Conditioning protein language models using high-throughput sequence-fitness data collection

Sonia C. Yuan, Jason Yang, Jinbei Li, Bastian Vogeli, Simon R. Krarup, Emily Roberts, Bjarke Erichsen, Vanessa H. Mujica, Kenan Jijakli, Søren Karst, Lei Yang, Alex T. Nielsen, Tyler Korman, & Frances H. Arnold

Abstract

Current generative models of protein sequences, such as protein language models (pLMs), can generate novel functional sequences, but most strategies do not integrate labeled fitness data from real-world experiments. In this study, we explore fitness-conditioned generation from an autoregressive pLM, capturing evolutionary information from a protein family, using direct preference optimization (DPO) with large amounts of real-world experimental data. Our method leverages MillionFull, a high-throughput method used to collect over 100,000 unique sequence-fitness pairs for O-methyltransferases (OMTs) that form isovanillic acid, a non-native reaction. Specifically, we finetune ProGen2 on natural OMTs, after which we use the MillionFull-collected labeled dataset to align the pLM to generate sequences with higher fitness. This DPO-conditioned model generates sequences with significantly higher predicted fitness than the pretrained model while maintaining high sequence diversity and mutational profiles consistent with top-performing experimental variants. Impressively, preliminary wet-lab validation confirms that the best-performing DPO variant has a 16-fold fitness increase from the parent sequence and a 3-fold increase from the top variant in the training data. Overall, we demonstrate a robust "lab-in-the-loop" framework capable of generating diverse, high-fitness enzyme variants for non-native functional targets.

1 Introduction

Protein language models (pLMs) are powerful generative models that can produce new functional sequences, including real-world validated enzymes (Lambert et al.; Nijkamp et al.). These generative approaches are powerful because they enable broad exploration of the protein design space, unlike local search approaches such as directed evolution (Arnold), or even machine learning-assisted directed evolution (MLDE) (Wu et al.). However, these models are unconditional in the sense that they only capture the distribution of natural proteins, whereas many protein engineering tasks involve engineering enzymatic activity and function that deviates from native activity. These tasks need assay-labeled fitness data measuring protein function as well as guidance strategies to incorporate such data to steer pLMs toward the desired function (Yang et al., b; Xiong et al.).

While existing sequence-fitness datasets have provided valuable insights into protein function, they are often constrained in scale and scope by the trade-offs between mutational depth and breadth. Current deep mutational scanning (DMS) or site-saturation mutagenesis (SSM) libraries typically explore low-order mutations throughout the protein or high-order mutations at a limited set of positions (Stiffler et al.; Lite et al.; Johnston et al.; Papkou et al.; Olson et al.). Thus, there are few datasets that capture high-order mutations across a range of positions. Moreover, measuring the enzymatic function at scale is a challenging task, with most assays relying on HPLC screening of products, which limits the throughput to 1,000 sequences, a tiny fraction of sequence space. Consequently, there remains a critical

