Functional Alignment of Protein Language Models via Reinforcement Learning with Experimental Feedback

Nathaniel Blalock*
Chemical and Biological Engineering
University of Wisconsin-Madison

Srinath Seshadri*
Biomedical Engineering
Duke University

Agrim Babber
Biochemistry
University of Wisconsin-Madison

Sarah Fahlberg
Biochemistry
University of Wisconsin-Madison

Philip Romero*
Biomedical Engineering
Duke University

Abstract

Protein language models (pLMs) have emerged as state-of-the-art tools for generative protein sequence design. pLMs however do not inherently design new sequences with function beyond what occurs in nature, demonstrating a misalignment with the protein engineering objective of redesigning a protein sequence with enhanced function. In the field of natural language processing, Reinforcement Learning with Human Feedback (RLHF) aligned the large language model ChatGPT towards preferred responses via supervised fine-tuning (SFT) and proximal policy optimization (PPO). We adapt SFT and PPO for the functional alignment of pLMs using experimental data and call this method Reinforcement Learning with Experimental Feedback (RLXF). We use RLXF to align ESM-2 and a generative variational autoencoder to design 5 mutant variants of the oxygen-independent fluorescent protein CreiLOV. We find a greater fraction of designs from aligned ESM-2 were active and at least half as bright as CreiLOV with *in vivo* fluorescence assays. We present RLXF as a versatile method to functionally align pLMs using experimental data for protein sequence redesign.

1 Introduction

pLMs have emerged as a state-of-art tool for generative protein sequence design. These models utilize transformer-based architectures [1, 2, 3, 4, 5] to learn complex evolutionary and structural dependencies from natural protein sequences [1, 4] and structures [5, 6, 7, 8]. This allows pLMs to construct rich representations of functional and structural protein design constraints [9, 10, 11] learned via masked language modeling (MLM) [1, 5, 7, 12, 13, 14, 15] or casual language modeling (CLM) [16, 17, 18, 19, 20] pre-training objectives. Recently, pLMs designed novel and functional GFP variants [5], lysozyme variants [16], and peptide binders [21, 22], amongst others. While the generation of functional GFP and lysozyme variants far in sequence identity from natural training examples is a remarkable stride towards *de novo* protein design, the function of these variants remained within the distribution of natural sequences. These results suggest pLMs are misaligned with an important protein engineering objective: redesigning protein sequences with enhanced

*Correspondence: nblalock@wisc.edu, sri.seshadri@duke.edu, philip.romero@duke.edu

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

function [23]. Thus, there is a need for a computational workflow that leverages experimental data to align pLMs for generative protein redesign.

Reinforcement learning with human feedback (RLHF) [24, 25] aligned large language models like ChatGPT towards human-preferred responses and away from offensive or harmful outputs, without catastrophic forgetting of pre-training knowledge [26, 27, 28]. Remarkably, RLHF outperformed supervised fine-tuning (SFT) alone [26, 25], allowed targeted and iterative enhancements of model responses [27, 28], and reduced compute demands by approximately 1000-fold compared to traditional reinforcement learning [24]. Our method Reinforcement Learning with Experimental Feedback (RLXF) is adapted from the RLHF workflow [26] and is guided by an ensemble of reward models trained with deep mutational scanning (DMS) data (Figure 1).

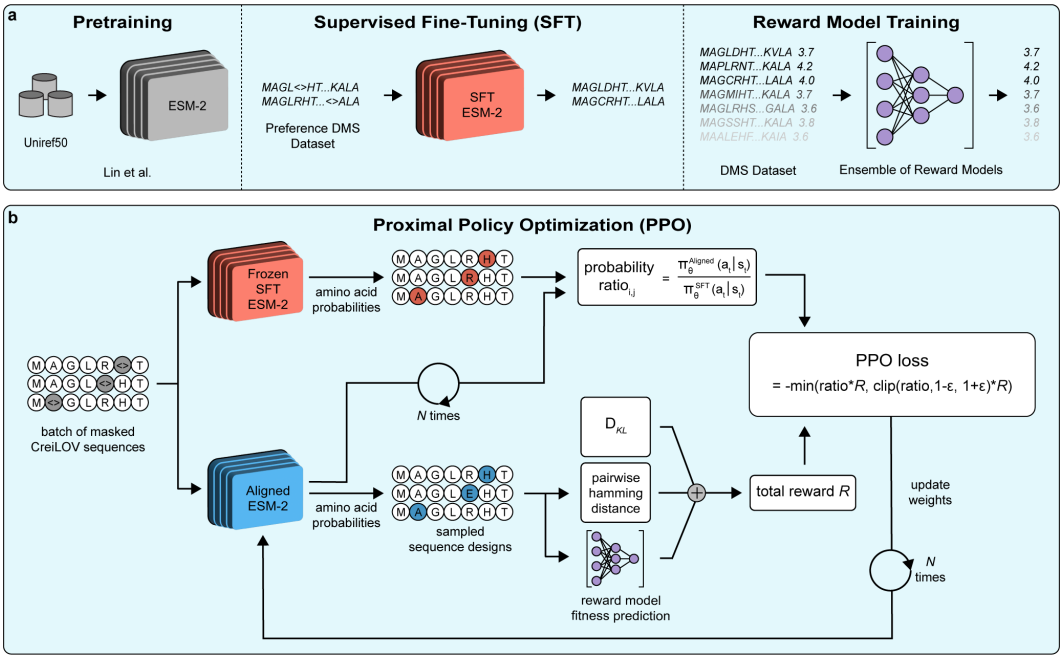

Figure 1: Reinforcement Learning with Experimental Feedback (RLXF). (a) RLXF consists of pre-training ESM-2 with natural protein sequence data [4], briefly supervised fine-tuning ESM-2 towards experimentally validated designs functionally similar to CreiLOV, and aligning ESM-2 using PPO guided by an ensemble of reward models trained with CreiLOV DMS data. (b) PPO iteratively updates ESM-2 weights N times each epoch to become more likely to generate brighter and diverse sequence designs. Backpropagation occurs through the probabilities from the aligned model for amino acids sampled by the aligned ESM-2 model found in the numerator of the probability ratio term. Probabilities from the pre-trained model for amino acids sampled by the aligned model are in the denominator. The total reward term consists of a normalized predicted log fluorescence score from our ensemble of reward models, a pairwise Hamming distance term to encourage sequence diversity, and a Kullback-Leibler divergence penalty (D_{KL}) to prevent drastic updates to ESM-2 weights that result in forgetting knowledge gained during pre-training. PPO loss is clipped with the epsilon (ϵ) term to stabilize training.

We align the pLM ESM-2 [4] and a generative protein sequence variational autoencoder (VAE), trained with a method previously shown to be capable of designing novel, diverse, and functional luciferases [29], to redesign the flavin-binding fluorescent protein CreiLOV [30, 31]. CreiLOV is an attractive target as a reporter protein for studying hypoxic/anaerobic environments such as gut microbiomes, tumor environments, and high-density fermentations [32] that the ubiquitous oxygen-dependent green fluorescent protein (GFP) cannot. Designing a more fluoresent CreiLOV variant would create a more practical reporter protein as CreiLOV is significantly less fluorescent than GFP.

2 Methods

We describe methods in the supplementary methods.

3 Results

We trained an ensemble of multi-layer perceptron reward models with 6,925 sequences containing 1-4 mutations from a CreiLOV DMS dataset [33] to predict the log mean fluorescence of CreiLOV variants (Table S1). The ensemble of reward models achieved a Spearman correlation of 0.93 for a test set containing 5 mutant variants of CreiLOV, indicating an ability to guide the alignment of ESM-2 and the VAE during PPO (Figure S1).

The VAE was pre-trained on a multiple sequence alignment of curated natural sequences related to CreiLOV (Figure S2-S3). The VAE could reconstruct CreiLOV and generate sequence designs similar in sequence identity to CreiLOV, suggesting the VAE did not require SFT to initialize parameters for alignment.

Pre-trained ESM-2 can generate sequence designs similar in sequence identity to CreiLOV (Figure S4).

We supervised fine-tuned ESM-2 to maximize the likelihood of the best mutations in the DMS dataset using pairs of prompts (*CreiLOV variants with beneficial amino acid mutations masked*) and responses (*CreiLOV variants with beneficial amino acid mutations unmasked*) (Figure S5).

Alignment with RLXF improved the likelihood of ESM-2 and the VAE to design CreiLOV variants with greater predicted log fluorescence than pre-trained models (Figure 2a-b, Figure S6-S7). Not only did RLXF shift the distribution of model sequence designs toward brighter designs, aligned models learned to avoid mutating flavin-binding residues and extrapolated beyond the mutational regime of wet lab experimental training data, suggesting effective exploration and exploitation of CreiLOV sequence design space (Figure 2c, Figure S8).

