its ability in a real situation. Since the predicted values of the new samples should span a tight region, the residuals of the predicted values and reference values should be assessed for normality and also to test that no residual exceeds ±3 standard deviations.
10. Once all of the validation procedures have been followed, the method is ready to deploy for real situation usage.

The previous steps are intended to be a guide for a user to follow when developing a new intact tablet (or any method in general). There will always be exceptions or additional steps required, but if this process is followed, reliable methods can be developed with full traceability. The articles by Ritchie [20, 21] and Moffat [64] are also an excellent source of reference for the development of quantitative methods in the pharmaceutical industry. Figure 71.39 shows the typical output of a chemometric package for the development of an intact tablet method.

Recently, a number of instrument vendors have utilized the diffuse reflectance sampling method by installing it into the tablet press and measuring 100% of tablets as they are pressed. This is a very high-speed application and can be used for trending purposes only. Other systems are available that automatically take a random tablet off the production line and measure a number of properties of the tablet, including NIR analysis for a more accurate assessment of the product state during tablet compression.

71.6.2.3 Process Analytical Technology The fundamental premise of the PAT initiative is to gain understanding of the process at the point of manufacture using state-of-the-art, modern process analysis technologies. NIR is an excellent example of a PAT. The following provides a brief list of applications where NIR spectroscopy has been used as a PAT:

1. Granulation: The use of NIR for granulation has been met with success and failure over the years. Optimal placement of probes into a high shear granulator is the key to success, where the measurement of moisture/solvent uniformity can be assessed. The overall granule size and distribution are of key importance. NIR

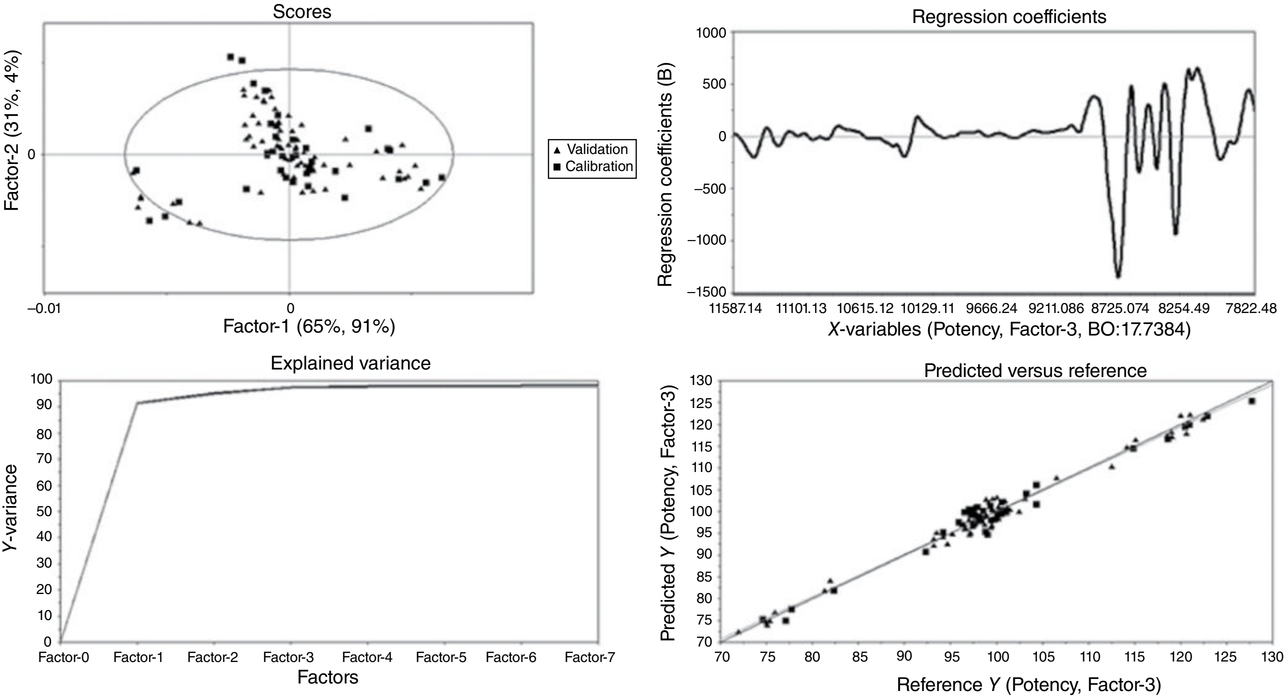


FIGURE 71.39 Results generated by a chemometrics package for the development of a quantitative intact tablet analysis method.

measurements are often supplemented with other measurement systems, including in-line particle size analyzers to provide a measure of uniformity and distribution simultaneously. With the current push by industry to adopt CM [18], granulation has changed from a batch process to a continuous one. This has simplified the measurement of granulation through the use of an analyzer at a fixed point where near 100% measurement of the granulation can be obtained. This all relates to the US FDA's push for QbD and the ability to adjust and adapt a process based on real-time measurement systems.

2. Fluid bed drying (FBD): NIR instrumentation can be integrated directly into an FBD operation to detect drying end points. This avoids overdrying of the granules and thus less potential problems during milling and compression operations. Models can be developed quantitatively (using procedures similar to those outlined in Section 71.6.2.2) or qualitatively by trend analysis, with conformity of PCA-based methods (including batch modeling). Refer to the literature for more information on this application [65].
3. Blend monitoring: NIR analyzers have been developed that integrate directly with tumble and continuous blender systems. They are designed to utilize wireless communication systems, and in the case of a tumble blender, each time the blender rotates a gravity switch or accelerometer triggers the instrument to take a scan when the powder is completely covering the sampling window. This application has been discussed by many authors in the literature [6].
4. Tablet compression: A number of instrument vendors have developed systems for the monitoring of tablets as they are being compressed. In one configuration, the instrument is placed just above the tablet ejection area, and diffuse reflectance measurements are made on tablets in a 100% inspection system. Another configuration utilizes an at-line approach that measures tablets for certain properties at regular intervals. This system was already briefly described in Section 71.6.2.2.
5. Tablet/granule coating: The thickness of a tablet coating is often a critical quality attribute of a pharmaceutical solid dose form for two main reasons. Firstly, for cosmetic reasons, tablet coating provides a quality assurance aspect to the end user when they see tablets evenly coated and thus are acceptable for use. Secondly, coatings may be functional, that is, they contain slow-release properties that may be absorbed by the different parts of the body. Functional coatings may also be applied to granules to achieve differing release profiles. NIR finds use in this application in fluidized bed and pan coaters. The placement of the NIR probe is again the critical factor in coating operations. The probe must be placed where it cannot be coated itself. As the coating builds up on a tablet or granule, the spectral properties of the uncoated tablet change to those of the coating material. Multivariate methods can then be used to classify tablets as being coated or uncoated, and in some cases, the coating thickness can also be monitored.

Figure 71.40 shows the changes in spectral character from coated to noncoated for selected samples during a tablet coating process. The figure also shows how the process can be monitored and controlled multivariately.

6. Packaging: NIR instrumentation can be placed in line with visual inspection tools to determine if the tablet not only has the right appearance but also has the right chemical composition. Samples moving through a packaging system do so at a fast rate; therefore, instrumentation has been developed that can measure 100% of tablets at a high rate and provide the required analysis of surface chemistry. These systems, although they do not provide any intrinsic process understanding in terms of tablet properties and functionality, are often overlooked as "nice to haves" in an operation; however, the detection of one different tablet in a package could be the difference between life and death for a patient in some cases, and the cost of product recall can be much greater than the investment cost of the packaging inspection system in the first place.

It is noted here that the trend toward 100% inspection tools is currently being driven by vendors and proponents of CM systems. A CM manufacturing line is an integrated manufacturing process, joining all of the unit operations together and using PAT to monitor and control the processes, ensuring that the quality of an intermediate leaving one operation is suitable for processing in the next operation. Another initiative in the

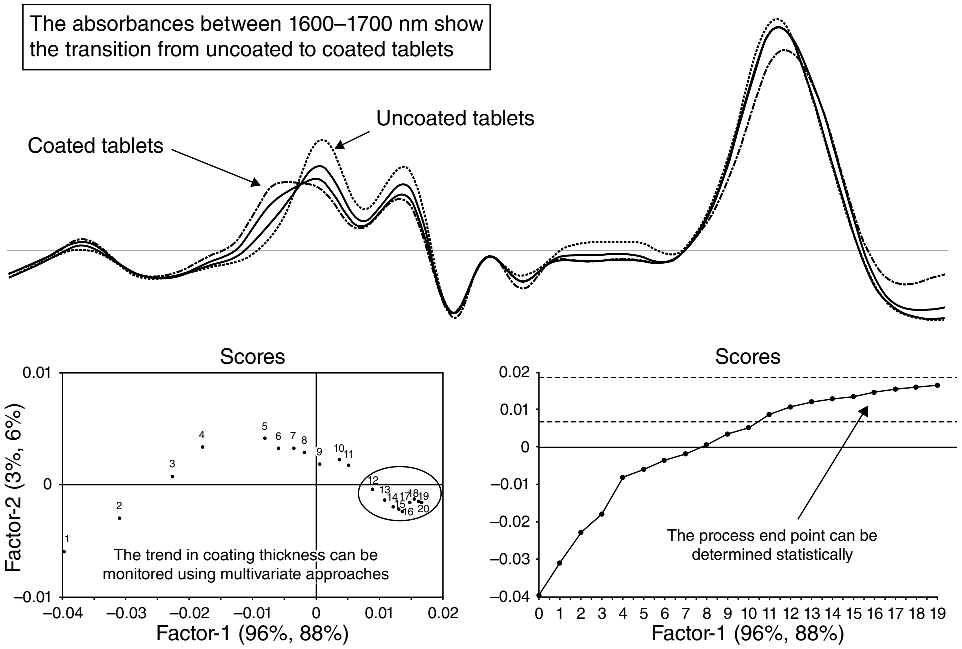


FIGURE 71.40 Spectral changes observed by NIR and monitored multivariately for a tablet coating operation.

pharmaceutical industry is known as QbD. A QbD process is designed to be adaptable to changing material attributes. This topic is briefly discussed in the review by Swarbrick [18], and the interested reader is referred to this article and the references in it.

71.6.2.4 Bioprocesses In recent times, the ability to measure bioprocesses has gained much interest, particularly fermentation reactions where the product value is high and the process can take up to several weeks to complete. There has been much effort directed toward the monitoring of such processes using a combination of process input values and PAT outputs for tracking the trajectory of the so-called golden batch [66].

Many of the bioprocesses currently employed in the biopharmaceutical industry are highly aqueous in nature, and this poses a problem for NIR when the concentration of the proteins being measured becomes too dilute. However, when a specific region has a good S/N ratio, the NIR method can be used to either predict constituent values during the progress of the reaction (by use of hierarchical modeling approaches) or the scores obtained by PCA projection can be combined with process data (using a process called data fusion) to monitor the processes via a physicochemical modeling approach [66]. The major benefit methods such as NIR provides to biopharmaceutical manufacturers are that the process can be controlled analytically and adjustments made to it, as required to bring it back on trajectory (see Figure 71.41 for an example).

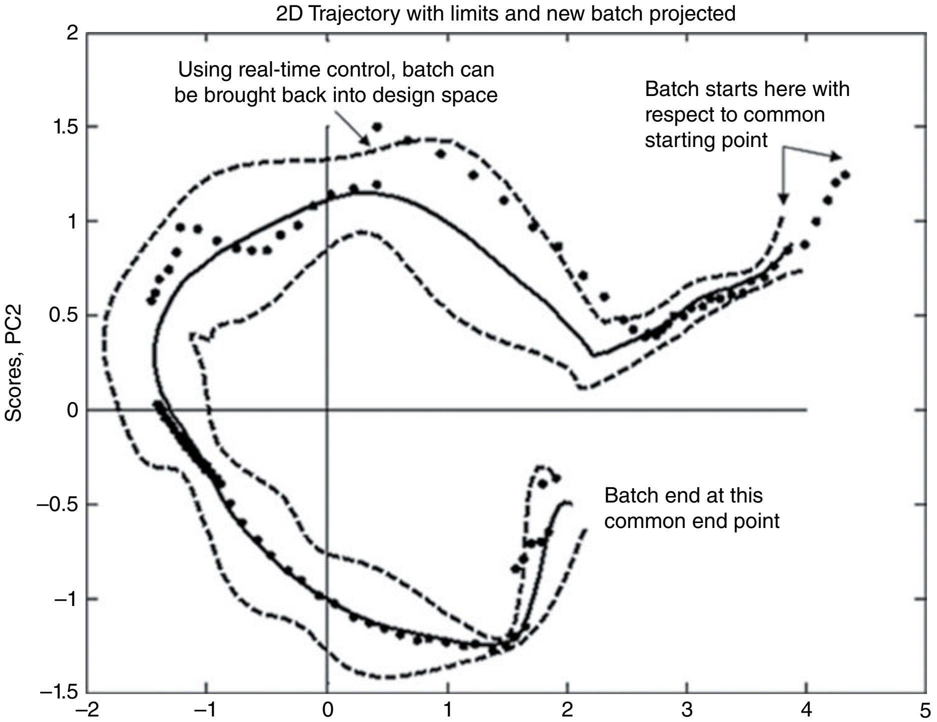


FIGURE 71.41 Process output design space for a fermentation reaction monitored using PAT.

The production of biofuels is also an application where NIR has found suitable application. Since the main product being produced is ethanol, monitoring the evolution of the O—H absorbance makes for an excellent tool for predicting yield and other quality parameters.

71.6.3 Applications in the Petrochemical and Refining Sectors

An oil refinery takes crude oil (a highly complex natural substance) and through the processes of distillation, cracking, isomerization, and alkylation generates base feedstock for many different common products used in everyday lives. Some of the products go to gasoline manufacture; others are sent for polymer production, and NIR has also played a part in the monitoring of such processes, bringing about efficiencies that were not possible without the use of the method.

71.6.3.1 Gasoline Blending With the quality of crude oil declining and the need for more efficiency in the oil refining sector, waste versus profit is a big concern (as it is for any industry). It has been quoted that every 0.2 octane numbers given away can result in a refinery losing anywhere between 50 and 150 K USD per day, based on the refinery output [67].

NIR has long been used to measure multiple properties of gasoline and diesel blends in a manner that reduces such octane (or cetane) giveaway. Depending on the crude oil quality, refineries produce intermediate gasoline products of various octane number values, in general:

1. Light straight run (LSR): This is formed primarily of straight chain alkanes of low octane number. Refineries aim to put as much of this intermediate into final blends as possible as it is a cheap component and can occupy storage tank space quickly if it is not utilized.
2. Aromatics: These are represented mainly by compounds such as benzene, toluene, and xylenes. They are particularly high in octane number, but due to environmental and health organization regulations, there is only a limited amount of aromatics that can be utilized in a gasoline blend. Refineries also like to add aromatics to a gasoline blend as much as legally possible because of their high octane numbers.
3. Isomerates: These are formed by the chemical reorganization of simple alkanes of lower octane number into highly branched compounds that can increase the octane number with respect to the starting materials. These are added to LSR to improve the rating of the final gasoline product.
4. Alkylates: These are highly branched aliphatic compounds that are formed by the reaction of smaller aliphatics (mainly butane) using concentrated hydrofluoric or sulfuric acid. This makes their production a highly dangerous (and expensive)

process, and these compounds have very high octane numbers. They are used in premium-grade gasoline because they provide high octane numbers with no aromatics to worry about.

The task of the refinery now is a balancing act, adds as much LSR and aromatics (within legal bounds), and minimizes the use of isomerates and alkylates such that gasoline of adequate performance characteristics can be produced. NIR has been used both in line and at line for many years to monitor blending of gasoline. Since the monitoring can be performed in real time, some blending systems can adjust the rates of the various components to minimize octane giveaway.

Octane number is not the only property that can be measured; others include benzene content, vapor pressure, alkene content, and many other properties. The analysis is simple; using a liquid transmission device for a laboratory-based system, a sample is typically measured as is, and analysis report is generated. However, when applied in or on line, the need for sample conditioning is sometimes required to remove excess water buildup that may influence the predicted results after application of chemometric models to the NIR spectra. Figure 71.42 shows a general implementation of NIR into a gasoline blending line for monitoring and control of the process.

71.6.3.2 Other Processes There are potentially many uses of NIR spectroscopy for the analysis, monitoring, and control of petrochemical processes. The primary

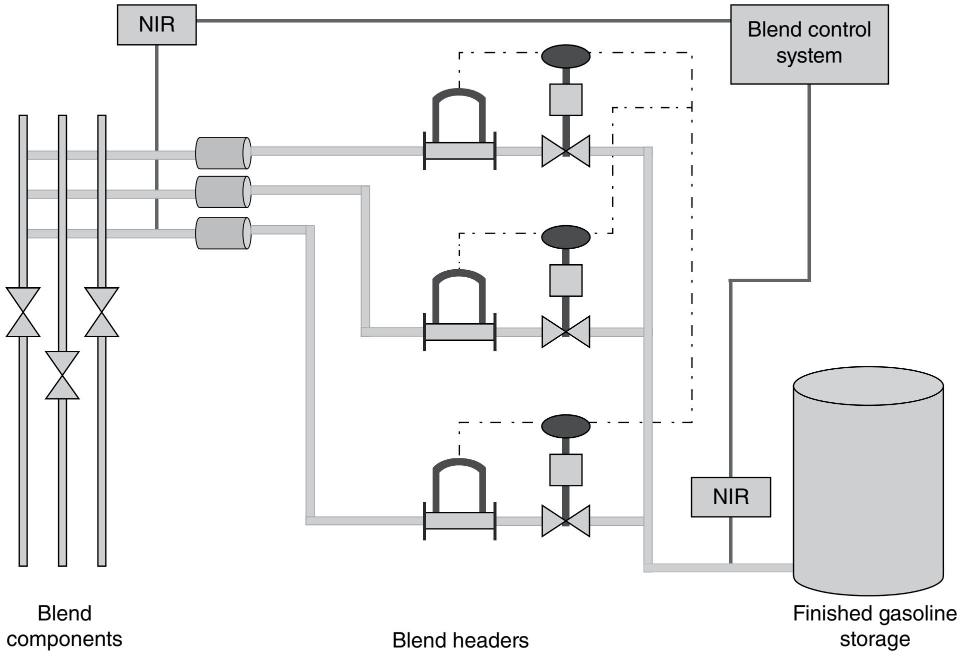


FIGURE 71.42 General implementation of NIR into a gasoline blending operation.

application that dominates in the industry is gasoline blending, but other applications include:

- Measurement of moisture content in hydrofluoric acid for alkylation.
- Measurement of crude oil properties during extraction (particularly for oil sands operations).
- Measurement of input feed, intermediate, and final product output from fluid catalytic cracking (FCC) operations.
- Measurement of sample outputs from the various stages of atmospheric distillation operations allowing assessment of the quality of the various cuts from the crude oil input.

As with the biopharmaceutical applications discussed in Section 71.6.2.4, NIR can be used to monitor chemical reactions in line for the production of specialty chemicals at a petrochemical plant, in particular the production of polymers and related products. The general approach is similar in all industries and applications, find an application suitable for NIR, optimize sampling, and monitor and control in real time.

71.6.4 Applications in the Food and Beverage Industries

Food and beverage applications of NIR are very similar to those discussed in the pharmaceutical industry (Section 71.6.2.). These include but are not limited to:

- Raw material identification and quality, particularly for flours used in baking and spices used in specialty mixtures.
- Product blending of cake and bread mixes and the exact combination of spices for repeatable product quality when used by consumers.
- Monitoring of moisture and food ingredient uniformity, particularly for dough mixing in breads and cake mixtures.

As these situations are very similar in nature to pharmaceutical operations, the reader is referred to the section on pharmaceutical applications for more details. This section will provide a short introduction to the use of NIR for food authentication.

Consumers around the world have come to expect the highest standards in food quality, and in particular, if they purchase what is said to be authentic on the product label, they expect it to be that product. However, in today’s society, there is a minor element who would substitute nonauthentic foods as authentic to gain superior profits from inferior products. NIR has been used for the authentication of foods and wines for a number of years now.

Gerard Downey of the Agriculture and Food Development Authority in Ireland (Teagasc) has been active for many years in the research of NIR for food authentication. The following provides a short list of the applications where NIR has been successfully used:

- Adulteration of olive oil with sunflower oil in Mediterranean countries [68]
- Authentication and origin of honey [69]
- Authentication of varietal origins for the classification of commercial white wines [70]
- Authentication of coffee bean varieties [71]

The previous list of selected applications shows that the main areas of food authentication are centered around luxury and fine foods. With the introduction of micro-NIR systems, the future may see the development of consumer-based NIR devices attached to smartphones that will be able to perform these types of analyses at the point of sale and verify that what is being paid for is what is being received.

71.7 SUMMARY AND FUTURE PERSPECTIVES

NIR instrumentation has seen a rapid rise in development and application particularly since the mid 1990s when computing power and chemometric methods allowed for the rapid capture and modeling of data for predictive and classification purposes.

Initially thought of as a tool for agricultural applications due to its nondestructive nature and ability to handle heterogeneous samples, other industries, particularly the pharmaceutical/biopharmaceutical and petrochemical industries, have taken full advantage of the in-line implementation options available.

In the pharmaceutical industry, the PAT initiative in its early days was based primarily around NIR for at-line and in-line applications. Other industries such as the food and beverage, fine chemical, textile, and renewable energy sectors have also utilized the benefits of NIR for business critical applications.

The future of NIR is looking very bright with the introduction of microspectrometers. These devices will find use in consumer applications where they will be integrated with mobile devices such as smartphones for applications such as personalized medicine, food authentication, and even gasoline integrity. By integrating these microanalyzers with smartphones, applications can be written that will access libraries over the Internet for the rapid identification of materials in an easy and reliable manner. This opens up many applications for national security, food integrity, process analysis, and medical applications. This vision of the future is the same vision Hirschfeld shared in his 1985 paper [24].

71.8 TERMINOLOGY

ANISG Australian Near Infrared Spectroscopy Group.

ANN Artificial neural network refers to a set of nonlinear mathematical techniques used for the development of classification and prediction models based on machine learning algorithms.

AOTF Acoustooptical tunable filter refers to a wavelength dispersion system that uses a tellurium oxide crystal and a radio-frequency source to selectively tune the crystal to measure sharp spectral bands in the NIR region.

API Active pharmaceutical ingredient refers to the active component(s) in a pharmaceutical formulation.

ASTM American Society for Testing and Materials.

ATR Attenuated total reflectance refers to a sampling device that utilizes a chemically inert crystal of precise design to make close contact with a sample surface. When electromagnetic radiation is passed through the crystal, a series of internal reflections are set up at the crystal boundary. The light at the surface of the crystal interacts with the sample, and the chemical information is transmitted back to the detector to generate an NIR spectrum.

AU Absorbance units refers to a unitless measure of spectral intensity measured on a logarithmic scale.

CM Continuous manufacturing refers to a flow production method used to manufacture, produce, or process materials without interruption.

EMA European Medicines Agency.

EMSC Extended Multiplicative Scatter/Signal Correction refers to a scatter correction technique that uses MSC as its foundation and extends it with wavelength-dependent correction terms and can also be extended to include the use of good and bad spectra, thus making the correction a physicochemical model.

FBD Fluid bed drying refers to a process where air of controlled temperature is passed through the bottom of a bed of wet powder in order to create a fluidized motion that simultaneously dries and regulates granule size.

FCC Fluid catalytic cracking refers to a process used in the petrochemical industry for taking heavy vacuum distillates and cleaving them through catalytic processes into smaller, more volatile fractions for use in gasoline blending.

FFT Fast Fourier transform refers to a mathematical procedure that converts a time-domain signal into a frequency-domain signal. In the case of NIR, the frequency-domain signal is a near-infrared spectrum.

FIR Far infrared refers to the region of the electromagnetic spectrum between 16,000 and 111,000 nm.

FT-NIR Fourier transform near-infrared spectroscopy refers to a method of spectral collection that utilizes an interferometer as the wavelength dispersion device. The signal sent to the detector is a time-dependent signal. The instrument uses a fast Fourier transform to convert the time-domain signal into a frequency-domain signal, thereby generating a near-infrared spectrum.

FWHM Full Width at Half Maximum refers to the width of a spectral band at the half-way point on the *y*-axis. This is used to understand the size of any smoothing window used in derivatives such that the band is not overly distorted or smoothed out.

GMP Good Manufacturing Practice refers to a set of guidelines used to ensure that a manufacturer consistently meets its quality and safety targets for the products it manufactures.

GUI Graphical user interface refers to a system that displays the results generated by an instrument in a form that is interpretable to an end user.

HPLC High-performance liquid chromatography refers to an analytical procedure that separates and quantifies specific components of a mixture using a purpose-designed separation column and an appropriate mobile phase to affect the separation.

ICNIRS International Conference on Near Infrared Spectroscopy

IR Infrared refers to any part of the electromagnetic spectrum that produces thermal energy, that is, NIR, MIR, and FIR.

LOD Limit of detection refers to a statistical analysis that determines at which point a sample signal is distinguishable from background noise. This is typically measured as 3σ above the baseline noise.

LOQ Limit of quantification refers to the statistical assessment of a sample signal that can be reliably used for building quantitative models. This is typically measured as 10σ above the baseline noise.

LSR Light straight run refers to a gasoline cut typically of low boiling point and low octane number. This is utilized in gasoline blending to minimize the use of expensive fractions such as isomerates and alkylates.

LVF Linear Variable Filter refers to a wavelength dispersion device that utilizes a wedge-shaped filter to separate polychromatic light into its monochromatic components, which are then detected on a diode array detector.

LW Long wave refers to the region of the NIR spectrum between 1100 and 2500 nm. This is primarily where diffuse reflectance measurements are performed.

MEMS Microelectromechanical spectrometer refers to a wavelength dispersion system that uses two parallel dielectric mirrors, built on a microchip substrate to generate a tunable Fabry–Perot filter. Through the application of electrical voltage, the distance between the dielectric mirrors can be controlled resulting in the generation of an NIR spectrum.

MIR Mid-infrared refers to the region of the electromagnetic spectrum between 2,500 and 16,000 nm.

MLR Multiple linear regression refers to a multivariate regression technique that fits a small number of terms (typically <20) in order to fit a linear model that can be used to predict properties of new samples when measured by the variable in the model.

MSC Multiplicative Scatter Correction, sometimes referred to a Multiplicative Signal Correction, refers to a scatter correction method that corrects for both additive and multiplicative effects by normalizing data to a mean or reference spectrum.

MVA Multivariate analysis refers to a set of mathematical tools used to analyze more than one variable simultaneously in order to understand variable and sample relationships and also to build regression and classification models that utilize the most important parts of a data set.

NIR Near infrared refers to the region of the electromagnetic spectrum between 700 and 2500 nm.

NMR Nuclear magnetic resonance refers to a type of spectroscopy that utilizes a strong magnetic field and radio wave pulses to excite typically protons in a molecule and measure their relaxation times. The result is a highly detailed spectrum that allows an analyst to structurally elucidate the molecule in the sample.

PASG Pharmaceutical Analytical Sciences Group.

PAT Process analytical technology refers to an initiative of the US FDA that encourages pharmaceutical and related industries to adopt novel and state-of-the-art process technologies for better understanding of processes and products. NIR has been the major PAT tool used to date.

PCA Principal Component Analysis refers to an exploratory data analysis method used to understand complex sample and variable relationships in multivariate data. The aim of PCA is to isolate those variables that most contribute to sample patterns observed. PCA models can be further used for developing classification models or predicting the state of new samples with respect to the original model.

PCR Principal Component Regression refers to a method for relating the variations in a response variable (*Y*-variable) to the variations of several predictors (*X*-variables). This method performs particularly well when the various *X*-variables express common information, that is, when there is a large amount of correlation. Principal Component Regression is a two-step method. First, a Principal Component Analysis is carried out on the *X*-variables. The principal components are then used as predictors in a multiple linear regression.

PLSR Partial Least Squares Regression refers to a method for relating the variations in one or several response variables (*Y*-variables) to the variations of several predictors (*X*-variables). This method performs particularly well when the various *X*-variables express common information, that is, when there is a large amount of correlation between the variables.

QbD Quality by design refers to a global pharmaceutical initiative that builds quality into the process such that the overall integrity of a batch can be assured through direct process monitoring and control. Changes to the process are allowed without any deviation reports being generated as long as the process is maintained in its design space. NIR is a PAT tool that enables organizations to monitor and control processes within the design space.

QC Quality control refers to protocols used to determine the fitness for purpose of a product before it is released to the next step of its life cycle.

RF Radio frequency refers to electromagnetic waves generated in the radio wave region of the spectrum, typically of wavelengths between centimeters and meters.

SEC Standard error of calibration refers to the variation in the precision of calibration sample predictions over several samples. SEC is computed as the standard deviation of the prediction residuals and is dependent of the validation method used.

SEL Standard error of laboratory refers to the baseline precision of a reference analytical method used to calibrate a secondary method such as NIR. The precision of the secondary method must be statistically equivalent to the SEL in order for it to be used as an alternative method.

SEP Standard error of prediction refers to the variation in the precision of predictions over several samples. SEP is computed as the standard deviation of the prediction residuals when test set validation is the validation method used.

S/N Signal-to-noise ratio refers to that part of a measurement signal that can be statistically differentiated from baseline noise. This is often used in limit of detection (LOD) and limit of quantification (LOQ) studies.

SNV Standard Normal Variate refers to a scatter correction technique commonly used in NIR spectroscopy. It used the spectrum itself to correct for both additive and multiplicative effects by subtracting the grand spectrum mean from all points and dividing them by the spectra overall standard deviation. The end result is a spectrum with its center of gravity at zero and a spread of absorbances between ±3 standard deviations around the mean.

SW Short wave refers to the region of the NIR spectrum between 700 and 1100 nm. Sometimes called the Herschel region, this is primarily the region where transmission spectra are collected.

