Designing active and thermostable enzymes with sequence-only predictive models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Data-driven models of protein fitness can be useful in designing novel proteins with improved properties, but many questions remain regarding how and in what settings they should be used. Here, we ask: How can we use predictive models of protein fitness, whose predictions we might not always trust, to design protein sequences enhanced for multiple fitness functions? We propose a general approach for doing so, and apply it to design novel variants of eight different acylphosphatase and lysozyme wild types, intended to be more thermostable and at least as catalytically active as the wild types. Our method does not require a structure, experimental measurements of activity, curation of homologous sequences, or family-specific thermostability data. Experimental characterizations of our designed sequences, as well as sequences designed by PROSS, a competitive baseline method for improving protein thermostability, are currently underway and forthcoming.

1 Introduction

The design of proteins with improved fitness has historically been accomplished in the wet lab with resource-intensive directed evolution campaigns, or *in silico* with compute-intensive biophysics simulations. Recent years, however, have seen increasing interest in augmenting or replacing such methods with data-driven approaches. One paradigm is to build a data-driven model of the fitness function of interest—for example, fluorescence or enzyme activity (1)—then deploy some optimization algorithm to identify novel sequences predicted by the model to exhibit higher fitness than known protein sequences (1; 2; 3; 4).

In this work, we aim to make progress on two open challenges in this paradigm. First, how can we design sequences that are enhanced for *multiple* fitness functions? Second, how can we leverage predictive models for this task, while accounting for the fact that they are not equally trustworthy everywhere in sequence space (5)? We propose one general approach for doing so, and apply it to design novel enzyme variants intended to be more thermostable and at least as catalytically active as wild type (WT) enzymes on their native reactions, for eight WTs from two different enzyme families: acylphosphatases (ACYPs) and lysozymes.

Engineering thermostable enzymes is a longstanding problem of broad interest, as it enables their deployment in the industrial conditions required for the production of food, biofuels, and pharmaceuticals, wastewater treatment, and a host of other industrial applications (6). However, it remains challenging, in part due to inherent trade-offs between enzyme thermostability and activity (7; 8). Our *in silico* results are promising: the melting temperatures of our designed variants are predicted to surpass those of both the respective WTs and variants designed by PROSS, a competitive baseline method for enhancing thermostability (9; 10), and for six out of the eight WTs, our designed variants are also predicted to be at least as active as the WT. More importantly, experimental characterization of the thermostabilities and activities of our designed sequences and those designed by PROSS are cur-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

rently underway. We look forward to sharing the forthcoming data, which will reveal whether—and more importantly, why—our proposed approach did or did not succeed.

Another goal of this work was to investigate the use of likelihoods of models such as PROGEN2 (11)—high-capacity models of the distribution of natural protein sequences, which we henceforth refer to as *natural protein sequence density models*—as a proxy for the catalytic activity of enzymes on their native reactions, in order to design active enzyme sequences. Likelihoods of such models have been successfully used to design catalytically active lysozymes (12) and high-affinity antibodies without the need for fitness-labeled data (13). We build upon these precedents by pursuing the enhancement of multiple fitness functions—catalytic activity and thermostability—for two very different enzyme families.

Finally, our predictive model for thermostability was trained on the "meltome" dataset, which contains the melting temperatures of more than 34k unique protein sequences from the proteomes of thirteen species (14), from which we excluded any sequences from our two families of interest (nine ACYP sequences and no lysozyme sequences). If our designed sequences prove to be both catalytically active and thermostable in the wet lab, commensurate with predictions, it means our method can be used to design active, thermostable enzymes without a structure, experimental measurements of activity, curation of homologous sequences, or family-specific thermostability data.

