Chapter 2

BASIC LANGUAGE MODELS ARE SKEWED MIRRORS OF THE PROTEIN UNIVERSE

2.1 Introduction

Protein language models (PLMs) — and more recently, their genomic language model (GLM) cousins — have become increasingly utilized as descriptors and generators of real and synthetic biological components (Ferruz et al., 2022; Hie et al., 2022; Hwang et al., 2024; Lin et al., 2023; Nguyen et al., 2024). That PLMs are consistent with natural protein-space as far as sequence statistics, structure statistics, and biochemical and biophysical properties is a critical prerequisite if we are to use and trust PLMs as vehicles to answer fundamental questions about the nature of protein-space, such as those posed in the preceding chapter. Likewise if PLMs are to prove dependable and formidable as engines for adding novelty in sequence, structure, or function to the protein universe (Ferruz and Höcker, 2022). In general, the ability of PLMs to implicitly internalize the relevant knowledge is assumed to follow from: (1) the sheer depth and volume of large training datasets such as UniRef50 ($\approx$ 50 million sequences), UniRef90 ($\approx$ 100M sequences), and UniRef 100 ($\approx$ 3 billion sequences); and (2) the application of well-benchmarked model architectures and unsupervised learning methods from natural language processing (NLP) (Chen et al., 2023; Suzek et al., 2007; Vaswani et al., 2017). This faith is broadly placed despite the fact that features such as tokenization scheme, vocabulary size, and loss functions, and hyperparameters including learning rate and masking fraction, are often ported directly from NLP work without adapting for the imperfect analogy between English and amino-acid "texts."[1] PLMs may indeed be learning and storing energy functions and co-evolutionary statistics deep within stacked and layered transformers, but does that manifest in the boundless synthetic protein "texts" — natural-like or novel — that can now be generated at the push of a button (Roney and Ovchinnikov, 2022; Zhang et al., 2024)?

Answering this question is complicated further by the fact that generating out of a

[1] Although it is beyond the scope of this thesis, intriguingly, protein "text" corpora may boast several advantages over human-created texts as far as language model training tractability, total-parameter scaling, and time to convergence. This points to an opportunity to systematically perturb training schemes and hyperparameter selection to craft bespoke PLMs with reduced compute overhead. For further discussion, see Frey et al. (2024).

PLM is not a single universal concept, but rather covers a multitude of approaches permitted by model architecture, application-specific factors, and personal philosophy. The idea and mechanism of generating a single sequence is intuitive with an autoregressive decoder-only model such as ProtGPT2 or ProGen — start with a blank slate, add one token (a single amino-acid or a short subsequence, depending on the model in question) at a time, moving from left-to-right, conditioned on whatever has come before (Ferruz et al., 2022; Madani et al., 2023). It is less straightforward for a model trained with a masked language modeling (MLM) objective and/or lacking a decoder. In such instances, obtaining a single sequence might be done via Gibbs sampling — filling in one sequence position at a time in a Markov Chain Monte Carlo (MCMC) process, left-to-right or randomly, often with additional model fine-tuning and/or a natural seed sequence (Garcia et al., 2024; Johnson et al., 2021). Another common choice, beam search, incorporates heuristics that land somewhere between greedy decoding and fully probabilistic sampling, while more indirect schemes might incorporate PLM-derived metrics (e.g. sequence likelihood, predicted 3D contact maps) into an external energy function facilitating MCMC search on a traditional sequence landscape in single amino-acid mutational steps (Elnaggar et al., 2022; Verkuil et al., 2022). Still others have trained supplemental decoders of ESM2 embeddings to map high-dimensional latent space representations back to amino-acid sequences in tailored use cases (Chen et al., 2024).

Given the potential of PLMs to reach into novel sequence-spaces as highlighted in Chapter 1, and the proliferation of PLMs and associated sampling strategies in the absence of detailed analysis of generative output (computational or experimental), we perform the first at-scale *in silico* statistical characterization of sequence and structure composition in PLM-generated amino-acid sequences. We demonstrate that not all models and sampling strategies are created equal. In particular, autoregressive sampling from ProtGPT2 dramatically outperforms Gibbs sampling from ESM2 in proposing realistic protein structures and achieving structural diversity. Despite outperforming ESM2, however, the structural coverage of ProtGPT2 sharply distorts the distribution of natural protein structures. Further, we discover that while ProtGPT2 displays an impressive ability to sample and assemble novel sequence motifs, maximizing sequence novelty through hyperparameter tuning exacerbates its already substantial shortcomings as far as preserving structural breadth. Together, our results identify a critical need for PLM-based generative strategies that accurately capture rare and novel protein features if we are to push the boundaries of fundamental biophysics and protein design.

