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

Arjuna M. Subramanian[1], Zachary A. Martinez[1], Alec L. Lourenço[1], Shichen Liu[1], Matt Thomson[1,*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA

[*]Corresponding author. Email: mthomson@caltech.edu

Abstract

The combinatorial scale of amino-acid sequence-space has traditionally precluded substantive study of the full protein sequence-structure map. It remains unknown, for instance, how much of the vast uncharted landscape of far-from-natural sequences encodes the familiar ensemble of natural folds in a fashion consistent with the laws of biophysics but seemingly untouched by evolution on Earth. The scale of sequence perturbations required to access these spaces exceeds the reach of even gold-standard experimental approaches such as directed evolution. We surpass this limitation guided by the innate capacity of protein language models (pLMs) to explore sequences outside their natural training data through generation and self-feedback. We recast pLMs as probes that explore into regions of protein "deep space" that possess little-to-no detectable homology to natural examples, while enforcing core structural constraints, in a novel sequence design approach that we term "foldtuning." We build a library of foldtuned pLMs for >700 natural folds in the SCOP database, covering numerous high-priority

targets for synthetic biology, including GPCRs and small GTPases, composable cell-surface-receptor and DNA-binding domains, and small signaling/regulatory domains. Candidate proteins generated by foldtuned pLMs reflect distinctive new "rules of language" for sequence innovation beyond detectable homology to any known protein and sample subtle structural alterations in a manner reminiscent of natural structural evolution and diversification. Experimental validation of two markedly different fold targets; the tyrosine-kinase- and small-GTPase-regulating SH3 domain and the bacterial RNase inhibitor barstar demonstrates that foldtuning proposes protein variants that express and fold stably *in vitro* and function *in vivo*. Foldtuning reveals protein sequence-structure information at scale outside of the context of evolution and promises to push forward the redesign and reconstitution of novel-to-nature synthetic biological systems for applications in health and catalysis.

Nature has likely sampled only a fraction of all protein sequences and structures allowed by the laws of biophysics (*1*). The 20 proteinogenic amino acids ensure a combinatorially vast sequence-space; to roughly comprehend this magnitude, consider that making one copy of each of the $\sim 10^{78}$ possible sequences for a small protein domain of length 60 would require more matter than exists in the visible universe. High-quality sequence databases, in contrast, contain $\sim 10^{9}$ unique protein sequences distributed across the tree of life (*2, 3*). The observed protein catalog likely reflects selection for factors such as favorable folding kinetics, cofactor usage, and binding/catalytic functions (*4–8*). However, these proteins, no matter how evolutionarily fit, are not the only solutions of the sequence-to-structure mapping problem. Hydrophobic/polar patterning schemes distinguish energetically-favorable three-dimensional structures and generate stable α-helical bundle proteins encoded by novel sequences (*9–14*). Deep multiple sequence alignments (MSAs) capture sparse co-evolutionary signals sufficient to generate artificial proteins with comparable stability to natural examples (*15, 16*). And measurements on random sequence libraries suggest that as many as 1-in-10^{11} amino-acid sequences may code for functional proteins, providing ample "sparks" for alternate protein populations beyond nature (*17, 18*). Systematically locating stable, functional proteins that reconstitute known structural motifs but lie in regions of sequence-space with no meaningful similarity to nature promises to unlock expanded repertoires of binding partners,

signaling interactions, and substrate scopes for synthetic biology, while revealing key amino-acid sequence rules and constraints undergirding the fundamental biophysics of molecular machines.

We posit that the problem of mining such "döppelganger" proteins can be met by a search strategy that balances large perturbations to sequence against small perturbations to backbone structure. Global sequence perturbations of this magnitude are not accessible to directed evolution – which searches sequence-space locally under strong stability and fitness restrictions – or to machine learning models trained on high-throughput but inescapably local fitness data collected in deep-mutational scanning (DMS) experiments (*19–21*). Inverse-folding structure-to-sequence design methods can diversify sequence more substantially, but enforce strict backbone constraints that preclude the sorts of small structural innovations and ornamentations that have conferred new and/or expanded functionalities throughout natural evolution (*22–25*). In contrast, protein language models (pLMs) explicitly learn sequence-level amino-acid dependencies, implicitly internalizing the information flow from sequence to structure (*26–28*). Furthermore, when used as protein *generators*, pLMs reach beyond natural sequences and structures (*27, 29, 30*). Given that pLMs understand the core determinants of sequence-to-structure mapping while retaining an innate explorative capacity, we introduce "foldtuning" as an approach that transforms pLMs into probes that trace structure-preserving paths through far-from-natural regions of protein sequence-space. Leveraging complementary features of pLMs, structure prediction models, and ultrafast structure-based search tools, foldtuning mines protein-space for examples that honor the sequence "grammar" of a target backbone fragment while exhibiting novel semantics (*26, 29, 31–35*). We successfully apply foldtuning to > 700 structural motifs of interest from the SCOP and InterPro databases, covering all four major tertiary topology classes (all-α, all-β, $\alpha + \beta$, and α/β) and wide-ranging functional families, including GPCRs, transcription factors, cell-to-cell signaling domains, and various cytokines. We show that, generally, successive rounds of foldtuning progressively reduce similarity between pLM-generated and wild-type sequences, reaching new-to-nature 'rules of language' for constructing proteins and simultaneously making incremental structural changes such as loop expansion/minimization and symmetry adaptation. High-throughput screening of foldtuned sequence libraries for two target folds – the tyrosine-kinase- and small-GTPase-regulating SH3 domain, and the barstar ribonuclease inhibitor – via mass-spectrometry-based proteomics and survival assays, respectively, identifies stable, functional candidate variants with 0-40% sequence identity to

their closest respective neighbors in the known protein universe. Ultimately, sequence remodeling through foldtuning stretches the limits of how far common protein folds can be diversified at the sequence level while preserving structure and function.

Sequence exploration with 'soft' structure constraints

We initially consider whether pretrained pLMs might have the innate capacity to mimic natural structural elements with novel sequences off-the-shelf, without the need for additional training. We generated $\sim 10^6$ sequences of 100 amino-acids (100aa) in length from two commonly-used transformer-based pLMs, ProtGPT2 and ESM2-150M, assigning ESMFold-predicted structures, to the nearest TMscore > 0.5 match among the 1579 labeled protein folds in the SCOP database wherever possible (fig. S1A) (*27, 29, 36*). Structural coverage is imperfect, with high-confidence assignments to just 668 (42.3%) and 356 (22.5%) SCOP folds for ProtGPT2 and ESM2-150M respectively. As sampling temperature and token pool size (topk) are increased, large fractions of sequences emitted by ProtGPT2 exhibit no detectable homology to any representative sequence in the UniRef50 database (tab. S1, tab. S2). However, this increased sequence novelty comes at the cost of a dramatic rise in prevalence of structurally non-assignable sequences and difficulty recapitulating topologies other than all-α bundles (fig. S1B-D).

Consequently, we sought a more robust method to access far-from-natural sequences coding for structurally diverse fold classes. In this approach, which we term "foldtuning," a pLM is first finetuned on AlphaFoldDB-sourced natural sequences that adopt a target backbone structure (contained in a custom database we refer to as SCOP-UniRef50 (*37*)), followed by several rounds of finetuning on *self-generated batches of sequences* predicted to adopt the target fold while differing maximally from the natural training examples (Fig. 1). The maximal dissimilarity criterion is enforced by selecting for finetuning the subset of 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 (*32*). Each round of foldtuning can be thought of as a step along a trajectory that drives a pLM to access subpopulations of progressively further-from-natural artificial sequences while preserving the determinants of a fixed target structure (Fig. 1A).