2 Proposed computational approach

Let $\mathcal{X}$ denote the set of all amino acid sequences of length L. We consider the general goal of generating sequences, $x \in \mathcal{X}$, which are as diverse as possible, while exhibiting high values of m different fitness variables, $Y^{(1)}, \ldots, Y^{(m)} \in \mathbb{R}$. For example, we are interested in generating diverse enzyme sequences that are both highly thermostable (*i.e.*, have high melting temperature) and exhibit high catalytic activity for a reaction of interest. Therefore, we let $Y^{(1)}$ denote melting temperature and let $Y^{(2)}$ denote catalytic activity, k_{cat}/K_M, which quantifies how efficiently an enzyme converts a particular substrate into a particular product. Toward this goal, we propose sampling from the sequence distribution, $p^\star$, that solves the following optimization problem:

$$\begin{aligned} \underset{p \in \mathcal{P}}{\arg\max} \quad & H(p) \\ \text{subject to} \quad & \mathbb{E}_p[f_1(x)] \geq \tau_1, \\ & \cdots \\ & \mathbb{E}_p[f_m(x)] \geq \tau_m, \\ & \text{support}(p) \subseteq \textstyle\bigcap_{j=1}^m \textsc{TrustRegion}_j \end{aligned} \tag{1}$$

where $\mathcal{P}$ is the set of all distributions over $\mathcal{X}$, $H(p)$ denotes the entropy of the distribution p, $\tau_1, \ldots, \tau_m \in \mathbb{R}$ are desired lower bounds on the expected predicted fitnesses of sequences drawn from $p^\star$, and $\textsc{TrustRegion}_j \subseteq \mathcal{X}$ denotes the region of sequence space on which we trust the predictions of the j-th predictive model, f_j (see examples in Section 2.1). Sequences sampled from the solution to this problem will simultaneously (i) have high average predicted fitnesses and (ii) belong to the region of sequence space on which we trust the predictions of all the models.

The distribution, $p^\star$, that solves Eq. (1) has a likelihood of the following form:

$$p^\star(x) \propto \begin{cases} \exp(\sum_{j=1}^m \lambda_j \cdot f_j(x)) & \text{if } x \in \bigcap_{j=1}^m \textsc{TrustRegion}_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\lambda_1, \ldots, \lambda_m > 0$ are Lagrange multipliers corresponding to the first m constraints in Eq. (1). Our proposed approach is therefore to sample from $p^\star$, where we treat $\lambda_1, \ldots, \lambda_m$ as hyperparameters that navigate a trade-off between the expected predicted fitnesses of the designed sequences and their diversity, as quantified by the entropy of the distribution: higher values of $\lambda_1, \ldots, \lambda_m$ lead to higher expected predicted fitnesses, but also sequence distributions with lower entropy.

2.1 Predictive models and trust regions for thermostability and catalytic activity

To design thermostable, active enzymes, we used the following two predictive models and corresponding trust regions. See the Appendix for more details.

Predictive model of catalytic activity. The fitness effect of adding mutations to a WT sequence has been shown to be positively correlated with the likelihood of the mutant under natural protein sequence density (NPSD) models (15; 11; 16), particularly for fitness functions related to the natural function of the WT. We wanted to investigate whether such likelihoods could therefore be used to design active enzymes without needing experimental measurements of catalytic activity. We set $f_1(x) = f_{\text{ProGen}}(x)$ to be the average log-likelihood of a sequence, x, under five PROGEN2 models of varying sizes: PROGEN2-base, PROGEN2-medium, PROGEN2-large, PROGEN2-xlarge, and PROGEN2-BFD90.

To avoid ambiguity, we will refer to f_{ProGen} as the PROGEN2 *log-likelihood*, and will refer to the (log-)likelihood given in Eq. (2) as the *(log-)likelihood under the target* or *designed distribution*.

Trust region for catalytic activity. Correlations between fitness and NPSD model likelihoods have been demonstrated primarily on datasets comprising sequences that are mutants of a wild type sequence (WT) known to be at least moderately fit (15; 11; 16). Based on this evidence, the region of sequence space on which we trust the PROGEN2 log-likelihood of an enzyme sequence to be correlated with its activity, and to which we therefore assign $\text{TRUSTREGION}_1 = \text{TRUSTREGION}_{\text{ProGen}}$, is the set of sequences within M mutations of an active WT enzyme sequence[1]. We applied our method to four ACYP and four lysozyme WT sequences, ranging from 91 to 177 residues in length (see Appendix for details). We used two different settings of M for $\text{TRUSTREGION}_{\text{ProGen}}$: we first set it to the maximum number of mutations, M_{PROSS}, introduced by the baseline method PROSS (9)—which we describe shortly—then also ran our method with $M = M_{\text{PROSS}} + 7$ for ACYPs and $M = M_{\text{PROSS}} + 10$ for lysozymes. These two different settings will expose the effect of more and less conservative trust regions on the PROGEN2 log-likelihood.

