Project 2 - Predicting protein melting temperature from amino acid sequences

Igor Pavlović
Communication Systems, EPFL

Nataša Jovanović
Electrical Engineering, EPFL

Jelisaveta Aleksić
Electrical Engineering, EPFL

***Abstract*—This project, conducted in collaboration with the Laboratory for Biomolecular Modeling [1] under the mentorship of Lucien Krapp, focuses on developing a predictive model to estimate protein melting temperatures from amino acid sequences. Two distinct methods were implemented: fine-tuning the pre-trained ESM-2 model [2], [3] and using the CARBonAra Architecture [4] together with the AlphaFold model [5]. A publicly available dataset [6] was used for training and evaluation. The final model demonstrated strong predictive performance, achieving a Pearson's correlation coefficient (PCC) of 0.77, Spearman's correlation coefficient (SCC) of 0.56, Root Mean Square Error (RMSE) of 7.7, and Mean Absolute Error (MAE) of 5.6.**

I. Introduction

Proteins are a large class of molecules essential for the proper functioning of living organisms, thus knowing their functions and functionalities is a key focus in clinical and research contexts. In recent decades, the prediction and analysis of protein stability have become a crucial focus in chemical, medical, and biological research.

It is known that thermodynamic stability is one of the key properties of proteins that influences their structure and function [7]. However, various factors, such as amino acid composition or the presence of other molecules, can affect protein thermal stability, which then leads to loss of function or the formation of toxic protein aggregates.

The protein's melting temperature (T_m) is the temperature at which the protein unfolds, losing its 3D structure in favor of a linear polymer due to thermal energy. This is one way of measuring thermal stability, and analyzing it can provide a better understanding of the thermal stability of a protein. In other words, the higher the melting temperature, the more stable the protein is. This has important implications in chemical, medical, and biological research. For example, industrial chemical processes benefit from engineered enzymes with high thermal stability, leading to more sustainable and environmentally friendly practices, often called green chemistry.

Many research groups are working on developing machine learning models to analyze proteins and their functionalities. The introduction of protein sequence embeddings has enabled researchers to capture essential protein properties directly from amino acid sequences. Furthermore, models such as ESM-2 have shown that large-scale pre-trained models designed for protein sequences can predict various protein characteristics, including stability and binding affinities [3], [8].

Despite these advancements, accurately predicting the melting temperature of proteins remains challenging because of the intricate interplay of factors that influence stability. This project seeks to build upon these foundational efforts by employing pre-trained models and novel architectures to predict protein melting temperature.

A. AlphaFold

AlphaFold, developed by DeepMind [9], is a deep learning-based model that predicts 3D protein structures directly from amino acid sequences with near-experimental accuracy. It employs a transformer-based architecture integrated with Evoformer blocks, which efficiently process sequence and pairwise residue information to learn spatial relationships and constraints. AlphaFold combines supervised learning with a multi-track attention mechanism to predict inter-residue distances, torsion angles, and a confidence score (pLDDT), enabling precise reconstruction of protein structures [10][11]. These predicted structures are useful for our task as an input to the CARBonAra model.

B. CARBonAra

CARBonAra (Context-aware Amino acid Recovery from Backbone Atoms and heteroatoms) [4] is a protein sequence generation model based on geometric transformers. It predicts amino acid probabilities for protein backbones while incorporating molecular contexts, such as surrounding ligands or ions, to enhance design precision.

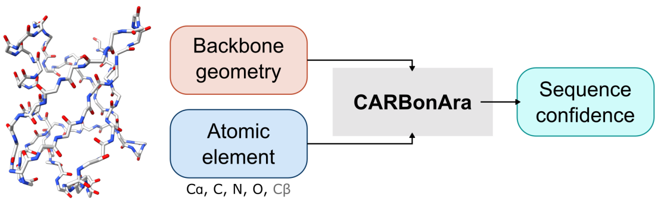

Figure 1. CARBonAra summary [12]

The model is trained on structural data from the Protein Data Bank (PDB) and achieves competitive sequence recovery rates compared to leading methods like ProteinMPNN [13] and ESM-IF1 [14] while being computationally efficient. It excels in tasks requiring molecular context, such as designing highly thermostable enzymes or engineering functional proteins.

C. ESM-2

The ESM-2 model, developed by *Meta AI* and available through *Hugging Face* is a deep-learning model used for protein sequence analysis. It is mainly used to predict protein sequences, properties, and interactions between amino acid sequences. It has a transformer architecture and works with strings of amino acids. As input, it has protein sequences which are tokenized. A self-attention mechanism is used to focus on different parts of the sequence. This model is pre-trained with large datasets, and thus we used it for our problem.

