In silico enzyme prediction and generation by fine tuning protein language models

Marco Nicolini
AnacletoLab - Dept. of Computer Science
Università degli Studi di Milano
marco.nicolini@unimi.it

Emanuele Saitto
AnacletoLab - Dept. of Computer Science
Università degli Studi di Milano

Ruben E. Jimenez
School of Applied Mathematics (EMAp)
FGV, Rio de Janeiro

Emanuele Cavalleri
AnacletoLab - Dept. of Computer Science
Università degli Studi di Milano

Aldo J. G. Alfonso
School of Applied Mathematics (EMAp)
FGV, Rio de Janeiro

Dario Malchiodi
AnacletoLab - Dept. of Computer Science
Università degli Studi di Milano

Alberto Paccanaro
School of Applied Mathematics (EMAp)
FGV, Rio de Janeiro

Peter N. Robinson
Berlin Institute of Health at Charité (BIH)
Berlin

Elena Casiraghi
AnacletoLab - Dept. of Computer Science
Università degli Studi di Milano

Giorgio Valentini
AnacletoLab - Dept. of Computer Science
Università degli Studi di Milano
valentini@di.unimi.it

Abstract

We present *Finenzyme*, a Protein Language Model (PLM) that integrates transfer learning from a decoder–based Transformer, conditional learning with functional keywords, and fine–tuning to predict and generate functionally characterized enzymes. Our experiments show that *Finenzyme* significantly enhances generalist PLMs like ProGen in the in silico prediction and generation of enzymes belonging to specific Enzyme Commission (EC) categories. Despite low sequence identity, *Finenzyme* generated proteins exhibit high structural similarity to natural enzymes. In silico functional characterization using the CLEAN tool confirms that the generated enzymes maintain the same EC functions as natural enzymes. 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, and can in perspective support targeted enzyme engineering tasks.

1 Introduction

Deep Learning and Large Language Models(LLM) have revolutionized molecular design and modeling, ranging from drug repurposing [1, 2], to active deep learning for drug discovery [3], to protein modeling and generation [4–7].

ML for molecules and materials in the era of LLMs (ELLIS workshop ML4Molecules, December 6, 2024).

In particular several studies showed that fine tuning "generalist" foundation models can improve performance in several modelling and downstream tasks. For instance, parameter-efficient fine-tuning of PLMs improved protein-protein interactions [8]. Lafita et al. [9] focused on fine-tuning ESM family models, showing that fine-tuning PLMs with deep mutational scanning improves variant effect prediction. Fine-tuning the ProtBert-BFD and Prot5-XL-Uniref50 Transformers [10] led to a significant improvement in Gene Ontology and Enzyme Commission (EC) number predictions [11], close to the top-model results of the CAFA3 challenge [12]. Our results on the fine-tuning of the ProGen model on the lysozyme family of proteins achieved a statistically significant improvement in terms of accuracy and perplexity [13], confirming previous results on a similar task [14].

Other studies have outlined the drawbacks and limitations of fine-tuning, demonstrating that this approach sometimes fails to achieve the desired results. In the context of PLMs, Schmirler et al. [15] demonstrated that fine-tuning ESM2 and ProstT5 [16] models boosts predictions in various tasks, ranging from disorder and mutation effects prediction to sub-cellular location prediction, while no statistically significant improvement can be achieved in other tasks, such as secondary structure prediction.

In this work, we explore the specific conditions under which fine-tuning improves the performance of pre-trained PLMs applied to enzyme prediction and generation. To this end we propose a new model, *Finenzyme*, based on a multi-faceted strategy to learn EC (Enzyme Commission) categories of enzymes, by combining PLM transfer learning, conditional learning and fine-tuning. We show that fine-tuning boosts prediction and generation of specific EC categories, whereas for more general EC categories the improvement is negligible. We characterize the enzymes generated by *Finenzyme* models by comparing the 3D representations of natural and generated enzymes using ESMfold [17] and Foldseek [18], showing that the primary and tertiary structure of the enzymes generated by *Finenzyme* models resemble those of natural ones. We further characterize the *Finenzyme*-generated enzymes by clustering the 3D structure representations of natural and generated enzymes, revealing that the resulting clusters are largely superposed. We provide source code to reproduce all experiments, scripts, and tutorials to allow users to fine-tune *Finenzyme* on any EC category or groups of functionally related EC categories.