2.2 Results & Discussion

Creating a fold-annotated database of PLM-generated protein structures

In order to characterize sequence and structure statistics, we initially constructed a database of AI-generated artificial protein sequences from a suite of representative models. We selected two commonly-used PLMs, both transformer-based but otherwise starkly contrasting in architecture and compatible sequence generation methods: (1) ProtGPT2, an autoregressive decoder-only model with 774M parameters; and (2) ESM2, a bidirectional encoder-only model with 150M parameters (Fig. 2.1) (Ferruz et al., 2022; Lin et al., 2023).[2] Sequence generation from ProtGPT2 involves stepwise addition of tokens drawn probabilistically from a SwissProt-extracted vocabulary of 50,256 short amino-acid subsequences, proceeding left-to-right while conditioning on the in-progress sequence to the left. Generation begins with a blank seed sequence and continues until either a prespecified number of tokens is reached or a STOP token is generated, whichever occurs first. For our database, we sampled 100,000 sequences with ProtGPT2, applying the default best-performing hyperparameters — sampling temperature 1, top_k 950, top_p 1.0, repetition penalty 1.2 — from the original study, and enforcing a stopping criterion after 40 tokens in the absence of a STOP token. Generated sequences were truncated to a maximum length of 100aa, and sequences containing rare or ambiguous amino acids (B=Asx, J=Ile/Leu, O=Pyl, U=Sec, X=Xaa, or Z=Glx) were filtered out, leaving 99,982 sequences for downstream analysis.

For ESM2-150M, we elected a left-to-right Gibbs sampling approach in single-token increments for ease of fair comparison to the autoregressive method and to align with existing benchmarks in the field (Johnson et al., 2021). In contrast to ProtGPT2 sampling, ESM2-150M uses the amino-acid alphabet (20 canonical AAs + 6 rare/ambiguous AAs) as its vocabulary and generates up until a fixed sequence length is reached.[3] We generated 148,500 sequences of length 100aa from ESM2-150M, with default hyperparameters of sampling temperature 1 and no repetition penalty, and applying the same filtering for rare/ambiguous amino acids as with ProtGPT2.

[2] ESM2 may refer to a family of associated language models of various transformer stack height and layer count, all trained on the 2021_09 release of UniRef50. Here, we use the 150M-parameter model to manage compute overhead on the generation task. The full ESM2 model collection includes versions with 8M, 35M, 150M, 650M, 3B, and 15B parameters.

[3] In theory, the ESM2 vocabulary also alows for an early STOP (end-of-sequence/<eos>) token to be generated, but we did not observe this in practice, and it is highly unlikely to occur given that the ESM2 models were trained without explicit <eos> tokens in training data clusters.

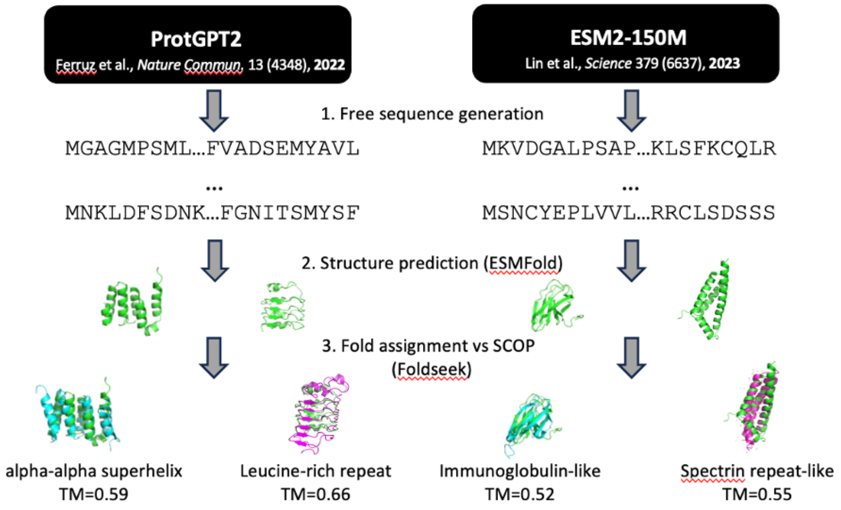


Figure 2.1: **Workflow for structural annotation of pLM-generated sequences.** Schematic overview of pLM generation, structure prediction, and structural search+assignment pipeline for representative models ProtGPT2 and ESM2-150M.

We also generated a control set of 74,250 random amino-acid sequences of fixed length 100aa, weighting sampling probability for each of the 20 canonical amino acids to be proportional to natural abundance in UniProt, i.e. preserving first-order sequence statistics but none of the second- and higher-order correlations between residues that transformers are expected to capture. Lastly, to benchmark against a completely different class of generative models, we added an inverse-folding comparison set comprised of 110,700 sequences, three per each of the 36,900 representative experimental structures in the **S**tructural **C**lassification **o**f **P**roteins (SCOP) database, designed from backbone-to-sequence inference by ESM-IF1 following default hyperparameters (Andreeva et al., 2020; Hsu et al., 2022). After filtering to exclude rare ligands in template structures and rare/ambiguous amino acids in outputs, the inverse-folding set was reduced to 104,591 sequences in total.