TOS Theory of Sampling refers to a set of protocols that ensure representative sampling techniques are implemented and adhered to for a particular measurement system.

USDA US Department of Agriculture.

US FDA US Food and Drug Administration.

REFERENCES

1. European Medicines Agency, “Guideline on the Use of Near Infrared Spectroscopy (NIRS) by the Pharmaceutical Industry and the Data Requirements for New Submissions and Variations”, EMEA/CHMP/CVMP/QWP/17760/2009 Rev2, 2014.
2. United States Food and Drug Administration, “Development and Submission of Near Infrared Analytical Procedures, Guidance for Industry”, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM440247.pdf accessed June 3, 2015.
3. Richardson, A. D., Reeves III, J. B., and Gregoire, T. G., “Multivariate analyses of visible/near infrared (VIS/NIR) absorbance spectra reveal underlying spectral differences among dried, ground conifer needle samples from different growth environments”, *New Phytol.*, 161(1), 2004, pp 291–301.
4. Curcio, J. A. and Petty, C. C., “The near infrared absorption spectrum of liquid water”, *J. Opt. Soc. Am.*, 41(5), 1951, pp 302.
5. Herschel, W., “Investigation of the powers of the prismatic colours to heat and illuminate objects; with remarks, that prove the different refrangibility of radiant heat. To which is added, an inquiry into the method of viewing the sun advantageously, with telescopes of large apertures and high magnifying powers”, *Phil. Trans. R. Soc.*, 90, 1800, pp 225.
6. Ciurczak, E. W. and Drennen III, J. K., *Pharmaceutical and Medical Applications of Near Infrared Spectroscopy*, Practical Spectroscopy Series Vol 31, Marcel Dekker Inc., New York, 2002, pp 1.
7. Hindle, P. H., “Historical Development”, in *Handbook of Near Infrared Analysis*, 3rd Ed, eds. Burns, D. A. and Ciurczak, E. W., CRC Press, Boca Raton, FL, 2008, pp 3.
8. Davies, A. M. C., “An Introduction to Near Infrared Spectroscopy”, https://www.impublications.com/content/introduction-near-infrared-nir-spectroscopy accessed June 3, 2015.
9. Coblentz, W. W., *Investigation of Infrared Spectra, Part I*, Carnegie Institution, Washington, DC, 1905.
10. Ben-Gera, I. and Norris, K., “Direct spectrophotometric determination of fat and moisture in meat products”, *J. Food Sci.*, 33, 1968, pp 64.
11. Burns, D. A. and Margoshes, M., “Historical Development”, in *Handbook of Near Infrared Spectroscopy*, Practical Spectroscopy Series Vol 13, eds. Burns, D. A. and Ciurczak, E. W., Marcel Dekker Inc., New York, 1992, pp 3.
12. Williams, P. C., “Implementation of Near-Infrared Spectroscopy”, in *Near Infrared Technology in the Agricultural and Food Industries*, 2nd Ed, eds. Norris, K. and Williams, P. C., AACC, Saint Paul, MN, 2001, pp 145–170.
13. Osborne, B. G., “Principles and practice of near-infrared (NIR) reflectance analysis”, *Int. J. Food. Sci. Tech.*, 16(1), 1981, pp 13–19.
14. Shenk, J. S. and Westerhaus, M. O., “Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy”, *Crop Sci.*, 31(2), 1990, pp 469–474.
15. Flinn, P. C. and Downes, G. J., “The importance of Near Infrared Spectroscopy in Deciding Appropriate Feeding Strategies for Australian Livestock”, in *Near Infrared Spectroscopy:*

the Future Waves, eds. Davies, A. M. C. and Williams, P., NIR Publications, Chichester, 1996, pp 512.

16. Blakeney, A. B. and Flinn, P. C., “Determination of non-starch polysaccharides in cereal grains with near-infrared reflectance spectroscopy”, *Mol. Nut. Food. Res.*, 49(6), 2005, pp 546–550.
17. Batten, G. D., “Forestry and the Environment: Challenges for Near Infrared Spectroscopy”, in *Near Infrared Spectroscopy: proceedings of the 11th International Conference*, eds. Davies, A. M. C. and Garrido-Varo, A., NIR Publications, Chichester, 2004, pp 749.
18. Swarbrick, B., “Review: advances in instrumental technology, industry guidance and data management systems enabling the widespread use of near infrared spectroscopy in the pharmaceutical/biopharmaceutical sector”, *J. Near Infrared Spectrosc.*, 22(3), 2014, pp 157–168.
19. Mark, H., *Principles and Practice of Spectroscopic Calibration*, John Wiley & Sons, New York, 1991.
20. Mark, H., Ritchie, G. E., Roller, R. W., Ciurczak, E. W., Tso, C., and MacDonald, S. A., “Validation of a near-infrared transmission spectroscopic procedure, part A: validation protocols”, *J. Pharm. Biomed. Anal.*, 28(2), 2002, pp 251–260.
21. Ritchie, G. E., Roller, R. W., Ciurczak, E. W., Mark, H., Tso, C., and MacDonald, S. A., “Validation of a near-infrared transmission spectroscopic procedure, part B: application to alternate content uniformity and release assay methods for pharmaceutical solid dosage forms”, *J. Pharm. Biomed. Anal.*, 29(1–2), 2002, pp 159–271.
22. International Council for Near Infrared Spectroscopy (ICNIRS), http://icnirs.org/ accessed June 3, 2015.
23. Shapiro, H. M., “In memoriam, Tomas Hirschfeld (1939–1986)”, *Cytometry*, 7(5), 1986, pp 399.
24. Hirschfeld, T., “Instrumentation in the next decade”, *Science*, 230, 1985, pp 286–291.
25. The Australian Near Infrared Spectroscopy Group, http://www.anisg.com.au/ accessed June 3, 2015.
26. Norris, K. and Williams, P. C., *Near Infrared Technology in the Agricultural and Food Industries*, 2nd Ed, AACC, Saint Paul, MN, 2001.
27. McClure, W. F. and Tsuchikawa, S., “Instruments”, in *Near-Infrared Spectroscopy in Food Science and Technology*, eds. Yukihiro Ozaki, W. F. M. and Alfred, A. C., John Wiley & Sons, Inc., Hoboken, NJ, 2006, pp 75–108.
28. Skoog, D. A., West, D. M., and Holler, F. J., *Fundamentals of Analytical Chemistry*, 5th Ed, W. B. Saunders Company, New York, 1988, pp 466–472.
29. Tipler, P. A., *Physics*, 2nd Ed, Worth Publishers Inc., New York, 1982, pp 921–923.
30. Crocombe, R. A., “Miniature optical spectrometers, Part III: conventional and laboratory near-infrared spectrometers”, *Spectroscopy*, 23(5), 2008, pp 40.
31. Jacquinot, P., “Caractères communs aux nouvelles mèthodes de spectroscopie interfèrentielle; Facteur de mèrite”, *J. Phys. Radium*, 19, 1958, pp 233.
32. Tipler, P. A., *Physics*, 2nd Ed, Worth Publishers Inc., New York, 1982, pp 905–906.
33. Brown, S. D., “Signal Processing and Digital Filtering”, in *Practical Guide to Chemometrics*, ed. Gemperline, P., Taylor & Francis Group, Boca Raton, FL, 2006, pp 389.

34. Fellgett, P., “A propos de la théorie du spectromètre interfèntiel multiplex”, *J. Phys. Radium*, 19, 1958, pp 187.

35. Connes, J. and Connes, P., “Near infrared planetary spectra by Fourier spectroscopy I, instruments and results”, *J. Opt. Soc. Am.*, 56(7), 1966, pp 896.

36. Berntsson, O., Danielsson, L.-G., and Folestad, S.,“Characterization of diffuse reflectance fiber probe sampling on moving solids using a Fourier transform near-infrared spectrometer”, *Anal. Chim. Acta*, 431(1), 2001, pp 125.

37. IMP, “NIR Products at Pittcon 2015”, https://www.impublications.com/content/nir-products-pittcon-2015 accessed June 3, 2015.

38. Harris, S. E. and Wallace, R. W., “Acousto-optic tuneable filter”, *J. Opt. Soc. Am.*, 59, 1969, pp 744.

39. Bragg, W. H. and Bragg, W. L., “The reflexion of X-rays by crystals”, *Proc. R. Soc. Lond. A*, 88(605), 1913, pp 428–438.

40. Crocombe, R. A., “Miniature optical spectrometers: the art of the possible, Part IV: new near-infrared technologies and spectrometers”, *Spectroscopy*, 23(6), 2009, pp 26.

41. Norris, K. H. and Hart, J. R., “Direct spectrophotometric determination of moisture content of grain and seeds”, *J. Near Infrared Spectrosc.*, 4, 1996, pp 23–30.

42. O’Brien, N. A., Hulse, C. A., Friedrich, D. M., Van Milligen, F. J, von Gunten, M. K., Pfeifer, F., and Siesler, H. W., “*Miniature Near-Infrared (NIR) Spectrometer Engine for Handheld Applications*”, eds. Druy, M. A. and Crocombe, R. A., Society of Photo-optical Instrumentation Engineers, Washington, DC, 2012, pp 837404-1-8.

43. Esbensen, K. H. and Paasch-Mortensen, P., “Process Sampling: Theory of Sampling, the Missing Link in Process Analytical Technologies (PAT)”, in *Process Analytical Technology*, ed. Bakeev, K., John Wiley & Sons, Ltd, Chichester, 2010, pp 37–80.

44. Savitzky, A. and Golay, M. J. E., “Smoothing and differentiation of data by simplified least squares procedures”, *Anal. Chem.*, 36, 1964, pp 1627–1639.

45. Steiner, J., Termonia, Y., and Deltour, J., “Comments on smoothing and differentiation of data by simplified least squares procedure”, *Anal. Chem.*, 44, 1972, pp 1906–1909.

46. Hopkins, D. W., “What is a Norris derivative?”, *NIR News*, 12(3), 2001, pp 3–5.

47. Norris, K., “Applying Norris derivatives. Understanding and correcting the factors which affect diffuse transmittance spectra”, *NIR News*, 12(3), 2001, pp 6.

48. Barnes, R. J., Dhanoa, M. S., and Lister, S. J., “Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra”, *Appl. Spectrosc.*, 43(5), 1989, pp 772–777.

49. Martens, H., Jensen, S. Å., and Geladi, P., *Multivariate linearity transformation for near-infrared reflectance spectrometry* in Proc. Nordic. Symp. Appl. Stat., Stockland Forlag Publ, Stavanger, Norway, 1983, pp 205–234.

50. Martens, H. and Stark, E., “Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy”, *J. Pharm. Biomed. Anal.*, 9, 1991, pp 625–635.

51. Martens, H. Nielsen, J. P., and Engelsen, S. B., “Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures”, *Anal. Chem.*, 75, 2003, pp 394–404.

52. Gemperline, P., "*Practical Guide to Chemometrics*", 2nd Ed, Taylor & Francis Group, Boca Raton, FL, 2006.
53. Hastie, T., Tibshirani, R., and Friedman, J., "*The Elements of Statistical Learning, Data Mining, Inference, and Prediction*", 2nd Ed, Springer Science + Business Media LLC, New York, 2009, pp 389–416.
54. Flinn, P., "NIR: A Key Component of the Premium Grains for Livestock Project", Grain Industries Center for NIR, 6th Annual Meeting for Participants, Canberra, 2001, pp 32.
55. Fox, G. and Manley, M., "Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals", *J. Sci. Food Agric.*, 94(2), 2014, pp 174–179.
56. Batten, G. D., Blakeney, A. B., Glennie-Holmes, M., Henry, R. J., McCaffery, A. C., Bacon, P. E., and Heenan, D. P., "Rapid determination of shoot nitrogen status in rice using near infrared reflectance spectroscopy", *J. Sci. Food Agric.*, 54(2), 1991, pp 191–197.
57. Dambergs, R. G., Cozzolino, D., Cynkar, W. U., Esler, M. B., Janik, L. J., Francis, I. L., and Gishen, M., "Strategies to Minimise Matrix-Related Error with Near Infrared Analysis of Wine Grape Quality Parameters", in *Near Infrared Spectroscopy: proceedings of the 11th International Conference*, eds. Davies, A. M. C. and Garrido-Varo, A., NIR Publications, Chichester, 2004, pp 183.
58. McBratney, A., Whelan, B., and Ancev, T., "Future directions of precision agriculture", *Precis. Agric.*, 6, 2005, pp 7–23.
59. JEFFRESS, "PRODUCTS—Cane Analyser—InfraCana® II IC02", http://www.jeffress.com.au/products_ic02.htm, accessed June 3, 2015.
60. US FDA, "Pharmaceutical CGMPs for the 21st Century—A Risk Based Approach, Final Report", 2004, http://www.fda.gov/drugs/developmentapprovalprocess/manufacturing/questionsandanswersoncurrentgoodmanufacturingpracticescgmpfordrugs/ucm137175.htm accessed June 3, 2015.
61. US FDA, "Guidance for Industry: PAT—a Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance", 2004, http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070305.pdf accessed June 3, 2015.
62. Ciurczak, E. W., "Following the progress of a pharmaceutical mixing study via near-infrared spectroscopy", *Pharm. Tech.*, 15(9), 1991, pp 141.
63. Broad, N., Graham, P., Hailey, P., Hardy, A., Holland, S., Hughes, S., Lee, D., Prebble, K., Salton, N., and Warren, P., "Guidelines for the Development and Validation of Near-Infrared Spectroscopic Methods in the Pharmaceutical Industry", in *Handbook of Vibrational Spectroscopy*, eds. Chalmers, J. M. and Griffiths, P. R., John Wiley & Sons, New York, 2002.
64. Moffat, A. C., Trafford, A. D., Jee, R. D., and Graham, P., "Meeting the international conference on harmonisation's guidelines on validation of analytical procedures: quantification as exemplified by a near-infrared reflectance assay of paracetamol in intact tablets", *Analyst*, 125, 2000, pp 1341.
65. Barla, V. S., Kumar, R., Nalluri, V. R., Gandhi, R. R. and Venkatesh, K., "A practical evaluation of qualitative and quantitative chemometric models for real-time monitoring of moisture content in a fluidized bed dryer using NIR technology", *J. Near Infrared Spectrosc.*, 22(3), 2014, pp 221–228.

66. Westad, F., Swarbrick, B., Gidskehaug, L., and Flaaten, G. R., “Assumption free modeling and monitoring of batch processes”, *Chemom. Intell. Lab. Syst.*, 149, 2015, pp 66–72.

67. Emerson Process Management, “Refining”, http://www2.emersonprocess.com/siteadmin center/PM%20Micro%20Motion%20Documents/Refining-Blending-PSG-MC-00796.pdf accessed June 3, 2015.

68. Downey, G., Mcintyre, P., and Davis, A. N., “Detecting and quantifying sunflower oil adulteration in extra virgin olive oils from the eastern mediterranean by visible and near-infrared spectroscopy”, *J. Agric. Food Chem.*, 50(20), 2002, pp 5520–5525.

69. Hennessy, S., Downey, G., and O’Donnnell, C., “Multivariate analysis of ATR/FT-IR spectroscopic data to confirm the origin of honeys”, *Appl. Spectrosc.*, 62(10), 2008, pp 1115–1123.

70. Cozzolino, D., Smyth, H. E., and Gishen, M., “Feasibility study on the use of visible and near infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins”, *J. Agric. Food Chem.*, 51, 2003, pp 7703–7708.

71. Downey, G. and Boussion, J., “Authentication of coffee bean variety by near-infrared reflectance spectroscopy of dried extract”, *J. Sci. Food Agric.*, 71(1), 1996, pp 41–49.

72

NANOMATERIALS PROPERTIES

Paul J. Simmonds
Departments of Physics & Materials Science and Engineering, Boise State University, Boise, ID, USA

72.1 INTRODUCTION

In general, a nanomaterial is defined as possessing functional structural features that are less than 100 nanometers (nm) in size (1 nm = 10^{-9} m). Nanomaterials exist widely in nature and give certain organisms their unique characteristics. For example, nanostructured hairs on the legs of spiders allow them to walk up walls, while waxy nanostructures give the leaves of plants such as the lotus their striking ability to repel water (Fig. 72.1).

It is, however, the recent development of artificial nanomaterials with tunable properties that has created so much interest and excitement in this field. Whole industries have grown up around nanomaterials, and entire university departments are devoted to their study. Nanomaterials are now essential to fields as disparate as telecommunications, medicine, renewable energy, and quantum computation. Arguably, materials containing structures on other length scales (e.g., milli- or micromaterials) have not made such a huge impact. So what is it about nanomaterials and their properties that make them so important? In order to answer this question, in this chapter we will explore the following three topics:

1. How and why certain properties of materials change as we reduce them in size from bulk (i.e., macroscopic scale) to the nanoscale.
2. How the behavior of certain materials is affected or dictated by the properties of the nanoscale features and structures that comprise them.
3. How nanomaterial properties can be controlled, for example, by engineering size, composition, or structure.

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

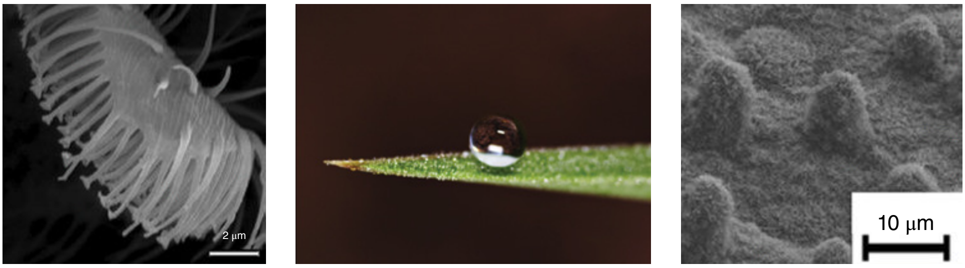


FIGURE 72.1 (Left) Hairs on the leg of a jumping spider, *Evarcha arcuata*. Each of these tiny hairs terminates in a contact pad nanostructure, which, by exploiting van der Waals attraction, allows the spider to crawl on ceilings and walls. Source: Kesel *et al.* [1]. Reproduced with permission of IOP Publishing. (Center) Water landing on the leaves of certain plants balls up into pseudospherical droplets that readily roll off, carrying with them any dirt particles that could otherwise inhibit photosynthesis. Source: http://upload.wikimedia.org/wikipedia/commons/8/8c/Dew_2.jpg. CC-BY-SA-3.0. (Right) This superhydrophobic property is a direct result of the presence of waxy nanostructures that cover the surface of the leaves: in this case, the lotus. Source: Cheng *et al.* [2]. Reproduced with permission of IOP Publishing.

Picture a cube of some material, say a metal or semiconductor, the sides of which are a few centimeters in size. We can easily pick up this macroscopic object and measure its various physical properties, for example, its Young's modulus, magnetic state, or melting point. Now imagine reducing the length of each side of that cube. First we would pass through the milliscale (10^{-3} m), followed by the microscale (10^{-6} m). Continuing to shrink the size of the cube, we would arrive at the *nano*scale. At each scale we could repeat the measurements to see if any of the material's properties had changed.

What we would find is that many of the cube's material properties do not change in any meaningful way at either the milli- or microscales. This is not to say that there are not plenty of unique and important applications for milli- and micromaterials. In general, however, their properties are simply scaled down from the macroworld, and it is simply the fact that they are smaller that makes them useful. For example, the first *micro*chips were created in the 1970s when transistors became small enough that they could be packed together in high-density arrays.

In contrast, upon reaching the nanoscale we see a much more abrupt and important shift in the cube's material properties. Certain characteristics become profoundly different from what they were when the object was larger, and the cube may take on new and sometimes surprising behaviors as a result. It is arguably these properties, that don't scale but change in a fundamental way when one descends to the nanoscale, which are the most interesting.

Due to the enormous breadth of modern nanomaterials, we cannot hope to discuss every family of nanomaterials in a single chapter. Inevitably this means that there are several major areas of current research that I have had to exclude in the interest of

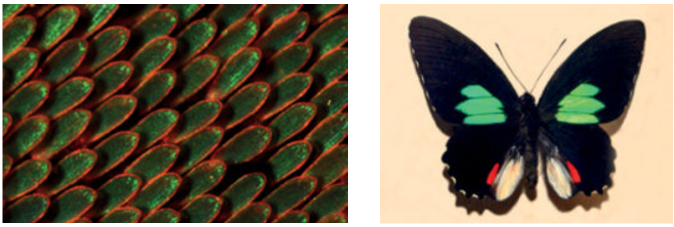


FIGURE 72.2 Naturally occurring photonic crystal nanostructures in the wings of a butterfly, *Parides sesostris zestos* (left). Source: Saranathan *et al.* [3]. Reproduced with permission of PNAS. The structure of this particular photonic crystal gives it the property of absorbing red and blue light, meaning that only green light is reflected. This gives the butterfly its colorful green markings, seen here (right) as three bright stripes on each wing. Source: http://commons.wikimedia.org/wiki/File%3APapilionidae_-_Parides_sesostris_zestos.JPG. CC-BY-SA-3.0.

brevity. Perhaps the most notable omission is the field of photonic bandgaps and optical metamaterials. These use periodic arrays of nanostructures to engineer the optical properties of a material, for example, its refractive index, reflectivity, or color. These take their cue from naturally occurring photonic structures in bird feathers and butterfly wings (Fig. 72.2). Nanostructures such as pillars or holes are created in arrays such that their separation is on the order of half the wavelength of the light of interest. The following references give an introduction to the exciting properties of these nanomaterials, further discussion of which is sadly outside of the scope of this chapter [4–9].

Rather than be an exhaustive list of nanomaterials, this chapter is instead intended to give a general overview of the most important properties that nanomaterials offer us. To do this, we will focus solely on those physical properties that change drastically as an object is reduced in size from the macroscale to the nanoscale. We now know, for example, that the physical and chemical properties of common materials such as carbon, silicon, or gold can be tuned simply by changing their size [10]. Color, magnetic and electronic behaviors, and even melting point can all be changed by shrinking a material down to the nanoscale.

We can broadly divide these significant changes in behavior at the nanoscale into two major groups. *The first group* contains those nanomaterial properties that come from a change in the relative influence of certain classical physical forces. For example, because of their low mass, gravity ceases to play a major role in the behavior of nanoscaled objects. In contrast, an enormous increase in the surface-to-volume ratio of nanomaterials means that electrostatic forces become extremely important. *The second group* contains those properties peculiar to nanomaterials that arise due to a fundamental shift in the way that nanoscaled objects interact with particles such as electrons and photons. As we will see, it is at the nanoscale that a quantum mechanical description of objects becomes most accurate. Compared with classical mechanics, with which we are all well accustomed at the macroscale, quantum physics famously makes some strange predictions about the way tiny objects behave. It is this quantum framework for looking at the world at the nanoscale that helps explain some of the unexpected properties of nanomaterials.

72.2 THE RISE OF NANOMATERIALS

The recent explosion in nanomaterials research has been driven by some key inventions that allow us to create and look at nanoscaled objects. In the 1970s, the development of molecular beam epitaxy [11], and metal organic chemical vapor deposition [12, 13], made possible the growth of ultrahigh purity crystalline materials with atomic-level control over their thickness and other dimensions. Using increasingly short wavelengths of light, modern photolithographic methods can now create patterns on the surface of a material that have features only a few tens of nanometers in size, a limit that continues to fall [14–18]. Other synthesis approaches that have been critical at some stage in the development of nanomaterials include nanostructured masks; advanced wet, dry, and selective etching techniques; nanoimprinting; nanostructure self-assembly; and self-alignment. As we look at the properties of various nanomaterials in this chapter and discuss how they are made, it is important to distinguish between "top-down" and "bottom-up" approaches [19]. Top-down methods start with a featureless surface and remove material to create nanostructures. Bottom-up approaches start with a featureless surface and add material to create, or "grow," the nanostructures.

The majority of technologically important nanomaterials are crystalline in nature, for example, most semiconductors and metals. Crystals form a subgroup of solid materials that are characterized by a highly ordered, periodic, symmetric atomic structure, where the interatomic distances are fixed and well known for a given material. Although noncrystalline nanomaterials exist (especially in nature), in the interests of space this chapter will mainly focus on examples of crystalline nanomaterials and their properties.

However, the ability to synthesize extremely high-quality materials with increasingly small sizes is useless unless we are able to look at and measure what we have made. To this end, some critical advances in microscopy were made during the 1980s with the advent of the scanning tunneling microscope and atomic force microscope. For the first time, these techniques allowed researchers to directly image and manipulate nanostructures or even individual atoms. Other key characterization techniques used (or sometimes specifically developed) to measure and interrogate the properties of nanomaterials include X-ray diffraction, photoluminescence, electroluminescence, transmission electron microscopy, and low-temperature magnetoresistance techniques. Many of these essential characterization tools and approaches are covered elsewhere in this book. Due to these tools and a whole host of other discoveries in the burgeoning field of nanotechnology, nanostructures can now be controllably synthesized from an extensive range of materials for a wide variety of applications.

72.3 NANOMATERIAL PROPERTIES RESULTING FROM HIGH SURFACE-AREA-TO-VOLUME RATIO

72.3.1 The Importance of Surfaces in Nanomaterials

It is important to note that the chemistry deep within the interior of a solid material can be very different from its chemistry at the surface. In a typical crystalline material, atoms in the interior (or bulk) are stable and relatively inert. The distances to their nearest neighbors are well defined in every direction and they exist within a symmetric potential field. The situation at the surface of the material is very different. Think of creating a fresh surface by slicing through the bulk of the crystal to expose a single atomic layer. This newly exposed surface would initially have the same atomic arrangement as did the atoms in the bulk. However, creating this surface required the breaking of many atomic bonds, freeing up a large number of unpaired bonding electrons. These unpaired electrons, often referred to as "dangling bonds," mean that the surface is extremely energetically unstable. To reduce this surface energy, the atoms almost immediately rearrange themselves into structures that are more stable, but also more complex, often involving double bonds or dimer pairs [20].

Although these new atomic reconstructions do help stabilize the surface, they also create a surface that is fundamentally distinct from the bulk for several reasons. First, the new surface reconstruction changes its chemical and physical properties with respect to that of a region deep in the bulk. Second, even though the surface energy is minimized following reconstruction, it is typically not lowered to its bulk value. Thirdly, and most obviously, atoms at the surface have nothing attached to them in at least one direction. This means that external species (atoms, molecules, etc.) can approach the atoms at the surface and interact with them. This is usually unlikely for atoms deep within the bulk of a material. Overall, the upshot is that surfaces are typically significantly more reactive, both chemically and physically, than the interior bulk of a material.

Going back to our example of a macroscopic cube, the number of atoms at the surface is only a small fraction of the total number of atoms within the object as a whole. As a result, the physical properties within the bulk of the material have a strong influence on the properties of the cube as a whole. However, as we reduce the size of the cube, its surface-area-to-volume (SAV) ratio begins to increase. Reducing the length of each side of the cube by a factor of 10 *increases* its SAV ratio by a factor of 10. Therefore, the SAV ratio of a cube with sides 1 nm will be 10,000,000 higher than a cube with sides 1 cm (Fig. 72.3). In fact, it could be argued that for some nanomaterials, such as single-walled carbon nanotubes or the 2D materials that consist of a single

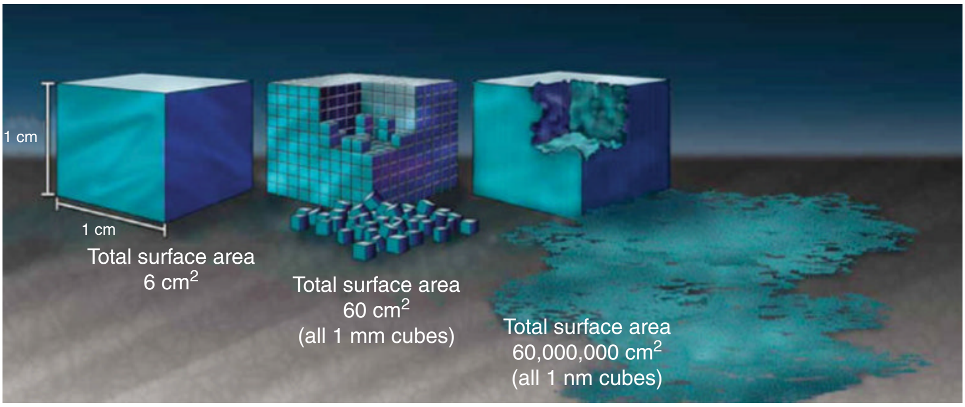


FIGURE 72.3 Illustration of the enormous increase in surface-area-to-volume (SAV) ratio exhibited by nanomaterials compared with macroscale objects. Source: http://www.nano.gov/nanotech-101/special. CC-BY-SA-3.0.

layer of atoms, the SAV ratio is infinite since they have no volume in the traditional sense of the word and are essentially all surface.

This enormous increase in SAV ratio is one of the root causes of the strikingly different properties of nanomaterials. Processes and interactions that are highly surface dependent will have a much stronger influence on behavior. Van der Waals forces, surface energy-driven phenomena such as capillarity [21], or surface-dependent interactions including catalysis [22], absorptive capacity, and chemical selectivity [23], can hence be intensified by many orders of magnitude at the nanoscale.

72.3.2 Electrostatic and Van der Waals Forces

The strength of electrostatic and van der Waals forces at the nanoscale means that they are used routinely for moving and placing nanostructures. It was these forces that allowed researchers in 1989 to carefully position 35 xenon atoms to spell out the IBM logo in one of the most iconic early demonstrations of nanotechnology [24]. More recently, in 2013, researchers used a similar approach to create the world's smallest stop-motion film "A Boy and His Atom" (http://www.research.ibm.com/articles/madewithatoms.shtml). Of course these accomplishments are entertaining (and good for advertising!), but they also illustrate just how precisely nanoscale objects with large SAV ratios can now be manipulated.

