Predicting Protein Stability Changes from Mutations via Deep Sequence and Structure Learning

Valentin Badea, Shyam Chandra, John Lin, Seshagiri (Seshu) Rao Mallina

# Abstract

The goal of our project is to predict variations in protein stability (measured as change in Gibbs free energy, or ) induced by single amino-acid mutations. We used data in the format of unique wild-type/mutated protein pairs from FireProtDB. Upon completion of the preprocessing step, we consider 4291 such pairs, spanning 141 unique proteins. Then, we tried several approaches to this regression task, including fully-sequence and fully-structure-based models. We encoded information from protein sequence inputs using ESM2, a Large Protein Language Model. Additionally, we used Graph Neural Networks to encode 3D protein structures. Overall, our models learn useful latent representations of changes between wild-type and mutants for both protein structural or sequence embeddings. These representations of the mutations are then linked to single predictions through a regression head. Our best performing model leveraged both structure and sequence latent representations of the protein mutation through a concatenation layer and achieved significantly better results than the two single-data-modality models. In conclusion, we show that a model predicting mutation-induced changes in protein stability can benefit from two complementary latent representations of the mutation through dedicated sequence and structure embeddings.

# Introduction

## Background

Single-point mutations can cause changes in protein stability by affecting the three-dimensional protein structure. , or the change in Gibbs free energy between wild-type and mutant protein, is a metric that can be used to characterize how a single-point mutation affects protein stability. Accurate calculation of can therefore facilitate the development and optimization of stable proteins for various applications, including drug discovery.1,2 Currently, is measured experimentally, but this process is often laborious, time-consuming, and low-throughput. Machine learning models have previously been developed to predict changes in protein stability, although performance was limited due to the use of manually calculated features.3,4

Recent deep learning approaches typically learn embeddings from either protein sequences5 or 3D structures6 to predict upon single-point mutations. Large language model (LLM)-based approaches utilize embeddings from protein language models and have achieved a root mean squared error (RMSE) value of 1.46 kcal/mol from the S669 dataset.7 In contrast, graph neural network (GNN)-based methods extract structure-based features and have reached an RMSE value of 1.19 kcal/mol in the S2648 dataset.8 However, neither of these methods simultaneously consider both sequence and structure information, and few models explicitly model both wild-type and mutant proteins. We hypothesize that integrating sequence- and structure-based embeddings may improve predictive performance capturing both evolutionary and structural contexts. Here, we develop a deep learning approach that combines information from protein sequences and 3D structures to predict changes in protein stability () due to single-point mutations.

## Research Question

Can we accurately predict changes in protein stability () upon single-point mutations by integrating information from protein sequence and 3D structure using transformers and graph neural networks?

# Methods

## Dataset

Protein sequence and stability data were downloaded as a CSV file from the FireProtDB website (<https://loschmidt.chemi.muni.cz/fireprotdb/>).9 FireProtDB is a curated database of experimentally validated values associated with single-point mutations, and it provides open-source access for academic research purposes. The full dataset includes 211 proteins, 6,375 single-point mutations, and 53,445 values. We then performed the following preprocessing steps. First, we removed all duplicate experiments (identified by experiment IDs). Then, we removed all entries with missing values, UniProt IDs, or protein names. We observed that some single-point mutations were assayed several times under different experimental conditions (e.g. temperature or pH), resulting in multiple values. To address this, we calculated the median value for each mutation. We also excluded all proteins that are longer than 1,024 amino acids to decrease computational cost. Finally, we removed all entries for which we were unable to retrieve the PDB structure for the wild-type protein. After preprocessing, the final dataset includes 141 proteins, 4,291 single-point mutations, and 4,291 values.