Using ProtGPT2 as the base pretrained pLM, we foldtuned models for 727 structural targets; 708 SCOP folds (out of the top 850 ranked by natural abundance, for an 83.3% success rate), plus 19 cytokines and chemokines of interest curated from InterPro. Foldtuned versions of ProtGPT2 are effective at matching the target backbone structure, with a median *structural hit rate* of 0.565 after two rounds of updates on far-from-natural artificial sequences (Fig. 2A). Target backbones span numerous classes of functional interest, including transcription factor DNA-binding domains, GPCR/small GTPase signaling systems, modular cell surface receptor domains, and defense proteins (e.g. antimicrobial peptides, toxins) (Fig. 2A). Sequence novelty relative to natural examples also increases steadily with additional update rounds, as evidenced by increases in both semantic change (median: 46.9 after two rounds) and what we term the *sequence escape rate*, the fraction of target structure matches that feature no detectable sequence homology to any known protein in the UniRef50 dataset (median: 0.211 after four update rounds) (Fig. 2A-B). Notably, foldtuning of ProtGPT2 in general simultaneously maximizes structural hit rate and sequence escape rate, or only requires an insignificant decline in the former to maximize the latter (Fig. 2C). A high structural hit rate and a high sequence escape rate jointly indicate that a fold tolerates substantial sequence plasticity without major perturbation to structure; that is, the fold in question is highly *designable*, being encoded by many variable sequences. Taking the product of structural hit rate and sequence escape rate as a proxy for "designability," we find that the right-handed β-helix, ribbon-helix-helix domain, TIM β/α-barrel, pleckstrin homology domain, and α/α toroid are ranked as the most-designable SCOP motifs (Fig. 2D).

New sequence motifs and structural innovations

Given the succesful generalization of foldtuning to several hundred targets covering structural and functional families of significant relevance to synthetic biology, we inspected trends and characteristics in sequence and structure properties among foldtuning-generated proteins. Dimensionality-reduction by PCA and UMAP on ESM2-650M embeddings (averaged final-layer hidden states), taking generated GPCRs and immunoglobulin domains as representative examples as generated examples of interest, indicate that with each additional round, foldtuned versions of ProtGPT2 propose sequences that drift further and further from natural training examples in abstract feature-

space while preserving structural fidelity to the target (Fig. 3A,D). For GPCRs, foldtuning rapidly converges on generating sequences with no detectable homology against UniRef50, dropping from a median sequence identity of 0.250 and a median fractional aligning region length of 0.380 after the initial evotuning round on natural examples to the median sequence having no detectable homologous region of any length after the first round of foldtuning, and maintaining that trend over four rounds (Fig. 3B). Foldtuning relaxes sequence constraints more slowly for immunoglobulins, going from a median sequence identity of 0.336 and a median fractional aligning region length of 0.695 after initial evotuning to an identical median sequence identity mark but a clear drop in median fractional aligning region length to 0.531 after a full four rounds (Fig. 3E). It also bears noting that while the median foldtuned immunoglobulin sequence does not drop past 33.6% sequence identity despite extensive foldtuning, this mark still represents a leap in sequence novelty inaccessible to purely experimental approaches and equivalent to separation over vast evolutionary timescales. All-against-all deep sequence alignment of foldtuned variants (2703 GPCRs, 3035 immunoglobulins) and SCOP-UniRef50 entries (34,327 GPCRs, 150,258 immunoglobulins) for the corresponding fold reveals that at the sequence level, foldtuned variants largely self-cluster (min. seq. iden. threshold 30%) into distinct subpopulations that infill regions of sequence-space not sampled by nature (Fig. 3C,F). Foldtuning-infilled clusters are more tightly linked with prominent clusters of natural sequences for the immunoglobulin-like fold than for GPCRs, consistent with the relative degrees of sequence homology observed. We also considered whether beyond exploring new sequence semantics at a global level, foldtuning might be favoring different "vocabularies" in its preferences for short local subsequences. To characterize vocabulary trends among foldtuning-generated sequences, we conducted a simple *n*gram-based analysis of foldtuned variants in comparison to SCOP-UniRef50 examples, splitting all sequences into sliding windows of length 1-4 and calculating the usage frequencies of the 20, 400, 8000, and 16000 possible 1-grams, 2-grams, 3-grams, and 4-grams respectively. Across disparate fold families of interest, we observed noticeable "vocabulary shifts" where foldtuned models favored different subsets of 4-grams relative to natural counterparts, with, for example, statistically significant usage changes in anywhere from 3.4% 4-gram "words" for $\beta\beta\alpha$-zinc fingers to 5.3% for immunoglobulins, to 13.0% for both GPCRs and associated small GTPases, highlighting different levels of attainable sequence relaxation (Fig. 3H).

On the structural side, we consistently noticed that over one or more rounds of foldtuning, without any explicit structural direction aside from the TMscore-based *in silico* filtering and validation step, subsets of predicted structures would tweak the relevant formal fold description from SCOP, with ornamentations both subtle (e.g. shortening disorded loops) or more substantial (e.g. reversing strand connectivity or altering shape symmetry). For instance, foldtuning on the TIM α/β -barrel fold – shared by sequentially and functionally diverse enzyme families and common to as many as 10% of all non-disordered natural proteins – underwent rampant structural exploration in the course of attaining impressive structural validation (0.298 after evotuning to 0.770 after four rounds of foldtuning) and sequence escape rates (0.621 after evotuning to 0.995 after four rounds). All-against-all global structural alignment and clustering of 3094 foldtuned TIM barrel variants, removing variants that remained isolated after applying a min. TMscore cutoff of 0.7, separated 799 foldtuned TIM barrels into six distinct clusters of structural diversification (Fig. 3G). One cluster matched the familiar 8-fold symmetry common among wild-type examples, four others featured 9-fold, 10-fold (2 clusters, differing slightly in the manner of barrel closure), and 11-fold symmetries, with the final cluster adding a pair of non-terminal surface β-strands resembling a natural feature seen in predicted structures of cofactor-F420-utilizing bacterial redox proteins.

Experimental validation of foldtuned proteins

Emboldened by the ability of foldtuning to readily propose plausible far-from-natural protein sequences – as prefiltered computationally by structure prediction, search, and assignment – we sought to validate selected examples experimentally for expression and function. From a roster of small folds ($\leq$ 84aa) for which coding DNA oligo pools could be easily synthesized, we focused first on the SH3-like barrel (SCOP ID: 2000090). The SH3 domain is a notable mediator of protein-protein interactions and regulator of signal transduction, particularly in tyrosine kinase pathways. Engineered SH3 domains have historically been desirable in synthetic biology for roles in designed artificial protein recognition and signaling cascades, but their utility has been plagued by β-barrel design difficulty and a lack of orthogonality to natural SH3s (*38*). Applying the standard evotuning+four rounds of foldtuning to ProtGPT2 with SH3s as the target led to 2593 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 51.9% and

31.0% respectively. In contrast to, e.g. deep-mutational scanning libraries, protein sequences in foldtuned variant libraries have low pairwise sequence similarities and unique proteolytic digestion signatures, enabling direct high-throughput characterization of protein expression and properties by mass-spectrometry-based proteomics (Fig. 4A) (*39*). 1347/2593 (51.9%) foldtuned SH3 variants are detected as expressing in a reconstituted transcription-translation system by mass-spectrometric profiling (Fig. 4B). Using length-normalized signal as a proxy for abundance, expression intensity does not reflect sequence similarity to natural SH3s and is not noticeably shifted w.r.t. the number of foldtuning cycles performed (Fig. 4B). To rule out cases where high cell-free expression intensity might mask solubility and/or aggregation issues from poor folding stability we compared foldtuned protein recovery under native and denaturing purification conditions via multiplexed proteomics; variants without folding pathologies are expected to show equivalent or greater signal in the native fraction relative to the denatured one (Fig. 4A,C). Indeed, we observe a subpopulation of 87 foldtuned SH3s that are both highly abundant in the initial expression assay and shifted away from the denatured fraction in the second solubility/aggreggation assay (Fig. 4B). We reasoned that foldtuned SH3 variants with high expressability and relative folding stability might recognize the proline-rich peptide motifs found in the binding partners of natural SH3 domains. *In silico* screening with AlphaFold3 predicts that, indeed, foldtuned SH3 variants – including a candidate with no detectable sequence homology to known proteins – can bind either class I or class II proline-rich ligands in a hydrophobic aromatic-sidechain-rich cleft analogous to the wild-type interface (Fig. 4D).