When sampling 1000 designs from each model, aligned ESM-2 generated design with the greatest maximum predicted log fluorescence and the aligned VAE generated designs with the greatest mean, median, and minimum predicted log fluorescence. The win rate of all models improved after alignment (Table 1).

Metric	Pre-trained ESM-2	SFT ESM-2	Aligned ESM-2	Pre-trained VAE	Aligned VAE
Mean	3.9735	4.0041	4.0036	4.0381	**4.0878**
Median	3.9898	4.0151	4.0143	4.0380	**4.0949**
Max	4.1006	4.1578	**4.1592**	4.1123	4.1248
Min	3.6266	3.7027	3.7027	3.9066	**3.9786**
Win Rate vs. Pre-trained ESM-2	-	0.623	0.620	0.740	**0.955**
Win Rate vs. SFT ESM-2	0.377	-	0.854	0.646	**0.891**
Win Rate vs. Pre-trained VAE	0.260	0.354	0.352	-	**0.902**

Table 1: RLXF aligns models to design 5 mutant variants with enhanced predicted fluorescence.

We validated designs from ESM-2 with *E. coli*-based fluorescence assays (Figure 3). We found that a greater fraction of designs from aligned models compared to pre-trained models were active and at least half as bright as CreiLOV. However, we were unable to design a sequence brighter than CreiLOV with the current implementation of SFT and PPO.

Interestingly, we found that increases in the top 10% recall metric correlated more strongly with improved generative performance of models than the widely used Spearman correlation metric (Table S2). This suggests the top 10% recall metric better indicates the functional alignment of pLMs for generative protein redesign.

4 Discussion

In this work, we propose RLXF to deploy a pLM aligned with SFT and PPO for the optimization of protein sequences with *in vivo* fluorescence beyond what occurs in nature. Our method provides

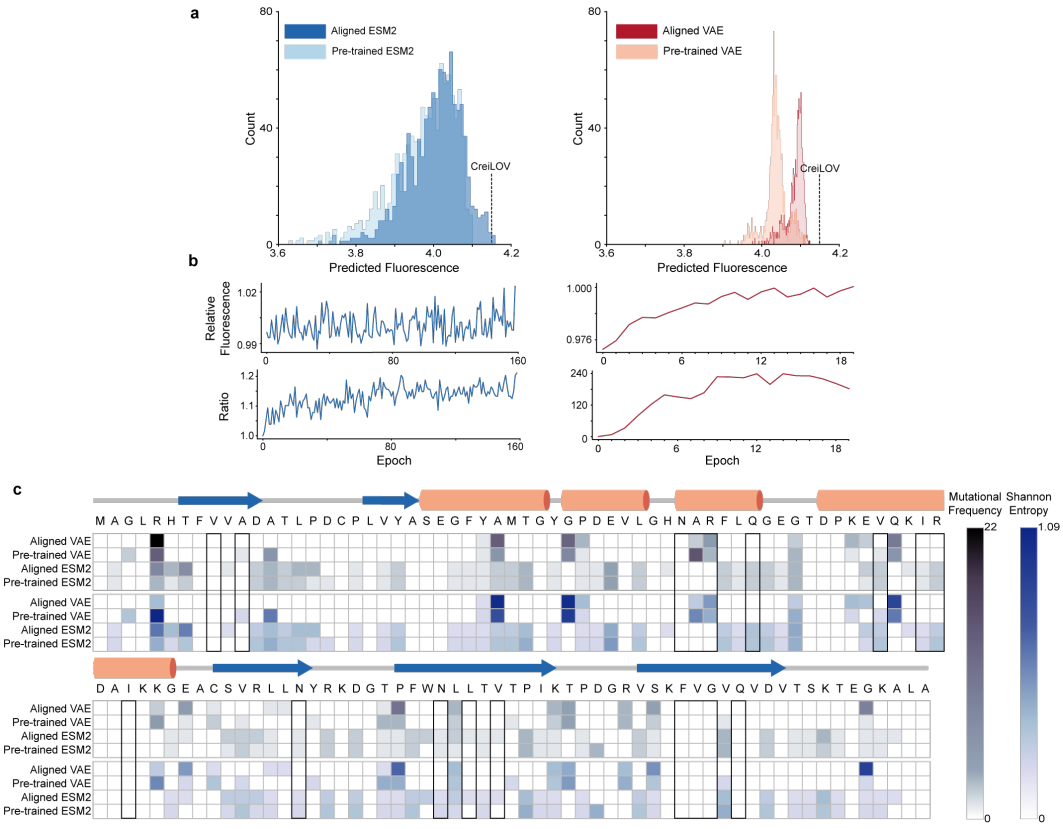

Figure 2: *In silico* evaluation of aligned models. (a) Histograms depicting the distribution of predicted log fluorescence for 1,000 designs generated by pre-trained and aligned models. The predicted fluorescence of CreiLOV is indicated for reference. (b) Predicted log fluorescence relative to CreiLOV and probability ratios between pre-trained and aligned model samples increased during training. (c) Mutational frequency and Shannon entropy indicate how often models mutate each position of CreiLOV and the diversity of these mutations. Residues outlined in black are flavin-binding residues.

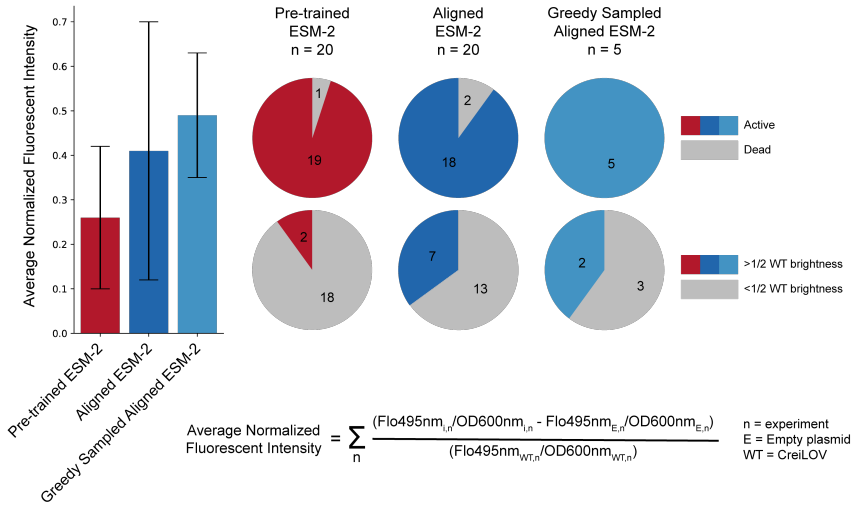

Figure 3: *In vivo* fluorescence assay data from pre-trained and aligned ESM-2.

an alternative to training new pLMs with intensive compute resources and large databases. Given that common pre-training objectives do not always scale with model size [5, 34], our method also provides an alternative to creating new pre-training tasks [35, 34].

We chose PPO rather than other policy optimization methods, such as direct preference optimization used in related studies [5, 1, 36, 37], because PPO excels at complex generative tasks [38, 39]. The *in vivo* fluorescence of our designs and the increasing number of studies using reinforcement learning techniques may signify a shift in protein engineering paradigms towards supplementing the knowledge already present in pLM representations with experimental data for domain-specific tasks. While our method is sensitive to hyperparameter selection during SFT and PPO, Optuna [32] provided an automated and efficient manner to optimize hyperparameters for stable alignment. We may be able to reduce the experimental data required to align pLMs with RLXF by deploying a VAE as a reward model, as our VAE appeared to learn local evolutionary patterns specific to CreiLOV functional requirements during pre-training.

The aligned VAE quickly converged on a few mutations during alignment (R5E, P35K, P82K, and G115A/S) while ESM-2 did not. Granted, the brightest designs from ESM-2 often had non-polar amino acids substituted with positively charged amino acids. We hypothesize this observation comes from differences in ESM-2 and VAE pre-training and model size. The VAE is pre-trained with an MSA containing local evolutionary information while ESM-2 is trained on protein sequences sampled across nature [4]. This biases ESM-2's per-residue likelihoods towards general trends across evolution rather than family or domain-specific mutations.

