Unexplored regions of the protein sequence-structure map revealed at scale by a library of "foldtuned" language models

Arjuna Subramanian
Division of Biology and Biological Engineering
California Institute of Technology
Pasadena, CA 91125
amsubram@caltech.edu

Matt Thomson
Division of Biology and Biological Engineering
California Institute of Technology
Pasadena, CA 91125
mthomson@caltech.edu

Abstract

Nature has likely sampled only a fraction of all protein sequences and structures allowed by the laws of biophysics. However, the combinatorial scale of amino-acid sequence-space has traditionally precluded substantive study of the full protein sequence-structure map. In particular, it remains unknown how much of the vast uncharted landscape of far-from-natural sequences consists of alternate ways to encode the familiar ensemble of natural folds; proteins in this category also represent an opportunity to diversify candidates for downstream applications. Here, we characterize sequence-structure mapping in far-from-natural regions of sequence-space guided by the capacity of protein language models (pLMs) to explore sequences outside their natural training data through generation. We demonstrate that pretrained generative pLMs sample a limited structural snapshot of the natural protein universe, including >300 common (sub)domain elements. Incorporating pLM, structure prediction, and structure-based search techniques, we surpass this limitation by developing a novel "foldtuning" strategy that pushes a pretrained pLM into a generative regime that maintains structural similarity to a target protein fold (e.g. TIM barrel, thioredoxin, etc) while maximizing dissimilarity to natural amino-acid sequences. We apply "foldtuning" to build a library of pLMs for >700 naturally-abundant folds in the SCOP database, accessing swaths of proteins that take familiar structures yet lie far from known sequences, spanning targets that include enzymes, immune ligands, and signaling proteins. By revealing protein sequence-structure information at scale outside of the context of evolution, we anticipate that this work will enable future systematic searches for wholly novel folds and facilitate more immediate protein design goals in catalysis and medicine.

1 Introduction

The collection of naturally occurring protein structural motifs ("protein folds") cataloged to date cannot reflect exhaustive sampling of all possible sequence-structure pairs – there are $20^{100} \approx 10^{130}$ choices for a small domain of length 100, dwarfing even the exploratory capacity of a few billion

NeurIPS 2023 Generative AI and Biology Workshop.

years of evolutionary time. Faced with such a daunting scale, biophysicists have long contemplated what sequences and structures fill the unseen parts of protein-space. One pervasive question is that of which protein folds are most "designable," that is, which structures tolerate the greatest sequence variation, and moreover, the most substantial departure from natural sequence space [Fontana, 2002, England and Shakhnovich, 2003]? The hidden degeneracy of the protein sequence-to-structure mapping (Figure 1**a**) holds implications for determining fundamental "rules" distinguishing stable well-folded proteins from gibberish amino-acid strings, accessing diverse candidates for protein design tasks, and even demystifying the roles of certain classes of proteins at the origins of life [Dupont et al., 2010, Alva et al., 2015].

Attempts to probe the designability question have historically been stymied by both the combinatorial complexity of sequence-space and the time-consuming nature of experimental protein structure determination. However, advances in deep learning methods for proteins now place characterizing the structural ensemble of far-from-natural sequences within reach. Transformer-based protein language models (pLMs) such as ProtGPT2, ProtT5-XL, and the ESM2 family can generalize to novel sequences and structures beyond their natural training data, suggesting that they might serve as guides into meaningful regions of far-from-natural sequence-space, skirting the high-dimensional sampling problem. [Ferruz et al., 2022, Elnaggar et al., 2022, Lin et al., 2022, Verkuil et al., 2022]. Likewise, rapid structure prediction via models such as AlphaFold2 and ESMFold makes assaying the structure side of the sequence-structure map computationally tractable [Jumper et al., 2021, Lin et al., 2022]. Combining pLMs with rapid protein structure prediction (ESMFold), we show that "off-the-shelf," pretrained pLMs possess a latent capacity to generate sequences beyond the natural protein universe that map onto roughly 300 known structural motifs; however, the resulting structure distributions are skewed relative to the natural case. Building on these findings, we introduce a new "foldtuning" algorithm that modifies a PLM to preserve generative fidelity to a target fold while moving progressively further into far-from-natural sequence-space; we apply this approach for >700 common folds, uncovering well-folded regions of sequence-space and offering preliminary insight into how designability varies between folds.

2 Results

2.1 Untuned pLMs access a subset of known protein structures

We initially assess the ability of two commonly-used pLMs, ProtGPT2 and ESM2-150M, to sample from the full global sequence-structure landscape [Ferruz et al., 2022, Lin et al., 2022]. We generate $\sim 10^6$ sequences of 100aa from each model, via L-to-R next-token prediction and Gibbs sampling for ProtGPT2 and ESM2-150M respectively. Generated sequences are fed into an ESMFold structure prediction step and predicted structures are queried against a custom Foldseek database comprised of the 36,900 representative experimental structures (covering 1579 labeled protein folds, each a structurally conserved unit) of the **S**tructural **C**lassification **O**f **P**roteins (SCOP) dataset in TMalign mode [van Kempen et al., 2023, Andreeva et al., 2020]. We validate this structure prediction and assignment workflow by assessing the capacity of ESMFold to generalize to far-from-natural sequences with recent experimental structures deposited in the Protein Data Bank, finding a median backbone alignment RMSD of 0.92 ± 0.14 Å on $n = 122$ sequences/structures satisfying basic quality control filters (Figure S1). Among ProtGPT2- and ESM2-150M-generated sequences, just 385 (24.4%) and 309 (19.6%) unique SCOP folds are represented respectively according to structure prediction and assignment (Figure 1**d**).

To determine where pLM-generated sequences lie with respect to natural sequence space, we extract the internal representations ("embeddings") of these sequences with the ESM2-650M model, reduce dimensionality to 2D using UMAP, and apply a rule-of-thumb that the embeddings of qualitatively similar sequences should co-localize [McInnes et al., 2018]. We observe that a subpopulation of ProtGPT2-generated sequences and most ESM2-150M-generated sequences co-localize more substantially with random amino-acid sequences than with a set of ≈ 1.5 million natural proteins (Figure 1**b-c**). Furthermore, many of the pLM-generated sequences that are assignable to SCOP folds do not co-localize with natural sequences (Figure 1**c**). In contrast, $\sim 10^6$ sequences generated with a SOTA inverse-folding model, ESM-IF1, achieve high predicted structural fidelity but remain ensconced in natural sequence space, co-localizing almost perfectly with natural sequences (Figure 1**b**) [Hsu et al., 2022]. Taken together, it is clear that, unlike off-the-shelf inverse-folding models,

