Reinforcement Learning for Protein Motif Scaffolding Design

Team Members: Jordan Cahoon, Yaowei Deng
Emails: cahoon@stanford.edu, yaowei@stanford.edu

Extended Abstract

Motivation

The motif-scaffolding problem is an open challenge in protein design where "scaffolds" are constructed around a user-defined "motif." Successfully performing this conditional generation will enable users to design proteins with specific functions with applications in medicine, sustainability, engineering, and more. However, due to the biophysical heterogeneity and expansive design space of all possible motifs, it has been challenging to develop and evaluate methods that perform this task well. Recently, MotifBench, a collection of hard-to-solve motif test cases, was released to provide a benchmark toward standardized evaluation for motif scaffolding. Current pre-trained protein language models (pLMs) are known to perform poorly on this benchmark, despite performing well on other protein tasks (e.g., unconditional backbone generation). In this project, we apply a popular paradigm of finetuning language models known as preference optimization to improve scaffold generation with pLMs by aligning generation to correct scaffolds.

Methods

We apply two preference optimization methods, direct preference optimization (DPO) and identity preference optimization (IPO) to the baseline ESM3 model (1.4B parameters): ESM3-open. We finetune ESM3-open for three select test cases from MotifBench. In the first stage of training, we align ESM3 on offline preferences for each motif. Then using the finetuned model, we generate online preferences and conduct a second stage of finetuning. To our knowledge, this is the first time preference optimization has been applied to the motif-scaffolding problem.

Implementation & Results

We generate offline and online preference datasets for three of MotifBench test cases. Optimal scaffolds are selected by motif preservation, biological plausibility, and sequence foldability. These scaffolds are paired with negative examples using the same motif coordinates. Each dataset contains 1.2-1.8k preference pairs. ESM3-open is finetuned using 8 epochs, using batch size 8 and 8 gradient accumulation steps. We also use gradient clipping to improve training stability.

After the first stage of finetuning, we observe scaffold success rate improve by 7% and 1% with IPO and DPO respectively. After a second stage of finetuning, we observe a drop in performance by -22% and -16% with IPO and DPO respectively compared to baseline. Additionally, we find that first-stage finetuned models improve motif preservation ($p < 0.001$).

Discussion & Conclusion

Our results demonstrate that preference optimization can improve scaffold generation with ESM3-open, however more work will be needed to evaluate this for pLMs with more parameters and more motifs. Despite this, our results show that finetuned ESM3-open has a higher scaffold success rate and that overall scaffolds better preserve motif function. However, we observe that two-stage finetuning has an adverse effect on performance on this task. While our project is a proof of concept, and we recognize that our results are highly sensitive to choice of motif, we believe that preference optimization may be a viable method to improve pLM scaffold generation and ultimately the ability to flexibly design proteins for a variety of domains.

Abstract

We present reinforcement learning (RL)-based methods for approaching the motif-scaffolding problem, a canonical conditional generation task in protein design. We apply direct preference optimization (DPO) and identity preference optimization (IPO) to ESM3-open for generating scaffolds to support three challenging test cases presented in MotifBench, a new benchmark for evaluating the performance of generative models on the motif-scaffolding problem. We curate 1.2-1.8k offline and online preference pairs for each motif test case and finetune ESM3 using these data. With single-stage finetuning, we see an average success rate increase of 7% and 1% with IPO and DPO, respectively. However, we see that performance drops drastically with the scaffolds generated from the two-stage finetuned models. These results suggest that two-stage finetuning may lead to the models overfitting to the entire backbone structure at the expense of motif fidelity. Our findings highlight a promising direction for motif scaffolding using reinforment learning techniques.

1 Introduction

Protein language models (pLMs) have revolutionized novel protein design. By conditioning generation of amino acid sequences and structure (scaffolds) on a set of high conserved regions (motifs) of the protein, researchers can design novel proteins with desired function or stability. This constitutes the motif-scaffolding problem: the goal of identifying diverse protein structures preserve the motif and maintain its geometry. However, there remains an open challenge of how these motifs should be placed in the surrounding scaffold generated by the pLM.

Some current methods [1, 2] perform this conditional generation by relying on user-defined motif placements around which to generate the scaffold. Others [3, 4] determine placements at sampling time from a set of possible arrangements along the generation prompt.