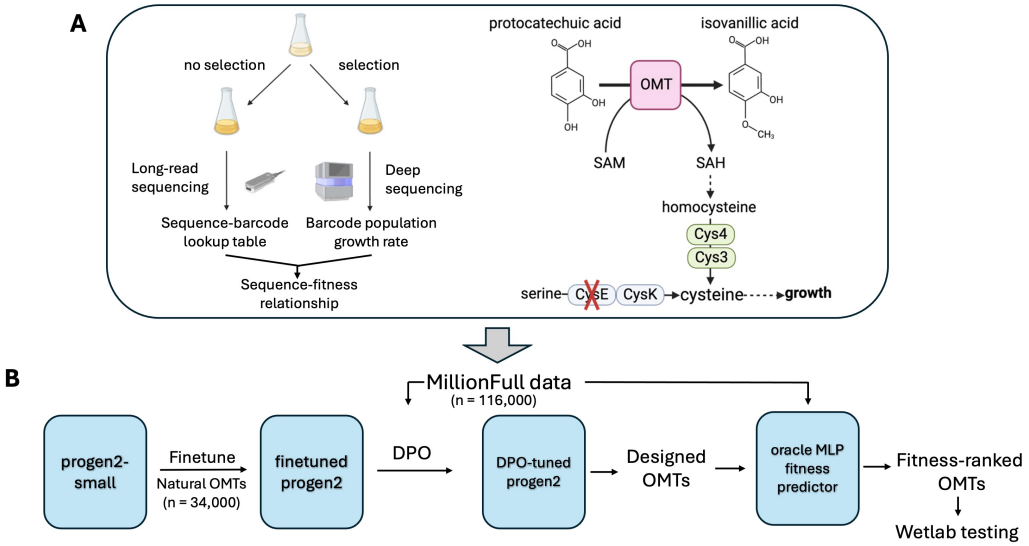

Figure 1: Summary of experimental and computational workflows. A) MillionFull enables collection of large-scale enzyme sequence-fitness data from growth-based selection assays. Left: Nanopore sequencing maps variant barcodes to sequence reads, while Illumina sequencing counts barcode frequencies in pre- and post-selection cultures which are used to calculate growth rates. Right: Cysteine pathway-knockout *E. coli* is dependent on OMT for growth. B) Machine-learning pipeline to align ProGen2 with evolutionary and experimental data

need for rich, large-scale datasets that capture high-order mutational information across a wide range of positions to fully navigate the enzymatic fitness landscape.

In addition to high-quality datasets, robust methods are required to integrate functional information into the pLM framework. To bridge the gap between unconditional generative modeling and specific functional requirements, researchers have increasingly utilized guidance strategies that steer pLMs toward desired fitness peaks (Yang et al., a; Blalock et al.; Stocco et al.; Thomas et al.). One such strategy is Direct Preference Optimization (DPO) (Rafailov et al.), a method related to reinforcement learning that, in the protein domain, involves finetuning protein language models based on feedback from labeled sequence-fitness pairs. In the DPO framework, a pLM is treated as a policy $\pi_\theta(x)$ and trained to shift its generative distribution from $p(x)$ toward a conditional distribution $p(x|y)$ by maximizing the likelihood of sampling high-fitness sequences. While DPO builds upon policy-based methods like PPO, it streamlines the optimization process by eliminating the requirement for an external reward model, instead utilizing the language model itself to implicitly represent the reward objective. Since its inception, DPO has been successfully applied to guide structure-conditioned pLMs toward increased stability (Widatalla et al.), reduce the immunogenicity of MHC-I epitopes (Gasser et al.), design EGFR binders (Stocco et al.), and antibodies (Vasan et al.). However, most existing applications rely on computationally generated preference signals and focus primarily on binding or structural objectives. Consequently, the use of DPO for enzyme engineering – conditioned on large-scale, mutationally diverse experimental data – remains largely unexplored.

To address these limitations, our work (1) advances the ability to collect real-world enzymatic fitness data in high throughput (MillionFull) and (2) explores alignment strategies for incorporating large amounts of such data into pLMs (DPO) while conserving evolutionary information from the protein family. Namely, we introduce MillionFull to construct a library of over 100,000 unique O-methyltransferase (OMT) sequence-fitness pairs targeting the non-native formation of isovanillic acid (Figure 1A). We then used this dataset to guide ProGen2 toward high-fitness variants (Figure 1B). After analyzing the proposed enzyme

variants *in silico*, we synthesize and test the top-predicted variants in the wet-lab using a high-throughput cell-free protein expression (CFPS) and screening methodology and find an ideal variant with a 16-fold boost in enzyme fitness compared to parent sequence.

2 Results

2.1 MillionFull creates a diverse library of OMTs with improved fitness relative to parent

We generated a rich dataset of OMTs variants with multiple mutations sampled across the entire protein sequence. Our library contains roughly 116,000 unique protein variants. Within the library, approximately ten unique amino acid substitutions are sampled for each position in the parent protein sequence and there is an even distribution of the identity of the mutated residue (Fig. 2A). The number of mutations per sequence follows a Poisson distribution with an average of five, and the maximum number of mutations per sequence is twenty-six (Fig. 2B, middle).

