Acta Pharmaceutica Sinica B 2025;15(6):3343−3345

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

HIGHLIGHT

PRIME, a temperature-guided language model revolutionizes protein engineering

KEY WORDS

Protein engineering;
Directed evolution;
Pre-trained language model;
Deep learning;
Thermal stability;
Enzyme activity;
Protein language model;
Mutant fitness prediction

Recently, a novel protein language model (PLM) was published by Liang Hong group in *Science Advances*[1], introducing PRIME (PRotein language model for Intelligent Masked pretraining and Environment prediction, Fig. 1). PRIME is a deep learning model designed to predict and improve protein stability and activity without relying on experimental mutagenesis data. This innovative approach leverages a vast dataset of 96 million protein sequences annotated with their host bacterial optimal growth temperatures (OGTs) to develop a model that effectively guides protein engineering across various applications.

Protein engineering for pharmaceutical and industrial applications faces several major challenges. Traditional methods, such as directed evolution and rational design, typically demand extensive experimental screening or deep mechanistic insights into protein structures and functions[2,3]. In recent years, PLMs have emerged as promising tools for protein engineering[4]. However, many existing PLMs struggle to recommend mutations that enhance both stability and activity, two critical properties for engineered proteins.

PRIME successfully addressed these challenges by offering a data-driven approach that predicts promising mutations to increase both stability and activity without relying on experimental data. The model's architecture is built on a transformer-based encoder, augmented with two specialized modules: one for Masked Language Modeling (MLM[5]) and another for OGT prediction[6]. This setup enables the model to capture the fundamental relationship between sequences and temperature-related attributes that are crucial for the stability and function of proteins, making it particularly advantageous for engineering industrial enzymes or proteins that need high-temperature tolerance and resilience in practical applications.

One of the most notable strengths of PRIME lies in its "zero-shot" capability, which allows it to identify beneficial mutations for a given protein without any experimental data. The authors compared PRIME's zero-shot performance against several state-of-the-art models, including deep learning approaches such as, SaProt[7] and Stability Oracle[8], as well as traditional computational methods like GEMME[9] and Rosetta[10].

Across 283 protein assays, PRIME demonstrated superior performance in predicting changes in melting temperature (ΔT_m) and excelled in the ProteinGym benchmark[11], which encompasses diverse protein properties including catalytic activity, binding affinity, stability, and fluorescence intensity. Notably, PRIME achieved a score of 0.486 on the ProteinGym benchmark, significantly surpassing the second-best model, SaProt, which scored 0.457 ($P = 1 \times 10^{-4}$, Wilcoxon test).

To validate PRIME's efficacy, the authors conducted wet-lab experiments on five distinct proteins: LbCas12a, T7 RNA polymerase, creatinase, nonnatural nucleic acid polymerase, and the variable domain of the heavy chain of a nano-antibody against growth hormone (VHH). PRIME was used to select top-ranking single-site mutants for each protein. Remarkably, over 30% of these mutations demonstrated notable improvements in physicochemical properties, such as thermostability, catalytic activity, binding affinity, or resilience to extreme alkaline conditions and the ability to polymerize nonnatural nucleic acids.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.04.010
2211-3835 © 2025 The Authors. Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