Another scientific breakthrough resulting from the electrostatic manipulation of nanomaterials was the discovery in 2004 of graphene. Graphene consists of a single sheet of carbon, one atom thick, with an sp^2 (honeycomb) bonding structure. Researchers used the van der Waals forces between scotch tape and pieces of graphite to peel off thin flakes of carbon. The flakes were then attached to silicon wafers, again using van der Waals attraction. Some of these flakes were found to be only one atom thick, and this was the first experimental demonstration of graphene. Graphene has since been shown to possess many remarkable properties from exceptional strength to very high electrical conductivity. Graphene also has a unique electronic structure that arises from the fact that it is only one atom thick. We will discuss the specific properties that result from this unusual electronic structure in Section 72.4.1.1.

72.3.3 Color

The increase in SAV ratio at the nanoscale can have surprising effects on certain fundamental physical properties. For example, we know gold as a shiny yellow metal, but even in medieval times, gold nanoparticles were being used unknowingly to create red and purple stained glass windows. This change in color arises from the way that light interacts with arrays of gold atoms. An oscillation is set up in free electrons at the metal surface, and this oscillation has a resonance called a surface plasmon. Compared with bulk gold, it is much easier for incoming light to excite this plasmon mode at the surface of a gold nanoparticle due to its much higher SAV ratio. The plasmon interaction results in strong absorption of photons with wavelengths close to 450 nm, which gives gold nanoparticles their characteristic red color [25].

72.3.4 Melting Point

The temperatures at which bulk materials melt are well known, typically with high precision. However, as materials are shrunk down to the nanoscale, researchers have shown that their melting points can change. For example, the usual melting points of the elements indium and gold are 157 and 1064°C, respectively. However, as these metals are reduced in size until they are clusters of atoms a few nanometers across, their melting points fall dramatically (Fig. 72.4) [26–28].

The reason for this change in such a fundamental physical property as melting point is once again the increase in SAV ratio at the nanoscale. During the melting of any material, the energy provided by heating enables the bonds between atoms in the solid to be broken to form a liquid. Atoms at the surface of a solid are less stable than atoms deep within the bulk since they are not bound to anything above. Less energy is therefore needed to pull an atom off the surface (i.e. to break the bonds with atoms below and beside it) than to remove an atom from within the bulk, where there are additional bonds above it to break. In a nanocluster, the high SAV ratio means that a far larger proportion of the total atoms are at the surface. As a result, less heat energy is needed to break all the bonds in the nanocluster and so the melting point is lower.

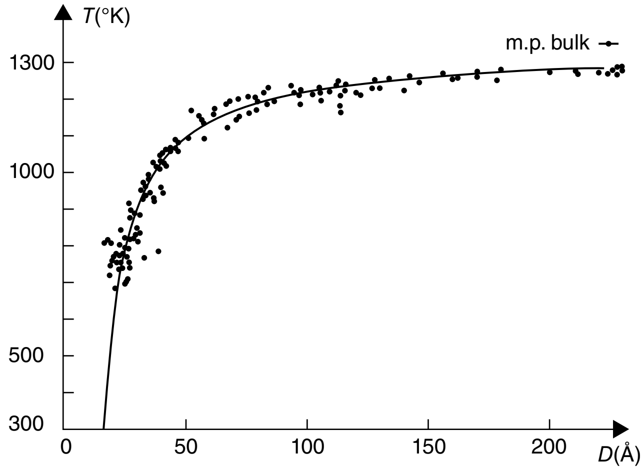


FIGURE 72.4 The change in the melting point of gold nanoclusters as a function of their diameter *D*. The unit of distance on the *x*-axis is the Angstrom, Å, (1Å = 0.1 nm). As the size of the gold nanoclusters increases, the melting temperature, *T* rises toward the bulk melting point of gold of 1337 K (1064°C). Source: Buffat and Borel [26]. Reproduced with permission of the American Physical Society.

72.3.5 Magnetism

One of the most surprising effects of shrinking materials to the nanoscale is perhaps that of turning certain nonmagnetic metals into magnets. Unlike the elements iron, cobalt, and nickel, noble metals such as gold, platinum, and palladium display no magnetic properties in the bulk. However, when nanoclusters are created from these metals, they can suddenly become magnetic [29, 30]. This effect is rather sensitive and appears to be strongly linked to nanocluster geometry [31] and the specific number of atoms involved (i.e., nanocluster radius) [30]. The origin of this effect seems to be highly surface dependent and hence another consequence of the large SAV ratio of nanomaterials [32]. Atoms at the surface of the nanocluster have bonding electrons with nothing outside of the nanoparticle to bond to. These unpaired surface electrons (which don't exist in the bulk metal), combined with the complex electronic structure of the nanoclusters, induce an overall magnetism in the nanoparticle [33]. As the size of nanoclusters is reduced, the SAV ratio rises, amplifying the effect of unpaired surface electrons, thereby increasing the magnetism (Fig. 72.5).

72.3.6 Hydrophobicity and Surface Energetics

High SAV ratios give researchers an opportunity to control a material's surface energy by creating nanostructured coatings. Controlling the surface energy influences the way that these coatings interact with external chemical species (e.g. water). Hydrophobic coatings have very low surface energies and so repel water via capillary

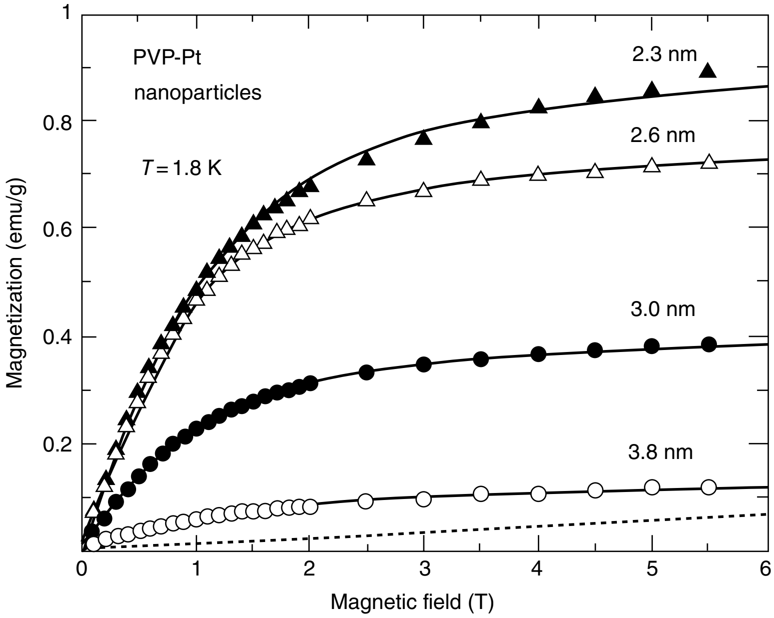


FIGURE 72.5 Dependence of the magnetism of platinum nanoclusters on their radius. As the radius of the nanoclusters is decreased from 3.8 to 2.3 nm, their magnetization increases. The dashed line shows bulk platinum metal and its relative lack of magnetic behavior. Source: Yamamoto *et al.* [30]. Reproduced with permission of Elsevier.

action to prevent it from wetting the surface. Hydrophobic surfaces abound in nature. A common example is the lotus plant (Fig. 72.1), the leaves of which, at the nanoscale, are covered in waxy bumps. Rather than wetting the surface, the waxy nanostructures cause water to ball up into droplets and roll off the leaves. The leaves are hence self-cleaning: the water droplets carry away any particles of dirt that could hinder photosynthesis.

The extent of hydrophobicity is quantified by measurement of the contact angle, θ_c, between a droplet of water and the surface (Fig. 72.6). Hydrophobic surfaces have θ_c less than 90° and the water is said to "wet" the surface; hydrophilic surfaces are characterized by θ_c greater than 90°. Superhydrophobic surfaces have values of $\theta_c > 150°$ and approach the limit of 180° (i.e., perfectly spherical water droplets) where the water is completely repelled by the surface.

Many applications for superhydrophobic nanomaterial coatings exist, from preventing icing of control surfaces on aircraft, to self-cleaning windows. The key to superhydrophobic coatings with low surface energies is that they consist of nanostructures with very sharp protrusions. By using chemical treatments to functionalize the sharp nanostructured surfaces, the water droplets are forced toward the tips of the points, while a layer of air trapped between the points actually prevents the water from reaching the surface (Fig. 72.7a). Incidentally, it is possible to reverse their function by

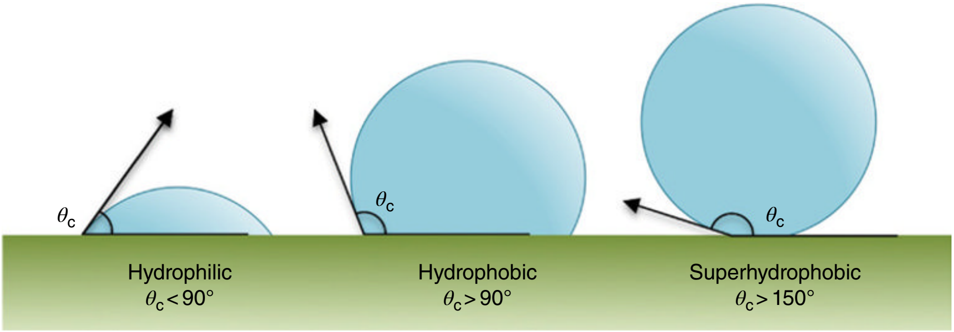


FIGURE 72.6 The degree of hydrophobicity is characterized by the contact angle, θ_c, formed between a water droplet and the surface on which it sits.

instead coating the spikes with a different chemistry that makes the surface superhydrophilic. When water hits a superhydrophilic surface, it forms an almost perfectly uniform, flat layer (see Section 72.3.8).

The nanoscale protrusions can be engineered using numerous approaches, including sharpening the tips of a bundle of silica optical fibers into nanocones (Fig. 72.7b) [35], selectively etching certain parts of a glass matrix (Fig. 72.7c) [36], or even using diatomaceous earth, which contains the skeletal remains of creatures that possess naturally occurring nanoscaled features (Fig. 72.7d) [34].

72.3.7 Nanofluidics

Another area dependent on the unusual capillary properties of nanomaterials is the field of nanofluidics. Nanofluidics is the study of how fluid flow changes at the nanoscale as a result of the increased SAV ratio of nanostructures and nanochannels. Capillary action is a manifestation of the interaction between two materials (often a liquid and a solid) at an interface. At the nanoscale, the number of molecules in the liquid becomes very low, while the effect of the solid surfaces on flow is greatly magnified due to the high SAV ratio. As a result, nanofluidic systems exhibit some unique properties [37–40].

The high SAV ratio in nanofluidic channels means that flow is governed predominantly by surface charge, and this can be used to distinguish between or separate different ions [39]. As a fluid moves across a surface, a double layer of charged particles builds up at the interface. The first layer consists of charged particles that are chemically attached to the surface in some way, and these may be positive or negative, depending on the situation. The second layer is made up of particles with the opposite charge that are coulombically attracted to (and electrically screen) the first layer. When the thickness of this double layer becomes comparable to the width of the nanochannel, only ions of one polarity (i.e., positive or negative) will be able to pass through [39].

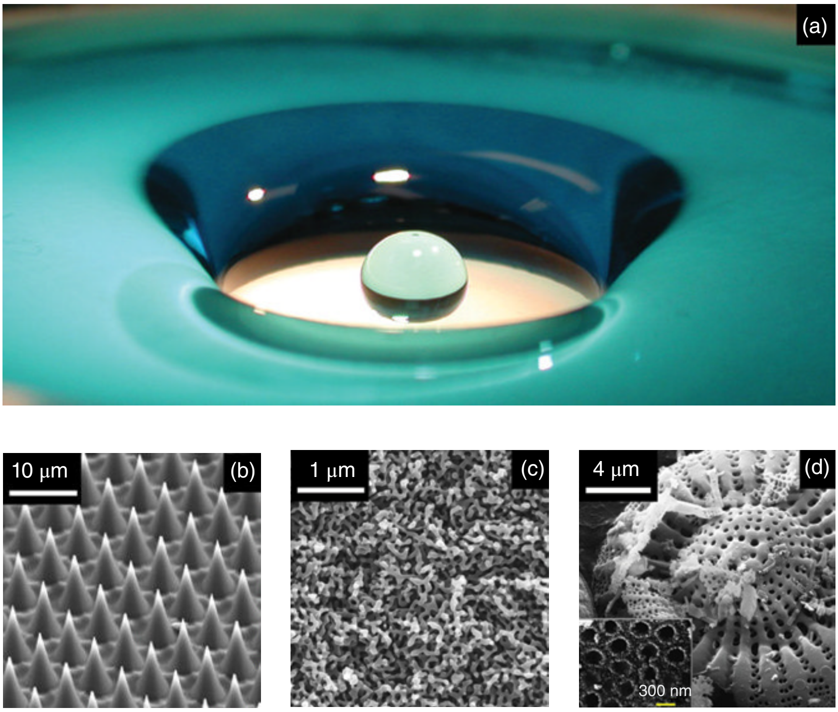


FIGURE 72.7 (a) The "Moses effect" where water is strongly repelled from a nanostructured superhydrophobic surface. A thin layer of air prevents the water from ever touching the structured surface. Source: Reproduced with permission of John T. Simpson. (b) An array of glass nanocones created by etching the ends of bundled and fused optical fibers. Source: Reproduced with permission of John T. Simpson. (c) A borosilicate glass film after spinodal decomposition and selective etching of sodium borate-rich regions, showing the porous nanoscale branched network of the silica-rich film. Source: Reproduced with permission of John T. Simpson. (d) Scanning electron microscopy image of the skeletal remains of a circular diatom (i.e., diatomaceous earth) characterized by micro- and nanoscale surface features. Source: Polizos *et al.* [34]. Reproduced with permission of Elsevier.

It is also the case that nanochannels may be comparable in size to molecules dissolved in the fluid. Larger molecules are thus physically blocked from passing through, while smaller molecules are transported freely [38]. In this way nanofluidic devices are able to act as electrostatic or physical filters, either separating different ions in an electrolytic fluid on the basis of their charge or distinguishing between molecules based on their size. This ability to filter nanoparticles and molecules from fluids is extremely appealing for biotechnology applications, particularly in the areas of DNA and protein analysis [40–42]. Nanofluidic channels can be made in a variety of ways, for example, using electron beam lithography or selective etching to make nanotubes [19, 41–43].

72.3.8 Nanoporosity

Closely aligned with both the nanostructured superhydrophobic surfaces and nanofluidic systems are nanoporous materials. Nanoscaled holes or pores can be used as molecular filters, antifogging coatings, desiccants, ion exchangers, and catalysts [44–47]. Let us take the example of antifogging coating as an example. Nanoporosity is used to engineer surface energies in a similar way to the superhydrophobic materials discussed in Section 72.3.6. The difference is that in this case the nanoporous material has a very high surface energy (either naturally or after chemical treatment) such that it is superhydrophilic (i.e., θ_c approaches 0°). Water is attracted to its surface, creating a continuous sheet of liquid rather than forming into discrete droplets that scatter light [46]. When applied to glass, these superhydrophilic nanoporous coatings hence prevent fogging and keep the glass clear (Fig. 72.8).

A very different application for nanoporous materials is in the technologically important area of terahertz (THz) optoelectronics. The efficient emission and detection of light at THz frequencies is critical to a wide range of fields from defense to biomedicine. Certain bulk semiconductor crystals emit THz radiation when excited with an ultrafast laser, but the strength of the THz signal is usually low [48]. Researchers have shown that increasing the surface area of the semiconductor crystal significantly enhances the THz emission intensity. Of course, with their high SAV ratio, nanostructured surfaces are ideally suited to this purpose. Groups began by demonstrating that in semiconductor nanowires, an enhancement of THz emission of up to 40 times could be achieved [49, 50]. They showed that the nanowires' high aspect ratio (length/diameter) geometry was critical for achieving the desired effect [49, 50]. Researchers then explored the use of nanoporous surfaces (Fig. 72.9a) [52, 53]. Compared with the bulk semiconductor crystal, THz emission enhancements of 100–1000 times have been reported for nanoporous GaP surfaces (Fig. 72.9b) [51].

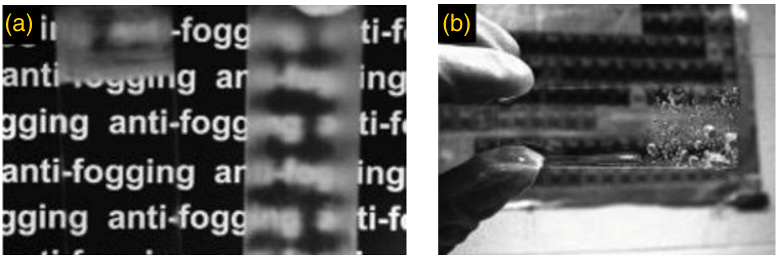


FIGURE 72.8 (a) Comparison of the fogging behavior of a bare glass slide (right-hand slide) and a slide with a superhydrophilic nanoporous coating (left-hand slide). For comparison, the fogged region at the top of the left-hand slide was not coated. (b) Glass slide illustrating the nonuniform water dewetting behavior of normal glass (right half) compared to the uniform wetting behavior of the surface with the superhydrophilic nanoporous coating (left half). Source: Cebeci *et al.* [46]. Reproduced with permission of the American Chemical Society.

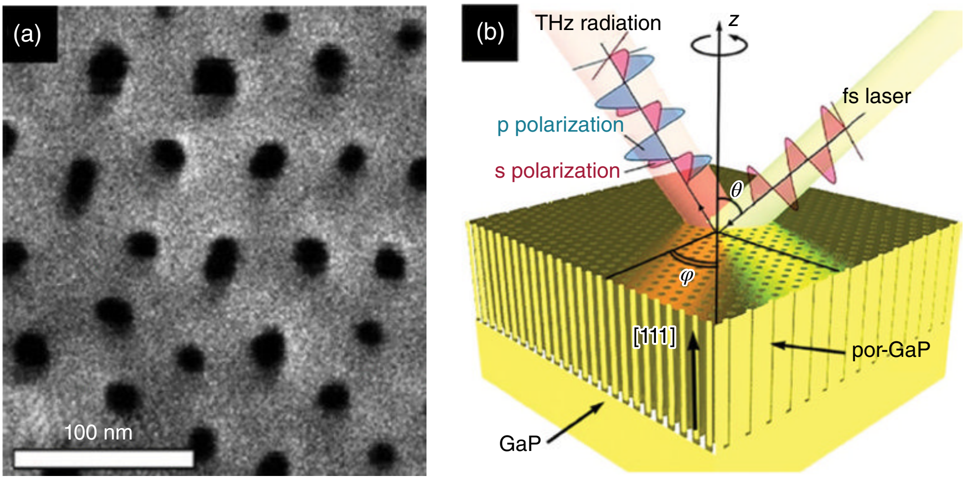


FIGURE 72.9 (a) The nanoporous GaP surface created by ion etching, where the nanopores have an aspect ratio of approximately 1500 [51]. (b) Schematic diagram of the experimental geometry for THz emission from a nanoporous material. The femtosecond (fs) laser coming in from the right excites the nanopatterned porous GaP (por-GaP) surface so that a THz beam is emitted on the left. Source: Atrashchenko *et al.* [51]. Reproduced with permission of AIP Publishing LLC.

A wide range of techniques have been adopted to create nanoporous materials, including the use of large molecules [44, 45], structured arrays [47], nanocasting [54], and dealloying and selective etching [55].

72.3.9 Nanomembranes

Researchers are able to create nanostructured membranes out of various electrically conductive materials. The conductivity of these nanomembranes change when different chemical species interact with their surface. The high SAV ratio of a nanomembrane provides a huge amount of surface area for these species to interact with. As a result, nanomembranes can be used as extremely sensitive chemical sensors [56].

Nanomembranes with thicknesses on the nanoscale can also be designed as micromechanical sensors [57]. A 25–70 nm thick polymer/gold composite nanomembrane is freely suspended above a comparatively large opening several hundred microns in diameter (Fig. 72.10) [58]. In response to stress, either from the tip of an atomic force microscope or pressurized gas, the nanomembrane bulges out and its total deflection is measured under different conditions. The nanomembranes have a high elastic modulus, such that deflections up 40 μm can be readily achieved without plastic deformation or rupture of the membrane [58]. After the load is removed, the nanomembranes return to their natural unstressed shape within a few seconds.

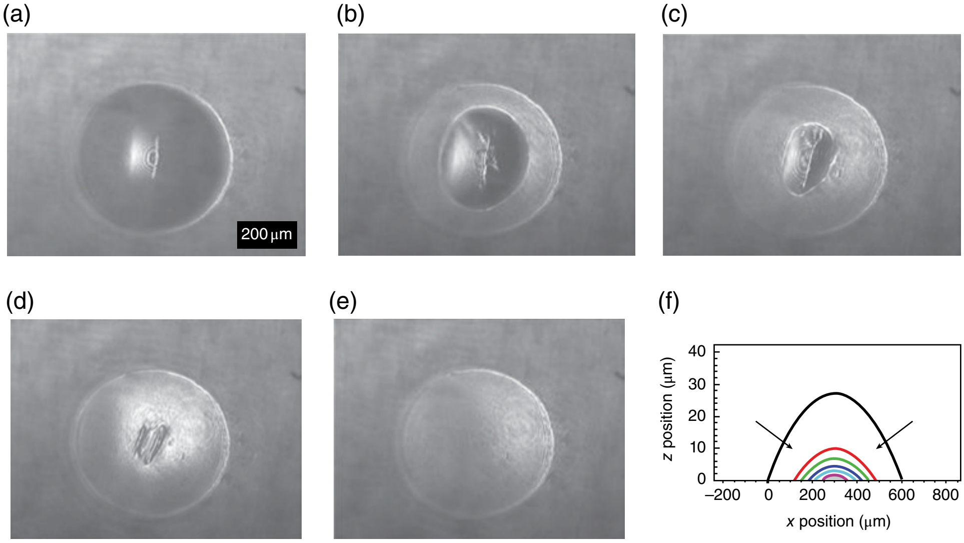


FIGURE 72.10 Auto recovery of a freely suspended nanomembrane subjected to high pressure and long loading time. (a) Fully distorted nanomembrane under high pressure (4 kPa). (b)–(e) Evolution of nanomembrane shape as a function of time after sudden pressure release—images taken at 2 s intervals. The overstretched central portion disappears and the nanomembrane becomes flat. (f) Cross sections through the nanomembrane during recovery. Source: Jiang *et al.* [58]. Reproduced with permission of Nature Publishing Group.

These nanomembranes have an enormous dynamic range: the highest pressure they can detect is 10^8 larger than the smallest measurable pressure [58]. As a result, nanomembranes like this are significantly more sensitive than comparable silicon membranes of the same diameter. Resonant frequencies of around 100 kHz causing vibrational amplitudes of 25 nm are measured for the nanomembranes. These nanomembranes could therefore be used as acoustic or even thermal sensors.

72.3.10 Nanocatalysis

The use of catalysts to start or speed up chemical reactions and to increase their efficiency or yield is ubiquitous. The chemical and oil industries are completely reliant on them, while consumers use them daily, for example, to reduce emissions from their vehicles. Many catalysts work by accelerating the reaction rate at the interface between a solid surface and a fluid. Reactant species adsorb to the surface of the catalyst, a process that lowers the energy barrier for the reaction to occur. The catalyst, often consisting of a transition metal element, does not take part in the reaction itself. The catalyst is therefore not consumed and so only a small amount is typically needed. Since adsorption of reactant species is entirely reliant on the availability of active sites on the surface of the catalyst, increasing a catalyst's surface area greatly increases its efficacy. Due to their high SAV ratios, catalytic nanomaterials can be extremely effective [22, 59, 60]. What is more, the activity of nanocatalysts can be further enhanced by careful engineering of their structural properties, for example, their shape, alloy composition, or nanostructure [22, 59, 60].

Increasingly, nanomaterials with an electrocatalytic behavior are sought for their ability to increase the reaction rate between hydrogen and oxygen, with the aim being to create lightweight but highly efficient fuel cells for use in vehicles [61–63]. Nanocatalysts are therefore at the forefront of the green energy revolution and an area of intensive research.

72.3.11 Further Increasing the SAV Ratio

Despite the fact that nanomaterials inherently possess very high SAV ratios, for some applications it is desirable to increase the SAV ratio even further. As an example, researchers wished to increase the surface area of a silicon nanowire (Fig. 72.11) so that it could be covered with a greater number of light-sensitive molecules for a detector. Rather than use complex photolithographic approaches to reduce feature size, they exposed the silicon nanowire to a metal-induced chemical etch to create silicon “nanoforests” (Fig. 72.11) [64]. Taking into account each silicon “tree,” it is clear that compared with the original planar nanowire, these nanoforests have an enormous surface area and hence a greatly increased SAV ratio. Similar approaches that try to increase SAV ratio in order to magnify various nanomaterial properties are widely used.

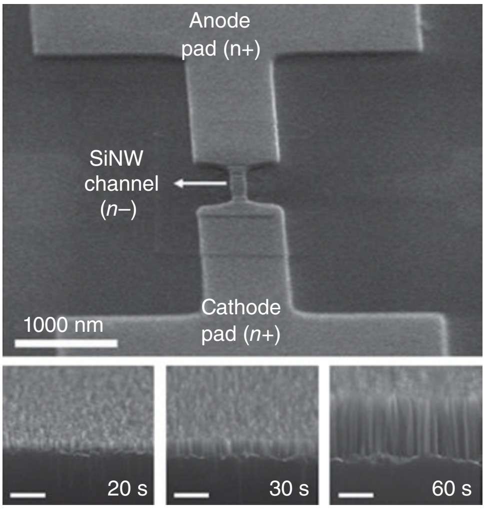


FIGURE 72.11 Top image shows a silicon nanowire (SiNW) connecting a cathode to an anode. The bottom three images show how the surface of the SiNW evolves into a Si nanoforest as the etch time is increased (scale bars = 200 nm). Source: Seol *et al.* [64]. Reproduced with permission of the American Chemical Society.

72.3.12 Nanopillars

The high aspect ratio nanoforests in Figure 72.11 are created using a top-down approach. Similar structures, known interchangeably as nanopillars or nanowires, can also be created using a bottom-up approach. Nanopillars can be made to grow vertically out of a flat substrate surface using various epitaxial (i.e., crystal growth) techniques. Depending on the desired application and the growth method used, these nanopillars are either randomly located across the substrate surface [65] or arranged in ordered arrays (Fig. 72.12) [66].

The versatility of the bottom-up approach means that the nanopillars can be created in arrays with predefined geometries [66, 67], even on traditionally incompatible substrate materials [65, 68], or combining multiple materials in the same pillar in so-called core–shell or axial geometries to build electronic or photonic device structures [69–71]. Randomly arranged pillars are typically created using metallic nanodroplets to catalyze the vertical growth [72, 73]. In contrast, nanopillar arrays are formed by first depositing a mask on the substrate surface. E-beam or nanoimprint lithography is used to open nanoholes in the mask to expose the substrate only in certain areas. During growth, the nanopillars form only in the holes and so by designing the arrangement of the nanoholes, nanopillar arrays with almost any geometry can be readily engineered [67, 74]. This can be put to use to create photonic crystals for light trapping or to act as a laser cavity [69, 75, 76].

Nanopillars typically have radii of 25–100 nm but can be several microns in length. This enormous aspect ratio exaggerates the already large SAV ratios possible in nanostructured materials and can be exploited in a wide range of applications including light-emitting diodes [71, 74], lasers [76, 77], photodetectors [78], solar cells [69, 79], transistors [80–82], and electrically gated modulator waveguides [83]. Taking solar cells as an example, the large SAV ratio of the pillars is particularly attractive for several reasons: the large surface area increases photon absorption probability; the light-trapping properties of the nanopillar array further improves photon capture; and

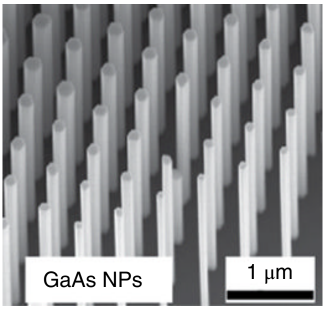


FIGURE 72.12 An ordered array of GaAs nanopillars grown via metal organic chemical vapor deposition on a nanopatterned GaAs substrate. Source: Lin *et al.* [66]. Reproduced with permission of IOP Publishing.

the ease of radial charge collection due to the high aspect ratio of the nanopillars raises the efficiency of the solar cell [69].

72.3.13 Nanomaterial Functionalization

Preceding sections have touched on the fact that, the large SAV ratio of nanostructured materials can be exploited through "functionalization." Chemical or biological species that serve some active, useful purpose are attached to the surface of nanomaterials. Because of the large surface areas, enormous numbers of these functional molecules can be attached, making their combined action highly effective. Functionalization hence allows the chemical or physical properties of a nanomaterial to be engineered for specific applications [84].

Functionalization has been demonstrated for various nanomaterial families, including graphene [85–87], carbon nanotubes [88–90], nanowires [91–93], and semiconductor nanoparticles known as colloidal quantum dots [94–96]. Applications for functionalized nanomaterials are enormously varied but include electrical devices [85], improving the structural properties of composite materials [86], and chemical sensing [91, 92]. However, it is the field of biotechnology that is finding some of the most exciting uses for functionalized nanomaterials. Biofunctionalized nanomaterials are used as light-emitting fluorophores for *in vivo* imaging, labeling, diagnostics, and sensing [94–96]; measuring DNA hybridization in real time [93]; and the delivery of drugs to specific areas of an organism for targeted disease treatment [87, 95].

Taking colloidal quantum dots as an example, these semiconductor nanoparticles are extremely promising candidates for cancer treatment. They exist in solution and are small enough that they can pass into an organism and interact with it at the cellular level. Quantum dots are first coated to prevent them from degrading once they enter the organism [95]. Molecular species that target a specific kind of cancer are then attached to their exterior surfaces. Finally, drugs that attack that cancer type are added. The size and selectivity of these functionalized quantum dots means that they are able to deliver the drugs directly to where they are needed. Importantly, the light-emitting properties of the quantum dots of their movements can be simultaneously tracked throughout the organism to verify that they end up in the correct location. In this way the quantum dots are also monitored to see whether they are eventually excreted by the organism or ultimately build up in certain locations (Fig. 72.13) [95].