2 Methods

2.1 Conditional Transformers for protein learning and generation

We used a pre-trained model, i.e. ProGen [14], that adopts the CTRL conditional Transformer architecture [19], which employs keywords to guide the generation of texts. Given a training instance, represented by a protein sequence $x = \{x_1, x_2, \ldots, x_m\}$ and its related keywords t, where $x_i, 1 \leq i \leq m$, represents the amino acids, the model learns through back-propagation the conditioned probability $p(x|t)$, that can be factorized using the chain rule of probability:

$$p(x|t) = \prod_{i=1}^{m} p(x_i|x_{<i}, t) \ , \tag{1}$$

where $p(x_i|x_{<i}, t)$ denotes the conditional probability of x_i given all preceding elements $x_1, \ldots, x_{i-1}$ and the functional tags t.

This formulation breaks down protein language modeling into a next-amino acid prediction task. Consequently, the pre-trained model with parameters θ can be trained to minimize the negative log-likelihood over a dataset of sequences $D = \{(t, x)^{k=1}, \ldots, (t, x)^{k=|D|}\}$:

$$\mathcal{L}(D) = -\sum_{k=1}^{|D|} \sum_{i=1}^{|x^k|} \log p_\theta(x_i^k|x_{<i}^k, t^k) \ . \tag{2}$$

Thus, by acquiring knowledge about the conditional probability distribution, the protein language model can generate new sequences $\tilde{x}$ of length m by sequentially sampling its components: $p_\theta(x_0|t)$, $p_\theta(x_1|\tilde{x}_0, t), \ldots, p_\theta(x_m|\tilde{x}_{<m}, t)$.

2.2 Fine-tuning of conditional Transformers

The goal of fine-tuning is to leverage the knowledge acquired from millions of protein sequences and transfer the "learned knowledge" encapsulated in the weights of the pre-trained ProGen model into a new model fine-tuned to learn the language of a specific EC category of enzymes.

We specialized the models to learn the enzyme characteristics of specific EC classes through a dual approach: a) Conditional decoder learning: generating enzymes conditioned to a specific EC tag; b) Transfer learning and fine-tuning of a pre-trained model on specific EC classes.

For conditioning the models to learn and generate functionally characterized enzymes we prefixed each enzyme sequence with a conditional tag representing the EC class to which the enzyme belongs to. The EC tags for enzyme classes, which were absent in the pre-trained ProGen model, have been encoded and added to the fine-tuned models.

We fine-tuned our models on different EC categories, considering both general top-level and very specific low-level categories. For each EC category, we randomly split the data downloaded from UniProtKB into 90% training and 10% test sets. In addition, for each EC category two test sets were prepared: 1) a full test set that includes all available test data; and b) a filtered test set, with examples having BLAST sequence similarity against the training set less than 70%.

During the prediction, top-k sampling with $k = 1$ was employed, i.e. the most probable next amino acid is selected at each prediction step. In the training phase, the learning rate was set at 0.0001, and the batch size of 2 was used, representing the number of training examples in one iteration. Additionally, a warm-up period of 1000 iterations was implemented, during which the learning rate was gradually increased. To prevent the "exploding gradient" issue common in deep neural networks, gradient norm clipping with a norm of 0.25 was applied. The Adam (Adaptive Moment Estimation) algorithm [20] was utilized to compute adaptive learning rates for each parameter, ensuring efficient optimization. For training and testing the models, we used two multi–processor servers equipped with 128 GB of RAM and an NVIDIA $A100$ GPU accelerator.

3 Results

We studied in which conditions fine–tuning can significantly enhance ProGen's ability to predict and generate sequences belonging to specific EC categories. The primary and tertiary structures of the natural and generated *Finenzyme* enzymes were compared using ESMFold [17], and FoldSeek [18]. We also applied CLEAN [21] to study whether their functions were preserved. We finally clustered natural and generated enzymes according to their predicted 3D structure to investigate whether and to what extent their clusters overlap.

