Language Models for Text-guided Protein Evolution

Zhanghan Ni[1,2*] **Shengchao Liu**[3*] **Hongyu Guo**[4] **Anima Anandkumar**[1]
[1]Caltech [2]Carleton College [3]Independent
[4]National Research Council Canada [*] Equal contribution
nit@carleton.edu, anima@caltech.edu

Abstract

Language models have demonstrated efficacy in protein design by capturing the distribution of amino acid sequences and structures. To advance protein representation learning, biomedical textual description has been integrated as an additional modality, complementing existing sequence and structure information. The textual modality is crucial as it provides insights into detailed molecular functions and cellular contexts in which proteins operate. Existing deep learning methods have built foundation models based on this modality, aiming for challenging protein design tasks, including text-to-protein generation and text-guided protein editing. Meanwhile, the capability of utilizing such multiple modalities to handle natural protein evolution remains an open question. In this work, we introduce two tasks: text-guided point mutation and text-guided Enzyme Commission number switching. These tasks enable a preliminary exploration of the boundaries of utilizing a multimodal foundation model to understand protein evolution process. Our results show that structure-based models outperform sequence-based ones by 24% in structure-oriented evolution tasks, despite exhibiting significant biases. We also find that models using free-form text more effectively design enzyme functions, achieving a 30.06% closer alignment to target functions by integrating evolutionary context. Code is available on this GitHub repository.

1 Introduction

Protein Evolution and Design Protein evolves over time. At the molecular level, the accumulation of mutations that confer an environmental fitness advantage leads to adaptations [7, 21]. This process is driven by natural selection, which increases the frequency of beneficial mutations and decreases that of detrimental ones within a population, and is influenced by genetic drift, which can cause unbiased random changes in gene frequencies irrespective of their impact on fitness [28]. Natural evolution has progressively shaped proteins, increasing their complexity and endowing them with novel biochemical activities, including catalytic functions, as well as enhanced biophysical properties such as stability.

Directed evolution utilizes the evolvability of proteins by consecutively introducing point mutations to optimize protein towards a functional objective. For instance, enzyme cofactor preference could be shifted to enhance the productivity of biosynthetic pathways [4, 12]. Instead of mimicking natural evolution on existing proteins, *de novo* protein design engineers novel proteins with desirable functions from first principles [22]. This allows us to create proteins with functions that have not yet been made by nature, while integrating engineering principles such as modularity and controllability into the design process [11]. For instance, secondary structures can be assembled into proteins with novel fold topologies, and binding motifs from known protein-protein interfaces can be incorporated into *de novo* designed supporting scaffolds [10, 18, 23]. These approaches offer solutions to the

Workshop on Foundation Models for Science, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

protein function design problem, also known as the inverse function problem, which involves finding the sequences and structures that can achieve a desired function [14].

While the results have proven effective, protein function design using laboratory methods often presents challenges. For directed evolution, the laboratory experimental process involves time-consuming trial and error without knowing the functional consequences of the mutations being introduced [5]. Specifically, screening out the beneficial mutations among the hundreds to thousands of random variants in each generation is the bottleneck of directed evolution efforts [24]. Therefore, directed evolution can hugely benefit from using functionally informed mutations instead of random ones to reduce the size of the mutant pool in each generation. Similarly, *de novo* design of proteins with novel functions is challenging due to the as-yet unclear mapping from sequence to function and because foldable and functional proteins occupy only a small proportion of the protein sequence space [24]. In addition, the design process often overlooks the protein's cellular context, which can result in incompatibility with the complex conditions of living systems and may lead to aggregation [5].

LLM-based Protein Design Recent advances in artificial intelligence have profoundly influenced the inverse protein function problem [9, 13, 26, 27, 16]. Large language models (LLMs) excel in learning complex protein representations and can generate realistic proteins [19, 17]. Among these, multimodal language models, which integrate modalities such as sequence, structure, and function, are particularly promising. For instance, works represented by ESM3 can generate proteins conditioned on functional keywords[8], while other works represented by ProteinDT can edit proteins based on free-text descriptions of protein structure and function [15]. Protein function is a modality with great potential to guide function design as it is highly informative of the precise molecular activity and the cellular context in which the protein operates. This aligns well with protein evolution's objective to optimize a set of functional attributes that increase life's environmental fitness, and large language models' reasoning ability can help in making better mutational choices than the random choices made by nature and directed evolution efforts.

Our Contributions: Text-guided Protein Evolution In this study, we extend the capabilities of foundation models by exploring their application in text-guided protein evolution. We specifically investigate how textual information can enhance directed evolution and solve inverse function problem. To this end, we introduce two novel tasks: *text-guided point mutation* and *text-guided Enzyme Commission (EC) number switching*. Respectively, these two tasks examine the ability of language models to facilitate protein evolution and their capacity to comprehend evolutionary information. Our findings indicate that a structure-based model outperforms a sequence-based model by 24% in structure-oriented evolution tasks, although the former exhibited significant biases. Additionally, we demonstrate the utility of using free text to provide evolutionary context, assisting in function design with a 30.06% improvement in alignment with the desired function.

2 Benchmark

2.1 Text-guided Point Mutation

Harnessing the principles of natural selection, directed evolution perform protein function design by introducing predominantly random amino acid substitutions and selecting the beneficial mutations that align with a fitness objective, thus gradually ascending the protein fitness landscape [24]. This iterative process could be significantly accelerated by employing deliberate mutational choices produced by LLMs. In this task, we investigate the potential of using textual information to guide the generation of more effective point mutations. Additionally, we assess trustworthiness of computational screening as a substitute for the laboratory synthesis of mutants.

