Final Project for the Prediction of the Ideal Dielectric Breakdown Based on First Principle Calculations of Common Insulators and Semiconductors

Team Members: Antoine Moats and Martijn Bindels

Motivation:
Insulating materials are used across many engineering disciplines as they are integral to capacitors, wire insulation and semiconductors. However, these materials are only insulating if the external electric field is below a certain threshold, known as the maximum electrical breakdown field. Above this threshold, the material becomes conducting and loses its useful properties. Predicting the maximum value of the electrical breakdown field through empirical means is difficult due to the formation and accumulation of defects which is variable based on ambient temperature, structural vacancies, and the duration of the exposure to the external electric field. Therefore, the use of ab initio methods such as DFT are important in the discussion of the dielectric breakdown for any material.

Objective:
The goal of the project is to determine the maximum intrinsic dielectric breakdown field based on 8 key parameters. The model will be trained on intrinsic dielectric breakdown fields from density functional theory (DFT) calculations based on the database provided here: https://www.kaggle.com/datasets/chaozhuang/dielectric-breakdown-prediction-dataset?resource=download

Deliverable:
We will develop a machine learning architecture that will predict the intrinsic dielectric breakdown for any material where the following eight characteristics are defined:

1. Experimental band gap
2. Phonon cutoff frequency
3. Mean phonon frequency
4. Electronic contribution of the dielectric constant
5. Total dielectric constant
6. Nearest neighbor distance
7. Density
8. Bulk modulus
9. Crystal structure

The accuracy of the machine learning architecture will be determined based on a certain tolerance value (such as within 5% of expected). In addition, a correlation matrix between certain parameters and the dielectric breakdown will be determined to give a better insight on whether certain parameters are important versus which can be ignored.

Research Plan

Develop a Deep Neural Network (DNN) that utilizes the eight aforementioned parameters (some of which came from experiments, DFT calculations, or a priori knowledge) in the database to predict the intrinsic dielectric breakdown. This architecture will have a set amount of width for each layer and will vary in # of layers to potentially determine a set that has the best predictive abilities. The DNN architecture is included in figure 1 to clarify its structure.