3.1 Fine tuning is effective only for low–level EC classes

We fine-tuned seven models on general top-level EC categories and seven on more specific low-level EC categories, and we tested their performance at predicting the next amino acid in a sequence. The general EC classes represent broad categories of enzyme functions, such as oxidoreductases (EC1), transferases (EC2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), ligases (EC 6), and translocases (EC7). Specific EC classes denote particular enzyme catalysed reactions and in our experiments we used alcohol dehydrogenase (EC 1.1.1.1), DNA methyltransferase (EC 2.1.1.37), cellulase (EC 3.2.1.4), ribulose-bisphosphate carboxylase (EC 4.1.1.39), chorismate mutase (EC 5.4.99.5), biotin ligase (EC 6.3.4.15), and proton-translocating transhydrogenase (EC 7.2.1.1).

We measured the performance of our fine-tuned models at predicting the next amino acid in a sequence in two scenarios : a) Teacher Forcing (TF) in which the LLM is "forced" to predict the next amino acid x_i given the correct previous amino acids, i.e. $x_{<i}$; b) PreFixed testing (PF) in which the LLM makes a prediction without teacher forcing, using a prefix string for the first $n = 20$ amino acids. In both scenarios, the most probable amino acid is selected at each next amino acid prediction step. We measured performance using three different metrics, namely mean accuracy per-token, mean soft accuracy per-token based on BLOSUM62 [22] amino acid substitution matrix, and perplexity. Fig. 1 in the appendix shows that *Finenzyme* largely outperforms ProGen on the most specific EC categories, while no significant improvements can be observed with the most general top-level EC categories.

3.2 *Finenzyme* generated sequences preserve the function and the tertiary structure of natural enzymes

For each specific low-level EC category, we generated 2,000 enzyme sequences using top-p filtering [23], resulting in 14,000 total sequences. After filtering duplicates, 6,885 unique sequences were identified. We compared the sequences generated by *Finenzyme* to natural enzymes in terms of predicted function, primary sequence and tertiary structure.

The generated sequences were classified into EC categories using CLEAN [21], a state–of–the–art tool for enzyme function prediction which compares sequence embeddings to clusters of known EC numbers. CLEAN's maximum separation method yielded consistent EC predictions for the generated sequences (Fig. 2a in Appendix A). This shows that the generated sequences effectively captured functional similarities with natural enzymes.

To evaluate tertiary structure, we predicted the 3D structures of *Finenzyme*–generated enzymes using ESMFold [17] and compared them to natural enzymes from the PDB using Foldseek [18], based on structural similarity scores. The TM–scores, which measure structural similarity, were centered around 0.9 (Fig. 2b in Appendix A), indicating that the structures of the generated enzymes closely resemble those of natural ones, despite differences in sequence similarity. Indeed, the sequence identity of top–hit Foldseek pairs varied more widely, with a mode between 0.3 and 0.4, reflecting the ability of *Finenzyme* to generate structurally accurate enzymes with divergent primary sequences (Pearson correlation between TM–score and sequence identity $\rho = 0.14$).

Further analysis revealed a strong correlation between the confidence of ESMFold predictions (quantified by pLDDT scores) and the TM–scores of the generated proteins, with a Pearson correlation of $\rho = 0.64$ (Fig. 2c in Appendix A). This suggests that the structural predictions of *Finenzyme* are robust, even for sequences with low primary similarity to natural proteins. We also examined specific cases where *Finenzyme*–generated sequences with low sequence similarity still preserved high structural similarity to natural enzymes. Fig. 2d,e in Appendix A highlights two such examples, where *Finenzyme*–generated proteins (in green) closely matched the 3D structure of natural enzymes (in yellow), including accurate predictions of binding sites, which were confirmed to provide sufficient space for ligands (highlighted in violet).