This approach is extremely versatile. Various quantum dots or nanoparticles (different sizes, light emission wavelengths, etc.) can be combined with a wide range of surface functionalizations (different drugs, target organs, etc.). The ability to control multiple behaviors in a single dose of nanoparticles enables multiple treatments or diagnostic experiments to be run simultaneously [95, 97–100].

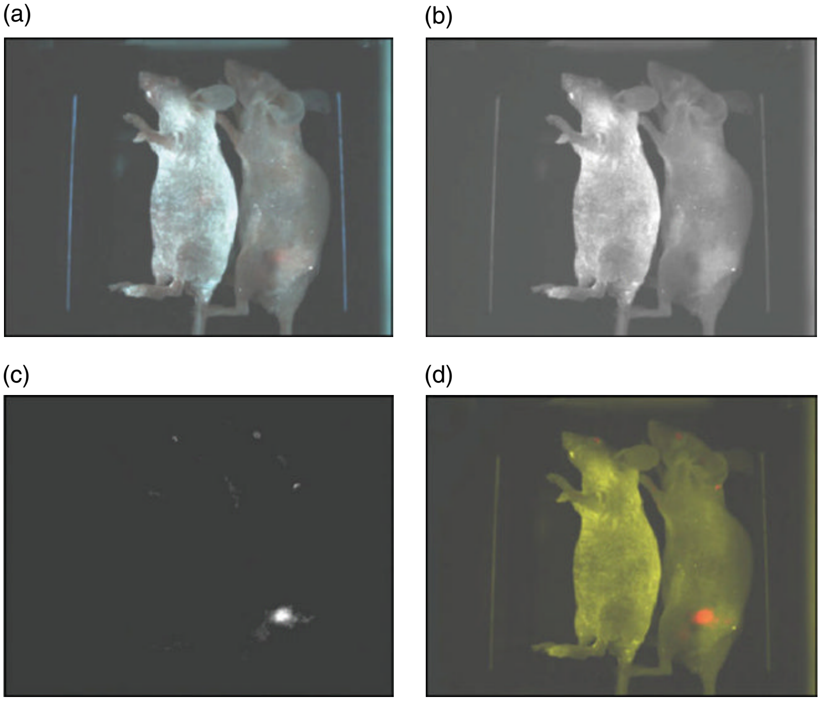


FIGURE 72.13 Spectral imaging of live animals injected with quantum dots that were functionalized to specifically target prostate cancer cells. Each image shows a healthy mouse (left) and a mouse with a prostate tumor (right). Fluorescence signals from the quantum dots, visible as brighter areas in the images of the sick mouse, indicate that they localize in the prostate tumor. The same dose of quantum dots injected into the healthy mouse showed no localized fluorescence signals. (a) Original image, (b) unmixed autofluorescence image, (c) unmixed quantum dot image, and (d) superimposed image. After *in vivo* imaging, the sick mouse was dissected to confirm that the quantum dot signals indeed came from an underlying tumor. Source: Gao *et al.* [95]. Reproduced with permission of Nature Publishing Group.

72.3.14 Other Applications for High SAV Ratio Nanomaterials

In addition to the applications already discussed, the high SAV ratio of nanomaterials (and the properties that result from this) makes them ideally suited for use in a wide range of technologically important areas. These applications include nanoengineered batteries [101–106], solar cells [107, 108], fuel cells [60–63], next-generation 3D computer chip architectures [109, 110], graphene membranes for electron microscopy [111, 112], and interaction with biological systems [113].

72.4 NANOMATERIAL PROPERTIES RESULTING FROM QUANTUM CONFINEMENT

So, an inherently large SAV ratio gives rise to nanomaterials with a dizzying array of properties. However, the second family of nanomaterials that we will look at acquires their unique characteristics simply because of their size. These nanomaterials are designed so that their interactions with electrons are fundamentally different than at the macroscale, and this gives them their extraordinary properties. Within semiconductor materials, the de Broglie wavelength of an electron (which is inversely proportional to the electron's momentum) is typically on the order of a few tens of nanometers. For example, in GaAs, the de Broglie electron wavelength at 300 K is 24 nm [114]. When nanomaterials are created with feature sizes smaller than the de Broglie wavelength, electrons within the material suddenly begin to behave very differently.

A different physical model is hence needed to describe and predict the electrons' strange behavior: quantum theory. Compared with classical mechanics with which we are all well accustomed at the macroscale, quantum physics famously makes some bizarre predictions about the way tiny objects behave. It is quantum physics that is responsible for giving this second family of nanomaterials their dramatically different properties. This quantum framework allows us to understand what is happening at the nanoscale and permits us to design and create semiconductor nanomaterials with novel properties that cannot be obtained by other means.

Probably the first realization of a structure specifically designed and engineered to have dimensionality on the nanoscale was the development of the quantum well in the 1970s by Bell Labs and other groups [115, 116]. A quantum well consists of a thin slice of semiconductor material with a low energy bandgap, sandwiched between two barriers made of another semiconductor with a larger bandgap (Fig. 72.14). So long as the thickness of the low bandgap material is on the order of the de Broglie electron wavelength, the energy of an electron confined to this region will become quantized. It should be noted that quantum mechanics is at play at all length scales; however, it is only below the de Broglie wavelength that the splitting between the quantized states becomes large enough that the energy levels no longer appear to be continuous.

What this means is that although the electron can have any arbitrary energy in the plane of the quantum well, in the nanoscaled out-of-plane direction the electron must occupy one of several discrete, fixed energy levels. The position of the energy levels is specific to a given quantum well nanostructure and is affected by the nanoscale width of the well and the bandgaps of the constituent semiconductor materials. The electron can move between these levels by absorbing or emitting packets (known as quanta) of energy. These quanta of energy are often absorbed or emitted in the form of a photon, that is, a particle of light with exactly the same energy. Since a photon's color is a function of its energy, this means that by simply adjusting the width of the quantum well nanostructure, we can control the color of light that it either absorbs or emits. This property of these nanomaterials led to one of the first applications of the quantum well: wavelength tunable semiconductor lasers [117].

A quantum well has one nanoscaled dimension, that is, its width. Following the development of quantum wells, semiconductor nanostructures that offer quantum confinement in two, and three dimensions have also since been demonstrated. These are, respectively, the quantum wire, which has two nanoscale dimensions (i.e., width and

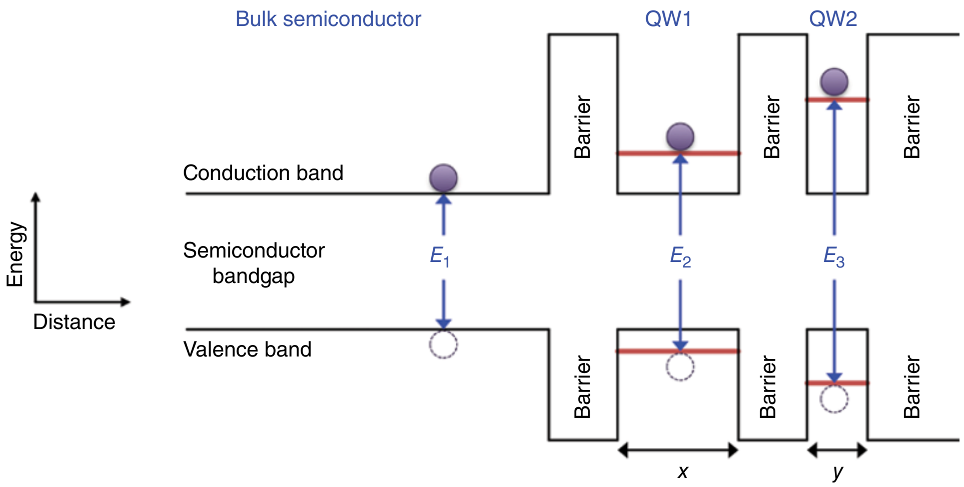


FIGURE 72.14 Schematic diagram showing the effect of quantization on electron energy levels. In a bulk semiconductor the electron transition is between the top of the valence band and the bottom of the conduction band. A photon emitted as a result of this transition will have energy E_1. In quantum well, QW1, of width x (where x is on the order of the de Broglie electron wavelength), the electron energy is now quantized perpendicular to the plane of the well. The quantized energy levels exist below (above) the valence (conduction) band edges. A photon emitted due to an electron transition in QW1 will thus have energy E_2, where $E_2 > E_1$. In quantum well, QW2, of width y (where $y < x$), the quantized electron energy levels are "squeezed" further below (above) the valence (conduction) band edges. A photon emitted due to an electron transition in QW2 will have energy E_3, where $E_3 > E_2 > E_1$.

height), and the quantum dot, which is nanoscaled in all three dimensions (i.e., width, height, and length) (Fig. 72.15).

In a quantum wire, the electron energy is continuous along the length of the wire and quantized in the two nanoscaled directions. Quantum dots are nanoscaled in all three spatial dimensions so that the electron's energy becomes fully quantized. In this respect, quantum dots are sometimes referred to as artificial atoms. Just as in an atom, the electrons in a quantum dot exist in well-defined energy levels, with rules about how multiple electrons must arrange themselves. The key difference between an atom and a quantum dot is that nature has fixed the atomic electron orbital energies, whereas we can readily tune the electron energy levels of a quantum dot simply by adjusting its size (Fig. 72.14).

As the number of nanoscaled dimensions increases from zero (bulk material), to one (quantum well), to two (quantum wire), to three (quantum dot), there is a big change in the density of states (DoS) (Fig. 72.16). The DoS is defined as the number of available "locations" that can be occupied by electrons at a given energy.

Together with the tunability of the electron energy levels, the capacity to change the DoS simply by adjusting the number of nanoscaled dimensions is perhaps the most

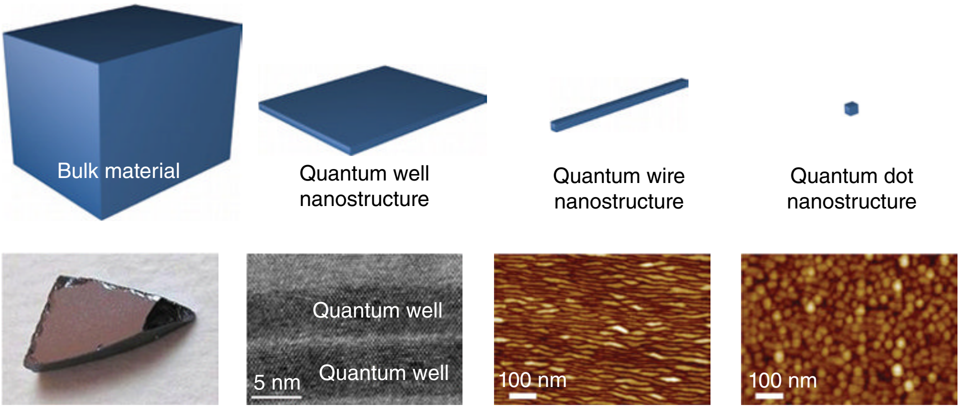

Bulk material
Quantum well nanostructure
Quantum wire nanostructure
Quantum dot nanostructure
Quantum well
Quantum well
5 nm
100 nm
100 nm

FIGURE 72.15 As we sequentially reduce each of the sides of a macroscopic bulk semiconductor crystal to the nanoscale, we create a specific family of quantum nanomaterials. (Top row): A quantum well offers quantum confinement in one dimension, a quantum wire in two, and a quantum dot has confinement in all three spatial dimensions. (Bottom row): Images of each of these quantum systems realized experimentally. Source: (bottom left) https://upload.wikimedia.org/wikipedia/commons/6/6f/Gallium_arsenide_crystal.jpg. CC-BY-SA 3.0; (bottom middle left) Paul [118]. Reproduced with permission of John Wiley & Sons, Inc. and (bottom middle right and bottom right) Li *et al.* [119]. Reproduced with permission of AIP Publishing LLC.

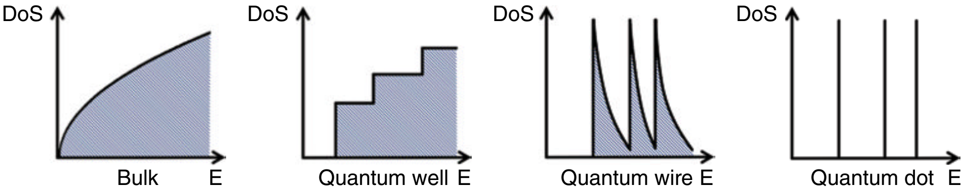


FIGURE 72.16 The change in density of states (DoS) in semiconductor nanomaterials as a function of the number of quantized dimensions. After [120].

attractive feature of semiconductor quantum nanostructures. This exquisite control over both the electron energy and DoS opens up new opportunities for designing materials for specific applications and for the discovery of new physics.

72.4.1 Quantum Well Nanostructures

Take the example of the semiconductor laser mentioned previously. Compared to a laser built from bulk semiconductor material, quantum well lasers have many superior characteristics [121, 122]. In a bulk semiconductor laser, the material chosen for the gain region dictates the emission wavelength. However, in a quantum well, quantization pushes the electron energy levels higher than the bulk semiconductor bandgap (Fig. 72.14).

This means that a photon emitted from a quantum well will have higher energy than a photon emitted from the same bulk semiconductor (in Fig. 72.14, $E_2 > E_1$). A photon with higher energy has a shorter wavelength. This property means that we can take a bulk semiconductor that emits light at a long wavelength and, by creating a quantum well out of it, make it emit light at a shorter, potentially more useful wavelength. For example, a bulk semiconductor that emits in the infrared region of the electromagnetic spectrum can be made into a quantum well laser that operates at visible wavelengths.

The wavelength of a bulk laser is fixed by the choice of semiconductor and can only be altered by changing the composition of the material. In contrast, in a quantum well laser, the emission wavelength is also a function of the width of the quantum well (Fig. 72.14). Higher energy (shorter wavelength) photons are emitted from thinner quantum wells, while wider quantum wells emit lower energy (longer wavelength) photons. Therefore simply by controlling the quantum well width, the lasing wavelength can be readily tuned ($E_3 > E_2$ in Fig. 72.14).

The ability to modify the electron DoS is the key to the extremely low threshold currents possible in quantum well lasers [123, 124]. The threshold current is the current above which stimulated emission exceeds spontaneous emission (i.e., the device begins to behave as a laser). Reducing the threshold current is highly desirable as it means that the laser uses much less power during its operation. In bulk material, the electron DoS increases with the square root of energy. In contrast, for a given quantized energy level in a quantum well, the electron DoS is independent of energy (i.e., constant). Compared with bulk material, a quantum well can therefore contain significantly more electrons with the same energy, which reduces the threshold current required to turn on a quantum well laser [125].

In addition, the threshold current in quantum well lasers is relatively insensitive to changes in temperature because of their step-like DoS (see Fig. 72.16) [122, 126]. Despite various cooling techniques, the temperature in a typical electronics rack can quickly rise, even under normal operating conditions. It is important therefore that a laser can cope with increasing temperatures and provide stable performance. The threshold current density, J_{th}, is related to the temperature of the laser, T, by the expression

$$J_{\mathrm{th}}(T) = J_{\mathrm{th}}(0)\exp(T / T_0)$$

The larger the value of the constant T_0, the less J_{th} varies as T is raised or lowered. Even early in their development, $T_0 = 437°\mathrm{C}$ was reported for quantum well lasers [126], which compares extremely favorably with $T_0 = 185°\mathrm{C}$ for bulk lasers developed around the same time [127].

A more recent development in lasers based on quantum well nanostructures is the quantum cascade laser [128–131]. In a single quantum cascade laser, hundreds of quantum wells are placed side by side. The quantum wells are carefully designed so that their quantized energy levels create a "staircase" for electrons to flow down (Fig. 72.17) and, in doing so, emit very long wavelength photons.

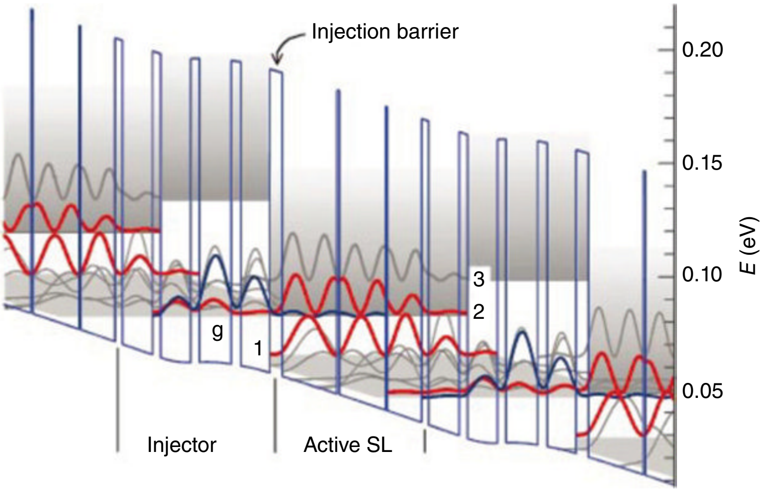


FIGURE 72.17 Calculated conduction band structure of a small part of a quantum cascade laser showing the multiple quantum wells with different widths. To achieve large gains, quantum cascade lasers will typically contain hundreds of repeats of these injector/active region sections. In this device, photons with energy 18 meV (4.35 THz) are emitted by electron transitions from state 2 to state 1 in the Active SL section. After traversing the injector ground state (g), the electron finds itself in the excited state 2 of the Active SL section and the photon is emitted. The electron is then injected into the next Active SL section and the process is repeated. Source: Köhler *et al.* [130]. Reproduced with permission of Nature Publishing Group.

Clearly, realizing a quantum cascade laser depends on the ability to create quantum wells with precise widths. This requires a growth technique that is stable over the many hours needed to grow these complex structures, most commonly, molecular beam epitaxy. These quantum cascade structures are currently the most effective ways of obtaining THz laser emission, the importance of which we discussed in Section 72.3.8.

In addition to enabling new devices, the ability to create quantum wells from ultrahigh electron mobility materials allowed researchers to directly test certain predictions made by quantum theory and led to the discovery of new physics. Perhaps most notable of these new findings was the discovery in 1980 of the integer quantum Hall effect, for which the Nobel prize in Physics was awarded in 1985 [132]. Discovery of the fractional quantum Hall effect followed in 1982 and led to the 1998 Nobel prize in Physics [133].

When a current is passed through a bulk conductive material in the presence of an orthogonally aligned magnetic field, a voltage is developed perpendicular to both. This is the classical Hall effect. The magnetic field curves the path of the electrons as they travel. As a result, a voltage is created by the accumulation of a net negative charge on one side of the conductor and positive charge on the other. The Hall coefficient, R_H, has an inverse linear dependence on the magnetic field strength.

To observe the quantum Hall effect, a quantum well is created in a high electron mobility semiconductor and cooled down, typically to 1.5 K or below, in the presence

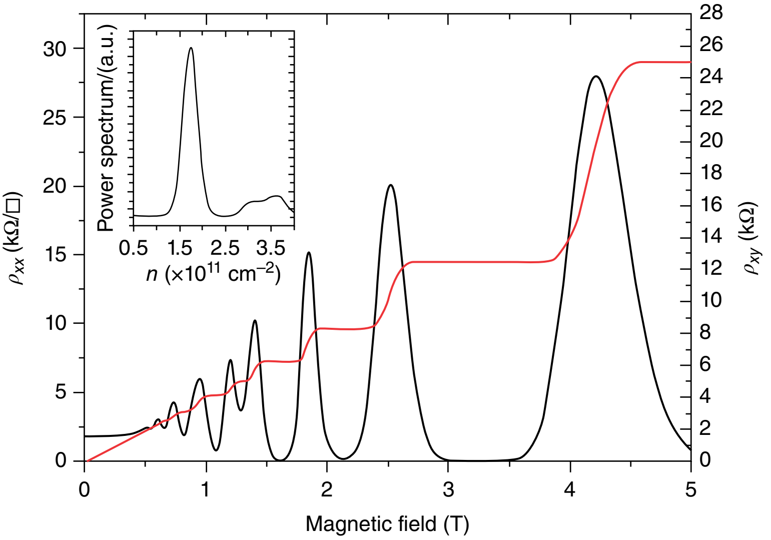


FIGURE 72.18 Measurement of longitudinal (ρ_{xx}) and Hall (ρ_{xy}) resistivities at 1.5 K as a function of magnetic field for an $In_{0.75}Ga_{0.25}As$ quantum well. The inset shows the fast Fourier transform of the longitudinal data, which gives an electron sheet density of 1.6×10^{11} cm^{-2}. The electron mobility in this sample was approximately 200,000 cm^2/V s. Source: Simmonds *et al.* [134]. Reproduced with permission of AIP Publishing LLC.

of a suitably large magnetic field. In contrast with the bulk case, when the magnetic field strength is swept, plateaus are observed in the linear Hall resistance (ρ_{xy} curve passing through the origin in Fig. 72.18). This is the signature of the quantum Hall effect [135, 136]. The longitudinal resistance in the direction of the current flow fluctuates between zero and finite values in a phenomenon known as Shubnikov–de Haas oscillations (curve oscillating between zero and non-zero values of ρ_{xx} in Fig. 72.18).

The magnetic field causes electrons in the plane of a quantum well to move in cyclotron orbits due to the Lorentz force. The orbits in which the confined electrons move have quantized radii, each of which is known as a Landau level [137]. As the strength of the magnetic field is increased, the degeneracy of these Landau levels (i.e., the number of electrons they can hold) rises. Electrons in higher Landau levels are therefore able to “fall” into lower Landau levels until the higher level is empty. At this point the Fermi energy (the highest energy level occupied by an electron) “jumps” down to the lower Landau level. This process is repeated as the magnetic field continues to rise, until all the electrons are in the lowest Landau level.

When the Fermi level is within a Landau level, the Hall resistance increases. However, when the Fermi level is between two Landau levels, the Hall resistance remains constant, and we see the characteristic plateaus of the quantum Hall effect (Fig. 72.18). The plateaus take values

$$\rho_{xy} = h/\nu e^2 \quad (\nu = 1,2,3,\ldots)$$

ν is referred to as the filling factor and is equivalent to the number of filled Landau levels in a spin–split system (where the magnetic field is strong enough to give electrons with opposite spins different energies). Note that at magnetic fields that aren't strong enough to lift the electron spin degeneracy, only the plateaus corresponding to even values of ν are seen. h/e^2 is the resistance quantum. As a result of quantum Hall effect measurements, its value (25.813… kΩ) is known with extreme precision and is now used as the international standard for resistance. The quantum Hall effect is used as a method for measuring the electron density in a quantum well and the electron mobility of the semiconductor material [134].

In extremely high mobility materials at very low temperatures in large magnetic fields, additional quantum Hall plateaus have been found at noninteger values of ν. This is the fractional quantum Hall effect [133]. The origins of the fractional quantum Hall effect are not yet fully understood, but experimental support is rising for a theory that invokes a new state of matter based on fractionally charged quasiparticles [138–141].

72.4.1.1 2D Materials Since the discovery of graphene in 2004 [142], research into atomically thin materials has exploded. Their 2D nature means that these materials are in some ways similar to the quantum wells discussed previously. However, the difference is that it is not simply the electrons that form a 2D sheet due to quantum confinement; here the atoms making up the material itself are arranged in a 2D sheet. This has a huge impact on the behavior of 2D materials, which are often drastically different from samples of the same materials consisting of multiple atomic layers. For example, graphite consists of many layers of graphene stacked on top of each other, but graphene has several unique properties that are not shared by graphite. Perhaps most important is graphene's highly unusual electronic structure. Electrons in graphene behave as if they have no mass and hence move through the 2D material at 1/300 the speed of light [143]. It is this property that explains graphene's excellent electrical and thermal conductivities [144]. Its hexagonal structure is incredibly strong, making it an ideal nanoscaffolding material. When rolled up, these sheets are the basis for carbon nanotubes (see Section 72.4.2.2). The quantum Hall effect was measured in graphene at low temperature very shortly after its discovery [143]. Successful measurement of the quantum Hall effect was then reported at room temperature [145]. This result is especially striking when one considers that previously the highest temperature at which the quantum Hall effect had been seen in any material was less than 50 K [146]. Given its highly attractive properties, a diverse range of applications are proposed for graphene from transparent, flexible electronics to hydrogen storage in fuel cells.

More recently, additional 2D materials have been developed, including hexagonal boron nitride, transition metal dichalcogenides, silicene, and germanene [147–152]. By stacking layers of different 2D materials (Fig. 72.19), it is hoped that their various properties can be combined in hybrid structures [152]. The properties of some of these

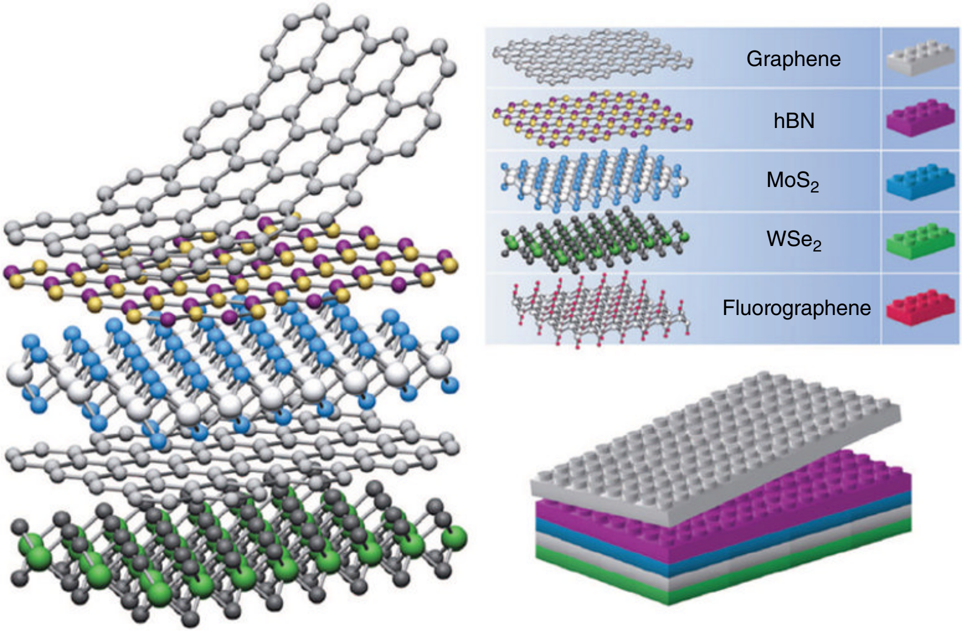


FIGURE 72.19 Hybrid structures composed from individual layers of 2D materials have been proposed. Loosely interconnected by van der Waals forces, the 2D layers will stack on top of one another like a child's building blocks. Source: Geim and Grigorieva [152]. Reproduced with permission of Nature Publishing Group.

materials are yet to be fully elucidated, but it is expected that they will complement, or in certain areas surpass, graphene. For example, silicene and germanene are expected to possess a small semiconductor bandgap (unlike graphene) [153]. This bandgap could enable these materials to be used as the channel material in a transistor, enhancing computer chip performance while reducing size.

72.4.2 Quantum Wire Nanostructures

Continuing from left to right in Figure 72.15, we come to quantum wire nanostructures. Quantum wires are 1D nanostructures that offer quantum confinement in two dimensions. From Figure 72.16 we see that there is another big change in the DoS when moving from quantum wells to quantum wires. To create a quantum wire, you can imagine taking a quantum well and "squeezing" down one of its long edges. If the length of this edge is reduced sufficiently, then the electron energy will also be quantized in this direction.

72.4.2.1 Gate-Defined Quantum Wires In fact, literally squeezing a quantum well was precisely the approach used to create the first quantum wires [154–156]. To do this, tiny metal pads (known as gates) are placed next to a quantum well nanostructure

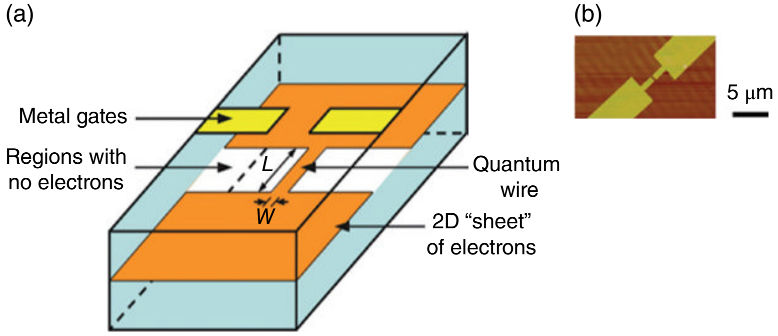


FIGURE 72.20 (a) Using electrostatic gates to create a quantum wire in a quantum well. A standard quantum well is created containing a 2D sheet of electrons. Metal gates are placed close to the quantum well. Applied gate voltage depletes electrons from underlying regions of the electron sheet. A quantum wire is formed by increasing the gate voltage to reduce the effective channel width (W) below the electron de Broglie wavelength. (b) A microscope image of a pair of gold gates on a sample of InGaAs containing a quantum well. Source: Simmonds *et al.* [157]. Reproduced with permission of the American Vacuum Society.

(Fig. 72.20a). The gates are electrically isolated from the quantum well by a thin insulating layer. Applying a negative voltage to the gates creates an electric field around each of them, and this field "depletes" the electrons from the immediately adjacent region of the quantum well. Depending on the shape and position of the metal gates, patterns can be created in the quantum well so that there are insulating areas where all the electrons have been removed and conductive areas where the electrons continue to exist as before.

Two parallel gates can be used to create a thin channel in the quantum well (Fig. 72.20a). Figure 72.20b shows a microscope image of a pair of these gates that were fabricated above a quantum well using electron beam lithography. In this example, the thin gap between the two gold finger-shaped gates is 500 nm. It is in this gap that the quantum wire is formed.