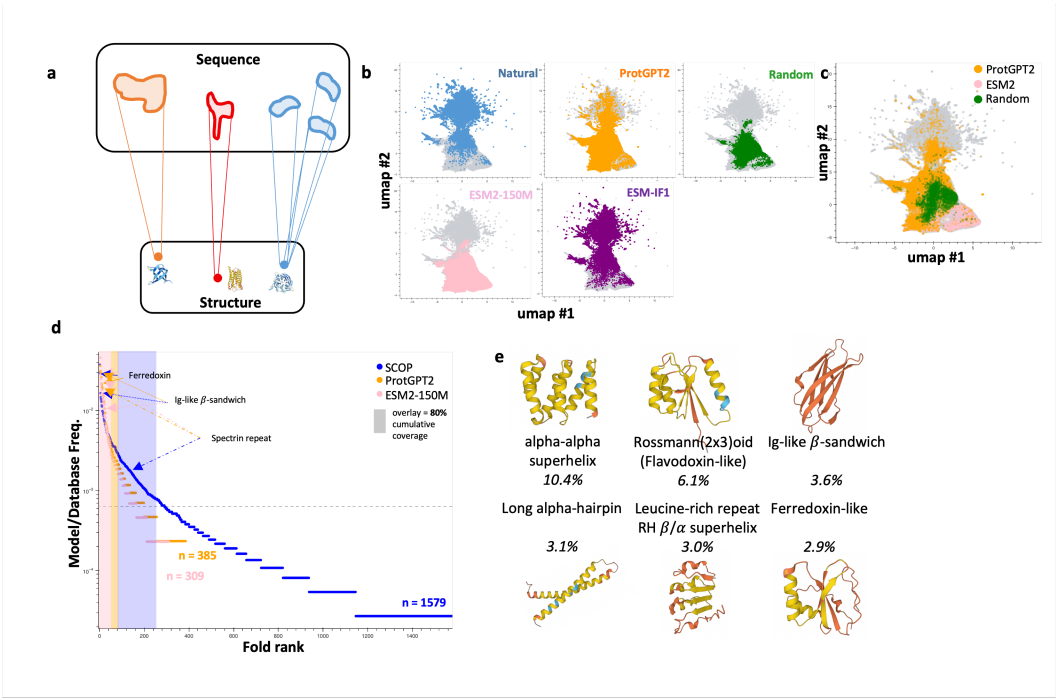

Figure 1: **Structural ensembles generated by pretrained language models are imperfect reflections of natural protein-space. a)** Subview of the global protein sequence-structure map; a given structure is encoded by multiple sequences, possibly "connected" in some informative space. **b)** Dimension-reduced UMAP representation of ESM2-150M embeddings of natural, random, inverse-folded, and pLM-generated sequences. **c)** UMAP representation of random and pLM-generated sequences assignable to a SCOP fold. **d)** Rank-ordered fold abundance plots for natural and pLM-generated sequences. **e)** The 6 most-common SCOP folds among ProtGPT2 outputs; representative structures are of far-from-natural sequences (no pBLAST hit with E-value < 10)

both pLMs generate sequences that are appreciably distinct in some statistical sense compared to natural sequences yet able to fold into many of the same 3D structures. Notably, the structural distributions emitted by the two pLMs indicate strong preferences for small subsets of folds at rates far exceeding their natural frequencies (Figure 1**d-e**). For ProtGPT2, which has a higher overall structural hit rate – 20.0% vs. 2.9% for ESM2-150M – overrepresented folds include α-α superhelices, Rossmann(2x3)oids, and immunoglobulin-like domains (Figure 1**e**).

2.1.1 Loosening pLM sampling constraints increases sequence novelty at the cost of structure hit rates and diversity

Foundational results from natural language processing suggest that protein structure and sequence diversity might be unlocked by changing sampling hyperparameters to increase generative options at each next-token prediction step. To determine whether this hypothesis holds for pLMs, we systematically varied two key sampling hyperparameters of ProtGPT2 – sampling temperature and top_k (the number of highest-probability tokens available to sample from at a given step) – and repeated the generation, structure prediction, and structure assignment workflow from Section 2.1 for batches of $\sim 10^6$ sequences. Consistent with the notion that "flattening the energy landscape" of sequence generation should boost novelty, we find that the fraction of generated sequences lacking detectable homology to natural protein sequences grows as temperature and top_k are increased (Table S1-S6, Figure S5). However, we concurrently observe that any gains in sequence novelty are obviated by marked losses on the structure generation front. As sampling temperature increases, the generation frequency of fold-assignable structures falls by a factor of roughly 2x, from 34.5% to 15.1% for the default top_k value of 950, and the number of unique SCOP folds identified plummets by $> 25\%$ (Table S1, Figure S2). Additionally, high-temperature generation dramatically favors

generation of proteins with an all-α global topology, *i.e.* α-helical bundles, at the notable expense of the functionally diverse α/β class (Figure S2-S4) [Choi and Kim, 2006]. While obtaining far-from-natural sequences for α-helical bundle proteins is useful for protein design writ large, the extreme structural biases introduced by pushing sampling hyperparameters into the regime necessary for sequence novelty indicate that a more robust method is required to access far-from-natural sequences for structurally diverse fold classes.

2.2 "Foldtuning" of a pLM maintains a target structure while escaping natural sequence space

Finding the structural reach of pretrained pLMs to be distorted, particularly when sequence novelty is an overriding goal, we introduce a new approach to push pLMs into far-from-natural sequence space. In this approach, which we term "foldtuning," a pLM undergoes initial finetuning on natural sequences corresponding to a given target fold, plus several rounds of finetuning on self-generated batches of sequences that are predicted to adopt the target fold while differing maximally from the natural training sequences (Algorithm 1). We achieve this by selecting for finetuning those structurally-validated sequences that maximize semantic change – defined for a generated sequence $s_k^{(i)}$ as the smallest L_1-distance between the ESM2-650M embeddings of $s_k^{(i)}$ and any of the natural training sequences [Hie et al., 2021]. Thus, foldtuning drives a pLM along a trajectory that accesses pockets of far-from-natural sequences while preserving the "grammar" of a fixed target fold.