Within our library, we find that approximately 10% of variants have a higher fitness than the parent and that improved variants occur across a range of mutational orders, from one up to fifteen (Fig. 2B, left, middle). The top-performing variants share a common set of eleven mutations spread across the protein sequence, and these variants are overwhelmingly mutated to the same residue at a specific position (Fig. 2B, right). These mutations are distributed both near and far from the active site (Fig. 2C), according to an AlphaFold3 structure prediction, making structurally-guided prediction of beneficial mutations extremely difficult. Moreover, due to the high-order mutational nature of our library, specific beneficial mutations tend to co-occur with each other, such as S28T and E50K, or the combination of L228F, E272V, and F277L. Critically, the combinations of mutations in top-performing variants are distinct from the top single mutants in the dataset (Appendix, Fig. A.2). Without the scale, even positional coverage, and high-order mutant information of our MillionFull-enabled library, these insights would not have been possible, underscoring the need for such rich mutational datasets to accelerate enzyme engineering.

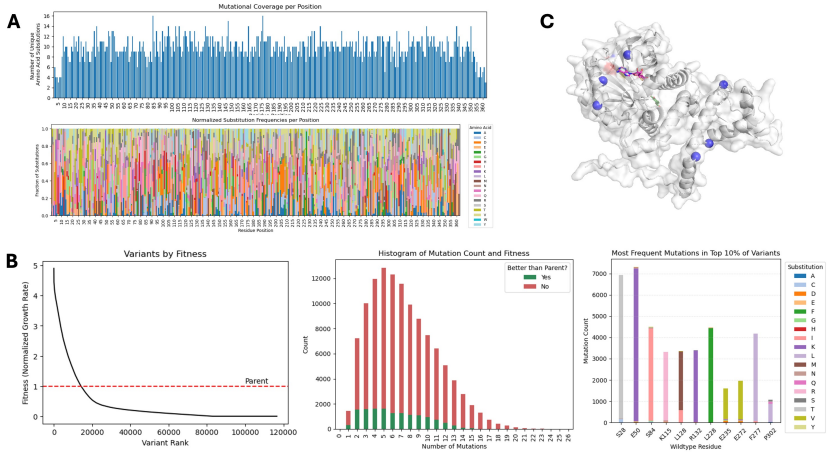

Figure 2: The MillionFull dataset. A) Top: the number of unique amino acid substitutions per position of the OMT protein. Bottom: the fraction of amino acid substitutions per position, colored by amino acid identity. B) Left: Variants ranked by fitness. Middle: Histogram of the number of mutations per sequence, colored if the variant fitness is higher (green) or lower (red) than the parent. Right: The most frequent mutations in the top 10% of variants, colored by amino acid identity. C) AlphaFold3 structure of the parent OMT. Blue spheres represent the positions of the most common mutations in top variants. The cofactor SAH is shown in magenta.

2.2 DPO enables generation of sequences with improved activity

Starting from a base pLM, namely ProGen2 finetuned on OMTs (Figure 1B), we performed DPO to align the base model based on the labeled sequence-fitness data from MillionFull. We then generated new preference-aligned sequences from the DPO-tuned model, hereby referred to as DPO-generated sequences. To evaluate the quality of DPO-generated sequences, we scored all sequences with a supervised oracle to obtain corresponding predicted fitness values. First, we ensured sufficient performance of our oracle on our evaluation set of sequences (Appendix, Fig. A.4). As a baseline, we compared the oracle-predicted fitness of randomly generated sequences, sequences generated after pretraining ProGen2 on natural OMTs, and DPO-generated sequences. Overall, the predicted fitnesses of DPO-generated sequences are significantly higher than sequences generated after pretraining and also random sequences (Fig. 3A). DPO-generated sequences have an average of 5 mutations from the parent, while pretrained variants have an average of approximately 240 mutations per sequence (Fig. 3B). This demonstrates that the process of conditioning a pLM on experimental data shifts its distribution to reflect high-quality examples in the training data. Indeed, the most frequent mutations in the top 10% of DPO-generated variants reflect an identical subset of mutations to the most frequent mutations in the training data (Fig. 3C). Overall, these *in silico* evaluations suggest that the process of conditioning a pLM on experimental data with DPO shifts generation toward high-fitness variants.