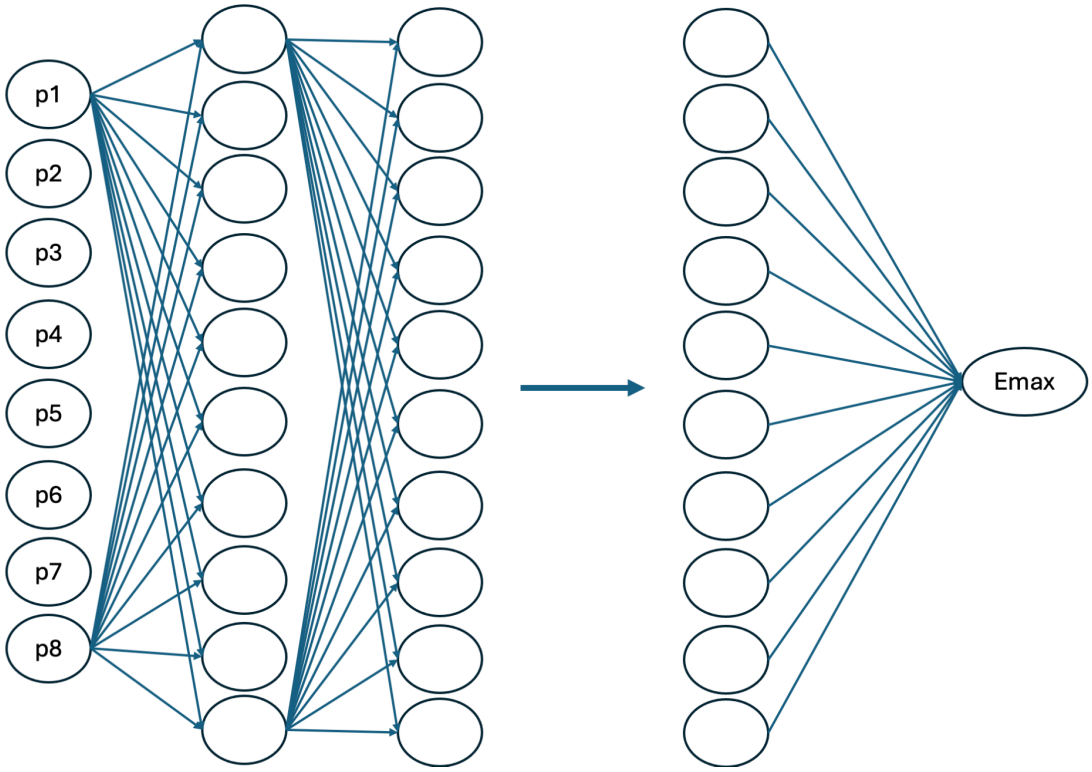

Figure 1: Deep neural network with a few of the nodal connections to demonstrate the feedforward system with constant width and variable # of layers

The loss function will simply be the mean squared error between the ML predicted dielectric breakdown value and the calculated dielectric breakdown value.

The 82 common insulators and semiconductors in the database will be split into a training set and a validation set based on the 80/20 rule i.e. ~65 will be used for training and 17 will be used for validation. It is evident that the DNN will need to be tuned based on several key hyperparameters. The hyperparameters that we need to tune are the activation functions (tanh v.s sinh), learning rate, regularization parameters, model width and layer #.

In order to determine the efficacy of our model, we will evaluate the accuracy and loss of our model with respect to other materials with the inputs and expected dielectric breakdown constant found in the materials project website. We will also compare our model to other architectures such as the random forest trained on the same dataset found here: https://pubs.acs.org/doi/10.1021/acs.chemmater.5b04109

Results and Discussion

A parameter sweep of the width and depth spanning multiple orders of magnitude was conducted to determine the regions where the optimal performance occured. The metric to be optimized was the percent of the predictions of the validation set that were within 5% of the actual values. Plots of the percent of the predictions that were within 5% demonstrated that increasing the width (and thus the number of connections) led to a decrease in performance.

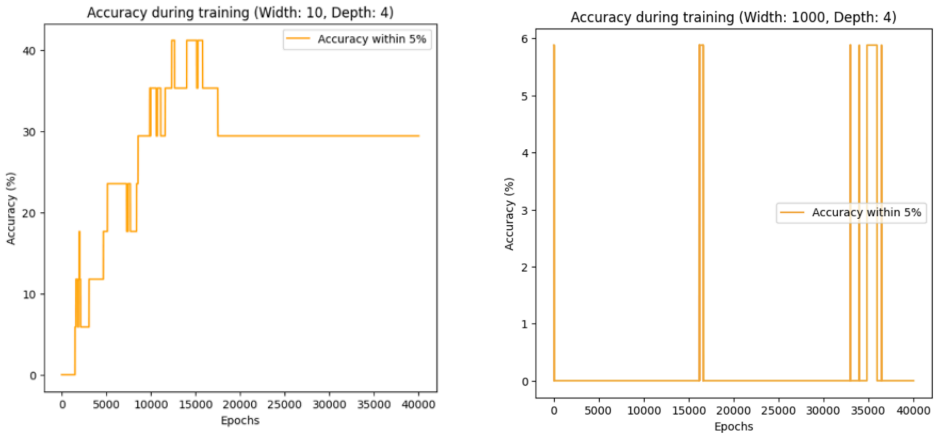

Figure 1: Plot of the percent accuracy of a tanh activation function DNN with a width 10 (left) vs width 1000 (right) and a constant depth of 4.

Based on the testing of widths > 50 which decreased performance, the parameter sweep of the width and layers was narrowed down to numbers between 4 and 10. In table 1, the loss of a DNN using the tanh activation function is recorded over several widths and layers. The analysis of the performance of the model was switched from the percent within 5% of expected for the validation data to the training loss to determine the relationship between number of connections and performance using a different metric. Using the loss as the metric for performance, it was evident that the parameter sweep resulted in a pair of optimal widths and depths for the tanh function.

Table 1: Loss versus width and layers when tanh activation function is used.

Width → / Layers↓	4	6	8	10
4	3.95E-04	1.76E-06	9.85E-06	3.08E-08
6	1.35E-02	1.52E-05	1.14E-09	8.59E-08
8	1.43E-04	2.39E-08	1.79E-09	9.89E-09
10	2.23E-04	2.63E-07	8.03E-10	3.59E-06

While the optimal width and depth were found for the training loss, the percent accuracy within 5% of the validation data still did not exceed 50%. This was a major concern as it appeared as though increasing the number of connections increased overfitting, reducing the

generalizability of the model. This led to the comparison of the percent accuracy for the sin and sigmoid activation functions for various widths at a depth of 6 to determine whether overfitting may be reduced using different activation functions.