3.3 Clustering of the 3D representation of natural and generated enzymes largely overlap

To further evaluate the relationship between the structure of the natural and *Finenzyme* generated sequences we clustered natural and generated enzymes using Single-Linkage Hierarchical Clustering focusing on EC 3.8.1 (hydrolases acting on halide bonds in C-halide compounds), a family of enzymes of particular interest in the chemical industry and bioremediation fields [24]. The metric used to construct the distance matrix was the E-value retrieved from two experiments that include both the natural and *Finenzyme* sequences: 1) Protein BLAST all-vs-all, which accounts for primary structure similarity; 2) Foldseek all-vs-all, which accounts for both primary and tertiary structure similarity.

In order to provide a more comprehensive database for structural analysis comparison, we predicted the structures for the entire natural dataset using ESMFold. The heatmaps of the adjacency matrices reordered according to [25] clustering are shown in Fig. 3 (see appendix). At the side of each heatmap, kernel-density estimates allow to visually compare the relative distributions of *Finenzyme* sequences and natural proteins. We can observe that each subclass contains sub-clusters, and clusters of natural and *Finenzyme* enzymes largely overlap.

4 Discussion

Previous studies have demonstrated that fine–tuning can improve protein language model (PLM) performance in specific tasks [8, 26, 10], though others have highlighted that it may not always yield better results [27, 15]. In our work, we showed how fine–tuning a conditional Transformer can significantly enhance enzyme generation when applied to specific EC categories. For broad, high–level EC categories, however, the improvements are negligible, as general models like ProGen already perform well on these overrepresented categories due to their extensive training on large datasets. In contrast, fine–tuning shines in underrepresented, functionally specific EC categories, where it allows models to focus on learning detailed and specialized features. Our analyses revealed that the primary sequences generated by *Finenzyme* often diverge from natural enzymes, but crucially,

the tertiary structures remain highly conserved. Additionally, in silico functional predictions using CLEAN confirmed that *Finenzyme*–generated enzymes retain the same functions as their natural counterparts, while clustering of the primary and tertiary structures of the *Finenzyme*-generated and natural EC 3.8.1 enzymes largely overlap.

The correlation between the log–likelihood of *Finenzyme*–generated sequences and structural metrics like TM–score, pLDDT, and sequence identity was significantly stronger than in ProGen, showing that fine–tuning boosts the reliability and confidence of predictions (Fig. 4 in the Appendix).

One limitation of this work is the pure in silico validation of the results, without a full wet-laboratory validation. Although wet-lab validation is planned as future work, we note that the ESMFold predicted structures of *Finenzyme*–generated enzymes are very similar to the natural ones, as measured by the TM-score, suggesting that *Finenzyme* captures essential structural features independently of sequence similarity.

Fine–tuning conditional Transformers through *Finenzyme* can be used to in silico generate specific functionally characterized enzymes, with relatively low computational resources (our models were trained on a server equipped with only one NVIDIA A100 GPU), offering a powerful in-silico tool to support the wet-lab design of specific enzymes.

The *Finenzyme* code, the scripts to reproduce the experiments and tutorials to fine–tune *Finenzyme* on any EC category are available from GitHub (the link has been anonymized).

Acknowledgments and Disclosure of Funding

This work was supported by National Center for Gene Therapy and Drugs Based on RNA Technology—MUR (Project no. CN_00000041) funded by NextGeneration EU program and by CONACYT Paraguay Grants PINV-108 and PINV-719. We declare that there are no competing interests.

References

[1] A. Bessadok and F. Grisoni. An explainable foundation model for drug repurposing. *Nat Med*, 2024. doi: 10.1038/s41591-024-03333-8.

[2] K. Huang, P. Chandak, Q. Wang, et al. A foundation model for clinician-centered drug repurposing. *Nat Med*, 2024. doi: 10.1038/s41591-024-03233-x.

[3] D. vanTilborg and F. Grisoni. Traversing chemical space with active deep learning for low-data drug discovery. *Nat Comput Sci*, 4(7407):786–796, 2024. doi: 10.1038/s43588-024-00697-2.