At a high level, PROSS first restricts candidate mutations at each site to amino acids observed at that site in homologous sequences, then filters out any predicted by Rosetta to destabilize the WT. It then returns the nine combinations of the remaining candidates that score best under the Rosetta energy function (9). PROSS has been used successfully to design numerous thermostable enzymes and other proteins (9; 10; 19) and represents a proven, competitive baseline. Note that PROSS begins by constraining the set of candidate mutations, due the computational bottleneck of modeling variants with many mutations with Rosetta. In contrast, because the cost of making predictions with our data-driven models is trivial, we can tractably evaluate orders of magnitude more sequences than PROSS, including mutants with more mutations away from the WT and mutants whose individual mutations alone may not be beneficial. Ongoing experimental characterization will reveal whether this more permissive exploration of sequence space enables our method to identify even more thermostable, active variants than PROSS.

Predictive model of thermostability. We set $f_2(x) = f_{\text{melting}}(x)$ to be the predicted melting temperature of a sequence x, according to a regression model fit to the "meltome" dataset, which contains the experimentally measured melting temperatures of more than 34k unique protein sequences (14). The regression model was an ensemble of ten two-layer neural networks, trained on ESM-1b (20) embeddings of each sequence, where we took the average of the ensemble's predictions.

Trust region for thermostability. We set $\text{TRUSTREGION}_2 = \text{TRUSTREGION}_{\text{melting}}$ to contain all sequences classified as in-distribution with the meltome sequences, according to an out-of-distribution detector we built following the k-nearest neighbors strategy in (21). Briefly, we took the outputs of the final hidden layer in the thermostability predictive model for all the meltome sequences, then classified a sequence as in-distribution if the Euclidean distance between its final-layer output and the k-th nearest meltome sequence final-layer output was less than a threshold, c. The hyperparameters k and c were set using three out-of-distribution sequence datasets (see the Appendix for details).

2.2 Metropolis-Hastings algorithm

We used a Metropolis-Hasting algorithm to sample from the distribution, $p^\star$, with the likelihood given in Eq. (2) (see the Appendix for details). We are working toward open-sourcing code implementing our approach, which will soon be available at `https://github.com/salesforce/designing-thermostable-enzymes`.

[1] See (17; 18) for emerging evidence of this correlation on other kinds of datasets. However, to be conservative, our trust region mimics the "WT-centered" datasets on which this correlation has been much more robustly documented.

3 *In silico* results

Predicted thermostability. For all eight WT sequences, our method designed variants with higher predicted melting temperatures than either the WT or PROSS-designed variants (by at least 2°C to 20°C over the latter), when constrained to the same number of mutations, M_{PROSS}, as the maximum number introduced by PROSS (Fig. 1, yellow and green diamonds compared to orange circles). When the constraint was relaxed, to $M_{\text{PROSS}} + 7$ for ACYPs and $M_{\text{PROSS}} + 10$ for lysozymes, our designed sequences achieved gains of at least 15°C to 30°C in predicted melting temperature over PROSS (Fig. 1, pink and purple diamonds compared to orange circles).

PROGEN2 log-likelihood. Examining the PROGEN2 log-likelihoods adds nuance to these results. Among the eight WTs, we observed two different *in silico* phenomena: for six WTs, both our method and PROSS designed variants with higher predicted melting temperatures and higher PROGEN2 log-likelihoods (Fig. 1, left column), while for two WTs, our method designed variants predicted to be more thermostable but at the cost of lower PROGEN2 log-likelihood (Fig. 1, right column). In the latter cases, the PROSS-designed variants also had lower PROGEN2 log-likelihoods and/or lower predicted melting temperatures (Fig. 1, right column), which suggests some reassuring consistency between our approach and PROSS: when our method cannot improve predicted thermostability without sacrificing PROGEN2 log-likelihood, neither can PROSS. Note that the two challenging WTs, human ACYP-2 and T4 lysozyme (Fig. 1, right column), are known to be particularly active (22). It is therefore plausible that they are close to a Pareto frontier between activity and thermostability (8).