As a second target for experimental characterization, we chose the barstar-like fold (SCOP ID: 2000624). With a single 3-stranded parallel β-sheet, the barstar-like fold is the simplest α/β unit and features a well-studied concerted folding pathway (*40*). It also offers an attractive opportunity for functional screening of a foldtuned variant library, as barstar's native function is to tightly bind and block the active site of barnase, a potent bacterial ribonuclease that is toxic in the absence of inhibition by barstar. We applied our standard foldtuning approach to barstar, yielding 1403 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 28.1% and 56.0% respectively. Variants were co-expressed with barnase from *B. amyloliquefaciens*; functional variants are expected to rescue host *E. coli* from the lethal effects of barnase expression in the absence of barstar (Fig. 4E). 11 foldtuned barstar variants were significantly enriched ($p < 0.05$)

relative to control under strong induction of barnase-barstar-variant co-expression, according to long-read sequencing of variant-coding amplicons, suggesting that they are sufficiently functional barstar mimics to mitigate the toxicity of barnase (Fig. 4F-G). Enrichment does not correlate with similarity to wild-type barstars, in fact, 7/11 of variants significantly enriched for barnase inhibition have no detectable homology to natural sequences (Fig. 4G). AlphaFold3 predictions of functionally enriched variants in complex with barnase further validate that four foldtuned variants are expected to bind barnase by inserting an α-helix into the barstar binding pocket (Fig. 4H). Comparison with a published experimental structure of the *B. amyloliquefaciens* barnase-barstar complex (pdb: 1BRS) reveals that most foldtuning-derived barstar mimics are presumed to adopt the canonical interaction interface with barnase, while some variants may explore alternate binding modes, as exemplified by variant #141 generated from the fourth-round ProtGPT2-Barstar model, a zero-homology variant that is rotated 90° relative to the wild-type (Fig. 4H).

Lastly, we note that foldtuning was recently applied outside of the scope of the 727 structural targets covered in this study to design far-from-natural mimics of insulin, a high-value translational target. Despite posing challenges for the foldtuning workflow due to the post-translational cleavage events required to transform inactive proinsulin into active insulin, ProtGPT2 was successfully foldtuned to generate single-chain insulin variants that specifically bind the insulin receptor while eliminating a conserved disulfide staple previously assumed to be obligatory for receptor-binding (*41*).

Conclusion

We introduced foldtuning as a solution to the challenges of finding and following sparks of viable protein signal in the gargantuan depths of amino-acid sequence-space. Through foldtuning, we constructed a library of ProtGPT2 derivatives for several hundred target folds of broad applicability in synthetic biology, and demonstrated that these models propose large perturbations to protein sequence while remaining within small perturbations of a target backbone structure. Sequences generated by foldtuned models are far-from-natural protein sequences that reflect wholly alien usage patterns in both local and global sequence contexts, often lacking any detectable similarity, even over a subfragment, to *any* of the > 60 million sequences cataloged in UniRef50. Experimen-

tally tested foldtuned protein variants evince characteristics of stable, well-folded, and functional artificial proteins at sufficient rates to guide future improvements to generative capacity and hint at unconventional binding and recognition modes. We envision that the foldtuning workflow will only grow in utility thanks to its modular nature, with potential modifications including replacement of the compute-intensive structural validation step by one or more scoring methods bespoke to the engineering problem at hand, end-to-end model updates based on experimental screening of foldtuned variants, and combinatorial diversification of protein domains and subdomains.

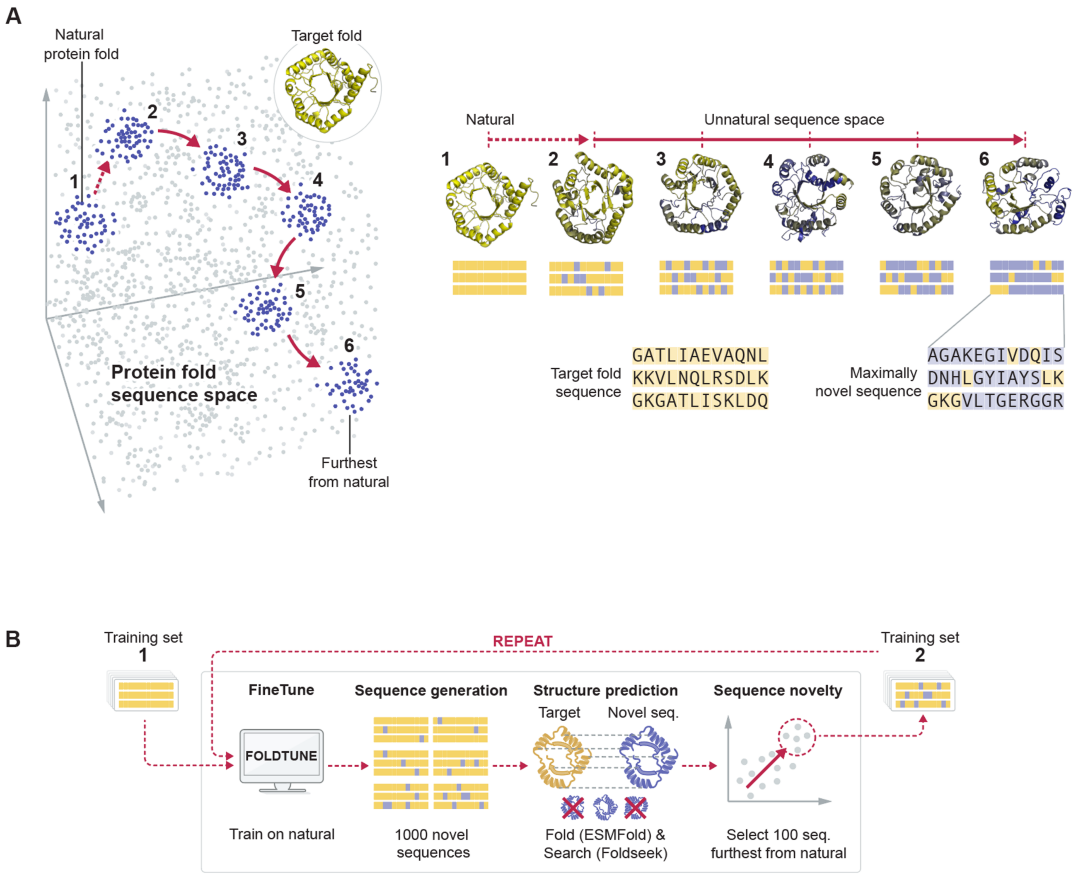

Figure 1: **Foldtuning explores far-from-natural sequences encoding alternate versions of natural protein structures.** (**A**) Conceptual overview of foldtuning. Beginning from natural protein sequences coding for a target backbone structure, foldtuning uses a protein language model (pLM)-based strategy to probe outwards in sequence-space, detecting subpopulations that maintain the target backbone while progressively decreasing sequence similarity to the closest natural example, passing below the detectable sequence homology threshold wherever possible. (**B**) Detailed schematic of the foldtuning workflow. For a provided backbone target fold, a pLM is initially fine-tuned (**1**) on 100 examples drawn from structural mining of the UniRef50 database (a representation of the natural protein universe clustered at 50% sequence similarity). In each subsequent round of foldtuning, a batch of artificial sequences are generated from the current pLM state and filtered for target backbone matching based on ESMFold structure prediction and Foldseek structure-based search (TMalign mode; tmscore threshold > 0.5); the pLM is then updated by finetuning on those filtered matches that maximize semantic change relative to the natural training examples (**2**).

