Understanding AlphaFold and Its Implications for a Deep Learning Approach to Protein-Compound Modeling

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Meghan Anderson

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Yanjun Qi, Department of Computer Science

Understanding AlphaFold and Its Implications for a Deep Learning Approach to Protein-Compound Modeling

CS4980

Meghan Anderson
Computer Science
The University of Virginia
mfa3fzv@virginia.edu

ABSTRACT

The AlphaFold program, which employs principles of artificial intelligence to generate predictive models of protein conformations based on amino acid sequences, has provided the scientific community with a solution to the protein-folding problem. The breakthrough in modeling may hold the key to solving a few emergent issues in bioinformatics relating to protein structure in response to different environmental conditions, binding chemical compounds, and binding other proteins. This project includes a survey of the successful approach to the problem of 3D modeling protein structures. The study also provides an analysis of current literature researching the implementation of deep learning methods to accurately predict compound-protein interactions.

While the implementation of a predictive model of protein-ligand conformations incorporating AlphaFold was not feasible in the time frame for the project, an investigation of contemporary methodologies predicting compound-protein interactions offers a window into how such a program could be established. The work is a small contribution to a burgeoning area of research with the intention of making the topic more accessible to computer scientists with a limited background in biology.

1 INTRODUCTION

The protein folding problem is a long-standing struggle among biological and medical researchers. Protein structure is highly correlated to protein function. Consequently, information about protein conformation can offer insights into how cells function as well as insights into the manipulation of proteins for medicinal purposes. The conformation of a protein can be determined experimentally, but the process is expensive, time-consuming, and considerable. Determining the structure of a single protein with X-ray crystallography could take up to three or five years [1], and the number only goes up when tackling an array of different proteins. Other laboratory methods have constraints as well. NMR spectroscopy is constrained to smaller proteins. Electron microscopy, on the other hand, is limited by high cost and maintenance, though it accommodates larger proteins. Thus, the search for a computational model capable of eliminating the need for intensive experiments has been ongoing.

1.1 AlphaFold

AlphaFold [2] has been the stand out of all of the programs developed to provide 3D modeling of proteins. The AlphaFold project executed by DeepMind was started in 2016, and the source code for the AI system was made available to the public in 2020. AlphaFold 2, the second, updated version of AlphaFold has successfully achieved protein structure accuracy comparable to that of experimental procedures in labs. Generally, an accuracy of above 90% is considered to be within the threshold required for a predicted structure to be considered reliable. No other predictive model system had accomplished this feat [3]. However, the inclination to declare the protein-folding problem solved has been cautioned.

1.1.1 How AlphaFold Works

Amino acid sequences, the primary structure of proteins, serve as input to the program. The idea is to enter the amino acid sequence derived from the genetic code of a protein and receive a 3D model of the protein as output. The program consists of a recurrent convolutional neural network (CNN) applied to sequence alignments between the amino acid sequence of the input and amino acid sequences of homologous proteins with known 3D structures from genetic databases. Iterative refinement is implemented to continually adjust the conformation as the algorithm runs.

AlphaFold has revolutionized 3D protein modeling, but the system is just the tip of the iceberg in terms of predicting the vast array of conformations individual proteins can take on.

1.2 Expansions to AlphaFold to Address Limitations

Despite its success, AlphaFold has its limitations. For one, most proteins do not take on a singular conformation, yet AlphaFold only predicts one conformation for each protein. Following this observation, AlphaFold does not address the issue of conformational changes in response to ligand binding. Moreover, in vivo proteins are generally exposed to a number of ions and cofactors that may alter the orientation of certain amino groups, but the model does not account for these interactions. The issue of predicting protein structure in the presence of other entities may not have a universal solution, but instead require multiple solutions to more niche problems. The research work scrutinized in this paper focuses mainly on the interactions between proteins and chemical compounds. An extension on the AlphaFold program for protein-protein modeling is, however, included to highlight how AlphaFold may be incorporated into the CPI prediction models.

1.2.1 The Compound-Protein Interaction (CPI) Problem

In the interest of the development of a deep learning CPI prediction program, a large volume of data must be obtained on compounds, protein targets, and their corresponding molecular interaction profiles. The problem is obscured by the fact multiple compounds may bind to the same protein, but at different locations. Because of this, the computational model would need to predict three things: (1) whether the protein and compound will bind, (2) the active site for compound docking, and (3) the structure of the protein-compound complex. Over the years, many experiments have been conducted with the hopes of assembling dependable models for CPIs. These experiments have been relatively successful, but none offer the generalizability required of a powerful predictor [5]. That is, they do not perform well for unknown compounds and proteins. The strides in protein-folding made by AlphaFold might be the missing piece to creating such a system, but an investigation into this expansion is minimal at this time.