Wild-type protein structures were retrieved as PDB files from the AlphaFold Protein Structure Database, which provides full-length structural predictions for proteins based on UniProt IDs.10,11 To model the structural effects of single-point mutations, we generated mutant protein structures using PyRosetta, a Python-enabled interface to the Rosetta macromolecular modeling suite.12 Briefly, for each single-point mutation, we loaded the corresponding wild-type PDB file into a pose object. Then, we replaced the residue at a single position in the pose object with a new amino acid and repacked any residues within 8.0 Å of the mutated residue’s center. Lastly, we minimized the pose to a local energy minimum by modifying the configuration of the backbone and side-chain atoms and saved this pose as a PDB file. We repeated this process for every unique single-point mutation in the filtered dataset to obtain paired wild-type and mutant protein structures for downstream modeling applications.

## Deep Learning Approach

We constructed three deep learning models: 1) A sequence-based model that uses a pre-trained protein language model with frozen weights, 2) A structure-based model that uses a GNN architecture, and 3) A multimodal model that integrates sequence- and structure-based embeddings (**Figure 1**). Each model predicts a scalar value given a wild-type protein and a single-point mutation. We split the filtered dataset into train (80%), validation (10%), and test (10%) splits. Training was carried out to minimize the RMSE between the predicted and actual values. Hyperparameters for all models were set at 10 epochs, batch size 16, and learning rate 10-4. A random seed was set at 42 to ensure reproducibility.

**Sequence Model.** First, we used a pre-trained protein language model (ESM2) with frozen weights to create contextual residue-level embeddings for the wild-type and mutant sequences.13 Then, we calculated the element-wise difference between the residue-level embeddings for the wild-type and mutant sequences to encode mutation-induced change in protein embeddings. This difference is processed by a 1D convolution layer to reduce dimensionality, passed through a transformer encoder block (with a multi head self attention layer, a position-feed-forward layer and some skip-connections), and summarized with both global max and average pooling layers. Finally, a fully-connected linear layer predicts a value.

**Structure Model.** We retrieved wild-type and mutant structures from the AlphaFold Protein Structure Database and PyRosetta, respectively. To convert protein structures into graph representations, each structure (wild-type or mutant) was separately parsed to extract the coordinates of the alpha carbon atoms. Residues were represented as nodes with one-hot encoded features corresponding to the standard 20 amino acids. Edges were constructed between residue pairs if the Euclidean distance between the alpha carbon atoms was below a recommended threshold of 8.0 Å, which encodes local spatial proximity and residue-level interactions relevant to structural stability.14 The resulting graphs include: (1) node features containing 20-dimensional one-hot vectors encoding residue identity, (2) edge indices describing residue pairs within a distance threshold, and (3) node positions storing the 3D coordinates of alpha carbon atoms to guide GNN message passing. Next, we used a GNN to create separate embeddings for the wild-type and mutant proteins and calculated the element-wise difference between the two embeddings. This difference is then processed by a fully-connected linear layer to predict a value. In contrast to the sequence-based model, the weights of the structure-based model’s GNN are learned during training.

**Sequence-Structure Fusion Model.** Our fusion model integrates sequence and structure representations to predict . For each mutation, we computed wild-type and mutant sequence embeddings using the pretrained ESM2 model with frozen weights and wild-type and mutant structure embeddings using the same GNN from our structure based model. We calculated the element-wise difference between wild-type and mutant embeddings for each modality, then concatenated the sequence and structure difference vectors to form a joint representation. This fused embedding was passed through a 1D-convolution layer, transformer block and fully-connected linear layer to predict . During training, the ESM2 weights remained fixed to preserve the evolutionary information encoded by the pretrained protein language model. On the other hand, the GNN’s, transformer block’s and regression head’s weights were updated via backpropagation, allowing the model to adapt structural embeddings to the task of predicting .

## Evaluation metrics

Evaluation metrics include Root Mean Squared Error (RMSE) and Pearson correlation coefficient (PCC).

# Results

## Exploratory Data Analysis

After preprocessing, the final dataset includes 141 proteins, 4,291 single-point mutations, and 4,291 values. **Figure 2** shows the distribution of values **(a)** and protein lengths **(b)** in the final dataset. The median value is 0.66, and the median protein length is 208 amino acids. The top 5 proteins represented in the dataset include Immunoglobulin G-binding protein G, Thermonuclease, Lysozyme C, Ribonuclease, Dihydrofolate reductase which collectively account for ~46% of entries in our final filtered dataset.