The model takes a protein sequence, represented as a string of amino acid residues, and converts each residue into a dense numerical vector using a learned embedding matrix. These embeddings, similarly to text sequence embeddings, capture semantic and structural properties of amino acids based on their positions and context within the sequence. These embeddings are used in various tasks, including protein structure prediction, function annotation, protein-protein interaction prediction, and other bioinformatics applications. [2]

II. Data

A. Training dataset

1) Description: The initial dataset used for training as well as the test dataset is from the publicly available Kaggle competition [6]. However, this dataset has since been updated and is available at [15]. In the updated file 2409 rows where all features are marked as NaN have been removed and 25 rows where the *pH* and *tm* values were transposed were updated. The updated dataset contains 28981 data points, each with 4 features - Sequence ID (*seq_id*), Protein sequence (*protein_sequence*), pH value (*pH*), and data source (*data_source*) as well as a corresponding melting temperature in degrees Celsius (*tm*).

The minimum sequence length is 5, the maximum sequence length is 8798, the median is 351 and the mean is 450. The distribution of protein sequence lengths can be seen in Figure 2. Number of different *data_source* values is 324. As there are many different sources of amino acids, this can influence the model in terms of having the stability and reliability of each data source. Furthermore, since almost all pH values are set to 7, we completely ignored this feature.

2) Grouping by Wildtype and Mutations: The data contains sequences but some of them are mutations of the wildtype sequence. To better evaluate the performance of our model, the training set is divided into groups.

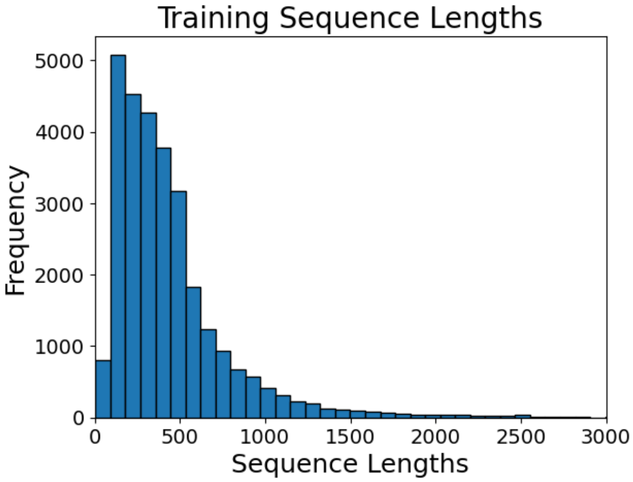

Figure 2. The distribution of protein sequence lengths inside the training data.

We implemented the grouping largely based on the code provided at [16] where the wildtype sequence for each group is estimated from all sequences in the dataset, and the corresponding mutations are identified. The most common sequence within a group is considered the wild type. Sequences that differ from the wild type by only one letter (amino acid) are considered to be mutations.

Groups with fewer than 3 sequences are excluded. The length of sequences within the same group can differ by at most 1 amino acid. After filtering, a total of 105 groups remain. Two files are generated; one with the updated training dataset containing the original grouped sequences along with the newly added *wildtype* feature and one with the sequences that do not meet the minimum group size criterion or do not fit into the defined groups. This division would ensure that during training and validation, all the mutations will be processed together within the group. Moreover, there would be better control of data leakage.

As per Figure 3, there are large differences in group sizes. The maximum group size is 708, the minimum is 1, and the mean is 41. Because of the small number of groups - 105 and the discrepancy between group sizes, further statistical and knowledge-based analysis would be required to accurately use the groups for splitting the data. To train our models, we opted for cross-validation with a random split between ungrouped data.

B. Test dataset

The test set comprises 2413 data points without the corresponding labels. However, the test data consists exclusively of many different mutations of the same protein with all the test sequences having almost the same length. Therefore, the primary focus was on evaluating the effects of protein features on model performance rather than achieving a high score on the Kaggle leaderboard [6], since most high-performing solutions make use of this bias in the test data.