Protein-compound binding has a myriad of practical applications. Perhaps the most important application lies in drug discovery. Formulating a fast-acting, effective drug rests on the ability to predict the interaction between the drug, a chemical compound, and proteins it will encounter in the body. Aside from pharmaceuticals, the matter of understanding chemoreceptors in the body, in general, would be facilitated by a program capable of diagramming protein relationships to compounds of interest.

2 BACKGROUND

The first steps to analyzing and understanding AlphaFold involved cloning the open-source code from the GitHub repository into a Jupyter Notebook. System requirements to run the program are outlined in the ReadMe file and include installing Docker, NVIDIA container tool kit, genetic databases, and model parameters. The code is written primarily in Python but contains some code in other languages: Jupyter Notebook, Shell, and DockerFile. AlphaFold was downloaded and run to both discover possible routes for approaching the protein-compound problem and garner a better understanding of the program itself. To test the code on sample data and garner a better understanding of how it functions, the program was run with amino acid sequences from FASTA files in the NIH database.

As no implementation of a CPI predictor integrating AlphaFold was constructed, the only additional information crucial to understanding the latter sections of this body rests on familiarity with fundamental machine learning concepts.

3 RELATED WORK

Great expectations for advancement with machine learning technologies in the world of drug discovery have led to several broad overviews of successful models. Two dissertations laid the groundwork for this inquiry into progress projecting probable CPIs.

Lim et al. [2021] realized an exhaustive report of modern models for CPI prediction. Their summary honed in on what databases were popular as well as the AI methods which had shown great success. The body mentions AlphaFold only briefly. It contends that AlphaFold’s utilization of evolutionary knowledge of proteins ushered in its triumph over the previous modeling. This observation could be advantageous to others hoping to predict CPIs, but the work does not detail the incorporation of the AlphaFold program or its accompanying database. The survey also details the older machine learning schemas such as tree-based models and SVMs. The work eventually outlines data representation and negative decision boundaries as the prominent challenges for future works implementing AI for CPI prediction.

The second body of work related to the objective of this paper concentrates specifically on drug interactions with protein targets rather than compounds and proteins in general [5]. The inspection contains a deep dive into both the machine learning methods and databases in vogue today, and, similar to previous work, enlists varying data representation methods across databases as a major challenge for CPI predictions. This article made no mention of AlphaFold.

Where these works provided broad overviews of CPI predictive models over the past decade, this paper will instead focus solely on deep learning models that have emerged in the past three years.

4 SYSTEM DESIGN

Many models have come to fruition in applying deep learning to CPIs. The models assessed employ neural networks on compound and protein data to output predictions on binding with no visual representation of the protein-compound complex.

4.1 Neural Networks for CPI

4.1.1 End-to-End Learning of Neural Networks for Graphs and Sequences

The first methodology to be reviewed makes use of both graph neural networks (GNN) and CNNs to predict CPIs [7]. The procedure set forth by the bioinformatics researchers follows end-to-end learning concepts of machine learning. To train the model, one-dimensional data of compounds were converted to two-dimensional graphical representations, where atoms are nodes and bonds are edges, with the software RDKit. A GNN was then applied resulting in a compound vector. Meanwhile, a CNN was trained on one-dimensional protein data to produce a protein vector. A neural attention mechanism then joined the two vectors to produce CPI predictions.

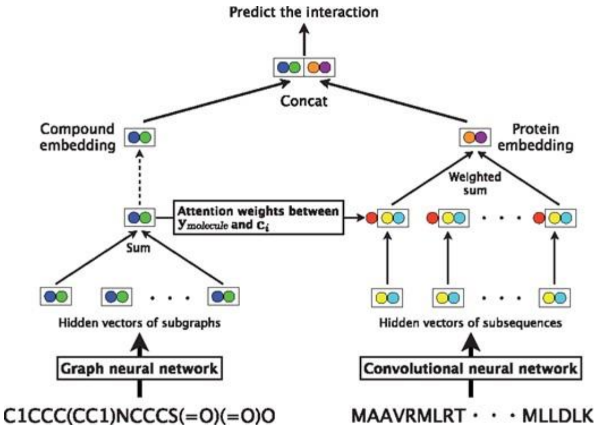

Figure 1: A high-level overview of the program that utilizes graph networks to predict CPI.

The work boasts a model prediction accuracy exceeding that of previous works which relied on SVM principles such as fixed data and perform particularly poorly on unbalanced datasets. Construction of 3D models of the compound-protein complex may be generated from the data the algorithm outputs; they are not an included feature of the system. The preeminent limitation of this procedure is its reliance on input for protein structure to be in 1D format. The amino acid sequences alone do not capture the mosaic of molecular interactions that dictate protein folding. The authors of the paper claim creating a GNN to accommodate 3D protein structures as input would likely increase the accuracy of the CPI predictions.