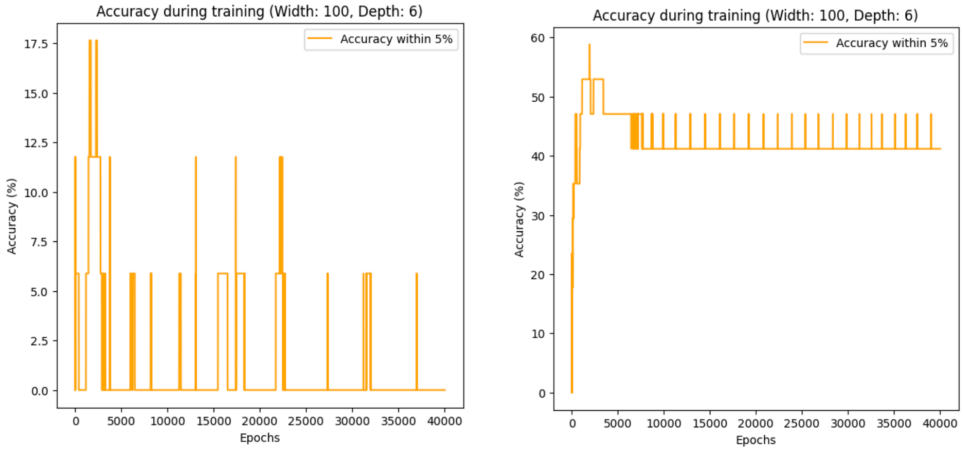

Figure 2: Plot of the percent accuracy for the sin activation function (left) and the sigmoid activation function (right) at the same width and depth.

The drastic fluctuations in the % accuracy shown in the left side of figure 2 is representative of most of the models that used the sin activation function. These same fluctuations were noted in the loss function and demonstrated that the sin activation function was not a promising activation function. However, the sigmoid emerged as a promising architecture as it resulted in ~40% accuracy for large widths (100) shown on the right side of figure 2 and small widths (10) not shown.

The sigmoid DNN resulting in approximately the same accuracy as well as training loss for a range of widths and depth necessitated the use of a different metric to characterize the performance of the model. After looking into the research paper that published the database, other metrics were found. By plotting the ML predicted and the DFT computed electrical breakdown fields for both the validation and the training data, the authors were able to make conclusions on the efficacy of their models.

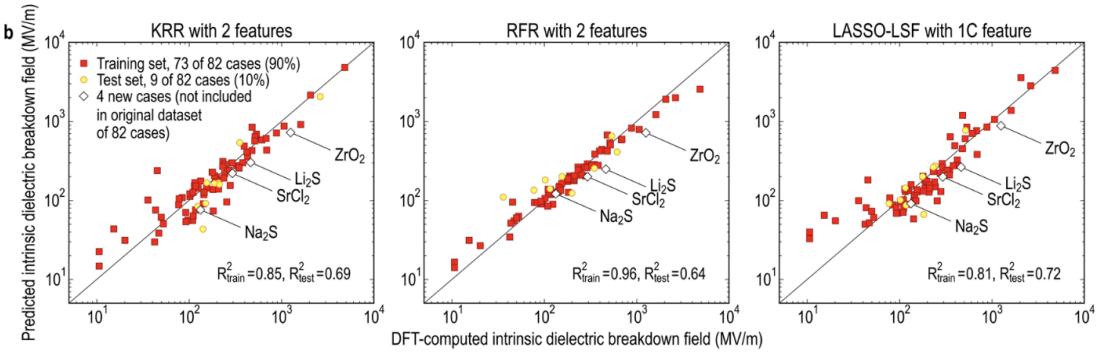

Figure 3: Plot of the predicted vs DFT computed electrical breakdown field for the kernel ridge regression (KRR), random forest regression (RFR) models, and the least absolute shrinkage and selection operator based least-squares fit (LASSO-LSF) [1].

Using this analysis, the tendency of the sigmoid DNN to result in overfitting for a range of widpth and depths was clear. As shown in figure 4, the R^2 for the training data is a perfect match. However, the R^2 for the validation data is much lower than 80% illustrating that these models can not be generalized to materials with different inputs.

