Precision in Mutation: Enhancing Drug Design with Advanced Protein Stability Prediction Tools

Team 4 | AlgoRxplorers

Karishma Thakrar, Jiangqin Ma, Max Diamond, Akash Patel

ABSTRACT

Understanding the impact of single-point amino acid mutations on protein stability is crucial for unlocking insights into disease mechanisms and advancing drug development. Protein function relies heavily on structural stability, influenced by changes in Gibbs free energy ($\Delta\Delta G$) resulting from these mutations. The challenge in predicting protein stability changes lies in the scarcity of data and the complexity of model interpretation. Our study proposes the use of deep neural networks to accurately predict $\Delta\Delta G$ values. By improving model interpretability and applying transfer learning, we intend to enhance prediction accuracy and generalizability. This approach promises not only to refine $\Delta\Delta G$ predictions but also to contribute to a deeper understanding of protein dynamics, potentially revolutionizing disease research and drug discovery.

1 INTRODUCTION

Many drugs target proteins to modulate their activity such that they may bind more efficiently to their targets and lead to more effective treatments [3]. A protein's activity is typically impacted by alterations in its sequence, leading to a significant change in its structure and function, thereby influencing protein stability. When mutations cause the protein to become unstable, it often leads to a range of diseases and cancers, underscoring the importance of maintaining protein structure integrity for cellular health. Meanwhile, mutations that enhance protein stability could lead to the development of more effective drugs, offering new avenues for treatment. In 2022 alone, the pharmaceutical companies spent nearly $244 billion on R&D, underscoring the importance of a more efficient drug discovery and development process [12]. We aim to reduce the time taken to identify viable new drug candidates by assessing the impact of protein point mutations on stability. This will be done by analyzing the change in Gibbs free energy ($\Delta\Delta$G) between naturally occurring, or wild-type, proteins and their mutated versions, which is a common way to assess stability changes. Our goal is to create an algorithm that predicts changes in Gibbs free energy caused by single-point amino acid mutations, effectively forecasting how specific mutations can alter protein stability.

2 OBJECTIVE & APPROACH

To realize our objective, we plan to harness advanced deep learning techniques, specifically integrating the spatial insights from ThermoMPNN, a cutting-edge graph neural network (GNN), with the evolutionary context captured by Meta's Evolutionary Scale Modeling (ESM-2), a pretrained protein transformer model. This hybrid innovative approach aims to leverage latent transfusion by combining the embeddings of both models to produce a comprehensive and intricate prediction model. We plan to utilize two data sources: FireProtDB, a database offering annotated data on protein stability changes due to mutations, and the RCSB Protein Data Bank (RCSB PDB), a repository providing 3D structures of proteins from experimental findings. We will create a predictive, explainable model that outperforms existing models that mainly depend on a single data source when quantifying the effects of amino acid mutations on protein stability.

Built with Python, HTML, and JavaScript, our end product is a cutting-edge tool designed for researchers and pharmaceutical companies, offering rapid and precise predictions of how mutations impact protein stability. The tool provides an interactive 3D visualization of proteins along with dynamic dashboards, enabling real-time exploration of the effects of mutations. This unique blend of interactive visuals and predictive capabilities is unprecedented, distinguishing our project in the field. In terms of logistics and deadlines, midterm evaluations of success will include successful EDA, at least one model that accurately predicts $\Delta\Delta G$, and sketches of three interactive and user-centric visuals. Final evaluations of success include material improvement upon baseline $\Delta\Delta G$ predictions as well as our finalized interactive visualizations. We aim to spend approximately three months this project, amounting to approximately $120 in total fees using Google Colab.

