Published in Computational Science – ICCS 2024, Málaga, Spain, 02-04 July 2024, which should be cited to refer to this work. DOI:10.1007/978-3-031-63759-9_23

EsmTemp - Transfer Learning Approach for Predicting Protein Thermostability

Adam Sułek[1,*], Jakub Jończyk[1,2], Patryk Orzechowski[1,3,4], Ahmed Abdeen Hamed[1], and Marek Wodziński[4,5,*]

[1] Sano Centre for Computational Medicine, 30-072 Kraków, Poland
[2] Jagiellonian University Medical College, 30-688 Kraków, Poland
[3] University of Pennsylvania, Philadelphia, PA 19104, USA
[4] AGH University of Science and Technology, 32059, Kraków, Poland
[5] University of Applied Sciences Western Switzerland, Sierre, Switzerland
* a.sulek@sanoscience.org, wodzinski@agh.edu.pl

Abstract. Protein thermostability is one of the most important features of bio-engineered proteins with significant scientific and industrial applications. Unfortunately, obtaining thermostable proteins is both expensive and complex. Recent advances in Protein Language Models (pLM) offer promising framework for sequence-to-sequence problems, especially in the realm of protein thermostability prediction. In this work, we present EsmTemp, a transfer learning model based on the ESM-2 pLM architecture. EsmTemp undergoes training on a meticulously curated dataset comprising 24,000 protein sequences with known melting temperatures. A rigorous evaluation, conducted through a 10-fold cross-validation, yields a coefficient of determination (R^2) of 0.70 and a mean absolute error of 4.3°C. These outcomes highlight how pLM has the potential to advance our understanding of protein thermostability and facilitate the rational design of enzymes for various applications.

Keywords: protein thermostability · protein language models · transfer learning · ESM-2 · artificial neural network.

1 Introduction

Thermostable proteins are important for biotechnology, medicine, and pharmacy [8, 18]. Key examples include the heat-resistant DNA polymerase employed in Polymerase Chain Reaction [16], proteases used in the synthesis of angiotensin-converting-enzyme inhibitors [2], or Ferritins harnessed for crafting nanocarriers. Enhanced enzyme heat resistance improves industrial processes, resulting in faster reactions, enhanced substrate solubility, and reduced contamination risks [12]. Thermostable proteins are often sourced from thermophilic organisms or engineered through mutagenesis and directed evolution. Despite their effectiveness, these methods have limitations, such as time investment and trial-and-error reliance. Determining the melting temperature (Tm) is a common method for assessing protein thermostability. It indicates the temperature at which half of

the protein unfolds. However, accurate experimental Tm determination demands substantial quantities of high-purity proteins or cell cultures, making it labor-intensive and costly. Hence, there's a critical need to develop a more efficient and cost-effective method for predicting protein Tm.

The protein structure is determined by the sequence of amino acids, which directly impact their function and properties. Proteins with similar amino acid sequences not only share corresponding biochemical functions, but also exhibit similar characteristics, even if they are found in different organisms. On the other hand, single amino acid changes can profoundly alter protein structure, interactions, and properties, as exemplified by the critical impact of point mutations on the behavior of some proteins [17]. This observation is essential for predicting protein structures and properties using computational techniques. Over the past few years, remarkable progress has been made in the study of proteins, largely due to the utilization of machine learning (ML) approaches in the field of natural language processing (NLP). Through the translation of the amino acid sequence into tokenized vectors, models become capable of capturing pertinent protein features. Sequence embeddings serve as a versatile tool for predicting various properties of proteins, both at the global and local levels [15]. Protein Language Models (pLM) developed so far, such as Evoformer, ProteinBERT, ESM-1b and ProtGPT2 are successfully used in automated function or structure predictions [1, 3, 6, 14]. Prediction of thermostability using ML and pLM offers insights for protein engineering.

Several classification algorithms and tools have been developed to distinguish proteins into thermostable and thermolabile based on their sequences. Support Vector Machines, Random Forests and Gradient Boosting have reported accuracy rates even up to 85% [9, 20]. Although certain classification-based predictors have shown satisfactory performance, they are not a suitable solution for accurately determining the precise Tm value. ProTstab2, developed with the LightGBM framework and protein sequence featurization, stands out among the available regression models [19].