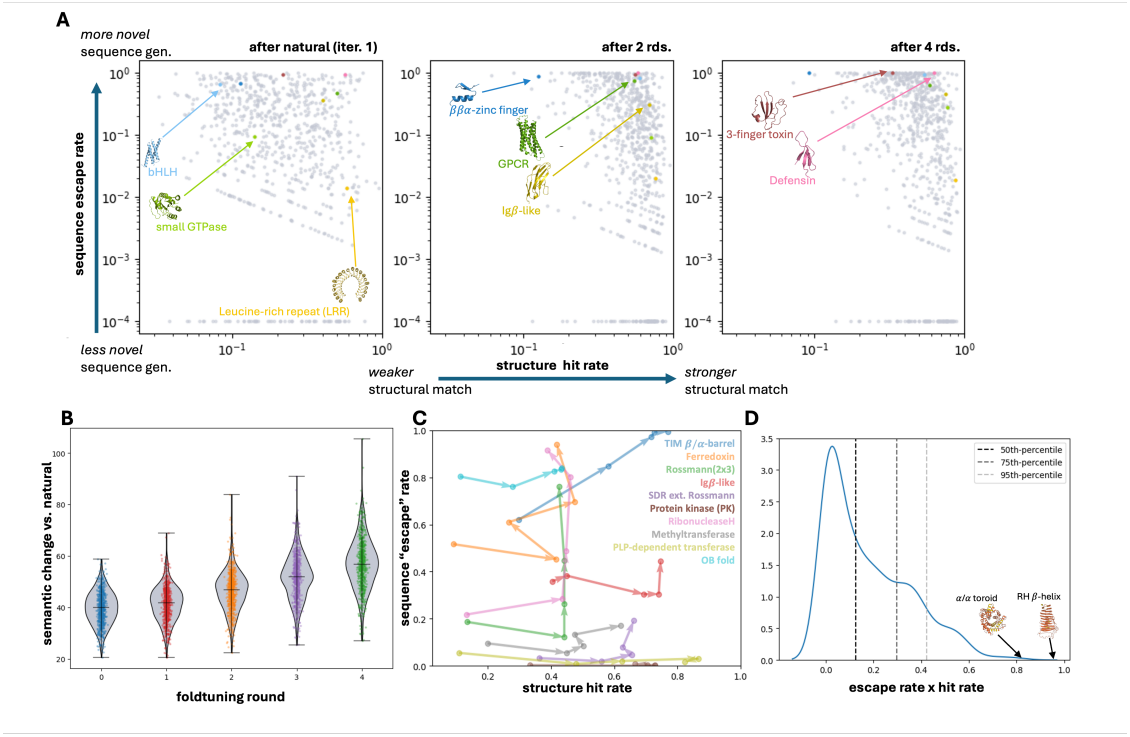

Figure 2: **Foldtuned models sample novel sequences for 727 diverse targets.** (**A**) Sequence escape vs. structural hit rates after natural-only evotuning or two or four rounds of foldtuning for 727 targets. Selected structural/functional targets are highlighted: transcription factors (blue), GPCRs/small GTPases (green), cell surface receptor domains (gold), and small antimicrobial/toxin proteins (red). (**B**) Semantic change, defined as the minimal L1-norm between the ESM2-650M embeddings of a generated sequence and any sequence in the natural training set, increases with additional rounds of foldtuning. (**C**) Over up to four rounds of foldtuning, structural hit and sequence escape rates are generally maximized simultaneously without explicit conditioning. (**D**) Target folds are ranked by sequence "designability", taking the product of structural hit and sequence escape rates as a proxy.

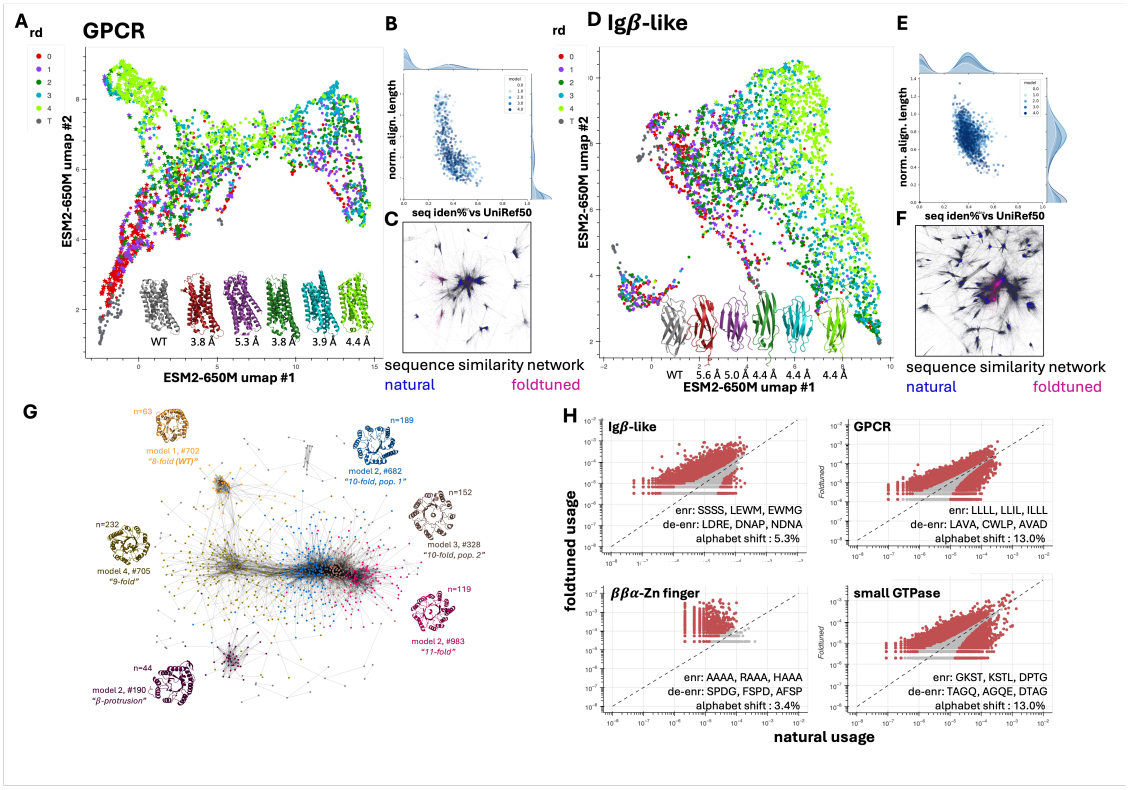

Figure 3: **Foldtuning accesses new sequence populations and structural innovations while 'fuzzily' preserving a target backbone.** (**A**) UMAP of round-by-round foldtuning sequence diversification captured by ESM2-650M final-layer hidden states with G-protein coupled receptors (GPCRs, SCOP ID: 2000339) as the target structure. (**B**) Sequence fragments aligning to UniRef50 cluster representatives decrease in length and identity with additional rounds of foldtuning with GPCRs as the target. (**C**) Network representation of similarity between natural (blue) and foldtuned (pink) GPCR sequences; edges connect pairs of sequences that align reciprocally with > 30% identity. (**D-F**) Same as (**A-C**), for Immunoglobulin-like domains (Igβ-like, SCOP ID: 2000051). (**G**) Network representation of structural similarity between foldtuned TIM barrel sequences; edges connect pairs of structures that globally align with TMscore> 0.7, node coloring reflects Louvain clustering assignments with representative cluster structures color-coded. (**H**) Usage patterns of amino-acid subsequences of length 4 ("4grams") in foldtuned sequences vs. natural sequences for Igβ-like domains (top left), GPCRs (top right), $\beta\beta\alpha$-zinc finger DNA-binding domains (SCOP ID: 2000684, bottom left), and small GTPases (SCOP ID: 2001260, bottom right); "alphabet shift " denotes the fraction of 4grams with a statistically significant usage shift (colored red, $p < 0.05$, Binyamini-Hochberg correction for correlated tests)