Next, we selected the top 300 DPO-generated variants, based on predicted fitness from our oracle, for wet-lab testing using CFPS to streamline the screening and validation process. Impressively, 132/300 sequences have a higher fitness than the parent, and of these, 24 sequences have a higher fitness than the top variant in the MillionFull dataset (Fig. 3D). Specifically, the best DPO-generated variant has a 16-fold fitness increase from the parent sequence and a 3-fold increase from the top variant in the training data (Fig. 3E). The best-performing DPO variants contain a combination of beneficial mutations in the training data (mutations in Fig. 2B right and Fig. 3C) with some surprising additions unaccounted for in the top variants, such as D76N, Q13H, T58S, among others (Fig. 3F). More work is needed to validate these wetlab results, interpret the model predictions, and understand the individual and combinatorial effects of mutations in different contexts.

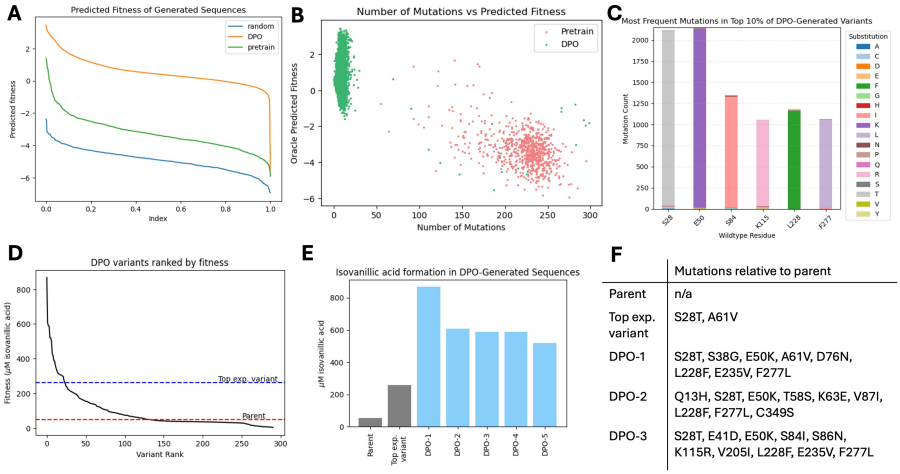

Figure 3: DPO-generated sequences. A) Variants ranked by oracle-predicted fitness of DPO-generated (orange), finetuned (green), and random (blue) sequences. B) Comparing the number of mutations vs. oracle predicted fitness of DPO-generated sequences vs pretraining sequences. C) The most frequent mutations in the top 10% of DPO-generated variants. D) Isovanillic acid activity in the 300 DPO sequences selected for wetlab testing. E) Isovanillic acid activity, μM, of top DPO-generated variants compared to the parent sequence and the top experimental variant. F) Mutations relative to the parent of the top three DPO-generated sequences and the top experimental variant.

3 Conclusion

Current pLM alignment strategies are often limited by a lack of real-world experimental data capturing the complex fitness landscapes required for enzyme engineering. We addressed this by developing MillionFull, a platform that generated over 100,000 unique sequence-fitness pairs for O-methyltransferase (OMT) activity on a non-native target. Unlike traditional DMS or SSM libraries, the mutational diversity of this data set captures the high-order combinatorial effects critical to identifying superior variants. By applying DPO to ProGen2 after finetuning on natural OMTs, we successfully aligned the generative model with this experimental landscape while incorporating evolutionary information, producing sequences with significantly higher predicted fitness than fine-tuned baselines. Preliminary wet-lab characterization demonstrates that DPO-generated sequences outperform the best variant in the training data and achieve substantially higher fitness than the parent sequence. These results show how large-scale, functional data effectively enable the conditioning of pLMs to generate proteins with desired target properties while preserving the distribution of natural sequences, and we expect that this strategy can be extended to a range of other protein engineering applications.

