Brazilian Symposium on Bioinformatics 2024

Enhancing Enzyme Generation with Fine-Tuned Conditional Transformers

Marco Nicolini[1], Emanuele Saitto[1], Rubén Jiménez[4], Emanuele Cavalleri[1], Marco Mesiti[1], Aldo Galeano[4], Dario Malchiodi[1], Alberto Paccanaro[4,5], Peter N. Robinson[2,3], Elena Casiraghi[1,2], and Giorgio Valentini[1,2]

1. AnacletoLab, Dipartimento di Informatica, Universita degli Studi di Milano, Italy
2. ELLIS – European Laboratory for Learning and Intelligent Systems
3. Berlin Institute of Health at Charité (BIH), Berlin, Germany
4. Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil
5. Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway University of London, Egham, UK

We introduce *Finenzyme*, a Protein Language Model (PLM) that models specific Enzyme Commission (EC) categories by integrating transfer learning from a decoder-based Transformer, conditional learning using specific functional keywords, and fine-tuning. Using *Finenzyme*, we analyze how fine-tuning improves the prediction and generation of EC categories. Our findings reveal a two-fold reduction in perplexity for EC-specific categories compared to a general model, showing that fine-tuning helps capture specialized enzymatic functions that are not well represented in general models. We evaluated the generated enzymes using state-of-the-art tools such as ESMFold [1] for structure prediction and Foldseek [2] for structural similarity assessment. Despite low sequence identity, the generated proteins exhibit high structural resemblance to natural enzymes. Functional characterization using the CLEAN [3] tool confirms that the generated enzymes maintain the same EC functions as natural enzymes. Additionally, we demonstrate that the embedded representations of the generated enzymes closely resemble those of natural ones, making them suitable for downstream tasks such as enzyme classification and functional annotation. Clustering analysis reveals that the generated enzymes form clusters that largely overlap with those of natural enzymes, indicating that *Finenzyme* effectively captures the structural and functional properties of target enzymes. Finally, we showcase a practical application of *Finenzyme* in generating enzymes with specific functions using in-silico directed evolution – a computationally efficient fine-tuning methodology that significantly enhances and can assist targeted enzyme engineering tasks.

Keywords: *Large Language Models, Protein Language Models, Conditional Transformers, Enzyme design and modelling, Protein engineering*

References

1. Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123–1130.
2. Van Kempen, M., et al. (2024). Fast and accurate protein structure search with Foldseek. Nature Biotechnology, 42(2), 243–246.
3. Yu, T., et al. (2023). Enzyme function prediction using contrastive learning. Science, 379(6639), 1358–1363.