Annealed Biological Sequence Optimization

Yuxuan Song* **Botian Wang*** **Hao Zhou** **Wei-Ying Ma**
Institute for AIR, Tsinghua University
{yxsong0816,thomaswbt123}@gmail.com,{zhouhao,Maweiying}@air.tsinghua.edu.cn

Abstract

Designing biological sequences with desired properties is an impactful research problem with various application scenarios such as protein engineering, anti-body design, and drug discovery. Machine learning algorithms could be applied either to fit the property landscape with supervised learning or generatively propose reasonable candidates to reduce wet lab efforts. From the learning perspective, the key challenges lie in the sharp property landscape, i.e. several mutations could dramatically change the protein property and the large biological sequence space. In this paper, we propose annealed sequence optimization (ANSO) and aim to simultaneously take the two main challenges into account by a paired surrogate model training paradigm and sequence sampling procedure. The extensive experiments on a series of protein sequence design tasks have demonstrated the effectiveness over several advanced baselines.

1 Introduction

Protein engineering strives to identify protein variants with superior or new biological functions, including enhanced fluorescence intensity [1], improved enzyme activity [2], and increased therapeutic efficacy [8]. Protein functions are encoded in their amino acid sequences, and this information is revealed through the spontaneous folding of the polypeptide chain into its specific three-dimensional structure [4, 16, 9]. It is believed that the proteins lie in a compact manifold in the large sequence space, *i.e.*, $20^{\|L\|}$. The protein fitness landscape, introduced by [22], has illuminated the constraints governing the distribution of protein sequences and enables us to optimize proteins for specific functions through sequence design and engineering.

From this perspective, *directed evolution* has become a favored method for optimizing protein sequences. By creating a multitude of variants and selecting those that demonstrate the highest fitness, this innovative technique mimics the natural process of evolution. With sufficient large enough cycle steps, the directed evolution is capable to gain desired proteins. However, the process requires extensive wet-lab validations which makes its utility limited. Recent advancements have sparked interest in using machine learning techniques to reduce the burden of laboratory experiments in directed evolution. It has been demonstrated the potential of such methods to significantly enhance protein engineering and generate innovative applications for this field.

Machine learning techniques for optimizing protein sequences encompass various theoretical frameworks, including Bayesian Optimization, Model-based Optimization, and Reinforcement Learning. However, these methods generally consist of two key components: the proxy module, which serves as a surrogate model for the fitness landscape to evaluate the candidate proposals, and the proposal/exploration module, which concerns how to generate novel candidates. In the task-specific scenario of protein sequence optimization, there are correspondingly key challenges for both two modules. On one hand, the protein fitness space is known to be sparse and sharp in the vast sequence space, *i.e.* a few mutations on the amino-acid could lead to a dramatic change in the fitness score. On

*Equal Contribution. Hao Zhou is the corresponding author.

ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023).

the other hand, the large sequence space makes it hard to search exhaustively. Besides, the intrinsic constraint requires the exploration lies in the relative compact protein manifolds, *e.g.* sequences with high mutation counts are laborious to synthesize for mass production.

In this paper, regarding the above-mentioned challenges, we propose annealed protein sequence optimization (ANSO). The proposed method aims to simultaneously take the two main challenges into account through a paired surrogate model training paradigm and sequence sampling procedure. Specifically, the proxy/surrogate module adopts a 'divide-and-conquer' strategy, *i.e.* decomposing the fitness landscape learning as several sub-classification tasks, to address the direct learning difficulty of the sharp and sparse protein fitness landscape. With the above-designed proxy module, we naturally designed an annealed sampling method, which utilizes intermediate guidance for climbing in the protein fitness landscape which leads to a stable and fast uphill path to the desired protein sequence domains. We conduct extensive empirical experiments to delve deeply into the property of protein optimization. Also, we demonstrate that the proposed ANSO could stably outperform several strong baselines on the eight benchmarked protein optimization tasks.