[4] G Valentini, D Malchiodi, J Gliozzo, M Mesiti, M Soto-Gomez, A Cabri, J Reese, E Casiraghi, and PN Robinson. The promises of large language models for protein design and modeling. *Frontiers in Bioinformatics*, 3:1304099, 2023.

[5] B.L. Hie, V.R. Shanker, D. Xu, et al. Efficient evolution of human antibodies from general protein language models. *Nat Biotechnol*, 2023. doi: 10.1038/s41587-023-01763-2.

[6] N. Ferruz and B. Schmidt, S. a nd Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nat Commun*, 13(4348), 2022. doi: 10.1038/s41467-022-32007-7.

[7] Jeffrey A Ruffolo and Ali Madani. Designing proteins with language models. *Nature Biotechnology*, 42(2):200–202, 2024.

[8] Samuel Sledzieski, Meghana Kshirsagar, Bonnie Berger, Rahul Dodhia, and Juan M. Lavista Ferres. Parameter-efficient fine-tuning of protein language models improves prediction of protein-protein interactions. In *Machine Learning for Structural Biology Workshop, NeurIPS*, 2023.

[9] Aleix Lafita, Ferran Gonzalez, Mahmoud Hossam, Paul Smyth, Jacob Deasy, Ari Allyn-Feuer, Daniel Seaton, and Stephen Young. Fine-tuning protein language models with deep mutational scanning improves variant effect prediction. *arXiv*, 2024.

[10] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

[11] Markus Wenzel, Erik Grüner, and Nils Strodthoff. Insights into the inner workings of transformer models for protein function prediction. *Bioinformatics*, 40(3):btae031, 01 2024. doi: 10.1093/bioinformatics/btae031.

[12] Zhou, N., Jiang, Y., Bergquist, T.R. and others. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol*, 20(244), 2019. doi: 10.1186/s13059-019-1835-8.

[13] M. Nicolini, D. Malchiodi, A. Cabri, E. Cavalleri, M. Mesiti, A. Paccanaro, PN Robinson, J Reese, E Casiraghi, and G Valentini. Fine-Tuning of Conditional Transformers Improves the Generation of Functionally Characterized Proteins. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSTEC*, pages 561–568, 2024.

[14] A Madani, B Krause, ER Greene, S Subramanian, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol.*, 26, 2023. doi: 10.1038/s41587-022-01618-2.

[15] R. Schmirler, M. Heinzinger, and B. Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat Commun*, 15(7407), 2024. doi: 10.1038/s41467-024-51844-2.

[16] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *bioRxiv*, 2024. doi: 10.1101/2023.07.23.550085.

[17] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

[18] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

[19] NS Keskar, B McCann, L Varshney, C Xiong, and R Socher. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[20] D Kingma and J Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `https://api.semanticscholar.org/CorpusID:6628106`.

[21] Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

[22] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv*, 2020. doi: 10.48550/arXiv.1904.09751.

[24] Tatsuo Kurihara and Nobuyoshi Esaki. Bacterial hydrolytic dehalogenases and related enzymes: occurrences, reaction mechanisms, and applications. *The chemical record*, 8(2):67–74, 2008.

[25] Ziv Bar-Joseph, David K Gifford, and Tommi S Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1):S22–S29, 2001.

[26] M Wang, J Patsenker, . Li, Y Kluger, and SH Kleinstein. Supervised fine-tuning of pre-trained antibody language models improves antigen specificity prediction. *bioRxiv*, May 2024. doi: 10.1101/2024.05.13.593807.

[27] Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729, 2023. doi: https://doi.org/10.1016/j.patter.2023.100729.