Increasing the negative gate bias raises both the strength and lateral spread of the electric field around the gates. By increasing the field we increase the size of the depleted regions of the quantum well where there are no electrons (Fig. 72.20a), which has the effect of reducing the channel width, *W*. Eventually, *W* will become small enough that it is on the order of the electron de Broglie wavelength. At this point the electron energy levels in the channel are quantized in two directions: firstly in the direction normal to the plane of the original quantum well and secondly in the direction between the two gates. Electron energy is however still continuous along the length of the channel making this a quantum wire nanostructure.

Quantum wires created in this way are typically a few hundred nanometers long. If the quantum well/wire is made from high-quality semiconductor material with a very low background of impurities, then the electron mean free path (the average distance electrons travel between collisions) can be several microns. In this situation,

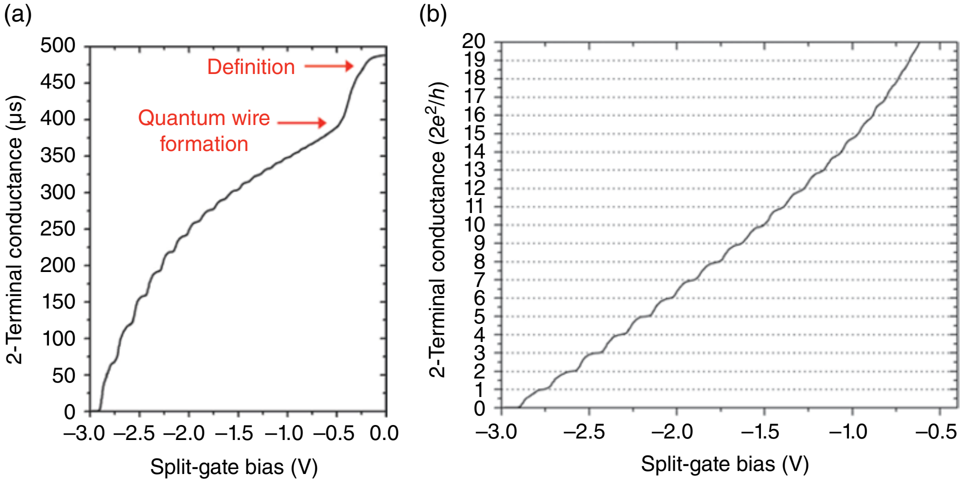


FIGURE 72.21 (a) The conductance as a function of gate bias for ballistic electron transport through a 1D channel defined in a GaAs quantum well. Below −0.5 V the conductance exhibits quantized steps. (b) The data in (a) normalized to the conductance quantum $2e^2/h$ [137].

electrons can traverse the entire length of the quantum wire without scattering, and this is known as ballistic transport. It was by studying these ultrapure 1D nanostructures that researchers in 1988 discovered new physics related to their unique quantum mechanical properties [155, 156].

As the negative gate bias is increased, the conductance of the channel initially undergoes a sharp but continuous decrease as W starts to reduce (Fig. 72.21a). This behavior is called “definition.” However, at some particular negative gate bias (around −0.5 V in Fig. 72.21a), the channel conductance abruptly changes slope signifying the formation of the quantum wire. At increasingly negative gate bias, the channel conductance is no longer continuous but becomes quantized into steps (Fig. 72.21a). When researchers first observed these plateaus in the conductance through the quantum wire it came as a surprise.

After subtracting any contact and series resistances from the measurement, researchers discovered that the height of each step was identical (Fig. 72.21b). The increase in conductance at each step is the conductance quantum $2e^2/h$ (where e is the charge of the electron and h is Planck’s constant) [155, 156]. This conductance quantum is the exact contribution to the conductance offered by electrons moving through a single 1D edge state. As the negative gate bias increases and W gets smaller, these 1D edge states are sequentially depopulated or “squeezed out” of the quantum wire and so the conductance goes down by $2e^2/h$. Using magnetic fields to separate the electrons into two populations depending on their spin, it is possible to observe additional conductance plateaus at e^2/h values [155, 157–159].

This is not the end of the story however. Below the last $2e^2/h$ plateau where the last 1D conductance channel has been depopulated and the conductance should go to zero,

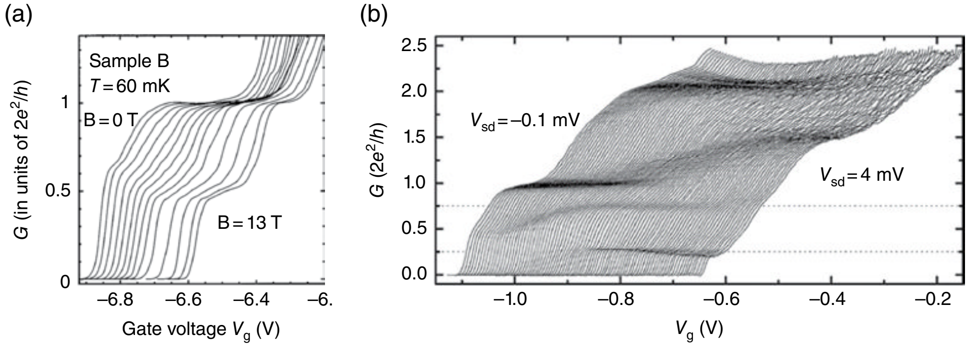


FIGURE 72.22 (a) Plateau in the conductance of a GaAs/AlGaAs quantum wire at $0.7(2e^2/h)$ showing how in the presence of an increasingly strong magnetic field it develops into a plateau at $0.5(2e^2/h)$. Source: Thomas *et al.* [160]. Reproduced with permission of the American Physical Society. (b) The evolution of sub-$(2e^2/h)$ conductance plateaus as an increasingly large voltage, V_{sd}, is applied along the length of an $In_{0.75}Ga_{0.25}As$ quantum wire (successive V_{sd} traces are offset laterally for clarity). Particularly striking is the emergence of plateau features at exactly $0.25(2e^2/h)$ and $0.75(2e^2/h)$. Source: Simmonds *et al.* [158]. Reproduced with permission of AIP Publishing LLC.

some strange behavior has been observed. Additional plateaus appear under certain measurement conditions. The first and most prominent is the so-called “0.7 structure,” which appears at $0.7(2e^2/h)$ (Fig. 72.22a) [160]. Other well-known sub-$2e^2/h$ features are seen at $0.20–0.25(2e^2/h)$ and $0.75–0.85(2e^2/h)$ (Fig 72.22b) [158, 161, 162]. These features are believed to arise due to spontaneous spin polarization in the quantum wires as a result of electron–electron interaction. Their precise origin remains the subject both of considerable debate and a great deal of research [158, 160–165].

72.4.2.2 Carbon Nanotubes There are other ways of creating quantum wires in addition to forming a gated channel in a quantum well. The first is to use carbon nanotubes. Discovered in 1991 by Iijima, carbon nanotubes consist of one or more sheets of graphene rolled up to form a cylinder [166]. As discussed in Section 72.4.1.1, graphene is a single atomic layer of carbon atoms arranged in a hexagonal sheet with purely in-plane sp^2 bonding [142]. The precise way in which the graphene sheet is rolled can have a big effect on both the shape (Fig. 72.23) and the properties of the carbon nanotube, for example, causing it to be metallic or semiconducting in nature [167, 168]. The walls of a carbon nanotube can be as thin as just one atom, and their diameter is typically in the range 0.4–5 nm (depending on the number of graphene sheets rolled up), but their length can be several microns. Carbon nanotubes therefore represent almost perfect 1D nanomaterials.

To measure their 1D electronic properties, carbon nanotubes must be contacted with metal electrodes at both ends. One way to do this is to physically place a carbon nanotube on an insulating surface so that it touches two gold electrodes. By sweeping the

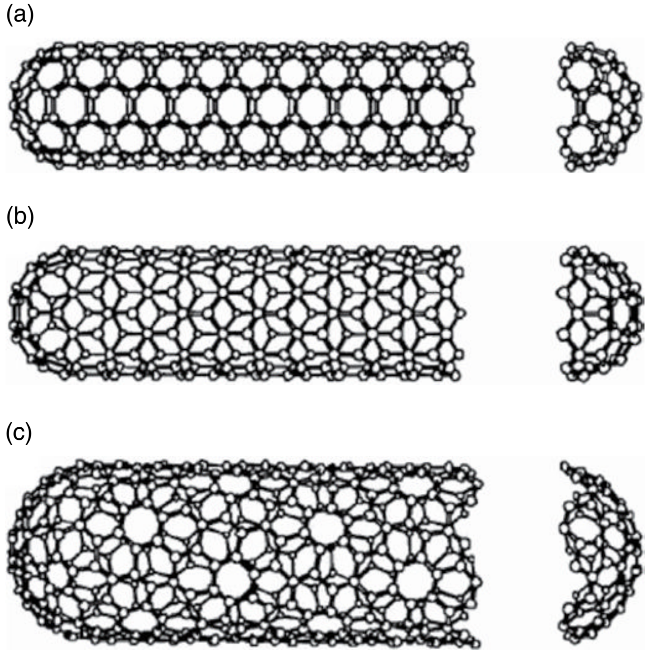


FIGURE 72.23 Three examples of the crystallographic arrangements that can be adopted by carbon nanotubes. The differences between them come from the angle at which the graphene sheet is cut before it is rolled up. The three examples previously of crystallographic arrangements adopted by carbon nanotubes are often referred to as (a) "armchair" nanotubes—graphene cut at 30°; (b) "zigzag" nanotubes—graphene cut at 0°; and (c) "helical" nanotubes—graphene cut at some intermediate angle. Source: Dresselhaus *et al.* [167]. Reproduced with permission of Springer.

voltage between the two electrodes and measuring the current that flows in the carbon nanotube, it is possible to map out the "ladder" of quantized 1D electron energy levels (Fig. 72.24).

As the voltage is raised, electrons in the left electrode become aligned in energy with a quantized 1D electron state in the carbon nanotube and a step in the current is observed. Each step in the current indicates that an additional 1D state in the carbon nanotube can now be accessed.

A group at the Georgia Institute of Technology used an ingenious method to measure quantized conductance in carbon nanotubes [170]. A fiber consisting of an enormous number of carbon nanotubes was created. At the tip of the fiber, a few individual carbon nanotubes were sticking out from the fiber further than their neighbors. The nanotube fiber was slowly brought down until it came into contact with a liquid metal (either mercury or gallium), which served as the second electrode. The conductance through the fiber was monitored throughout the experiment. Initially, when there is no contact, the conductance is zero. However, once the first carbon nanotube contacts the liquid metal electrode, there is a sudden jump in the conductance by G_0, where

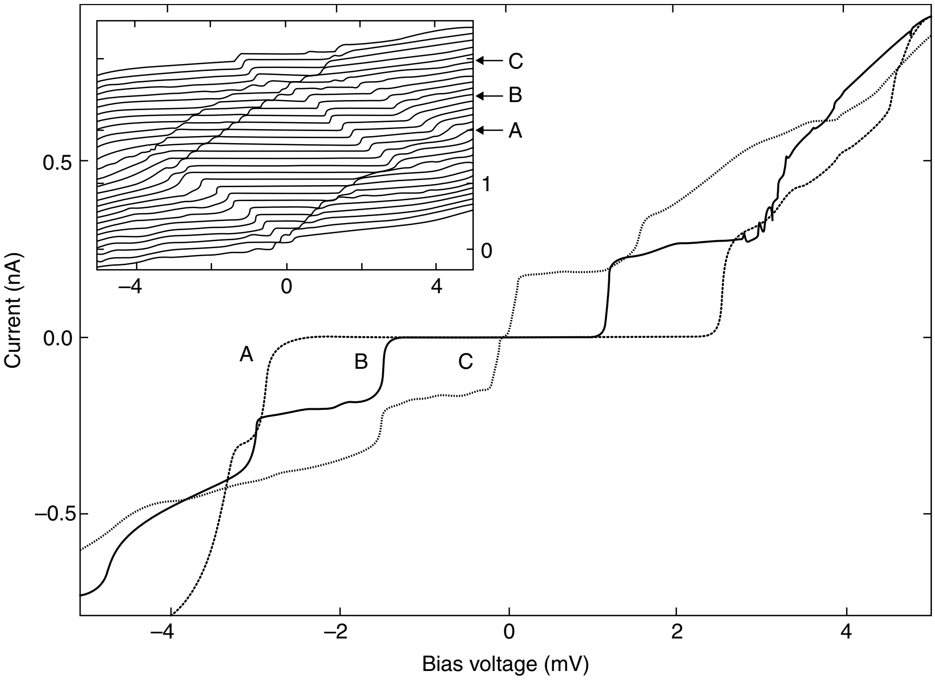


FIGURE 72.24 Current–voltage measurements of a carbon nanotube with electrical contacts at each end reveal its quantized electron levels. A third electrode is used to apply different gate bias (curves A, B, and C). Source: Tans *et al.* [169]. Reproduced with permission of Nature Publishing Group.

$G_0 = 2e^2/h$. As the fiber is pushed down further, additional nanotubes touch the metal, and, each time, a step of G_0 in the conductance is seen. $2e^2/h$ is the same conductance quantum that was measured for each 1D edge state in the gate-defined quantum wires earlier. This experiment proves that carbon nanotubes behave as perfect 1D quantum wires [167, 168, 170].

In addition to their 1D electrical properties, carbon nanotubes are incredibly useful for a very wide range of applications [171]. Due to their sp^2 bonding, they couple enormous mechanical strength under tension with very low weight, making them an excellent additive for strengthening anything from construction materials to clothing. They can be grown in macroscopic lengths of more than 10 cm, which, given their nanoscale diameter, corresponds to extraordinary aspect ratios in the hundreds of millions [172]. Given their excellent electrical properties, carbon nanotubes with such enormous lengths could be used to directly build circuits and devices. They can be used as tips in various scanning probe microscopy techniques and are even a highly promising candidate for hydrogen storage, which is a key challenge in the development of practical fuel cell technology. Their diverse properties mean that current applications for carbon nanotubes range from biomedicine to clean energy [173, 174].

72.4.2.3 Quantum Wires in Self-Assembled Nanopillars In an offshoot from the work on self-assembled nanopillars discussed in Section 72.3.12, researchers have shown that individual nanopillars can be turned into quantum wires.

Individual semiconductor nanopillars are grown, snapped off and placed on an insulating substrate, and finally contacted with electrodes similar to the carbon nanotube in Figure 72.24a. Since self-assembled nanopillars typically have minimum diameters in the region of 50 nm, gate electrodes are used to electrostatically reduce the effective pillar width until it falls below the electron de Broglie wavelength and quantization occurs [175]. Evidence exists that exotic new quasiparticles, called Majorana fermions, are observed when a quantum wire formed in an InSb nanopillar is brought into contact with a superconductor due to the strong spin–orbit coupling in the semiconductor [176].

1D conductance quantization has also been demonstrated in atomically thin metal wires. Quantum wires are created by mechanically pulling apart nanoscale gold contacts while continually measuring the electrical conductance. Chains of at least four gold atoms form in which the highest measured conductance is $2e^2/h$, proving that they are 1D in nature. Electrons traverse these gold atomic chains without scattering meaning that they could be very useful for nanoscale electronics [177].

72.4.2.4 Epitaxial Self-Assembled Quantum Wires Finally, arrays of self-assembled quantum wires can be created during the growth of certain compressively strained semiconductor heterostructures. When InAs is deposited on InP(001) (or several of its lattice-matched alloys: InAlAs, InGaAs, AlAsSb, and GaAsSb), it spontaneously forms thin "dash-like" nanostructures parallel to the [−110] direction (Fig. 72.25) [178–180]. These nanostructures are typically less than 5 nm high and 15–20 nm wide but can be 500 nm or more in length fulfilling the requirements for a quantum wire.

The InAs quantum wires self-assemble via the Stranski–Krastanov mode where growth initially proceeds by the formation of a flat 2D wetting layer. However, after

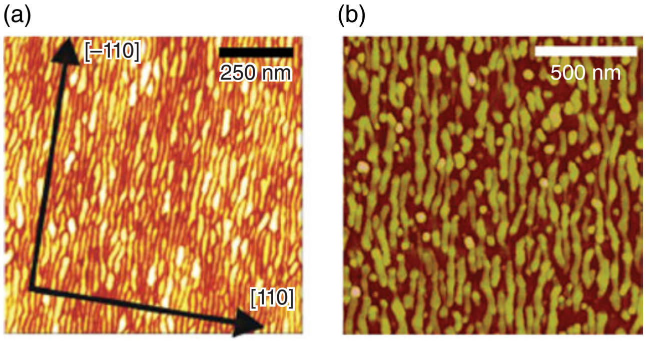


FIGURE 72.25 Self-assembled InAs quantum wires formed on (a) InAlAs (Source: Simmonds *et al.* [178]. Reproduced with permission of the American Vacuum Society) and (b) GaAsSb surfaces (Source: Simmonds *et al.* [179]. Reproduced with permission of AIP Publishing LLC). Regardless of the buffer material, the InAs wires form parallel to the [−110] direction.

the deposition of some critical thickness, there is a sudden transition to 3D growth due to the accumulation of compressive strain in the InAs layer and the quantum wires form. The anisotropic self-assembly of the wires is attributed to the fact that growth is energetically more favorable in the [−110] direction [178, 180].

This common InAs wetting layer and the fact that the InAs nanostructures are closely packed into arrays mean that measuring the electrical properties of a single quantum wire is not possible. Quantized conductance cannot therefore be demonstrated in these quantum wires, but their 1D nature has been confirmed by mapping out their density of states (Fig. 72.26). A qualitative comparison of the 10 K curves in Figure 72.26 with

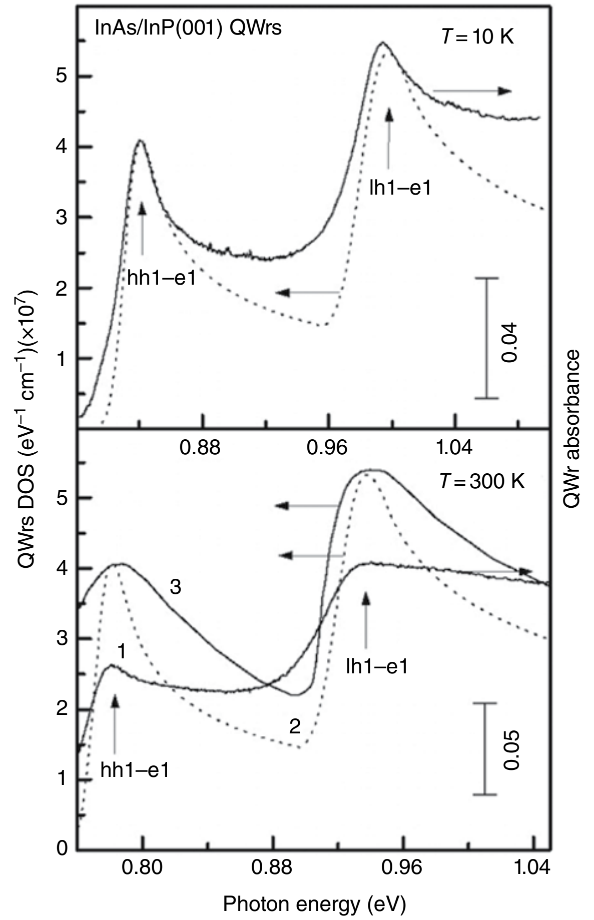


FIGURE 72.26 Comparison between simulated density of states for self-assembled InAs quantum wires on InP(001) and their experimental optical absorption spectra at 10 K (top) and 300 K (bottom). In the top panel the simulated (experimental) data are the dotted (solid) lines. In the bottom panel the experimental data are line 1, while two density of states curves simulated to take into account different broadening mechanisms appear as lines 2 and 3. Source: Mazur *et al.* [181]. Reproduced with permission of IOP Publishing.

the 1D DoS in Figure 72.16 confirms the characteristic shape of a 1D quantum system and indicates that these are indeed quantum wires.

Self-assembled InAs quantum wires have found multiple applications. Their high density of states at each quantized electron energy level is helpful for reducing the threshold current density in lasers [181–183]. What is more, the wavelength of light emitted by InAs quantum wires on InP and its alloys is highly tunable. Changing the size of the wires by changing the amount of InAs deposited changes the position of the quantized energy levels (in the same way as we saw for quantum well nanostructures). The emission wavelength can readily be tuned over a very wide range from 1.1 to 1.8 μm (Fig. 72.27). This range covers the 1.3 and 1.55 μm wavelengths that are critically important to fiber optic-based telecommunications [181].

Until very recently, the only way to create self-assembled quantum wires was to use compressive strain. However, a method for creating self-assembled quantum wires from tensile-strained semiconductors has recently been discovered [184]. The underlying growth processes by which these tensile-strained nanostructures self-assemble are very similar to those in traditional compressively strained self-assembly [185–187]. However, the tensile strain has some interesting and potentially very useful effects on the properties of the quantum wires. Perhaps the most striking result is that unlike compressive strain, which increases the semiconductor bandgap of the quantum wire, tensile strain *reduces* the bandgap. As a result, tensile-strained quantum wires emit light at longer wavelengths than can be reached in either bulk semiconductor material or in

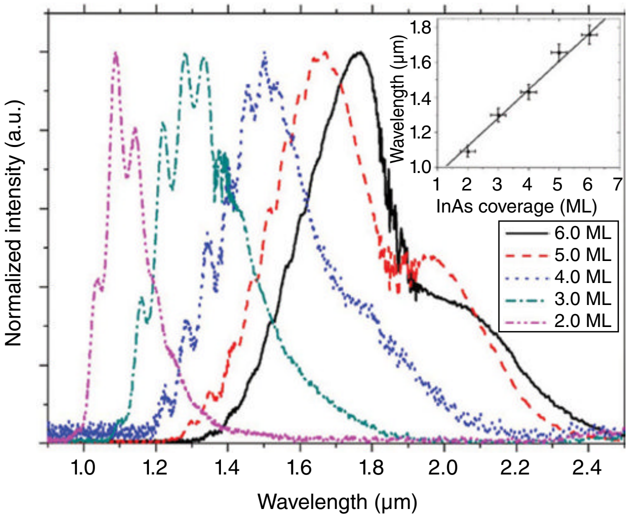


FIGURE 72.27 Photoluminescence from InAs quantum wires measured at 4.2 K showing straightforward control of the emission wavelength with InAs deposition thickness in monolayers (ML). (Inset) Peak emission wavelength from the 5 PL spectra plotted against InAs coverage revealing a linear dependence of 170 nm/ML. Source: Simmonds *et al.* [178]. Reproduced with permission of the American Vacuum Society.

compressive-strained quantum wires. This push toward the infrared end of the electromagnetic spectrum could have important implications for technologies including night vision, astronomy, and trace gas sensing [184].

72.4.3 Quantum Dot Nanostructures

Taking the progression from left to right in Figure 72.15 to its logical conclusion, we reach quantum dots, in which electrons experience quantum confinement in all three spatial dimensions [188]. The theoretical 0D DoS consists of a series of delta functions at discrete energies, although in reality each energy level is somewhat broadened due to inhomogeneity and thermal effects. This 0D DoS has many practical benefits for device applications. Using the example of the laser again, quantum dot lasers exhibit low threshold currents, high gains, and large T_0 values (i.e., excellent thermal stability of the threshold current) [189]. In addition, the discrete electron energy levels in a quantum dot should result in highly monochromatic light emission. By changing the quantum dot size, the emission wavelength can be tuned, as for both quantum wells and wires. The unique electronic structure of quantum dots places them at the forefront of research into a wide variety of cutting-edge technologies, including quantum computation and quantum cryptography [190–192].

72.4.3.1 Gate-Defined Quantum Dots A 2D sheet of electrons in a quantum well can be patterned into quantum dots using electrostatic gates (in the same way as we saw for quantum wires) [193, 194]. In this case, the gates are shaped to create quasicircular regions in the electron sheet. Controlling the gate voltages allows the size of the circular region to be shrunk until its size is less than the de Broglie wavelength in both in-plane directions and the out-of-plane direction of the quantum well. Once inside the quantum dot, the energy of an electron is quantized in all three directions.

Electrons can even be pushed in and out of the quantum dots using additional gates as "plungers." Indeed, due to the ease with which additional gates can be added, arrays of multiple quantum dots can be created in almost any geometry. Such versatile systems serve as a rich toolbox for the exploration of phenomena such as coupling between adjacent quantum dots, quantum computation, or even as part of a quantum refrigerator (Fig. 72.28) [195–199].

72.4.3.2 Colloidal Quantum Dots In Section 72.3.13 we discussed this family of quantum dot nanoparticles in relation to their applications for nanomedicine [95]. However, as well as their size and the ease with which they can be functionalized, the quantum mechanical properties of colloidal dots are also of great interest. These solution-based quantum dots are created using chemical reactions between various metal oxide and metal organic precursors [200]. This approach is extremely versatile. Colloidal quantum dot nanoparticles can be created from a range of different materials (most commonly the II–VI family of semiconductors, such as CdSe and CdTe) [201].

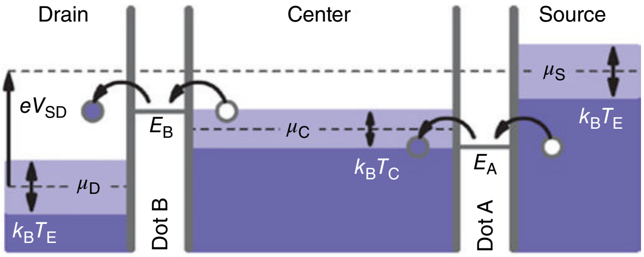


FIGURE 72.28 A quantum dot refrigerator created by connecting three 2D sheets of electrons in quantum wells with two quantum dots. An electron flows from the Source into the Center via the quantized energy level E_A in dot A. The quantized energy of the Center 2D electron sheet is thermally broadened. The quantized energy level E_B of dot B is made to be coincident with that of the hottest electrons in the Center so that these are selected to flow out to the Drain. In summary, electrons are injected into the Center with low energy and removed with higher energy such that the overall amount of thermal energy removed is E_B-E_A and this cools the Center region. Source: Prance *et al.* [195]. Reproduced with permission of the American Physical Society.

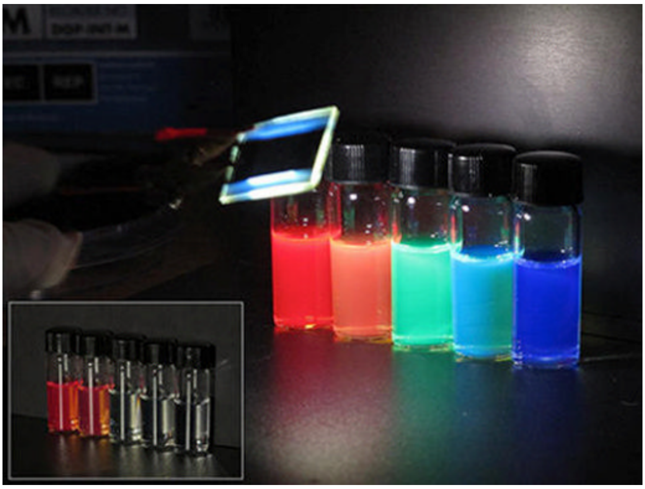


FIGURE 72.29 Depending on their size, colloidal quantum dots dispersed in hexane emit light across the entire visible spectrum (wavelengths 430–620 nm) when viewed under UV light. The inset image shows the QDs illuminated by the tungsten room light without UV irradiation. Source: Kwak *et al.* [203]. Reproduced with permission of the American Chemical Society.

Coating a low bandgap nanoparticle with a larger bandgap shell creates the conditions necessary for quantum confinement of electrons [202]. Light emission wavelength can be readily tuned by controlling the core diameter of these core–shell quantum dots (Fig. 72.29) [203, 204].

As we have seen, colloidal quantum dots are ideally suited to functionalization for use in biomedical applications [95, 97–100]. However, these colloidal quantum dots can additionally be used as building blocks and packed together into 3D superlattices [205]. Their solution-based nature means that thin films of these quantum dots can be easily prepared using standard spin-coating or inkjet printing techniques [206]. These simple production techniques are significantly cheaper than other synthesis techniques such as molecular beam epitaxy or metal organic chemical vapor deposition. The attractive properties and cost-effectiveness of colloidal quantum dots mean that they are being developed for use in a wide range of electronic and optoelectronic devices. For a comprehensive review of colloidal quantum dots, their properties, and their applications, see Ref. 207.

72.4.3.3 Epitaxial Self-Assembled Quantum Dots Similar to the self-assembled nanowires discussed in Section 72.4.2.4, the strain in semiconductor materials can be harnessed to drive the creation of quantum dot nanostructures. Traditionally, *compressive* strain has been used as the driving force for the growth of these nanomaterials. The two most commonly explored self-assembled systems are In(Ga)As quantum dots in a GaAs matrix [208] and Ge quantum dots in a Si matrix [209]. Since their discovery in the early 1990s, these and other quantum dot systems have been explored extensively, and their properties employed in a wide range of device applications, including lasers, solar cells, and quantum computers [119, 179, 192, 210–215].

Self-assembled quantum dots usually nucleate and grow at random locations across the substrate surface (Fig. 72.30 (left)). However, by patterning the substrate before growth, quantum dots can be made to nucleate at only at predetermined positions (Fig. 72.30 (right)) [216–218]. This technique can be used to generate quantum dot molecules or to assist with locating the quantum dots during subsequent device fabrication. For example, one area of current interest is to create photonic crystal cavities

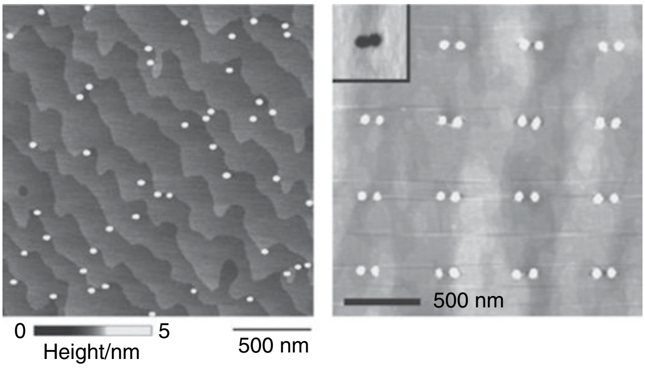


FIGURE 72.30 (Left) InAs self-assembled quantum dots typically form at random locations across the surface of a GaAs substrate. (Right) If small pits are patterned into the GaAs surface prior to InAs deposition, the quantum dots can be made to form only in the location of the pits, leading to an ordered array. In this case the growth conditions have been controlled such that the quantum dots assemble in pairs. Source: Atkinson *et al.* [216]. Reproduced with permission of Elsevier.