We aim to apply reinforcement learning (RL) to the motif-scaffolding problem to determine optimal motif coordinates and guide scaffold generation to preserve motif functions while prioritizing correctness and novelty. We will use test cases presented in MotifBench [5] for examples of motifs to scaffold and ESM3 [6] to perform the generation. We will score our generated designs using the evaluation pipeline presented in MotifBench.

2 Related Work

2.1 pLMs

Pre-trained pLMs excel in identifying patterns and relationships within protein subunits, known as *residues*, and can be used for a variety of downstream tasks including, but not limited to, *de novo* protein design, structure prediction, functional annotation, and backbone generation. The input to these models is a set of residues with coordinates, with the exact position of the residues impacting the resulting generated structures. Our work would be one of the first to consider residue placement for conditional generation with pLMs.

2.2 RL & pLMs.

In recent years, RL has been used to guide pLMs to design *de novo* proteins with specific properties. The developers of the pre-trained pLM, ESM3[6], have demonstrated that direct preference optimization (DPO) [7] can generate out-of-distribution structures that deviate from training instances. Other works have expanded on this idea by using both online and offline RL that optimize function while maintaining biological plausibility [8] [9] [10]. While these methods improve the novelty of generated proteins, our work would be the first to apply RL specifically for conditional generation, which requires preserving the functionality of the motif. ProteinRL generates structures for both single- and multi- objective design functions using online RL to fine-tune using a prespecified reward function to optimize for desired properties. DPO_pLM is a offline RL framework that uses an oracle that uses fitness and fold similarity to guide sequence generation. In contrast, ESM-PF trains an efficient proxy reward model to a mutation policy to generate biologically plausible *de novo* sequences.

2.3 Motif-scaffolding Problem.

Protein structure generation conditioned on motif segments is a canonical problem in the protein design field. To accommodate motifs that catalyze specific reactions, generative methods [1] [6] use residue-level information (protein geometry, amino acid identity, position, side-chain rotamer) and sequence indices as input. Recent work [11] takes a different approach: the model uses only the atomic coordinates of the catalytic site and infers indices and side-chain rotamers dynamically. Acknowledging the importance of motif placement, our work will take advantage of RL's ability to explore large combinatorial spaces to optimize motif coordinates for generation.

3 Methods

To utilize recent advancements in reinforcement learning with human feedback (RLHF), we formulate this problem as a preference optimization task where we align ESM3 to prefer generating good scaffolds y_{good} over bad scaffolds (y_{bad}). We focus on two methods, Direct Preference Optimization (DPO) ([8]) and Identity Preference Optimization (IPO) ([12]). These methods take a dataset $\mathcal{D}$ $(y_{good}, y_{bad}|x)$ which contains a prompt x and a pair of good and bad completions, (y_{good}, y_{bad}). Using the dataset and a reference policy π_{ref}, the target policy π_θ is optimized. DPO optimizes the the target policy π_θ using the following loss:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_{good},y_{bad})\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_{good}|x)}{\pi_{ref}(y_{good}|x)} - \beta\log\frac{\pi_\theta(y_{bad}|x)}{\pi_{ref}(y_{bad}|x)}\right)\right]$$

IPO optimizes the following loss:

$$\mathcal{L}_{IPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_{good},y_{bad})\sim\mathcal{D}}\left[\left(\log\frac{\pi_\theta(y_{good}|x)}{\pi_{ref}(y_{good}|x)} - \log\frac{\pi_\theta(y_{bad}|x)}{\pi_{ref}(y_{bad}|x)} - \frac{\beta^{-1}}{2}\right)^2\right]$$

Both techniques train a policy that increases the likelihood of generating positive examples, y_{good}, and decreases likelihood of generating negative examples y_{bad} (Figure 1). Unlike DPO, IPO adds regresses the gap between positive and negative examples to a constant, preventing overfitting to the preference dataset. In our particular application y refers to the generated scaffold and x refers to the motif coordinates.