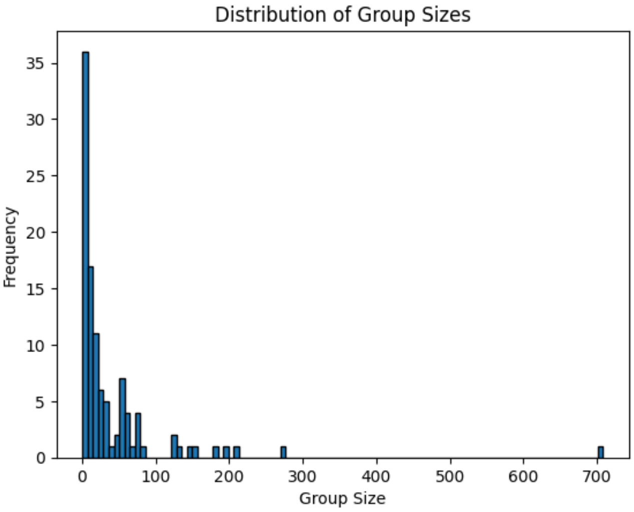

Figure 3. The frequency of different group sizes

III. Methods

A. Model 1 - Pretrained ESM-2 Model

This model from *Hugging Face* is used to generate sequence embeddings. These embeddings are used to train a neural network to predict the melting temperature (T_m) values. Tokenization of sequences is performed on both train and test sequences with a maximum length of 512. Because of the large size of datasets, data are loaded in batches of size 8. Adam optimizer is used with the learning rate $5 \cdot 10^{-5}$ and the number of epochs is 10. As mentioned, the neural network has weights that are used from the pre-trained ESM-2 model. Data is trained on this model and then tested on the validation dataset.

B. Model 2 - Carbonara Model

To extract additional features related to the protein structure, we first generated the five most likely protein structures using the AlphaFold Database, and then we used the CARBonAra architecture [12] embeddings generated from the protein structure (precisely, the output from the penultimate layer). These embeddings are represented by a sequence of vectors (one vector per amino acid).

After generating the embeddings, two models were used to evaluate the usefulness of these features. The first model performs pooling on the sequence vectors to map these embeddings into a fixed number of features that could be used to train a multilayer perceptron with two layers and 64 nodes in the hidden layer. The pooling layer extracts the following values from the sequence of carbonara embeddings: min, max, mean, median, std and the values from the first and the last embeddings vector. The second model trains a recurrent neural network on the embeddings extracted directly from CARBonAra. It uses a single LSTM layer, along with a single linear layer. Both models are trained on 100 epochs using the Adam optimizer with the learning rate of 10^{-3}.

Furthermore, we explored additional features such as pLDDT scores, which reflect local confidence of the Alpha Fold prediction of the structure. Since a low confidence in protein structure might indicate a less stable protein and therefore a lower melting temperature.

C. Evaluation metrics

Since the test dataset is small and includes only one protein sequence with its mutations, we primarily rely on *k-fold cross-validation* to evaluate our models. The data is randomly split into train and validation sets using 5-fold cross-validation, providing a more robust and reliable estimate of model performance.

For training each of our models, a simple root mean squared error loss function (RMSE) was used to compute the gradients. Furthermore, we also computed the mean absolute error (MAE), Pearson correlation coefficient (PCC), and the Spearman correlation coefficient (SCC) to evaluate the model performance on the validation data using 5-fold cross-validation.

IV. Experiments and Results

A. Experimental setup

We used *PyTorch* for model implementation and training. For sequence-based modeling, we fine-tuned the *esm2_t6_8M_UR50D* model from Hugging Face, a pre-trained transformer with 6 layers and 8 million parameters. For this purpose, *transformers* was used. Protein 3D structures were generated using AlphaFold, while structural features were extracted using CARBonAra. Additionally, *scikit-learn* was used for evaluation metrics as well as *scipy*. For sequence grouping *statistics*, *collections* and *operator* are used. For visualization, we used *matplotlib* and *seaborn* and for working with datasets, *pandas*. For working with HDF5 files, *h5py* was used.

B. Model 1 performance

Figure 4 presents the relation between predicted and true values of protein melting temperature with the model explained in III-A on the validation dataset after 10 epochs. It can be seen that they have a positive correlation. The range between [40-60]°C is the densest, therefore the largest number of points have a value of T_m within that range. This graph does not give a perfectly linear dependence of true and predicted values from which we can conclude that there are some wrong predictions as seen in calculated metrics.

We obtained the following results on the validation set: the Pearson correlation coefficient is 0.77, the Spearman correlation coefficient is 0.56, the MAE is 5.6 and the RMSE is 7.7.