4 Acknowledgments

We would like to express our sincere thanks to: Dr. Le Yuan, Dr. Feiran Li, Dr. Bruce Wittmann, and Jonathan Funk for the valuable discussions that motivated the study; Prof. Paul Jensen and Dr. Viji Kandasamy for their contributions to the method design; Dr. André Faure and Dr. Arsenios Vlassis for their assistance with Illumina sequencing and data analysis; Dr. Linda Ahonen and Adrian Frey for their support with chemical analytics; Dr. Søren Karst, Dr. Scott Quinoo, Troels Hansen, Dr. Tue Jørgensen, Dr. Zofia Jarczyńska, and Keyan Liu for their contributions to nanopore experiments and data analysis; Dr. Se Hyeuk Kim, Dr. Christoffer Rode, Christina Lenhard, and Lena Heer for their help with molecular biology experiments. We would like to thank Dr. Bruce Wittmann for providing comments and edits to this manuscript, as well as Dr. Debbie Marks for insightful discussion about the results.

References

Frances H. Arnold. Directed evolution: Bringing new chemistry to life. 57(16):4143–4148. ISSN 1433-7851. doi: 10.1002/anie.201708408. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5901037/.

Nathaniel Blalock, Srinath Seshadri, Agrim Babbar, Sarah A. Fahlberg, Ameya Kulkarni, and Philip A. Romero. Functional alignment of protein language models via reinforcement learning. URL https://www.biorxiv.org/content/10.1101/2025.05.02.651993v1. Pages: 2025.05.02.651993 Section: New Results.

Andre J. Faure, Jörn M. Schmiedel, Pablo Baeza-Centurion, and Ben Lehner. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. 21(1):207. ISSN 1474-760X. doi: 10.1186/s13059-020-02091-3.

Hans-Christof Gasser, Diego A Oyarzún, Javier Antonio Alfaro, and Ajitha Rajan. Tuning ProteinMPNN to reduce protein visibility via MHC class i through direct preference optimization. 38:gzaf003. ISSN 1741-0126. doi: 10.1093/protein/gzaf003. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11970896/.

Kadina E. Johnston, Patrick J. Almhjell, Ella J. Watkins-Dulaney, Grace Liu, Nicholas J. Porter, Jason Yang, and Frances H. Arnold. A combinatorially complete epistatic fitness landscape in an enzyme active site. 121(32):e2400439121. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2400439121. URL https://pnas.org/doi/10.1073/pnas.2400439121.

Théophile Lambert, Amin Tavakoli, Gautham Dharuman, Jason Yang, Vignesh Bhethanabotla, Sukhvinder Kaur, Matthew Hill, Arvind Ramanathan, Anima Anandkumar, and Frances H. Arnold. Sequence-based generative AI design of versatile tryptophan synthases. ISSN 2041-1723. doi: 10.1038/s41467-026-68384-6. URL https://www.nature.com/articles/s41467-026-68384-6.

Thuy-Lan V Lite, Robert A Grant, Isabel Nocedal, Megan L Littlehale, Monica S Guo, and Michael T Laub. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. 9:e60924. ISSN 2050-084X. doi: 10.7554/eLife.60924. URL https://doi.org/10.7554/eLife.60924.

Hao Luo, Anne Sofie L. Hansen, Lei Yang, Konstantin Schneider, Mette Kristensen, Ulla Christensen, Hanne B. Christensen, Bin Du, Emre Özdemir, Adam M. Feist, Jay D. Keasling, Michael K. Jensen, Markus J. Herrgård, and Bernhard O. Palsson. Coupling s-adenosylmethionine–dependent methylation to growth: Design and uses. 17(3):e2007050. ISSN 1545-7885. doi: 10.1371/journal.pbio.2007050. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2007050.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. 14(11):968–978.e3. ISSN 24054712. doi: 10.1016/j.cels.2023.10.002. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471223002727.

C. Anders Olson, Nicholas C. Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. 24(22):2643–2651. ISSN 1879-0445. doi: 10.1016/j.cub.2014.09.072.

Andrei Papkou, Lucia Garcia-Pastor, José Antonio Escudero, and Andreas Wagner. A rugged yet easily navigable fitness landscape. 382(6673):eadh3860. doi: 10.1126/science.adh3860. URL https://www.science.org/doi/10.1126/science.adh3860.

Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, and Robert D Finn. HMMER web server: 2018 update. 46:W200–W204. ISSN 0305-1048. doi: 10.1093/nar/gky448. URL https://doi.org/10.1093/nar/gky448.