4 Experiments

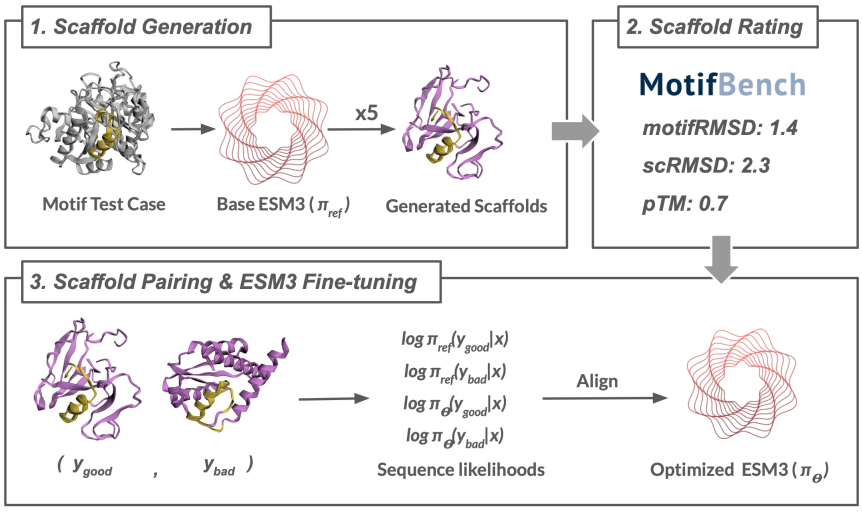
4.1 Offline dataset generation.

We focused on three test cases of varying difficulties from MotifBench [5], 1LDB, 6E6R, and 3TQB. We conditioned on these motifs to generate scaffolds by sampling random motif placements for protein sequence and structure generation with ESM3-open (1.4B parameters) ([6]). For each motif placement, we batch generate five scaffolds with different sequences. We repeat this sample-then-generate process 1000 times to create a total of 5000 structures for each motif (Figure 1).

4.2 Assigning preferences

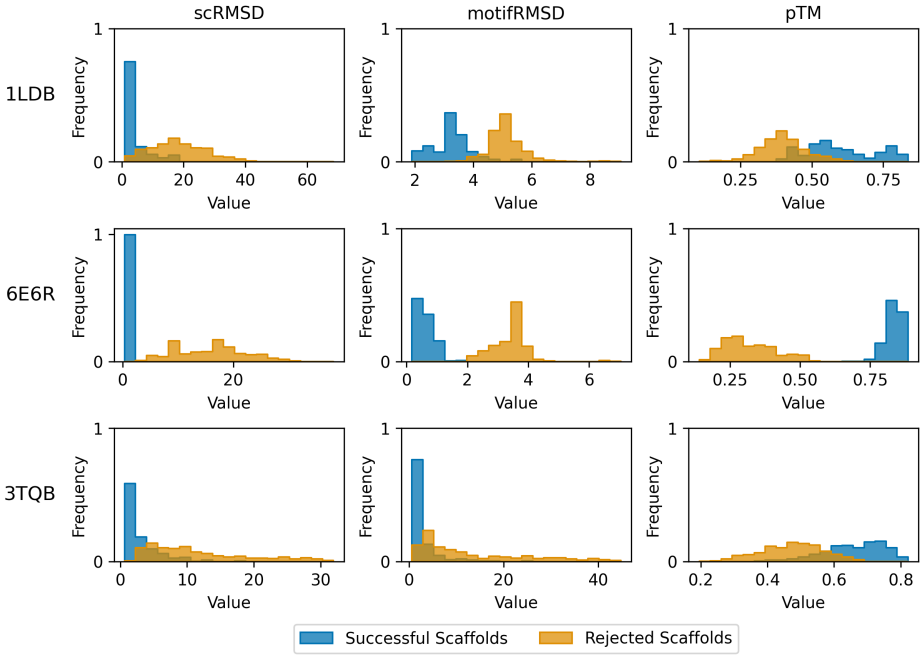
We assign preferences based on scaffold correctness due to the poor performance of ESM3-open on the test cases. Based on previous work, preference pairs (y_{good}, y_{bad}) are determined by the following metrics: motif maintenance ($motifRMSD$), scaffold validity (self-consistency RMSD, or $scRMSD$) ([13]), and predicted template modeling (pTM). Successful scaffolds, y_{good}, have the following properties $pTM_{good} \geq 0.5$, $motifRMSD_{good} \leq 3.5$, and $scRMSD_{good} \leq 3$ for the first test case (1LDB), $pTM_{good} \geq 0.8$, $motifRMSD_{good} \leq 1$, and $scRMSD_{good} \leq 2$ for the second test case (6E6R), $pTM_{good} \geq 0.6$, $motifRMSD_{good} \leq 2$, and $scRMSD_{good} \leq 2.5$ for the third test case (3TQB). Different thresholds are used to account for the relative difficulty of the test case. For each pair (y_{good}, y_{bad}), we enforce gaps to ensure the positive sample is sufficiently better

Figure 1: Preference optimization pipeline



than the negative sample, which are sampled from the same motif placements as the positive examples (Figure 1). Overall, successful scaffolds have improved rewards across all metrics compared to that of rejected scaffolds (Figure 2). In total, the preference datasets contains 1.2k, 1.8k, 1.6k pairs, respectively for 1LDB, 6E6R, and 3TQB.