Dataset Construction ProteinGym is a collection of benchmarks for evaluating the effects of protein substitutions, insertions, and deletion, with data sourced from deep mutational scanning (DMS) studies and clinical cases [20]. We repurposed this dataset by extracting the sequences from DMS substitution studies and the corresponding quantitative evaluations of the mutant for a functional objective.

Task Design As illustrated in Figure 1(a), an ancestral protein sequence x_a from ProteinGym is masked at all possible positions to simulate a DMS study with single point mutations. A textual

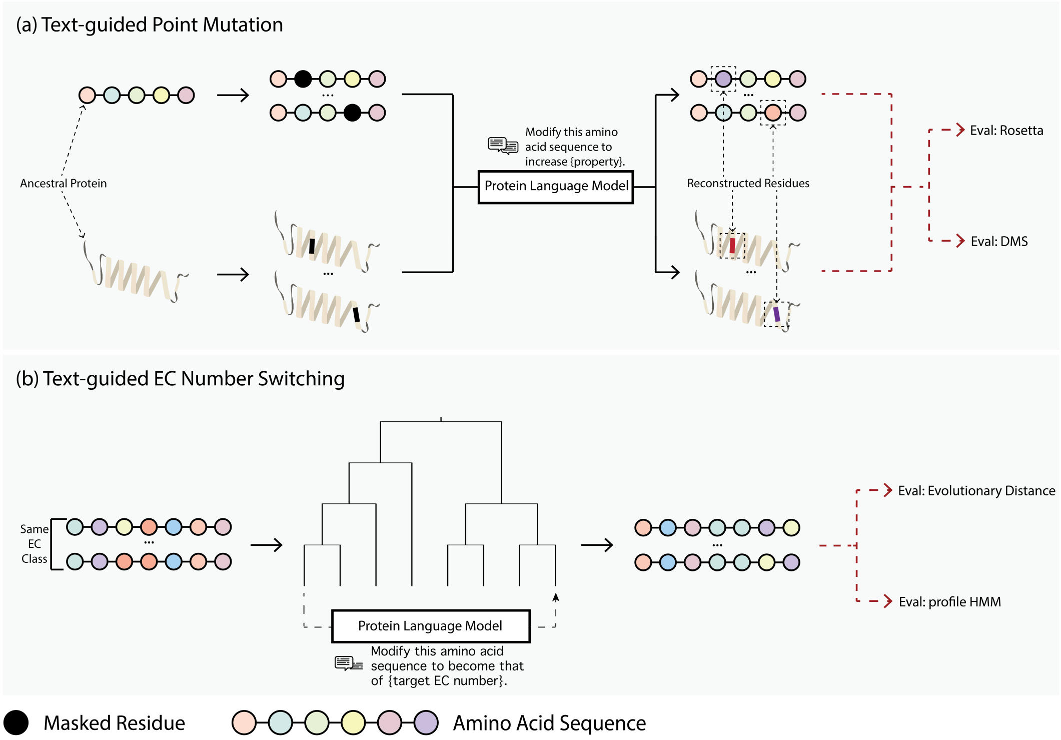
(a) Text-guided Point Mutation
Ancestral Protein
Modify this amino acid sequence to increase {property}.
Protein Language Model
Reconstructed Residues
Eval: Rosetta
Eval: DMS
(b) Text-guided EC Number Switching
Same EC Class
Protein Language Model
Modify this amino acid sequence to become that of {target EC number}.
Eval: Evolutionary Distance
Eval: profile HMM
Masked Residue
Amino Acid Sequence

Figure 1: Benchmarking framework for protein language models in protein evolution tasks. (a) Visualization for text-guided point mutation, showing two tracks: sequence mutation and structure mutation (b) Visualization for text-guided EC number switching. The phylogenetic tree contains evolutionary relationships between different EC numbers at Level 4.

prompt x_t is applied to guide the generation of mutants x_m with protein language models, with $\tilde{x}_m \sim p(x_m \mid x_t, x_a)$. The prompts used are specified in Appendix C.

Evaluation As shown in Figure 1(a), in this task, we mutate ancestral sequences used in a DMS study, which assess a specific functional objective [20]. We employ two metrics to evaluate the generated mutants. The first is a computational oracle for the same functional objective as assessed in the DMS study. Since DMS examines all possible single amino acid changes in a protein sequence, the data from the study can be directly utilized to evaluate the generated mutants, serving as the second metric.

2.2 Text-guided EC Number Switching

Design of novel protein functions necessitates substantial jumps in sequence space, where functional sequences are generally sparse yet often situated near other functional sequences [24]. Consequently, rather than designing a protein with a target function *de novo*, it could be more effective to evolve towards the target protein from ones with similar functions. In this task, we evaluate the ability of LLMs to solve the inverse enzyme function problem through text-guided design, with a particular focus on their ability to leverage textual evolutionary information to facilitate this evolution.