A Appendix: figures

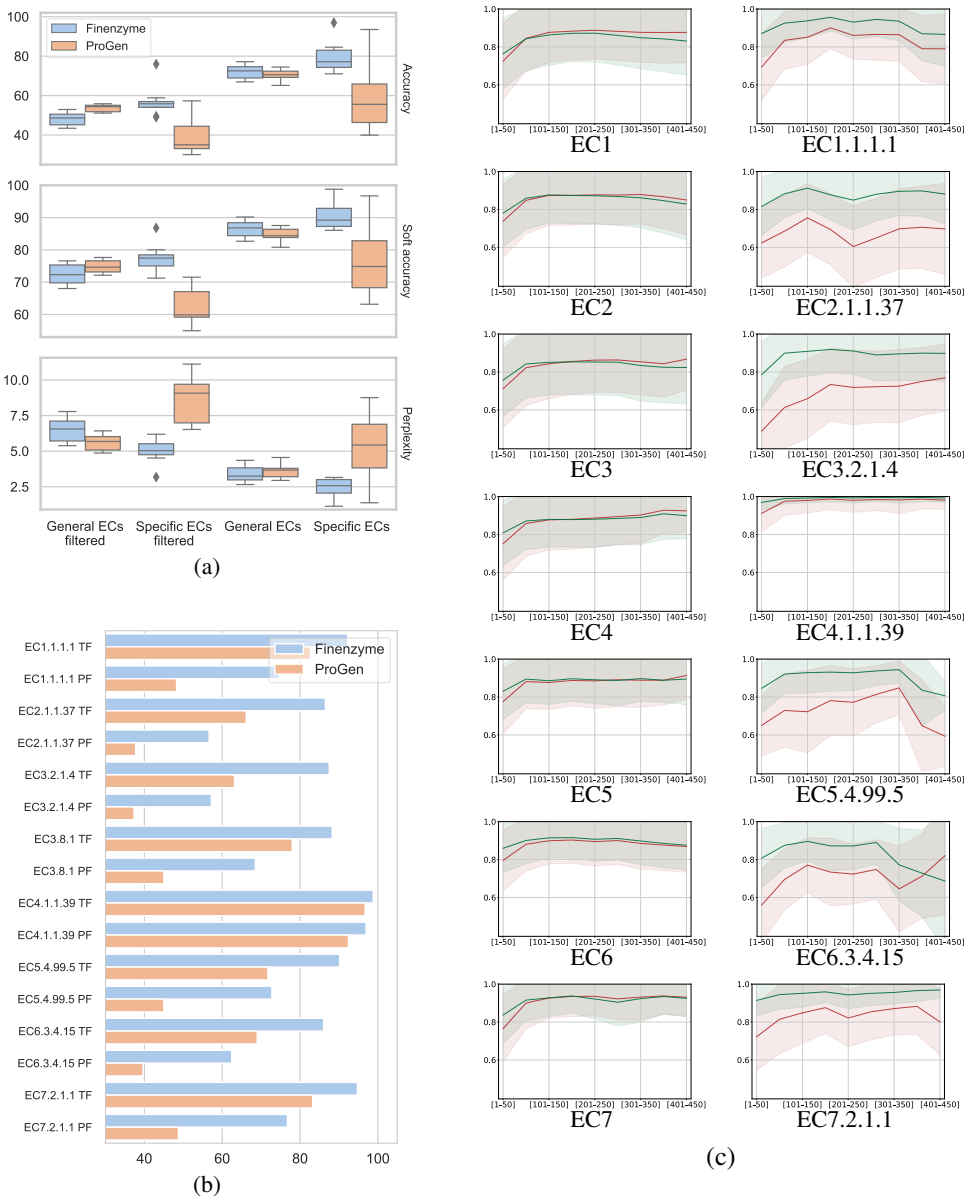

Figure 1: Comparison of ProGen and *Finenzyme* on general and specific EC classes. (a) accuracy (top), soft accuracy (center), perplexity (bottom) compared between general and specific EC classes. "Filtered" denotes whether the test dataset was filtered using BLAST against the training set to retain only sequences with less than 70% identity. (b) soft accuracy of the specific EC classes. "TF" denotes teacher forcing, "PF" testing without teacher forcing with a prefixed chain of 20 amino acids. (c) soft accuracy comparison of *Finenzyme* and ProGen on general (left) and specific (right) EC classes on the full test set. The x-axis reports the position of the amino acid and the y-axis the corresponding average soft-accuracy of the predicted enzymes in ProGen (red line) and *Finenzyme* (green) across the enzymes of the EC category. Shadows represent the standard deviation.