Concluding remarks. Wet-lab characterization of the sequences designed using both our approach and PROSS are currently underway. We look forward to sharing the results to advance the field's understanding of how to employ data-driven models to design proteins enhanced for multiple fitness functions.

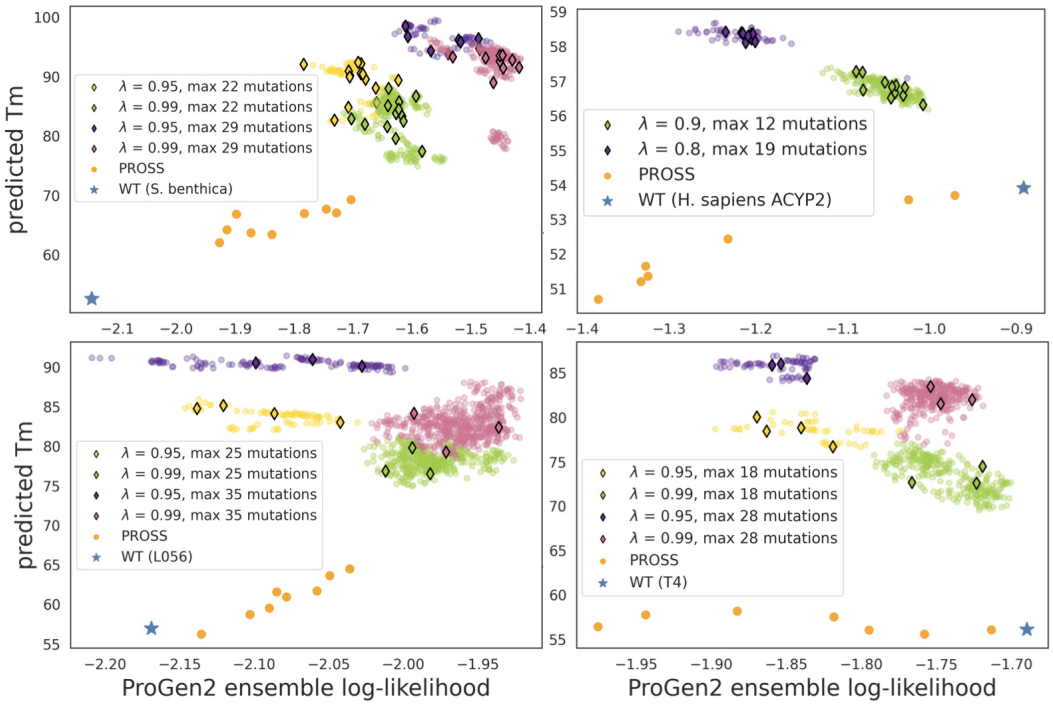

Figure 1: PROGEN2 log-likelihoods and predicted melting temperatures (C°) of variants designed using our approach and PROSS, for four example WT sequences. Top row: results for the ACYP WTs in *S. benthica* (left) and human (right). Bottom row: results for a designed lysozyme, L056 (12) (left), and the phage T4 lysozyme (right). In all subplots, the WT is the blue star, PROSS variants are orange circles, and the remaining circles are our designed sequences, where the colors correspond to different hyperparameter settings. The diamonds are random samples selected as our final designed variants, whose experimental characterizations are currently underway. The legend references two hyperparameters: the maximum number of mutations used for $\text{TRUSTREGION}_{\text{ProGen}}$, and λ, for which higher values correspond to higher PROGEN2 log-likelihoods.

References

[1] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church, "Low-N protein engineering with data-efficient deep learning," *Nature Methods*, vol. 18, no. 4, pp. 389–396, 2021.

[2] B. L. Hie and K. K. Yang, "Adaptive machine learning for protein engineering," *Curr. Opin. Struct. Biol.*, vol. 72, pp. 145–152, Dec. 2021.

[3] Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, and F. H. Arnold, "Machine learning-assisted directed protein evolution with combinatorial libraries," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, pp. 8852–8858, Apr. 2019.

[4] K. K. Yang, Z. Wu, and F. H. Arnold, "Machine-learning-guided directed evolution for protein engineering," *Nat. Methods*, vol. 16, pp. 687–694, Aug. 2019.