Dataset Construction EC number is a numerical system that classifies enzymes based on the reaction they catalyze, with four levels of progressively finer classification that is highly informative for enzyme function [1, 29]. We extracted enzyme sequences and corresponding metadata from UniProt Knowledgebase [3]. To understand the evolutionary relationship between different EC classes, phylogenetic analysis were conducted on all four EC subclasses. Results suggested that functional closeness does not equate to evolutionary closeness, and the intra-class sequence gap is substantial. Only sequences that share the same third-level EC number are evolutionarily closely related, as shown

Table 1: Prompts for EC number switching

Prompt Type	Information
Type 1	Name of the target enzyme
Type 2	Descriptions of the target EC number at four levels
Type 3	Descriptions of the target EC number at four levels, plus the reaction it catalyzes
Type 4	Descriptions of the target EC number at four levels, plus a list of EC families diverged along the evolutionary path from source to target

in Figure 2. Therefore, the scope of our work is constrained to performing text-guided switching of the fourth-level EC number.

Task Design As depicted in Figure 1(b), a set of ancestral enzymes x_a sharing the same Level 4 EC number is collected. Our objective is to evolve these enzymes into variants that retain the same Level 3 EC number but transition to a different Level 4 EC number. This evolution is facilitated by language models under the guidance of textual prompts x_t, with $\tilde{x}_e \sim p(x_e \mid x_t, x_a)$. The specific prompts are detailed in Appendix C. We categorize these prompts into four types, each incorporating increasing levels of evolutionary and functional information, as summarized in Table 1. The Type 1 prompt includes only the full enzyme name associated with an EC number entry, while the Type 2 prompt also incorporates information spanning from Level 1 to Level 3 [1]. Given the descriptive power of text in detailing precise protein functions, the Type 3 prompt integrates the precise chemical reaction that the enzyme catalyzes. Finally, recognizing the substantial sequence divergence between EC classes, we introduce a Type 4 prompt designed to facilitate EC number transitions by providing descriptions of evolutionarily intermediate EC classes within a phylogenetic tree. This aims to test the models' ability to assimilate evolutionary knowledge and effectively navigate the EC space.

Evaluation We compare the generated enzymes with the target EC family in terms of both evolutionary and functional aspects. To quantify the evolutionary proximity of generated sequences to target enzyme number, we compute the pairwise evolutionary distances between the generated enzymes and the ones from the target EC class. Additionally, we compute the number of generated enzymes that reside inside the monophyletic clade of the target EC class, as shown in Figure 3 in Appendix A. To evaluate whether the generated enzymes possess the function of the target EC class, we derived a profile hidden Markov model (profile HMM) from all sequences associated with the target EC class in the NCBI Protein database [2]. This profile was employed to search for homologs among the generated enzymes. Compared to tree-based metrics, profile HMMs are more sensitive to detecting remote homologies—those that have significantly diverged from the ancestral sequence but retain their functional roles. This is because profile HMMs are particularly effective in identifying conserved functional domains and motifs within a family [30].

3 Experiments

To benchmark the proposed downstream tasks, we evaluate two language models. The first is ESM3, a generative masked language model that incorporates sequence, structure, and text modalities. For our evaluation, we focus specifically on the sequence and structure tracks [8]. The second model is ProteinDT, a text-guided protein sequence editing framework that offers two editing methods: latent interpolation and latent optimization [15]. Both models were assessed across the two tasks. To ensure comparability between sequence-based and structure-based results, we used ESMFold to predict the protein structures from the sequences [13].

3.1 Text-guided point mutation

Here, we present a case study using the eukaryotic caltractin protein, which facilitates the proper assembly and stabilization of microtubules in a wide range of organisms, from yeast and algae to humans [25]. We aim to mimic a single generation in directed evolution studies by introducing single point mutations that enhances the structural stability of the calctractin sequence. Rosetta was used as a computational oracle to compute the energy after structure prediction [6].

Table 2: Text-guided single point mutations evaluated with computational oracle and experimental data. The **best** results are marked. $(+)$ indicates that a prompt for stability increase was used. $(+)$ indicates that a prompt for stability decrease was used. $(\emptyset)$ indicates that no prompt was used to guide the generation. Success rate is defined as the proportion of mutants with stability increased. REU stands for Rosetta Energy Units, with lower value indicating higher stability. DMS score is an experimental measurement of protein stability, with higher value indicating higher stability. Synthesizability is defined as the proportion of mutants that is non-lethal and could be synthesized experimentally. In random mutation, each position was mutated to every possible amino acid residue.

Model	**Computational Oracle**		**Experimental Evaluation**		
	Success Rate	**Δ REU**	**Success Rate**	**Δ DMS Score**	**Synthesizability**
ESM3 Sequence $(+)$	**1.00**	-109.57	**0.94**	0.07	0.65
ESM3 Structure $(+)$	0.96	-92.30	**0.94**	**0.12**	0.42
ESM3 Sequence $(-)$	**1.00**	**-109.59**	**0.94**	0.03	0.67
ESM3 Structure $(-)$	0.97	-93.93	0.93	0.11	0.44
ESM3 Sequence $(\emptyset)$	**1.00**	-109.56	0.93	0.04	0.67
ESM3 Structure $(\emptyset)$	0.97	-93.93	**0.94**	0.11	0.43
ProteinDT Latent Interpolation $(+)$	0.45	28.23	0.63	-0.53	0.93
ProteinDT Latent Optimization $(+)$	0.12	10.06	0.70	-0.40	0.59
Random Mutation $(\emptyset)$	0.40	24.86	0.67	-0.42	**0.98**

The evaluation of the generated mutants is presented in Table 2. Under computational oracle evaluation, all variants of ESM3 consistently outperform ProteinDT in enhancing structural stability, achieving up to a 55% higher success rate and 488% greater reduction in Rosetta Energy Units. In experimental evaluations using data from the original DMS study, the non-lethal mutants exhibit a 24% higher success rate under ESM3, with a 130% greater increase in DMS scores. However, ESM3 exhibits a discernible bias towards stability optimization, as evidenced by increased stability even absent a prompt or against a prompt to decrease stability.