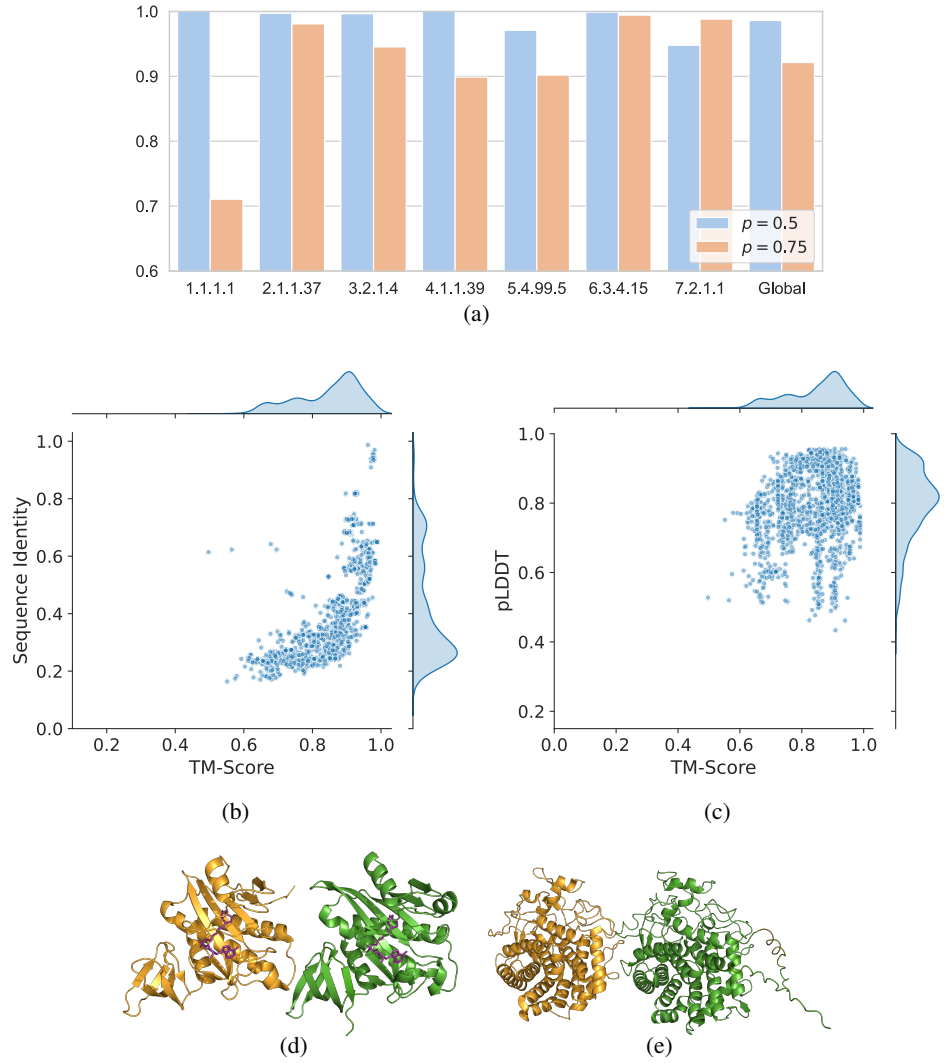

Figure 2: Analysis of the function and tertiary structure of *Finenzyme* generated enzymes. (a): Results of CLEAN EC category predictions. Bar plots show the F1 scores for EC number predictions for the generated enzymes, with blue bars representing results of enzymes generated with $p = 0.50$ and orange bars with $p = 0.75$. "Global" represents the weighted average F1 score when considering all EC numbers combined. (b) Correlation between structural similarity (TM–score) and sequence identity (BLAST Max ID) between generated and natural enzymes retrieved from PDB across all low–level EC classes. (c) Correlation between ESMFold prediction confidence (pLDDT) and structural similarity to known proteins in the PDB (TM–score). Blue scatterplots (b, c) refer to top-$p = 0.5$ nucleus filtering. (d), (e): Comparison of the tertiary structure of *Finenzyme*–generated sequences (green) and natural enzymes (yellow). Enzyme generated from (d) EC family 6.3.4.15 (biotin ligase). The target found is "2e41", a biotin protein ligase (UniProt accession O57883) from *Pyrococcus horikoshii*: sequence similarity $= 34.9\%$, TMscore $= 0.94$, pLDDT (ESMFold) $= 0.95$; the binding site of the enzyme is predicted with high confidence and provides adequate space for the ligand (highlighted in violet). (e) EC 3.2.1.4 (cellulase). The PDB target found is "8ihw", an endoglucanase from *Eisenia fetida*: similarity $= 39.6\%$, TMscore $= 0.95$, pLDDT (ESMFold) $= 0.92$.