Alignment responses may scale with model size [5]. However, it is not clear if this is the result of larger pLMs learning additional general protein design constraints during pre-training or the increase in model size preventing rapid over-fitting to local maxima during alignment. Interestingly, we found that aligning ESM-2 8M was significantly more challenging than aligning ESM-2 35M. In addition, the predicted log fluorescence of sequence designs from pre-trained ESM-2 models increases with model size up to ESM-2 650M (Figure S9). We hypothesize that larger pLMs than the ESM-2 35M we aligned in this study can more effectively explore complex sequence-function mappings, often called a protein fitness landscape, to design protein sequences with the greatest functional enhancements. In addition, RLXF can be adapted for iterative, multi-objective functional alignment of pLMs, perhaps to explore vast design landscapes using a self-driving lab [40] to efficiently obtain experimental data.

We hypothesize we were unable to design a variant brighter than CreiLOV as a result of pre-training knowledge lost during SFT. The single mutant logit space from the supervised fine-tuned ESM-2 revealed supervised fine-tuned ESM-2 could no longer reconstruct CreiLOV (Figure S10). Interestingly, ESM-2 still learned several beneficial mutations during SFT and PPO to increase the fraction of designs from ESM-2 that were active and at least half as bright as CreiLOV.

We are investigating these hypotheses to design variant sequences brighter than CreiLOV and further establish guidelines for the functional alignment of pLMs.

5 Acknowledgements

The authors thank Tom Sercu for his insight into evaluating ESM-2 models and Adithyan Unni, Sam Gelman, and Chase Freschlin for their ideas and discussions that aided this work. We also thank Sameer D'Costa for server management and the High Throughput Computing Center at UW-Madison for compute resources. Finally, we acknowledge the NSF (Award Number: 2226383) and NIH NIGMS (Grant Number: 5R01GM150929-02) for funding this work.

References

[1] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, Chiming Liu, Aohan Zeng, Yuxiao Dong, Jie Tang, and Le Song. xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein, January 2024. arXiv:2401.06199 [cs, q-bio].

[2] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik,

and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[3] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling, January 2023. arXiv:2301.06568 [cs, q-bio].

[4] 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. Publisher: American Association for the Advancement of Science.

[5] 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. Pages: 2024.07.01.600583 Section: New Results.

[6] Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Accurate and robust protein sequence design with CarbonDesign. *Nature Machine Intelligence*, 6(5):536–547, May 2024. Publisher: Nature Publishing Group.

[7] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary, April 2024. Pages: 2023.10.01.560349 Section: New Results.

[8] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual Language Model for Protein Sequence and Structure, March 2024. Pages: 2023.07.23.550085 Section: New Results.

[9] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021.

[10] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, December 2019. Publisher: Nature Publishing Group.

[11] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723, December 2019.

[12] Zhangzhi Peng, Benjamin Schussheim, and Pranam Chatterjee. PTM-Mamba: A PTM-Aware Protein Language Model with Bidirectional Gated Mamba Blocks, February 2024. Pages: 2024.02.28.581983 Section: New Results.

[13] Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294.e2, March 2024. Publisher: Elsevier.

[14] Carlos Outeiral and Charlotte M. Deane. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence*, 6(2):170–179, February 2024. Publisher: Nature Publishing Group.

[15] Sophia Vincoff, Shrey Goel, Kseniia Kholina, Rishab Pulugurta, Pranay Vure, and Pranam Chatterjee. FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Focused Probabilistic Masking, June 2024.

[16] 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. Publisher: Nature Publishing Group.

[17] Timothy F. Truong Jr and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences, November 2023. arXiv:2306.06156 [cs, q-bio].

[18] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. RITA: a Study on Scaling Up Generative Protein Sequence Models, July 2022. arXiv:2205.05789 [cs, q-bio].

[19] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022. Publisher: Nature Publishing Group.

[20] Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. ProtMamba: a homology-aware but alignment-free protein state space model, May 2024. Pages: 2024.05.24.595730 Section: New Results.

[21] Tianlai Chen, Sarah Pertsemlidis, Rio Watson, Venkata Srikar Kavirayuni, Ashley Hsu, Pranay Vure, Rishab Pulugurta, Sophia Vincoff, Lauren Hong, Tian Wang, Vivian Yudistyra, Elena Haarer, Lin Zhao, and Pranam Chatterjee. PepMLM: Target Sequence-Conditioned Generation of Peptide Binders via Masked Language Modeling, November 2023. arXiv:2310.03842 [q-bio].

[22] Suhaas Bhat, Kalyan Palepu, Vivian Yudistyra, Lauren Hong, Venkata Srikar Kavirayuni, Tianlai Chen, Lin Zhao, Tian Wang, Sophia Vincoff, and Pranam Chatterjee. De Novo Generation and Prioritization of Target-Binding Peptide Motifs from Sequence Alone, June 2023. Pages: 2023.06.26.546591 Section: New Results.

[23] Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. *Nature Biotechnology*, 42(2):216–228, February 2024. Publisher: Nature Publishing Group.

[24] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[25] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020.

[26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. arXiv:2203.02155 [cs].

[27] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. arXiv:2204.05862 [cs].

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian

Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. arXiv:2307.09288 [cs].

[29] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLOS Computational Biology*, 17(2):e1008736, February 2021. Publisher: Public Library of Science.

[30] Arnab Mukherjee, Kevin B. Weyant, Utsav Agrawal, Joshua Walker, Isaac K. O. Cann, and Charles M. Schroeder. Engineering and Characterization of New LOV-Based Fluorescent Proteins from Chlamydomonas reinhardtii and Vaucheria frigida. *ACS Synthetic Biology*, 4(4):371–377, April 2015. Publisher: American Chemical Society.

[31] Arnab Mukherjee, Joshua Walker, Kevin B. Weyant, and Charles M. Schroeder. Characterization of Flavin-Based Fluorescent Proteins: An Emerging Class of Fluorescent Reporters. *PLOS ONE*, 8(5):e64753, May 2013. Publisher: Public Library of Science.

[32] Harun F. Ozbakir, Nolan T. Anderson, Kang-Ching Fan, and Arnab Mukherjee. Beyond the Green Fluorescent Protein: Biomolecular Reporters for Anaerobic and Deep-Tissue Imaging. *Bioconjugate Chemistry*, 31(2):293–302, February 2020. Publisher: American Chemical Society.

[33] Yongcan Chen, Ruyun Hu, Keyi Li, Yating Zhang, Lihao Fu, Jianzhi Zhang, and Tong Si. Deep Mutational Scanning of an Oxygen-Independent Fluorescent Protein CreiLOV for Comprehensive Profiling of Mutational and Epistatic Effects. *ACS Synthetic Biology*, 12(5):1461–1473, May 2023. Publisher: American Chemical Society.

[34] Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature Reuse and Scaling: Understanding Transfer Learning with Protein Language Models, February 2024. Pages: 2024.02.05.578959 Section: New Results.

[35] Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022. arXiv:2206.13517 [cs, q-bio].

[36] Ziyi Zhou, Liang Zhang, Yuanxi Yu, Banghao Wu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications*, 15(1):5566, July 2024. Publisher: Nature Publishing Group.

[37] Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via Direct Preference Optimization, May 2024. Pages: 2024.05.20.595026 Section: New Results.

[38] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study, April 2024. arXiv:2404.10719 [cs].

[39] Christof Angermueller, David Belanger, Kevin Murphy, David Dohan, Ramya Deshpande, and Lucy Colwell. MODEL-BASED REINFORCEMENT LEARNING FOR BIO- LOGICAL SEQUENCE DESIGN, 2020.

[40] Jacob T. Rapp, Bennett J. Bremer, and Philip A. Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1):97–107, January 2024. Number: 1 Publisher: Nature Publishing Group.

[41] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023.

[42] L. Steven Johnson, Sean R. Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):431, August 2010.

[43] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017. Number: 2 Publisher: Nature Publishing Group.

[44] Thomas A Hopf, Charlotta P I Schärfe, João P G L M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M J J Bonvin, and Debora S Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3:e03430, September 2014. Publisher: eLife Sciences Publications, Ltd.

[45] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, October 2018. Number: 10 Publisher: Nature Publishing Group.

[46] Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature Communications*, 13(1):1914, April 2022. Publisher: Nature Publishing Group.

[47] Xinqiang Ding, Zhengting Zou, and Charles L. Brooks III. Deciphering protein evolution and fitness landscapes with latent space models. *Nature Communications*, 10(1):5644, December 2019. Number: 1 Publisher: Nature Publishing Group.

[48] Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *NeurIPS Track on Datasets and Benchmarks*, 2023.

[49] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, 2020. arXiv:1910.14659 [cs, eess, stat].