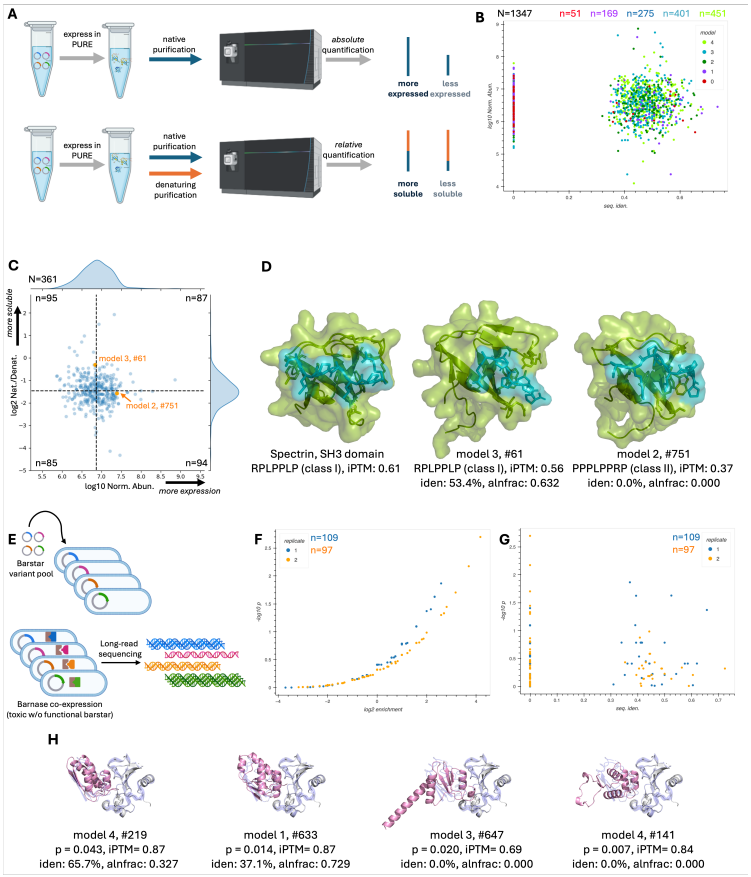

Figure 4: **Foldtuning-generated SH3 and barstar variants are expressable, stable, and functional.** (**A**) Schematic of mass-spec proteomics assays for variant library expression and folding stability. (**B**) SH3 expression assay signal intensity normalized by expected tryptic peptide count vs. sequence identity to closest natural hit in UniRef50 for foldtuning-generated variants (total $N = 1347$). (**C**) SH3 folding stability (measured by relative abundance ratio between natural and denaturing purification fractions) vs. expression assay results (as in (**B**)) for $N = 361$ variants. (**D**) AlphaFold3 predicted structures and iPTM scores for selected SH3 variants (green) bound to a classI/II proline-rich peptide (teal), compared to the wildtype *G. gallus* spectrin SH3 domain. (**E**) Schematic of barnase-inhibition survival assay for barstar variant library stability and function. (**F**) Survival assay p-value rank plot for barstar variants. For a given variant, enrichment is calculated as the ratio of amplicon sequencing reads with and without induction of co-expression of the lethal binding-partner barnase. (**G**) Survival assay p-values from (**F**) vs. barstar variant sequence identity to most-similar natural hit in UniRef50. (**H**) AlphaFold3 predicted structures and iPTM scores for selected barstar variants (pink) in complex with barnase (white). An experimental crystal structure of the wildtype barnase-barstar complex from *B. amyloliquefaciens* (pdb: 1BRS) is overlaid in blue.

References and Notes

1. J. Maynard Smith, Natural Selection and the Concept of a Protein Space. *Nature* **225** (5232), 563–564 (1970), publisher: Nature Publishing Group, doi:10.1038/225563a0, https://www.nature.com/articles/225563a0.

2. M. Mirdita, *et al.*, ColabFold: making protein folding accessible to all. *Nat Methods* **19** (6), 679–682 (2022), publisher: Nature Publishing Group, doi:10.1038/s41592-022-01488-1, https://www.nature.com/articles/s41592-022-01488-1.

3. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596** (7873), 583–589 (2021), number: 7873 Publisher: Nature Publishing Group, doi:10.1038/s41586-021-03819-2, https://www.nature.com/articles/s41586-021-03819-2.

4. D. Baker, A surprising simplicity to protein folding. *Nature* **405** (6782), 39–42 (2000), number: 6782 Publisher: Nature Publishing Group, doi:10.1038/35011000, https://www.nature.com/articles/35011000.

5. A. L. Watters, *et al.*, The Highly Cooperative Folding of Small Naturally Occurring Proteins Is Likely the Result of Natural Selection. *Cell* **128** (3), 613–624 (2007), publisher: Elsevier, doi:10.1016/j.cell.2006.12.042, https://www.cell.com/cell/abstract/S0092-8674(07)00117-1.

6. C. L. Dupont, A. Butcher, R. E. Valas, P. E. Bourne, 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 (2010), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.0912491107, https://www.pnas.org/doi/abs/10.1073/pnas.0912491107.

7. V. Alva, J. Söding, A. N. Lupas, A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **4**, e09410 (2015), publisher: eLife Sciences Publications, Ltd, doi:10.7554/eLife.09410, https://doi.org/10.7554/eLife.09410.

8. P. Vyas, *et al.*, Helicase-like functions in phosphate loop containing beta-alpha polypeptides. *Proceedings of the National Academy of Sciences* **118** (16), e2016131118 (2021), publisher:

Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2016131118, https://www.pnas.org/doi/abs/10.1073/pnas.2016131118.

9. K. Yue, K. A. Dill, Forces of tertiary structural organization in globular proteins. *Proceedings of the National Academy of Sciences* **92** (1), 146–150 (1995), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.92.1.146, https://www.pnas.org/doi/abs/10.1073/pnas.92.1.146.

10. E. Bornberg-Bauer, How are model protein structures distributed in sequence space? *Biophysical Journal* **73** (5), 2393–2403 (1997), doi:10.1016/S0006-3495(97)78268-7, https://linkinghub.elsevier.com/retrieve/pii/S0006349597782687.

11. H. Li, R. Helling, C. Tang, N. Wingreen, Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science* **273** (5275), 666–669 (1996), publisher: American Association for the Advancement of Science, doi:10.1126/science.273.5275.666, https://www.science.org/doi/abs/10.1126/science.273.5275.666.

12. R. Helling, *et al.*, The designability of protein structures. *Journal of Molecular Graphics and Modelling* **19** (1), 157–167 (2001), doi:10.1016/S1093-3263(00)00137-6, https://www.sciencedirect.com/science/article/pii/S1093326300001376.

13. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids. *Science* **262** (5140), 1680–1685 (1993), publisher: American Association for the Advancement of Science, doi:10.1126/science.8259512, https://www.science.org/doi/10.1126/science.8259512.

14. M. H. Hecht, *et al.*, De novo heme proteins from designed combinatorial libraries. *Protein Science* **6** (12), 2512–2524 (1997), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560061204, doi:10.1002/pro.5560061204, https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.5560061204.

15. M. Socolich, *et al.*, Evolutionary information for specifying a protein fold. *Nature* **437** (7058), 512–518 (2005), publisher: Nature Publishing Group, doi:10.1038/nature03991, https://www.nature.com/articles/nature03991.

16. S. W. Lockless, R. Ranganathan, Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **286** (5438), 295–299 (1999), publisher: American Association for the Advancement of Science, doi:10.1126/science.286.5438.295, https://www.science.org/doi/10.1126/science.286.5438.295.

17. A. D. Keefe, J. W. Szostak, Functional proteins from a random-sequence library. *Nature* **410** (6829), 715–718 (2001), number: 6829 Publisher: Nature Publishing Group, doi:10.1038/35070613, https://www.nature.com/articles/35070613.

18. C. L. Tong, K.-H. Lee, B. Seelig, *De novo* proteins from random sequences through *in vitro* evolution. *Current Opinion in Structural Biology* **68**, 129–134 (2021), doi:10.1016/j.sbi.2020.12.014, https://www.sciencedirect.com/science/article/pii/S0959440X21000026.

19. P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10** (12), 866–876 (2009), publisher: Nature Publishing Group, doi:10.1038/nrm2805, https://www.nature.com/articles/nrm2805.

20. A. E. Wilson, W. M. Kosater, D. A. Liberles, Evolutionary Processes and Biophysical Mechanisms: Revisiting Why Evolved Proteins Are Marginally Stable. *J Mol Evol* **88** (5), 415–417 (2020), doi:10.1007/s00239-020-09948-y, https://doi.org/10.1007/s00239-020-09948-y.

21. S. A. Fahlberg, C. R. Freschlin, P. Heinzelman, P. A. Romero, Neural network extrapolation to distant regions of the protein fitness landscape (2023), doi:10.1101/2023.11.08.566287, https://www.biorxiv.org/content/10.1101/2023.11.08.566287v1, pages: 2023.11.08.566287 Section: New Results.

22. C. Hsu, *et al.*, Learning inverse folding from millions of predicted structures, in *Proceedings of the 39th International Conference on Machine Learning* (PMLR) (2022), pp. 8946–8970, https://proceedings.mlr.press/v162/hsu22a.html, iSSN: 2640-3498.

23. J. Dauparas, *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378** (6615), 49–56 (2022), publisher: American Association for the Advancement of Science, doi:10.1126/science.add2187, https://www.science.org/doi/10.1126/science.add2187.