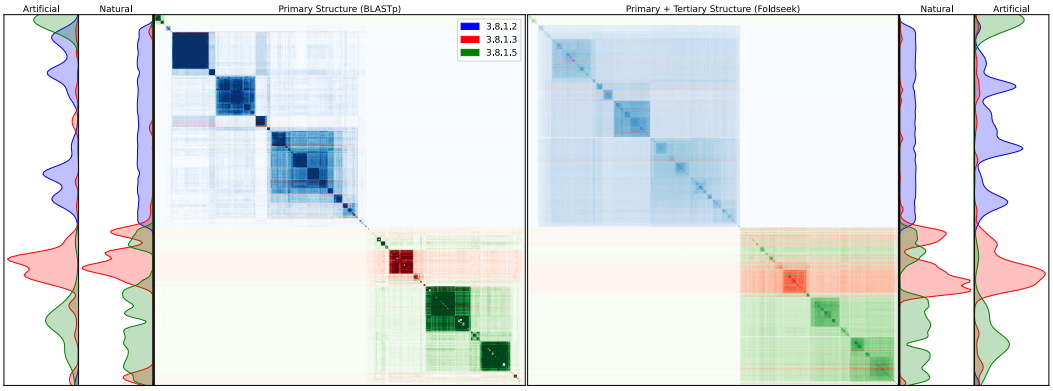

Figure 3: Heatmaps of the clustered natural and *Finenzyme* generated "artificial" enzymes. Different colors highlight different EC categories. Heatmaps are obtained from the adjacency matrix computed through BLAST (left) and Foldseek (right) E-values. On the sides: kernel density estimation of the distribution of natural and *Finenzyme* generated proteins.

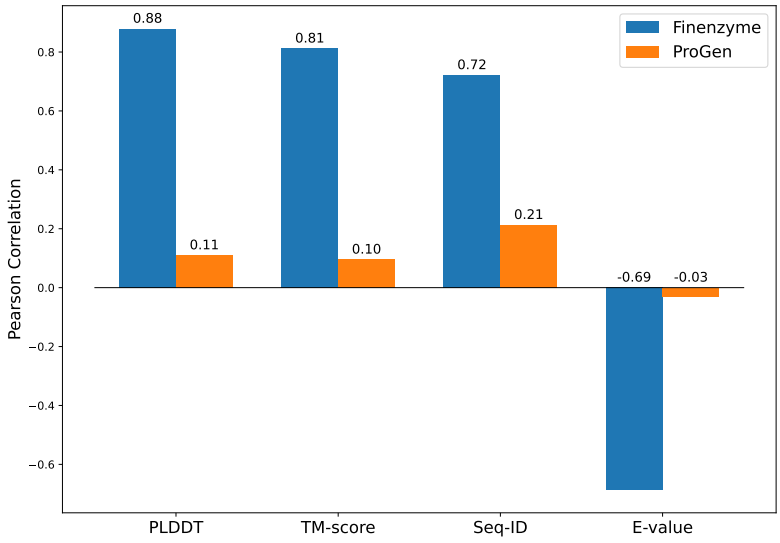

Figure 4: Pearson correlation between the log-likelihood scores of *Finenzyme* (blue) and ProGen (orange) against pLDDT, TM-score, sequence identity and E-value. The metrics are computed through ESMFold and Foldseek.