[50] Grégory Boël, Reka Letso, Helen Neely, W. Nicholson Price, Kam-Ho Wong, Min Su, Jon Luff, Mayank Valecha, John K. Everett, Thomas B. Acton, Rong Xiao, Gaetano T. Montelione, Daniel P. Aalberts, and John F. Hunt. Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, 529(7586):358–363, January 2016.

6 Supplementary Methods

6.1 Data Curation for Reward Models and VAE

We used data to train reward models from an extensive CreiLOV DMS dataset [33]. The training split contains 2,204 single mutants (92.6% coverage), 176 double mutants, 978 triple mutants, and 3,565 four mutation variants. We curated validation and test sets using 75/25 data splits for 9,603 five mutation variants.

We curated 260,349 natural sequences related to CreiLOV from the protein database UniRef90 [41] with the Hidden-Markov model homology search tool Jackhmmer [42] to obtain a multiple-sequence alignment (MSA) [43, 44, 45] containing proteins related to CreiLOV with a maximum of 2 iterations (N=2). We removed sequences less than 75% of the length of CreiLOV [45], removed sequences with an amino acid repeating 10 times in a row, and removed positions of the MSA not corresponding to CreiLOV [29]. We reweighted the remaining 243,682 sequences with neighbors classified as having a Hamming distance/length of sequence greater than 0.8 to reduce phylogenetic bias from uneven sampling [45, 46, 47]. We withheld 100 sequences from the reweighted MSA to later assess VAE overfitting to the training set as a pseudo-test set given that these sequences are unlabeled. We randomly sampled the remaining 243,582 sequences with a 90/10 split for the training and validation sets.

6.2 Training Reward Models

We trained 100 multi-layer perceptrons with fixed training, validation, and test data splits to predict the log mean fluorescence of CreiLOV variants in a supervised manner using a mean squared error (MSE) loss objective. Each MLP in the ensemble was initialized with a different seed. Input sequences were one-hot encoded. Hyperparameters for the ensemble of multi-layer perceptron reward models are included in Section 6.8 of the supplementary methods.

6.3 VAE Pre-training

The VAE is pre-trained by maximizing a modified version of the Evidence Lower Bound (ELBO) that effectively minimizes the Kullback–Leibler divergence between the variational approximation and the true posterior distribution [45]:

$$\mathcal{L}(\phi, \theta; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \tag{1}$$

The first term can be considered a reconstruction loss that is computed using cross-entropy between an input one hot-encoded sequence and output likelihoods. The second term is a Kullback-Leibler divergence term (D_{KL}) with the prior distribution $p(\mathbf{z})$ of $\mathcal{N}(0, \mathbf{I})$). β is the weight for the KL divergence term.

The VAE essentially receives batches of one-hot encoded sequences from the reweighted MSA, compresses the one-hot encoded sequences into a regularized latent space during encoding, and reconstructs these natural sequences from the latent space during decoding. Hyperparameters were selected by a grid search of 673 combinations are included in Section 6.8 of the supplementary methods.

6.4 Supervised Fine-tuning ESM-2

We briefly finetune ESM-2 35M in a supervised manner with the top 512 scoring variants in the training set, using the standard masked language modeling (MLM) training objective. We chose ESM-2 35M to make our method practical to those without intensive compute resources.

For input into ESM-2, we masked the mutations relative to wild type CreiLOV for all 512 sequences. This strategy is also referred to as instruction tuning in the literature.

During training, we minimize the cross-entropy between this preference data, *i*, and ESM-2 predictions, *j*:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(p_{ij}) \tag{2}$$

We conduct 1,000 trials of hyperparameter optimization using the Optuna Tree-Parzen Estimator (TPE) sampler and a custom PyTorch callback to terminate trials if model parameters became non-finite (i.e., NaN or infinity) after each training batch. We utilized two NVIDIA GeForce RTX 4090 GPUs (48 GB of VRAM). Hyperparameters selected are included in Section 6.8 of the supplementary methods.

6.5 Policy Proximal Optimization

Our PPO implementation utilizes two copies of SFT ESM-2 or the pre-trained VAE weights. One copy remains frozen while the aligned model has trainable parameters. In agreement with reinforcement learning terminology, we refer to a model as a policy in this section.

Proximal policy optimization relies on the minimization of a clipped surrogate objective function often called a clipped PPO loss:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[-\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{3}$$

The probability ratio $r_t(\theta)$ is between the trainable policy π_θ and the frozen policy $\pi_{\theta_{old}}$.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{4}$$

$\hat{A}_t$ is the advantage estimate. The hyperparameter ϵ establishes a pessimistic bound on model updates, preventing excessively large model updates to help balance exploration and exploitation. We estimate A_t using the total reward $R(x, y)$ similar to [25] with sequence prompts x to models and sampled sequence designs y:

$$R(x, y) = r_\theta(x, y) + \gamma H(y) - \beta D_{\text{KL}} \tag{5}$$

The total reward $R(x, y)$ consists of three components: a normalized predicted log mean fluorescence term $r_\theta(x, y)$ from our ensemble of reward models, a weighted pairwise Hamming distance term $\gamma H(y)$ between sampled sequence outputs to encourage the generation of diverse sequences, and a weighted Kullback–Leibler divergence penalty D_{KL} between the trainable and frozen policies.

$$D_{\text{KL}} = \log \left[\frac{\pi_\phi^{\text{trainable}}(y|x)}{\pi^{\text{frozen}}(y|x)} \right] \tag{6}$$

For $r_\theta(x, y)$, we normalized the 5th percentile predicted log fluorescence from the ensemble of reward models relative to CreiLOV or pre-trained designs (See section 6.8). When aligning ESM-2, we masked each position of CreiLOV one-at-a-time and input the sequence with one mask into the fixed ESM-2 model to generate the frozen log probabilities for single mutations prior to training. This time-consuming calculation required a forward pass for each position of CreiLOV but does not need to be repeated during alignment. For each epoch with PPO, we calculated new log probabilities for masked and mutated positions with the trainable ESM-2 policy. We used this method to simplify and expedite the calculation of D_{KL} each epoch. When aligning the VAE, we considered the likelihood matrices ($21 \times L$ with 21 referring to the 20 amino acids and gap tokens and L being the length of CreiLOV) from the trainable and frozen policies for D_{KL}.

Each PPO epoch involves the following steps:

1. Recursively identify the number of masks or noise to generate sequences with an average batch Hamming distance of 5 from CreiLOV.
2. Generate log probabilities for amino acids with the frozen policy.
3. Generate log probabilities for amino acids with the trainable policy.
4. Sample log probabilities from the trainable policy to obtain sequence designs with an average batch Hamming distance of 5 from CreiLOV.
5. Calculate the total reward R.
6. Calculate probability ratio $r_t(\theta)$.
7. Calculate the clipped PPO loss objective function.
8. Update trainable policy weights.
9. Repeat steps 3, 6, 7, and 8 for $N - 1$ more steps to complete a trajectory.

For ESM-2, we iteratively masked the CreiLOV sequence to identify the number of masks to generate sequences with an average batch Hamming distance of 5 from CreiLOV. For VAE alignment, we recursively applied noise to the latent representation of CreiLOV to generate sequences with an average batch Hamming distance of 5 from CreiLOV.

6.6 Computing *in silico* Evaluation Metrics

We utilized several techniques from literature to calculate top 10% recall and Spearman correlations for models across alignment with RLXF. These metrics are commonly used to evaluate the zero-shot predictions of models [48].

We calculated masked and mutant marginal scores as described in [9]. The masked marginal score is defined as:

$$\text{Masked Marginal} = \sum_{i \in M} \left[\log p(x_i = x_i^{mt} | x_{-M}) - \log p(x_i = x_i^{wt} | x_{-M}) \right] \tag{7}$$

where M is the set of mutated positions, x_i^{mt} is the mutant amino acid at position i, x_i^{wt} is the wildtype amino acid at position i, and x_{-M} represents the sequence with masks at the mutated positions. The mutant marginal score is similarly calculated as:

$$\text{Mutant Marginal} = \sum_{i \in M} \left[\log p(x_i = x_i^{mt} | x^{mt}) - \log p(x_i = x_i^{wt} | x^{mt}) \right] \tag{8}$$

where x^{mt} is the full mutant sequence.

We calculated pseudo-perplexity scores as described in [4, 49]. The pseudo-perplexity score is defined as:

$$\text{Pseudo-perplexity}(x) = \exp \left\{ -\frac{1}{L} \sum_{i=1}^{L} \log p(x_i | x_{j \neq i}) \right\} \tag{9}$$

where L is the length of the sequence, and $p(x_i | x_{j \neq i})$ is the probability of amino acid x_i at position i given all other amino acids in the sequence.