Our results suggest that LLMs are not yet able to accelerate directed evolution efforts. In the evaluation of non-lethal mutants, while there is consistency observed between the computational and experimental evaluations for ESM3 mutants, ProteinDT demonstrates higher success rates in experimental evaluation, with its latent optimization showing a 58% increase in success rate. Further analysis of the computational oracle's confidence is detailed in Table 4. In the evaluation of lethal mutants, the current computational oracle is unable to identify any non-lethal mutants. Although the random mutation baseline exhaustively explores the mutant space, revealing that 98 percent of mutations are non-lethal and indicating high mutation tolerance in caltractin, a large percentage of mutants generated by ESM3 are lethal. ESM3 tends to sample the small proportion of lethal mutants, resulting in a synthesizability 26% lower than that of ProteinDT. As a result, although the lethal mutants proposed by ESM3 all lead to increased stability, on average, 46.33% of the time a completely non-synthesizable mutant will be proposed by ESM3. While ProteinDT's latent interpolation method proposes 93% lethal mutants, its success rate is 4% lower than the random mutation among the lethal ones it proposes. Therefore, neither model could directly aid directed evolution in proposing better mutational choices.

3.2 Text-guided EC number switching

Here, we present a case study of the two models applied to the more challenging protein evolution task: EC number switching. We focus on the third-level EC class 1.1.1, which are oxidoreductases that acts on the CH-OH group of donors and has NAD^+ or $NADP^+$ as acceptor [1]. EC 1.1.1.1 was selected as the source EC class and three other classes of varying evolutionary distances were selected as the target EC classes.

The generated enzymes were evaluated based on their evolutionary relationships to sequences from the target EC class, as shown in Table 3. Notably, ESM3-generated enzymes displayed the furthest evolutionary distance from the target EC class while achieving the highest monophyletic ratio. This suggests that ESM3 proteins capture a clade-specific pattern of mutational accumulation that positions them within the target clade, despite diverging from the target EC family in other conserved regions by sampling underexplored regions of sequence space. This finding is consistent with the lack of significant hits in profile HMM searches for the ESM3 enzymes, indicating that conserved

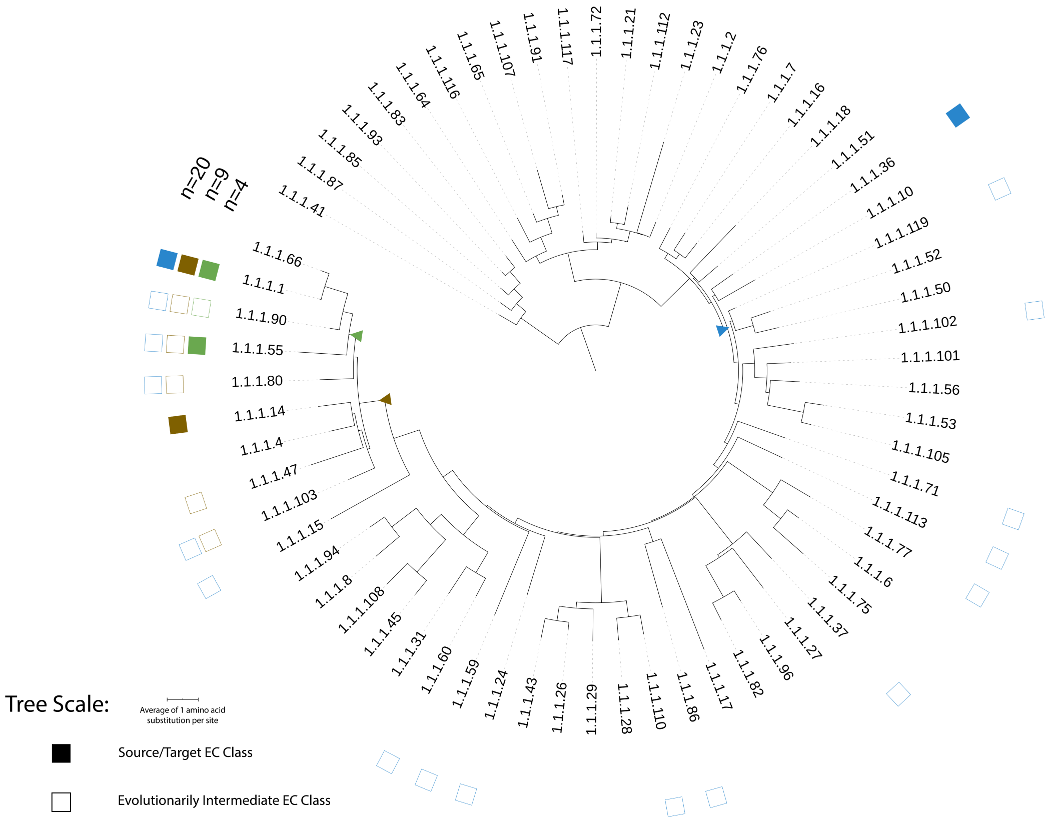

Figure 2: Evolutionary relationships of enzymes with EC numbers starting with 1.1.1. Triangles signify the most recent common ancestors of each pair of EC classes. Each color corresponds to a different EC number switching case study, showcasing three case studies with varying evolutionary distances. The label *n* denotes the number of bifurcations required to transition from the source EC to the target EC class.

local motifs, which are essential for enzyme function and typically captured by profile HMMs, are not represented in the sequences generated by ESM3. In contrast, ProteinDT-generated enzymes demonstrated closer evolutionary distances to proteins from the target EC family, with significant profile HMM hits. This alignment suggests that ProteinDT captures key conserved motifs across the entire target EC class, resulting in high overall sequence similarity and, consequently, a shorter evolutionary distance. However, the proteins exhibited a low monophyletic ratio, indicating a distinct sequence of mutational accumulation that does not fully align with the evolutionary path specific to the target EC protein family—an essential feature for placement within the monophyletic group.