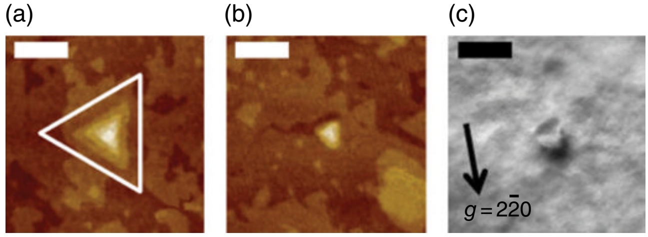


FIGURE 72.31 (a) Atomic force micrograph of a large tensile-strained quantum dot showing its symmetric triangular shape (the white equilateral triangle is drawn for comparison). (b) A tensile quantum dot of smaller, more typical size. (c) Plan-view transmission electron micrograph of a capped quantum dot showing that it is under tensile strain but free from strain-related defects. All scale bars are 100 nm. Source: Yerino *et al.* [221]. Reproduced with permission of AIP Publishing LLC.

that contain a single quantum dot. Interesting optical effects occur when the emission from the quantum dot coincides with the resonance of the photonic crystal cavity. These hybrid structures can be created from randomly dispersed quantum dot layers by creating many photonic crystal cavities and then looking for one that contains a single dot [219]. However, a more reliable, efficient, and high-yield approach is first to know precisely where the quantum dots are and then to create the photonic crystals around them [216].

More recently a method that uses *tensile* strain to drive quantum dot self-assembly has also been discovered [220, 221]. Like the tensile-strained quantum wires discussed at the end of Section 72.4.2.4, the residual tensile strain in these quantum dots causes them to emit light at longer wavelengths than is possible in either the bulk material or compressively strained quantum dots. This new way of creating self-assembled quantum dots opens up the possibility of creating families of quantum nanomaterials with small bandgaps, a feat that previously has been very difficult. What is more, these tensile-strained quantum dots can be grown with extremely high symmetry (Fig. 72.31), which gives them a very low fine structure splitting. This property makes them ideally suited to quantum information applications, such as quantum computing and quantum cryptography [221].

It is also possible to grow quantum dot nanostructures without using strain. The most common approach is known as droplet epitaxy. The first step is to deposit a tiny amount of metal onto the surface, which organizes itself into discrete liquid nanodroplets dispersed across the surface. The second step is to crystalize the semiconductor material out of these metal droplets by annealing. For example, to create an unstrained InAs quantum dot on GaAs, one begins by depositing pure indium metal to form the nanodroplets. Then, to turn the metal into InAs, the nanodroplets are annealed under an arsenic atmosphere, and the III–V semiconductor crystalizes. This approach allows quantum dots to be created on traditionally incompatible substrates where the strain-driven approach fails [222, 223]. In addition, this approach can also be used to create quantum dot molecules that may find application in quantum computation [224].

72.5 CONCLUSIONS

Given the enormous scope of current research into the development of nanomaterials with new and exciting properties, in this chapter it has been impossible to do more than give a flavor of some of the major areas of exploration. As a result there are many omissions and highly simplified or abridged descriptions of complex topics that I would have loved to address more completely given more space. Nevertheless, I trust that the interested reader will be able to use the references provided as a starting point from which to explore these areas in more detail.

Despite these caveats, it is my hope that the foregoing sections have provided a useful introduction to nanomaterials and their properties. In particular, I have focused on the influence of the high surface-to-volume ratio and the role of quantum confinement on nanomaterial properties. In doing so, I have attempted to explain why material properties at the nanoscale can be, at times, very different from those at the macroscale. From color and magnetic behavior to electronic structure, nanomaterials provide us with an opportunity to engineer and fine-tune the fundamental properties of a wide range of materials. As we have seen, numerous exciting scientific discoveries and hugely influential technological advances owe everything to the unique properties of nanomaterials. Ongoing research into nanomaterials will drive future innovation.

REFERENCES

1. A. B. Kesel, A. Martin, T. Seidl, Getting a grip on spider attachment: an AFM approach to microstructure adhesion in arthropods, *Smart Mater. Struct.* 13, 512–518 (2004).
2. Y. T. Cheng, D. E. Rodak, C. A. Wong, C. A. Hayden, Effects of micro- and nano-structures on the self-cleaning behaviour of lotus leaves, *Nanotechnology* 17, 1359–1362 (2006).
3. V. Saranathan *et al.*, Structure, function, and self-assembly of single network gyroid (I4132) photonic crystals in butterfly wing scales, *Proc. Natl. Acad. Sci.* 107, 11676–11681 (2010).
4. R. A. Shelby, D. R. Smith, S. Schultz, Experimental verification of a negative index of refraction, *Science* 292, 77–80 (2001).
5. J. D. Joannopoulos, P. R. Villeneuve, S. Fan, Photonic crystals: putting a new twist on light, *Nature* 386, 143–149 (1997).
6. E. Yablonovitch, Inhibited spontaneous emission in solid-state physics and electronics, *Phys. Rev. Lett.* 58, 2059–2062 (1987).
7. E. Yablonovitch, Photonic band-gap structures, *J. Opt. Soc. Am. B* 10, 283–295 (1993).
8. D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, S. Schultz, Composite medium with simultaneously negative permeability and permittivity, *Phys. Rev. Lett.* 84, 4184–4187 (2000).
9. D. Schurig *et al.*, Metamaterial electromagnetic cloak at microwave frequencies, *Science* 314, 977–980 (2006).
10. E. Roduner, Size matters: why nanomaterials are different, *Chem. Soc. Rev.* 35, 583–592 (2006).
11. A. Y. Cho, J. R. Arthur, Molecular beam epitaxy, *Prog. Solid State Chem.* 10, 157–191 (1975).

12. H. M. Manasevit, Single-crystal gallium arsenide on insulating substrates, *Appl. Phys. Lett.* 12, 156–159 (1968).
13. H. M. Manasevit, W. I. Simpson, The use of metal-organics in the preparation of semiconductor materials I. Epitaxial gallium-V compounds, *J. Electrochem. Soc.* 116, 1725–1732 (1969).
14. W.-D. Li, W. Wu, R. S. Williams, *Single-Digit Nanometer Nanoimprint Templates*, SPIE Newsroom, Bellingham, WC, 2013.
15. X. Wen, L. M. Traverso, P. Srisungsitthisunti, X. Xu, E. E. Moon, Optical nanolithography with λ/15 resolution using bowtie aperture array, *Appl. Phys. A* 117, 307–311 (2014).
16. V. Auzelyte *et al.*, Extreme ultraviolet interference lithography at the Paul Scherrer Institut, *J. Micro/Nanolithography, MEMS, MOEMS* 8, 021204 (2009).
17. L. Pan *et al.*, Maskless plasmonic lithography at 22 nm resolution, *Sci. Rep.* 1, 175 (2011).
18. Y. Kyoung Ryu, P. Aitor Postigo, F. Garcia, R. Garcia, Fabrication of sub-12 nm thick silicon nanowires by processing scanning probe lithography masks, *Appl. Phys. Lett.* 104, 223112 (2014).
19. D. Mijatovic, J. Eijkel, A. van den Berg, Technologies for nanofluidic systems: top-down vs. bottom-up—a review, *Lab Chip* 5, 492–500 (2005).
20. B. A. Joyce, D. D. Vvedensky, Self-organized growth on GaAs surfaces, *Mater. Sci. Eng. R* 46, 127–176 (2004).
21. J. W. van Honschoten, N. Brunets, N. R. Tas, Capillarity at the nanoscale, *Chem. Soc. Rev.* 39, 1096–1114 (2010).
22. P. Serp, K. Philippot, Eds., *Nanomaterials in Catalysis*, Wiley-VCH, Weinheim, Germany, 2013.
23. X. D. Wang, C. J. Summers, Z. L. Wang, Mesoporous single-crystal ZnO nanowires epitaxially sheathed with Zn2SiO4, *Adv. Mater.* 16, 1215–1218 (2004).
24. D. M. Eigler, E. K. Schweizer, Positioning single atoms with a scanning tunnelling microscope, *Nature* 344, 524–526 (1990).
25. U. Kreibig, M. Vollmer, *Optical Properties of Metal Clusters*, Springer, Berlin, 1995.
26. P. Buffat, J.-P. Borel, Size effect on the melting temperature of gold particles, *Phys. Rev. A* 13, 2287–2298 (1976).
27. K. M. Unruh, T. E. Huber, C. A. Huber, Melting and freezing behavior of indium metal in porous glasses, *Phys. Rev. B* 48, 9021–9027 (1993).
28. K. Koga, T. Ikeshoji, K. Sugawara, Size- and temperature-dependent structural transitions in gold nanoparticles, *Phys. Rev. Lett.* 92, 115507 (2004).
29. Y. Nakae *et al.*, Anomalous spin polarization in Pd and Au nano-particles, *Phys. B* 284–288, 1758–1759 (2000).
30. Y. Yamamoto *et al.*, Magnetic properties of the noble metal nanoparticles protected by polymer, *Phys. B* 329–333, 1183–1184 (2003).
31. N. Watari, S. Ohnishi, Atomic and electronic structures of Pd_{13} and Pt_{13} clusters, *Phys. Rev. B* 58, 1665–1677 (1998).
32. Y. Sakamoto *et al.*, Ferromagnetism of Pt nanoparticles induced by surface chemisorption, *Phys. Rev. B* 83, 104420 (2011).

33. C. Q. Sun, Dominance of broken bonds and nonbonding electrons at the nanoscale, *Nanoscale* 2, 1930–1961 (2010).
34. G. Polizos *et al.*, Scalable superhydrophobic coatings based on fluorinated diatomaceous earth: abrasion resistance versus particle geometry, *Appl. Surf. Sci.* 292, 563–569 (2014).
35. B. D'Urso, J. T. Simpson, M. Kalyanaraman, Nanocone array glass, *J. Micromech. Microeng.* 17, 717–721 (2007).
36. T. Aytug *et al.*, Optically transparent, mechanically durable, nanostructured superhydrophobic surfaces enabled by spinodally phase-separated glass thin films, *Nanotechnology* 24, 315602 (2013).
37. W. Sparreboom, A. van den Berg, J. C. T. Eijkel, Transport in nanofluidic systems: a review of theory and applications, *New J. Phys.* 12, 015004 (2010).
38. J. M. Oh, T. Faez, S. de Beer, F. Mugele, Capillarity-driven dynamics of water–alcohol mixtures in nanofluidic channels, *Microfluid. Nanofluid.* 9, 123–129 (2009).
39. R. B. Schoch, J. Han, P. Renaud, Transport phenomena in nanofluidics, *Rev. Mod. Phys.* 80, 839–883 (2008).
40. J. C. T. Eijkel, A. van den Berg, Nanofluidics: what is it and what can we expect from it? *Microfluid. Nanofluid.* 1, 249–267 (2005).
41. A. H. J. Yang *et al.*, Optical manipulation of nanoparticles and biomolecules in subwavelength slot waveguides, *Nature* 457, 71–75 (2009).
42. J. O. Tegenfeldt *et al.*, Micro- and nanofluidics for DNA analysis, *Anal. Bioanal. Chem.* 378, 1678–1692 (2004).
43. J. Goldberger, R. Fan, P. Yang, Inorganic nanotubes: a novel platform for nanofluidics, *Acc. Chem. Res.* 39, 239–248 (2006).
44. S. S.-Y. Chui, S. M.-F. Lo, J. P. H. Charmant, A. G. Orpen, I. D. Williams, A chemically functionalizable nanoporous material $[Cu_3(TMA)_2(H_2O)_3]_n$, *Science* 283, 1148–1150 (1999).
45. D. Bradshaw, J. B. Claridge, E. J. Cussen, T. J. Prior, M. J. Rosseinsky, Design, chirality, and flexibility in nanoporous molecule-based materials, *Acc. Chem. Res.* 38, 273–282 (2005).
46. F. Ç. Cebeci, Z. Wu, L. Zhai, R. E. Cohen, M. F. Rubner, Nanoporosity-driven superhydrophilicity: a means to create multifunctional antifogging coatings, *Langmuir* 22, 2856–2862 (2006).
47. C. Sanchez, C. Boissière, D. Grosso, C. Laberty, L. Nicole, Design, synthesis, and properties of inorganic and hybrid thin films having periodically organized nanoporosity, *Chem. Mater.* 20, 682–737 (2008).
48. A. Krotkus, Semiconductors for terahertz photonics applications, *J. Phys. D. Appl. Phys.* 43, 273001 (2010).
49. D. V. Seletskiy *et al.*, Efficient terahertz emission from InAs nanowires, *Phys. Rev. B* 84, 115421 (2011).
50. V. N. Trukhin *et al.*, Terahertz generation by GaAs nanowires, *Appl. Phys. Lett.* 103, 072108 (2013).
51. A. Atrashchenko *et al.*, Giant enhancement of terahertz emission from nanoporous GaP, *Appl. Phys. Lett.* 105, 191905 (2014).

52. M. Reid, I. V. Cravetchi, R. Fedosejevs, I. M. Tiginyanu, L. Sirbu, Enhanced terahertz emission from porous InP (111) membranes, *Appl. Phys. Lett.* 86, 021904 (2005).

53. R. Adomavičius *et al.*, Terahertz pulse emission from nanostructured (311) surfaces of GaAs, *J. Infrared Millimeter Terahertz Waves* 33, 599–604 (2012).

54. A.-H. Lu, F. Schüth, Nanocasting: a versatile strategy for creating nanostructured porous materials, *Adv. Mater.* 18, 1793–1805 (2006).

55. J. Erlebacher, M. J. Aziz, A. Karma, N. Dimitrov, K. Sieradzki, Evolution of nanoporosity in dealloying, *Nature* 410, 450–453 (2001).

56. X. Zhao *et al.*, Influence of surface properties on the electrical conductivity of silicon nanomembranes, *Nanoscale Res. Lett.* 6, 402 (2011).

57. C. Jiang, V. V. Tsukruk, Freestanding nanostructures via layer-by-layer assembly, *Adv. Mater.* 18, 829–840 (2006).

58. C. Jiang, S. Markutsya, Y. Pikus, V. V. Tsukruk, Freely suspended nanocomposite membranes as highly sensitive sensors, *Nat. Mater.* 3, 721–728 (2004).

59. R. Narayanan, M. A. El-Sayed, Catalysis with transition metal nanoparticles in colloidal solution: nanoparticle shape dependence and stability, *J. Phys. Chem. B* 109, 12663–12676 (2005).

60. Z. Peng, H. Yang, Designer platinum nanoparticles: control of shape, composition in alloy, nanostructure and electrocatalytic property, *Nano Today* 4, 143–164 (2009).

61. C.-J. Zhong *et al.*, Fuel cell technology: nano-engineered multimetallic catalysts, *Energy Environ. Sci.* 1, 454–466 (2008).

62. B. Seger, P. V. Kamat, Electrocatalytically active graphene-platinum nanocomposites. Role of 2-D carbon support in PEM fuel cells, *J. Phys. Chem. C* 113, 7990–7995 (2009).

63. P. H. Matter, L. Zhang, U. S. Ozkan, The role of nanostructure in nitrogen-containing carbon catalysts for the oxygen reduction reaction, *J. Catal.* 239, 83–96 (2006).

64. M.-L. Seol, J.-H. Ahn, J.-M. Choi, S.-J. Choi, Y.-K. Choi, Self-aligned nanoforest in silicon nanowire for sensitive conductance modulation., *Nano Lett.* 12, 5603–5608 (2012).

65. K. A. Bertness, N. A. Sanford, A. V. Davydov, GaN nanowires grown by molecular beam epitaxy, *IEEE J. Sel. Top. Quantum Electron.* 17, 847–858 (2011).

66. A. Lin *et al.*, Extracting transport parameters in GaAs nanopillars grown by selective-area epitaxy, *Nanotechnology* 23, 105701 (2012).

67. A. C. Scofield *et al.*, Bottom-up photonic crystal cavities formed by patterned III-V nanopillars, *Nano Lett.* 11, 2242 (2011).

68. T. Mårtensson *et al.*, Epitaxial III–V nanowires on silicon, *Nano Lett.* 4, 1987–1990 (2004).

69. G. Mariani, A. Scofield, C.-H. Hung, D. L. Huffaker, GaAs nanopillar-array solar cells employing in situ surface passivation, *Nat. Commun.* 4, 1497 (2013).

70. R. Yan, D. Gargas, P. Yang, Nanowire photonics, *Nat. Photonics* 3, 569–576 (2009).

71. F. Qian, S. Gradečak, Y. Li, C.-Y. Wen, C. M. Lieber, Core/multishell nanowire heterostructures as multicolor, high-efficiency light-emitting diodes, *Nano Lett.* 5, 2287–2291 (2005).

72. M. Borgström, K. Deppert, L. Samuelson, W. Seifert, Size- and shape-controlled GaAs nano-whiskers grown by MOVPE: a growth study, *J. Cryst. Growth* 260, 18–22 (2004).

73. B. Mandl *et al.*, Growth mechanism of self-catalyzed group III–V nanowires, *Nano Lett.* 10, 4443–4449 (2010).

74. H. Sekiguchi, K. Kishino, A. Kikuchi, Emission color control from blue to red with nanocolumn diameter of InGaN/GaN nanocolumn arrays grown on same substrate, *Appl. Phys. Lett.* 96, 231104 (2010).

75. Z. Fan *et al.*, Ordered arrays of dual-diameter nanopillars for maximized optical absorption, *Nano Lett.* 10, 3823–3827 (2010).

76. A. C. Scofield *et al.*, Bottom-up photonic crystal lasers, *Nano Lett.* 11, 5387–5390 (2011).

77. J. C. Johnson *et al.*, Single gallium nitride nanowire lasers, *Nat. Mater.* 1, 106–110 (2002).

78. P. Senanayake *et al.*, Photoconductive gain in patterned nanopillar photodetector arrays, *Appl. Phys. Lett.* 97, 203108 (2010).

79. G. Mariani *et al.*, Hybrid conjugated polymer solar cells using patterned GaAs nanopillars, *Appl. Phys. Lett.* 97, 013107 (2010).

80. K. Tomioka, M. Yoshimura, T. Fukui, A III–V nanowire channel on silicon for high-performance vertical transistors, *Nature* 488, 189–192 (2012).

81. T. Tanaka *et al.*, Vertical surrounding gate transistors using single InAs nanowires grown on Si substrates, *Appl. Phys. Express* 3, 025003 (2010).

82. Z. L. Wang, Piezopotential gated nanowire devices: piezotronics and piezo-phototronics, *Nanotoday* 5, 540–552 (2010).

83. C. J. Barrelet, A. B. Greytak, C. M. Lieber, Nanowire photonic circuit elements, *Nano Lett.* 4, 1981–1985 (2004).

84. F. Caruso, Nanoengineering of particle surfaces, *Adv. Mater.* 13, 11–22 (2001).

85. M. D. Stoller, S. Park, Y. Zhu, J. An, R. S. Ruoff, Graphene-based ultracapacitors, *Nano Lett.* 8, 3498–3502 (2008).

86. T. Ramanathan *et al.*, Functionalized graphene sheets for polymer nanocomposites, *Nat. Nanotechnol.* 3, 327–331 (2008).

87. Z. Liu, J. T. Robinson, X. Sun, H. Dai, PEGylated nanographene oxide for delivery of water-insoluble cancer drugs, *J. Am. Chem. Soc.* 130, 10876–10877 (2008).

88. D. Tasis, N. Tagmatarchis, A. Bianco, M. Prato, Chemistry of carbon nanotubes, *Chem. Rev.* 106, 1105–1136 (2006).

89. J. Chen *et al.*, Solution properties of single-walled carbon nanotubes, *Science* 282, 95–98 (1998).

90. R. J. Chen, Y. Zhang, D. Wang, D. Hongjie, Noncovalent sidewall functionalization of single-walled carbon nanotubes for protein immobilization, *J. Am. Chem. Soc.* 123, 3838–3839 (2001).

91. Y. Cui, Q. Wei, H. Park, C. M. Lieber, Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species, *Science* 293, 1289–1292 (2001).

92. A. Kolmakov, D. O. Klenov, Y. Lilach, S. Stemmer, M. Moskovits, Enhanced gas sensing by individual SnO_2 nanowires and nanobelts functionalized with Pd catalyst particles, *Nano Lett.* 5, 667–673 (2005).

93. Y. L. Bunimovich *et al.*, Quantitative real-time measurements of DNA hybridization with alkylated nonoxidized silicon nanowires in electrolyte solution, *J. Am. Chem. Soc.* 128, 16323–16331 (2006).

94. X. Michalet *et al.*, Quantum dots for live cells, in vivo imaging, and diagnostics, *Science* 307, 538–544 (2005).

95. X. Gao, Y. Cui, R. M. Levenson, L. W. K. Chung, S. Nie, In vivo cancer targeting and imaging with semiconductor quantum dots, *Nat. Biotechnol.* 22, 969–976 (2004).
96. I. L. Medintz, H. T. Uyeda, E. R. Goldman, H. Mattoussi, Quantum dot bioconjugates for imaging, labelling and sensing, *Nat. Mater.* 4, 435–446 (2005).
97. J. Gao, H. Gu, B. Xu, Multifunctional magnetic nanoparticles: design, synthesis, and biomedical applications, *Acc. Chem. Res.* 42, 1097–1107 (2009).
98. V. Bagalkot *et al.*, Quantum dot-aptamer conjugates for synchronous cancer imaging, therapy, and sensing of drug delivery based on bi-fluorescence resonance energy transfer, *Nano Lett.* 7, 3065–3070 (2007).
99. C. Minelli, S. B. Lowe, M. M. Stevens, Engineering nanocomposite materials for cancer therapy, *Small* 6, 2336–2357 (2010).
100. D. Kim, Y. Y. Jeong, S. Jon, A drug-loaded aptamer-gold nanoparticle bioconjugate for combined CT imaging and therapy of prostate cancer, *ACS Nano* 4, 3689–3696 (2010).
101. R. Teki *et al.*, Nanostructured silicon anodes for lithium ion rechargeable batteries, *Small* 5, 2236–2242 (2009).
102. P. Simon, Y. Gogotsi, Materials for electrochemical capacitors, *Nat. Mater.* 7, 845–854 (2008).
103. A. S. Aricò, P. Bruce, B. Scrosati, J.-M. Tarascon, W. van Schalkwijk, Nanostructured materials for advanced energy conversion and storage devices, *Nat. Mater.* 4, 366–377 (2005).
104. C. K. Chan *et al.*, High-performance lithium battery anodes using silicon nanowires, *Nat. Nanotechnol.* 3, 31–35 (2008).
105. B. Kang, G. Ceder, Battery materials for ultrafast charging and discharging, *Nature* 458, 190–193 (2009).
106. X. W. (David) Lou, L. A. Archer, Z. Yang, Hollow micro-/nanostructures: synthesis and applications, *Adv. Mater.* 20, 3987–4019 (2008).
107. P. V. Kamat, Meeting the clean energy demand: nanostructure architectures for solar energy conversion, *J. Phys. Chem. C* 111, 2834–2860 (2007).
108. M.-L. Kuo *et al.*, Realization of a near-perfect antireflection coating for silicon solar energy utilization, *Opt. Lett.* 33, 2527–2529 (2008).
109. E. Buitrago *et al.*, The top-down fabrication of a 3D-integrated, fully CMOS-compatible FET biosensor based on vertically stacked SiNWs and FinFETs, *Sensors Actuators B Chem.* 193, 400–412 (2014).
110. W. K. Choi *et al.*, Synthesis of silicon nanowires and nanofin arrays using interference lithography and catalytic etching, *Nano Lett.* 8, 3799–3802 (2008).
111. B. Alemán *et al.*, Transfer-free batch fabrication of large-area suspended graphene membranes, *ACS Nano* 4, 4762–4768 (2010).
112. R. R. Nair *et al.*, Graphene as a transparent conductive support for studying biological molecules by transmission electron microscopy, *Appl. Phys. Lett.* 97, 153102 (2010).
113. J. R. Morones *et al.*, The bactericidal effect of silver nanoparticles, *Nanotechnology* 16, 2346–2353 (2005).
114. IOFFE, Basic properties of GaAs at 300 K, Ioffe Physico-Technical Insitute (available at http://www.ioffe.rssi.ru/SVA/NSM/Semicond/GaAs/basic.html, accessed on October 30, 2015).

115. R. Dingle, A. C. Gossard, W. Wiegmann, Direct observation of superlattice formation in a semiconductor heterostructure, *Phys. Rev. Lett.* 34, 1327–1330 (1975).

116. N. Holonyak Jr., R. M. Kolbas, R. D. Dupuis, P. D. Dapkus, Quantum-well heterostructure lasers, *IEEE J. Quantum Electron.* 16, 170–186 (1980).

117. J. P. van der Ziel, R. Dingle, R. C. Miller, W. Wiegmann, W. A. Nordland Jr., Laser oscillation from quantum states in very thin GaAs-$Al_{0.2}Ga_{0.8}As$ multilayer structures, *Appl. Phys. Lett.* 26, 463–465 (1975).

118. D. J. Paul, The progress towards terahertz quantum cascade lasers on silicon substrates, *Laser Photon. Rev.* 4, 610–632 (2010).

119. H. W. Li *et al.*, Quantum dot resonant tunneling diode for telecommunication wavelength single photon detection, *Appl. Phys. Lett.* 91, 073516 (2007).

120. M. Asada, Y. Miyamoto, Y. Suematsu, Gain and the threshold of three-dimensional quantum-box lasers, *IEEE J. Quantum Electron.* 22, 1915–1921 (1986).

121. Y. Arakawa, A. Yariv, Quantum well lasers—gain, spectra, dynamics, *IEEE J. Quantum Electron.* 22, 1887–1899 (1986).

122. Y. Arakawa, H. Sakaki, Multidimensional quantum well laser and temperature dependence of its threshold current, *Appl. Phys. Lett.* 40, 939 (1982).

123. T. Fujii, S. Yamakoshi, K. Nanbu, O. Wada, S. Hiyamizu, Very low threshold current GaAs-AIGaAs GRIN-SCH lasers grown by MBE for OEIC applications, *J. Vac. Sci. Technol. B* 2, 259–261 (1984).

124. W. T. Tsang, Extremely low threshold (AlGa) As modified multiquantum well heterostructure lasers grown by molecular-beam epitaxy, *Appl. Phys. Lett.* 39, 786–788 (1981).

125. D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications: An Applied Approach*, Springer, New York, 1st ed., 2013.

126. R. Chin *et al.*, Temperature dependence of threshold current for quantum-well $Al_xGa_{1-x}As$-GaAs heterostructure laser diodes, *Appl. Phys. Lett.* 36, 19–21 (1980).

127. M. Ettenberg, C. J. Nuese, H. Kressel, The temperature dependence of threshold current for double-heterojunction lasers, *J. Appl. Phys.* 50, 2949–2950 (1979).

128. J. Faist *et al.*, Quantum cascade laser, *Science* 264, 553–555 (1994).

129. C. Gmachl, F. Capasso, D. L. Sivco, A. Y. Cho, Recent progress in quantum cascade lasers and applications, *Reports Prog. Phys.* 64, 1533–1601 (2001).

130. R. Köhler *et al.*, Terahertz semiconductor-heterostructure laser, *Nature* 417, 156–159 (2002).

131. B. S. Williams, Terahertz quantum-cascade lasers, *Nat. Photonics* 1, 517–525 (2007).

132. K. V. Klitzing, G. Dorda, M. Pepper, New method for high-accuracy determination of the fine-structure constant based on quantized Hall resistance, *Phys. Rev. Lett.* 45, 494–497 (1980).

133. D. C. Tsui, H. L. Stormer, A. C. Gossard, Two-dimensional magnetotransport in the extreme quantum limit, *Phys. Rev. Lett.* 48, 1559–1562 (1982).

134. P. J. Simmonds, H. E. Beere, D. A. Ritchie, S. N. Holmes, Growth-temperature optimization for low-carrier-density In0.75Ga0.25As-based high electron mobility transistors on InP, *J. Appl. Phys.* 102, 083518 (2007).

135. R. E. Prange, S. M. Girvin, Eds., *The Quantum Hall Effect*, Springer-Verlag, New York, 2nd ed., 1989.

136. H. L. Stormer, D. C. Tsui, The quantized hall effect, *Science* 220, 1241–1246 (1983).

137. P. J. Simmonds, Molecular beam epitaxy of InGaAs and InAlAs for low-dimensional electrical transport, Ph.D. thesis, University of Cambridge, 2007.

138. R. B. Laughlin, Anomalous quantum hall effect: an incompressible quantum fluid with fractionally charged excitations, *Phys. Rev. Lett.* 50, 1395–1398 (1983).

139. F. D. M. Haldane, Fractional quantization of the hall effect: a hierarchy of incompressible quantum fluid states, *Phys. Rev. Lett.* 51, 605–608 (1983).

140. J. K. Jain, Composite-Fermion approach for the fractional quantum hall effect, *Phys. Rev. Lett.* 63, 199–202 (1989).

