Research Assessment #13

Name: Aiswharyaa Lalgudi Nagarajan

Date: 15 November, 2024

Subject: ISM (Independent Study and Mentorship)

MLA citation:

Ouyang, John F., et al. “Deep Learning Models Will Shape the Future of Stem Cell Research.” *Stem Cell Reports*, vol. 18, no. 1, Jan. 2023, pp. 6–12. https://doi.org/10.1016/j.stemcr.2022.11.007.

Assessment: Research Assessment #12

The most important part in the article are about how deep learning (DL) is revolutionizing stem cell research. The article mentions the increasing relevance of DL in analyzing and interpreting massive amounts of biological data. The article continues to outline how DL models have been used to reconstruct developmental trajectories, predict differentiation potential, and combine multi-modal data sets such as transcriptomics and imaging data. An illustrative example that stands out is AlphaFold2, which has transformed protein structure prediction, providing a model for the future effect of DL in stem cell science. Furthermore, the article emphasizes that while DL has excellent potential, challenges such as batch effects, dataset standardization, and model interpretability must be addressed to ensure meaningful biological applications.

### 

This information is highly relevant to my research in computational biology, particularly for my RNA-seq + CNN-RNN model that I presented at Regeneron ISEF. The article corroborates my approach by validating DL's need to analyze intricate biological data. It explains how single-cell transcriptomics has been developed with the help of DL usage, complete with case studies of models that are particularly good at integrating high-dimensional datasets. It aids the further development of my model to minimize false positives and maximize predictive ability. In addition, the article also talks about the significance of incorporating standardized datasets from sources like ENCODE, FANTOM, and the Human Cell Atlas (HCA). This made me consider incorporating these repositories into my dataset preprocessing pipeline to develop more solid models. Moreover, the issue of DL model interpretability makes me consider looking at explainable AI techniques to achieve the maximum transparency possible in my model's predictions, which will be a critical point to ensure while presenting to ISEF judges.

The information in the article can be broken down into four core themes: (1) the necessity of high-quality, standardized datasets, (2) the various applications of DL in stem cell research, (3) the challenges related to DL interpretability, and (4) the future of integrating DL into experimental workflows. One of the key takeaways is that while DL models have significantly improved our ability to read complex biological data, their value is often stymied by batch effects, dataset variability, and the requirement for robust validation techniques. The article also clarifies that while consortia-led data generation efforts (e.g., ENCODE, FANTOM, HCA) have mitigated some of these challenges, they do not eliminate variations arising from different experimental protocols. This directly relates to my research, where batch effects in RNA-seq data have been a concern, affecting the reliability of model predictions. Furthermore, the discussion on the limitations of black-box models in biology provides insights into the necessity of model explainability, reinforcing the importance of techniques such as attention mechanisms and feature attribution methods.

With this information, I will improve my research approach by embracing better practices in data preprocessing, including utilizing standardized repositories for training data and developing techniques to mitigate batch effects. A specific step is embracing sophisticated data harmonization techniques, including domain adaptation and transfer learning, to increase the generalizability of my CNN-RNN model. Additionally, I will explore the addition of multi-modal datasets (e.g., transcriptomics, epigenomics, and imaging data) to enhance accuracy, which was the article’s main emphasis. Another primary consideration is the challenge of model interpretability. Inspired by the article’s discussion on making DL models more biologically meaningful, I will investigate explainability frameworks such as SHAP (Shapley Additive Explanations) or Grad-CAM (Gradient-weighted Class Activation Mapping) to highlight which features contribute to predictions. This will provide scientific credibility to my project and help me present strong evidence validating my project’s accuracy and novelty. Lastly, this research raises a fundamental question: How do we bridge the gap between in silico predictions and actual biological confirmation? This prompts me to consider potential experimental collaborations that could validate my model’s predictions in a wet lab setting.

This paper effectively reiterated the importance of high-quality data and cross-disciplinary collaboration in DL applications. It provided theoretical and practical data that directly impacted my research, so it was a worthwhile paper to read. The discussion on the importance of robust data preprocessing techniques was particularly insightful because it has motivated me to seek additional ways to improve the consistency of my RNA-seq data. Furthermore, the paper raised points I had not carefully considered, such as the necessity of in silico to in vitro verification. The importance of the findings in the paper has maintained my interest and passion in my work. Overall, the article has been thought-provoking and inspiring, and it has made me anticipate refining my methodology and implementing novel techniques in my project.