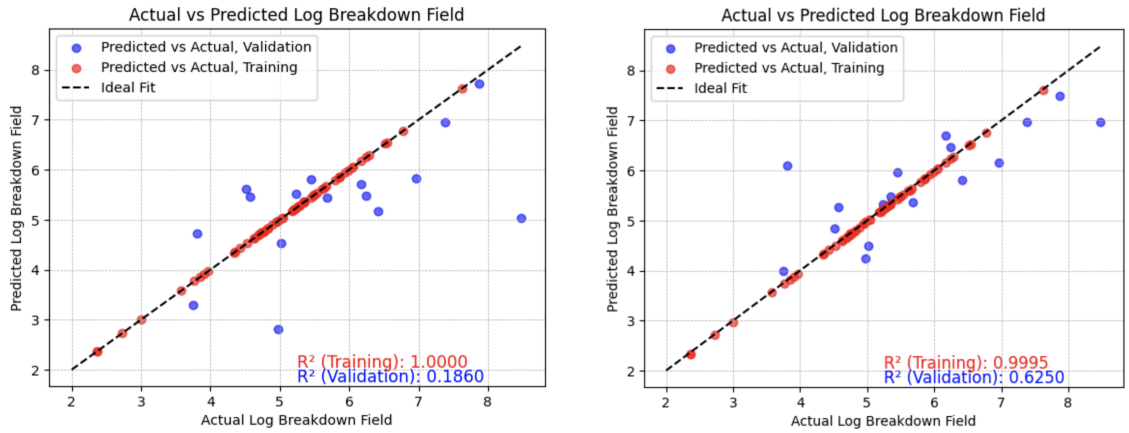

Figure 4: Plot of the predicted vs DFT computed electrical breakdown field for the sigmoid activation with a width of 10 (left) and width of 100 (right).

Using the R^2 analysis for the training and validation data sets demonstrated that while the sigmoid performed well using the percent within 5% metric it resulted in overfitting. This led to the same analysis being conducted for the tanh activation function for the parameters in table 1 which utilized relatively small width and depths. As shown in figure 5, a neural network with a width of 2 and depth of 2 provides a good model fit for both the training and the validation data while the model with more parameters underperforms. Using more complex models for the tanh leads to overfitting, which can be seen in figure 5 as the spread of the validation data is larger than the spread of the training data.

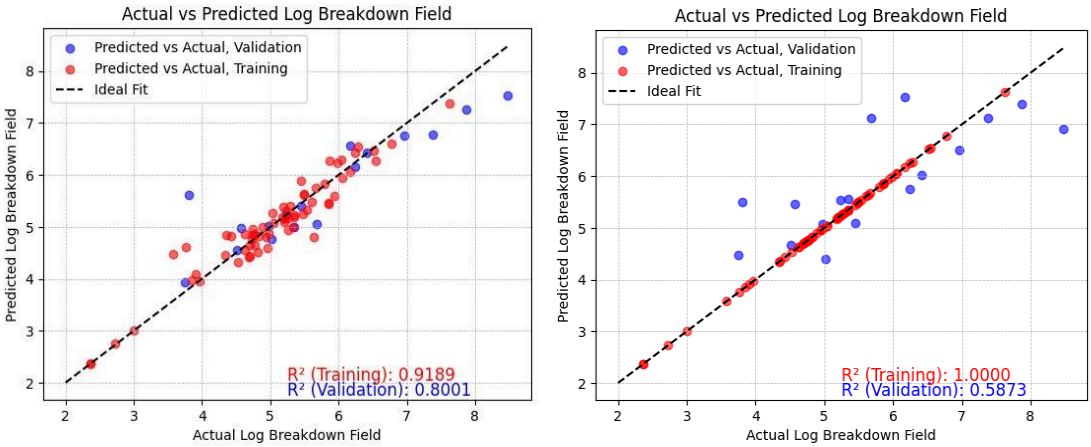

Figure 5: Actual vs Predicted Log Breakdown Field for a width of 2 and 2 layers shown on the left, and for a width of 6 and 6 layers shown on the right.