Overall, using prompts that are more informative of functional and evolutionary context reduces the evolutionary distance between generated enzymes and those from the target EC class. Providing information about the catalyzed reaction offers a marginal benefit in this regard. Specifically, Type 3 prompts, which include additional chemical reaction information compared to Type 2 prompts, result in a 3.4% reduction in evolutionary distance. In contrast, incorporating evolutionary information is more beneficial. The evolutionary context provided by Type 4 prompts, for EC numbers 1.1.1.55 and 1.1.1.36, facilitated the generation of sequences with the shortest evolutionary distances to the target EC class among the four prompt types. This suggests that providing additional evolutionary context may guide the model to generate sequences that are evolutionarily closer to the target class. However, for EC number 1.1.1.14, the prompt containing only the enzyme name was most effective. This may indicate that, for certain enzymes, the name itself is sufficiently informative, specifying the type of reaction the enzyme catalyzes, thus making additional functional information in Types 2 and 3 prompts unnecessary. Nonetheless, on average, providing evolutionary context to ProteinDT with Type 4 prompts resulted in enzymes with a 12.31% shorter evolutionary distance than when

Table 3: Phylogenetic analysis and profile HMM evaluation of text-guided EC number switching. Enzymes from the source EC class 1.1.1.1 were attempted to be switched to three different target EC classes of varying evolutionary distance under the guidance of four types of prompts in Table 1. In phylogenetric tree based evaluation, evolutionary distance is defined as the average pairwise distance between the sequences of the two classes in a tree, with number of amino acid substitutions per site as unit. A shorter distance is better. Monophyletic ratio is defined as the proportion of generated enzymes that resides in the clade of the target EC class in a phylogenetic tree as shown in Figure3. In profile HMM based evaluation, a lower E-value indicates a higher confidence for the enzyme being a homologous sequence of the target EC class and a higher score indicates a better sequence alignment of the generated enzyme with the target EC sequences. A hyphen indicates no significant hit was found by the profile HMM search.The **best** results are marked.

Model	**Target EC Class**	**Prompt Type**	**Phylogenetic Tree**		**Profile HMM**	
			Evolutionary Distance	**Monophyletic Ratio**	**E-value**	**Score**
ProteinDT	1.1.1.55	1	2.12	0.00	1.81E-08	61.68
ProteinDT	1.1.1.55	2	3.28	0.00	1.81E-08	61.68
ProteinDT	1.1.1.55	3	3.11	0.00	1.75E-08	61.71
ProteinDT	1.1.1.55	4	**2.10**	0.00	1.81E-08	61.68
ESM3	1.1.1.55	1	3.37	**0.28**	-	-
ProteinDT	1.1.1.14	1	**1.93**	0.28	0.042	77.72
ProteinDT	1.1.1.14	2	2.07	0.16	0.042	77.73
ProteinDT	1.1.1.14	3	2.02	0.28	0.042	77.72
ProteinDT	1.1.1.14	4	2.31	0.16	0.042	77.72
ESM3	1.1.1.14	1	2.63	**0.44**	-	-
ProteinDT	1.1.1.36	1	2.82	0.08	-	-
ProteinDT	1.1.1.36	2	2.85	0.08	-	-
ProteinDT	1.1.1.36	3	2.80	0.08	-	-
ProteinDT	1.1.1.36	4	**2.78**	0.08	-	-
ESM3	1.1.1.36	1	4.28	**0.40**	-	-

no evolutionary context was provided in Type 2 prompts. Comparing the two models, ProteinDT-generated enzymes are 30.06% closer to the target EC family compared to those generated by ESM3 without such context.

Additionally, it seems challenging for ProteinDT to switch to a distant EC class (e.g., n=20), as there were no significant hits at all from profile HMM searches when attempting to switch to 1.1.1.36. This difficulty may arise because large evolutionary and functional distances between the source and target EC classes demands substantial changes in secondary and tertiary structures. While capable of making localized sequence alterations required by switching to closely related EC, ProteinDT may not be adept at capturing complex structural and functional relationships with protein sequences.

4 Conclusion

In this work, we introduced two novel protein evolution tasks: text-guided point mutation and text-guided EC number switching. In the point mutation task, among synthesizable mutants, ESM3 models that incorporate protein structure modality outperformed the sequence-based model, ProteinDT, in structure-oriented evolution tasks but demonstrated a strong bias toward increasing protein stability. However, neither of the two models is sufficiently effective to assist in directed evolution, as neither achieves both a high synthesizability and a high success rate. In the EC number switching task, ESM3 excelled at exploring novel sequence spaces while possessing clade-specific mutational patterns. In contrast, ProteinDT generated sequences with motifs that are highly conserved among proteins in the target EC class and displayed a closer evolutionary distance to these proteins. Our findings suggest the benefits of leveraging evolutionary information for more effective enzyme function design, which is a capability afforded by models supporting free-text modality. Moving forward, developing models that integrate both free-text and structure modalities could enhance performance in protein engineering tasks.