3 CURRENT PRACTICE

Our project draws upon a diverse range of recent scientific research to explore innovative protein stability prediction techniques. Cao [5], Heyrati [13], Kuhlman [10], Dieckhaus [7], M. Baek [3], Chandra [6] and Wei [8] introduce neural network models for predicting protein characteristics, ranging from stability change to bioactivity and even protein structure. Cao's [5] DeepDDG model, while directly related to our project, is hard to implement given the code isn't open to the public but its DeepDDG server [1] could be used for benchmarking purposes. While Heyrati's [13] study with a sophisticated Siamese neural network allowed us to learn about feature extraction and similarity learning which our algorithms will also be leveraging, their focus on bioactivity classification offered little overlap with ours. Although it didn't examine proteins beyond their structure, we found Kuhlman's [10] combined use of neural networks along with traditional machine learning methods like SVM on embeddings interesting since we'll also be implementing these methods. Although M. Baek's [3] work utilizes amino acids to predict proteins structure, which is valuable for our research, it failed to generalize across diverse protein families. Chandra's [6] use of a Transformer and encoded features to predict protein structure was valuable for our research although the training dataset was limited. Wei's [8] paper offered a comprehensive overview of applying machine learning techniques to materials science, making otherwise time-consuming tasks more efficient, which resonates with our work. However, the models offered little interpretability which is crucial for trust and adoption. J. Baek [2] introduces the Graph Multiset Transformer (GMT) to predict molecular properties. The paper is valuable for us in learning more about the intricate relationships among amino acids, the building blocks of proteins, despite the limited training dataset. M. Baek [3], Maziarka [11], Dieckhaus [7] and Wang [15] extend the research body to graph neural networks (GNNs), which help compress complex protein features while preserving key structural interactions. Maziarka [11] introduces the Molecule Attention Transformer designed for small molecules which would require significant modifications for us given the complexity of protein structures; the paper was still useful to learn of the various Transformer architectures in the domain. While Wang[15] offered a small training dataset, the GNNs employed of protein structures for property prediction offered a useful perspective. Dieckhaus [7] introduces ThermoMPNN, a model combining GNNs and transfer learning, to predict protein stability changes. ThermoMPNN is relevant for enhancing prediction accuracy by learning from extensive protein behaviors, but the model could be simplified for computational efficiency. Baselious [4] and Rives [14] similarly leverage transfer learning to predict protein structure and properties, respectively. Baselious [4] discusses the use of an advanced protein structure prediction algorithm for proteins that play a major role in epigenetic regulation, developing our understanding of the impact of mutations on protein structure and function despite have inconsistencies in accurately predicting the dynamic regions of proteins. Rives [14] introduces ESM-2, a pretrained protein transformer model, which clusters similar proteins. The features generated by ESM-2, while underfitting the data slightly, will assist our model's ability to develop a rich understanding of proteins and can pair with a simplified ThermoMPNN architecture to generate strong protein stability change predictions. By leveraging insights across these studies, we will improve upon existing limitations by focusing on building a precise and understandable model.

4 METHODOLOGY

We leverage a combination of existing and novel techniques to improve upon the state of the art in protein thermostability prediction. We tackle this complex problem in two parts - modeling and visualizations - in order to enhance the drug discovery workflow. The modeling component combines ThermoMPNN, a cutting-edge graph neural network (GNN), with the evolutionary context captured by Meta's Evolutionary Scale Modeling (ESM-2), a pretrained protein transformer model, in order to predict the change in protein stability. The visualization enables researchers to explore protein structures throughout the mutation process along with key metrics related to protein stability. The two analytical facets work together to dramatically reduce the time and cost associated with drug discovery.

Before modeling, we explored the data, starting with a confusion matrix comparing wild-type and mutation cases. From the plot it was observed that the amino acid Valine(V) and Leucine(L) in the wild-type were frequently replaced by Alanine(A) due to mutation. Additionally, we implemented k-means algorithm to both the ThermoNPNN and XGBoost embedding dataset, noting that all data

points fit perfectly into clusters. Prior to k-mean, PCA was implemented to reduce the dimensionality of the data. With a cluster size of 12, all data points were allocated to their appropriate clusters. The Immunoglobulin G-binding protein, G, is one of the most prominent protein in the dataset with the count of 600.

The modeling can be further broken down into two components - the Evolutionary Scale Model and the state of the art message passing neural network, ThermoMPNN. The Evolutionary Scale Model data supplements the FireProtDB with the RCSB Protein Data Bank (RCSB PDB), which includes key information about a protein's structure. This is combined with a BLOSUM and DeMaSk substitution matrices, which quantifies whether amino acid substitutions in a protein chain will increase or decrease stability. Next, Meta's state of the art transformer protein language model, ESM-2, is leveraged to generate features related to protein structure and function. The ESM-2 model generates over 1280 new features localized around the mutation. This data is compatible with and combined with ThermoMPNN to predict the change in Gibbs free energy ($\Delta\Delta G$) due to a single point mutation (i.e. an amino acid substitution). We then extract the feature weights and importance for use later.