Figure 2: Distribution of each metric for scaffolds generated by base ESM3 for the test cases, stratified by scaffold success.



4.3 Aligning Scaffold Generation to Motifs

We finetune ESM3-open for each of our select motifs with IPO and DPO. After the first stage of preference optimization, we collect more samples using the finetuned policy and conduct a second

round of training with the new preferences. We use the same cutoffs as described in previous sections and preference optimization techniques, but with preferences generated using the finetuned policies. For each stage of training, we finetune the ESM3-open model for 8 epochs with batch size 8 and 8 gradient accumulation steps. We also use gradient clipping to improve training stability. This results in a total of four models per motif, for the two stages of training and the two methods of preference optimization.

4.4 Evaluation

We evaluated our trained models on a set of 100 motif placements for the three test cases presented in MotifBench: 1LDB, 6E6R, and 3TQB. We then run the MotifBench scoring pipeline to evaluate the samples generated from our RL-finetuned ESM3. To capture the ability of the finetuned generative model, we use the MotifBench scoring:

$$\text{MotifBench score} = \frac{1}{\text{\# test cases}} \sum_{i=1}^{\text{\# test cases}} (100+\alpha)\frac{\text{\# unique solutions for case } i}{\alpha + \text{\# unique solutions for case } i},$$

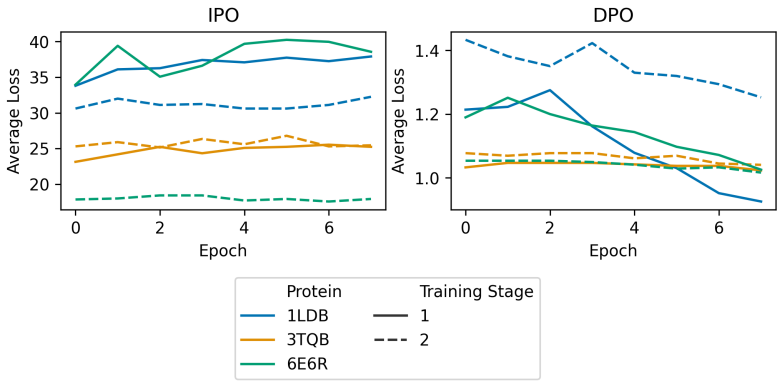
where $\alpha = 5$. Successful scaffolds are defined as $motifRMSD_{good} \leq 1$, and $scRMSD_{good} \leq 2$. We report the success rate as the number of successful scaffolds out of the 100.

5 Results

5.1 Preference Optimization Improves Scaffold Generation

Because evaluation through MotifBench is costly, we evaluate the trained model after completing 8 epochs of training. We see that training decreases average DPO loss for both test cases, but is not stable for IPO loss (Figure 3). This could be explained by the small difference between the log-likelihoods for successful and rejected scaffolds. After a second stage of finetuning with the online preference pairs, we notice that the average loss for the second stages of training is lower for 6E6R, about the same of 3TQB, and variable for 1LDB compared to the first stages of training. We observe the loss for 3TQB is generally lower than that of the other test cases likely due to the noisy preferences where the reward metrics are overlapping (Figure 2).