We predicted structures for all $\sim 430,000$ sequences with ESMFold. ESMFold has been shown to exceed the prediction accuracy of AlphaFold2 in the absence of deep multiple sequence alignment (MSA) information, which far-from-natural sequences lack by definition (Jumper et al., 2021; Lin et al., 2023). ESMFold's MSA-free single-sequence transformer architecture is additionally suitable for efficient inference in large-scale structure prediction tasks and for more transparent

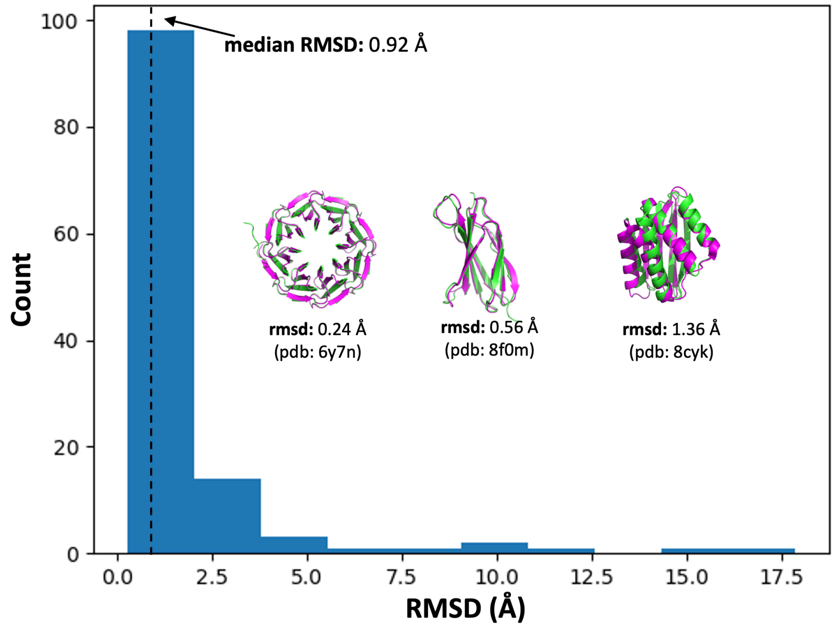


Figure 2.2: **ESMFold achieves single-angstrom structure prediction accuracy on *de novo* designed sequences.** Backbone atom root-mean-square deviation (RMSD; median= 0.92±0.14) for ESMFold predicted structures of $n = 122$ *de novo* designed proteins vs. experimental ground-truth structures, covering α, β, and mixed-$\alpha\beta$ global topologies, and including designs obtained from physics-based and generative AI models. All sequences in the validation set had experimental structures deposited in the Protein Data Bank (PDB) *after* the ESMFold training cutoff date of 05-01-2020.

analysis of model behavior. Bolstering this contention, we assessed the accuracy of ESMFold structural prediction on out-of-distribution samples by evaluating 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 the original ESMFold publication, we filtered out structures with resolution > 9 Å, length ≤ 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 and coverage of diverse structure topology classes, indicating sufficient generalization of

ESMFold beyond natural training data for use as a structure prediction oracle on PLM-generated sequences (Fig. 2.2).

Finally, to identify common natural structural motifs in PLM-generated, inverse-folded, and control variants, we annotated our database of predicted structures at the "fold" level of the SCOP classification, covering 1579 possible fold labels. Each predicted structure was assigned to a consensus fold label by performing a structure-based search against all SCOP representative PDB structures (n = 36900; the same structures used as backbone templates for inverse-folding) with Foldseek in accelerated TMalign mode and selecting the SCOP fold accounting for the most hits satisfying TM-score > 0.5 and max(query coverage, target coverage) > 0.8; in the absence of a hit satisfying these criteria, the predicted structure in question was labeled as un-assignable. The full generation, folding, and annotation workflow is summarized in Fig 2.1.

PLM-generated sequences are protein-like

For a first-pass analysis, we consider whether generated sequences and their corresponding (predicted) structures recapitulate the global characteristics of natural proteins. To determine where pLM-generated sequences lie with respect to natural sequence-space, we extract the ESM2-150M final hidden-layer internal representations ("embeddings") of all >400,000 generated sequences and 100,000 diverse natural sequences coding for SCOP fold examples mined from the AlphaFoldDB (Varadi et al., 2022).[4] We 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 ProtGPT2-generated sequences separate into two subpopulations, one co-localizing with natural sequences, and a second co-localizing with random sequences (Fig. 2.3A). In contrast, ESM2-150M-generated sequences co-localize most substantially with random sequences. Inverse-folded sequences from ESM-IF1 largely mirror the distribution of natural sequences, implying that they do not represent any significant departure from natural protein-space.

Turning towards coarse structural properties, the compactness/globularity of predicted structures for pLM-generated sequences — estimated as the fractional burial of all amino-acid surface area relative to the linear polypeptide chain — does not map onto whether generated variants are co-localizing with natural vs. random

[4] Except where otherwise specified, natural sequences/structures are drawn uniformly from a custom SCOP-UniRef50 database for which assembly details may be found in Section 2.4.