ProThermDB held about 10k entries on protein melting temperatures, serving as the main thermostability data repository [13]. The Meltome Atlas database enriched existing data by including Tm information for 48k proteins across 13 species and a human cell line dataset of 13k cases from 14 cell lines [5]. Still, the thermostability dataset is considerably smaller than >200M protein sequences stored in UniProt Knowledgebase, underscoring a substantial data deficit.

The introduction of pLM-based transfer learning for evaluating protein thermostability through sequence analysis was first presented in the DeepSTABp tool [7]. This model combines multilayer perceptron with a pre-trained transformer-based model, ProtTrans-XL. In addition to protein sequences, it considers measurement context (cell lysate or whole cell) and the source organism's Optimal Growth Temperature (OGT).

ESM-2 protein language models are constructed using the transformer architecture, specifically tailored for training on protein sequences [4, 11]. Recognizing amino acid sequences as a unique language, transformer models, similar

to BERT-like architectures, have demonstrated proficiency in capturing crucial features of proteins [1]. The input exclusively comprised protein sequences, encompassing the 20 standard amino acids. Our tool EsmTemp predicts Tm values for various proteins by leveraging pre-trained ESM2 (650M) embeddings and fine-tuning. By combining the data from Meltome Atlas and ProThermDB, we curated a comprehensive dataset characterized by a broad Tm value range and standardized experimental protocols. EsmTemp's effectiveness has been validated in down-stream task pertaining to protein thermostability, demonstrating its usability in amino acid sequence analyses. The code, the data and additional figures are provided here: `https://github.com/SanoScience/esm_temp`

2 Materials and methods

The dataset combined data from ProThermDB and Meltome Atlas [5, 13]. The aggregated data underwent preprocessing to ensure consistency and relevance for subsequent analysis. Experimental protocols were standardized by retaining only measurements conducted on cell lysates. The dataset was refined by excluding protein sequences with Tm values outside the range of 30°C to 98°C and lengths below 100 or exceeding 900 amino acids. After eliminating duplicate entries, the remaining sequences underwent clustering via the CD-HIT algorithm [10] to ensure that the training and test sets do not have identical or near identical examples. To minimize redundancy, we set a sequence similarity threshold of 0.7 and a minimum cluster size of 3, while keeping all other parameters at their default values. As a result of this process, we obtained a curated dataset comprising 24,472 distinct amino acid sequences. Each sequence is accompanied by its respective source organism and Tm value.

We utilized ESM-2, a model equipped with 650M parameters and 1280-dimensional embeddings for amino acids. This allowed us to represent each protein through pooling the embeddings from the amino acids, with predictions made using an output layer containing a single neuron [11]. The network weights were optimized using AdamW with a learning rate of 1e-4. MAE quantifies a model's accuracy in predicting continuous variables by assessing the average error magnitude. The model's performance was assessed with (R^2), determining the variance in the dependent variable that can be predicted by the independent variable.

Due to GPU memory constraints (40GB, NVIDIA A100 units), a pseudo-batch approach was employed, where only 1 sequence was included in each batch. The gradient updates in a backward step were performed after accumulating the loss over a mini-batch containing 16 proteins. Our investigation involved gradually unfreezing three layers of the ESM-2 model to analyze the advantages of fine-tuning in transfer learning. This strategy aims to refine the pre-trained model's parameters for the Tm prediction task. We used 10-fold cross-validation to address training variations.

3 Results

In this study, three separate hypotheses were tested. First, we verified how data scaling affected the performance of the model. Secondly, we explored potential benefits from fine-tuning process on the range of unfrozen layers (Figure 1). Thirdly, we evaluated the model's predictive performance both on a broad scale and within specific organisms.