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. URL http://arxiv.org/abs/2305.18290.

Michael A. Stiffler, Doeke R. Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in TEM-1 -lactamase. 160(5):882–892. ISSN 1097-4172. doi: 10.1016/j.cell.2015.01.035.

Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding generative protein language models with reinforcement learning. URL http://arxiv.org/abs/2412.12979.

Neil Thomas, David Belanger, Chenling Xu, Hanson Lee, Kathleen Hirano, Kosuke Iwai, Vanja Polic, Kendra D. Nyberg, Kevin G. Hoff, Lucas Frenz, Charlie A. Emrich, Jun W. Kim, Mariya Chavarha, Abi Ramanan, Jeremy J. Agresti, and Lucy J. Colwell. Engineering highly active nuclease enzymes with machine learning and high-throughput screening. 16(3). ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2025.101236. URL https://www.cell.com/cell-systems/abstract/S2405-4712(25)00069-9.

Archit Vasan, Gautham Dharuman, Ozan Gokdemir, Heng Ma, and Arvind Ramanathan. Antibody design using preference optimization and structural inference. URL https://openreview.net/forum?id=sB0IRUc5YJ.

Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. URL https://www.biorxiv.org/content/10.1101/2024.05.20.595026v1. Pages: 2024.05.20.595026 Section: New Results.

Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. 116(18):8852–8858. doi: 10.1073/pnas.1901979116. URL https://www.pnas.org/doi/10.1073/pnas.1901979116.

Junhao Xiong, Ishan Gaur, Maria Lukarska, Hunter Nisonoff, Luke M. Oltrogge, David F. Savage, and Jennifer Listgarten. ProteinGuide: On-the-fly property guidance for protein sequence generative models. URL http://arxiv.org/abs/2505.04823.

Jason Yang, Wenda Chu, Daniel Khalil, Raul Astudillo, Bruce J. Wittmann, Frances H. Arnold, and Yisong Yue. Steering generative models with experimental data for protein fitness optimization, a. URL http://arxiv.org/abs/2505.15093. version: 2.

Jason Yang, Francesca-Zhoufan Li, Yueming Long, and Frances H. Arnold. Illuminating the universe of enzyme catalysis in the era of artificial intelligence. 0(0), b. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2025.101372. URL https://www.cell.com/cell-systems/abstract/S2405-4712(25)00205-4.

A Appendix

A.1 Data and code availability

The MillionFull dataset is released at `https://zenodo.org/records/17282389`. Code for this project will be released upon publication.

A.2 Methods

To generate the MillionFull dataset, we utilized a growth-based selection assay in *E. coli* that couples O-methyltransferase (OMT) activity to cysteine biosynthesis, where the methylation byproduct S-adenosyl-L-homocysteine (SAH) serves as the essential precursor for cell growth (Luo et al.). Starting from the promiscuous A. thaliana OMT1 template (UniProtID Q9FK25), we employed two rounds of error-prone PCR and an intermediate selection step to produce a high-diversity final library. We implemented a barcoding strategy – mapping 25bp random tags to protein sequences via Nanopore long-read sequencing – and quantified variant frequencies across triplicate growth selections using Illumina deep-sequencing. After filtering for quality, removing indels, and processing growth rates through the DiMSum suite (Faure et al.), we averaged the growth rates of degenerate nucleotide sequences to define a fitness value for over 100,000 unique protein sequences, normalized to the parent growth rate. We validated the MillionFull fitness values by measuring individual growth rates and HPLC-quantified isovanillic acid production from 34 variants across a range of fitness values (Appendix, Fig. A.1). For additional details, see Li et. al.