2 Methodology

2.1 Likelihood-free formulation of Surrogate Model

The surrogate model tends to fit the protein landscape and serves as the scoring function to distinguish different candidates. It is important to make the surrogate model take the uncertainty of fitness estimation into account from either the generalization perspective or the expected utility perspective. Under the probabilistic framework, fitting the protein landscape could be formatted as either a regression or a classification task. While taking uncertainty into account, the regression-based methods usually need to specify an inductive bias on the output space, *e.g.*, Gaussian hypothesis, which could limit the expressiveness of the surrogate model. Hence, the classification formulation is then applied to tradeoff the exploration and exploitation in the surrogate model. A widely used approach is to define a conditional likelihood function [17]:

$$P(\mathbb{S} \mid x) = \begin{cases} 1, & f(x) \geq \lambda \\ 0, & f(x) < \lambda \end{cases} \tag{1}$$

(f denotes the oracle fitness function.) And estimating such a conditional likelihood function is essentially a classification problem, which leads to a more general and natural interpretation of the density ratio estimation perspective:

$$\mathcal{L}_\theta = \mathbb{E}_{(x,y)\sim P_{\text{data}}} P_\theta(S|x) = \mathbb{E}_{(x,y)\sim P_{\text{data}}} u(y;\tau)\log C(\mathbf{x}) + u(y;\tau)\log(1 - C(\mathbf{x})) \tag{2}$$

Here $u(y;\tau) := \mathbb{I}(y - \tau > 0)$. And we task this a step further such an objective is equivalent to a density ratio estimation objective under some proper scoring rule [19]:

$$\mathbb{E}_{(x,y)\sim P_{\text{data}}} u(y;\tau)\log C(\mathbf{x}) + u(y;\tau)\log(1 - C(\mathbf{x})) = \mathbb{E}_{(x,y)\sim P_{\text{data}}} \frac{P(x|f(x) > \lambda)}{P(x|f(x) < \lambda)} * \frac{N_{f(x)>\lambda}}{N_{f(x)<\lambda}} \tag{3}$$

Note the above objective does not involve the explicit parameterization of the likelihood function, we refer to the above parameterization as a likelihood-free formulation.

2.2 On the Difficulty of Likelihood-free Surrogate Estimation

The density estimation perspective provides a convenient tool to help us deeply understand the challenges of fitting a likelihood-free surrogate model in the protein sequence optimization tasks. We provide the statistics of fitness scores in the Green Fluorescent Protein (avGFP) dataset, where the goal is to design sequences with higher log-fluorescence intensity values, in Fig. 1. In this task, we are interested in providing a more accurate surrogate estimation of the domains with fitness scores upon the wild type, *i.e.* the red line which could be actually seen as the outliers. As shown in this case, from the density ratio estimation perspective(Eq. 3), the two distributions could be far away from each other which could lead to a challenging ratio estimation task [13]. To tackle this challenge, we could borrow the spirit of 'divide-and-conquer' by the telescoping density ratio estimation technique [13].

$$\frac{p_0(\mathbf{x})}{p_m(\mathbf{x})} = \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})}\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \cdots \frac{p_{m-2}(\mathbf{x})}{p_{m-1}(\mathbf{x})}\frac{p_{m-1}(\mathbf{x})}{p_m(\mathbf{x})} \tag{4}$$