141. J. Martin *et al.*, Localization of fractionally charged quasi-particles, *Science* 305, 980–984 (2004).

142. K. S. Novoselov *et al.*, Electric field effect in atomically thin carbon films, *Science* 306, 666 (2004).

143. K. S. Novoselov *et al.*, Two-dimensional gas of massless Dirac fermions in graphene, *Nature* 438, 197–200 (2005).

144. S. V. Morozov *et al.*, Giant intrinsic carrier mobilities in graphene and its bilayer, *Phys. Rev. Lett.* 100, 016602 (2008).

145. K. S. Novoselov *et al.*, Room-temperature quantum hall effect in graphene, *Science* 315, 1379 (2007).

146. G. Landwehr *et al.*, Quantum transport in n-type and p-type modulation-doped mercury telluride quantum wells, *Phys. E* 6, 713–717 (2000).

147. K. S. Novoselov *et al.*, Two-dimensional atomic crystals, *Proc. Natl. Acad. Sci.* 102, 10451–10453 (2005).

148. L. Song *et al.*, Large scale growth and characterization of atomic hexagonal boron nitride layers, *Nano Lett.* 10, 3209–3215 (2010).

149. S. Z. Butler *et al.*, Progress, challenges, and opportunities in two-dimensional materials beyond graphene, *ACS Nano* 7, 2898–2926 (2013).

150. P. Vogt *et al.*, Silicene: compelling experimental evidence for graphene like two-dimensional silicon, *Phys. Rev. Lett.* 108, 155501 (2012).

151. M. E. Dávila, L. Xian, S. Cahangirov, A. Rubio, G. Le Lay, Germanene: a novel two-dimensional germanium allotrope akin to graphene and silicene, *New J. Phys.* 16, 095002 (2014).

152. A. K. Geim, I. V. Grigorieva, Van der Waals heterostructures, *Nature* 499, 419–425 (2013).

153. Z. Ni *et al.*, Tunable bandgap in silicene and germanene, *Nano Lett.* 12, 113–118 (2012).

154. T. J. Thornton, M. Pepper, H. Ahmed, D. Andrews, G. J. Davies, One-dimensional conduction in the 2D electron gas of a GaAs-AlGaAs heterojunction, *Phys. Rev. Lett.* 56, 1198–1201 (1986).

155. D. A. Wharam *et al.*, One-dimensional transport and the quantisation of the ballistic resistance, *J. Phys. C Solid State Phys.* 21, L209–L214 (1988).

156. B. J. van Wees *et al.*, Quantized conductance of point contacts in a two-dimensional electron gas, *Phys. Rev. Lett.* 60, 848–850 (1988).

157. P. J. Simmonds *et al.*, Molecular beam epitaxy of high mobility In0.75Ga0.25As for electron spin transport applications, *J. Vac. Sci. Technol. B* 27, 2066–2070 (2009).

158. P. J. Simmonds *et al.*, Quantum transport in In0.75Ga0.25As quantum wires, *Appl. Phys. Lett.* 92, 152108 (2008).

159. N. K. Patel *et al.*, Properties of a ballistic quasi-one-dimensional constriction in a parallel high magnetic field, *Phys. Rev. B* 44, 10973–10975 (1991).

160. K. Thomas *et al.*, Possible spin polarization in a one-dimensional electron gas, *Phys. Rev. Lett.* 77, 135–138 (1996).

161. N. K. Patel *et al.*, Evolution of half plateaus as a function of electric field in a ballistic quasi-one-dimensional constriction, *Phys. Rev. B* 44, 13549–13555 (1991).

162. A. Kristensen *et al.*, Bias and temperature dependence of the 0.7 conductance anomaly in quantum point contacts, *Phys. Rev. B* 62, 10950–10957 (2000).

163. D. J. Reilly, Y. Zhang, L. DiCarlo, Phenomenology of the 0.7 conductance feature, *Phys. E* 34, 27–30 (2006).

164. A. C. Graham, M. Pepper, M. Y. Simmons, D. A. Ritchie, New interaction effects in quantum point contacts at high magnetic fields, *Phys. E* 34, 588–591 (2006).

165. S. M. Cronenwett *et al.*, Low-temperature fate of the 0.7 structure in a point contact: a kondo-like correlated state in an open system, *Phys. Rev. Lett.* 88, 226805 (2002).

166. S. Iijima, Helical microtubules of graphitic carbon, *Nature* 354, 56–58 (1991).

167. M. S. Dresselhaus, G. Dresselhaus, P. C. Eklund, A. M. Rao, Carbon nanotubes, in *The Physics of Fullerene-Based and Fullerene-Related Materials*, W. Andreoni, Ed. Springer, Dordrecht, the Netherlands, 2000, pp. 331–379.

168. C. Dekker, Carbon nanotubes as molecular quantum wires, *Phys. Today* May, 22–28 (1999).

169. S. J. Tans *et al.*, Individual single-wall carbon nanotubes as quantum wires, *Nature* 386, 474–478 (1997).

170. S. Frank, P. Poncharal, Z. L. Wang, W. A. de Heer, Carbon nanotube quantum resistors, *Science* 280, 1744–1746 (1998).

171. S. Iijima, Carbon nanotubes: past, present, and future, *Phys. B* 323, 1–5 (2002).

172. X. Wang *et al.*, Fabrication of ultralong and electrically uniform single-walled carbon nanotubes on clean substrates, *Nano Lett.* 9, 3137–3141 (2009).

173. X. Shi *et al.*, Fabrication of porous ultra-short single-walled carbon nanotube nanocomposite scaffolds for bone tissue engineering, *Biomaterials* 28, 4078–4090 (2007).

174. Y. Jung, X. Li, N. K. Rajan, A. D. Taylor, M. A. Reed, Record high efficiency single-walled carbon nanotube/silicon $p-n$ junction solar cells, *Nano Lett.* 13, 95–99 (2013).

175. I. van Weperen, S. R. Plissard, E. P. A. M. Bakkers, S. M. Frolov, L. P. Kouwenhoven, Quantized conductance in an InSb nanowire, *Nano Lett.* 13, 387–391 (2013).

176. V. Mourik *et al.*, Signatures of majorana fermions in hybrid superconductor-semiconductor nanowire devices, *Science* 336, 1003–1007 (2011).

177. A. I. Yanson, G. Rubio Bollinger, H. E. van den Brom, N. Agraït, J. M. van Ruitenbeek, Formation and manipulation of a metallic wire of single gold atoms, *Nature* 395, 783–785 (1998).

178. P. J. Simmonds *et al.*, Growth by molecular beam epitaxy of self-assembled InAs quantum dots on InAlAs and InGaAs lattice-matched to InP, *J. Vac. Sci. Technol. B* 25, 1044–1048 (2007).

179. P. J. Simmonds *et al.*, Structural and optical properties of InAs/AlAsSb quantum dots with GaAs(Sb) cladding layers, *Appl. Phys. Lett.* 100, 243108 (2012).

180. O. Bierwagen, W. T. Masselink, Self-organized growth of InAs quantum wires and dots on InP(001): the role of vicinal substrates, *Appl. Phys. Lett.* 86, 113110 (2005).

181. Y. I. Mazur *et al.*, Spectroscopy of shallow InAs/InP quantum wire nanostructures, *Nanotechnology* 20, 065401 (2009).

182. F. Lelarge *et al.*, Recent advances on InAs/InP quantum dash based semiconductor lasers and optical amplifiers operating at 1.55 μm, *IEEE J. Sel. Top. Quantum Electron.* 13, 111–124 (2007).

183. R. H. Wang *et al.*, Room-temperature operation of InAs quantum-dash lasers on InP (001), *IEEE Photonics Technol. Lett.* 13, 767–769 (2001).

184. C. D. Yerino *et al.*, Tensile GaAs(111) quantum dashes with tunable luminescence below the bulk bandgap, *Appl. Phys. Lett.* 105, 071912 (2014).

185. P. J. Simmonds, M. L. Lee, Tensile strained island growth at step-edges on GaAs(110), *Appl. Phys. Lett.* 97, 153101 (2010).

186. P. J. Simmonds, M. L. Lee, Self-assembly on (111)-oriented III-V surfaces, *Appl. Phys. Lett.* 99, 123111 (2011).

187. P. J. Simmonds, M. L. Lee, Tensile-strained growth on low-index GaAs, *J. Appl. Phys.* 112, 054313 (2012).

188. A. P. Alivisatos, Semiconductor clusters, nanocrystals, and quantum dots, *Science* 271, 933–937 (1996).

189. M. V. Maximov, N. N. Ledentsov, Quantum dot lasers, in *Dekker Encyclopedia of Nanoscience and Nanotechnology*, Vol. 4, J. A. Schwarz, C. I. Contescu, K. Putyera, Eds. Marcel Dekker. Inc., New York, 2004, pp. 3109–3126.

190. A. J. Shields, Semiconductor quantum light sources, *Nat. Photonics* 1, 215–223 (2007).

191. A. Shields, Quantum logic with light, glass, and mirrors, *Science* 297, 1821 (2002).

192. D. Kim, S. G. Carter, A. Greilich, A. S. Bracker, D. Gammon, Ultrafast optical control of entanglement between two quantum-dot spins, *Nat. Phys.* 7, 223–229 (2011).

193. M. A. Reed *et al.*, Observation of discrete electronic state in a zero-dimensional semiconductor nanostructure, *Phys. Rev. Lett.* 60, 535–537 (1988).

194. T. P. Smith III, K. Y. Lee, C. M. Knoedler, J. M. Hong, D. P. Kern, Electronic spectroscopy of zero-dimensional systems, *Phys. Rev. B* 38, 2172–2176 (1988).

195. J. R. Prance *et al.*, Electronic refrigeration of a two-dimensional electron gas, *Phys. Rev. Lett.* 102, 146602 (2009).

196. F. R. Waugh *et al.*, Single-electron charging in double and triple quantum dots with tunable coupling, *Phys. Rev. Lett.* 75, 705–708 (1995).

197. L. P. Kouwenhoven *et al.*, Transport through a finite one-dimensional crystal, *Phys. Rev. Lett.* 65, 361–364 (1990).

198. D. Loss, D. P. DiVincenzo, Quantum computation with quantum dots, *Phys. Rev. A* 57, 120 (1998).

199. J. R. Petta *et al.*, Coherent manipulation of coupled electron spins in semiconductor quantum dots, *Science* 309, 2180–2184 (2005).

200. J. M. Pietryga *et al.*, Utilizing the lability of lead selenide to produce heterostructured nanocrystals with bright, stable infrared emission, *J. Am. Chem. Soc.* 130, 4879–4885 (2008).

201. C. B. Murray, D. J. Norris, M. G. Bawendi, Synthesis and characterization of nearly monodisperse CdE (E = S, Se, Te) semiconductor nanocrystallites, *J. Am. Chem. Soc.* 115, 8706–8715 (1993).

202. J. J. Li *et al.*, Large-scale synthesis of nearly monodisperse CdSe/CdS core/shell nanocrystals using air-stable reagents via successive ion layer adsorption and reaction, *J. Am. Chem. Soc.* 125, 12567–12575 (2003).

203. J. Kwak *et al.*, High-power genuine ultraviolet light-emitting diodes based On colloidal nanocrystal quantum dots, *Nano Lett.* 15, 3793–3799 (2015).

204. B. O. Dabbousi *et al.*, (CdSe)ZnS core-shell quantum dots: synthesis and characterization of a size series of highly luminescent nanocrystallites, *J. Phys. Chem. B* 101, 9463–9475 (1997).

205. C. B. Murray, C. R. Kagan, M. G. Bawendi, Self-organization of CdSe nanocrystallites into three-dimensional quantum dot superlattices, *Science* 270, 1335–1338 (1995).

206. J. Zhao *et al.*, Efficient CdSe/CdS quantum dot light-emitting diodes using a thermally polymerized hole transport layer, *Nano Lett.* 6, 463–467 (2006).

207. D. V. Talapin, J.-S. Lee, M. V. Kovalenko, E. V. Shevchenko, Prospects of colloidal nanocrystals for electronic and optoelectronic applications, *Chem. Rev.* 110, 389–458 (2010).

208. D. Leonard, M. Krishnamurthy, C. M. Reaves, S. P. Denbaars, P. M. Petroff, Direct formation of quantum-sized dots from uniform coherent islands of InGaAs on GaAs surfaces, *Appl. Phys. Lett.* 63, 3203–3205 (1993).

209. D. J. Eaglesham, M. Cerullo, Dislocation-free Stranski-Krastanow growth of Ge on Si(100), *Phys. Rev. Lett.* 64, 1943–1946 (1990).

210. D. L. Huffaker, G. Park, Z. Zou, O. B. Shchekin, D. G. Deppe, 1.3 μm room-temperature GaAs-based quantum-dot laser, *Appl. Phys. Lett.* 73, 2564 (1998).

211. C. G. Bailey, D. V. Forbes, R. P. Raffaelle, S. M. Hubbard, Near 1 V open circuit voltage InAs/GaAs quantum dot solar cells, *Appl. Phys. Lett.* 98, 163105 (2011).

212. Y. Song, P. J. Simmonds, M. L. Lee, Self-assembled In0.5Ga0.5As quantum dots on GaP, *Appl. Phys. Lett.* 97, 223110 (2010).

213. K. Jacobi, Atomic structure of InAs quantum dots on GaAs, *Prog. Surf. Sci.* 71, 185–215 (2003).

214. B. Damilano, N. Grandjean, F. Semond, J. Massies, M. Leroux, From visible to white light emission by GaN quantum dots on Si(111) substrate, *Appl. Phys. Lett.* 75, 962 (1999).

215. D. Bimberg *et al.*, InGaAs-GaAs quantum-dot lasers, *IEEE J. Sel. Top. Quantum Electron.* 3, 196–205 (1997).

216. P. Atkinson, O. G. Schmidt, S. P. Bremner, D. A. Ritchie, Formation and ordering of epitaxial quantum dots, *Comptes Rendus Phys.* 9, 788–803 (2008).

217. Y. Nakamura *et al.*, Regular array of InGaAs quantum dots with 100-nm-periodicity formed on patterned GaAs substrates, *Phys. E* 21, 551–554 (2004).

218. S. Kiravittaya, H. Heidemeyer, O. Schmidt, Growth of three-dimensional quantum dot crystals on patterned GaAs (001) substrates, *Phys. E* 23, 253–259 (2004).

219. K. Rivoire, S. Buckley, Y. Song, M. L. Lee, J. Vučković, Photoluminescence from In0.5Ga0.5As/GaP quantum dots coupled to photonic crystal cavities, *Phys. Rev. B* 85, 045319 (2012).

220. P. J. Simmonds *et al.*, Tuning quantum dot luminescence below the bulk band gap using tensile strain, *ACS Nano* 7, 5017–5023 (2013).

221. C. D. Yerino *et al.*, Strain-driven growth of GaAs(111) quantum dots with low fine structure splitting, *Appl. Phys. Lett.* 105, 251901 (2014).

222. E. Stock *et al.*, Single-photon emission from InGaAs quantum dots grown on (111) GaAs, *Appl. Phys. Lett.* 96, 093112 (2010).

223. T. Kondo, K. Saitoh, Y. Yamamoto, T. Maruyama, S. Naritsuka, Fabrication of GaN dot structures on Si substrates by droplet epitaxy, *Phys. Status Solidi A* 203, 1700–1703 (2006).

224. B.-L. Liang *et al.*, Energy transfer within ultralow density twin InAs quantum dots grown by droplet epitaxy, *ACS Nano* 2, 2219–2224 (2008).

73

CHEMICAL SENSING

W. RUDOLF SEITZ
Department of Chemistry, University of New Hampshire, Durham, NH, USA

73.1 INTRODUCTION

Chemical sensing involves measuring the concentration of an analyte in a sample. It has the following two essential features: (i) the measurement is made in the sample or at the site of the sample rather than in a separate laboratory and (ii) the concentration of interest is measured either continuously or repeatedly such that changes in concentration with time can be followed. This chapter will review analytical measurement technologies that meet these requirements with emphasis on those methods that are not covered elsewhere in this book.

Most chemical sensors are used in one of two contexts. They can be part of a system for some type of process control. For example, a pH electrode can be used to sense the pH of a fermentation process. If the pH starts to deviate from the desired value, it sends a signal that causes acid or base to be dispensed into the fermentation to restore the pH to the desired value. Another example involves glucose measurements by diabetics, which are used to determine how much insulin should be dispensed. The ideal system for this would involve an implantable glucose sensor coupled to an insulin pump with the appropriate control algorithm. In practice, it hasn't yet proved feasible to develop a glucose sensor that functions long enough to warrant implantation. Instead, glucose is measured intermittently and the resulting value is used to establish the need for insulin.

The other context where sensors are frequently used involves alarm monitoring. The sensor is designed to produce a signal if a concentration of some analyte becomes high enough to cause a problem. Carbon monoxide sensors in the home fall into this

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.

category. Environmental sensing is often undertaken to make sure concentrations don't exceed a predetermined value where the analyte may prove problematic.

The ideal sensor would be perfectly selective for the analyte of interest and would respond continuously without maintenance. In practice, these goals are difficult to realize. Selectivity is often difficult to achieve or to combine with continuous response. Biosensors, described in another chapter, have been developed because biological reagents like enzymes and antibodies offer excellent selectivity. However, few biosensors offer continuous response and those that do have a limited lifetime because biological reagents are not indefinitely stable. The development of stable receptors that offer selectivity approaching that of biological reagents with improved stability is an active area of research that is likely to lead to improved sensors in the future.

Sensor arrays offer an alternative approach to selectivity. The idea is to use multiple sensors that have different selectivity patterns. While none is completely selective for a particular analyte, pattern recognition techniques can be applied to the responses from the array to distinguish different analytes. Gas sensor arrays have become known as "electronic noses" and have proven themselves to be well suited for certain types of problems. Sensor arrays designed for measurements in solution are known as "electronic tongues." These remain a subject of research and have yet to have the same practical impact as electronic noses.

Longevity and calibration are issues for any sensor. Reagent stability can limit sensor lifetime. For example, this is an issue with continuous biosensors, for example, the glucose electrode based on the enzyme glucose oxidase. Another problem involves sensor fouling, that is, accumulation of unwanted material on the surface of a sensor that affects its response. When applying a chemical sensor to a new type of sample, a series of control experiments needs to be undertaken to establish the length of time that a sensor will provide accurate response. Based on these experiments, a maintenance protocol that involves cleaning and recalibration can be established to assure that the sensor is responding with the desired accuracy at all times. Most physical sensors need little if anything in the way of maintenance. However, this is not the case for many chemical sensing technologies.

Ideally chemical sensors respond continuously to changes in analyte concentration. Measurement technologies such as in situ Raman spectroscopy that directly measure analyte concentration intrinsically provide continuous response. However, most chemical sensors involve an interaction between that analyte and the sensor that modifies the properties of a sensing element. If this interaction is reversible, then response is continuous. Other "continuous" sensors involve steady-state measurements that involve analyte consumption at such a slow rate that it does significantly change the analyte concentration.

As we go through various sensor technologies, we will consider the issues that affect that particular technology. Other sources of information will be referenced. The goal will be to provide a sense of the state of the art for each sensing technology plus an indication of the main thrusts of current research and how they may be expected lead to improved sensor technology in the future.

The field of chemical sensing encompasses a variety of analytical technologies, many of them quite different. Electrical engineers, chemists, and physicists are all active in chemical sensor research. There is an excellent book on chemical sensing that considers general sensing issues more in depth than this chapter and covers many of the same measurement technologies more in depth than this chapter [1]. There is also a recent comprehensive book that covers both chemical and biological sensors [2]. However, for the most part, it's easier to find more detailed information on a particular type of sensor than it is to find general coverage of all chemical sensing approaches.

The chemical sensing technologies that we will consider in this chapter are categorized as electrical, optical, and mass, depending on the type of signal that is measured. Several of the sensing technologies covered can be coupled to biological recognition agents to make biosensors.

73.2 ELECTRICAL METHODS

73.2.1 Potentiometry

73.2.1.1 Solution Phase Measurements Potentiometry is an established technique for measuring the activity of ions in solution. The measurement of pH with a glass electrode is by far the most widely applied and best known example of potentiometry. However, electrodes are available for sensing the activities of a wide variety of other cations and anions.

The basic arrangement for potentiometry is shown in Figure 73.1. What is actually measured is the difference in potential between two electrodes. The reference electrode is designed so that its potential is independent of solution composition. The saturated calomel electrode is most commonly used as a reference. The other electrode is known as the indicator electrode. Its potential varies with the activity of the ion of interest.

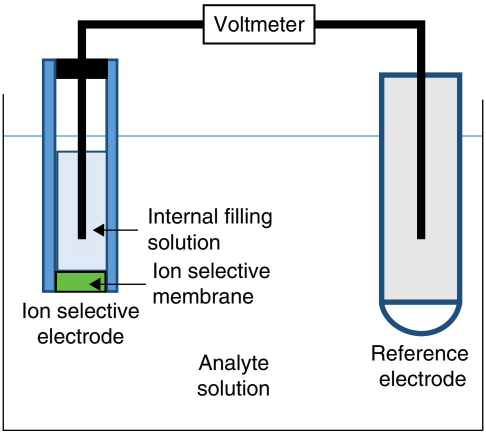


FIGURE 73.1 Instrumental arrangement for potentiometry.

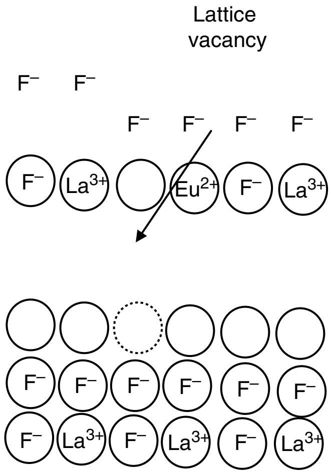


FIGURE 73.2 Lanthanum fluoride membrane used in ion selective electrode that responds selectively to fluoride. The membrane is doped with a +2 ion Eu(II) creating lattice vacancies because Eu(II) only needs to be surrounded by 2 rather than 3 fluoride ions. Fluoride ions can move into the vacancy and thus can be transported through the membrane. Transport is selective because other ions don't fit into the lanthanum fluoride crystal lattice.

In the case of pH measurements, the two electrodes are often combined onto a single housing known as a combination electrode.

Most indicator electrodes are membrane electrodes. The potential that is measured arises because there is a difference in the concentration of the ion of interest on either side of a membrane. To make an electrode that is selective for a particular ion, you need a membrane that is selectively permeable to that ion. We illustrate this with the ion selective electrode for fluoride, the second most widely applied ion selective electrode after the glass pH electrode. The membrane for this electrode, lanthanum fluoride doped with Eu(II), is shown in Figure 73.2. The presence of a +2 ion in LaF_3 leaves a vacancy in the crystal lattice. This results in a solid-state material that is permeable to fluoride because fluoride can hop from site to site. Selective fluoride transport arises because other anions are too large to conveniently fit into the lanthanum fluoride lattice. When the fluoride ion concentration in the internal filling solution of the ion selective electrode is higher than the fluoride concentration in the external solution, there is net transport of fluoride from the internal filling solution. The transport rate is so slow that this does not produce a large change in fluoride concentration in the external solution. However, it does result in an electrode potential that depends on the ratio of fluoride concentrations on either side of the lanthanum fluoride membrane.

Lanthanum fluoride is an example of a solid-state membrane. Another common type of ion selective membrane involves a compound known as an ionophore that selectively binds a particular ion dissolved in plasticized polyvinyl chloride.

The equation for ion selective electrode response is shown as follows:

$$E = k + \beta \frac{0.0592}{z} \log\left(\left[A\right] + K_{AI}\left[I\right]\right)$$

E is the measured electrochemical cell potential in volts. "k" and "β" are constants that must be determined by calibration. "z" is the charge on the ion that is being measured. For example, if a sulfide (S^{2-}) ion selective electrode is being used, the value of "z" is –2. [A] is the concentration of the analyte ion. [I] is the concentration of an interfering ion. K_{AI} is the selectivity coefficient for that particular interfering ion. The smaller the value of the selectivity coefficient, the more selective the electrode for analyte A compared to interference I. When the product K_{AI} [I] is significant compared to [A], the interfering ion starts to contribute to the measured potential.

The two constants, k and β, are determined by measuring electrode potential at two or more analyte concentrations in the absence of interfering ions. The theoretical value of β is 1.00 at 25°C. In practice it is usually slightly smaller. The value of β is determined from the slope of the response. Once β is known, the value of k can be calculated.

When you acquire an ion selective electrode, it will come with a series of selectivity coefficients. These can be used to determine whether a particular ion can be sensed without interference in a given context.

One important characteristic of ion selective electrodes is that they measure ion activities rather than total concentrations. Activity is a physical chemistry concept that may not be familiar to all readers. Analyte activity equals the analyte concentration times an activity coefficient. The value of the activity coefficient is 1.00 at infinite dilution but decreases as the ionic strength of the solution increases. To measure concentration, the usual practice is to adjust the ionic strength of the standards used to calibrate the electrode so that it equals the ionic strength of the sample. In this case, the activity coefficient is the same in both standards and sample, and the measured parameter is the concentration.

Another way of viewing this is that ion selective electrodes only measure the concentration of a particular ion. Thus, a Cu(II) electrode only measures the concentration of Cu(II) in solution. It does not respond to total Cu. Cu in the form of a complex ion, for example, $Cu(NH_3)_4^{2+}$, will not be sensed by an electrode. The glass pH electrode only measures the concentration of H^+ ion. It does not respond to total acid.

The instrumentation for potentiometry is nothing more than a voltmeter with a high input impedance. The ubiquitous pH meter is nothing more than such a voltmeter designed to directly read out in pH if it is appropriately calibrated using standard pH buffers.

Electrodes are available that respond selectively to a large number of ions including F^-, Br^-, Cl^-, I^-, CN^-, SCN^-, Ag^+, Cu^{2+}, Cd^{2+}, Pb^{2+}, Na^+, Li^+, K^+, NH_4^+, and S^{2-}.

Potentiometry is also widely used to measure the concentrations of acidic or basic dissolved gases, particularly carbon dioxide and ammonia. To make a carbon dioxide-sensitive electrode, a pH electrode is covered by a CO_2 permeable membrane. A thin layer of a bicarbonate solution is contained between the membrane and the pH sensor. The dissolved CO_2 diffuses through the gas permeable membrane into the bicarbonate solution. The CO_2 reacts with water-forming carbonic acid, H_2CO_3, which changes the pH of the solution. The measured potential due to pH changes in a systematic known manner with changing carbon dioxide partial pressures. Ammonia is sensed in a similar manner using an ammonia permeable membrane and a thin layer of an ammonium ion solution over a pH electrode.

Potentiometry is a mature technology. While work continues on new membrane formulations that may offer improved selectivity for certain applications, only incremental improvements can be expected. More in-depth information on potentiometry and its applications may be found in common analytical chemistry texts such as *Quantitative Chemical Analysis* by Daniel Harris.

73.2.1.2 Gas Phase Measurements There are also potentiometric gas sensors. The basic configuration is exemplified by the widely applied oxygen sensor based on zirconia shown in Figure 73.3. As with potentiometric ion selective electrodes, the measured parameter is a potential across a membrane resulting from a concentration gradient. One side contains a known reference concentration of the analyte, and the other side contains the sample of interest. In the zirconia sensor the reference side is air, which contains a constant percentage of oxygen.

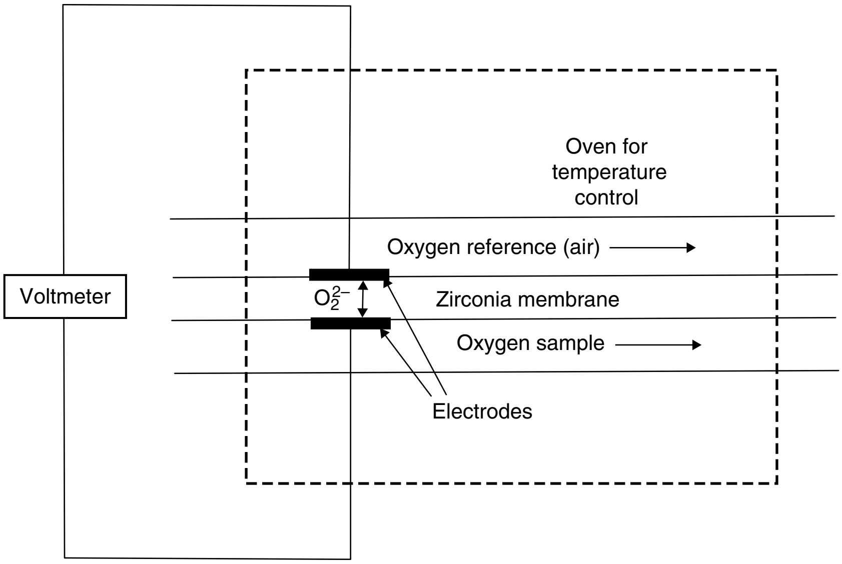


FIGURE 73.3 Schematic of high-temperature zirconia oxygen sensor. The reference gas is normally air. Oxygen is transported across the membrane as O^{2-}. The potential across the membrane is measured directly using electrodes applied to either side of the membrane.

The potential across the zirconia membrane is measured directly using two platinum electrodes applied directly to the membrane surface. While potentials develop at the interfaces between the platinum and the zirconia, the values of these two potentials are equal in magnitude and opposite in sign, so they cancel out leaving only the potential due to the difference in analyte concentration on either side of the metal oxide. This potential is due to selective oxygen transport across the zirconia. The oxygen reacts with zirconia to form O^{2-} ions that can hop from site to site. The transport mechanism is analogous to that observed with lanthanum fluoride membranes used for aqueous phase fluoride sensing. As with the other type of membrane electrode, the potential across the electrode varies with the logarithm of the ratio of oxygen concentrations on either side of the membrane.

The zirconia-based oxygen sensor is widely used to sense the oxygen concentrations in fuel–air mixtures in internal combustion engines. They are incorporated into feedback control systems to maintain oxygen concentrations at the level required for efficient combustion reactions. Current research on this type of oxygen sensor has been recently reviewed [3].

This approach is applicable to other oxidizing gases such as chlorine.