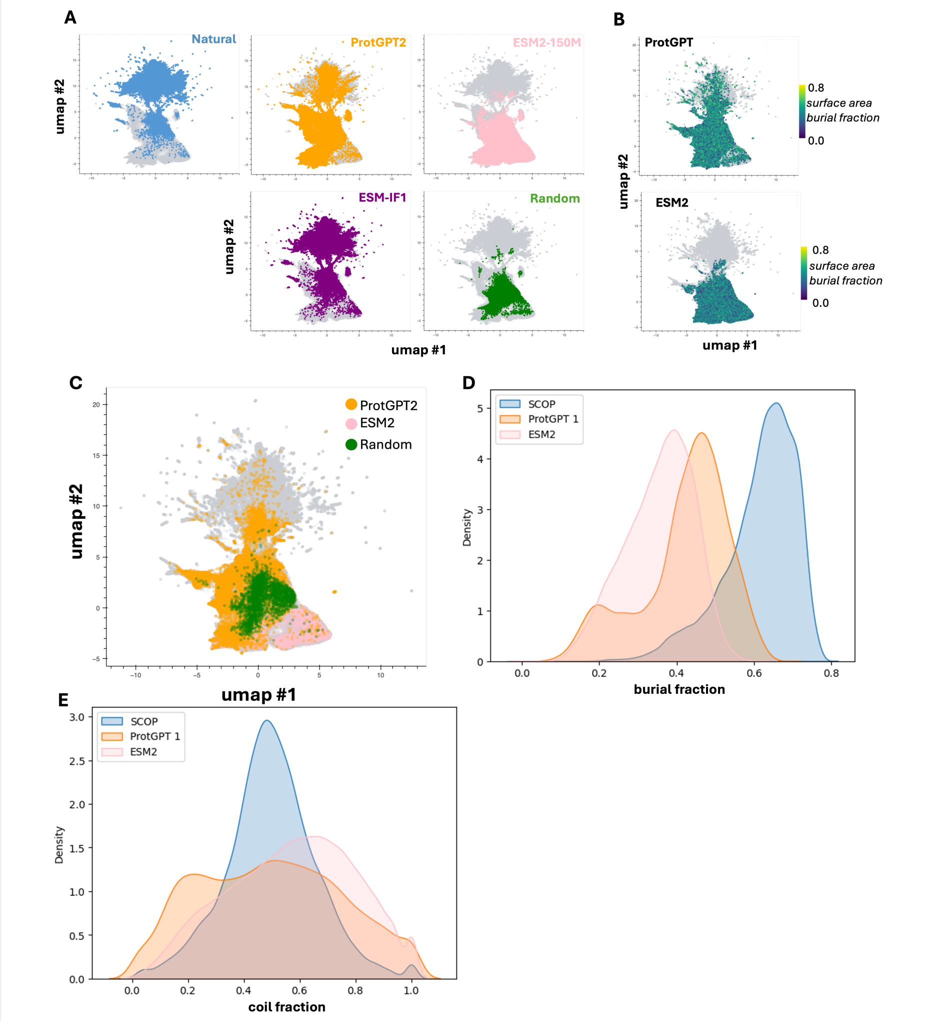


Figure 2.3: **PLM-generated sequences reflect the basic properties of compact, globular proteins.** (**A**) Dimension-reduced UMAP representation of ESM2-150M embeddings of natural, pLM-generated, inverse-folded, and random control sequences. (**B**) UMAP representation of pLM-generated sequences, colored by the fraction of amino-acid surface area buried (a measure of protein compactness). (**C**) UMAP representation of pLM-generated and random sequences assignable to a SCOP fold. (**D**) Fraction of amino-acid surface area buried for natural and pLM-generated sequences. (**E**) Fraction of residues annotated as random coils by DSSP for natural and pLM-generated sequences.

sequences (Fig. 2.3B). Similarly, SCOP folds are confidently assigned for large swaths of ProtGPT2-generated and ESM2-150M-generated sequences that do not co-localize with natural proteins (Fig. 2.3C). This suggests that both ProtGPT2 and ESM2-150M can emit sequences that are distinct in some statistical sense from natural ones yet able to fold into plausible and familiar 3D structures. However, this finding is tempered slightly by the realization that sequences from both PLMs are predicted to adopt structures that are less compact and less rich in secondary structure content (α-helix and β-sheet) on average than natural proteins in the SCOP reference set, implying that PLM output may tilt towards disordered regions (Fig. 2.3D-E).

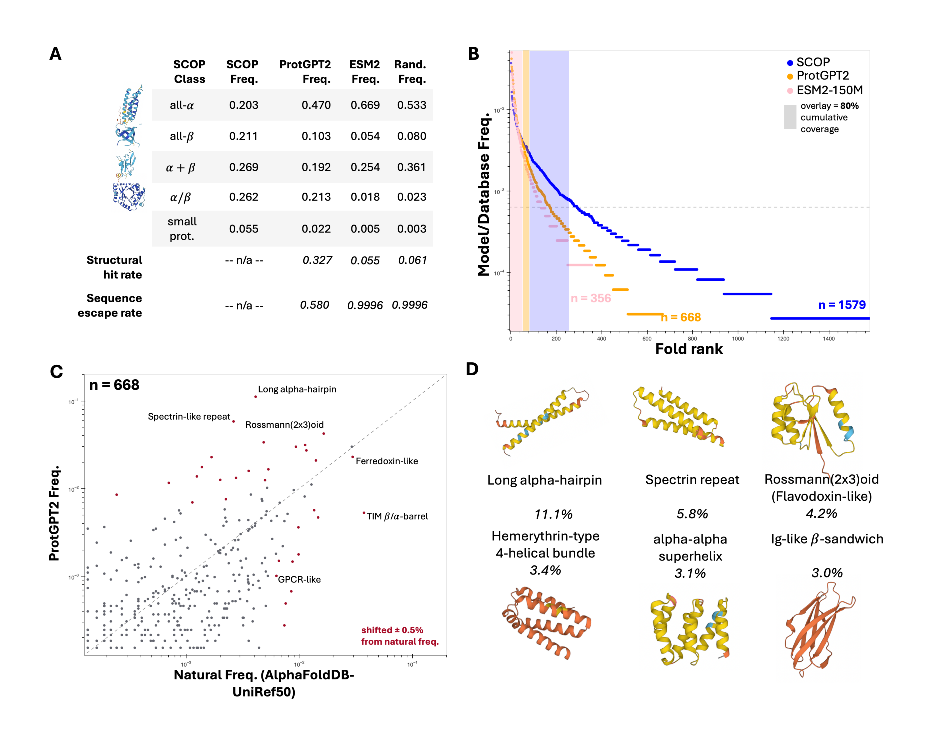
PLM-generated structures do not follow the natural distribution