## Model Performance

We trained a sequence-based model, a structure-based model, and a fusion model as detailed above. **Figure 3** shows RMSE between predicted and actual value from the training and validation datasets for each model over 10 epochs. For the sequence-based model, the training and validation RMSE at the 10th epoch are 1.34 and 1.54, respectively. For the structure-based model, the training and validation RSME at the 10th epoch are 1.37 and 1.52, respectively. For the fusion model, the training and validation RMSE at the 10th epoch are 0.98 and 1.39, respectively. The Pearson correlation coefficient for the sequence-based model, structure-based model, and fusion model are 0.61, 0.53, and 0.68, respectively (**Figure 4**). On the held-out test dataset, our fusion model performed the best with an RMSE of 1.40, while the sequence-only model has an RMSE of 1.47 and the structure-only model has an RMSE of 1.58.

## Comparison with Non-Deep Learning Approaches

We evaluated three non-deep learning baselines on the held-out test set. First, we built a dummy regressor that predicts the mean from the training set for all samples regardless of input. This serves as a sanity check and naive baseline. Second, we build a linear regression model that learns a linear relationship between a small set of handcrafted features and . Finally, we build a gradient boosted tree model from the same handcrafted features to predict . This tree model works by building a series of shallow decision trees, each trained to correct the residuals of the previous. This allows it to capture nonlinear relationships between features and outcomes.

To make the baselines more relevant, we incorporated biologically-meaningful features. This includes a BLOSUM62 substitution score matrix, which reflects evolutionary tolerance to specific amino acid changes. We also used the Kyte-Doolittle scale to capture hydropathy differences from shifts in amino acid polarity that can affect protein folding and stability. Finally, we used relative mutation positions within the sequence.

**Figure 4** visualizes the predicted vs true values for each model. The dummy and linear models fail to capture any variation, producing relatively flat predictions clustered around the mean. The gradient boosted tree approach slightly improves performance by capturing some nonlinear trends, but remains limited by minimal input features. As shown in **Table 1**, the gradient boosted tree model outperformed the dummy and linear models with a test RMSE of 1.63 and a PCC of 0.48. Overall, however, all baseline models significantly underperformed compared to the deep learning models. This suggests the value of learned representations from more sophisticated structure and sequence protein representations.

## 

## Error Analysis

As shown in **Figure 4**, our model tends to underperform on extreme values (both highly stabilizing and highly destabilizing mutations). This is evident in the consistent misprediction of outlier cases near the upper and lower bounds of the range regardless of model type. Incorporating more examples of extreme values or applying loss reweighting might help mitigate this.

To better understand where the fusion model underperforms, we analyzed prediction errors using absolute error. While RMSE is suitable for aggregate performance, absolute error is more interpretable at the sample level and allows for identification of outliers. We selected a threshold of 2.0 kcal/mol, which would correspond to a large enough error to misclassify a stabilizing vs destabilizing mutation. Out of all test samples, 52 were identified as having errors above this threshold. We then plotted absolute error against protein length (**Figure 5**). We also performed statistical analysis using a Mann-Whitney U test, which revealed no significant difference in protein lengths between high and low-error groups (). These findings suggest that protein size is not a primary driver of error in the fusion model as we hypothesized. Other factors could involve structural irregularities, limited training data, or complex mutation contexts.

# Discussion

## Interpretation & Implications

The combined structure and sequence model outperformed all other models, including the baseline, sequence-only, and structure-only models on the held-out test dataset. This model achieved lower RMSE and higher PCC, indicating that predicted values were more similar to actual values of . Our findings suggest that combining structure and evolutionary history representations allow the model to reach more accurate and generalizable results to predict protein stability. Although sequence-based models capture evolutionary features and functional motifs, incorporating structural information can provide insights to spatial context and interactions between 3D amino acid structures. Thus, our results validate incorporating structural and sequence data allowing models to learn richer representations and fine grained information. This promotes future work to integrate structural and sequential information in tasks relating to predicting protein attributes. By predicting protein stability more accurately, we can facilitate the design of novel proteins with stronger therapeutic potential.