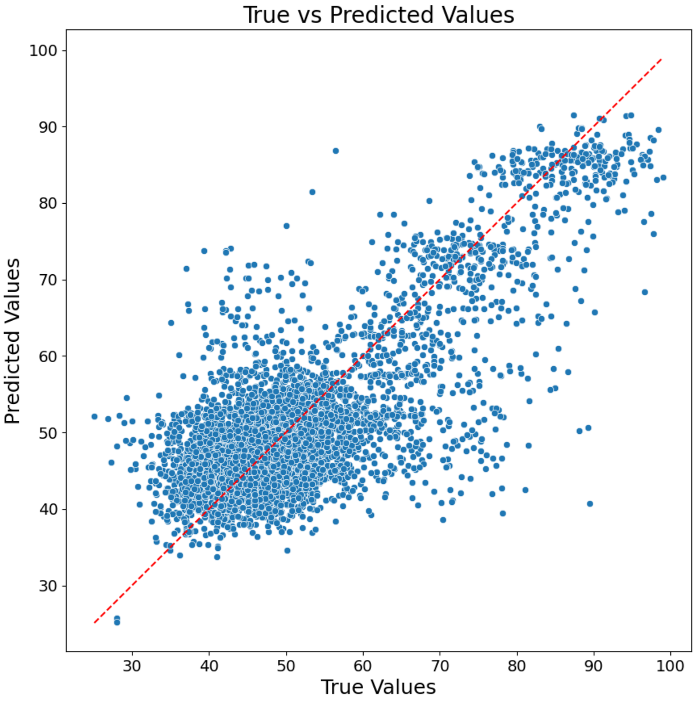

Figure 4. True vs Predicted values for the ESM model after 10 epochs

C. Model 2 performance

The performance of the two aforementioned models was evaluated on the CARBonAra output with and without the additional pLDDT scores in order to assess the impact of those features. Including pLDDT scores allows the model to include the reliability of results from AlphaFold. Higher pLDDT can give a better prediction of the protein melting temperature because it can better account for structural uncertainties. The performance of these models is presented in Table I.

Model	PCC	SCC	RMSE	MAE
MLP (w/o pLDDT)	0.48 ± 0.02	0.36 ± 0.02	12.1 ± 0.7	8.9 ± 0.7
MLP (w pLDDT)	**0.49 ± 0.01**	**0.37 ± 0.01**	**11.6 ± 0.2**	**8.7 ± 0.3**
RNN (w/o pLDDT)	0.27 ± 0.02	0.20 ± 0.03	12.6 ± 0.1	9.3 ± 0.1
RNN (w pLDDT)	0.28 ± 0.01	0.21 ± 0.01	12.5 ± 0.1	9.2 ± 0.2

Table I
MODEL 2 PERFORMANCES AND RESULTS WITH AND WITHOUT PLDDT

Table I shows a negligible effect pLDDT has on the performance of these two models. Comparing all the models and their respective metrics, the MLP model without the pLDDT factor achieves the best performance with the highest PCC and SCC as well as the lowest MAE and RMSE. Therefore, this model is the best derived from CARBonAra and AlphaFold.

D. Overall performance

In this table, we compare the performance of the implemented models - Model 1 (ESM) and the best Model 2 (MLP with pLDDT factor).

Model	PCC	SCC	RMSE	MAE
ESM	**0.77 ± 0.01**	**0.56 ± 0.01**	**7.7 ± 0.1**	**5.6 ± 0.1**
MLP (with pLDDT)	0.49 ± 0.01	0.37 ± 0.01	11.6 ± 0.2	8.7 ± 0.3

Table II
COMPARISON OF MODEL 1 (ESM) AND MODEL 2 (MLP WITH PLDDT) PERFORMANCES

We observe that Model 1 (ESM) outperforms Model 2 (MLP with pLDDT) in all key performance metrics. Specifically: - PCC and SCC are significantly higher in Model 1, indicating that the ESM model captures the linear and monotonic relationships between true and predicted values much better. - The RMSE and MAE for Model 1 are both lower, suggesting that its predictions are more accurate with fewer errors compared to Model 2. This analysis shows that the pre-trained ESM-2 model (Model 1) provides better overall performance in terms of both accuracy and correlation, making it the best model for this project.

V. SUMMARY

In this project, our focus was on predicting protein melting temperatures based on amino acid sequences. Two models were developed: one based on the pre-trained ESM-2 model and another incorporating the Carbonara architecture.

Model 1, which uses the pre-trained ESM-2 model, outperformed Model 2 (Carbonara with AlphaFold), with higher PCC and SCC, as well as lower RMSE and MAE. Specifically, Model 1 achieved a PCC of 0.77, SCC of 0.56, RMSE of 7.7, and MAE of 5.6, demonstrating better predictive accuracy and correlation.