For a finer-grained perspective on structure, we look to the SCOP fold label assignment procedure and observe that while a respectable 32.7% of ProtGPT2-generated sequences are assignable to a fold label, this is only the case for 5.5% of ESM2-150M-generated sequences, on par with the 6.1% assignment rate for random sequences (Fig. 2.4A). That the ESM2-150M fold assignment rate is no improvement over a control approach that includes first-order sequence statistics sparks doubt as to whether Gibbs sampling can reflect the higher-order sequence correlations presumably learned by ESM2-150M without manipulating the generation task to mimic the increased availability of contextual information during the training task.

Fold label assignments for both ESM2-150M and random sequences also skew heavily towards all-α topologies like helical bundles and $\alpha + \beta$ topologies like ferredoxins (Fig. 2.4A). SCOP topology class coverage with ProtGPT2 bears more resemblance to the natural distribution, especially as far as reaching the α/β folds that include most enzymatic diversity, but still overweights all-α content (Fig. 2.4A) (Choi and Kim, 2006). These trends in structural coverage breadth propagate to the fold level; 668/1579 (42.3%) of SCOP folds are detected in ProtGPT2 output, or $\sim$ 1.9x the 356/1579 (22.5%) represented in ESM2-150M output (Fig. 2.4B). Focusing on ProtGPT2, overrepresented folds include several flavors of α-helical bundles, Rossmann(2x3)oids, and the all-β immunoglobulin-like domain, while underrepresented folds include ubiquitous and diversified functional folds such as TIM β/α barrels, G-protein coupled receptors (GPCRs), and ferredoxins (Fig. 2.4C-D, Table S2.2). Evidently, plucked off the shelf, PLMs do not reproduce the natural frequencies of known protein folds.

Prioritizing sequence novelty shrinks accessible structure-space

While the structural ensembles sampled by PLMs fail to cover the breadth of natural structural-space and distort frequencies in the corners that they do touch, the plausible structures that they *do* access come with notable sequence novelty. In



	SCOP Class	SCOP Freq.	ProtGPT2 Freq.	ESM2 Freq.	Rand. Freq.
	all-α	0.203	0.470	0.669	0.533
	all-β	0.211	0.103	0.054	0.080
	$\alpha+\beta$	0.269	0.192	0.254	0.361
	α/β	0.262	0.213	0.018	0.023
	small prot.	0.055	0.022	0.005	0.003
Structural hit rate		-- n/a --	*0.327*	*0.055*	*0.061*
Sequence escape rate		-- n/a --	*0.580*	*0.9996*	*0.9996*

Figure 2.4: **Structural ensembles generated by pretrained language models cover natural protein-space imperfectly.** (**A**) Comparison of global protein topology preferences of natural, pLM-generated, and random sequences. (**B**)) Rank-ordered fold ensemble frequency plots for natural and pLM-generated sequences. (**C**) Fold ensemble comparison of ProtGPT2-generated sequences vs natural (SCOP-UnitRef50) sequences. (**D**) The six most-common SCOP folds among ProtGPT2 outputs; representative structures are of far-from-natural sequences (no MMseqs2 hit with E-value < 0.01).

particular, out of the 32,694/99,982 (32.7%) ProtGPT2-generated sequences with a fold label assignment, a further 18,962 (58.0% of assignable; 19.0% of all) have *no detectable homology* to any of the ~ 50 million representative protein sequences in UniRef50, a phenomenon that we dub sequence "escape" (Fig. 2.4A, Table 2.1). One hypothesis, inspired by typical NLP approaches, is that higher rates of sequence escape, and perhaps some of the missing structure coverage, might be reached by loosening sampling hyperparameters to encourage diversity in generated text. Continuing with ProtGPT2, the two critical and tunable hyperparameters are top_k and sampling temperature — increasing top_k allows for more tokens to be considered for sampling at a given step, while increasing temperature flattens the probability distribution over the token pool under consideration — both leading in theory to greater diversity in sequence output.

Table 2.1: **Base ProtGPT2 sequence and structure generation performance depends on sampling hyperparameters.** As sampling temperature and vocabulary size increase, generated sequences are more likely to lack homology to natural proteins, but also more likely to be unstructured and/or unassignable to any categorized SCOP fold label.