We calculated log ratio as described in [45, 43]. The log ratio is defined as:

$$\log \frac{p(x_{\text{Mutant}} | \theta)}{p(x_{\text{wildtype}} | \theta)} \tag{10}$$

where x refers to mutant or wildtype sequence and θ refers to the pre-trained VAE parameters.

6.7 *In vivo* Characterization of RLXF Designs

All designs were codon optimized for expression in *E. coli* [50] and ordered from Twist Bioscience as clonal genes, in which all design inserts were integrated into the pET-28a(+) expression vector. 100 ng of each design, in addition to wild type and empty plasmid controls, was transformed into chemically competent BL21 *E. coli* aliquots. 100 μL of transformation mix was plated on LB plates supplemented with 50 μg/mL of kanamycin, and grown at 37°C overnight. Individual colonies were picked and inoculated in 5 mL LB broth supplemented with kanamycin, and grown overnight at 37°C, shaking at 225 rpm, overnight. In a 96-well optical bottom plate, overnight cultures were diluted 1:100 in 200 μL of LB+kanamycin. Using an Agilent BioTek Synergy H1 plate reader, the cultures were grown to an OD_{600} of at least 0.4, at 37°C with shaking at 225 rpm, before being induced with 0.5 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG). Upon induction, the plate was incubated at 37°C with orbital shaking, while concurrently measuring OD_{600} and fluorescence emission at 495 nm in 30 minute intervals for 8 hours. For fluorescence excitation and emission, the minimum bandwidth of 9 nm was used, the gain was set at 142, and the Z-position was set to 4.75 mm. To obtain the final normalized fluorescence values for each design, the final raw fluorescence readings were normalized by the corresponding OD_{600}. The normalized empty plasmid control value was subtracted from all other normalized fluorescence values to account for background signal, and all experimental designs were then normalized against wild type values to obtain the final fluorescence reading relative to wild type. Sequences for which cells did not grow were removed from subsequent analysis. Consequently, we could not obtain data for 2 pre-trained ESM-2 designs, 4 greedy aligned ESM-2 designs, 3 pre-trained VAE designs, and 1 unconditional aligned VAE design. Efforts are currently underway to obtain replicates of the experimental data reported in this manuscript, as well as further *in vitro* characterization of the best designs including quantum yield assays of the purified ten brightest designs across all models evaluated.

6.8 Model Architecture and Hyperparameter Details

Reward Model Hyperparameters	
Loss	MSE
Learning Rate	1×10^{-6}
Batch Size	128
Epochs	2000
Dropout	0.1
Patience	400
Number of Models	100
Hidden Layer Dim.	400
Activation	ReLU
Optimizer	Adam
Embedding Type	One-Hot

Table 2: Hyperparameters for each multi-layer perceptron reward model.

Pre-trained VAE Hyperparameters	
Learning Rate	1×10^{-4}
Batch Size	32
Epochs	1000
Patience	100
Weighted D_{KL}	Cyclical Annealing Schedule
Number of Cycles	1
Embedding Type	One-Hot
1st Convolutional Layer	21 input channels, 21*16 output channels
2nd Convolutional Layer	21*16 input channels, 21 output channels
Kernel size	17
Padding	1
Fully Connected Layer Dimensions	400
Latent Space Dimensions	64
Optimizer	Adam
Activation	ReLU, LeakyReLU

Table 3: Hyperparameters identified by grid search to train VAE. ReLU activations were applied to fully connected layers. LeakyReLU activations were applied to convolutional layers.

SFT ESM-2 Hyperparameters	
ESM-2 Version	35M
Learning Rate	0.0401
Learning Rate Multiplier	0.9830
Learning Rate Multiplier Factor	0.9664
Batch Size	36
Epochs	2
Initial Number of Unfrozen Layers	27
Number of Layers to Unfreeze Each Epoch	10
Maximum Number of Unfrozen Layers	82
Training Positional Embedding	True
Weights for Cross Entropy Loss	Norm. Log Fluorescence Values (0-1)
Weight Decay	0.0010
Gradient Clipping Threshold	1
Optimizer	Adam
Warm Restart	Cosine Annealing
Scheduler	Cosine Annealing
Quantization	fp16

Table 4: Hyperparameters for SFT ESM-2 that were identified using Optuna.

RLXF-Aligned ESM-2 Hyperparameters	
Learning Rate	5×10^{-4}
Learning Rate Multiplier	0.9487
Learning Rate Multiplier Factor	0.9835
Batch Size	12
Increment to Increase Batch Size	3
Maximum Batch Size	20
Epochs	159
Iterations	3
Initial Number of Unfrozen Layers	36
Number of Layers to Unfreeze Each Epoch	15
Maximum Number of Unfrozen Layers	82
Training Positional Embedding	False
Fitness Advantage	Rel. to Pre-trained Ouputs
Fitness Batch Norm.	Max
Batch PPO Loss Norm.	Mean
Init. D_{KL} Weight	1×10^{-8}
D_{KL} Weight	1×10^{-7}
Pairwise HD Averaging Factor	60.4553
Number of Reward Models	100
Epsilon	0.2512
Weight Decay	3.1×10^{-3}
D_{KL} Backpropagation	False
Gradient Clipping Threshold	1.5727
Gradient Clipping Threshold Factor per Epoch	1.2230
Initial Number of Masks	5
Exponential Moving Average	True
Optimizer	Adam
Warm Restart	Cosine Annealing
Learning Rate Scheduler	Cosine Annealing
Quantization	fp16

Table 5: Hyperparameters for aligning SFT ESM-2 with PPO that were identified with Optuna.

RLXF-Aligned VAE Hyperparameters	
Learning Rate	7×10^{-4}
Learning Rate Multiplier	0.8821
Learning Rate Multiplier Factor	0.9718
Batch Size	29
Increment to Increase Batch Size	9
Maximum Batch Size	64
Epochs	27
Iterations	4
Initial Number of Unfrozen Layers	6
Number of Layers to Unfreeze Each Epoch	0
Amino Acid Sampling	Max Likelihood
Fitness Advantage	Rel. to CreiLOV Fitness
Fitness Batch Norm.	Max
Batch PPO Loss Norm.	Mean
Initial D_{KL} Weight	1×10^{-8}
D_{KL} Weight	1×10^{-7}
Pairwise HD Weighting Factor	89.8218
Epsilon	0.1826
Weight Decay	2.9920×10^{-6}
D_{KL} Backpropagation	False
Gradient Clipping Threshold	2.7783
Gradient Clipping Threshold Factor per Epoch	2
Target Hamming Distance from CreiLOV	5
Exponential Moving Average	True
Optimizer	Adam
Warm Restart	Cosine Annealing
Learning Rate Scheduler	Cosine Annealing

Table 6: Hyperparameters to align the VAE with RLXF identified with the Optuna.

7 Supplementary Tables and Figures

Number of Mutations	Count	Data split
0	2	Training split
1	2,204	Training split
2	176	Training split
3	978	Training split
4	3,565	Training split
5	9,603	Val./Test splits
6-15	151,085	Not used

Table S1: Distribution of mutations across data splits for the CreiLOV DMS dataset used to train the ensemble reward model.

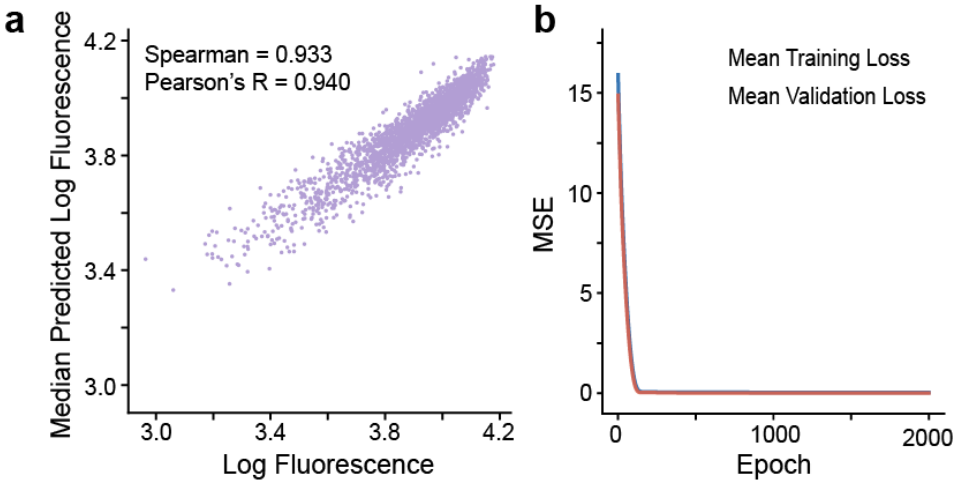

Figure S1: Ensemble of reward models accurately rank fluorescence of withheld sequences with 5 mutations relative to CreiLOV. (a) The ensemble of reward models predict the log fluorescence of test set amino acid sequences with 5 mutations relative to CreiLOV. (b) Mean multiple- squared error loss curves across reward model training.