4.1.2 Deep Learning Integration of Molecular and Interactome Data

Researchers of Keio University [8] constructed their model similarly, but theirs involves a more rigorous course of training. Rather than training simply on protein and compound information, the model incorporates both protein-protein networks and compound-compound networks for the sake of more robust vectors for each. The design favors the management of compound features via ECFP. ECFP is an algorithm which recursively identifies partial structures surrounding an atom, in turn precisely modeling the ever-changing molecular environment. One-dimensional protein data is passed through a CNN. The Node2vec algorithm was run on both sets of network data to create graphical representations of the data with edges representing the reliability of the data.

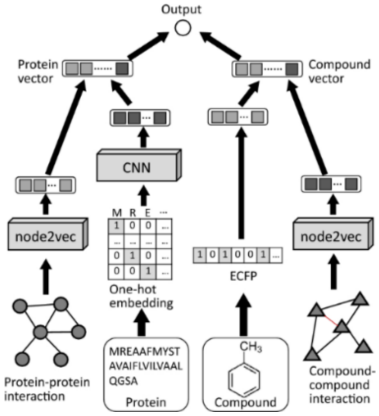

Figure 2: A high-level overview of the overall learning architecture for the program developed by researchers at Keio University with molecular and interactome data.

All the features of the compound data and all the features of protein data were concatenated separately and mapped onto the same latent space. This formulates the output indicating whether or not a given protein and compound will bind. SSGraphCPI [9] utilizes an extremely similar framework reliant on structural protein features.

It was concluded that the incorporation of network data contributed to the higher accuracy of CPI predictions when compared to the single-modal model. Again, the output of the algorithm simply provides information on whether a compound and protein will bind, and does not directly provide a 3D visualization for the interaction.

4.1.3 MONN

Li et al. [2020] established a Multi-Objective Neural Network (MONN) to forecast CPIs. Primarily relying on graphical representations of compounds and amino acid sequences, this framework employs both GNN and CNN to establish features. The objectives of the model are (1) non-covalent interaction prediction and (2) binding affinity prediction as a numerical value.

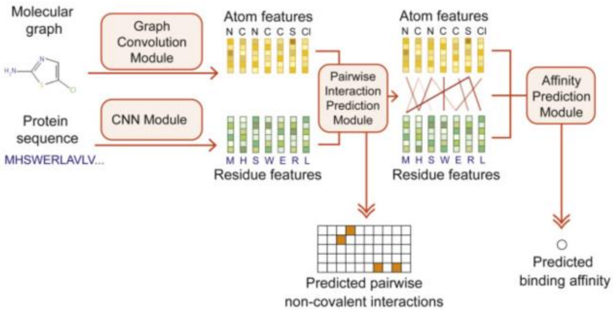

Figure 3: The Network Architecture of MONN.

Graphical representations of compounds are favored, as are amino acid sequences as input for proteins. Some structural properties of proteins are added later in the program to hone the model on protein features.

The work claims to fill in a gap in previous models whose neural attention could not capture the intricate non-covalent interactions between compound and protein.

5 CONCLUSIONS

5.1 Common Themes in Deep Learning CPI Modeling

Since 2019, various machine learning models have been generated to predict CPIs. Each of the models implements CNNs and/or GNNs to generate features on compounds and proteins. Along the same vein, each model trains on compound and protein data separately and feeds these features into a separate module for the prediction of CPIs. Future works may find better results by imitating the machine learning models described in this paper over other machine learning methods.

Another common theme in these models is the selection of databases whose representations of compounds are one-dimensional and may be transformed into two-dimensional data. Graphical representations of compounds may be highly important to feature training as the intricacies of atomic bonding within a molecule may be extrapolated. As for protein representations, the models all take in one-dimensional data (amino acid sequences from databases). However, the incorporation of protein structure data as well as protein-protein interaction data has proven to be effectful in generating a more robust model. Some other aspect to note that all these models share is their dependence on a neural attention mechanism to aid in prediction. Attention mechanisms amplify certain subsets of input data over others by weighting each piece of information. This concentration on distinct data may be integral to achieving high accuracy in predictive models and is an increasingly popular convention in machine learning.

The output of each model is not characterized as three-dimensional models of protein-compound complexes. This sheds light on a potential limitation solvable with fusion with AlphaFold designed to output 3D information. The models in this study were selected because they seemed to mimic the general structure of the AlphaFold neural architecture. This could be extremely helpful to the task of joining AlphaFold and CPI models to create a singular program.

5.2 The Future of CPI Prediction

As the current model schema for CPI prediction involves separate training of protein and compound data, incorporation of AlphaFold into any of the frameworks addressed could be achieved. Strides have been made in predicting protein-protein interactions with the modification of the AlphaFold program [13]. Learning from these successes as well as drawing from contemporary CPI models could reveal a way for CPI modeling and AlphaFold to coalesce. It is even possible the binding affinities produced by these models could be added as additional information to the AlphaFold system to enable the production of compound-protein complexes.