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

We systematically vary both temperature ($T = 0.8, 1.0, 1.2, 1.5, 2.5, 5.0$) and top_k ($N_k = 600, 950, 1500, 2400, 4000$), generating 100,000 sequences from ProtGPT2 for each of the 30 hyperparameter pairs on this grid and following the same truncation, filtering, structure prediction, and annotation workflow described previously.

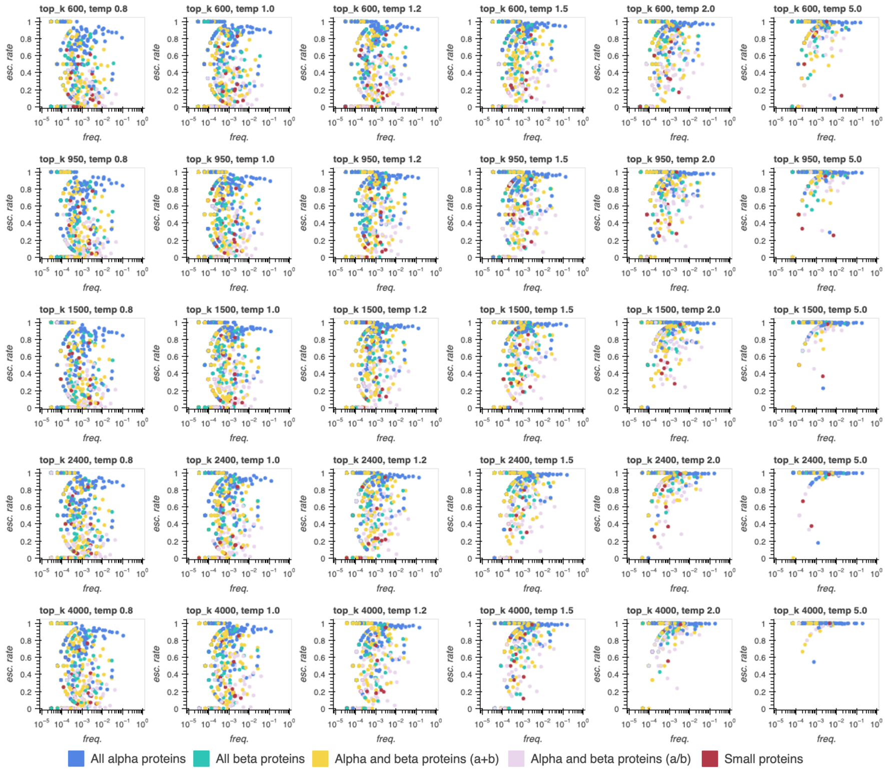


Figure 2.5: **Sequence escape rates increase across most folds as sampling temperature increases, at the cost of a shift towards all-α topologies.** Sequence escape rates for all assigned SCOP folds generated from ProtGPT2 within batches of 100k sequences for several sampling temperatures (0.8, 1, 1.2, 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.

Consistent with the NLP hypothesis, we see that sequence escape rates increase dramatically when temperature or top_k is increased, and approach 100% of assignable structures when both are increased simultaneously; this trend holds in aggregate and at the level of individual fold classes (Table 2.1, Fig. 2.5). However, far from rescuing the missing structural breadth, boosting sequence novelty exacerbates the issue. As temperature and/or top_k are increased, the number of unique SCOP folds detected plummets, the fraction of assignable structures (the "structural hit rate") falls precipitously, and topology class representation vanishes in favor of all-α helical bundles, largely at the expense of α/β proteins (Table 2.1, Fig. S2.1-S2.2). Again, these trends propagate down to individual fold classes, with a handful of helical bundles dominating the generative space, albeit with impressive sequence escape rates (Tables S2.1-S2.5). While obtaining far-from-natural versions of helical bundles could yet prove useful for protein design writ large (e.g. in minibinder design campaigns), the structural biases accentuated by by prioritizing sequence novelty reinforce the reality that without additional tuning or optimization, pre-trained PLMs are at best flawed mirrors of natural protein-space thanks to severe structural dropout.

2.3 Conclusion

We showed that, after "seeing" tens of millions of real protein sequences, PLMs are sufficiently aware of the sequence and structure statistics of natural proteins to yield realistic proteins that pass the *in silico* biophysical smell test — compact, globular, containing familiar secondary elements, and often bearing a passing resemblance to known structural motifs. In the case of ProtGPT2, this capacity emerges seamlessly in a free generation task that echoes the model's training task. We also demonstrated that ProtGPT2 is a powerful instrument for accessing sequence novelty, specifically sequences devoid of measurable homology to natural proteins even under highly sensitive search conditions. However, this sequence novelty comes at a substantial cost. Namely, limited structural breadth in model output, sacrificing much of the richness of nature's structural landscape. This presents as a fundamental tradeoff. The more sequence novelty is pursued by tuning sampling hyperparameters to explore the vastness of sequence-space, the more complete the collapse to a small collection of structural modes, often biophysically simple α-helical bundles.