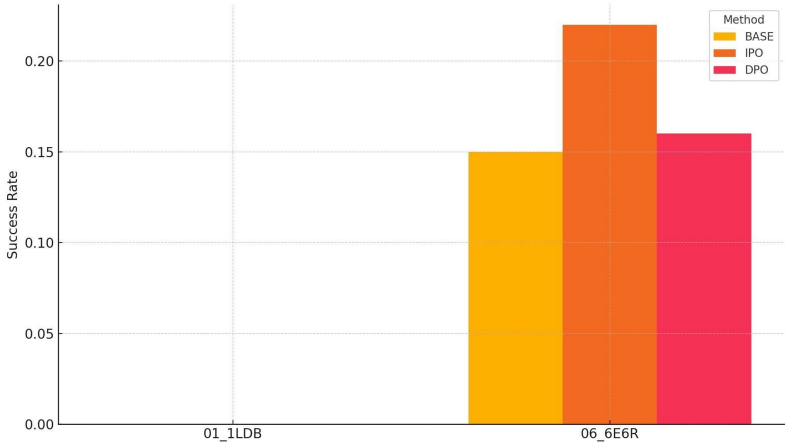
Figure 3: Training loss curves across 8 epochs for the motif test cases. Training stage prefers to the round of finetuning the model is currently on.



We observe that the number of correct scaffolds is greater for the DPO and IPO than the base model for the motif, 6E6R, which is expected due to the relatively lower difficulty level (Figure 4). However, there are no successful scaffolds for 1LDB after applying preference optimization. This is likely because no successful demonstrations existed in the preference dataset. When examining the individual metrics for the evaluation placements, we see $scRMSD$ is higher for DPO and IPO methods compared to that of the baseline (Table 1). We suspect that this is due to the preference for higher sequence compatibility versus geometric compatibility. For 6E6R, we observe a significant

decrease in $motifRMSD$ compared to baseline, whereas for 1LDB we do not see this effect (Table 1).

Figure 4: Success rate by method and test case



(a) 01_1LDB

(b) 06_6E6R

Metric	Base	IPO	DPO
RMSD	13.95	15.56*	15.20**
Motif RMSD	5.01	5.00	5.02
pTM	0.44	0.42	0.43

Metric	Base	IPO	DPO
RMSD	6.85	8.32***	7.89**
Motif RMSD	2.64	2.37***	2.47***
pTM	0.54	0.54	0.54

Table 1: Mean metric values for each method and test case. Asterisks denote significance vs. Baseline: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.2 Two-stage Training with Online Preferences

In a second stage of finetuning, we incorporate online preferences. After evaluating the set of two-stage-trained samples, we observe further reductions in DPO loss but at the expense of motif preservation. Average $motifRMSD$ rose in all cases, resulting in zero "successes" as defined by the MotifBench metric thresholds (Figure 5). Interestingly, although we see an increase in $motifRMSD$, we observe that, for certain test cases, there is a significant decrease in backbone (global) $RMSD$ (Figure 6). For test case 6E6R in particular, the distribution (IQR) of $RMSD$ becomes much narrower as a result of the integration of online preferences. In contrast, we see that two-stage training has a negligible effect on test case 3TQB global $RMSD$ (Table 2).

6 Discussion

6.1 Limitations

While results are promising, our project has a few limitations. Because we focus on three of the test cases from MotifBench, our evaluation does not represent all the heterogeneity in the different possible motifs. Furthermore, due to limited compute resources, we only generate a limited number of samples (<2k) to create the training set for our three select motifs and only evaluate the smallest ESM3 model (ESM3-open). Scaling to larger datasets and models may qualitatively change the results. Despite these limitations, we believe that our project demonstrates a proof of concept that preference optimization may be an additional method that can be used to refine existing pretrained pLMs to optimize conditional generation.

Figure 5: Success rate for each motif test case for base ESM3 and all finetuned models.