References

[1] Enzyme Nomenclature, . URL https://iubmb.qmul.ac.uk/enzyme/.

[2] Proteins - NCBI, . URL https://www.ncbi.nlm.nih.gov/home/proteins/.

[3] UniProt, . URL https://www.uniprot.org/help/uniprotkb.

[4] Sabine Brinkmann-Chen, Tilman Flock, Jackson K. B. Cahn, Christopher D. Snow, Eric M. Brustad, John A. McIntosh, Peter Meinhold, Liang Zhang, and Frances H. Arnold. General approach to reversing ketol-acid reductoisomerase cofactor dependence from NADPH to NADH. *Proceedings of the National Academy of Sciences of the United States of America*, 110(27): 10946–10951, July 2013. ISSN 1091-6490. doi: 10.1073/pnas.1306073110.

[5] H. Adrian Bunzel, J. L. Ross Anderson, and Adrian J. Mulholland. Designing better enzymes: Insights from directed evolution. *Current Opinion in Structural Biology*, 67:212–218, April 2021. ISSN 0959-440X. doi: 10.1016/j.sbi.2020.12.015. URL https://www.sciencedirect.com/science/article/pii/S0959440X21000075.

[6] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics (Oxford, England)*, 26(5):689–691, March 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq007.

[7] Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.

[8] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1. Pages: 2024.07.01.600583 Section: New Results.

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

[10] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker. Principles for designing ideal protein structures. *Nature*, 491(7423):222–227, November 2012. ISSN 0028-0836. doi: 10.1038/nature11600. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3705962/.

[11] Tanja Kortemme. De novo protein design—From new structures to programmable functions. *Cell*, 187(3):526–544, February 2024. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2023.12.028. URL https://www.cell.com/cell/abstract/S0092-8674(23)01402-2. Publisher: Elsevier.

[12] Yanjing Li, Hannan Xu, Haiteng Zhao, Hongyu Guo, and Shengchao Liu. Chatpathway: Conversational large language models for biology pathway detection. In *NeurIPS 2023 AI for Science Workshop*, 2023.

[13] 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, March 2023.

doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

[14] Dina Listov, Casper A. Goverde, Bruno E. Correia, and Sarel Jacob Fleishman. Opportunities and challenges in design and optimization of protein function. *Nature Reviews Molecular Cell Biology*, 25(8):639–653, August 2024. ISSN 1471-0080. doi: 10.1038/s41580-024-00718-y. URL https://www.nature.com/articles/s41580-024-00718-y. Publisher: Nature Publishing Group.

[15] Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A Text-guided Protein Design Framework, February 2023. URL https://arxiv.org/abs/2302.04611v3.

[16] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023.

[17] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, August 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2. Publisher: Nature Publishing Group.

[18] Enrique Marcos, Tamuka M. Chidyausiku, Andrew C. McShan, Thomas Evangelidis, Santrupti Nerli, Lauren Carter, Lucas G. Nivón, Audrey Davis, Gustav Oberdorfer, Konstantinos Tripsianes, Nikolaos G. Sgourakis, and David Baker. De novo design of a non-local -sheet protein with high stability and accuracy. *Nature structural & molecular biology*, 25(11): 1028–1034, November 2018. ISSN 1545-9993. doi: 10.1038/s41594-018-0141-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219906/.

[19] Lewis Moffat, Shaun M. Kandathil, and David T. Jones. Design in the DARK: Learning Deep Generative Models for De Novo Protein Design, January 2022. URL https://www.biorxiv.org/content/10.1101/2022.01.27.478087v1.

[20] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *Advances in Neural Information Processing Systems*, 36:64331–64379, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html.

[21] H. Allen Orr. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. *Evolution*, 52(4):935–949, 1998. ISSN 0014-3820. doi: 10.2307/2411226. URL https://www.jstor.org/stable/2411226. Publisher: Oxford University Press.

[22] L. Regan and W. F. DeGrado. Characterization of a helical protein designed from first principles. *Science (New York, N.Y.)*, 241(4868):976–978, August 1988. ISSN 0036-8075. doi: 10.1126/science.3043666.

[23] Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science (New York, N.Y.)*, 357(6347):168–175, July 2017. ISSN 1095-9203. doi: 10.1126/science.aan0693.

[24] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews. Molecular Cell Biology*, 10(12):866–876, December 2009. ISSN 1471-0080. doi: 10.1038/nrm2805.

[25] Sudha Veeraraghavan, Patricia A. Fagan, Haitao Hu, Vincent Lee, Jeffrey F. Harper, Bessie Huang, and Walter J. Chazin. Structural Independence of the Two EF-hand Domains of Caltractin *. *Journal of Biological Chemistry*, 277(32):28564–28571, August 2002. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M112232200. URL `https://www.jbc.org/article/S0021-9258(20)70251-6/abstract`. Publisher: Elsevier.

[26] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. Scaffolding protein functional sites using deep learning. *Science (New York, N.Y.)*, 377(6604):387–394, July 2022. ISSN 1095-9203. doi: 10.1126/science.abn2100.

[27] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL `https://www.nature.com/articles/s41586-023-06415-8`. Publisher: Nature Publishing Group.

[28] S. Wright. The roles of mutations, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the 6th International Congress of Genetics*, volume 1, pages 356–366, Menasha, WI, 1932. Brooklyn Botanical Garden.