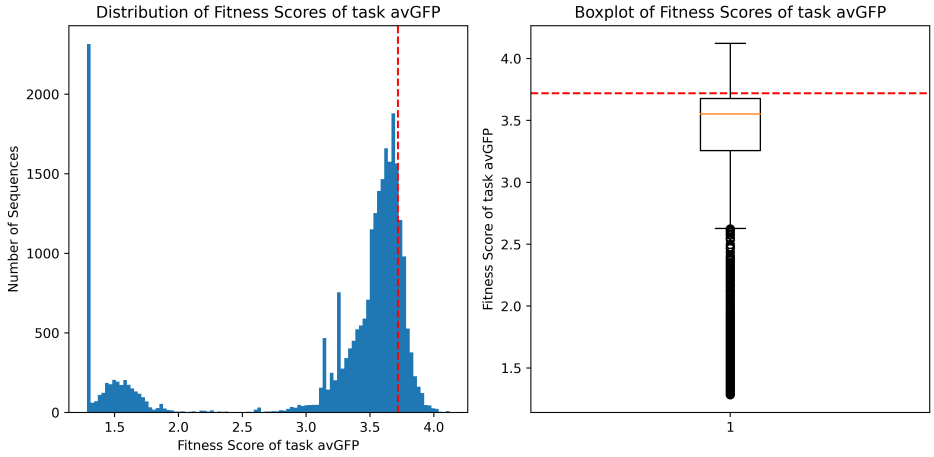

Figure 1: The statistics of the fitness score on the avGFP dataset. The red dash line stands for the wild type.

The construction of the conditional distribution in our setting, *i.e.*, $p_m(x) = p(x|f(x) > \lambda_m)$), naturally provides a convenient way to construct such distribution, by slicing the distribution with different $\lambda_0, \cdots, \lambda_m$.

2.3 Annealed Sampling

Intuitively, the proposed surrogate model training implies a bridge from the initial distribution to the top-scored distribution which we are interested in. To obtain the candidate from the estimated distribution, we could sample following the bridge and which results in a sampling procedure similar to the Annealed Importance Sampling [11]. With an estimated density ratio function as:

$$r(\mathbf{x};\boldsymbol{\theta}) = \prod_{k=0}^{m-1} r_k\left(\mathbf{x};\boldsymbol{\theta}_k\right) \approx \prod_{k=0}^{m-1} \frac{p_k(\mathbf{x})}{p_{k+1}(\mathbf{x})} = \frac{p_0(\mathbf{x})}{p_m(\mathbf{x})} \tag{5}$$

We tend to sample from the distribution p_0^θ which is implied by p_m and $r(x;\theta)$. And procedure will be decomposed as the following sampling chain:

$$p_m \to p_{m-1}^\theta \to p_{m-2}^\theta \to \cdots \to p_0^\theta \tag{6}$$

Note for each p_k^θ, we obtain samples from it by sampling from $p_{k+1}^\theta r_k(\theta_k)$.

In practice, the imbalanced proportion of the empirical samples from the intermediate distributions, *e.g.*, the limited number of samples from the top-5% distribution, imposes a great challenge in accurately estimating the density ratio. Therefore, we propose to construct the intermediate targets by ensembling the estimated density ratio function $r(x;\theta)$.

$$\hat{p}_n^\theta = p_m(x) \prod_{k=0}^{m-1} r_k(x;\theta)^{\gamma_k^n} \tag{7}$$

where γ_k^n is a parameter that trade-off the effect of density ratio estimator from different timestep. A special case is $\gamma_0^i = \cdots = \gamma_{m-1}^i$ for every timestep i, in this case, we essentially construct a single intermediate target, and as shown in the experiment we found this special case holds advantages in the offline settings.

2.3.1 A Mask-predicted Proposal

To conduct sampling, we choose a mask-predicted proposal to generate candidate protein sequences and use a Metropolis Hastings test to make sure the samples follow the intermediate distribution.

The proposal is trained by the pseudo-likelihood objective on the protein sequence to constrain the candidate still lie in the general protein distribution:

$$\mathcal{L}_\phi = \mathbb{E}_{\mathbf{x}\sim\mathbf{p}_{\text{data}}} - \frac{1}{N}\sum_{i=0}^{N-1} \log q_\phi(x_i|\mathbf{x}_{-\mathbf{i}}) \tag{8}$$

The whole sampling algorithm with MH-test is demonstrated in Algorithm. 1.