Algorithm 1 pLM "Foldtuning"

given a pretrained base model M_{-1} and target fold f
for round $k = 0, 1, 2 \ldots N$ **do**
  if $k = 0$ **then**
    let training set S_k contain n (default: $n = 100$) natural examples of fold f
  else
    let training set S_k contain all $s_{k-1}^{(i)}$ s.t. $z_{k-1}^{(i)}$ is among the n-largest values (highest semantic change) of the *(k-1)*-th round (see line 13)
  end if
  finetune M_{k-1} on S_k for 1 epoch, outputting updated model M_k
  generate N (default: $N = 1000$) sequences $s_k^{(0)}, s_k^{(1)}, \ldots$ from M_k
  fold generated sequences (ESMFold)
  assign fold labels by structure-based search (Foldseek)
  for all $\{s_k^{(i)}\}$ assigned to fold f **do**
    let semantic change $z_k^{(i)} = \min_j \|x^{(i)} - x_{train}^{(j)}\|_1$, where $s_k^{(i)} \mapsto x_k^{(i)} \in \mathbb{R}^{1280}$ via embedding with ESM2-650M
  end for
end for

2.2.1 "Foldtuned" models emit far-from-natural sequences for >700 target folds, including enzymes, cytokines, and signaling proteins

We "foldtune" ProtGPT2 (the best-performing model from Section 2.1) as described in Algorithm 1 for 727 total target folds; 708 SCOP folds (out of the top 850 ranked by natural abundance, for an 83.3% success rate), plus 19 cytokines/chemokines curated from InterPro. Model performance is assessed via two metrics; the fraction of generated sequences predicted to fold to the target structure, aka *structure hit rate*; and the fraction of structural hits with no sequence homology to any protein in the UniRef50 database (per MMseqs2 search, max $E = 0.01$), aka *sequence "escape rate"*. Considering all 727 models, two rounds of "foldtuning" increase the median structure hit rate from to 0.565 from 0.203 after initial finetuning on the natural target fold; median sequence escape rate increases markedly after four "foldtuning" rounds, to 0.211 from 0.134 after the initial finetuning phase (Figure 2**a**). As a second measure of sequence novelty, semantic change w.r.t. natural fold members increases steadily with each "foldtuning" round (Figure 2**b**). With most models, exemplified by many of the top-10 most-abundant SCOP folds, maximizing sequence escape rate does not require any significant decline in structure hit rate (Figure 2**c**). Exceptions to this rule include the cytokine tumour necrosis factor (TNF) and G protein-coupled receptors (GPCRs), which exhibit model performance tradeoff between structural fidelity and sequence novelty (Figure 2**d**). Lastly, taking the product of structure hit and sequence escape rates as a proxy for global designability splits the 727 folds into at least three subpopulations, with the right-handed β-helix, ribbon-helix-helix domain, TIM β/α-barrel, pleckstrin homology domain, and α/α toroid ranked as the most-designable folds (Figure 2**e**).

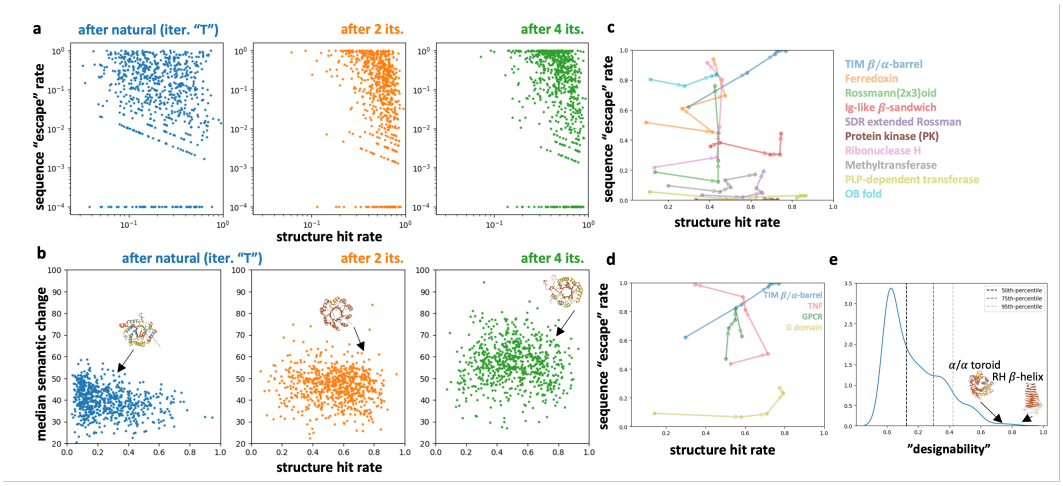

Figure 2: **727 "foldtuned" models achieve high structure hit and sequence escape rates. a-b)** Structure hit rates after initial ('T'), 2, and 4 rounds of "foldtuning", paired with **a,** sequence "escape" rates; **b,** sequence median semantic change w.r.t. natural fold members (embeddings extracted from ESM2-650M model). **c-d)** Hit/escape rate trajectories over 4 rounds of "foldtuning" for **c,** the 10 most-common natural folds; **d,** selected enzymatic, immune modulation, and signaling examples. **e**) Distribution of fold "designability" based on product of structure hit and sequence escape rates

3 Discussion

Knowing the features of the global protein sequence-structure map would unlock virtually limitless possibilities for protein design. We demonstrated that protein language models trained on the natural portions of this map can access far-from-natural sequence space, albeit with biases in preferred structures. We developed and deployed a "foldtuning" strategy to systemically explore deep into the far-from-natural corners of this map for 727 diverse targets including enzymes and signaling ligands/receptors. Beyond serving translational goals in protein design for health and catalysis, we expect that with tweaks to selection criteria, "foldtuning" will be readily repurposed to search for novel protein structures unseen in nature and complete the sequence-structure map.

Acknowledgments and Disclosure of Funding

We thank Steve Mayo, Carl Pabo, Zach Martinez, Alec Lourenco, Lucas Schaus, Blade Olson, Joe Boktor, as well as all members of the Thomson Lab for helpful discussions.

This work was supported by National Institutes of Health under award number R01GM150125, the Moore Foundation, the Packard Foundation, and the Heritage Medical Research Institute. The authors have no competing interests to disclose.

References

V. Alva, J. Söding, and A. N. Lupas. A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, 4:e09410, Dec. 2015. ISSN 2050-084X. doi: 10.7554/eLife.09410. URL https://doi.org/10.7554/eLife.09410. Publisher: eLife Sciences Publications, Ltd.

A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1):D376–D382, Jan. 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz1064. URL https://doi.org/10.1093/nar/gkz1064.

I.-G. Choi and S.-H. Kim. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences*, 103(38):14056–14061, Sept. 2006.

doi: 10.1073/pnas.0606239103. URL https://www.pnas.org/doi/abs/10.1073/pnas.0606239103. Publisher: Proceedings of the National Academy of Sciences.