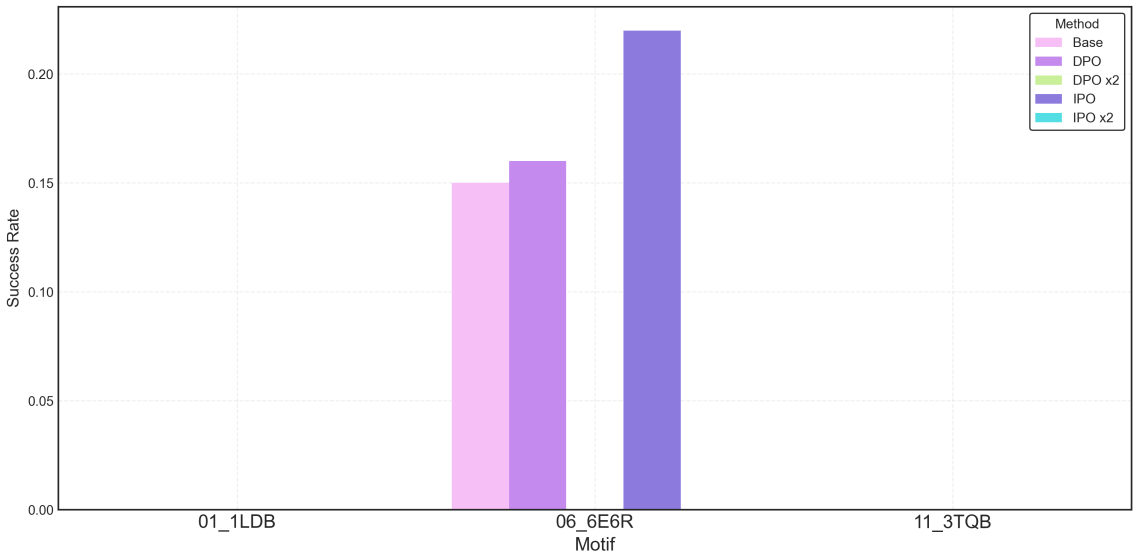


Figure 6: Distributions of backbone (global) RMSD conditioned on each motif test case for base EMS3 and all finetuned models on set of self-consistency (ProteinMPNN + ESMFold) samples.

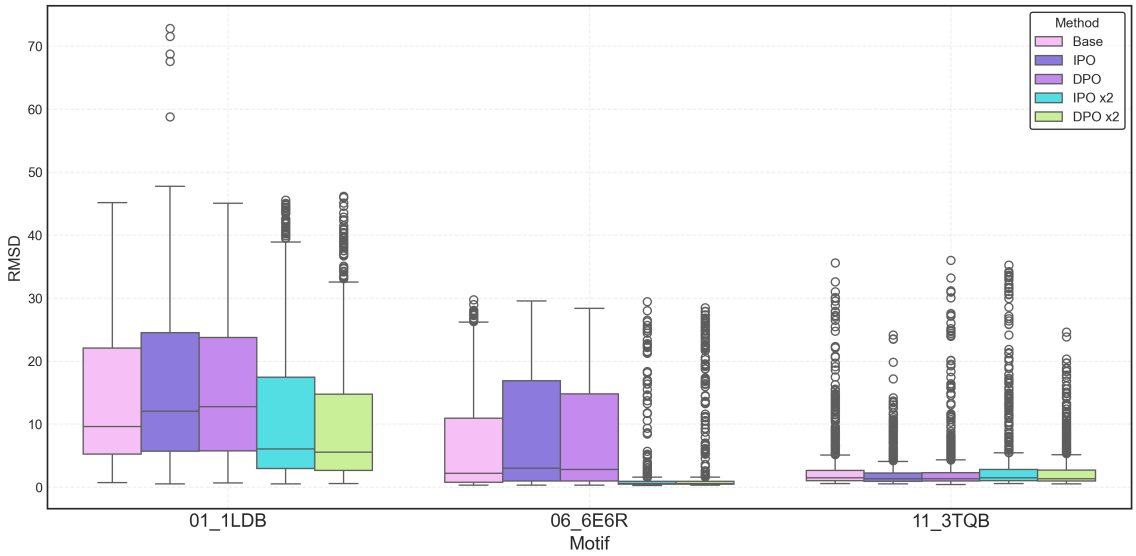


Table 2: Mean ± SD of global RMSD and motif RMSD (in Å) by method and MotifBench test case.

Motif	Method	RMSD (Å)	Motif RMSD (Å)
1LDB	Base	13.95 ± 11.00	5.01 ± 0.86
	DPO	15.20 ± 10.88	5.02 ± 0.93
	IPO	15.56 ± 12.04	5.00 ± 0.99
	DPO x2	10.31 ± 10.37	5.13 ± 0.70
	IPO x2	11.27 ± 11.55	5.13 ± 0.75
6E6R	Base	6.85 ± 8.45	2.64 ± 1.22
	DPO	7.89 ± 8.81	2.47 ± 1.22
	IPO	8.32 ± 9.15	2.37 ± 1.25
	DPO x2	2.75 ± 6.18	3.73 ± 0.35
	IPO x2	1.89 ± 4.64	3.69 ± 0.07
3TQB	Base	3.21 ± 4.91	10.93 ± 3.25
	DPO	2.96 ± 4.79	11.41 ± 4.96
	IPO	2.36 ± 2.88	10.49 ± 3.13
	DPO x2	2.73 ± 3.49	11.52 ± 6.37
	IPO x2	3.70 ± 6.01	11.45 ± 5.60