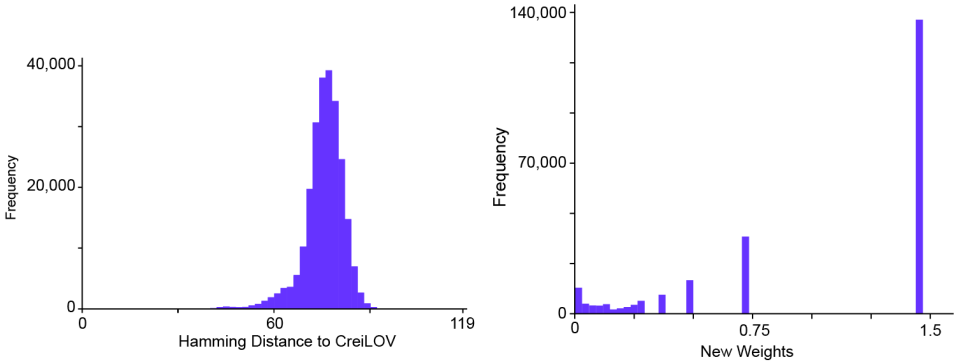

Figure S2: Distribution of sequences in the curated MSA used to pretrain the convolutional VAE before and after reweighting sequences to account for phylogenetic biases.

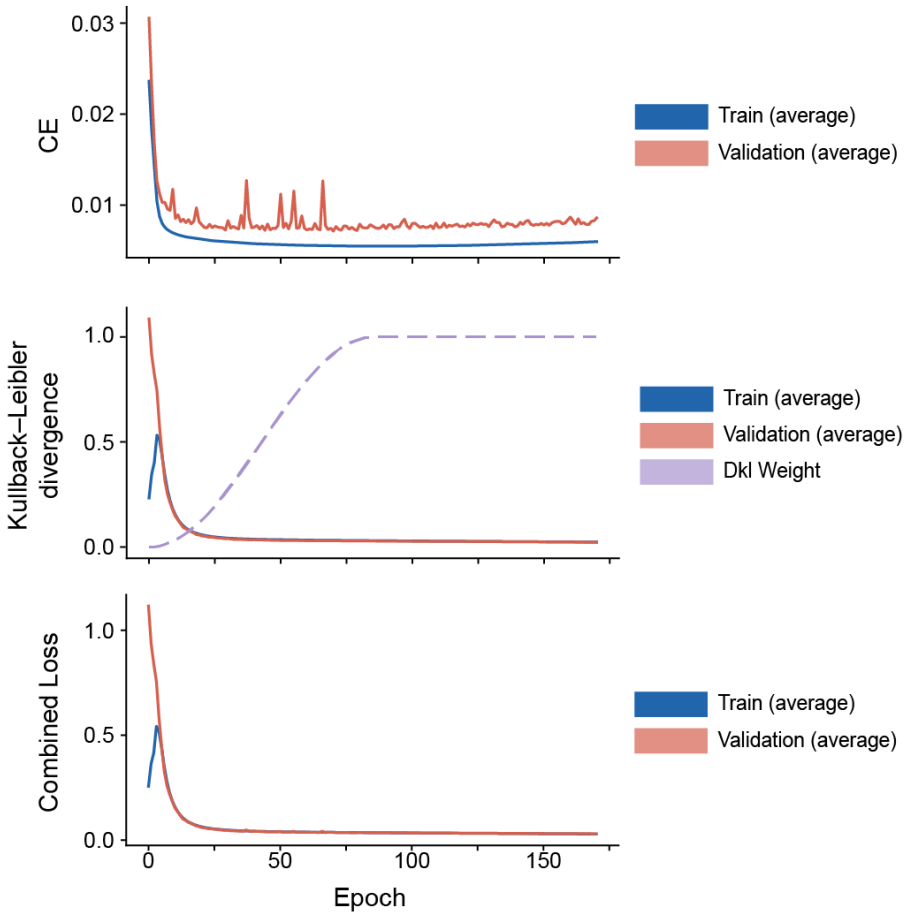

Figure S3: Training curves during pre-training of the VAE showing cross-entropy loss, D_{KL}, and the combined reconstruction loss.

Model	Method	Spearman Correlation	Top 10% Recall
Reward Model	Median Predicted Log Fluorescence	0.933	0.705
Pre-trained VAE	Log Ratio	0.344	0.162
Aligned VAE	Log Ratio	0.345	0.170
Pre-trained ESM-2	Pseudo-perplexity	0.689	0.332
SFT ESM2	Pseudo-perplexity	0.659	0.490
Aligned ESM2	Pseudo-perplexity	0.679	0.444
Pre-trained ESM-2	Mutant Marginal	0.746	0.365
SFT ESM2	Mutant Marginal	0.667	0.477
Aligned ESM2	Mutant Marginal	0.681	0.444
Pre-trained ESM-2	Masked Marginal	0.759	0.365
SFT ESM2	Masked Marginal	0.667	0.485
Aligned ESM2	Masked Marginal	0.682	0.444

Table S2: Model performance according to various *in silico* metrics

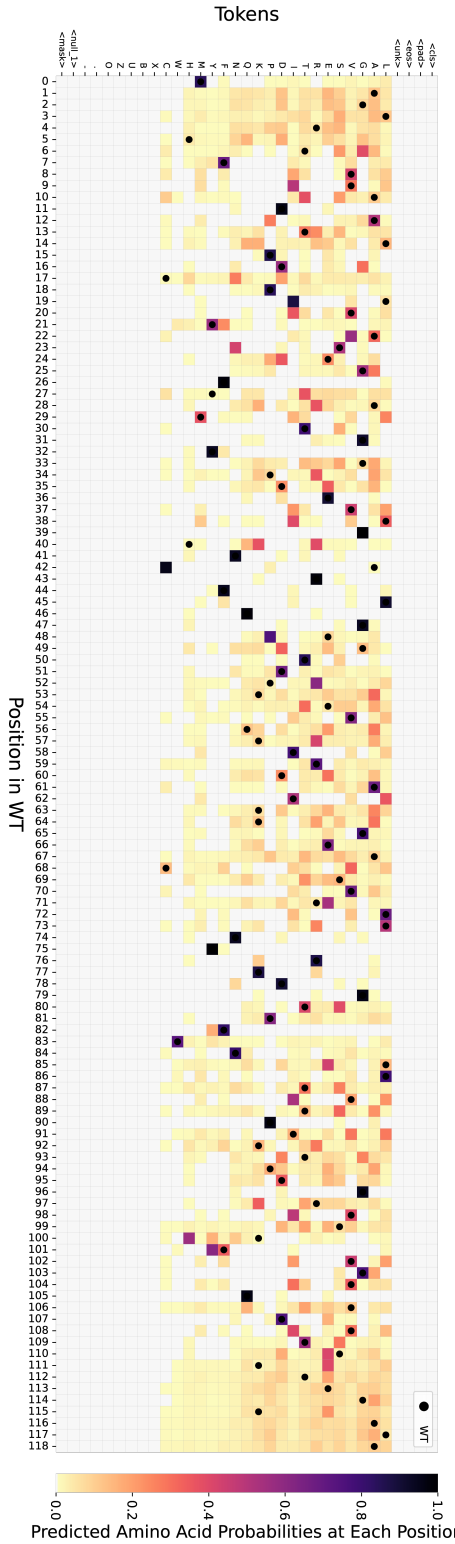

Figure S4: Single mutant logit heatmap for WT (CreiLOV) with ESM-2 (35M) by masking CreiLOV one position at a time with ESM-2 and saving the logits for the masked position.