C. L. Dupont, A. Butcher, R. E. Valas, P. E. Bourne, and G. Caetano-Anollés. History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proceedings of the National Academy of Sciences*, 107(23):10567–10572, June 2010. doi: 10.1073/pnas.0912491107. URL https://www.pnas.org/doi/abs/10.1073/pnas.0912491107. Publisher: Proceedings of the National Academy of Sciences.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, Oct. 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

J. L. England and E. I. Shakhnovich. Structural Determinant of Protein Designability. *Physical Review Letters*, 90(21):218101, May 2003. doi: 10.1103/PhysRevLett.90.218101. URL https://link.aps.org/doi/10.1103/PhysRevLett.90.218101. Publisher: American Physical Society.

N. Ferruz, S. Schmidt, and B. Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32007-7. URL https://www.nature.com/articles/s41467-022-32007-7. Number: 1 Publisher: Nature Publishing Group.

W. Fontana. Modelling 'evo-devo' with RNA. *BioEssays*, 24(12):1164–1177, 2002. ISSN 1521-1878. doi: 10.1002/bies.10190. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.10190. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.10190.

B. Hie, E. D. Zhong, B. Berger, and B. Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, Jan. 2021. doi: 10.1126/science.abd7331. URL https://www.science.org/doi/full/10.1126/science.abd7331. Publisher: American Association for the Advancement of Science.

C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8946–8970. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hsu22a.html. ISSN: 2640-3498.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. d. S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. Technical report, bioRxiv, Dec. 2022. URL https://www.biorxiv.org/content/10.1101/2022.07.20.500902v3. Section: New Results Type: article.

Z. A. Martinez, R. M. Murray, and M. W. Thomson. TRILL: Orchestrating Modular Deep-Learning Workflows for Democratized, Scalable Protein Analysis and Engineering, Oct. 2023. URL https://www.biorxiv.org/content/10.1101/2023.10.24.563881v1. Pages: 2023.10.24.563881 Section: New Results.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Feb. 2018. URL https://arxiv.org/abs/1802.03426v3.

M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, and M. Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, pages 1–4, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://www.nature.com/articles/s41587-023-01773-0. Publisher: Nature Publishing Group.

R. Verkuil, O. Kabeli, Y. Du, B. I. M. Wicky, L. F. Milles, J. Dauparas, D. Baker, S. Ovchinnikov, T. Sercu, and A. Rives. Language models generalize beyond natural proteins. Technical report, bioRxiv, Dec. 2022. URL https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1. Section: New Results Type: article.

4 Appendix

4.1 Methods

Except where otherwise noted, interfacing with all models was via the TRILL software package as described in Martinez et al. [2023]. Sections 4.1.1-4.1.4 provide further implementation details for the "foldtuning" steps described in Algorithm 1.

4.1.1 Sampling from Base Models

For ProtGPT2, we sampled 119,067 sequences by L-to-R next-token prediction with the default best-performing hyperparameters from Ferruz et al. [2022]. (sampling temperature 1, top_k 950, top_p 1.0, repetition penalty 1.2), terminating after 40 tokens or the first STOP token, whichever came first, and truncating sequences to the first 100aa as necessary. For ESM2-150M, we sampled 148,500 sequences from L-to-R using Gibbs sampling for next-token prediction with a default sampling temperature of 1, no repetition penalty, and allowing for sampling from all tokens, terminating after 100aa or the first STOP token, whichever came first.

The random-sequence control set was generated by position-independent sampling of 74,250 sequences of length 100aa from the 20 proteinogenic amino acids, with sampling probability for each amino acid proportional to its natural abundance (first-order statistics). The inverse-folding control set was constructed by generating three sequences from ESM-IF1 with each of the 36,900 representative structures in the SCOP database as a backbone template, for 110,700 sequences in total. Default hyperparameters for sampling were taken as in Hsu et al. [2022].

4.1.2 Finetuning of ProtGPT2 and Sampling from Finetuned Models

All finetuning of ProtGPT2 was performed with the Adam optimizer using a learning rate of 0.0001 and next-token prediction as the causal language modeling task. For "foldtuning" on a target fold f, the base ProtGPT2 model was finetuned in the initial "T" round for 1-3 epochs on 100 natural sequences belonging to fold f and selected randomly among deduplicated hits from a Foldseek-TMalign search of the SCOP database of superfamily representative PDB structures ($n = 36900$) against the AlphaFold-UniRef50 predicted structure database. Identical optimizer parameters were used for subsequent foldtuning rounds, finetuning for 1 epoch on 100 semantic-change-maximizing sequences assigned to fold f.

Sampling from finetuned ProtGPT2 models followed the same general procedures and hyperparameters as in 4.1.1, with 1000 sequences generated per finetuned model. Inference batch size on a single A100-80GB GPU ranged from 125-500 sequences per batch depending on target sequence length.

4.1.3 Structure Prediction and Assignment

Structures were predicted for all generated sequences – from control, base, and finetuned models – that passed a quality control check for absence of rare or ambiguous amino acid characters (B, J, O, U, X, Z). Sequences were truncated to a max length of 100aa (base or control models) or the median length of natural sequences for target fold f (finetuned models). All structures were predicted with ESMFold as described in Lin et al. [2022]. Inference batch size on a single A100-80GB GPU ranged from 10-500 sequences per batch depending on target sequence length.

If possible, each predicted structure was assigned a fold label by searching against the SCOP database of superfamily representative PDB structures ($n = 36900$) with Foldseek in accelerated TMalign mode as described in van Kempen et al. [2023] and selecting the SCOP fold accounting for the most hits satisfying TM-score > 0.5 and max(query coverage, target coverage) > 0.8 ("consensus hit").

4.1.4 Sequence Selection for Foldtuning

For each target fold f and foldtuning round $k = 1, 2, ...N$, the semantic change was calculated for all sequences $\{s_k^{(i)}\}$ assigned to fold f (as described in Section 4.1.3) as $z_k^{(i)} = \min_j \|x^{(i)} - x_{train}^{(j)}\|_1$, where $s_k^{(i)} \mapsto x_k^{(i)} \in \mathbb{R}^{1280}$ via embedding with ESM2-650M. The finetuning sequence set for

subsequent round $k+1$, S_k, was constructed by ranking the $\{s_k^{(i)}\}$ by their z_k in descending-order and taking the top 100 corresponding $s_k^{(i)} \in S_k$.

4.1.5 ESMFold Validation on Far-From-Natural Sequences

To assess the accuracy of ESMFold structural prediction on out-of-distribution samples, we evaluated model performance on *de novo* proteins with structures deposited in the Protein Data Bank (PDB) on-or-after the ESMFold training cutoff date of 05-01-2020. Mirroring the training set construction process described in Lin et al. [2022], we filtered out structures with Resolution > 9 Å, length $\leq$ 20aa, rare or ambiguous amino acids (BJOUXZ), or containing > 20% sequence composition of any one amino acid, and clustered remaining sequences at the 40% identity level, obtaining a validation set of $n = 122$ sequences. For each of the 122 sequences, the backbone RMSD was calculated between the ESMFold predicted structure and the ground-truth PDB experimental structure, with a median alignment RMSD of 0.92 ± 0.14 Å, indicating successful generalization of ESMFold beyond natural training data (Figure S1).