[5] D. H. Brookes, H. Park, and J. Listgarten, "Conditioning by adaptive sampling for robust design," in *Proc. of the International Conference on Machine Learning (ICML)*, 2019.

[6] F. Rigoldi, S. Donini, A. Redaelli, E. Parisini, and A. Gautieri, "Review: Engineering of thermostable enzymes for industrial applications," *APL Bioeng*, vol. 2, p. 011501, Mar. 2018.

[7] M. M. Pinney, D. A. Mokhtari, E. Akiva, F. Yabukarski, D. M. Sanchez, R. Liang, T. Doukov, T. J. Martinez, P. C. Babbitt, and D. Herschlag, "Parallel molecular mechanisms for enzyme temperature adaptation," *Science*, vol. 371, Mar. 2021.

[8] W. A. Baase, L. Liu, D. E. Tronrud, and B. W. Matthews, "Lessons from the lysozyme of phage T4," *Protein Sci.*, vol. 19, pp. 631–641, Apr. 2010.

[9] A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik, and S. J. Fleishman, "Automated structure- and Sequence-Based design of proteins for high bacterial expression and stability," *Mol. Cell*, vol. 63, pp. 337–346, July 2016.

[10] Y. Peleg, R. Vincentelli, B. M. Collins, K.-E. Chen, E. K. Livingstone, S. Weeratunga, N. Leneva, Q. Guo, K. Remans, K. Perez, G. E. K. Bjerga, Ø. Larsen, O. Vaněk, O. Skořepa, S. Jacquemin, A. Poterszman, S. Kjær, E. Christodoulou, S. Albeck, O. Dym, E. Ainbinder, T. Unger, A. Schuetz, S. Matthes, M. Bader, A. de Marco, P. Storici, M. S. Semrau, P. Stolt-Bergner, C. Aigner, S. Suppmann, A. Goldenzweig, and S. J. Fleishman, "Community-Wide experimental evaluation of the PROSS Stability-Design method," *J. Mol. Biol.*, vol. 433, p. 166964, June 2021.

[11] E. Nijkamp, J. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani, "ProGen2: Exploring the boundaries of protein language models." arXiv preprint 2206.13517, 2022.

[12] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, "Deep neural language modeling enables functional protein generation across families." bioRxiv preprint 2021.07.18.452833, July 2021.

[13] B. L. Hie, D. Xu, V. R. Shanker, T. U. J. Bruun, P. A. Weidenbacher, S. Tang, and P. S. Kim, "Efficient evolution of human antibodies from general protein language models and sequence information alone." bioRxiv preprint 2022.04.10.487811, 2022.

[14] A. Jarzab, N. Kurzawa, T. Hopf, M. Moerch, J. Zecha, N. Leijten, Y. Bian, E. Musiol, M. Maschberger, G. Stoehr, I. Becher, C. Daly, P. Samaras, J. Mergner, B. Spanier, A. Angelov, T. Werner, M. Bantscheff, M. Wilhelm, M. Klingenspor, S. Lemeer, W. Liebl, H. Hahne, M. M. Savitski, and B. Kuster, "Meltome atlas—thermal proteome stability across the tree of life," *Nat. Methods*, Apr. 2020.

[15] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 29287–29303, Curran Associates, Inc., 2021.

[16] P. Notin, M. Dias, J. Frazer, J. Marchena-Hurtado, A. Gomez, D. S. Marks, and Y. Gal, "Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval," in *Proc. of the International Conference on Machine Learning (ICML)*, 2022.

[17] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan, "An evolution-based model for designing chorismate mutase enzymes," *Science*, vol. 369, pp. 440–445, July 2020.

[18] C. Dallago, J. Mou, K. E. Johnston, B. J. Wittmann, N. Bhattacharya, S. Goldman, A. Madani, and K. K. Yang, "Flip: Benchmark tasks in fitness landscape inference for proteins." bioRxiv preprint 2021.11.09.467890, 2021.

[19] A. Goldenzweig and S. J. Fleishman, "Principles of protein stability and their application in computational design," *Annu. Rev. Biochem.*, vol. 87, pp. 105–129, June 2018.

[20] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, Apr. 2021.

[21] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proc. of the International Conference on Machine Learning (ICML)*, 2022.