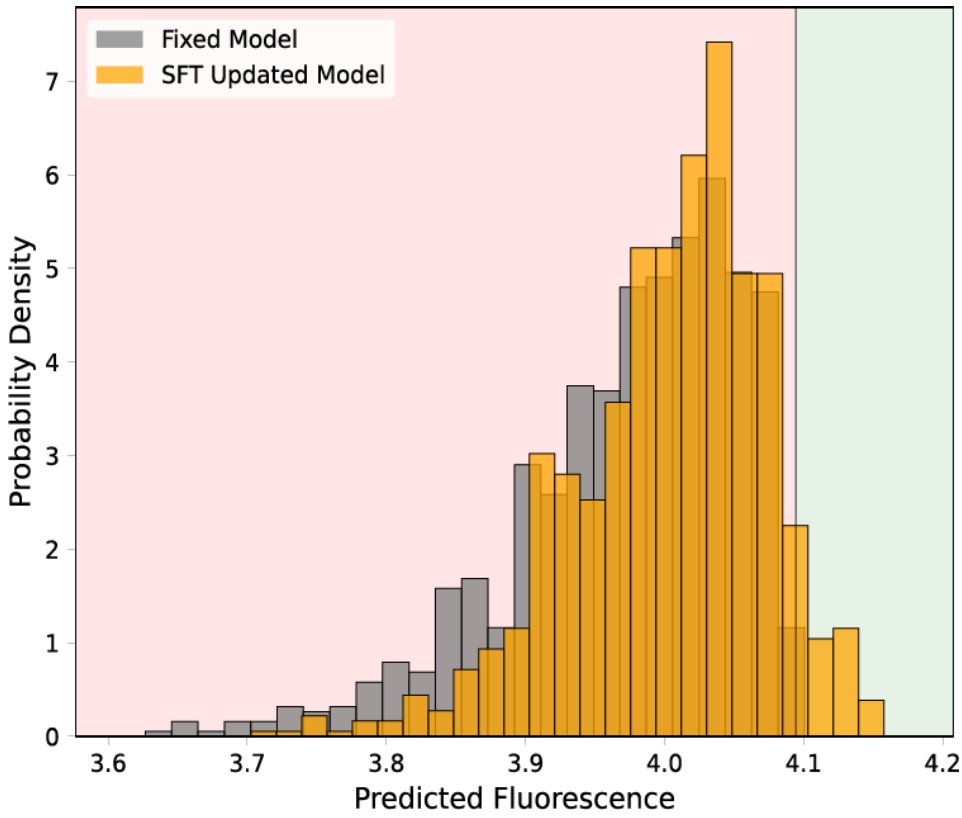

Figure S5: Reward model predicted log fluorescence of 1000 sequence designs from the SFT ESM-2 model. Designs in the green regime indicate designs with greater than wild type predicted fluorescence.

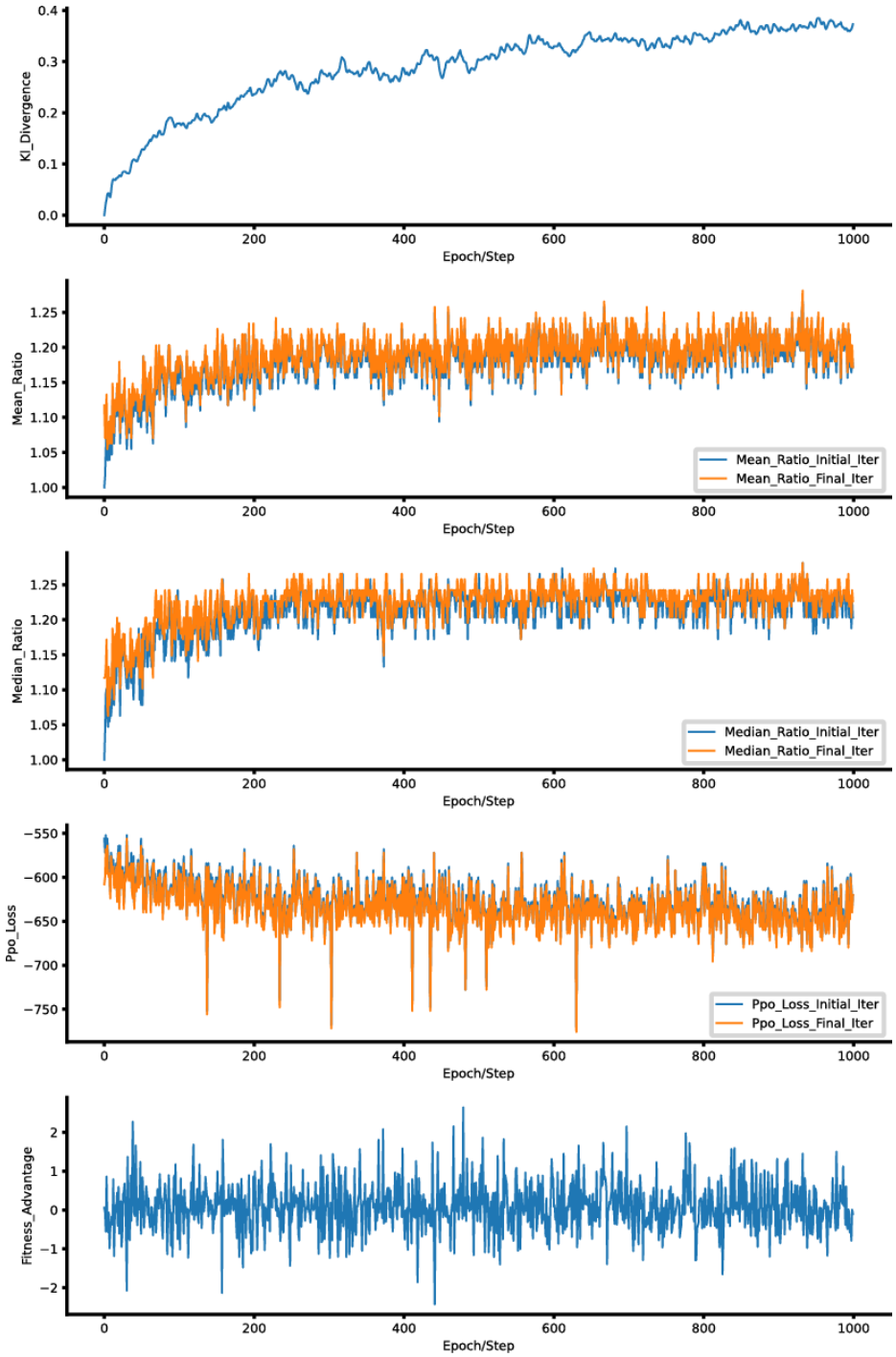

Figure S6: Training curves during alignment of ESM-2 showing the progression of D_{KL}, mean and median ratios of aligned model likelihoods relative to SFT, PPO clipped surrogate objective loss, and fitness advantage of generated designs.