This work provides an updated literature review to highlight the potential for future CPI predictive models with the remarkable strides of AlphaFold. Three-dimensional visualizations can enable scientists to scrutinize the results of CPI prediction more effectively. Furthermore, visual modeling of a protein in the presence of a compound may present drug researchers with more information on whether compound linkage will desirably modify protein conformation. It is evident there is an appreciable amount of work left to be done to authenticate a predictive, computational CPI model. Whether the AlphaFold program itself is repackaged to include CPI prediction capabilities or the database constructed is integrated, it appears AlphaFold will have a great impact on CPI modeling in years to come. The overview executed in this paper may serve as a jumping-off point for more scientists to begin researching machine learning models essential to CPI prediction, particularly computer scientists with little familiarity with biological and chemical processes.

6 FUTURE WORK

The bulk of this project consisted of understanding the databases and machine learning algorithms in use today to predict protein structures. The work is simply a high-level overview of a convoluted field of study. Additional training and more experience with bioinformatics would have made me better equipped to tackle this project. Given significantly more time and access to resources, a workflow could be designed to predict CPIs that make use of the advancements in the field with AlphaFold. Nevertheless, more research into budding machine learning models and developments in the world of AI would allow for a deeper investigation into predictive models for drug discovery. AI is developing at a rapid pace, so rapid developments in protein-compound modeling are sure to be on the horizon.

7 REFERENCES

[1] Shikha Agnihotry, Rajesh Kumar Pathak, Dev Bukhsh Singh, Apoorv Tiwari, and Imran Hussain. 2022. Chapter 11 - Protein structure prediction. In Bioinformatics, Dev Bukhsh Singh and Rajesh Kumar Pathak (eds.). Academic Press, 177–188. (2022). DOI:https://doi.org/https://doi.org/10.1016/B978-0-323-89775-4.00023-7

[2] Jumper, J., Evans, R., Pritzel, A. *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021). DOI:https://doi.org/10.1038/s41586-021-03819-2

[3] Deng, H., Jia, Y., & Zhang, Y. 2018. Protein structure prediction. *International journal of modern physics. B, 32* (2018), 1840009. DOI:https://doi.org/10.1142/S021797921840009X

[4] Lim, S., Lu, Y., Cho, C. Y., Sung, I., Kim, J., Kim, Y., Park, S., & Kim, S. 2021. A review on compound-protein interaction prediction methods: Data, format, representation and model. *Computational and structural biotechnology journal*, *19*, 1541–1556. (2021). DOI:https://doi.org/10.1016/j.csbj.2021.03.004

[5] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. 2020. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Briefings in Bioinformatics 22, 1 (January 2020), 247–269. DOI:https://doi.org/10.1093/bib/bbz157

[6] Jian Wang and Nikolay V. Dokholyan. 2021. Yuel: Compound-Protein Interaction Prediction with High Generalizability. bioRxiv (2021). DOI:https://doi.org/10.1101/2021.07.06.451043

[7] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. 2018. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35, 2 (July 2018), 309–318. DOI:https://doi.org/10.1093/bioinformatics/bty535

[8] Watanabe, N., Ohnuki, Y. & Sakakibara, Y. 2021. Deep learning integration of molecular and interactome data for protein–compound interaction prediction. *J Cheminform* **13,** 36 (2021). DOI:https://doi.org/10.1186/s13321-021-00513-3

[9] Wang, X., Liu, J., Zhang, C., & Wang, S. 2022. SSGraphCPI: A Novel Model for Predicting Compound Protein Interactions Based on Deep Learning. *International journal of molecular sciences*, *23*(7), 3780. (2022). DOI:https://doi.org/10.3390/ijms23073780

[10] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. 2020. MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities. Cell Systems 10, 4 (2020), 308-322.e11. DOI:https://doi.org/https://doi.org/10.1016/j.cels.2020.03.002

[11] Bryant, P., Pozzati, G. & Elofsson, A. 2022. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13,** 1265 (2022). DOI:https://doi.org/10.1038/s41467-022-28865-w

[12] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. 2022. Protein complex prediction with AlphaFold-Multimer. bioRxiv (2022). DOI:https://doi.org/10.1101/2021.10.04.463034

[13] Zhan Deng, Claudio Chuaqui, and Juswinder Singh. 2004. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. Journal of Medicinal Chemistry 47, 2 (2004), 337–344. DOI:https://doi.org/10.1021/jm030331x

[13] Bryant, P., Pozzati, G. & Elofsson, A. 2022. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13,** 1265 (2022). DOI:https://doi.org/10.1038/s41467-022-28865-w