24. A. Tóth-Petróczy, D. S. Tawfik, The robustness and innovability of protein folds. *Current Opinion in Structural Biology* **26**, 131–138 (2014), doi:10.1016/j.sbi.2014.06.007, https://www.sciencedirect.com/science/article/pii/S0959440X14000724.

25. X. Pan, *et al.*, Expanding the space of protein geometries by computational design of de novo fold families. *Science* **369** (6507), 1132–1136 (2020), publisher: American Association for the Advancement of Science, doi:10.1126/science.abc0881, https://www.science.org/doi/full/10.1126/science.abc0881.

26. Z. Lin, *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379** (6637), 1123–1130 (2023), publisher: American Association for the Advancement of Science, doi:10.1126/science.ade2574, https://www.science.org/doi/10.1126/science.ade2574.

27. R. Verkuil, *et al.*, *Language models generalize beyond natural proteins*, Tech. rep., bioRxiv (2022), doi:10.1101/2022.12.21.521521, https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1, section: New Results Type: article.

28. Z. Zhang, *et al.*, Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences* **121** (45), e2406285121 (2024), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2406285121, https://www.pnas.org/doi/abs/10.1073/pnas.2406285121.

29. N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* **13** (1), 4348 (2022), number: 1 Publisher: Nature Publishing Group, doi:10.1038/s41467-022-32007-7, https://www.nature.com/articles/s41467-022-32007-7.

30. A. Madani, *et al.*, Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* pp. 1–8 (2023), publisher: Nature Publishing Group, doi:10.1038/s41587-022-01618-2, https://www.nature.com/articles/s41587-022-01618-2.

31. N. Chomsky, On certain formal properties of grammars. *Information and Control* **2** (2), 137–167 (1959), doi:10.1016/S0019-9958(59)90362-6, https://www.sciencedirect.com/science/article/pii/S0019995859903626.

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

33. M. van Kempen, *et al.*, Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* pp. 1–4 (2023), publisher: Nature Publishing Group, doi:10.1038/s41587-023-01773-0, https://www.nature.com/articles/s41587-023-01773-0.

34. I. Barrio-Hernandez, *et al.*, Clustering predicted structures at the scale of the known protein universe. *Nature* **622** (7983), 637–645 (2023), number: 7983 Publisher: Nature Publishing Group, doi:10.1038/s41586-023-06510-w, https://www.nature.com/articles/s41586-023-06510-w.

35. G. A. Pavlopoulos, *et al.*, Unraveling the functional dark matter through global metagenomics. *Nature* **622** (7983), 594–602 (2023), number: 7983 Publisher: Nature Publishing Group, doi:10.1038/s41586-023-06583-7, https://www.nature.com/articles/s41586-023-06583-7.

36. A. Andreeva, E. Kulesha, J. Gough, 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 (2020), doi:10.1093/nar/gkz1064, https://doi.org/10.1093/nar/gkz1064.

37. Materials and methods are available as supplementary material.

38. D. E. Kim, *et al.*, De novo design of small beta barrel proteins. *Proceedings of the National Academy of Sciences* **120** (11), e2207974120 (2023), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2207974120, https://www.pnas.org/doi/abs/10.1073/pnas.2207974120.

39. K. Tsuboyama, *et al.*, Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* pp. 1–11 (2023), publisher: Nature Publishing Group, doi:10.1038/s41586-023-06328-6, https://www.nature.com/articles/s41586-023-06328-6.

40. G. Schreiber, A. R. Fersht, Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles. *Journal of Molecular Biology* **248** (2), 478–486 (1995), doi:10.1016/S0022-2836(95)80064-6, https://www.sciencedirect.com/science/article/pii/S0022283695800646.

41. A. L. Lourenco, *et al.*, Protein CREATE enables closed-loop design of de novo synthetic protein binders (2025), doi:10.1101/2024.12.20.629847, https://www.biorxiv.org/content/10.1101/2024.12.20.629847v2, pages: 2024.12.20.629847 Section: New Results.

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

43. M. Varadi, *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50** (D1), D439–D444 (2022), doi:10.1093/nar/gkab1061, https://doi.org/10.1093/nar/gkab1061.

Acknowledgments

We thank Steve Mayo, Richard Murray, Carl Pabo, Erik Winfree, and Tsui-Fen Chou, as well as all members of the Thomson Lab for helpful discussions and feedback. We thank Ashwin Rakhra, John Ng, and their colleagues at the Oracle Corporation for generous cloud computing support.

Funding: This work was supported by the National Institutes of Health under award number R01-GM150125, the Gordon and Betty Moore Foundation, the Caltech Rosen Bioengineering Center, and the David and Lucille Packard Foundation under a Packard Fellowship to M.W.T.

Author contributions: A.M.S. and M.W.T. conceptualized and developed the foldtuning framework. A.M.S. performed computational experiments and trained the foldtuned model library. Z.A.M. carried out software development and benchmarking. A.M.S., A.L.L., and S.L. designed and performed wet-lab experiments. A.M.S. and M.W.T. wrote the manuscript with input from all authors.

Competing interests: There are no competing interests to declare.

Data and materials availability: A streamlined implementation of foldtuning is distributed in TRILL (v1.8.3 and later; https://pypi.org/project/trill-proteins/). Other inquiries should be directed to the corresponding author.

Materials and Methods

Except where otherwise specified, all model access and interfacing was via TRILL v1.3.11 (*42*).

Sequence Generation from Pretrained Base pLMs

Sequences ($n = 100,000$) were sampled from ProtGPT2 by L-to-R next-token prediction with the default best-performing hyperparameters from (*29*); 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, whichever came first; 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 for downstream analysis. Sequences were sampled from ESM2-150M ($n = 148,500$), from L-to-R 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, whichever came first. Filtering was applied as for ProtGPT2.

pLM Finetuning and Sampling

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. We constructed a custom sequence fragmment database 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 of the AlphaFoldDB, filtering for reciprocal hits with fractional query and target coverage > 0.8 and TMscore> 0.5, and clustering the filtered fragments at 100% identity (*43*). We refer to this database as SCOP-UniRef50. For the preliminary foldtuning round, given a target fold f, the base ProtGPT2 model was initially finetuned for 1-3 epochs on 100 natural sequences belonging to fold f, selected at random from SCOP-UniRef50. The number of epochs required was determined by a pre-screen in which ProtGPT2 was finetuned for 1-5 epochs, generating 100 sequences per epoch, predicting and assigning structures as described below, and finding the minimum epoch such that at $\geq 7\%$ of sequences were assigned to fold f (to ensure sufficient synthetic data to initiate foldtuning). For the separate set of chemokines, cytokines, and growth factors, database

construction was abbreviated, with query sequences taken directly from InterPro curation and clustered at 100% identity. In subsequent foldtuning rounds, finetuning was performed with the same optimizer parameters, for 1 epoch only, on the top-100 previous-round sequences assigned to fold f ranked by semantic-change as described in the main text and below.

Sampling from finetuned ProtGPT2 models followed the same general procedures, hyperparameters, and processing steps as for sampling from the base pretrained ProtGPT2 model, with the following differences: (1) In each road of foldtuning, 1000 sequences were generated from the appropriate finetuned model; (2) termination was after $0.4 \times M$ tokens, where M is the median length of SCOP-UniRef50 natural sequences for target fold f, or the first STOP token; (3) truncation was applied at Maa. Inference batch size on a single NVIDIA A100-80G GPU ranged from 125-500 sequences depending on target sequence length.

Structure Prediction and Assignment

All structures (for filtered, truncated sequences as described above) were predicted with default ESMFold inference parameters (*26*). Inference batch size on a single A100-80GB GPU ranged from 10-500 sequences per batch depending on target sequence length, in order to mantain optimal memory utilization.

Predicted structures and their corresponding sequences were assigned a fold label by searching against the SCOP database of superfamily representative PDB structures ($n = 36900$) with Foldseek in fast 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, *i.e.* the "consensus hit" (*33*). Sequence-structure pairs without a hit satisfying the TM-score and coverage constraints were deemed non-assignable and excluded from subsequent consideration or analysis.