4.2 Supplemental Figures

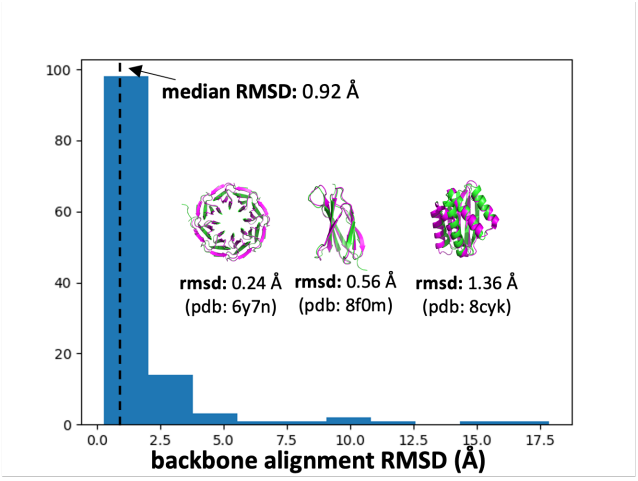

Figure S1: **ESMFold achieves high structure prediction accuracy on far-from-natural sequences.** Backbone aligment RMSD for ESMFold structures of $n = 122$ *de novo* proteins vs. experimental ground-truth structures, covering various global topologies. All sequences in the validation set had experimental structures deposited in the Protein Data Bank on-or-after the ESMFold training cutoff of 05-01-2020.

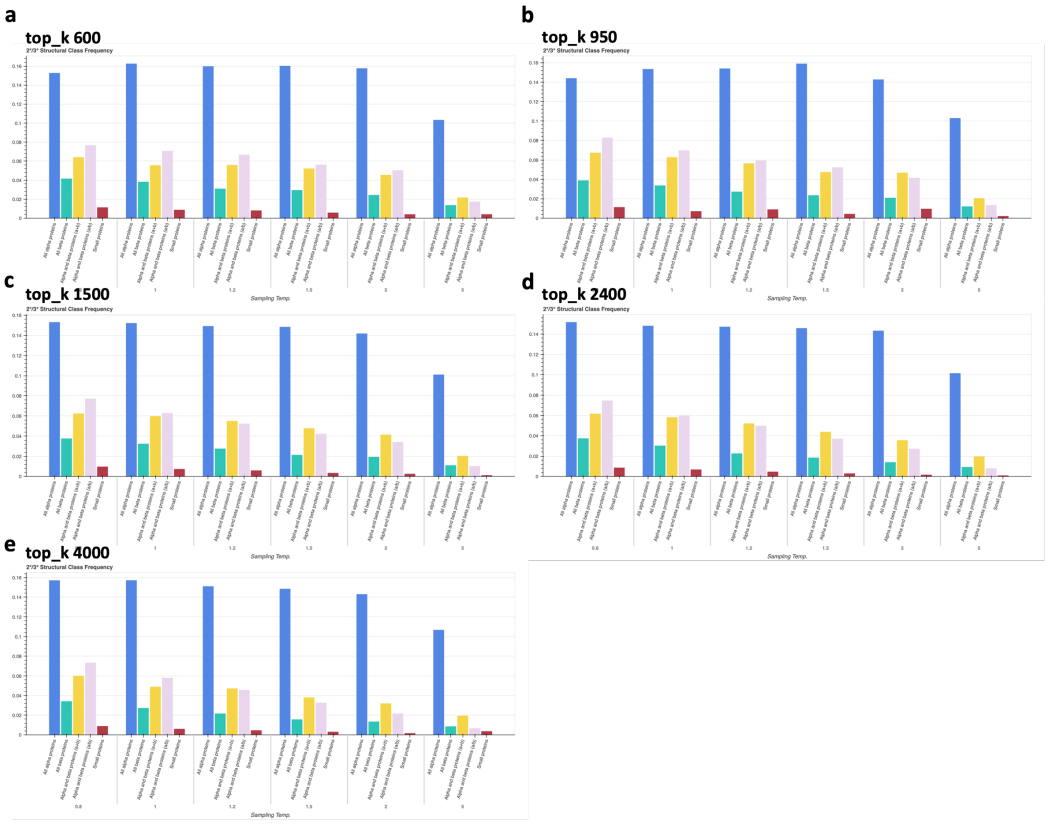

Figure S2: **Structure hit rates from base ProtGPT2 decrease as sampling temperature and top_k increase. a-e)** Structure hit rates from batches of 100k sequences generated from ProtGPT2 for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) and top_k values (number of highest-probability tokens considered in sampling out of 50,256 total) – **a,** 600, **b,** 950, **c,** 1500, **d,** 2400, **e,** 4000; broken down by protein global topology class (α, β, $\alpha + \beta$, α/β, or "small / minimal 2° structure")