Algorithm 1 Sampling Algorithm of ANSO

Input: current protein sequence $\mathbf{x} = (x_0, \cdots, x_N)$, estimated density ratio r_θ, proposal q_ϕ
uniform sample $i = \lfloor \text{rand}(N+1) \rfloor$.
get proposal $\hat{x_i}$ according to $q_\phi(\cdot|\mathbf{x}_{-\mathbf{i}})$
compute $A = \min\left\{1, \frac{r_\theta(\hat{\mathbf{x}})q_\phi(x_i|\mathbf{x}_{-\mathbf{i}})}{r_\theta(\mathbf{x})q_\phi(\hat{x_i}|\mathbf{x}_{-\mathbf{i}})}\right\}$
if $\text{rand}(0,1) < A$ **then**
    $x_i = \hat{x_i}$
else
    $x_i = x_i$
end if

3 Experiments

Following [12], we evaluate our method on the following eight protein engineering benchmarks with both the online and offline settings:
(1) **Green Fluorescent Protein (avGFP)**[15]. (2) **Adeno-Associated Viruses (AAV)** [3]. (3) **TEM-1 β-Lactamase (TEM)** [5]. (4) **Ubiquitination Factor Ube4b (E4B)** [18]. (5) **Aliphatic Amide Hydrolase (AMIE)** [21]. (6) **Levoglucosan Kinase (LGK)** [7]. (7) **Poly(A)-binding Protein (Pab1)** [10]. (8) **SUMO E2 Conjugase (UBE2I)** [20].

We compare the proposed method with several strong baselines following [12]. We conduct experiments in both online and offline settings and design various algorithmic variants. The results could be found in Tab. 1. It should be noted that in all experiments, we have standardized the telescope size to 3 and generated the intermediate distribution by choosing the highest 10%, 30%, and 50% of the data points. The mask proposal is initialized by the ESM model following [17]. More implementation details could be found in appendix A.

Table 1: Maximum fitness scores of all models on eight protein datasets. The best performance in each task for the online and offline settings are **bolded** respectively.

Method / Task	avGFP	AAV	TEM	E4B	AMIE	LGK	Pab1	UBE2I
ANSO (offline)	3.44	5.61	**0.72**	2.59	0.15	0.01	**0.77**	2.89
ANSO (offline regression)	**3.71**	**5.62**	0.30	**5.39**	**0.21**	**0.03**	0.65	2.69
ANSO (offline w/o ESM)	3.23	0.89	0.01	1.13	0.15	-0.01	0.62	**2.94**
ANSO (online)	3.63	6.59	0.01	2.44	**0.22**	0.02	1.00	**2.98**
ANSO (online w/o mh)	**3.79**	**7.59**	0.03	**3.87**	0.19	0.01	**1.41**	2.84
PEX	3.12	4.45	**0.27**	1.21	0.16	**0.04**	1.23	2.97
AdaLead	1.88	3.58	0.15	1.94	0.16	0.04	1.25	2.96
DyNA PPO	1.86	-3.33	0.02	-0.23	-2.15	-0.52	0.49	2.08
DbAS	1.90	2.24	0.09	0.61	-0.61	-0.03	0.96	2.90
CbAS	2.08	2.22	0.09	0.47	-0.51	0.04	0.83	2.93
Wildtype	1.34	-10.06	0.00	-2.70	-8.37	-1.71	0.02	-0.22

From the above results, we could find the proposed methods, with different implementation strategies, could considerably improve upon advanced baselines on all optimization tasks with different variants. Due to the data shift between different tasks, the model with the same configuration shows huge dynamics among the tasks. And we also have a single configuration that improves on 5 of the 8 tasks.

We can also draw interesting observations when comparing the algorithmic variants. For the offline setting, we note that the classification mechanism and the pretrained proposal module all have a great impact on the empirical performance. For the online setting, we find that the MH sampling module actually impairs performance in some tasks. We hypothesize that the iterative exploration mechanism in the online setting is beneficial for reducing uncertainty in estimating the protein landscape. However, adopting the MH sampling procedure tends to prioritize exploitation, potentially limiting the scope of exploration and thus leading to performance drop. We leave further investigations as future work.

4 Conclusion

Our paper introduces ANSO, a pioneering framework that leverages intermediate density ratio estimation tasks to enhance protein sequence design. By using this approach, we are able to improve surrogate model training and facilitate an annealed sampling procedure that leads to better results. Our density ratio perspective is highly flexible and holds promise for even broader objectives, which we plan to explore in future research.