## Limitations

One limitation of our study is the small sample size of our dataset. After preprocessing, our dataset includes only 141 proteins and 4,297 single-point mutations. This may limit the generalizability of our models. Our dataset is also biased toward destabilizing mutation (). To balance the dataset, we can reverse a subset of variations to switch the sign of experimental values.

Another limitation of our current approach is that we consider only single-point mutations during training. This model is therefore unlikely to be able to predict *ΔΔG* upon multi-point variations. We can consider the PTmul dataset, which includes 91 proteins and 914 multi-point mutations.4 Our model also does not consider non-point mutations such as indels. This is because we calculated the element-wise difference between the wild-type and mutant embeddings, which must share the same length.

## Challenges

While training the structure-based model, we ran into OutOfMemory errors. We found that, by default, the tokenizer from ESM2 pads the input to a maximum length that is extremely large. We resolved the memory issue by setting the maximum length to 1024, which retained approximately 80% of the data while improving computation times.

Moreover, the FireProtDB database required preprocessing since it included duplicate entries, inconsistent annotations, and missing values. Unfortunately, there was also limited documentation for FireProtDB, leading us to spend a lot of time on data cleaning. Furthermore, although associated PDB IDs were included, FireProtDB does not specify whether they correspond to wild-type or mutant forms, and many structures are fragments rather than full-length proteins. As a result, we opted not to use these PDB IDs and instead relied on full-length structural predictions from the AlphaFold Protein Structure Database for the wild-type proteins.

Finally, the Rosetta-based mutant PDB generation was time-consuming. To manage this, we split the dataset into subsets and ran multiple jobs in separate notebooks. This reduced runtime and allowed us to progress, but required manual dataset stitching and duplicate removal to ensure all generated structures were included in the deep learning models.

## Future Directions

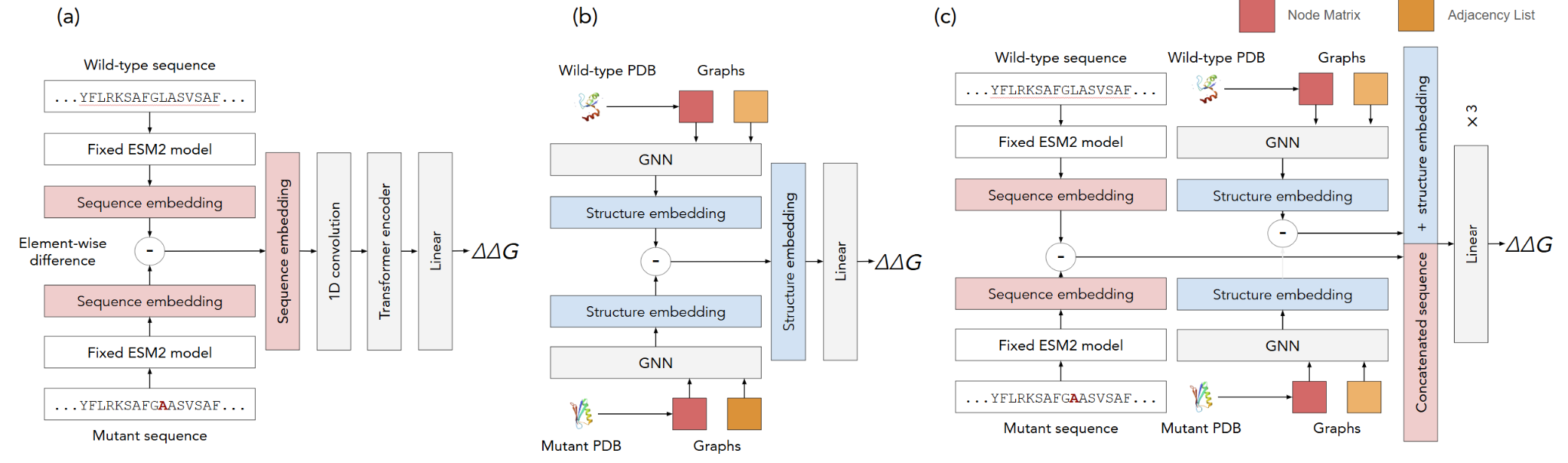
We will continue to optimize the architecture of our deep learning models. Other ongoing work includes hyperparameter tuning and weight initialization optimization. To add on, we want to test the three models that we built on new unseen datasets, such as S26484 and S66915, to see if our models are generalizable. The S2648 dataset includes 132 proteins and 2,648 single-point mutations (with experimentally validated values), whereas the S669 dataset includes 87 proteins and 669 single-point mutations (with experimentally validated values).

