BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Protein Stability Prediction Using Fine-Tuned Protein Language Models

Jakub Vlk*

Abstract

Protein stability is a fundamental physical property determining a protein's resistance to environmental factors. However, experimental measurement of stability remains resource-intensive, prone to high variance, and inherently slow. This creates a critical need for computational tools capable of predicting changes in stability ($\Delta\Delta G$) following mutations, rather than focusing on absolute stability values. A significant bottleneck in this field is the scarcity of large-scale datasets containing direct $\Delta\Delta G$ measurements. This work addresses this limitation by fine-tuning Protein Language Models (PLMs) on a massive, curated dataset of 864,033 mutation records, utilizing normalized proxy values to overcome data sparsity. Our ProtBERT-based model achieved a Pearson correlation of 0.57 on an independent test set, outperforming ESM-2-based architectures. These results demonstrate that fine-tuning PLMs on large-scale, silver-standard data is a highly effective strategy for scalable prediction of mutational effects, providing a robust tool for bioinformatics and therapeutic design.

*xvlkja07@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Proteins are fundamental to life, and their stability is essential for proper function. Single-point mutations can drastically alter stability, leading to dysfunction or, in rare cases, functional improvements. Measuring protein stability or stability changes experimentally in the wet lab (e.g., via traditional methods or Deep Mutational Scanning) is resource-intensive, creating a need for accurate computational predictors to accelerate research in biomedicine and biotechnology.

From a machine learning perspective, protein stability change prediction can be framed as a regression problem. A robust solution must generalize across diverse protein families, where stability is governed by many different physical, chemical, and quantum phenomena. While these phenomena can be simulated, doing so for proteins comprising thousands of amino acids quickly becomes computationally prohibitive, even on supercomputers.

Traditional computational methods include physics-based tools like FoldX [1] and Rosetta [2, 3], and more recent machine learning approaches. State-of-the-art increasingly leverages Protein Language Models (PLMs) pre-trained on vast sequence databases. Models like ThermoMPNN [4] show strong performance, but the optimal architecture and training strategy for stability prediction remain active research areas.

We propose fine-tuning two prominent PLM architectures—ProtBERT [5] and ESM-2 [6] on a newly constructed, mega-scale dataset. The ProtBERT model is adapted to natively compare wild-type and mutant sequences within a single context window, while the ESM-2 model processes sequences separately with shared weights. Both are equipped with custom regression heads and trained using advanced optimization techniques.

The main contributions are: (1) The creation and publication of a large, curated dataset for stability prediction, merging and normalizing data from multiple sources . (2) The implementation and comparative evaluation of two fine-tuned PLM-based architectures. (3) Empirical demonstration that the ProtBERT-based model, due to its inherent paired-sequence processing, achieves superior predictive performance on independent benchmark dataset.

2. Dataset Construction

A high-quality dataset is foundational. We merged two major experimental sources: the Megascale and Human Domainome databases, resulting in 864,033 point mutation records. To harmonize different experimental

readouts, a non-linear, double-sided sigmoid normalization was applied to project all $\Delta\Delta G$ values into a $[-1, 1]$ interval (see Figure 2 on the poster).

To prevent data leakage and ensure generalization, the dataset was split into training, validation sets based on protein domain classification using the CATH homology. This guarantees that proteins from the same structural family do not appear in different splits. An embedding-based clustering analysis using ESM-2 was also performed to visualize the diversity and coverage of the dataset (see Figure 2).

3. Model Architectures and Training

Two model families were implemented and compared.

ProtBERT-Based Model: The ProtBERT architecture naturally accepts paired sequences separated by *[SEP]* tokens. We feed sliding windows of 255 amino acids around the mutation site for both wild-type and mutant sequences. The final embeddings of both sequences are mean-pooled, their absolute difference is computed, and these features are concatenated and passed through a four-layer MLP regression head (see Figure 1 on the poster).

ESM-2-Based Model: The ESM-2 model offers a larger context window but does not natively support paired sequences. Our solution processes the wild-type and mutant sequences separately using the same model with shared weights but with reset positional encodings for each sequence. Various pooling strategies (mean and attention pooling) were tested to aggregate the sequence representations. The same but smaller regression layer was used for final prediction.

Training Details: Models were fine-tuned on the LUMI supercomputer using the scheduler with warm-up. The training dataset was balanced by including both forward (WT→MUT) and reverse (MUT→WT) mutations.

4. Results and Evaluation

Models were evaluated on a held-out test set and on three independent public benchmarks: S350, PonSol[1], and BenchStab.

The ProtBERT-based model consistently outperformed the ESM-2 variants. On our test set, it achieved a Pearson correlation of 0.55 and a Spearman correlation of 0.42 . The ESM-2 models, while training faster, achieved correlations around 0.37 on validation datasest.

On external benchmarks, the trend held. For the BenchStab dataset, ProtBERT achieved a Pearson correlation of 0.57, compared to -0.10 for ESM-2 with mean pooling. Interestingly, ESM-2 showed slightly better performance on the PonSol solubility dataset, suggesting different PLMs may be optimal for related but distinct prediction tasks.

A comparison with state-of-the-art tools (FoldX, Rosetta, ThermoMPNN) shows that our ProtBERT model provides competitive performance, particularly on the more complex BenchStab dataset (see Table 2).

5. Conclusions and Future Work

This work successfully developed and evaluated deep learning models for predicting protein stability changes upon mutation. The key finding is that fine-tuning Protein Language Models is a effective strategy with lower biase of the training data, with the ProtBERT architecture proving superior for this specific task due to its inherent design for comparing paired sequences.

The created large-scale dataset and the implemented models are publicly available, providing a resource for the community. Future work could involve integrating additional PLMs (e.g., ProtT5), expanding the dataset with more stabilizing mutations, and applying explainability methods to interpret the models' predictions.

Acknowledgements

I would like to thank my supervisor Miloš Musil, Ph.D., and my consultant Antonín Jarolím, Ing., for their invaluable guidance, biological insights, and technical support throughout this project. Thanks also to the administrators of the LUMI supercomputer for their computational resources.

References

[1] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic Acids Research*, 33(suppl_2):W382–W388, 07 2005.

[2] Elizabeth H. Kellogg, Adrienne Leaver-Fay, and David Baker. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, 79(3):830–838, 2011. Kellogg et al. explore methods for computing the stability change corresponding to a point mutation in a protein with variable conformational sampling and scoring function selection.

[3] Tanja Kortemme, Diana E. Kim, and David Baker. Computational alanine scanning of protein–protein interfaces. *Sci STKE*, 2004(pl2):pl2, 2004. Kortemme et al. exhibit an alanine scanning

[1] Protein solubility dataset for comparison

algorithm for protein–protein interfaces that correctly predicts 79% of hot spot residues.

[4] 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.

[5] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *CoRR*, abs/2007.06225, 2020.

[6] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.