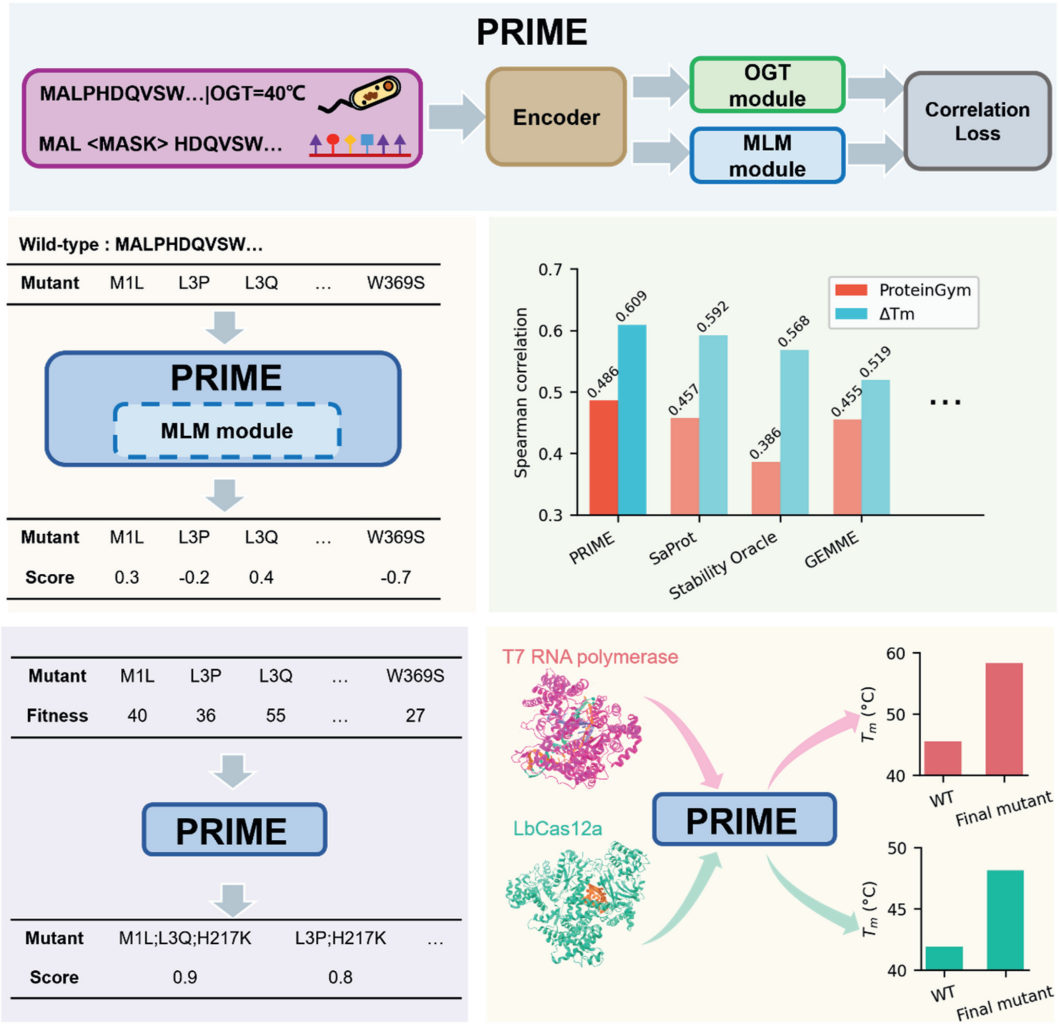

Figure 1 Training with optimal growth temperature data, PRIME achieves state-of-the-art protein mutation fitness prediction. It has been successfully applied to the directed evolution of multiple proteins, enhancing their stability and activity.

The effectiveness of PRIME was further demonstrated through the engineering of LbCas12a and T7 RNA polymerase. For LbCas12a, a complex multidomain protein with 1228 amino acids, PRIME guided an iterative optimization process through three rounds of mutagenesis and experimental validation. In the final round, all 30 multisite mutants exhibited higher melting temperatures (T_m) than the wild type. The best-performing eight-site mutant achieved a Tm of 48.15 °C, representing a significant 6.25 °C improvement over the wild type. The engineering of T7 RNA polymerase further showcased PRIME's capabilities. Aiming to enhance the enzyme's thermostability for applications such as mRNA vaccine production and isothermal amplification detection techniques, the team conducted the AI-guided mutagenesis and wet-lab validation of 95 mutants. This process successfully yielded a 12-site mutant with a melting temperature 12.8 °C higher than the wild type.

Notably, in both the LbCas12a and T7 RNA polymerase projects, PRIME demonstrated the ability to selectively combine certain individually negative single-site mutations into positive multi-site mutants. Such epistatic insights are typically elusive in conventional protein engineering but proved crucial for generating superior variants.

These case studies illustrate PRIME's efficiency in protein engineering. PRIME was able to guide the development of notable improved enzyme variants in just a few rounds of mutagenesis. This approach not only enhances the precision of protein engineering but also substantially reduces the time and resources required for experimental validation.

Still, several limitations warrant further exploration. The reliance of PRIME on bacterial OGTs may restrict its applicability to certain protein families. Additionally, integrating structural information or combining PRIME with other computational methods could expand its applications in drug development, enzyme design, and synthetic biology. As researchers continue to refine and adapt PRIME, it holds great promise for transforming how we discover, design, and optimize proteins in a growing range of industrial and pharmaceutical applications.

Author contributions

Yuanxi Yu: Writing – original draft. Qianhui Wang: Writing – review & editing. Yike Zou: Writing – review & editing, Conceptualization.

References

1. Jiang F, Li M, Dong J, Yu Y, Sun X, Wu B, et al. A general temperature-guided language model to design proteins of enhanced stability and activity. *Sc Adv* 2024;**10**:eadr2641.
2. Jiang K, Yan Z, Di Bernardo M, Sgrizzi SR, Villiger L, Kayabolen A, et al. Rapid *in silico* directed evolution by a protein language model with EVOLVEpro. *Science* 2025;**387**:eadr6006.
3. Woolfson DN. A brief history of *de novo* protein design: minimal, rational, and computational. *J Mol Biol* 2021;**433**:167160.
4. Ruffolo JA, Madani A. Designing proteins with language models. *Nat Biotechnol* 2024;**42**:200−2.
5. Devlin J, Chang MW, Lee K, outanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171−86.
6. Li G, Rabe KS, Nielsen J, Engqvist MK. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth Biol* 2019;**8**:1411−20.
7. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F. Saprot: protein language modeling with structure-aware vocabulary. *bioRxiv* 2023. https://doi.org/10.1101/2023.10.01.560349.
8. Diaz DJ, Gong C, Ouyang-Zhang J, Loy JM, Wells J, Yang D, et al. Stability oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nat Commun* 2024;**15**:6170.
9. Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol Biol Evol* 2019;**36**:2604−19.
10. Das R, Baker D. Macromolecular modeling with rosetta. *Annu Rev Biochem* 2008;**77**:363−82.
11. Notin P, Dias M, Frazer J, Marchena-Hurtado J, Gomez AN, Marks D, et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In: *Proceedings of the 39th International Conference on Machine Learning, PMLR*. **162**; 2022. p. 16990−7017.

Yuanxi Yu[b], Qianhui Wang[a], Yike Zou[a,b,*]

[a]*School of Pharmaceutical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*

[b]*Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai 201203, China*

*Corresponding author.

E-mail address: zouyike@sjtu.edu.cn (Yike Zou)

Received 6 January 2025

Received in revised form 25 January 2025

Accepted 26 January 2025