A Experimental Setting and Implementation Details

For the offline setting, we create 128 sample chains in total. For each chain, we sample 2 steps after a 15-step burn-in stage (5 for each proxy). Hence, 256 new protein sequences are proposed and evaluated in this setting. For the online setting, we base our implementation on the PEX [12] codebase. 10 rounds of optimization are conducted, with 100 new sequences proposed for each round. We adopt the MuFacNet proposed in [12] as the proxy architecture. For the training of the proxy model in the online and offline settings, we set the learning rate as 1e-3 and 1e-4, respectively, and use the Adam optimizer [6] to optimize the Binary Cross Entropy loss. We stop training when the loss fails to decrease for 10 epochs. Note that less new sequences are evaluated in the offline setting (256 v.s. 1000), but the training of the offline proxy requires the entire labeled dataset. The two settings are therefore not directly comparable but we show the results of both settings in table 1 to intuitively compare their empirical performance.

For computational efficiency, the $q_\phi(\cdot|\mathbf{x}_{-\mathbf{i}})$ term in Algorithm. 1 is neglected in implementation. We use the 6-layer ESM-1b model [14] as the proposal model. Both online and offline algorithms take approximately 1 hour to run on a Linux server with 1 Nvidia GeForce RTX 3090 GPU. All experiments are repeated for 40 times in the online settings and 3 times in the offline settings.

References

[1] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

[2] Jesse D Bloom and Frances H Arnold. In the light of directed evolution: pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences*, 106(supplement_1):9995–10000, 2009.

[3] Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

[4] Cyrus Chothia. Principles that determine the structure of proteins. *Annual review of biochemistry*, 53(1):537–572, 1984.

[5] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular biology and evolution*, 31(6):1581–1592, 2014.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014.

[7] Justin R Klesmith, John-Paul Bacik, Ryszard Michalczyk, and Timothy A Whitehead. Comprehensive sequence-flux mapping of a levoglucosan utilization pathway in e. coli. *ACS synthetic biology*, 4(11):1235–1243, 2015.

[8] HA Daniel Lagassé, Aikaterini Alexaki, Vijaya L Simhadri, Nobuko H Katagiri, Wojciech Jankowski, Zuben E Sauna, and Chava Kimchi-Sarfaty. Recent advances in (therapeutic protein) drug development. *F1000Research*, 6, 2017.

[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. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[10] Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *Rna*, 19(11):1537–1551, 2013.

[11] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

[12] Zhizhou Ren, Jiahan Li, Fan Ding, Yuan Zhou, Jianzhu Ma, and Jian Peng. Proximal exploration for model-guided protein sequence design. *bioRxiv*, 2022.

[13] Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

[14] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019.

[15] Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

[16] Benjamin S Schuster, Gregory L Dignon, Wai Shing Tang, Fleurie M Kelley, Aishwarya Kanchi Ranganath, Craig N Jahnke, Alison G Simpkins, Roshan Mammen Regy, Daniel A Hammer, Matthew C Good, et al. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proceedings of the National Academy of Sciences*, 117(21):11421–11431, 2020.

[17] Zhenqiao Song and Lei Li. Importance weighted expectation-maximization for protein sequence design. *arXiv preprint arXiv:2305.00386*, 2023.

[18] Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit. Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 2013.

[19] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[20] Jochen Weile, Song Sun, Atina G Cote, Jennifer Knapp, Marta Verby, Joseph C Mellor, Yingzhou Wu, Carles Pons, Cassandra Wong, Natascha van Lieshout, et al. A framework for exhaustively mapping functional missense variants. *Molecular systems biology*, 13(12):957, 2017.

[21] Emily E Wrenbeck, Laura R Azouz, and Timothy A Whitehead. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature communications*, 8(1):1–10, 2017.

[22] Caroline F Wright, Sarah A Teichmann, Jane Clarke, and Christopher M Dobson. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, 438(7069):878–881, 2005.