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

Team 4 | AlgoRxplorers

Karishma Thakrar, Jiangqin Ma, Max Diamond, Akash Patel

1 OBJECTIVE & APPROACH

Alterations in a protein's sequence can significantly change its structure and functional, influencing protein stability, a key factor in health and disease. 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. Analyzing the change in Gibbs free energy (ΔΔG) between naturally occurring, or wild-type, proteins and their mutated versions 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.

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), 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.

2 CURRENT PRACTICE

Our project draws upon a diverse range of recent scientific research to explore innovative protein stability prediction techniques. Cao [5], Heyrati [12], Kuhlman [9], 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 [12] 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 [9] 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 [10], Dieckhaus [7] and Wang [14] extend the research body to graph neural networks (GNNs), which help compress complex protein features while preserving key structural interactions. Maziarka [10] 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[14] 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 [13] 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 [13] introduces ESM, a pretrained protein transformer model, which clusters similar proteins. The features generated by ESM, 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.

3 CHALLENGES AND IMPACT

Many drugs target proteins to modulate their activity, so understanding how mutations affect protein stability can help in designing more effective drugs that bind more efficiently to their targets [3]. Also, understanding mutations that reduce protein stability is crucial for comprehending disease mechanisms and developing new protein therapies. As such, medical professionals, patients and pharmaceutical company would benefit from this research. Research shows in 2022 alone, pharmaceutical companies have spent close to $244 billion on R&D, underscoring the importance of more efficient drug development processes [11]. Successfully predicting enzyme thermal stability will significantly reduce the time taken to discover new drugs. The impact can be measured analytically using the Root Mean Square Error (RMSE). In this case, we will evaluate the RMSE between the experimentally measured $\Delta\Delta G$ values (the observed changes in protein stability due to mutations) and the $\Delta\Delta G$ values predicted by our model. We plan to evaluate our tool's usability through user studies with a diverse audience to gain valuable insights, even if direct access to target users like drug development professionals and bioinformaticians is limited. A key challenge in performing latent transfusion is the input of structural protein data from our data source Protein Data Bank (PDB) and retrieval of embeddings from two state-of-the-art models, which requires training both models. This process demands careful research, extensive computational resources, and time due to the need to develop and train two complex models in order to generate the necessary embeddings for infusion. Risks encompass the reliance on current, high-quality data sources, ensuring the model's predictions apply to new compounds, and integrating our UI tool into existing drug development workflows. However, the benefits are substantial, including more accurate predictions of protein stability changes, which lead to reduced drug discovery costs and faster development of new drugs.

4 WORK DIVISION

Tasks	Start Date	Duration	Contribution Members
Research	1/17/24	3 weeks	K. Thakrar, M. Diamond
Planning	2/7/24	1 week	A. Patel, J. Ma
EDA	2/7/24	3 weeks	M. Diamond, A. Patel
Modeling	2/28/24	4 weeks	K. Thakrar, J. Ma
UI/Visuals	3/27/24	3 weeks	A. Patel, J. Ma

All team members have contributed a similar amount of effort.

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

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

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

[12] 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.

[13] 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.

[14] 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.