Foldtuning Sequence Selection

For each target fold f and foldtuning round $k = 1, 2, \ldots N$, the semantic change was calculated for all generated sequences $\{s_k^{(i)}\}$ structurally assigned to fold f 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, and the "train" subscript denotes sequences selected from SCOP-UniRef50 for the initial foldtuning round. To construct the finetuning set

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

In Silico Evaluation of Foldtuned Models & Outputs

For a given foldtuned model with target fold f, structural hit rate was computed as the fraction of generated sequences with successful structure assignment to f. Sequence escape rate was computed as the fraction of *those sequences structurally assigned to the target* that do not return an alignment of any length to any cluster representative from UniRef50 in an MMSEQS2 search with default easy-search parameters and maximum e-value 0.01. The "designability" of a fold f was computed as the product of the corresponding structural hit and sequence escape rates.

Structural clustering analysis for a fold f was carried out by conducting an all-against-all structural alignment of successfully assigned variants with Foldseek in fast TM-align mode. Missing values (no alignment passing filters) were imputed as having a tmscore of 0. Results were represented as a graph with individual variants as nodes, and an edge joining any pair of nodes with a reciprocal average tmscore ¿ 0.7. Louvain clustering was performed with NETWORKX with default parameters. Isolated nodes were excluded from clustering and visualization. Sequence network analysis was carried out by separately preclustering foldtuned sequences and SCOP-UniRef50 sequence fragments assigned to fold f at 50% identity, using MMSEQS2 easy-cluster in covariance mode 1. Preclustered sequence sets were merged and searched all-against-all using MMSEQS2 easy-search with maximum e-value 0.01. Graph representations were constructed with preclustered sequences as nodes and edges joining pairs of nodes with reciprocal alignments of any length satisfying a minimum identity threshold of 30%. Isolated nodes were excluded from visualization. N-gram vocabulary analysis was carried out by splitting foldtuned sequences and SCOP-UniRef50 sequence fragments assigned to fold f into subsequences of length 4 and computing their respective frequency distributions and fold-change for foldtuned variants vs. SCOP-UniRef50 fragments. Significance testing for word frequency change was performed treating the SCOP-UniRef50 sequence subset for fold f as the null distribution (1000 samples drawn), setting a significance level $\alpha = 0.05$, and applying the Binyamini-Hochberg correction for correlated tests.

Oligo Pool Design and Preparation

Foldtuning-generated sequences selected for experimental characterization were truncated to remove disorded N- and C-terminal tail regions as predicted by ESMFold and identified in C_α contact maps computed with BIOTITE. Coding DNA sequences were designed by reverse translation with DNACHISEL, codon-optimizing for *E. coli*, with additional constraints on GC content (global $\geq 0.25, \leq 0.65$; never ≤ 0.19 or ≥ 0.71 over any subsequence of length 50) and homopolymers (restricted to < 14nt). Constant flanks – GACTACAAGGACGACGATGACAAG (5') and GGTTCCCACCATCATCACCATCAT (3') were added to code for a 5' FLAG tag and a 3' GSHHHHHH tag.

Oligo pools were ordered from Twist Biosciences as ssDNA fragments for sequences ≤ 300nt or as dsDNA fragments for sequences > 300bp and PCR-amplified with Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0494S) according to manufacturer instructions. T7RNA promoter, ribosome binding site, start codon, stop codon, and T7 terminator elements were added in a subsequent PCR-amplification step with the same reagents, and purified, concentrated, and resuspended in ultra-pure water using the Monarch Spin PCR & DNA Cleanup Kit (NEB, T1130S) according to manufacturer instructions.

In vitro Expression Measurements

Foldtuned variant pools were expressed *in vitro* with PURExpress (NEB, E6800) following the manufacturer's protocol, with 500ng template dsDNA per 50 µL reaction volume, incubating 18hrs at 29 °C. Expressed protein was purified under native conditions by His-tag pulldown using NEBExpress Ni Spin Columns (NEB, S1427L); 400 µL of eluate was washed and concentrated with Amicon Ultra Centrifugal Filters, 3 kDa MWCO (Millipore, UFC5003) 4x with 400 µL phosphate-buffered saline pH7.4, centrifuging at 14,000*g* for 30min per exchange, and 50 µL of concentrate recovered by reverse spin (1000*g* for 2min).

Concentrated purified protein samples were digested in an S-Trap micro spin column (Protifi, USA) according to the manufacturer's instructions and analyzed on Q-Exactive HF mass spectrometer coupled to EASY-nLC 1200. Peptides were separated on an Aurora UHPLC Column (25 cm × 75 µm, 1.7 µm C18, AUR3-25075C18-TS, Ion Opticks) with a flow rate of 0.35 µL/min for a

total duration of 1hr and ionized at 2.2 kV in the positive ion mode. Raw data files were searched against the Uniprot Escherichia coli proteome (UP000531813) and foldtuned variant sequences. Searches used the Proteome Discoverer 2.5 software based on the Sequest HT algorithm. Oxidation / +15.995 Da (M), deamidation / +0.984 Da (N), and acetylation / +42.011 Da(N-term) were set as dynamic modifications; carbamidomethylation / +57.021 Da (C) was set as fixed modification. The precursor mass tolerance was set to 10 ppm, whereas fragment mass tolerance was set to 0.05 Da. The maximum false peptide discovery rate was specified as 0.01 using the Percolator Node validated by q-value. Absolute abundance signal intensities were scaled by dividing by the expected peptide count from simulated tryptic digestion.

In vitro Folding Stability Measurements

Foldtuned variant pools were expressed, purified, washed, and concentrated as for the expression assay, as described above, with the modification that the reaction volume was split post-expression into 2 × 25 µL aliquots, one purified under native conditions and the other under denaturing conditions (6 M guanidinium chloride) following manufacturer instructions.

Concentrated purified protein samples were analyzed by Eclipse mass spectrometer coupled to Vanquish Neo. 1ug of peptides from S-trap based digestion with TPCK-treated trypsin were injected and separated on an Aurora UHPLC Column (25 cm × 75 µm, 1.7 µm C18, AUR3-25075C18-TS, Ion Opticks) with a flow rate of 0.35 µL/min for a total duration of 1 hour and ionized at 1.8 kV in the positive ion mode. Raw data files were searched against the Escherichia coli (strain B / BL21-DE3) proteome (UP000002032) foldtuned variant sequences using the Proteome Discoverer(PD) 2.5 software based on the SequestHT algorithm. Oxidation / +15.995 Da (M), Deamidated / +0.984 Da (N, Q), acetylation / +42.011 Da (protein N-term) and Met-loss / -131.040 Da (protein N-term, M) were set as dynamic modifications, and carbamidomethylation / +57.021 Da (C) was fixed modification. The precursor mass tolerance was set to 10 ppm, whereas fragment mass tolerance was set to 0.6 Da. The maximum false peptide discovery rate was specified as 0.01 using the Percolator Node validated by q-value. Enrichment was calculated as the abundance ratio of the natural channel relative to the denatured channel.

Barstar-Barnase Survival Assay

The barstar-like foldtuned variant pool was designed, ordered, and amplified to add regulatory elements as described above. Barstar variants were cloned as a single pool into barnase-barstar expression vector pMT416 (gift from Robert Hartley, Addgene plasmid #8607; http://n2t.net/addgene:8607; RRID:Addgene_8607), replacing the wild-type barstar-coding region, using NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621S) according to manufacturer's instructions. 1 µL of assembly product was transformed into 10 µL 5-alpha F'Iq Competent *E. coli* (NEB, C2992I) following the standard manufacturer heat-shock protocol. Outgrowth product was used to seed 2mL LB cultures at 1-in-200 dilution and incubated overnight at 37 °C, 250 rpm with carbenicillin as the selection marker. Upon reaching an OD600 of 0.6, cultures were split into two 1 mL aliquots; 1mM IPTG was added to one aliquot, the other was kept as an untreated control; all aliquots were incubated at 37 °C for 3hrs to strongly induce protein expression. Barstar-variant-coding regions were amplified directly from 0.2 µL of culture using Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0494S). PCR product was purified as described above, diluted to 5 ng/µL, and Premium PCR Sequencing performed by Plasmidsaurus using Oxford Nanopore Technology with custom analysis and annotation.

Reads were translated and filtered to retain only protein sequences containing the expected N- and C-terminal tag leader sequences and not prematurely truncated by a misplaced STOP codon. Translated reads were mapped back to the foldtuning-generating barstar variant sequences with MMSEQS2, requiring an aligned region of $>$ 80aa with a minimum sequence identity of 98%. Variant enrichment was calculated as the ratio of mapped reads under barnase-barstar induction vs the uninduced control. P-values were computed non-parametrically by assuming a null model of random read allocation, drawing 10^6 samples.

