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

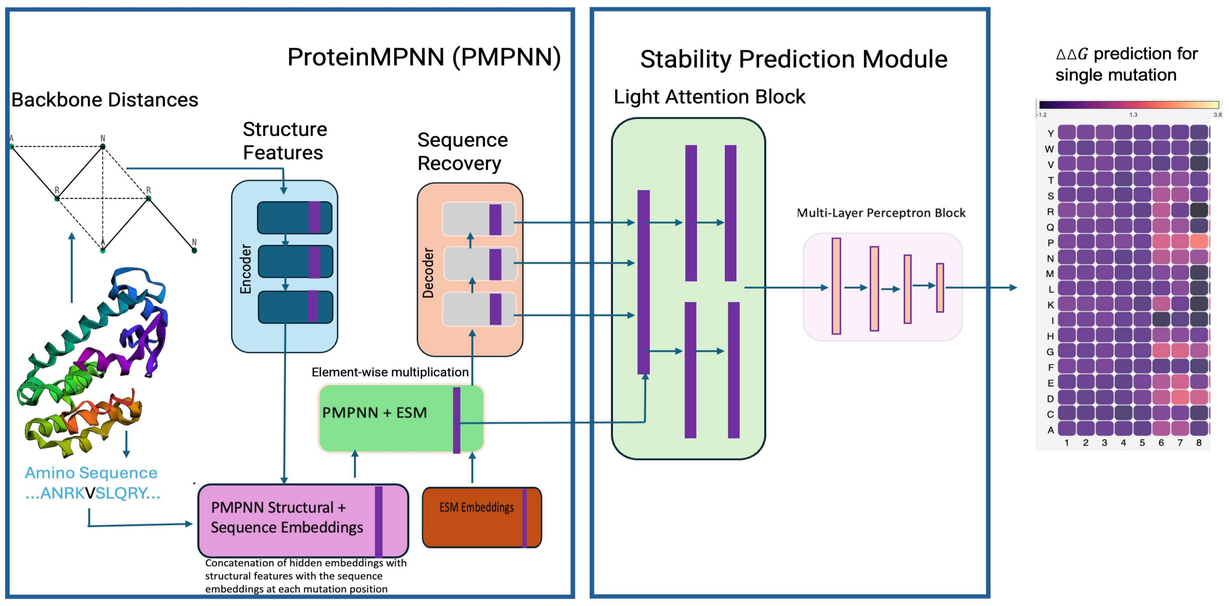
Team 4 | AlgoRxplorers
Karishma Thakrar, Jiangqin Ma, Max Diamond, Akash Patel

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. 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, pharmaceutical companies spent nearly $244 billion on R&D, underscoring the importance of a more efficient drug discovery and development process. 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

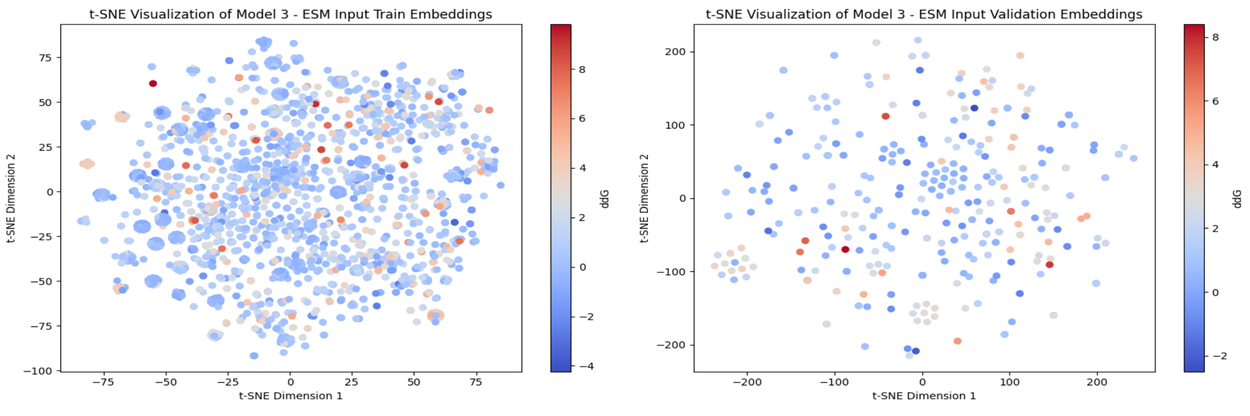
Approaches

Our research builds upon ThermoMPNN, a graph neural network (GNN) model that utilizes transfer learning to extract structural features and sequence embeddings from the ProteinMPNN model. We build four models based on ThermoMPNN. Our best model ThermoMPNN+, combines structural features and sequence embeddings from the ProteinMPNN model with ESM embeddings through element-wise multiplication, capturing interactions between corresponding feature embeddings. We use the AlphaFold2 algorithm to predict protein 3d structure to understand protein better.



Model	Inputs
Model 1 (ThermoMPNN)	Original ThermoMPNN model inputs: FireProtDB data with duplicates removed
Model 2	ESM embeddings and trained ThermoMPNN embeddings
Model 3 (ThermoMPNN+)	**ThermoMPNN inputs (structural features and sequence embeddings from ProteinMPNN) combined with ESM embeddings**
Model 4	Trained embeddings generated using domain-specific knowledge combined with ThermoMPNN trained embeddings

ESM embeddings are integral to our ThermoMPNN models, enriching them with a nuanced understanding of protein behavior that drives predictive performance.



Data

We use FireProtDB, which is downloaded from the website. It is a database with annotated protein stability changes caused by mutations and 3D protein structures from experiments. It includes the following data points for each protein-mutation pair:

Datapoint	Description
PDB	PDB ID to uniquely identify protein
Wildtype	Amino acid before mutation
Position	Position number where mutation took place
Mutation	Amino acid after mutation
$\Delta\Delta G$	Gibbs free energy
Sequence	Protein Sequence
Mutant_Seq	Mutated protein sequence
Coordinates	3-D coordinates of the backbone atoms, N, Cα, C, and O

Experiments and Results

The experiment is designed to determine the most accurate method of predicting the impact of amino acid changes on protein stability by integrating information from different models. Results show that the enhanced ThermoMPNN+ model performs best, indicating improved predictive power. Upon visualizing embeddings from ThermoMPNN+ model, we discovered that transforming the problem into a classification task yields to even more promising results.

Model	MSE	R^2	Spearman
Model 1 (ThermoMPNN)	2.1946	0.168	0.542
Model 2	2.397	0.064	0.387
Model 3 (ThermoMPNN+)	**2.057**	**0.196**	**0.552**
Model 4	2.490	0.027	0.354

ThermoMPNN+ excels at classifying mutation impacts, achieving 1.0 recall and 0.924 F1 score, outperforming ThermoMPNN. Its higher F1 indicates better class separability in embeddings for assessing stabilizing vs. destabilizing mutations.

Model	Acc.	Prec.	Recall	F1
ThermoMPNN	0.770	0.876	0.889	0.882
ThermoMPNN+	**0.859**	**0.859**	**1.000**	**0.924**

Conclusion

Our novel deep learning model, ThermoMPNN+, improves upon the state-of-the-art by incorporating features from a robust protein transformer model. We utilize the AlphaFold2 algorithm to accurately predict how proteins will look in 3D based on the order of their building blocks, which are amino acids.