[22] A. V. Gribenko, M. M. Patel, J. Liu, S. A. McCallum, C. Wang, and G. I. Makhatadze, "Rational stabilization of enzymes by computational redesign of surface charge-charge interactions," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, pp. 2601–2606, Feb. 2009.

[23] P. Leuenberger, S. Ganscha, A. Kahraman, V. Cappelletti, P. J. Boersema, C. von Mering, M. Claassen, and P. Picotti, "Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability," *Science*, vol. 355, Feb. 2017.

[24] E. J. Walker, J. Q. Bettinger, K. A. Welle, J. R. Hryhorenko, and S. Ghaemmaghami, "Global analysis of methionine oxidation provides a census of folding stabilities for the human proteome," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, pp. 6081–6090, Mar. 2019.

4 Appendix

4.1 Wild type sequences

We applied our method to design variants of the WT ACYP sequences in humans, the thermophile *Pyrococcus horikoshii*, the psychrophile *Shewanella benthica*, and the ACYP-like domain in the hypF WT in *Caldanaerobacter subterraneus*, as well as the WT lysozyme in phage T4 (8), two active lysozymes denoted L056 and L070 designed using an NPSD model (12), and a glycoside hydrolase WT found in *Brucella intermedia*. These WTs were chosen to span a wide range in both activity and thermostability, and had lengths ranging from 91 to 108 residues for ACYPs and 164 to 177 residues for lysozymes.

4.2 Predictive model of thermostability

The melting temperature predictive model can provide reasonably accurate predictions on a held-out enzyme family (Fig. 2). Forthcoming experimental data will clarify whether this predictive accuracy is sufficient to design enzyme variants with increased melting temperatures.

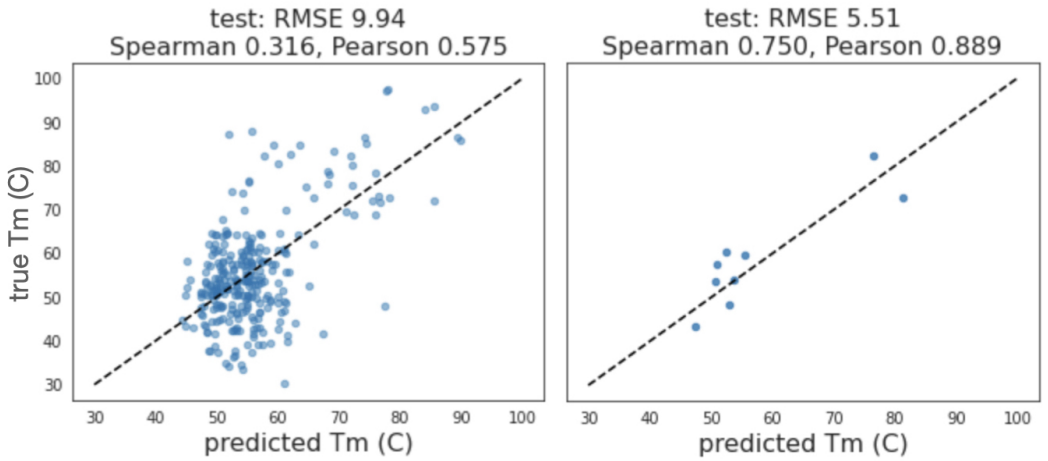

Figure 2: Performance of the thermostability predictive model on held-out ACYP-like sequences. Left: To assess the model's performance when the training and test sequences are very different, the meltome sequences were partitioned into two sets: those with sequence identity of less than 50% with each of the nine ACYP sequences in the meltome, and those with sequence identity of 50% or more with any of these ACYPs. The model was trained on the former, and predictions on the latter are shown. Right: The model was trained on all meltome sequences other than the nine ACYP sequences, and predictions on those nine ACYPs are shown. This was the final predictive model used to design ACYP variants.