73.2.2 Voltammetry

Voltammetry refers to methods where the potential of a working electrode is controlled relative to a reference electrode, and the resulting current is measured. This current is due to oxidation or reduction of a solute. The magnitude of the current is related to solute concentration. Voltammetry can be implemented directly in aqueous samples that and, therefore, can be used for continuous sensing. In practice, such applications are few. The one requirement is that the aqueous solution contains enough dissolved salt to be electrically conducting.

In practice voltammetry is not widely used for continuous chemical sensing. One problem is that most of the complex samples that we deal with in practical contexts contain components that deposit on the electrode surface. This blocks the electrode surface, reducing its effective area. Since current is proportional to electrode area, this causes an interference. The result is the voltammetry requires frequent electrode maintenance to the point where it is usually deemed impractical. A second problem is that it is often difficult to find a potential at which the analyte of interest can be oxidized or reduced without also oxidizing or reducing other solutes as well.

There are two strategies that have led to successful voltammetric sensing. One is to determine gases that can be oxidized or reduced using a membrane to protect the electrodes. By far the most important application involves the measurement of dissolved oxygen. The voltammetric oxygen electrode is shown in Figure 73.4. The voltammetric electrode is separated from the analytical sample by a membrane that is permeable to gases but not to solution. This eliminates interferences from all nonvolatile solutes. The most common membrane is Teflon, a substance that does not adhere strongly to

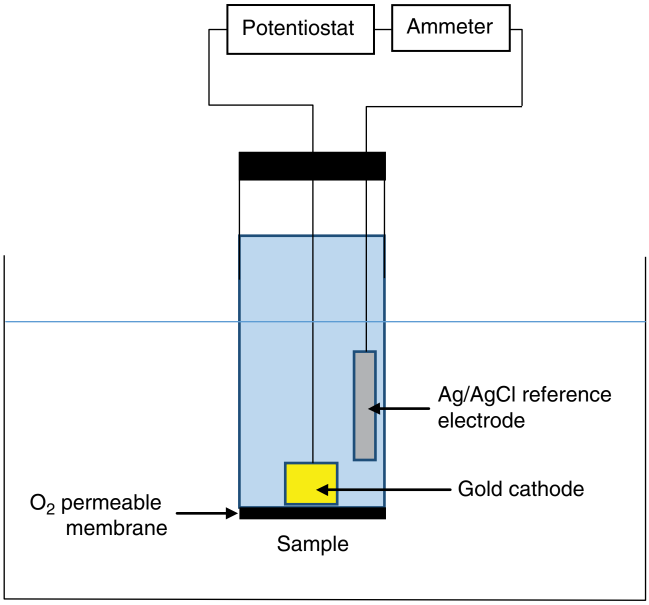


FIGURE 73.4 Schematic of voltammetric oxygen electrode. The measured parameter is the current due to the oxygen reduction at the gold cathode. The potential of this electrode is controlled by the potentiostat at a value sufficiently cathodic relative to the silver/silver chloride reference electrode to reduce oxygen.

anything. As a result, the dissolved oxygen electrode is not subject to fouling by solutes in the sample of interest. The dissolved oxygen electrode is sometimes known as a Clark electrode, after its inventor, Leland Clark.

There is a thin layer of a salt solution between the membrane and the electrode. This is required so that the potential of the working electrode can be controlled relative to the potential of a reference electrode. This voltage is held at a sufficiently negative value to reduce dissolved oxygen to water, $O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$. The resulting current depends on the concentration of dissolved oxygen.

The oxygen electrode is operated in a steady-state mode. The electrode potential is held a sufficiently negative value to reduce all oxygen at the surface of the electrode. This creates a concentration gradient. Oxygen from the sample diffuses across the membrane and the thin solution layer to reach the working electrode surface where it is reduced. If the sample is stirred, the sample will be homogeneous, and the oxygen concentration at the membrane surface will stay essentially constant. While the oxygen that is reduced at the working electrode surface reduces the solution concentration, this amount of oxygen that is reduced is a tiny percentage of the total oxygen amount so that concentration changes are too small to have a significant effect on oxygen concentration.

This approach is applicable to gases that can be oxidized or reduced. Examples include H_2, CO, NO_2, NO, H_2S, and SO_2. More information on voltammetric gas sensing is available from an excellent review article [4].

A second strategy for voltammetric sensing involves disposable electrodes prepared by screen printing. The most important example involves glucose sensing using a strip consisting of disposable electrodes covered by a layer of the immobilized enzyme glucose oxidase that converts glucose to a product that can be determined by voltammetry. This is a biosensor, covered in another chapter of this handbook.

73.2.3 Chemiresistors

The resistance of many substances depends on their environment. Changes in resistance can be used to sense this environment. Figure 73.5 shows two ways of configuring chemiresistors.

The first widely applied chemiresistor was developed by Taguchi for propane sensing. The sensing element is SnO_2. Propane adsorbing on the surface of the tin oxide donates electrons to the tin oxide, reducing its electrical resistance. This particular device is widely applied as an alarm sensor to detect gas leaks.

Metal oxides are one type of chemiresistive material. Response involves adsorption of gases on a surface. The change in resistance per mole of adsorbed gas will depend on the surface to volume ratio of the resistive material. Hence, to maximize sensitivity, it is desirable to use porous materials with high surface to volume ratios. By providing routes to the preparation of highly porous materials, nanotechnology is leading to improvements in the sensitivity of gas sensing chemiresistors.

The mechanism of metal oxide chemiresistor response involves changes in the number of charge carriers in a semiconductor due to adsorption of a gas at the semiconductor surface. Oxidizing gases like oxygen tend to attract electrons. Adsorption of oxygen on the surface of an n-type semiconductor attracts the negatively charged electrons, leading to an increase in electrical resistance. Adsorption of a reducing gas will have the opposite effect, increasing the number of charge carriers. If the resistive element is a p-type semiconductor, the opposite effects will be observed. By attracting

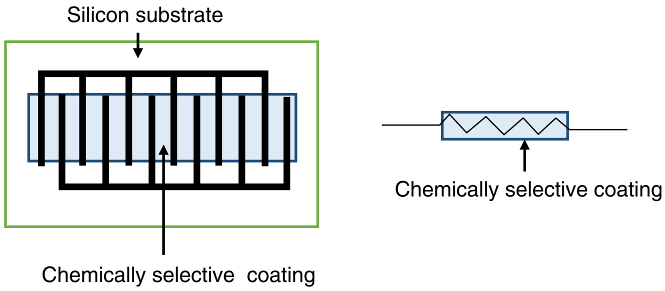


FIGURE 73.5 Two configurations for chemiresistors. In the configuration on the left, the chemically sensitive layer is deposited over interdigitated electrodes on the surface of a silicon semiconductor substrate. On the right, a chemically sensitive coating is coated directly onto a resistor.

electrons, adsorption of an oxidizing gas will increase the number of holes available to carry charge causing a decrease in resistance.

Organic conductors such as polyaniline also function as chemiresistors. In this case, gases partition into the organic material causing a change in resistance. Equilibration times depend on the thickness of the organic material.

Chemiresistors are true sensors in that they involve an equilibrium between the concentration of a gas and number of adsorbed gas molecules or atoms on the surface of the resistive elements. Temperature will affect this equilibrium and needs to be controlled. Selectivity depends on both the affinities of the surface for different gases and the abilities of different gases to affect resistance.

A different way of preparing a chemically sensitive resistor is to deposit small particles of a conducting material such as graphite into a polymer. When volatile organic compounds interact with the polymer, they induce swelling, which increases the distance between conducting particles, causing resistance to increase. In this case, selectivity depends on the coefficient for gas partitioning into the polymer.

Further information on chemiresistors may be obtained from a book on this subject [5]. The properties of metal oxides that are sensitive to gas concentrations are also described in a separate book [6].

The selectivities of chemiresistors are limited. However, their cost is very low. To increase selectivity, arrays of chemiresistors with different response characteristics have been constructed. As discussed later in the chapter, the combination of sensor arrays with the appropriate mathematics has proven to be a useful method for identifying complex mixtures of gases. Such devices are commonly known as electronic noses. A recent review article summarizes the state of the art with respect to chemiresistors in electronic noses [7].

73.2.4 Field Effect Transistors

Field effect transistors can be used to sense either gases or ions in solution. The basic arrangement is shown in Figure 73.6. In a field effect transistor, charge is applied to a gate that is separated from doped semiconductor by a thin layer of silicon dioxide. A positive charge applied to the gate attracts negative charge carriers to the surface of p-type semiconductor material creating a channel where current can flow from the source to the drain. The magnitude of the current flow depends on the size of the charge. If the charge is related to a chemical concentration, then the field effect transistor will function as a chemical sensor.

One class of applications involves ion sensing. The gate is coated with a membrane that selectively binds the ion of interest. This produces an ion selective field effect transistor (ISFET). For stable response, a reference electrode is required. It’s also necessary to encapsulate the device is some sort of polymer that protects all components of the field effect transistor except the membrane-coated gate from exposure to the external aqueous samples.

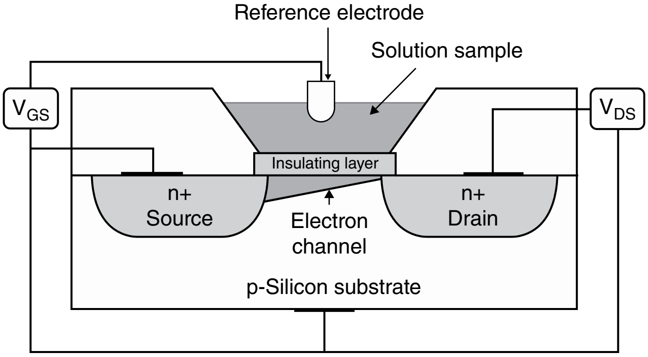


FIGURE 73.6 A solution phase chemically sensitive field effect transistor. The buildup of positive charge on the top surface of the insulating layer attracts electrons to the bottom layer. This affects the magnitude of the current between the source and the drain.

The membranes used for ISFETs involve the same polymers and ionophores that are used for ion selective electrode membranes. ISFETs are alternatives to ion selective electrodes with similar selectivities. Given that ISFETs are amenable to mass production with extremely low cost per device, it is expected that they will eventually replace ion selective electrodes. However, at present, this has not happened, and most potentiometric measurements are carried out with conventional electrodes.

Gas sensing is also possible using field effect transistors [8]. In this case, gas adsorption on the surface of the insulating layer affects the number of charge carriers in the channel that conducts current from the source to the drain.

73.3 OPTICAL METHODS

73.3.1 In situ Optical Measurements

Optical fibers consist of a high refractive index core surrounded by a lower refractive index cladding so that total internal reflection occurs when light strikes the core cladding interface at an angle greater than the critical angle. They provide a convenient method for transporting light to and from a sample. Both plastic and glass fibers are available, transmitting light in the ultraviolet, visible, and near-infrared regions of the electromagnetic spectrum. Both transmission and emission measurements are readily realized. One common configuration is shown in Figure 73.7. A central fiber carries light from a source into a sample. It is combined with six emission fibers that surround the excitation fiber and carry light back to a detection system. If a reflector is placed at the common end of the fiber bundle, this arrangement serves to measure sample absorbance. Without a reflector this arrangement measures scattered or emitted light either at the same wavelength or at a different wavelength.

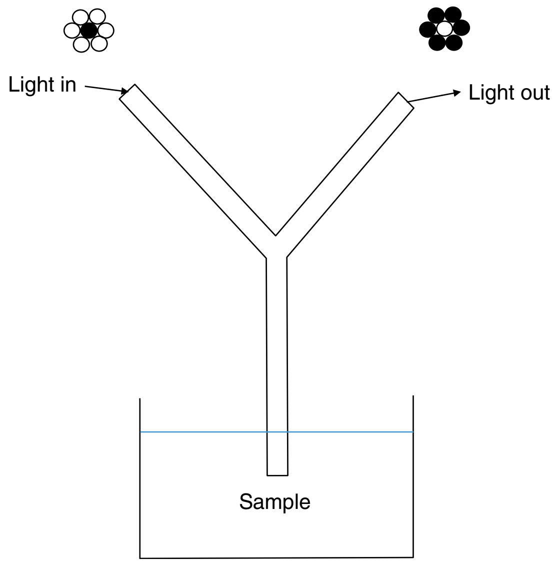


FIGURE 73.7 Common fiber optic arrangement for in situ spectroscopy. A single central fiber conducts light to a sample. A set of six fibers around the central fiber collect light from the sample and transport it back to a detection system.

Another common approach is to remove the cladding from an optical fiber and to directly expose the core to a sample. If the sample refractive index is lower than the fiber core refractive index, total internal reflection occurs at the core/sample interface. However, some of the light enters into the sample where it can be absorbed. This is an alternate method for measuring sample absorbance, essentially a form of internal reflection spectroscopy. If the absorbed light excites fluorescence, this can be captured within the fiber and transmitted to a detection system.

Several types of spectroscopy can be carried out within liquid samples. Absorption measurements in the ultraviolet and visible are readily performed using the fiber optic arrangement of Figure 73.7 with a reflector. Fiber optic attachments are readily available to perform this type of measurement.

Near-infrared spectroscopy is very conveniently performed through optical fibers. This form of spectroscopy is described in another chapter of this handbook and will not be considered further here.

Fluorescence and Raman spectra are also readily measured in situ. Fluorescence is limited to analytes that fluoresce. Raman spectroscopy is more generally applicable to a variety of sensing applications and will be considered further in the next section of this chapter.

A recent review provides more detailed descriptions of the various types of spectroscopy that can be implemented through optical fibers [9].

73.3.2 Raman Spectroscopy

Scattering refers to light that changes direction when it interacts with matter. Rayleigh scattering refers to light that changes direction without changing wavelength. Raman scattering refers to light that changes wavelength as well as changing direction. As shown in Figure 73.8, scattering may be viewed as resulting from an interaction between matter and a photon of light that promotes matter to a higher-energy "virtual state." This state does not involve a specific energy level of the excited matter and is extremely short lived. Normally, the virtual state returns to the original state, reemitting a photon of light with the same energy. However, it is observed that a small percentage of virtual states return to a different vibrational energy level, resulting in a shift of photon energy. Most often the initial state is a ground energy state, and the final state is a higher-energy vibrational state. This is known as Stokes scattering. The resulting Raman emission is at a longer wavelength than the incident wavelength. The difference in photon energies corresponds to the energy of a vibrational transition. Anti-Stokes scattering involves photons that originate in a higher-energy vibrational state and return to a ground state. Because the vast majority of molecules are initially in their ground vibrational state at room temperature, the Stokes lines are much more intense than the anti-Stokes line. Both are much weaker than Rayleigh scattering. The Raman signal is approximately 1 million times weaker than the Rayleigh signal.

The instrumental challenge of Raman spectroscopy is to measure weak emission at wavelengths close to the much stronger emission due to Rayleigh scattering. For many years, this required expensive instrumentation limiting practical applications. Recently, however, advances in optical technology have greatly reduced the cost of Raman spectroscopy. These include semiconductor diode lasers as high-intensity monochromatic sources that produce stronger Raman signals, notch filters for selectively removing the high-intensity Rayleigh line that produces less background, and charge coupled device

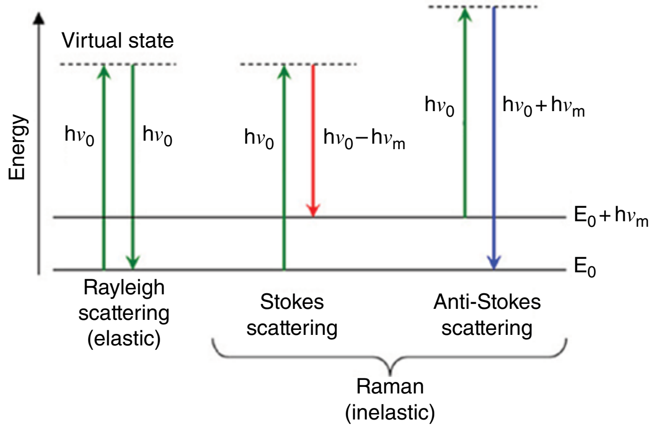


FIGURE 73.8 Processes involved in Raman scattering.

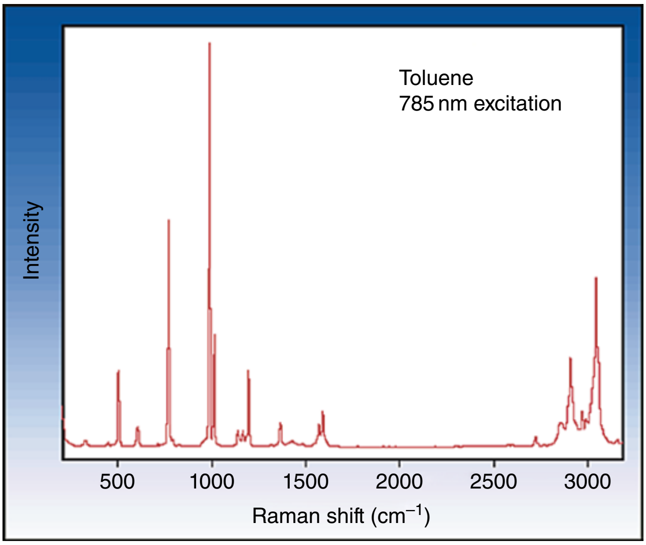


FIGURE 73.9 Raman spectrum of toluene.

detector arrays for simultaneously acquiring the complete Raman spectrum with high sensitivity.

When excited in the ultraviolet or visible region of the electromagnetic spectrum, many samples emit fluorescence. Fluorescence emission occurs at longer wavelengths than the excitation source and will overlap Raman spectra, often completely obscuring it. The probability of fluorescence can be greatly reduced by exciting at a sufficiently long wavelength in the visible or near-infrared region of the electromagnetic spectrum.

Figure 73.9 shows the Raman spectrum of toluene obtained using 785 nm semiconductor laser as the excitation source. The *x*-axis is plotted as the shift in energy compared to the Rayleigh band. This way Raman spectra acquired using excitation at different wavelengths will appear the same. The actual wavelengths depend on the wavelength of the excitation source and the magnitude of the Raman shift.

Raman spectra include bands due to both stretching and bending vibrations. The wavelength of the stretching vibrations depends on the particular type of bond involved in the stretch and may be used to identify functional groups present in the sample. The bending vibrations are sensitive to overall structure. When chemicals react, bonds in the reactants break forming new bonds in the products. Thus, all chemical reactions will give rise to changing spectra and can be followed by Raman spectroscopy. It is important to note that Raman spectra will include not just the sample of interest but also a large signal from the solvent used for a reaction, which may obscure otherwise useful spectral regions.

Raman spectroscopy is already establishing itself as a useful method for process control in the chemical and pharmaceutical industries, a trend that is likely to continue [10].

The importance of Raman spectroscopy for chemical sensing is likely to increase in the near future since several low-cost instruments are now available for this type of application.

73.3.3 Indicator-Based Optical Sensors

Indicators are reagents that change optical properties upon reversibly interacting with an analyte. The best known indicators are those that respond to pH. The reaction is shown in the following:

$$\mathrm{In}_{\text{base form}} + \mathrm{H}^{+} \longleftrightarrow \mathrm{In}_{\text{acid form}}$$

The range of pH that can be sensed by a particular indicator is given by the Henderson–Hasselbalch equation:

$$\mathrm{pH} = \mathrm{p}K_{\mathrm{a}} + \log\frac{\left[\mathrm{In}_{\text{base form}}\right]}{[\mathrm{In}_{\text{acid form}}]}$$

The pK_{a} is the negative log of the acid ionization constant, K_{a}, for the acid form of the indicator. The most common indicators are different colors in the acid and base form. The most familiar application context involves locating titration end points by observing the exact volume required to change indicator color.

Acid–base indicators can also be used to sense pH. The ratio of the concentration of the acid form of the indicator to the base form of the indicator can be determined instrumentally. One way of accomplishing this is to chemically bind the indicator to a solid substrate, which is then confined to the end of a fiber optic bundle similar to that shown in Figure 73.7. The ratio of the reflected intensity at a wavelength absorbed by the acid form of the indicator to the reflected intensity at a wavelength absorbed by the base form of the indicator will change with pH over the range $\mathrm{pH}=\mathrm{p}K_{\mathrm{a}}\pm 1$. Outside this range of pH, the indicator will be essentially completely in either the acid or base form.

For continuous sensing purposes it's more advantageous to use indicators that fluoresce in both the acid and base forms. Because fluorescence measurements are highly sensitive, this makes it possible to work with very low indicator amounts. By relating pH to the ratio of two intensities, it's possible to make pH measurements in environments such as within a cell where calibration of a single intensity measurement would be difficult if not impossible.

An important requirement for sensing with an indicator is that the indicator concentration be low enough so that it does not significantly perturb the concentration of analyte. In the case of an acid–base indicator, this means that the number of protons that come from the indicator itself does not significantly perturb the pH of the solution.

The same concept can be applied to metal ion detection. In this case the indicator is a ligand that changes optical properties upon complexation with a metal ion. For example, there are commercially available Zn(II) indicators that shift fluorescence emission wavelength upon complexation. Just as the response range of a pH indicator depends on indicator K_a, the response range of a metal ion indicator depends on formation constant for metal ion binding.

Like potentiometric electrodes, indicators measure metal ion activity, that is, the concentration of metal ion the form of the free metal ion, rather than measuring total metal. In environmental and biological samples, the metal ion activity rather than total metal concentration determines the effects of a given metal ions.

Metal ion indicators have been successfully developed for specialized applications. Ratiometric fluorescent indicators for Ca(II) and Zn(II) have revolutionized the study of the biology of these metal ions. Research to develop useful indicators for other metal ions is ongoing.

Another important type of optical indicator involves fluorescence quenching. One important example involves oxygen, a common quencher. The indicator in this case is an efficient fluorophore that is subject to oxygen quenching. The decrease in fluorescence intensity can be related to oxygen partial pressure. An even better approach is to measure the change in fluorescence lifetime that results from quenching. The advantage of lifetime measurements is that they do not depend on the amount of fluorophore and thus are not affected by slow fluorophore photodegradation. Nitro compounds, particularly explosives like 2,4,6-trinitrotoluene, are efficient fluorescent quenchers and may be sensed by low levels by fluorescent indicators.

A recent review on optical sensors includes work on indicator-based optical sensors [9].

73.4 MASS SENSORS

There are two widely used types of mass sensor, the quartz crystal microbalance and the surface acoustic wave device. Both are based on the piezoelectric effect in quartz. Application of an electric potential causes the quartz to deform slightly. Application of a varying electrical potential creates an acoustic wave in the quartz. In the case of the quartz crystal microbalance, electrodes are attached to opposite sides of a quartz crystal. An oscillating potential leads to a shear wave that propagates through the crystal as shown in Figure 73.10. In the case of the surface acoustic wave device, the wave propagates along the surface of the quartz as shown in Figure 73.11.

These devices are incorporated into electrical circuits that allow them to oscillate at a resonant frequency. This frequency depends on the mass of the quartz. Changes in mass cause a shift in the resonant frequency that is proportional to the change in mass. The proportionality constant will depend on the viscoelastic properties of the coating that is changing mass.

These waves propagate freely in air but are damped by liquids. The damping is more severe for surface waves. The surface acoustic wave is more sensitive to mass loading and is able to detect mass changes in the small picogram range. However, because it is more subject to damping by liquids, it is mainly used for gas sensing. The quartz

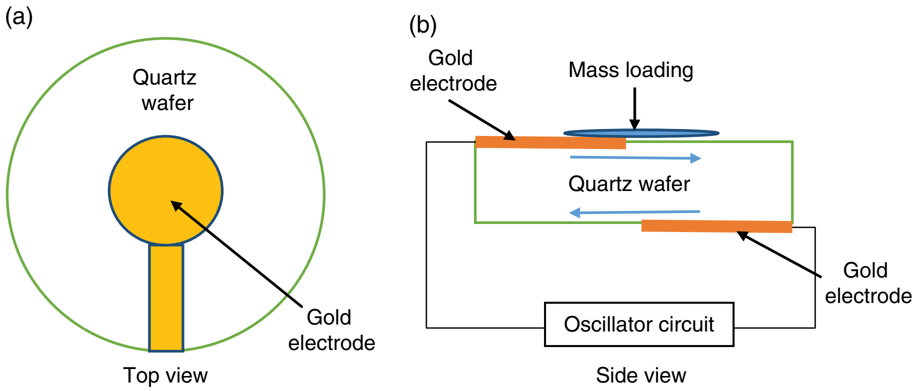


FIGURE 73.10 Schematic of quartz crystal microbalance. Electrical energy propagates as a shear wave through the quartz wafer. The oscillation circuit assumes the resonant frequency, which shifts in a linear fashion with mass loading. The arrows in the quartz wafer represent the mass displacement accompanying shear wave propagation through the quartz.

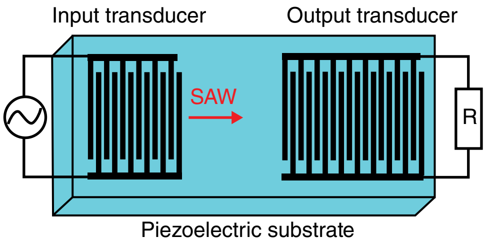


FIGURE 73.11 Arrangement for producing surface acoustic waves.

crystal balance is more easily operated in liquids with the ability to detect mass changes in the low nanogram range. It is more widely used for solution phase measurements.

Mass sensors can be coupled with recognition phases to make chemical sensors. What is detected is the shift in frequency accompanying analyte binding. One type of application involves gas sensors based on thin polymer coatings on surface acoustic wave devices. Sensitivity depends on the coefficient for gas partitioning into the polymer. Selectivity is low, depending on the relative affinity of the polymer for different gases. However, these sensors are good candidates for sensor arrays using several SAW devices each coated with a different polymer with a different selectivity pattern. This type of array is known as an electronic nose and will be covered in the last part of this chapter.

Quartz crystal microbalances are well suited for solution phase measurements. They are best suited for detecting large analytes such as proteins because of the larger mass change per analyte molecule. This type of device generally employs a biological recognition agent such as an antibody and is, therefore, classified as a biosensor.

An excellent book provides further details on acoustic sensors [11]. Recent activity in this area is summarized in a recent review article [12].

73.5 SENSOR ARRAYS (ELECTRONIC NOSE)

Many of the sensor technologies covered in this chapter are insufficiently selective to be useful for monitoring a specific analyte in a complex sample. However, they are sufficiently inexpensive that it is practical to prepare an array of sensors each of which responds with a somewhat different selectivity. This approach has proven to be practical for gas analysis. The chemically selective layers are polymers. They can be coated onto several of the base sensors described in this chapter including chemiresistors, surface acoustic wave devices, and field effect transistors. These sensor arrays are widely known as electronic noses because their response is similar to olfaction. Several electronic noses are available commercially.

The same concept can be applied to solution measurements. This type of array has been called an electronic tongue. However, while research is ongoing, it has yet to prove useful for practical problems and is unlikely to do so soon.

Electronic noses do not directly measure concentrations. Instead they are trained to recognize a particular type of sample. Determining the freshness of fish is an example of the kind of problem that can be addressed using an electronic nose. It would be initially exposed to a series a fish of known freshness. Then it would be used to assess the freshness of unknown fish using pattern recognition techniques to compare the vapor signature of the unknown fish to the signal from the known fish. Electronic noses are widely applicable to a variety of vapor detection problems [13–15].

REFERENCES

1. Janata, J.; *Principles of Chemical Sensors*, Second edition, Springer Verlag, 2009.
2. Banica, F.-G.; *Chemical Sensors and Biosensors: Fundamentals and Applications*, John Wiley & Sons Inc., 2012.
3. Moos, R.; Izu, N.; Rettig, F.; Reiss, S.; Shin, W.; Matsubara, I.; Resistive Oxygen Sensors for Harsh Environments, *Sensors* 11 (2011) 3439–3065.
4. Stetter, J.R.; Li, J.; Amperometric Gas Sensing – A Review, *Chem. Rev.* 108 (2008) 352–366.
5. Aswal, D.K.; Gupta, S.K.; *Science and Technology of Chemiresistor Gas Sensors*, Nova Publishers, 2007.
6. Eranna, G.; *Metal Oxide Nanostructures as Gas Sensing Devices*, CRC Press, 2011.
7. Chiu, S.-W.; Tang, K.-T.; Towards a Chemiresistive Sensor – Integrated Electronic Nose: A Review, *Sensors* 13 (2013) 14214–14247.
8. Zhang, C.; Chen, P.; Hu, W.; Organic Field Effect Transistor-Based Gas Sensors, *Chem. Soc. Rev.* 44 (2015) 2087–2107.
9. Qazi, H.H.; Mohammad, A.B.; Akram,M.; Recent Progress in Optical Sensors, Sensors 12 (2012) 16522–16556.

10. Gala, U.; Chauhan, H.; Principles and Applications of Raman Spectroscopy in Pharmaceutical Drug Discovery and Development, *Expert Opin. Drug Discov.* 2 (2015) 187–206.
11. Ballantine Jr., D.S.; Martin, S.J.; Ricco, A.J.; Frye, G.C.; Wohltjen, H.; White, R.M.; Zellers, E.T.; *Acoustic Wave Sensors, Theory, Design, and Physico-Chemical Applications*, Academic Press, 1996.
12. Vashist, S.K.; Vashist, P.; Recent Advances in Quartz Crystal-Based Sensors, *J. Sens.* 2011 (2011) 13 pages.
13. Gardner, J.; Bartlett, P.N.; Eds., *Sensors and Sensory Systems for an Electronic Nose*, Springer, 2013.
14. Bartlett, P.N.; *Electronic Noses: Principles and Applications*, Oxford University Press, 1999.
15. Patel, H.K.; *The Electronic Nose: Artificial Olfaction Technology*, Springer Verlag, 2014.

INDEX

Handbook of Measurement in Science and Engineering, First Edition. Edited by Myer Kutz.