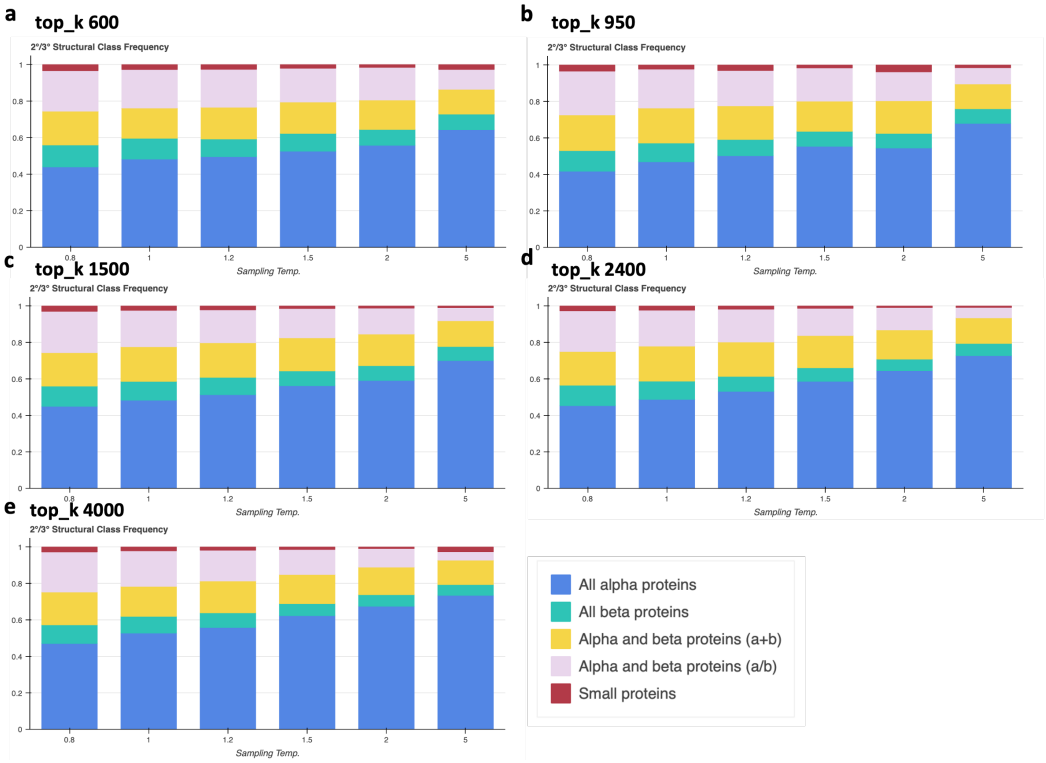

Figure S3: **Generated fold distributions shift towards all-α proteins and away from α/β proteins as sampling temperature increases. a-e)** Frequency of each protein global topology class (α, β, $\alpha + \beta$, α/β, or "small / minimal 2° structure") among all structure hits within batches of 100k sequences generated from ProtGPT2 for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) and top_k values (number of highest-probability tokens considered in sampling out of 50,256 total) – **a,** 600, **b,** 950, **c,** 1500, **d,** 2400, **e,** 4000

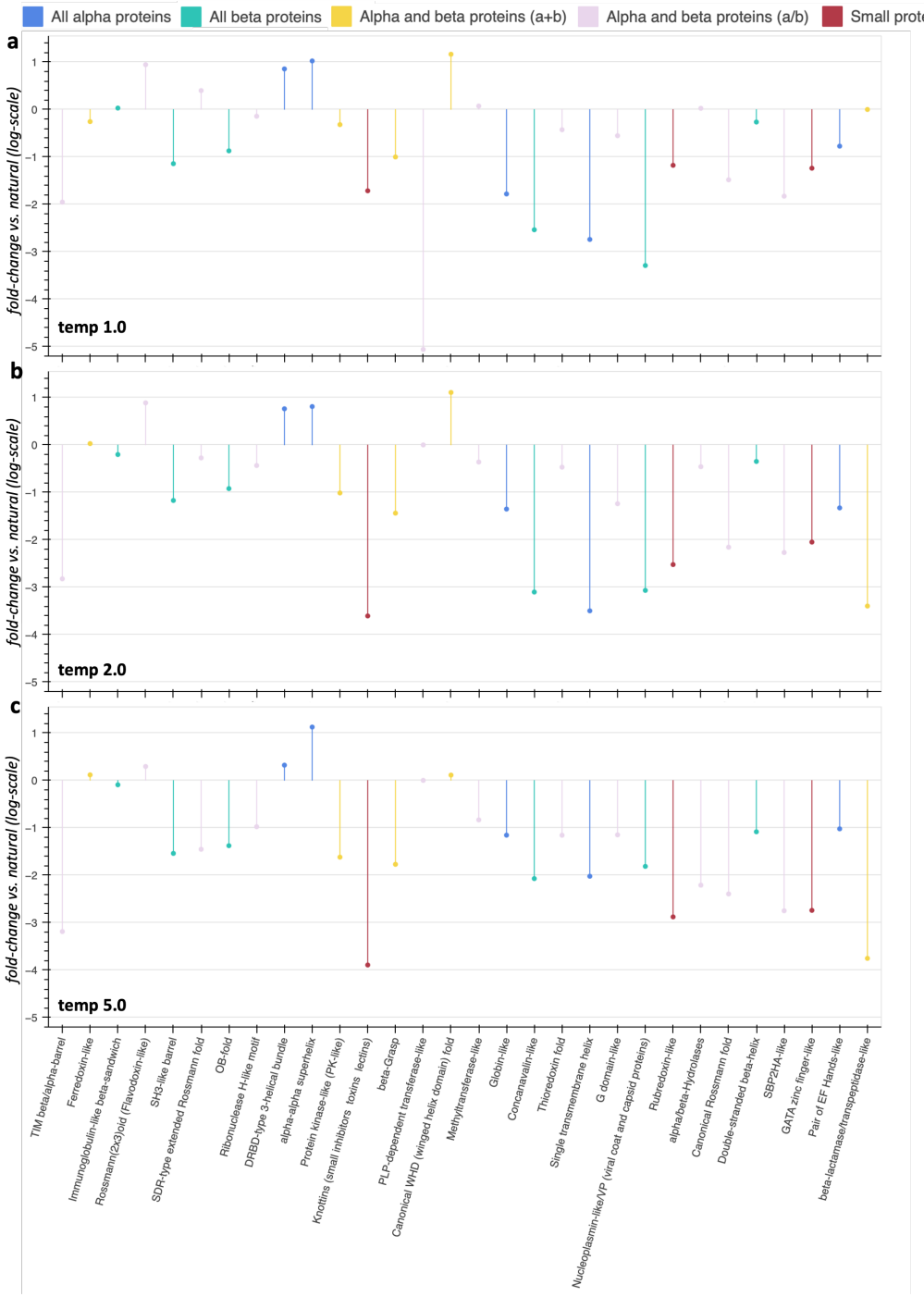

Figure S4: **Generated fold distributions differ substantially from the natural fold distribution across sampling temperatures. a-c)** Log-scale enrichment among ProtGPT2-generated sequences assigned to the 30 most-abundant SCOP folds with top_k 950 and sampling temperature **a,** 1, **b,** 2, or **c,** 5.

