# 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], Sonia C. Yuan[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 three markedly different fold targets; the tyrosine-kinase- and small-GTPase-regulating SH3 domain, the bacterial RNase inhibitor barstar, and the peptide hormone insulin 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 $\alpha$-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 inter-

actions, 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. Drawing conceptual inspiration from the counterfeiting games played by generative adversarial networks (GANs), foldtuning leverages competition and complementation between PLMs and structure prediction models to mine protein-space for examples that honor the sequence "grammar" of a target backbone fragment while exhibiting novel semantics (*26, 29, 31–36*). We successfully apply foldtuning to > 700 structural motifs of interest from the SCOP and InterPro databases, covering all four major tertiary topology classes (all-$\alpha$, all-$\beta$, $\alpha+\beta$, and $\alpha/\beta$) 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 three target folds – the SH3 adaptor domain, barstar, and insulin – 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 reflects a ”novelty first” ethos that stretches the limits of how far common protein folds can be diversified at the sequence level while extracting and preserving critical minimal rules of structure and function.

## Sequence exploration with ’soft’ structure constraints

In order to robustly access far-from-natural sequences coding for many structurally diverse fold classes – a feat beyond the reach of off-the-shelf pretrained PLMs, which are vulnerable to dramatic mode collapse – we develop ”foldtuning,” a structure-oriented algorithm that drives a PLM to sample extreme sequence novelty (generation ”in the limit”) while holding to a target fold class, summarized in Fig. 1A. The PLM of choice is first finetuned on natural protein fragments (sourced from a custom SCOP-UniRef50 database or InterPro PDB-derived metadata depending on the target, as described in (*37*)) that adopt the target backbone structure of interest; this initial step is analogous to ”evotuning” on a functional family as has been done in PLM-based enzyme design (*30*). Following this extra fold-specific pretraining, foldtuning proceeds through alternating rounds of: (1) sequence generation out of the current model state, and (2) model update by finetuning on a subset of self-generated artificial sequences that are predicted to coarsely adopt the target fold while differing maximally from natural counterparts in terms of sequence (Fig. 1B-C). Selection for preserving the target fold is achieved by predicting each structure with ESMFold and assigning a SCOP or InterPro label with Foldseek-TMalign search; this is a ”soft” structural constraint, using a TMscore $> 0.5$ global alignment threshold best understood as placing the generated candidate within the target fold *family* or *distribution*. Selection for sequence dissimilarity is enforced by ranking all structurally-validated sequences by 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 – in decreasing order, and taking the top 100 as the next synthetic training data for model updating. Dimension-reduced views of these embeddings for a representative subset of target folds suggests that ESM2-650M captures – and foldtuning navigates along – a representation of the sequence→structure map where structural classes (grouping corresponding pairs of natural and foldtuned artificial sequences) largely separate from one another, with artificial sequences drifting from their natural parents along concerted trajectories in the embedding-space (Fig. 1D,

Fig. S1- S2). Each foldtuning cycle can be thought of as a step along a path that drives a PLM to access subpopulations of progressively further-from-natural artificial sequences while preserving the broad form of the fixed target structure.