[29] Jason Yang, Ariane Mora, Shengchao Liu, Bruce J. Wittmann, Anima Anandkumar, Frances H. Arnold, and Yisong Yue. CARE: a Benchmark Suite for the Classification and Retrieval of Enzymes, June 2024. URL `http://arxiv.org/abs/2406.15669`. arXiv:2406.15669 [q-bio].

[30] Byung-Jun Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, 10(6):402–415, September 2009. ISSN 1875-5488. doi: 10.2174/138920209789177575.

A Appendix: Phylogenetic Anlaysis

Phylogenetic trees were used to investigate the evolutionary relationship between the generated, source, and target enzymes in the text-guided EC number switching task. Enzyme sequences were downloaded from NCBI Protein, and then aligned and trimmed using Muscle v5.1 and TrimAl v1.4. Maximum likelihood trees were inferred using `IQ-TREE v2.0.3` with 1,000 ultrafast bootstraps and a substitution model selected by `ModelFinder`. To ensure inclusion of all mixture models, the following flags were used: `-mrate E,I,G,I+G,R` and `-madd C10,C20,C30,C40,C50,C60,EX2,EX3,EHO,UL2,UL3,EX_EHO,LG4M,LG4X,CF4`. EC numbers that do not form monophyletic groups were pruned to ensure the uniqueness of ancestry for each EC number represented in the tree.

In the text-guided EC number switching task, the evolutionary relationship between the generated enzymes and the target enzymes are analyzed. A monophyletic ratio is defined as the proportion of generated enzymes that resides inside the clade of the target EC sequences. For example, Output 10 and Output 11 reside in the monophyletic EC 1.1.1.36 clade, as shown in Figure 3.

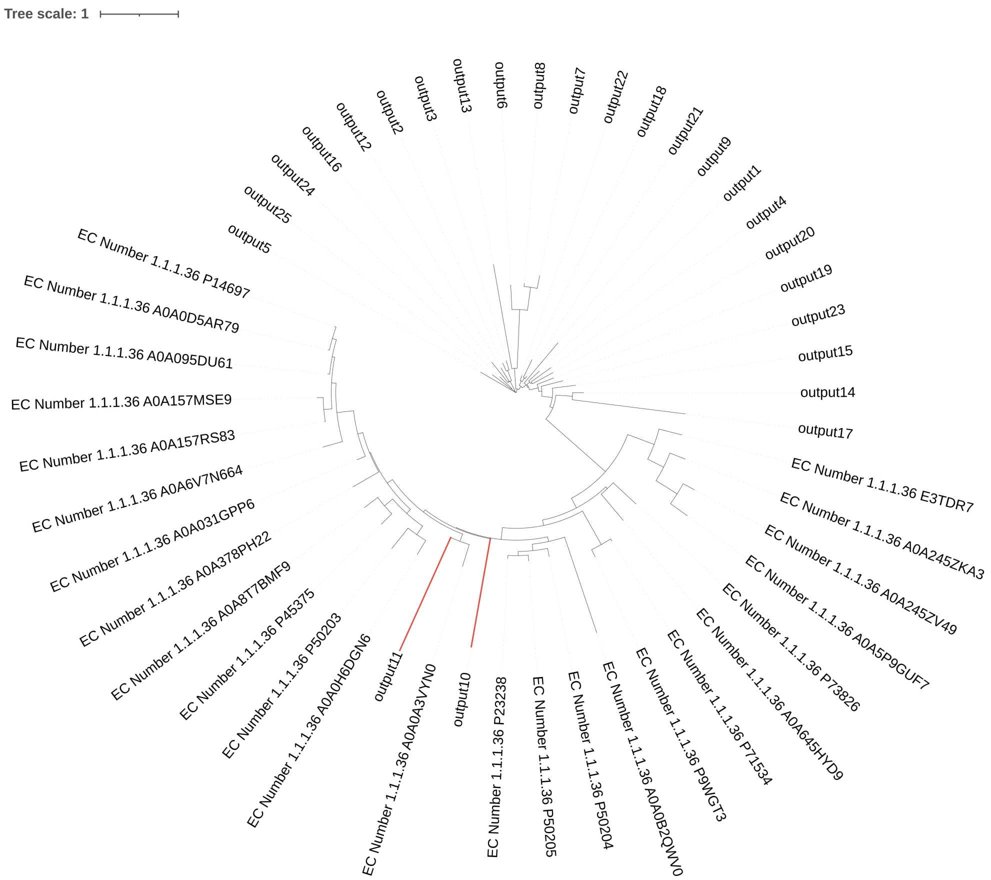
Tree scale: 1
output5
output25
output24
output16
output12
output2
output3
output13
output6
output8
output7
output22
output18
output21
output9
output1
output4
output20
output19
output23
output15
output14
output17
EC Number 1.1.1.36 E3TDR7
EC Number 1.1.1.36 A0A245ZKA3
EC Number 1.1.1.36 A0A245ZV49
EC Number 1.1.1.36 A0A5P9GUF7
EC Number 1.1.1.36 P73826
EC Number 1.1.1.36 A0A645HYD9
EC Number 1.1.1.36 P71534
EC Number 1.1.1.36 P9WGT3
EC Number 1.1.1.36 A0A0B2QWV0
EC Number 1.1.1.36 P50204
EC Number 1.1.1.36 P50205
EC Number 1.1.1.36 P23238
output10
EC Number 1.1.1.36 A0A0A3VYN0
output11
EC Number 1.1.1.36 A0A0H6DGN6
EC Number 1.1.1.36 P50203
EC Number 1.1.1.36 P45375
EC Number 1.1.1.36 A0A8T7BMF9
EC Number 1.1.1.36 A0A378PH22
EC Number 1.1.1.36 A0A031GPP6
EC Number 1.1.1.36 A0A6V7N664
EC Number 1.1.1.36 A0A157RS83
EC Number 1.1.1.36 A0A157MSE9
EC Number 1.1.1.36 A0A095DU61
EC Number 1.1.1.36 A0A0D5AR79
EC Number 1.1.1.36 P14697

Figure 3: Visualization of monophyletic EC class clade. Output 10 and Output 11, which are highlighted in red, reside within the monophyletic EC 1.1.1.36 clade.