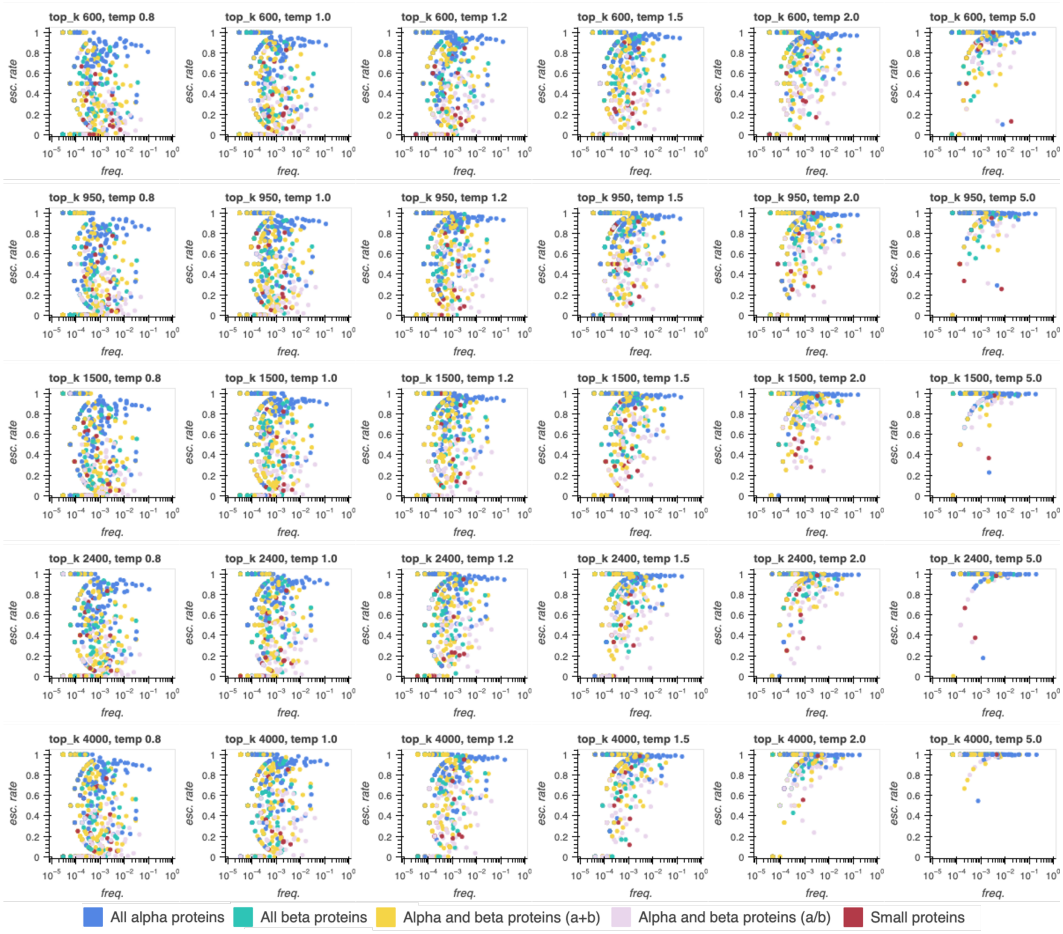

Figure S5: **Sequence escape rates increase across most folds as sampling temperature increases, at the cost of a shift towards all-α topologies.** Sequence escape rates for all assigned SCOP folds generated from ProtGPT2 within batches of 100k sequences for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) x several top_k values (number of highest-probability tokens considered in sampling out of 50,256 total; 600, 950, 1500, 2400, 4000.

4.3 Supplemental Tables

Table S1: **Base ProtGPT2 sequence and structure generation performance depends on sampling hyperparameters.**

Hyperparams		Results			
top_k	temp	Valid Seq.	# Folds	Struct. Hit	Seq. Esc.
600	0.800	1.000	658	0.347	0.445
600	1.000	1.000	635	0.336	0.545
600	1.200	1.000	645	0.322	0.629
600	1.500	1.000	617	0.304	0.717
600	2.000	0.999	606	0.282	0.797
600	5.000	0.981	513	0.160	0.912
950	0.800	1.000	643	0.345	0.466
950	1.000	1.000	668	0.327	0.580
950	1.200	0.999	620	0.306	0.674
950	1.500	1.000	625	0.287	0.766
950	2.000	0.998	587	0.262	0.855
950	5.000	0.985	473	0.151	0.958
1500	0.800	1.000	649	0.340	0.484
1500	1.000	1.000	646	0.315	0.609
1500	1.200	0.999	627	0.290	0.708
1500	1.500	1.000	608	0.263	0.816
1500	2.000	0.998	577	0.239	0.903
1500	5.000	0.988	476	0.144	0.981
2400	0.800	1.000	634	0.334	0.493
2400	1.000	1.000	634	0.303	0.628
2400	1.200	1.000	617	0.277	0.742
2400	1.500	1.000	588	0.248	0.857
2400	2.000	0.998	542	0.222	0.944
2400	5.000	0.991	460	0.139	0.993
4000	0.800	1.000	662	0.334	0.510
4000	1.000	1.000	644	0.298	0.650
4000	1.200	0.999	618	0.271	0.778
4000	1.500	1.000	574	0.238	0.894
4000	2.000	0.998	540	0.212	0.968
4000	5.000	0.993	442	0.145	0.998

Table S2: **Top SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k 600.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.095	0.033	0.810
Spectrin repeat-like	α	0.050	0.017	0.871
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.048	0.017	0.146
Immunoglobulin-like beta-sandwich	β	0.036	0.012	0.510
alpha-alpha superhelix	α	0.033	0.011	0.386
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.112	0.038	0.874
Spectrin repeat-like	α	0.056	0.019	0.907
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.014	0.252
Immunoglobulin-like beta-sandwich	β	0.034	0.012	0.596
Hemerythrin-type up-and-down 4-helical bundle	α	0.032	0.011	0.903
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.122	0.039	0.908
Spectrin repeat-like	α	0.063	0.020	0.929
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.014	0.330
Hemerythrin-type up-and-down 4-helical bundle	α	0.036	0.012	0.937
Immunoglobulin-like beta-sandwich	β	0.030	0.010	0.696
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.130	0.040	0.943
Spectrin repeat-like	α	0.068	0.021	0.950
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.013	0.424
Hemerythrin-type up-and-down 4-helical bundle	α	0.040	0.012	0.958
Immunoglobulin/albumin-binding domain-like	α	0.033	0.010	0.955
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.141	0.040	0.969
Spectrin repeat-like	α	0.075	0.021	0.968
Hemerythrin-type up-and-down 4-helical bundle	α	0.044	0.013	0.978
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.012	0.500
Immunoglobulin/albumin-binding domain-like	α	0.032	0.009	0.966
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.154	0.025	0.989
Spectrin repeat-like	α	0.084	0.014	0.989
Hemerythrin-type up-and-down 4-helical bundle	α	0.060	0.010	0.984
alpha-alpha superhelix	α	0.040	0.006	0.905
Immunoglobulin/albumin-binding domain-like	α	0.033	0.005	0.987