6.2 Challenges

Throughout this project, we faced many challenges. Particularly, we took care with creating the preference dataset and determining metrics that accurately reflect the preferences of our optimized model. ESM3-open preforms relatively poorly on the MotifBench test cases so we relaxed what we defined as a "good" scaffold to increase the number of preference pairs we could attain for our dataset. Beyond developing motif-specific models, we attempted to optimize one model to perform well on many MotifBench motifs but were not succcessful. We believe this is likely due to the heterogeneity and complexity of the motifs, many of which involve multiple segments, which creates a noisy preference dataset. Additionally, since these motif test cases are canonically challenging, it is difficult to generate "good" or "successful" designs for many of the motifs, as evidenced by the public leaderboard hosted by the authors of MotifBench. Perhaps more data or different fine-tuning methods are needed to learn from a preference dataset of using all MotifBench test cases.

7 Conclusion

7.1 Summary

In this project, we apply preference optimization methods to pLMs to improve conditional generation for the motif-scaffolding problem with test cases from MotifBench. We demonstrate that preference optimization can potentially improve scaffold generation with ESM3-open when fine-tuned on offline preferences, particularly when correct demonstrations exist in the dataset. Finally, we explore the potential to use online samples to finetune ESM3, and although the resulting success rates for the two-stage samples were lower than those of the offline-only models, further investigation on more robust integration of online preferences would be valuable. We see that for some test cases (6E6R in particular), incorporating online preferences may lead to lower global backbone $RMSD$, suggesting that the designed structure is more "foldable" (protein-like) under the self-consistency evaluation pipeline.

7.2 Future Work

To address our limitations, we would like to scale to larger training datasets and models. In particular, we are interested in developing a method that can fine-tune one pLM to improve scaffold generation in a motif-agnostic manner. The scaffolds generated using this hypothetical model will conserve the motif function regardless of the chemistry, leading to better overall performance and large applicability. This would likely involve incorporating data capturing biochemical and biophysical properties of the motif structure.

8 Team Contributions

- **Wei Deng**: Generated evaluation samples, ran evaluation pipeline on models using Motif-Bench and conducted analysis on results.
- **Jordan Cahoon**: Generated, scored, and matched scaffolds for the offline and online preference datasets, implemented preference optimization, and trained models.

9 Acknowledgements

We thank Brian L. Trippe and Tianyu Lu for valuable discussions that improved our understanding of the problem and the methods.

References

[1] 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.

[2] Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem, 2023. arXiv:2206.04119 [q-bio].

[3] Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of Many, One: Designing and Scaffolding Proteins at the Scale of the Structural Universe with Genie 2, 2024.

[4] Ke Liu, Weian Mao, Shuaike Shen, Xiaoran Jiao, Zheng Sun, Hao Chen, and Chunhua Shen. Floating Anchor Diffusion Model for Multi-motif Scaffolding, 2024. arXiv:2406.03141 [q-bio].

[5] Zhuoqi Zheng, Bo Zhang, Kieran Didi, Kevin K. Yang, Jason Yim, Joseph L. Watson, Hai-Feng Chen, and Brian L. Trippe. Motifbench: A standardized protein design benchmark for motif-scaffolding problems. February 2025.

[6] 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 A. 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. *Science*, 387(6736):850–858, February 2025. Publisher: American Association for the Advancement of Science.

[7] Matt Sternke and Joel Karpiak. ProteinRL: Reinforcement learning with generative protein language models for property-directed sequence design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[9] Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding generative protein language models with reinforcement learning, 2025.

[10] Jithendaraa Subramanian, Shivakanth Sujit, Niloy Irtisam, Umong Sain, Riashat Islam, Derek Nowrouzezahrai, and Samira Ebrahimi Kahou. Reinforcement learning for sequence design leveraging protein language models, 2024.

[11] Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M. Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, Magnus Bauer, Regina Barzilay, Tommi S. Jaakkola, Rohith Krishna, and David Baker. Atom level enzyme active site scaffolding using rfdiffusion2. *bioRxiv*, 2025.

[12] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

[13] Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem, 2023.