In addition to comparing different parameters of the DNN, a random forest neural network was developed to act as a comparison. While a DNN is characterized by its width and depth, a random forest model is characterized by the number of trees. In order to determine the number of trees that might result in the lowest loss, the RMSE vs number of trees was plotted as shown in figure 6. This analysis demonstrated that the loss plateaus once the number of trees exceeds ~20. Another tunable hyperparameter of random forest models is the seed which introduces randomness into the model. Iterating through a few random seed values illustrated that a random seed of 40 yielded the best results. Using a random seed of 40 and 50 trees, the plot of the predicted vs DFT calculated electrical breakdown field is included in figure 6.

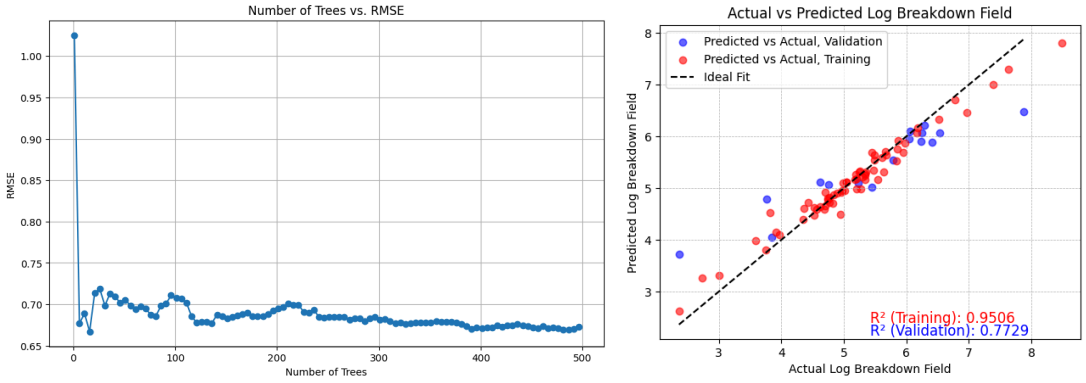

Figure 6: Number of trees versus root mean squared error (left) and Actual vs Predicted electrical breakdown (right).

Conclusions and Future Work

In this report the ability of a deep neural network to predict the maximum electrical breakdown field based on eight key characteristics is quantified. Several metrics were used to conduct parameter sweeps including the training loss, % of the predictions within 5 percent of the expected, and an R^2 analysis of the predicted vs actual electrical breakdown fields. While not explicitly stated earlier in the report, as they had the least amount of impact on the accuracy of the model, DNNs with regularization values of [1e-4, 1e-5, 1e-7, 1e-9] and learning rates of [1e-3, 1e-4, 1e-5] were tested. A regularization of 1e-7 and 1e-3 yielded the >0.91 R^2 for the training set and >0.80 R^2 for the validation set shown in figure 5. These results demonstrate that using a simple NN with a width of 2 and a depth of 2 can be used to predict the maximum electrical breakdown field. Furthermore, it was found that for the tanh activation function, increasing the number of connections increased overfitting. This observation is not consistent for the sigmoid activation function which illustrated that as the width increased the R^2 for the validation set (figure 4). However, the potential increase in the R^2 for the validation set vs increase in computational cost did not justify larger neural networks to be tested as the much smaller tanh NN performed well. In addition to performing parameter sweeps for a DNN, a random forest model was developed based on the same technique used in the original paper by Kim et al. (2016). The random forest model using a random seed of 40 and 50 trees yielded a >0.95 R^2 for the training set and >0.77 R^2 for the validation set; these results are comparable to the performance of the tanh DNN.

Going forward, the DNN model could be used to compute the expected electrical breakdown field in a manner that is quicker and less computationally expensive than DFT computations and easier to implement compared to RFR and other more complicated ML models. Future work could entail the use of characteristics not concluded from DFT calculations. Because the objective is to reduce computational cost through these models, training the models on data available from other means would be more useful. For example, someone who wanted to use this model to predict the electrical breakdown field for a new material would need to know the mean phonon frequency and electronic contribution of the dielectric constant which are determined through DFT calculations. However, conducting these calculations is computationally expensive and may be more work than calculating the maximum electrical breakdown field using DFT.

Works Cited

[1] Kim, Chiho, et al. “From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown.” *Chemistry of Materials*, vol. 28, no. 5, 8 Mar. 2016, pp. 1245–1588, https://doi.org/10.1021/acs.chemmater.5b04109.s001.