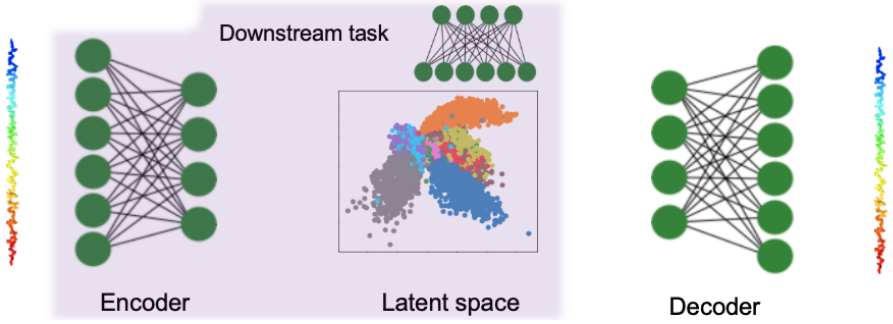

Fig. 1. Leveraging a transfer learning approach with the pLM autoencoder model involves utilizing a pre-trained latent space and its application to the downstream task.

Normalization techniques enhance the convergence and stability of optimization algorithms used in regression tasks. We examined how normalization and scaling methods, such as log scaling, square scaling, and min-max normalization, affect model prediction performance compared to non-scaling labels. Log and square scaling had no positive impact on the model's performance, as shown by Table 1's MAE and R^2 values. Normalizing the target variable consistently improved training outcomes, so we adopted min-max scaled labels for subsequent experiments.

The ESM-2 model has a complex architecture with multiple layers. By unfreezing specific layers, sequence embeddings can be fine-tuned to enhance model performance. Carefully controlling unfrozen layers is crucial to prevent gradient decreases and ensure effective learning. We used selective layer unfreezing in transfer learning framework and an extra Fully Connected layer to tailor the model for the study down-stream task. The effectiveness of ESM-2 in transfer learning for thermostability prediction was discerned by utilizing raw embeddings, which were subsequently passed as input to the regression task. We noticed a relatively strong correlation (R^2=0.64), though accompanied by a high mean absolute error. By unfreezing the final layers of ESM-2, it becomes possible to fine-tune not only the raw embeddings but also specific layers within the model, resulting in improved predictions. The results showed enhanced R^2 values and a corresponding decrease in MAE, as illustrated in Table 1.

Table 1. The results of sequence stability prediction with ESM-2 (650M parameters) model with different transfer learning and scaling procedures.

Method	MAE	R^2	Scaling Function
ESM-2 raw embeddings	8.86 ± 0.105	0.64 ± 0.012	Min-max
Unfrozen 0	4.90 ± 0.061	0.63 ± 0.018	Min-max
Unfrozen 1	4.90 ± 0.060	0.63 ± 0.015	Min-max
Unfrozen 2	4.42 ± 0.096	0.70 ± 0.025	Min-max
Unfrozen 3	4.33 ± 0.076	0.70 ± 0.018	Min-max
Unfrozen 3	4.57 ± 0.051	0.67 ± 0.005	Non-scaling
Unfrozen 3	4.50 ± 0.069	0.68 ± 0.019	Log
Unfrozen 3	4.54 ± 0.076	0.67 ± 0.018	sqrt

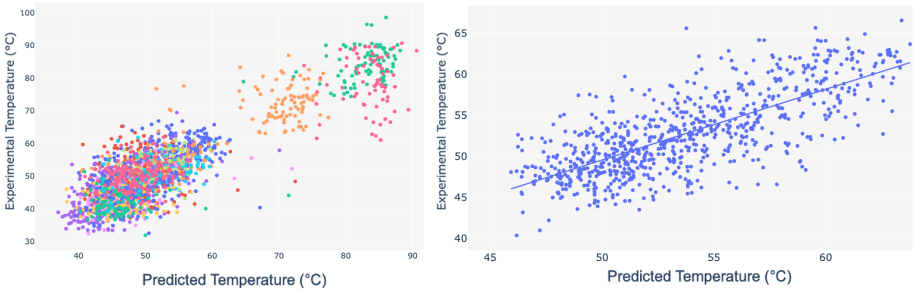

Fig. 2. The correlation analysis depicts the relationship between melting temperatures with the prediction for the best fine-tune ESM-2 model across specific organisms (A) and with a focus on fine-tuning exclusively on Human lysates (B).