4.3 Trust region for thermostability

To build the out-of-distribution (OOD) detector, following (21), we computed the final hidden layer outputs of the thermostability predictive model for all sequences in the training data—that is, all 34k unique protein sequences in the meltome dataset, other than the nine excluded ACYP sequences. Let $z_i = \phi(x_i)/\|\phi(x_i)\|_2$ denote the normalized final-layer representation of the i-th training sequence, x_i. Then given a test sequence, x', let $z_{(i)}$ denote the training sequence representation that is i-th closest to $z' = \phi(x')/\|\phi(x')\|_2$ in Euclidean distance. The test sequence is then classified as OOD if $\|z' - z_{(k)}\|_2 > c$, where k and c are hyperparameters that we set as follows. Following (21), for any value of k, the threshold, c, was set to the smallest value such that 95% of the training sequences are classified as in-distribution. The integer k was then set to the value that maximized the rate of OOD classification on sequences from three different OOD sequence datasets: (i) domain-length peptide sequences from (23), (ii) domain-length peptide sequences from (24), and (iii) full-length enzyme sequences from (7). The resulting hyperparameter values were $k = 10$ and $c = 3 \times 10^{-5}$.

4.4 Hyperparameter settings

The hyperparameters $\lambda_1 = \lambda_{\text{Progen}}$ and $\lambda_2 = \lambda_{\text{melting}}$ in Eq. (2) were set as $\lambda_{\text{Progen}} = \lambda_{\text{scale}}\lambda$ and $\lambda_{\text{melting}} = \lambda_{\text{scale}}(1-\lambda)$, where $\lambda_{\text{scale}} = 300$ and we used two different settings of λ for each WT sequence. For all WTs except for human ACYP-2, we ran our method with $\lambda \in \{0.99, 0.95\}$. For human ACYP-2, we used $\lambda \in \{0.8, 0.9\}$.

4.5 Metropolis-Hastings algorithm

We now describe the proposal distribution, $q(x' \mid x)$, that we developed for the Metropolis-Hastings (MH) algorithm, where $q(x' \mid x)$ denotes the probability of proposing the sequence x' given the current sequence, x, at each iteration of the MH algorithm. We found this proposal distribution to achieve an effective trade-off between producing sequences with high likelihood under the target distribution, thereby allowing the sampling algorithm to mix and converge in fewer iterations, and being computationally lightweight. To motivate its construction, we first observed that with naive proposal distributions, such as sampling mutants of the current sequence uniformly at random, it was often more difficult to propose a sequence, x', with higher $f_{\text{ProGen}}(x')$ than one with higher $f_{\text{melting}}(x')$. Therefore, we developed a proposal distribution intended to more efficiently propose mutants of the current sequence with higher $f_{\text{ProGen}}(x')$. The high-level intuition is that, given a multiple sequence alignment (MSA) of homologous sequences of a WT sequence, the position weight matrix (PWM) containing the frequencies of the twenty amino acids observed at each site should provide some first-order information as to which nearby mutants have high PROGEN2 likelihoods. However, we did not want practitioners to have to manually curate MSAs in order to use our method. As an alternative, we found that for many WT sequences, the entries of the following *likelihood-based position weight matrix* (L-PWM), $A \in \mathbb{R}^{L \times |\mathcal{A}|}$, are strongly correlated with the PWM computed from an MSA:

$$A_{la} = \frac{1}{Z_l} \exp(\gamma \cdot (f_{\text{ProGen}}(x_{l,a}) - f_{\text{ProGen}}(x_{\text{WT}}))),$$
$$Z_l = \sum_{a' \in \mathcal{A}} \exp(\gamma \cdot (f_{\text{ProGen}}(x_{l,a'}) - f_{\text{ProGen}}(x_{\text{WT}})))$$

where we set the hyperparameter $\gamma = 80$ to the value that maximized the average Pearson correlation between the entries, A_{la}, and the corresponding entries of the MSA-based PWM across WTs from several different protein families.

At each iteration of the MH algorithm, given the current sequence, x, we generate a proposal, x', using the following procedure:

1. Each position in the sequence x has an independent probability, $p_{\text{mutate}} = 0.02$, of being mutated.
2. For each position, $l \in \{1, \ldots, L\}$, that is to be mutated, the probability of mutating to the a-th amino acid is given by A_{la}.

We then accept or reject the proposal as prescribed by the MH algorithm. We ran the algorithm for 10k iterations and observed that chains appeared to mix and converge by at most 6k iterations. Therefore, we took a random sample of the sequences after 6k iterations as our final designed sequences.