Some wild-type sequences in our dataset were not available in the AlphaFold database. Discarding a single wild-type protein would also eliminate all associated mutations, reducing our dataset size substantially. To address this, we generated predicted wild-type structures from the amino acid sequence using AlphaFold2 locally, and then used these structures as templates for predicted mutant PDBs via PyRosetta. This strategy minimizes data loss but introduces uncertainty by compounding any prediction error from AlphaFold into subsequent modeling steps. Future directions would involve incorporating these predicted structures into the final dataset.

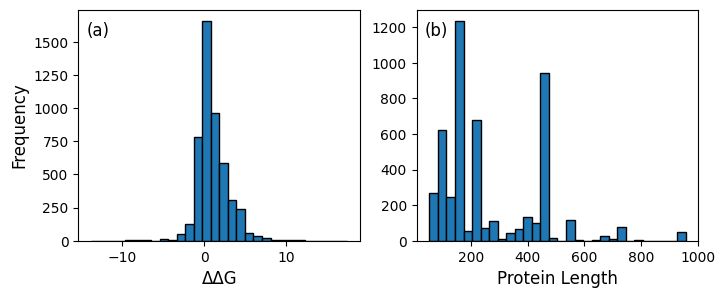
# Appendix: Tables and Figures

| **Model** | **RMSE** | **PCC** |
| --- | --- | --- |
| Dummy (mean) | 1.85 | - |
| Linear regression | 1.75 | 0.32 |
| Gradient boosted tree | 1.63 | 0.48 |
| Sequence | 1.47 | 0.61 |
| Structure | 1.59 | 0.53 |
| Fusion | **1.40** | **0.66** |

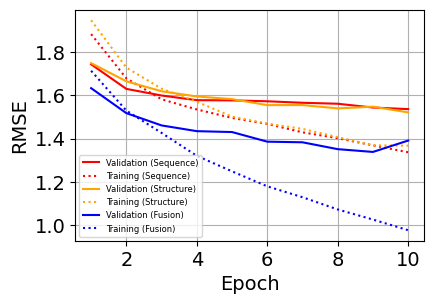
**Table 1.** Model performance on held-out test dataset. PCC is undefined for the dummy model due to zero variance. RMSE: Root mean square error, PCC: Pearson correlation coefficient