To implement the machine-learning pipeline, we first fine-tuned the ProGen2-small model on a curated set of 37,544 natural OMT homologs identified via HMMER (Potter et al.) search of the parent protein against the UniRef90 database. To account for phylogenetic bias, sequences were clustered using MMseqs2 (0.4 identity threshold) and weighted by cluster size during training. For functional alignment, we applied DPO using a ranked-loss objective on a training set of approximately 112,000 variants from the MillionFull dataset, reserving a 5,000-sequence universal test set (variants with ≥ 5 mutations) for evaluation. Optimal guidance parameters were determined through a grid search of β values, where $\beta = 0.5$ maximized the predicted fitness of generated variants (Appendix, Fig. A.5). Finally, we trained a supervised multi-layer perceptron (MLP) oracle on one-hot encoded sequences using the same train-test split to score in-silico designs. All DPO-generated sequences were filtered to ensure uniqueness and novelty relative to the training data.

To experimentally characterize the DPO-generated sequences, we scored all sequences with our oracle and ordered the top 300 variants as linear expression templates (LETs) containing flanking regions of pJL1. The LETs were amplified using PCR and then expressed using cell-free protein synthesis (CFPS) in a 384 well plate according to Landwehr et al. CFPS was then purified using magnetic Nickel-NTA beads in a 384 well plate and assayed in 96 well plates with 5mM PCA, 10mM SAM, 2mg/mL GsMTN, 20mM KCL, and 10mM $MgCl_2$. The reaction was run in 100mM TrisHCl at pH 8 and 25C for 18 hrs. The isovanillic product was screened via HPLC using UV absorbance at 290nm.

A.3 Supplemental Figures

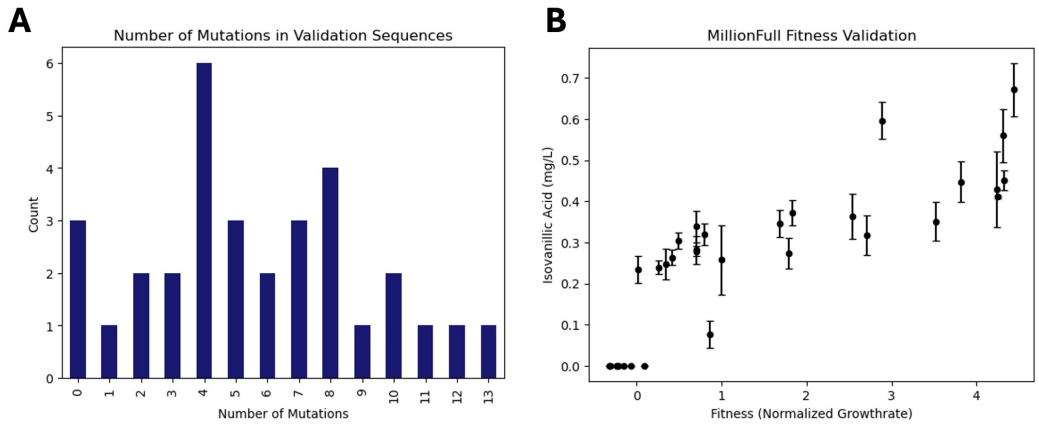

Figure A.1: MillionFull fitness validation. A) The number of mutations per sequence in the 34 sequences chosen for validation. B) MillionFull fitness, as normalized growthrate, vs isovanillic acid.