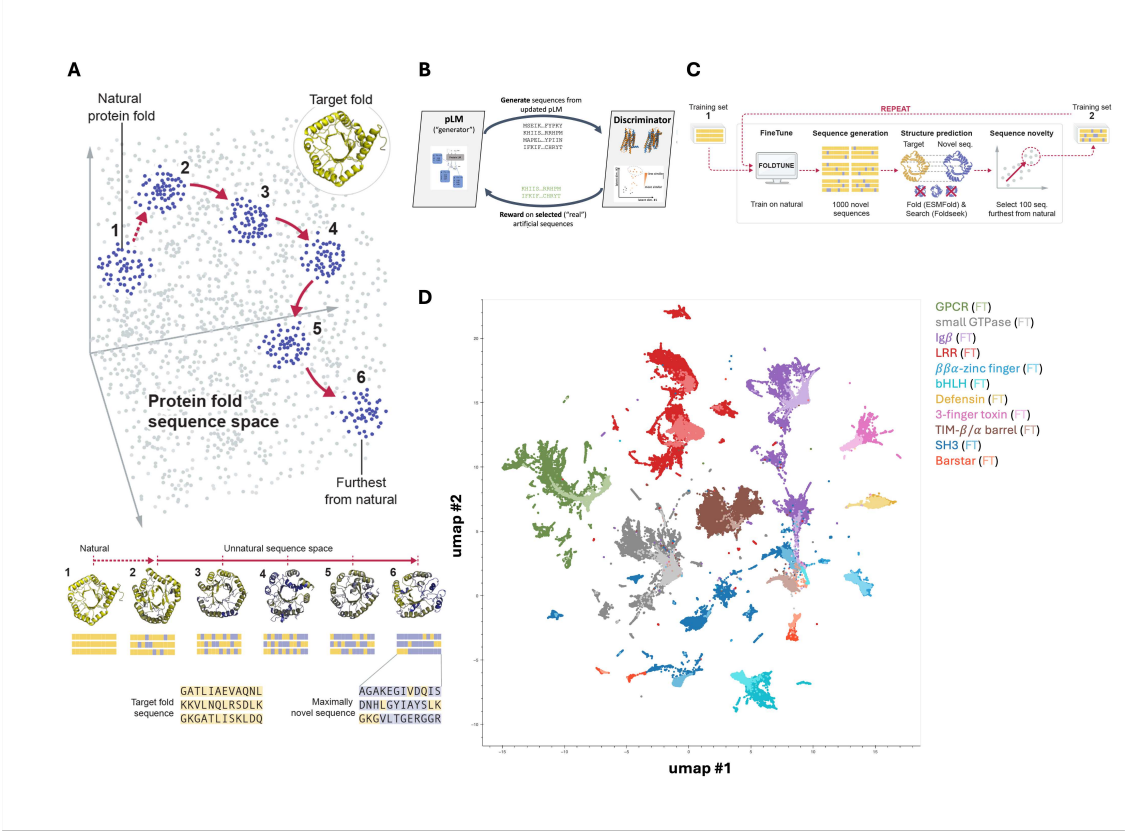



**Figure 1**: **Foldtuning explores far-from-natural sequences encoding alternate versions of natural protein structures.** (**A**) Foldtuning uses a protein language model (PLM)-based strategy to probe outwards in sequence-space, detecting subpopulations of sequences with progressively decreasing similarity to natural examples coding for a given target backbone structure. (**B**) Foldtuning alternates between sequence generation and discrimination/selection rounds in a closed-loop, inspired by the architecture of generative adversarial networks (GANs). (**C**) For a provided backbone target fold, a PLM (ProtGPT2) is initially finetuned (**1**) on target fold examples harvested from deep mining of natural sequence/structure data; in subsequently rounds the PLM is finetuned on self-generated artificial sequences validated by structure prediction (ESMFold) and structure-based search (Foldseek) and selected for maximization of semantic change w.r.t. natural examples (**2**). (**D**) 2D UMAP representation of ESM2-650M embeddings of natural (dark) and foldtuned (light) sequence examples for eleven representative target fold classes.

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. Succesfully foldtuned SCOP targets span numerous classes of functional interest for synthetic biology applications, including transcription factor DNA-binding domains, GPCR/small GTPase signaling components, modular cell surface receptor domains, and defense proteins (e.g. antimicrobial peptides, toxins). Foldtuned versions of ProtGPT2 are effective at landing near the target backbone fold, increasing from a a median *structural hit rate* of 0.203 after evotuning alone to 0.565 after two rounds of updates on far-from-natural artificial sequences, falling slightly to 0.509 after four rounds (Fig. 2A). Sequence novelty relative to natural examples increases with additional update rounds; the *sequence escape rate* – the fraction of target structure matches that do not feature any detectable sequence homology to any protein in UniRef50 – does not change significantly from evotuning (0.134) through two rounds of foldtuning (0.135), but grows steadily to 0.211 after four update rounds (Fig. 2A). When sequences do exhibit homology to natural proteins, the lengths of the aligning subsequences tend to decrease with each additional round of foldtuning, supporting the contention that foldtuning gradually relaxes sequence constraints even when the target structure appears more tightly restrained (Fig. S3). Fold-by-fold semantic change also captures a clear and steady progression away from natural sequences, from a median value of 39.9 following evotuning, to 46.9 after two rounds, to 56.8 after four (Fig. 2B). Notably, at least up to four rounds, foldtuning does not display any significant tradeoff between structural hit rate and sequence escape rate. In many cases, these metrics can be simultaneously maximized (e.g. TIM $\beta/\alpha$ barrels, Ig$\beta$-like domains); or in others, a substantial leap in sequence escape rate is gained with a minimal drop in structural hit rate (e.g. Ferredoxins, Rossman(2x3)oids) (Fig. 2C).

