Poster

Tuning LLM models to predict protein's 3D structure

Varun Kohli[1]

[1]*Google, Zurich, Switzerland*

varunkohli@google.com

Predicting a protein's 3D structure from its amino acid sequence remains a fundamental challenge in structural biology, with wide-ranging applications in drug discovery and disease understanding. An impressive milestone has been reached in protein research, with the structures of around 100,000 unique proteins now determined [1] through a substantial experimental effort [2, 3]. This achievement paves the way for further advancements in understanding protein function and opens up vast possibilities for drug discovery and disease treatment. While the billions of known protein sequences still hold many secrets to be uncovered [4], this significant progress demonstrates the incredible potential of continued research in this field. While AlphaFold1 [5] and AlphaFold2 [6] have demonstrated remarkable success in this domain, the potential of Large Language Models (LLMs) like Gemini, a cutting-edge model developed by Google, remains largely untapped.

This work explores the tuning of Gemini to enhance its ability to predict protein 3D structures. Building upon the foundational work of AlphaFold2 and leveraging the Protein Data Bank (PDB) as a comprehensive source of experimentally determined protein structures for training and validation, we investigate fine-tuning and various parameter-efficient fine-tuning (PEFT) strategies, including Low-Rank Adaptation (LoRA) and other adapter-based methods. We incorporate protein sequence representations, structural motifs, evolutionary information, and domain-specific knowledge relevant to protein folding and stability.

The performance of our fine-tuned Gemini model is rigorously evaluated against AlphaFold2 and other state-of-the-art methods using benchmark protein datasets. We hypothesize that the unique architecture and capabilities of Gemini, combined with our specialized fine-tuning techniques, will lead to improved protein structure prediction accuracy and efficiency.

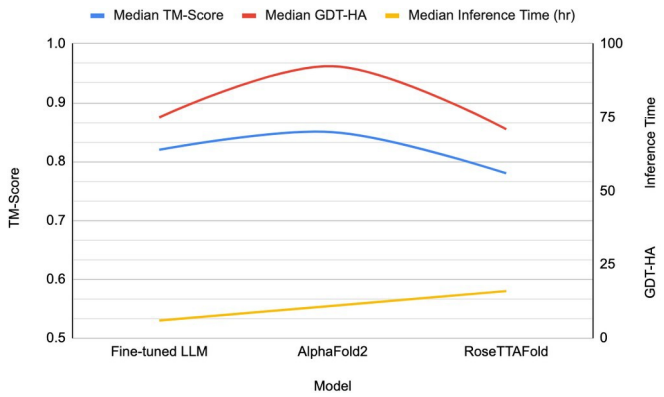

Figure 1. Performance comparison between fine-tuned LLM model and other protein structure prediction methods.

We anticipate that our findings will not only contribute to a deeper understanding of the relationship between protein sequences and their 3D structures but also unlock the full potential of Gemini and Large Language Models for protein structure prediction. This advancement promises to accelerate progress in diverse fields, including drug design and personalized medicine.

[1] Berman, H. M., et al. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235-242.

[2] Thompson, M. C., Yeates, T. O. & Rodriguez, J. A. Advances in methods for atomic resolution macromolecular structure determination. F1000Res. 9, 667 (2020).

[3] Bai, X.-C., McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. Trends Biochem. Sci. 40, 49– 57 (2015).

[4] Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 48, D570–D578 (2020).

[5] Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

[6] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, et al. Improved protein structure prediction using potentials from deep learning. Nature, 577 (2020), pp. 706-710.