Binding Mode Prediction

AlphaFold3 was used for all structure prediction tasks involving protein-protein or protein-peptide complexes, via the AlphaFold-Server interface (https://alphafoldserver.com). For the SH3 domain, predicted complex structures were computed for foldtuning-generated putative SH3 variants in the presence of a representative class I (RPLPPLP) or class II (PPP:PPRP) proline-rich pep-

tide mofif. For the barstar-like fold, predicted complex structures were computed for foldtuning-generated putative barstar variants in the presence of wild-type barnase from *B. amyloliquefaciens*(uniprot:P00648). Predicted structures were compared to a wild-type reference, either the spectric SH3 domain from *Gallus gallus* or the barnase-barstar complex from *Bacillus amyloliquefaciens* (pdb: 1brs).

ESMFold Validation on Out-of-Distribution Sequences

ESMFold structural prediction model performance on *de novo* proteins was evaluated on 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 (*26*), we filtered out structures with resolution > 9Å, 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 using C_α carbons only, with a median alignment RMSD of 0.92 ± 0.14 Å.

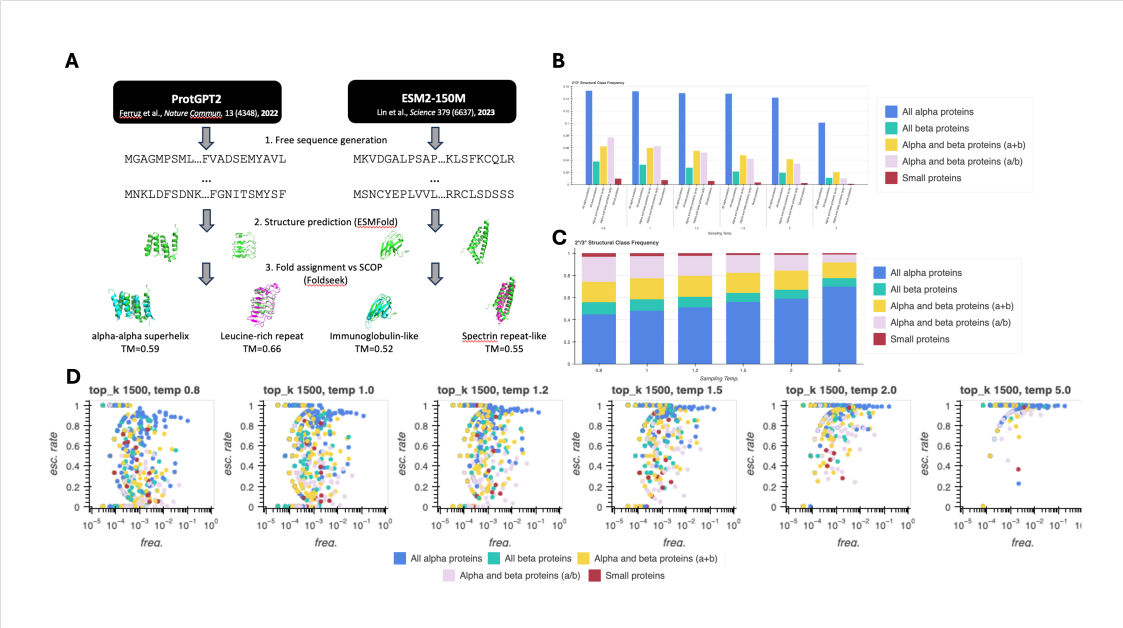

Figure S1: **Increased sampling temperature improves ProtGPT2-generated sequence novelty at the cost of structural coverage.** (**A**) Overview of pLM generation, structure prediction, and structural search+assignment pipeline for representative models ProtGPT2 and ESM2-150M. (**B**) Fraction of ProtGPT2-generated sequences assigned to various fold topology classes (all-α, all-β, $\alpha+\beta$, and α/β) at different sampling temperatures, with top_k fixed at 1500. (**C**) Same data as in (**B**), normalized to remove non-assignable predcted structures. (**D**) Fraction of ProtGPT2-generated sequences assigned to different SCOP folds vs. fraction of assigned sequences with no detectable homology to any cluster representative in UniRef50, at different sampling temperatures, with top_k fixed at 1500, and colored by topology class.

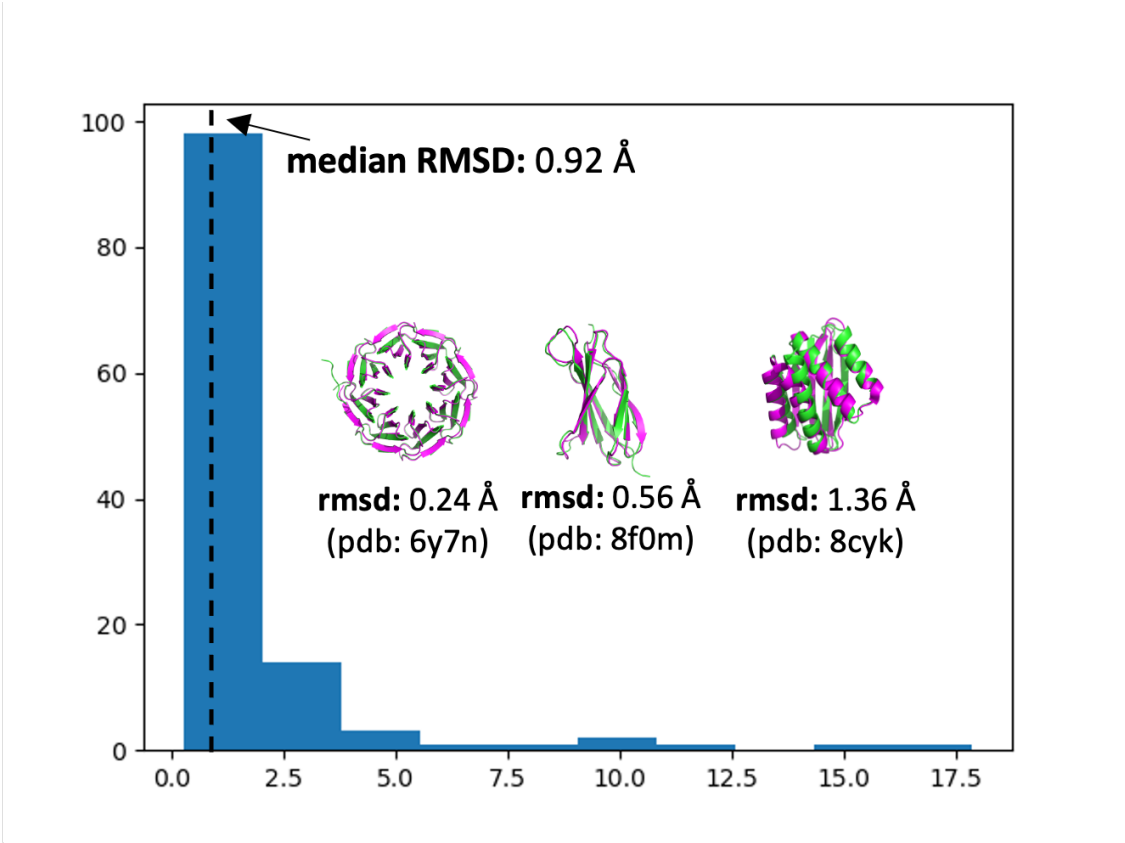

Figure S2: **ESMFold achieves high structure prediction accuracy on far-from-natural sequences.** Backbone aligment RMSD for ESMFold-predicted structures of *de novo* designed proteins vs. experimental ground-truth (n=122, median=0.92 Å); experimental structures restricted to PDB deposition date on-or-after the ESMFold training cutoff of 05-01-2020.

Table S1: **Base ProtGPT2 sequence and structure generation performance depends on sampling hyperparameters (top_k=** $600, 950, 1500$**).**

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

Table S2: **Base ProtGPT2 sequence and structure generation performance depends on sampling hyperparameters (top_k=** $2400, 4000$**).**

top_k	temp	Valid Seq.	# Folds	Struct. Hit	Seq. Esc.
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