The ThermoMPNN model is a message-passing neural network used to predict $\Delta\Delta G$ due to point mutations in a protein. The ThermoMPNN model uses data from FireProtDB, a compilation of protein sequence and structural information. Rather than using the pretrained ThermoMPNN model, we leverage the ESM-2 embeddings to train our own innovative ThermoMPNN-inspired model. More specifically, we combine the ESM-2 embeddings with the ThermoMPNN model embeddings through latent transfusion. Latent Transfusion is a conceptual process where latent features or hidden representations in one model are infused into another model. This technique is used to enhance the second model's ability to learn complex patterns by leveraging the distilled knowledge from the first model. After combining the embeddings by aligning the mutation positions and protein identifiers, we feed them through a one-stage training process, allowing the model to spend more time learning meaningful representations of protein structure and function through the augmented set of embeddings.

The ThermonMPNN modeling is done to enhance predictions of protein thermostability changes, and is supplemented with an innovative web app. AlphaFold2 predicts the 3D structure of proteins with high accuracy based on their amino acid sequence, which is a state-of-the-art artificial intelligence program developed by DeepMind [9]. We utilize AlphaFold2 model to predict the 3D structure of mutated protein. The web app allows users to visualize this protein's 3D structures of wild-type and mutated, and interact with the structure of a protein in 3D, shown below in Figure 1. The protein structure is displayed using a spectrum color scheme, where the sequence is colored in a gradient transitioning from blue at the N-terminus to red at the C-terminus, highlighting the orientation and folding of the protein chain.

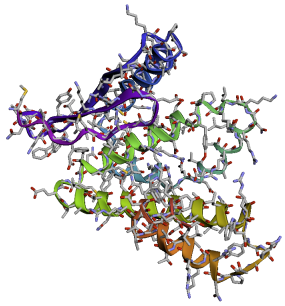

Figure 1: An interactive 3D protein structure

Users will also be able to explore the change in structure and Gibbs free energy due to single point mutations in real-time. Users can select from a wide array of proteins and mutations to find the best combinations based on their particular use cases.

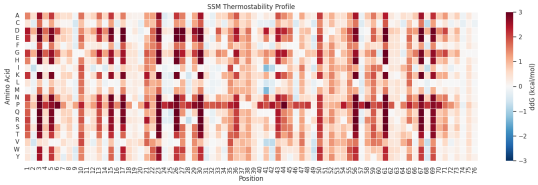

Figure 2: Predicted $\Delta\Delta G$ matrix by mutation type and position

Figure 2 illustrates a matrix summarizing the predicted $\Delta\Delta G$ due to a substitution of the amino acid at position i with the amino acid in column j, with darker shades of red indicating a larger increase than lighter shades.

In summary, our approach introduces the following innovative techniques:

(1) Latent transfusion of embeddings from two SOTA protein models - ESM-2 and ThermoMPNN

(2) Intuitive and interactive web app detailing protein structure and thermostability changes due to single point mutations

5 EXPERIMENTS

First, we trained a version of ThermoMPNN with feature importances extracted from XGBoost model using ESM-2; results can be found in Table 1. Next, we developed the latent transfusion of embeddings of ESM-2 and combined them with ThermoMPNN. Model training has yet to be completed. Then we will be augmenting our dataset with additional input data, such as the Megascale protein dataset, which will take further testing and validation. Finally, we will be exploring further model architecture fine-tuning and development to ensure we are able to achieve optimal results, reducing the mean squared error and increasing the r-squared. Interpretability of deep learning models is always a challenge, which we will seek to address through ablation studies to determine the features that were most important in $\Delta\Delta G$ predictions. Lastly, we will seek to address the overfitting problem experienced by both models by introducing regularization.