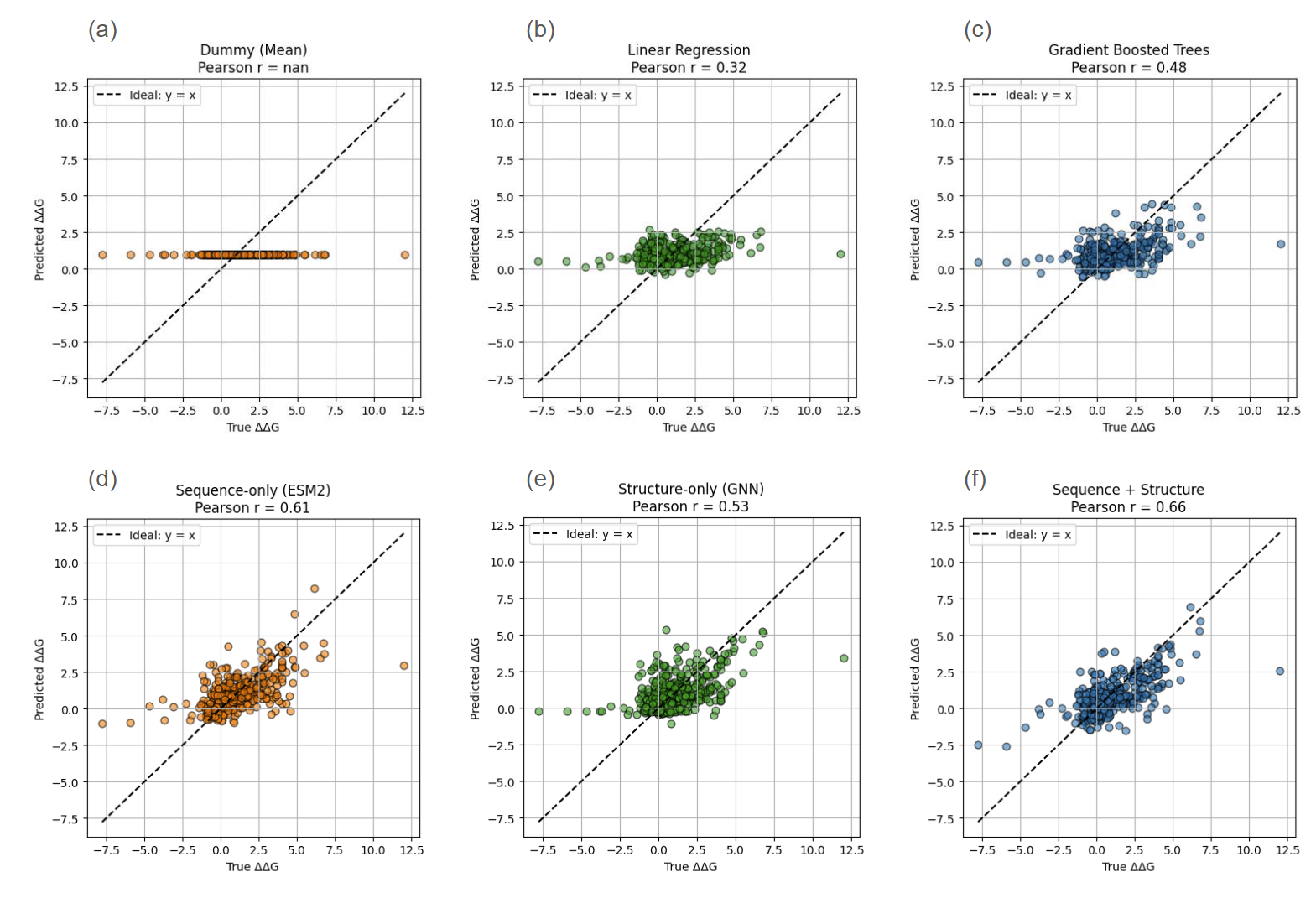
**Figure 1.** Model architecture of sequence (a), structure (b), and fusion (c) models.



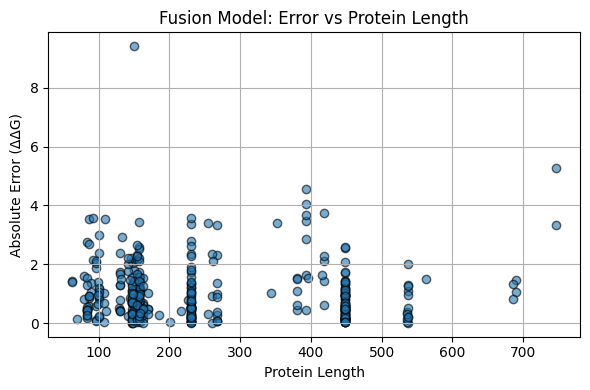
**Figure 2.** Histogram of values (a) and protein lengths (b) in FireProtDB



**Figure 3.** Model performance on training and validation datasets. RMSE: Root mean square error



**Figure 4.** Correlation between predicted vs actual values on held-out test dataset for dummy (a), linear regression (b), gradient boosted tree (c), sequence (d), structure (e), and fusion (f) models. The first row represents the baseline models and the second row involves the three deep learning models that were developed. PCC: Pearson correlation coefficient



**Figure 5.** Distribution of absolute error across protein length (number of residues) for fusion model on held-out test dataset.