In contrast to situations encountered in foundational ML subfields including natural language processing and computer vision, this collapse to a subset of modes occurs *without* obvious training data contamination and only weakly reflects the relative frequencies of these modes in the UniRef50 training data common to both ProtGPT2 and the ESM2 model family.[5] Put in plainer biological terms, while the long alpha-hairpin and the spectrin repeat come to dominate model output, it's the TIM β/α barrels (or stable subsectors thereof) and ferredoxins that ought to carry the day if natural abundance were the guiding factor. Instead, this behavior may well stem from a combination of limitations baked into model architecture (e.g. the unidirectional context window of ProtGPT2) and mechanistic discordance between training and generation tasks (e.g. 15% vs 100% sequence masking in training vs. generation

[5] Although ProtGPT2 and ESM2 were trained on different versions of UniRef50 (ProtGPT2: 2021_04, ESM2: 2021_09), with distinct train-test partitioning approaches, it is unlikely that this would translate into any significant difference in database composition or contamination.

contexts respectively for ESM2). The absence of many rare and/or functionally relevant structural motifs from generative PLM output could prove deleterious for future AI-driven protein design campaigns. Further, this shortcoming suggests that alternative forcing strategies will be required for harnessing sequence novelty and reliably sampling functional protein populations from PLMs, particularly in the pursuit of "linguistically consistent" proteins well beyond the confines of natural sequence-space. Tailored strategies for achieving these goals are explored in the subsequent chapters.

2.4 Methods

Except where otherwise specified, all model access and interfacing was via TRILL v1.3.11 (Martinez et al., 2023).

Sequence Generation from Protein Language Models

For the model comparison experiment, sequences (n = 100000) were sampled from ProtGPT2 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. The termination condition was set following the 40th token or the first STOP token occurring prior to the 40th token; sequences longer than 100aa were truncated to 100aa as the maximal length. Sequences containing rare or ambiguous amino acids (B, J, O, U, X, or Z) were filtered out as invalid, leaving 99,982 sequences. Sequences were sampled from ESM2-150M (n = 148500), from L-to-R with next-token prediction with Gibbs sampling, with a default sampling temperature of 1, no repetition penalty, and allowing for sampling from the full token distribution. The termination condition was set following the 100th amino-acid or the first STOP token occurring prior to the 100th amino-acid. Truncation and filtering were applied as for ProtGPT2.

For the hyperparameter scan experiment, sequences (n = 100000 per configuration) were generated from ProtGPT2 by L-to-R next-token prediction with top_p 1.0 and repetition penalty 1.2 fixed, and a grid search over 30 (temperature, top_k) pairs derived from six possible temperatures ($T = 0.8, 1.0, 1.2, 1.5, 2.0, 5.0$) x five possible top_k pool sizes ($N_k = 600, 950, 1500, 2400, 4000$). Truncation and filtering were applied as in the model comparison experiment.

Sequence Generation from Control Models

The random-sequence control set was generated by position-independent sampling of $n = 74250$ sequences of length 100aa from the 20 proteinogenic amino acids, with sampling probability for each amino acid proportional to its natural abundance. As sequence length was fixed and the rare/ambiguous amino acids B, J, O, U, X, and Z excluded, no filtering or truncation steps were required.

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 $n = 110700$ sequences in total. Pre- and post-processing for rare ligands in templates and rare/ambiguous amino acids in outputs, respectively, reduced inverse-folding output to 104,591 sequences. Default hyperparameters for sampling were taken as in Hsu et al. (2022).

Structure Prediction and Assignment

All structures (for filtered, truncated sequences as described above) were predicted with default ESMFold inference parameters as in Lin et al. (2023). For the model comparison experiment, structures were singly-inferenced (batch size 1), with compute resource collaboration with Yurts AI (now Legion Intelligence). For the hyperparameter scan experiment, structures were batch-inferenced with batch size 100 to optimally utilize memory allocation on A100-80GB GPUs, with compute resource collaboration through Oracle Cloud Infrastructure (OCI).

Predicted structures were annotated to SCOP fold labels via FOLDSEEK structure-based search against the custom SCOP-UniRef50 database (construction described in a standalone subsection) running in accelerated TMalign mode. The consensus SCOP fold was defined as the fold accounting for the most hits with TMscore > 0.5 and max(query_coverage, target_coverage) > 0.8.

Sensitive Sequence Search and Novelty Characterization

In both the model comparison and hyperparameter scan experiments, PLM-generated and control sequences were searched against UniRef50 using MMSEQS2 with default easy-search parameters and maximum e-value 0.01. Sequence escape rate was computed as the fraction of sequences not returning an alignment hit of any length to any cluster representative from UniRef50 at the specified e-value threshold.

Construction of the SCOP-UniRef50 Sequence-Structure Database

The SCOP-UniRef50 custom sequence-structure fragment database was constructed by performing reciprocal Foldseek searches (in fast TM-align mode) of the SCOP database of superfamily representative PDB structures ($n = 36900$) against the UniRef50 portion (based on the 2021_04 release) included in the July 2022 update to the AlphaFoldDB as first reported in Varadi et al. (2022) and made available as a precompiled Foldseek database in van Kempen et al. (2023), filtering for reciprocal hits with fractional query and target coverage > 0.8 and TMscore> 0.5, and clustering the filtered fragments at 100% identity.