Table 1: Evaluation results

$\Delta\Delta G$ prediction metrics

	Metric			
	ThermoMPNN		ThermoMPNN+	
	Train	*Val*	*Train*	*Val*
MSE	0.9935	2.1946	0.7702	2.0877
R^2	0.6798	0.1452	0.7517	0.1868

6 PLAN OF ACTIVITIES

Original

Tasks	Start	Duration	Members
Research	1/17/24	3 weeks	KT, MD
Planning	2/7/24	1 week	AP, JM
EDA	2/7/24	3 weeks	MD, AP
Modeling	2/28/24	4 weeks	KT, JM
UI/Visuals	3/27/24	3 weeks	AP, JM

Revised

Tasks	Start	Duration	Members
Research	1/17/24	3 weeks	KT, MD
Planning	2/7/24	1 week	AP, JM
EDA	2/7/24	3 weeks	MD, AP
Modeling	2/28/24	5 weeks	KT
UI/Visuals	3/19/24	4 weeks	AP, JM, MD

All team members have contributed a similar amount of effort.

7 CONCLUSIONS

We set out with the goal of improving the costly and time-consuming drug discovery workflow. To do so, we focused on accurately predicting protein thermostability changes due to single-point mutations. Accurate prediction of protein thermostability after a mutation through the change in Gibbs free energy is key in predicting a new drug's potential efficacy. To tackle this problem, we developed a novel deep learning model, ThermoMPNN+, which materially improves upon a state-of-the-art ThermoMPNN model by incorporating features from a robust protein transformer model used to predict $\Delta\Delta G$ due to protein mutations. We also built an intuitive UI that lets users interact with 3D protein structures and explore mutations' impacts to thermostability in real-time. With these new tools in hand, researchers can expect a dramatic reduction in time spent evaluating proteins for new drugs as well as the cost associated with drug discovery R&D. This will culminate in faster iteration through the drug discovery process and ultimately to better outcomes of new drugs developed with the assistance of our tools.

Our future work will include improving the ThermoMPNN+ model with a more developed architecture, training it on a larger, integrated dataset and fine-tuning hyperparameters. We will perform ablation studies to better grasp feature importance, and continue improving the web app with a focus on user experience and utility.

REFERENCES

[1] Deepddg server. https://protein.org.cn/ddg.html. Accessed: [2024-02-29]. Access to the server, not a paper.

[2] Jinheon Baek, Minki Kang, and Sung Ju Hwang. Accurate learning of graph representations with graph multiset pooling. 2021.

[3] Minkyung Baek, Frank DiMaio, and Ivan et al. Anishchenko. Accurate prediction of protein structures and interactions using a 3-track neural network. *Science*, 373:871–876, 2022.

[4] Fady Baselious, Dina Robaa, and Wolfgang Sippl. Utilization of alphafold models for drug discovery: Feasibility and challenges. histone deacetylase 11 as a case study. *Computers in Biology and Medicine*, 167:107700, 2023.

[5] Huali Cao, Jingxue Wang, Liping He, Yifei Qi, and John Z Zhang. Deepddg: predicting the stability change of protein point mutations using neural networks. *Journal of chemical information and modeling*, 59(4):1508–1514, 2019.

[6] Abel Chandra, Laura Tunnermann, Tommy Lofstedt, and Regina Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *National Library of Medicine*, 12, 2023.

[7] Henry Dieckhaus, Michael Brocidiacono, Nicholas Z. Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121(6):e2314853121, 2024.

[8] Xiang-Yu Sun Kun Xu Hui-Xiong Deng Jigen Chen Zhongming Wei Ming Lei Jing Wei, Xuan Chu. Machine learning in materials science. pages 338–358, 2019.

[9] John Jumper, Richard Evans, Alexander Pritzel, and Tim et al. Green. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

[10] Bradley Kuhlman, B. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20:681–697, 2019.

[11] Ukasz Maziarka and Dawid Majchrowski. Relative molecule self-attention transformer. *Journal of Cheminformatics*, 16(1):3, 2024.

[12] Matej Mikulic. Rd spending as revenue share of leading 10 pharmaceutical companies in 2022, 2023. Extra paper but not peer-reviewed.

[13] Mehdi Paykan Heyrati, Zahra Ghorbanali, Mohammad Akbari, Ghasem Pishgahi, and Fatemeh Zare-Mirakabad. Bioact-het: A heterogeneous siamese neural network for bioactivity prediction using novel bioactivity representation. *ACS Omega*, 8(47):44757–44772, 2023.

[14] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[15] Shuyu Wang, Hongzhou Tang, Yuliang Zhao, and Lei Zuo. Bayestab: Predicting effects of mutations on protein stability with uncertainty quantification. *Protein Science*, 31(11):e4467, 2022.