FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2024

Exploring Protein Thermostability Using LLMs

Lukáš Kuhajda[1], Tomáš Honzík, Daniel Georgiev[2]

1 Introduction

Proteins are the functional units of living organisms, performing a vast array of functions. Organisms on our planet thrive in a huge range of temperatures, from -20°C to +122°C, with human proteins optimized to work at 36-37°C. Some proteins begin to lose activity or denature even with a slight deviation from this standard temperature.

Protein sequences consist of 20 amino acids represented by letters. Proteins sharing over 80% sequence similarity may differ in their optimal activity temperature by more than 20°C. Understanding the complexity of proteins and how mutations can affect their natural function and structure exceeds human capabilities, as do the current artificial tools.

The global enzyme market, comprising proteins acting as catalysts in chemical reactions, was valued at $6.4 billion in 2021 and is projected to reach $8.7 billion in 2026. Efforts are underway globally to screen for new sources of enzymes capable of functioning in extreme conditions, reflecting a significant endeavor to better understand and engineer enzymes for different temperature ranges.

Proteins can be categorized into four groups based on their optimal activity temperature: psychrophilic (<20°C), mesophilic (20-50°C), thermophilic (50-80°C), and hyperthermophilic (>80°C). The goal of this project is to develop a cutting-edge neural network model for classifying proteins into temperature groups and to advance our understanding of thermostability beyond current limits.

2 Methods

In the years 2021-2023, numerous studies have emerged in the realm of protein thermostability classification (Ahmed (2022), Charoenkwan (2022), Zhao (2023)). These works commonly categorize proteins into only two classes: non-thermophilic and thermophilic. However, this approach yields less nuanced results and offers diminished reporting value. Furthermore, they rely on outdated machine learning techniques and training datasets typically comprising thousands to tens of thousands of labeled protein sequences.

Training deep neural networks typically demands much larger datasets. Therefore, this study generated a novel dataset consisting of 3 million protein sequences to enhance the model's generalizability and robustness.

A pretrained large language transformer model for protein sequences, trained on unmasking (ESM), was acquired and fine-tuned to function as a classifier for the mentioned four temperature classes. Simultaneously, it was fine-tuned to classify proteins into the two temperature classes (non-thermophilic, thermophilic) for comparison with prior publications.

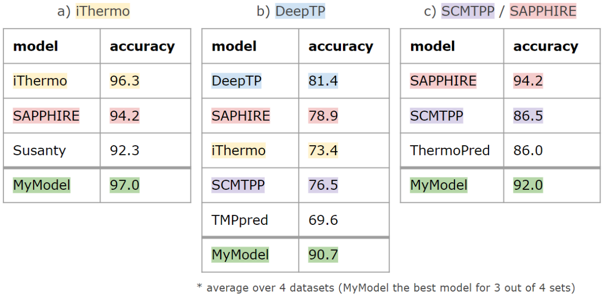
[1] PhD student of Applied Sciences and Informatics, field of study Cybernetics, specialization Artificial Intelligence for Synthetic Biology, e-mail: kuhajdal@ntis.zcu.cz

[2] University of West Bohemia, Faculty of Applied Sciences, email:thonzik@ntis.zcu.cz, georgiev@kky.zcu.cz

3 Results

A state-of-the-art classifier, named MyModel, was developed. Figure 1 illustrates a comparison between MyModel and published models. Each published model demonstrates optimal performance on its respective dataset when compared to other models. For instance, the iThermo model exhibits the highest accuracy on its test dataset (see Fig. 1a), but its performance notably declines on the DeepTP dataset compared to other models (see Fig. 1b). In Fig. 1b, the iThermo model performs much worse than the SAPPHIRE model, despite its superior performance on its own dataset.

MyModel consistently demonstrates superior performance across most of the testing datasets. Notably, all proteins in the datasets used for comparison were excluded from the training dataset.



a) iThermo

model	accuracy
iThermo	96.3
SAPPHIRE	94.2
Susanty	92.3
MyModel	97.0

b) DeepTP

model	accuracy
DeepTP	81.4
SAPHIRE	78.9
iThermo	73.4
SCMTPP	76.5
TMPpred	69.6
MyModel	90.7

c) SCMTPP / SAPPHIRE

model	accuracy
SAPPHIRE	94.2
SCMTPP	86.5
ThermoPred	86.0
MyModel	92.0

* average over 4 datasets (MyModel the best model for 3 out of 4 sets)

Figure 1: Accuracy comparison on multiple test datasets from published papers. A table title corresponds to the publication the dataset was provided by.

4 Conclusion

A newly fine-tuned transformer model for protein thermostability classification underwent testing on published datasets and was compared to other classification models. The model exhibited the most robust performance in classifying proteins into two classes. However, the primary emphasis during fine-tuning was on classification into four classes, which offers more detailed insights into protein characteristics and can be leveraged for subsequent downstream tasks.

Acknowledgement

This work was supported by the University of West Bohemia grant, project No. SGS-2022-022 and with the support of ELIXIR CZ Research Infrastructure (ID LM2023055, MEYS CR).

References

Ahmed Z et al. iThermo: A Sequence-Based Model for Identifying Thermophilic Proteins Using a Multi-Feature Fusion Strategy. *Front Microbiol. 2022 Feb 22;13:790063.*

Zhao J, Yan W, Yang Y. DeepTP: A Deep Learning Model for Thermophilic Protein Prediction. *Int J Mol Sci. 2023 Jan 22;24(3):2217.*

Charoenkwan P, Schaduangrat N, Moni MA, Lio' P, Manavalan B, Shoombuatong W. SAPPHIRE: A stacking-based ensemble learning framework for accurate prediction of thermophilic proteins. *Comput Biol Med. 2022 Jul;146:105704.*