Table S3: **Top SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k 950.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.101	0.035	0.841
Spectrin repeat-like	α	0.050	0.017	0.872
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.048	0.016	0.177
Immunoglobulin-like beta-sandwich	β	0.032	0.011	0.534
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.031	0.011	0.342
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.111	0.036	0.893
Spectrin repeat-like	α	0.058	0.019	0.918
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.014	0.273
Hemerythrin-type up-and-down 4-helical bundle	α	0.034	0.011	0.926
alpha-alpha superhelix	α	0.031	0.010	0.571
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.126	0.039	0.930
Spectrin repeat-like	α	0.065	0.020	0.946
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.041	0.013	0.345
Hemerythrin-type up-and-down 4-helical bundle	α	0.038	0.012	0.948
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.030	0.009	0.530
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.136	0.039	0.943
Spectrin repeat-like	α	0.069	0.020	0.969
Hemerythrin-type up-and-down 4-helical bundle	α	0.046	0.013	0.960
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.012	0.491
Immunoglobulin/albumin-binding domain-like	α	0.030	0.009	0.960
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.149	0.039	0.978
Spectrin repeat-like	α	0.076	0.020	0.984
Hemerythrin-type up-and-down 4-helical bundle	α	0.045	0.012	0.976
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.010	0.596
Immunoglobulin/albumin-binding domain-like	α	0.035	0.009	0.974
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.178	0.027	0.991
Spectrin repeat-like	α	0.090	0.014	0.996
Hemerythrin-type up-and-down 4-helical bundle	α	0.064	0.010	0.989
Immunoglobulin/albumin-binding domain-like	α	0.038	0.006	0.986
alpha-alpha superhelix	α	0.035	0.005	0.934

Table S4: **Top SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k 1500.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.101	0.035	0.847
Spectrin repeat-like	α	0.052	0.018	0.878
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.045	0.015	0.210
Immunoglobulin-like beta-sandwich	β	0.033	0.011	0.555
alpha-alpha superhelix	α	0.031	0.010	0.426
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.119	0.037	0.895
Spectrin repeat-like	α	0.059	0.019	0.918
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.013	0.304
Hemerythrin-type up-and-down 4-helical bundle	α	0.035	0.011	0.930
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.029	0.009	0.472
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.129	0.038	0.930
Spectrin repeat-like	α	0.067	0.019	0.956
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.012	0.425
Hemerythrin-type up-and-down 4-helical bundle	α	0.040	0.012	0.951
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.029	0.008	0.528
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.145	0.038	0.963
Spectrin repeat-like	α	0.077	0.020	0.984
Hemerythrin-type up-and-down 4-helical bundle	α	0.047	0.012	0.976
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.011	0.566
Immunoglobulin/albumin-binding domain-like	α	0.033	0.009	0.968
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.158	0.038	0.988
Spectrin repeat-like	α	0.078	0.019	0.989
Hemerythrin-type up-and-down 4-helical bundle	α	0.050	0.012	0.986
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.039	0.009	0.708
Immunoglobulin/albumin-binding domain-like	α	0.034	0.008	0.987
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.182	0.026	0.993
Spectrin repeat-like	α	0.094	0.014	0.999
Hemerythrin-type up-and-down 4-helical bundle	α	0.070	0.010	0.994
Ferredoxin-like	$\alpha+\beta$	0.040	0.006	0.984
Immunoglobulin/albumin-binding domain-like	α	0.038	0.005	0.993

Table S5: **Top SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k 2400.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.106	0.036	0.850
Spectrin repeat-like	α	0.052	0.018	0.894
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.014	0.210
alpha-alpha superhelix	α	0.032	0.011	0.435
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.031	0.011	0.358
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.125	0.038	0.905
Spectrin repeat-like	α	0.062	0.019	0.939
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.012	0.353
Hemerythrin-type up-and-down 4-helical bundle	α	0.038	0.011	0.917
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.028	0.008	0.446
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.138	0.038	0.945
Spectrin repeat-like	α	0.071	0.020	0.959
Hemerythrin-type up-and-down 4-helical bundle	α	0.043	0.012	0.957
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.041	0.011	0.456
Ferredoxin-like	$\alpha+\beta$	0.030	0.008	0.792
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.158	0.039	0.976
Spectrin repeat-like	α	0.077	0.019	0.985
Hemerythrin-type up-and-down 4-helical bundle	α	0.052	0.013	0.981
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.038	0.010	0.601
Ferredoxin-like	$\alpha+\beta$	0.033	0.008	0.888
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.167	0.037	0.994
Spectrin repeat-like	α	0.086	0.019	0.992
Hemerythrin-type up-and-down 4-helical bundle	α	0.056	0.012	0.991
Immunoglobulin/albumin-binding domain-like	α	0.041	0.009	0.991
Ferredoxin-like	$\alpha+\beta$	0.036	0.008	0.956
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.190	0.027	0.998
Spectrin repeat-like	α	0.095	0.013	0.998
Hemerythrin-type up-and-down 4-helical bundle	α	0.069	0.010	0.998
Ferredoxin-like	$\alpha+\beta$	0.041	0.006	0.993
Immunoglobulin/albumin-binding domain-like	α	0.041	0.006	0.996

Table S6: **Top SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k 4000.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.108	0.036	0.855
Spectrin repeat-like	α	0.054	0.018	0.892
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.014	0.217
alpha-alpha superhelix	α	0.031	0.010	0.448
Hemerythrin-type up-and-down 4-helical bundle	α	0.031	0.010	0.896
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.123	0.037	0.904
Spectrin repeat-like	α	0.065	0.019	0.939
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.039	0.012	0.377
Hemerythrin-type up-and-down 4-helical bundle	α	0.038	0.011	0.930
alpha-alpha superhelix	α	0.028	0.008	0.609
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.146	0.039	0.949
Spectrin repeat-like	α	0.071	0.019	0.974
Hemerythrin-type up-and-down 4-helical bundle	α	0.046	0.012	0.967
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.041	0.011	0.544
Ferredoxin-like	$\alpha+\beta$	0.031	0.008	0.812
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.161	0.038	0.981
Spectrin repeat-like	α	0.086	0.020	0.991
Hemerythrin-type up-and-down 4-helical bundle	α	0.054	0.013	0.982
Immunoglobulin/albumin-binding domain-like	α	0.039	0.009	0.983
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.035	0.008	0.699
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.183	0.039	0.997
Spectrin repeat-like	α	0.092	0.019	0.994
Hemerythrin-type up-and-down 4-helical bundle	α	0.062	0.013	0.998
Immunoglobulin/albumin-binding domain-like	α	0.038	0.008	0.995
Ferredoxin-like	$\alpha+\beta$	0.038	0.008	0.970
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.196	0.029	0.999
Spectrin repeat-like	α	0.097	0.014	1.000
Hemerythrin-type up-and-down 4-helical bundle	α	0.079	0.011	1.000
Ferredoxin-like	$\alpha+\beta$	0.040	0.006	0.998
Immunoglobulin/albumin-binding domain-like	α	0.038	0.005	1.000