In this study, we experimented with unfreezing up to three consecutive layers. We found that unfreezing three layers yielded the most favorable results, leading us to adopt this approach for subsequent experiments. The final model achieves MAE of 4.3°C and an R^2 score of 0.704 on a 10-fold validation procedure (Table 1, Fig. 3A.). While the model prediction demonstrates a relatively strong correlation during global assessment of protein thermostability over the full temperature range, the visualization with color coding for each species reveals lower or even no correlation within the species (Table 2, Fig. 3A). To further explore the targeted applicability of EsmTemp, aiming specifically for improved correlation within organisms and potential applications for point mutations, we conducted fine-tuning within a single organism, utilizing the most effective strategies from previous experiments. It shows that the availability of more data may facilitate the training of a more accurate model within the species. Considering the extensive datasets on protein thermostability across diverse organisms, the variation in protein sequences, which can be attributed to their unique biological functions, presents difficulties in accessing information regarding their thermostability. The

results summarized in Table 2 clearly showed that despite a similar temperature range and a very similar Tm distribution, the model is able to accurately capture the relationship between the amino acid sequence and thermostability only in the case of human proteins.

Table 2. The results of Tm prediction R^2 using the optimal fine-tune ESM-2 model for specific organisms.

Lysate from organism	**Correlation R^2**	**Number of cases**
Human	0.49	7997
Arabidopsis thaliana seedling	0.26	2260
Caenorhabditis elegans	0.15	2917
Mus musculus	0.10	3870

EsmTemp model was compared to ProtStab2 and DeepSTABp, both trained with Meltome Atlas data [19, 7]. However, this comparison may not be entirely equal. Similarities between unfiltered test and training data can compromise valid comparisons. Furthermore, our method ensured robustness and repeatability by using k-fold validation, while the other models used randomly selected test sets. Methodological differences make k-fold validation more conservative and reliable than single random test sets. To address this issue, we focused on comparing the R^2 correlation values in the publications. The R^2 score for ProtStab2 on a blind test subset was 0.58, while DeepSTABp performed better with a score of 0.80. While EsmTemp achieved an R^2 score of 0.70, surpassing ProtStab2, it did not demonstrate the same level of correlation as DeepSTABp.

4 Conclusions

This study examines the possibilities and constraints of employing transfer learning from ESM-2 embeddings and ESM-2 fine-tuning for the prediction of protein melting temperature (Tm). Protein language models are advanced neural networks that undergo training using vast databases of protein sequences such as UniProt. Typically, pLMs are trained by predicting missing amino acids or motifs in protein sequences, enhancing their understanding of protein properties. As a result, pLMs serve as valuable repositories of parameterized data, adept at addressing various tasks within their domain. Such knowledge transfer allows to obtain predictive models that are more effective than the original model training on a limited data resource. Our results demonstrate the model's accurate prediction of Tm for a vast landscape of proteins, with a mean absolute error of 4.3°C. Furthermore, our model does not necessitate external features, such as the source of the experiment or the OGT. The limitations of our approach lie in predicting the effects of point mutations on Tm, as our model relies on the global sequence information and does not account for the local structural features. Taking this into consideration, our future plans involve implementing

more advanced feature extraction techniques, like graph convolutions, to enhance the precision and practicality of Tm prediction.

Acknowledgements The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533. This publication is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. This research was possible due to the support of PLGrid infrastructure grant plgsano4-gpu and PLG/2023/016261 on the Athena supercomputer cluster.