Future work: To achieve better model performance, it should be tested on a larger dataset with, for example, different values of *pH* values and to take group split - training and validating considering wildtype sequences and their mutations. This would give a more general model that can better predict melting temperatures of unknown protein sequences. Furthermore, we could employ an architecture based on the one from [17], that uses GNN on a combination of protein structures, sequences and knowledge-based features, or an architecture from [18] that uses light-attention and MLP blocks as an addition to embeddings from the pre-trained model.

VI. Ethical risk

Description: Since the dataset used for training and evaluating our models is publicly available and does not contain any sensitive information, there are no ethical risks regarding the data itself. However, the ethical risk we have identified in our project is the potential misuse of our protein melting temperature predictions in medical applications, such as for therapeutic protein development [19]. Predictions can be misinterpreted or inaccurate which could lead to ineffective or unsafe protein-based treatments. This could lead to a risk to patient safety and delayed therapeutic development.

Impacted Stakeholders: The impacted groups would be medical researchers, pharmaceutical companies, and patients.

Significance of the risk: The risk is moderately severe, as incorrect predictions can have significant repercussions in the medical field. The likelihood of this issue occurring is medium due to the robust validation processes typically followed in medical research.

Evaluation of the Risk: To assess this risk, we examined cases where machine learning models have been misused or produced inaccurate predictions in the medical field. We consulted our mentor to understand the consequences of incorrect protein stability predictions and how they might impact clinical applications.

Barriers: We have not been able to take it into account. We can not control how some users will apply the predictions. There are constraints due to the complexity of protein behavior and the accuracy.

References

[1] Laboratory of Biological Modeling (LBM), EPFL, "Lbm laboratory website," 2024. [Online]. Available: https://www.epfl.ch/labs/lbm/

[2] Hugging Face, "Esm model documentation," 2024. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/esm

[3] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[4] L. F. Krapp, F. A. Meireles, L. A. Abriata, J. Devillard, S. Vacle, M. J. Marcaida, and M. Dal Peraro, "Context-aware geometric deep learning for protein sequence design," *Nature Communications*, vol. 15, no. 1, p. 6273, 2024.

[5] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, K. Tunyasuvunakool, P. Kohli, D. Hassabis, and et al., "Alphafold protein structure database," pp. D439–D444, 2022. [Online]. Available: https://alphafold.ebi.ac.uk/

[6] Kaggle, "Novozymes enzyme stability prediction dataset," 2024. [Online]. Available: https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/data?select=train.csv

[7] A. Jarzab, N. Kurzawa, T. Hopf *et al.*, "Meltome atlas—thermal proteome stability across the tree of life," 2020. [Online]. Available: https://doi.org/10.1038/s41592-020-0801-4

[8] F. Jung, K. Frey, D. Zimmer, and T. Mühlhaus, "Deepstabp: A deep learning approach for the prediction of thermal protein stability," Apr 18 2023.

[9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[10] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon *et al.*, "Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic acids research*, vol. 50, no. D1, pp. D439–D444, 2022.

[11] M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo *et al.*, "Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences," *Nucleic acids research*, vol. 52, no. D1, pp. D368–D375, 2024.

[12] LBM-EPFL, "Carbonara: Computational analysis of rna binding proteins," 2024. [Online]. Available: https://github.com/LBM-EPFL/CARBonAra/tree/main

[13] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel *et al.*, "Robust deep learning–based protein sequence design using proteinmpnn," *Science*, vol. 378, no. 6615, pp. 49–56, 2022.

[14] B. Hie, S. Candido, Z. Lin, O. Kabeli, R. Rao, N. Smetanin, T. Sercu, and A. Rives, "A high-level programming language for generative protein design," *bioRxiv*, pp. 2022–12, 2022.

[15] Kaggle, "Discussion: Novozymes enzyme stability prediction," 2024. [Online]. Available: https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/discussion/356251

[16] R. Hatch, "Novotrain: Data contains wildtype groups," 2023. [Online]. Available: https://www.kaggle.com/code/roberthatch/novo-train-data-contains-wildtype-groups/notebook

[17] G. Li, S. Yao, and L. Fan, "Prostage: Predicting effects of mutations on protein stability by using protein embeddings and graph convolutional networks," Jan 2 2024, open Access.

[18] H. Dieckhaus, M. Brocidiacono, N. Z. Randolph, and B. Kuhlman, "Transfer learning to leverage larger datasets for improved prediction of protein stability changes," 2024. [Online]. Available: https://doi.org/10.1073/pnas.2314853121

[19] J. B. Tucker and C. Hooper, “Protein engineering: Security implications: The increasing ability to manipulate protein toxins for hostile purposes has prompted calls for regulation,” *EMBO reports*, vol. 7, no. S1, pp. S14–S17, 2006.