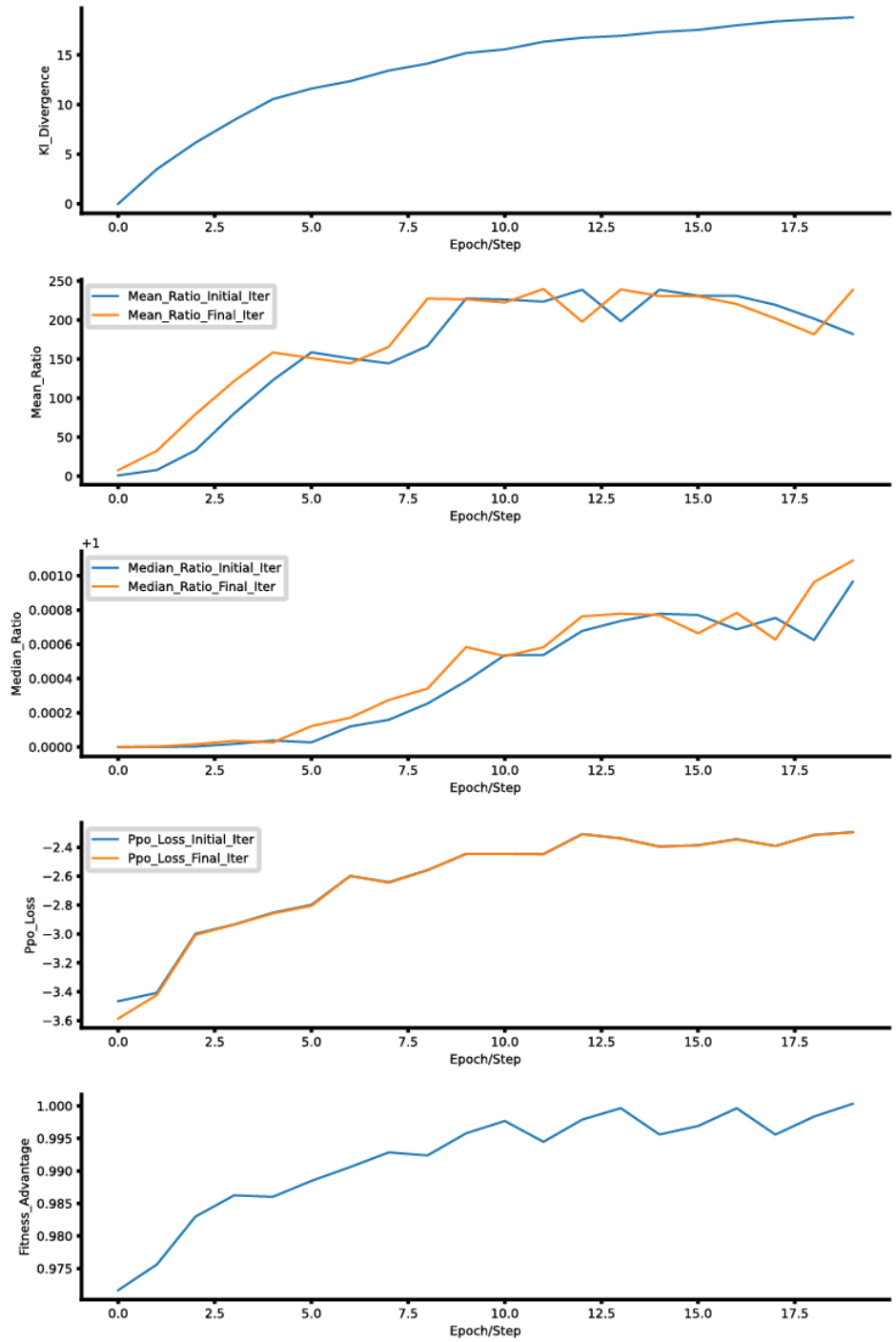

Figure S7: Training curves during alignment of the VAE showing the progression of D_{KL}, mean and median ratios of aligned model likelihoods relative to SFT, PPO clipped surrogate objective loss, and fitness advantage of generated designs.

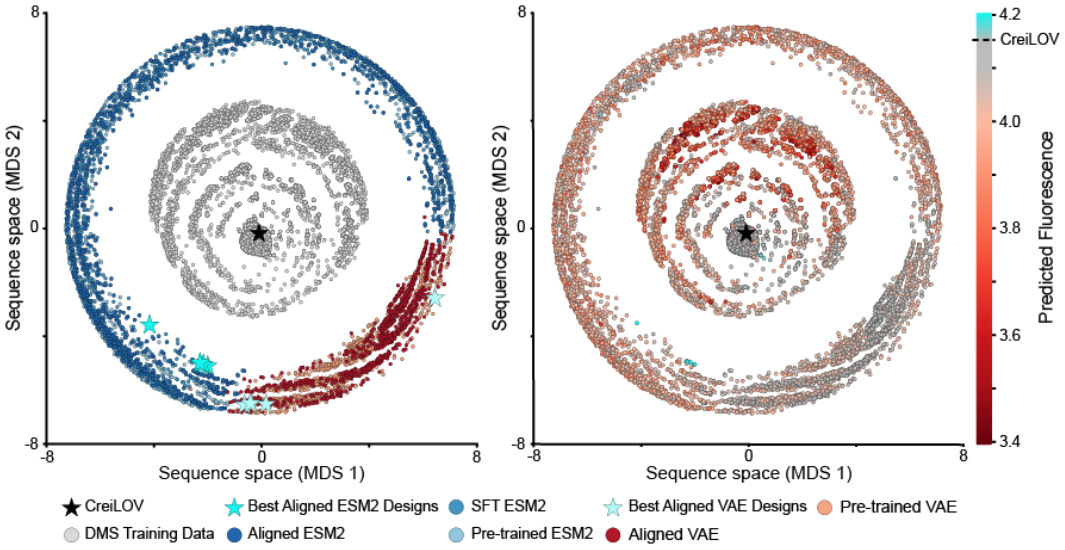

Figure S8: Multi-dimensional scaling plots visualizing how aligned models extrapolate beyond the mutation regimes present in training data to design brighter sequence variants.

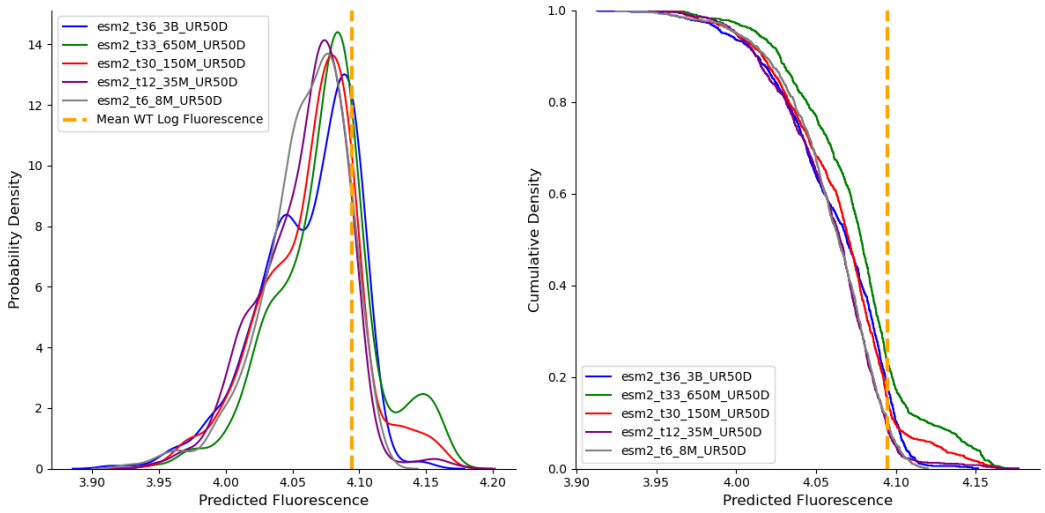

Figure S9: Kernel density and cumulative density plots of 1000 sequence designs from pre-trained ESM-2 models.

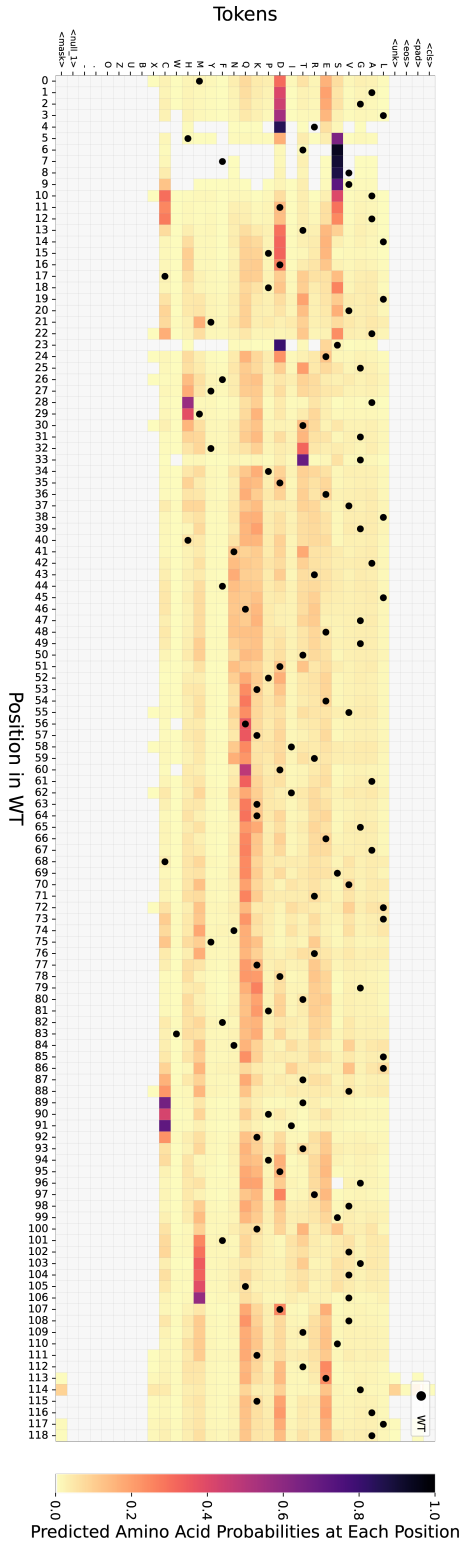

Figure S10: Single mutant logit heatmap for WT (CreiLOV) with SFT ESM-2 (35M) by masking CreiLOV one position at a time with ESM-2 and saving the logits for the masked position. We see SFT ESM2 learned several beneifical mutations such as T7S but resulted in an inability to reconstruct CreiLOV after SFT.