For the model comparison experiments, $n = 100000$ natural sequences were uniformly sampled from SCOP-UniRef50 and jointly embedded along with PLM-generated and control sequences using ESM2-150M. This choice was made vs. sampling directly from SCOP in order to (1) obtain a similar number of natural sequences ($\sim 10^5$) to model-generated and control batches, and (2) draw sequence fragments with representative taxonomic coverage for evolutionarily conserved folds, as opposed to the narrower taxonomic coverage in SCOP, itself a function of skewed taxonomic coverage in the Protein Data Bank (Andreeva et al., 2020).

Basic Chemical Property Calculations

Amino-acid surface area burial fraction was calculated using custom code and reference individual amino-acid surface areas (HMS Bionumbers: 103239). Secondary structure annotations were assigned with DSSP via the corresponding PyMOL v3.1.0 wrapper.

2.5 Supplemental Material

Supplemental Figures

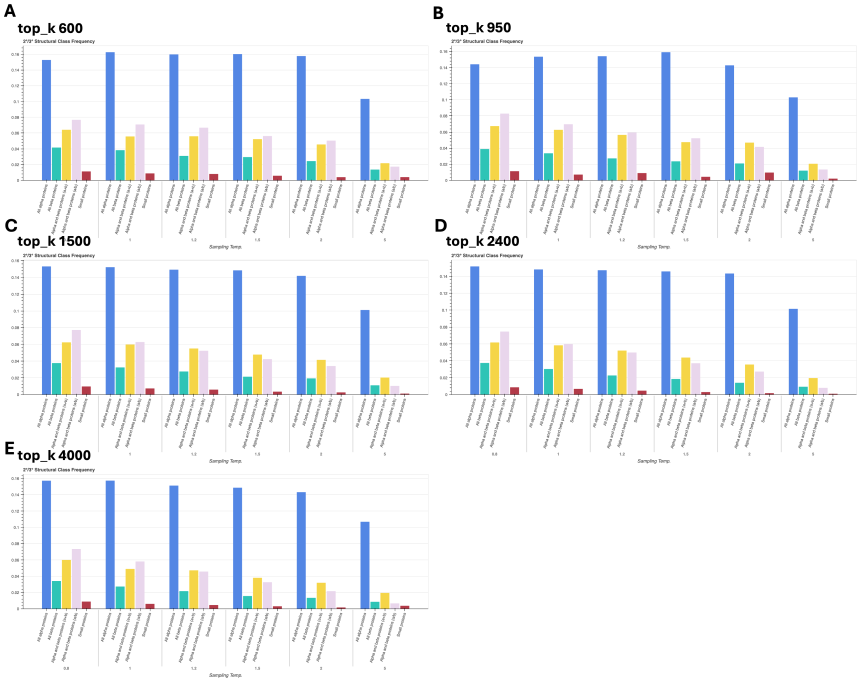


Figure S2.1: **Structure hit rates from base ProtGPT2 decrease as sampling temperature and top_k increase.** Structure hit rates from batches of 100k sequences generated from ProtGPT2 for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) and top_k values (number of highest-probability tokens considered 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").

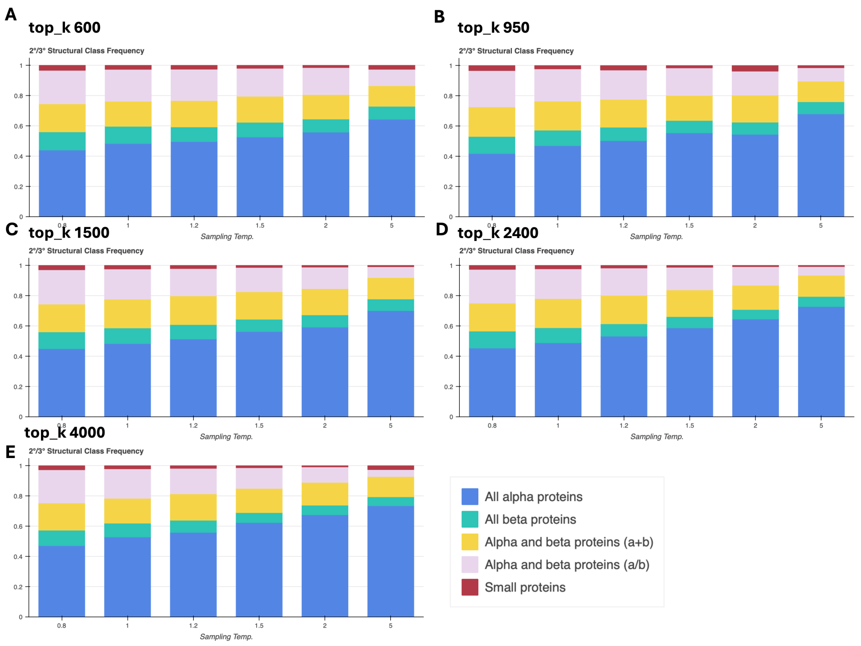


Figure S2.2: **Generated fold distributions shift towards all-α proteins and away from α/β proteins as sampling temperature increases.** Frequency of each protein global topology class (α, β, $\alpha + \beta$, α/β, or "small / minimal 2° structure") among all structure hits within batches of 100k sequences 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.

Supplemental Tables

Table S2.1: **Most common 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 S2.2: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 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 S2.3: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 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 S2.4: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 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 S2.5: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 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