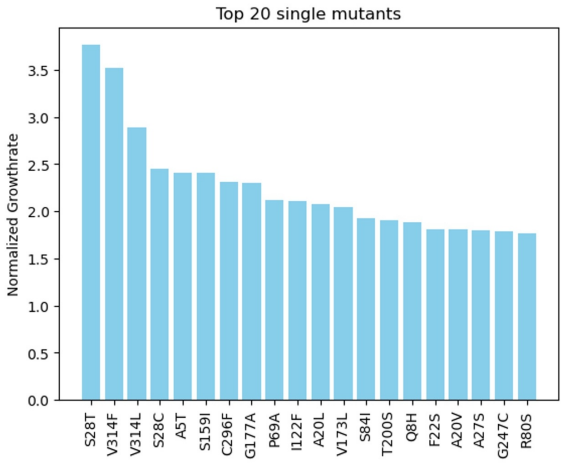

Figure A.2: Top 20 single mutants in the MillionFull dataset

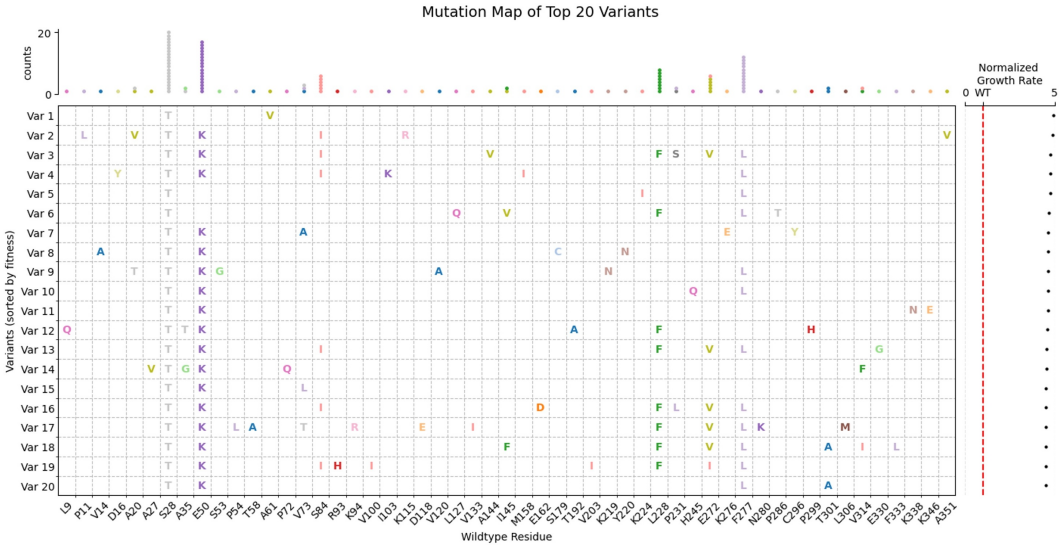

Figure A.3: Mutation map of the top 20 variants in the MillionFull dataset

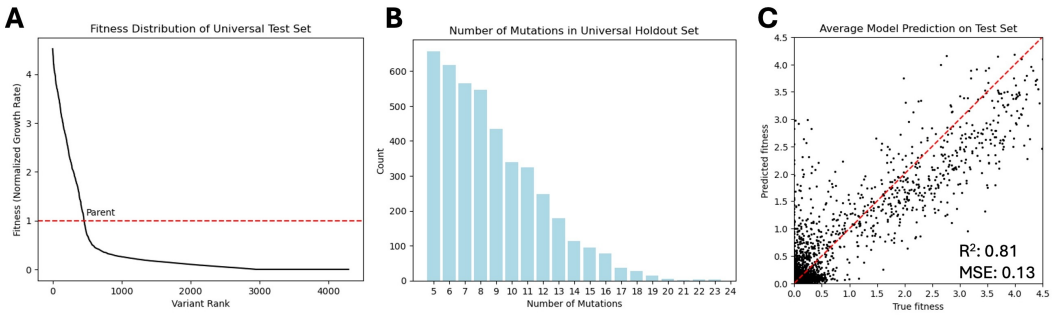

Figure A.4: Oracle evaluation. A) The fitness distribution of the 5,000 variants in the test set matches the overall distribution of the data. B) The test set consists of variants with five or more mutations. C) Oracle performance on the test set yields a R^2 score of 0.81 and a MSE of 0.13.

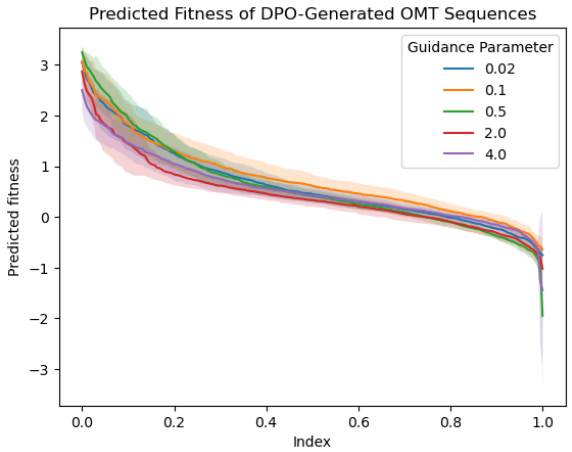

Figure A.5: Guidance parameter search for DPO.