References

1. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M.: Proteinbert: a universal deep-learning model of protein sequence and function. Bioinformatics **38**, 2102–2110 (2022). https://doi.org/10.1093/bioinformatics/btac020
2. Cheung, I.W.Y., Nakayama, S., Hsu, M.N.K., Samaranayaka, A.G.P., Li-Chan, E.C.Y.: Angiotensin-i converting enzyme inhibitory activity of hydrolysates from oat (avena sativa) proteins by in silico and in vitro analyses. J. Agric. Food Chem. **57**, 9234–9242 (2009). https://doi.org/10.1021/jf9018245
3. Ferruz, N., Schmidt, S., Höcker, B.: Protgpt2 is a deep unsupervised language model for protein design. Nat Commun **13**, 4348 (2022). https://doi.org/10.1038/s41467-022-32007-7
4. Hu, M., Yuan, F., Yang, K., Ju, F., Su, J., Wang, H., Yang, F., Ding, Q.: Exploring evolution-aware & -free protein language models as protein function predictors (2022), `http://arxiv.org/abs/2206.06583`
5. Jarzab, A., Kurzawa, N., Hopf, T., Moerch, M., Zecha, J., Leijten, N., Bian, Y., Musiol, E., Maschberger, M., Stoehr, G., Becher, I., Daly, C., Samaras, P., Mergner, J., Spanier, B., Angelov, A., Werner, T., Bantscheff, M., Wilhelm, M., Klingenspor, M., Lemeer, S., Liebl, W., Hahne, H., Savitski, M., Kuster, B.: Meltome atlas-thermal proteome stability across the tree of life. Nat Methods **17**, 495–503 (2020). https://doi.org/10.1038/s41592-020-0801-4
6. 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., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with alphafold. Nature **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
7. Jung, F., Frey, K., Zimmer, D., Mühlhaus, T.: Deepstabp: A deep learning approach for the prediction of thermal protein stability. International Journal of Molecular Sciences **24**, 7444 (2023). https://doi.org/10.3390/ijms24087444
8. Kaneko, H., Minagawa, H., Shimada, J.: Rational design of thermostable lactate oxidase by analyzing quaternary structure and prevention of deamidation. Biotechnology letters **27**, 1777–1784 (2005). https://doi.org/10.1007/s10529-005-3555-2

9. Ku, T., Lu, P., Chan, C., Wang, T., Lai, S., Lyu, P., Hsiao, N.: Predicting melting temperature directly from protein sequences. Computational Biology and Chemistry **33**, 445–450 (2009). https://doi.org/10.1016/j.compbiolchem.2009.10.002
10. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**, 1658–1659 (2006), `https://doi.org/10.1093/bioinformatics/btl158`
11. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**, 1123–1130 (2023). https://doi.org/10.1126/science.ade2574
12. Mesbah, N.: Editorial: Enzymes from extreme environments, volume ii. Frontiers in Bioengineering and Biotechnology **9**, 799426 (2021). https://doi.org/10.3389/fbioe.2021.799426
13. Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D., Gromiha, M.: Prothermdb: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Research **49**, D420–D424 (2021). https://doi.org/10.1093/nar/gkaa1035
14. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C., Ma, J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences **118**, e2016239118 (2021). https://doi.org/10.1073/pnas.2016239118
15. Saar, K.L., Morgunov, A.S., Qi, R., Arter, W.E., Krainer, G., Lee, A.A., Knowles, T.P.J.: Learning the molecular grammar of protein condensates from sequence determinants and embeddings. Proc Natl Acad Sci U S A **118**(e2019053118) (2021). https://doi.org/10.1073/pnas.2019053118
16. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H.A.: Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. Science **239**(4839), 487–491 (1988). https://doi.org/10.1126/science.2448875
17. Schaefer, C., Rost, B.: Predict impact of single amino acid change upon protein structure. BMC Genomics **13**(S4), S4 (2012). https://doi.org/10.1186/1471-2164-13-S4-S4
18. Schilling, J., Jost, C., Ilie, I.M., Schnabl, J., Buechi, O., Eapen, R.S., Truffer, R., Caflisch, A., Forrer, P.: Thermostable designed ankyrin repeat proteins (darpins) as building blocks for innovative drugs. Journal of Biological Chemistry **298**(1) (2022). https://doi.org/10.1016/j.jbc.2021.101403
19. Yang, Y., Zhao, J., Zeng, L., Vihinen, M.: Protstab2 for prediction of protein thermal stabilities. International Journal of Molecular Sciences **23**, 10798 (2022). https://doi.org/10.3390/ijms231810798
20. Zhang, G., Fang, B.: Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. Protein Pept Lett **13**, 965–970 (2006). https://doi.org/10.2174/092986606778777560