# Group Member Contributions

*Valentin Badea*

Three models’ architectures, Sequence-based model development, Fusion-model development, Data Processing

*Shyam Chandra*

Fusion-model development, dataset aggregation, training and fine-tuning final models, baseline-model development, final report drafting

*John Lin*

Data preprocessing, PyRosetta pipeline, data analysis, final report drafting, poster development

*Seshu Mallina*

Seshu contributed to the project by creating the pipeline to create the graph structures for wild-type and mutant sequences, implementing the structure-based model, creating a held out test dataset, and assisted in writing the final paper report.

# References

1. Goldenzweig, A., & Fleishman, S. J. (2018). Principles of Protein Stability and Their Application in Computational Design. *Annual Review of Biochemistry*, *87*, 105–129. https://doi.org/10.1146/annurev-biochem-062917-012102

2. Notin, P., Rollins, N., Gal, Y., Sander, C., & Marks, D. (2024). Machine learning for functional protein design. *Nature Biotechnology*, *42*(2), 216–228. https://doi.org/10.1038/s41587-024-02127-0

3. Teng, S., Srivastava, A. K., & Wang, L. (2010). Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics*, *11*(2), S5. https://doi.org/10.1186/1471-2164-11-S2-S5

4. Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N., & Fariselli, P. (2019). DDGun: An untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics*, *20*(Suppl 14), 335. https://doi.org/10.1186/s12859-019-2923-1

5. Savojardo, C., Manfredi, M., Martelli, P. L., & Casadio, R. (2025). DDGemb: Predicting protein stability change upon single- and multi-point variations with embeddings and deep learning. *Bioinformatics*, *41*(1), btaf019. https://doi.org/10.1093/bioinformatics/btaf019

6. Wang, S., Tang, H., Shan, P., Wu, Z., & Zuo, L. (2023). ProS-GNN: Predicting effects of mutations on protein stability using graph neural networks. *Computational Biology and Chemistry*, *107*, 107952. https://doi.org/10.1016/j.compbiolchem.2023.107952

7. Zhang, Y., Gao, Z., Tan, C., & Li, S. Z. (2023). *Efficiently Predicting Protein Stability Changes Upon Single-point Mutation with Large Language Models* (No. arXiv:2312.04019). arXiv. https://doi.org/10.48550/arXiv.2312.04019

8. Wang, S., Tang, H., Zhao, Y., & Zuo, L. (2022). BayeStab: Predicting effects of mutations on protein stability with uncertainty quantification. *Protein Science : A Publication of the Protein Society*, *31*(11), e4467. https://doi.org/10.1002/pro.4467

9. Stourac, J., Dubrava, J., Musil, M., Horackova, J., Damborsky, J., Mazurenko, S., & Bednar, D. (2021). FireProtDB: Database of manually curated protein stability data. *Nucleic Acids Research*, *49*(D1), D319–D324. https://doi.org/10.1093/nar/gkaa981

10. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

11. Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., … Velankar, S. (2024). AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, *52*(D1), D368–D375. https://doi.org/10.1093/nar/gkad1011

12. Chaudhury, S., Lyskov, S., & Gray, J. J. (2010). PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics (Oxford, England)*, *26*(5), 689–691. https://doi.org/10.1093/bioinformatics/btq007

13. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2020). *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences* (p. 622803). bioRxiv. https://doi.org/10.1101/622803

14. Pietal, M. J., Bujnicki, J. M., & Kozlowski, L. P. (2015). GDFuzz3D: A method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. *Bioinformatics*, *31*(21), 3499–3505. https://doi.org/10.1093/bioinformatics/btv390

15. Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., Capriotti, E., & Fariselli, P. (2022). Predicting protein stability changes upon single-point mutation: A thorough comparison of the available tools on a new dataset. *Briefings in Bioinformatics*, *23*(2), bbab555. https://doi.org/10.1093/bib/bbab555