B Appendix: Oracle Confidence

Table 4: Confidence of REU Oracle. Concordance is the percentage of results from the computational oracle that are consistent with experimental outcomes in DMS study. $(+)$ indicates a prompt for stability increase, $(-)$ indicates a prompt for stability decrease, and $(\emptyset)$ indicates that no prompt was used to guide the generation.

Model	**Concordance**
ESM3 Sequence $(+)$ REU	0.674
ESM3 Structure $(+)$ REU	0.673
ESM3 Sequence $(-)$ REU	0.674
ESM3 Structure $(-)$ REU	0.677
ESM3 Sequence $(\emptyset)$ REU	0.674
ESM3 Structure $(\emptyset)$ REU	0.676
ProteinDT Latent Interpolation $(+)$ REU	0.561
ProteinDT Latent Optimization $(+)$ REU	0.401
Random Mutation $(\emptyset)$ REU	0.556

C Appendix: Prompts and Generation Condition

Text-guided point mutation: ESM-3 A textual description is applied to specific sequence tokens or structural tokens at the masked mutation position(s), as shown in Table 5. The generation was conditioned either on sequence, structure, and text, or solely on sequence and structure. Two editing methods were utilized: sequence generation followed by structure prediction, and direct structure generation.

Text-guided point mutation: ProteinDT A textual description is applied to the entire sequence, as shown in Table 5, and the protein is edited using latent optimization and latent interpolation.

Table 5: Prompts for Text-guided Point Mutation

Model	**Prompt**
ESM-3	Functional Keywords: Increase; Stability.
ProteinDT	Modify the amino acid sequence to have higher stability.

Text-guided EC number switching Prompts used in this task are shown in Table 6.

Table 6: Prompts for Text-guided EC number switching

EC Number	Type	Prompt
EC1.1.1.14	Type1	Modify this amino acid sequence to become that of sorbitol dehydrogenase.
	Type2	Modify this amino acid sequence to become that of sorbitol dehydrogenase, which also is an enzyme classified as an oxidoreductase, acting on the CH-OH group of donors, with NAD or NADP as acceptor.
	Type3	Modify this amino acid sequence to become that of sorbitol dehydrogenase, which catalyzes the conversion of L-iditol to L-sorbose.
	Type4	Modify this amino acid sequence to that of sorbitol dehydrogenase. The following enzymes become increasingly similar to sorbitol dehydrogenase: aryl-alcohol dehydrogenase, lactaldehyde reductase, isopropanol dehydrogenase, L-threonine 3-dehydrogenase, and glucose 1-dehydrogenase.
	ESM3	sorbitol; dehydrogenase
EC1.1.1.55	Type1	Modify this amino acid sequence to become that of lactaldehyde reductase.
	Type2	Modify this amino acid sequence to become that of lactaldehyde reductase, which also is an enzyme classified as an oxidoreductase, acting on the CH-OH group of donors, with NAD+ or NADP+ as acceptor.
	Type3	Modify this amino acid sequence to become that of lactaldehyde reductase, which catalyzes the conversion of propane-1,2-diol to L-lactaldehyde.
	Type4	Modify this amino acid sequence to that of lactaldehyde reductase, considering that aryl-alcohol dehydrogenase is more similar to lactaldehyde reductase than the original sequence.
	ESM3	lactaldehyde; reductase
EC1.1.1.36	Type1	Modify this amino acid sequence to become that of acetoacetyl coenzyme A reductase.
	Type2	Modify this amino acid sequence to become that of acetoacetyl coenzyme A reductase, which also is an enzyme classified as an oxidoreductase, acting on the CH-OH group of donors, with NAD+ or NADP+ as acceptor.
	Type3	Modify this amino acid sequence to become that of acetoacetyl coenzyme A reductase, which catalyzes the conversion of (R)-3-hydroxyacyl-CoA to 3-oxoacyl-CoA.
	Type4	Modify this amino acid sequence to that of acetoacetyl coenzyme A reductase. The following enzymes become increasingly similar to acetoacetyl coenzyme A reductase: aryl-alcohol dehydrogenase, lactaldehyde reductase, isopropanol dehydrogenase, L-threonine 3-dehydrogenase, D-iditol 2-dehydrogenase, 2-hydroxy-3-oxopropionate reductase, 3-hydroxypropionate dehydrogenase, quinate dehydrogenase, aromatic 2-oxoacid reductase, ketol-acid reductoisomerase, malate dehydrogenase, lactaldehyde reductase, propanediol:NAD oxidoreductase, L-xylose 1-dehydrogenase, alcohol dehydrogenase, 3-dehydrosphinganine reductase, and glucose 1-dehydrogenase.
	ESM3	acetoacetyl coa; reductase