Having a high structural hit rate and a high sequence escape rate would suggest that a fold tolerates substantial sequence plasticity without major disruption 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 $\beta$-helix, ribbon-helix-helix (RHH) domain, TIM $\beta/\alpha$-barrel, anti-parallel $\beta/\alpha$ (PT) barrel, and $\alpha/\alpha$ toroid are ranked as the most designable SCOP motifs, followed by transmembrane $\beta$-barrels, Sm-like barrels, defensins, the winged helix domain, and the POU domain (Fig. 2D, Table S1). Five of these ten motifs are symmetric or periodic in structure; three are transcription factor DNA-binding domains; two have ancient non-specific functions (RNA-binding and antimicrobial

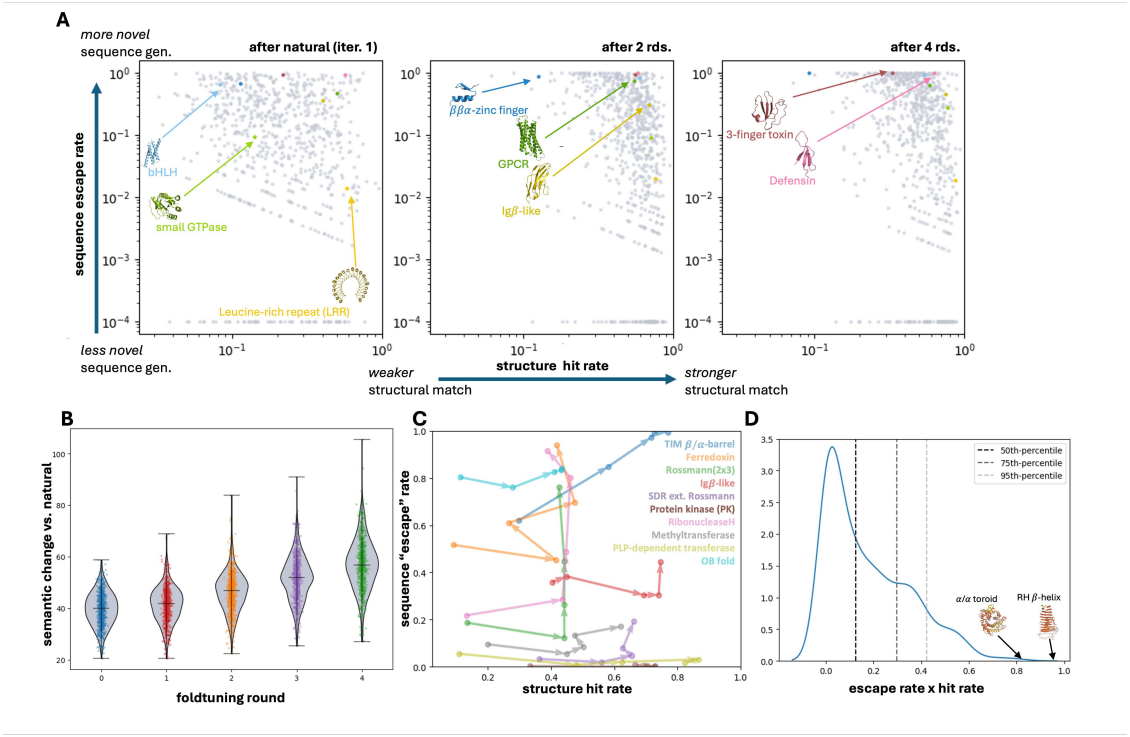



**Figure 2**: **Foldtuned models readily sample novel sequences for** > 700 **targets.** (**A**) Sequence escape vs. structural hit rates after natural-only evotuning or two or four rounds of foldtuning for 727 targets. Selected structural/functional targets are highlighted: transcription factors (blue), GPCRs/small GTPases (green), cell surface receptor domains (gold), and small antimicrobial/toxin proteins (red). (**B**) Semantic change between generated and natural sequences increases with additional rounds of foldtuning. (**C**) Sequence escape rate is maximized without compromising structural hit rate for 10 naturally-abundant target folds. (**D**) Ranking of target folds by sequence "designability", as estimated by the product of structural hit and sequence escape rates.

activity by membrane disruption for Sm and defensins, respectively). Each of these structural and functional traits appears to be a general feature of designable folds, which span the four standard topology classes, not just the all-$\alpha$ helical bundles that are commonly presumed to follow the simplest sequence rules and that are favored by PLMs in the absence of tuning or steering. Furthermore, natural fold abundance in SCOP-UniRef50 is only weakly explanatory of designability, indicating that foldtuning is detecting inherent fold-to-fold variation in the strictness of sequence constraints on a level removed from how evolution has sampled and diversified sequences (Fig. S4).

## Foldtuning explores new sequence rules and populationss

Given the readiness with which foldtuning generalizes to several hundred targets covering structural and functional families of significant relevance to synthetic biology, we turn our attention to sequence features of foldtuning-generated proteins. Taking generated G-protein coupled receptors (GPCRs) and immunoglobin domains (Ig$\beta$-like) as representative examples of interest, we return to PCA→UMAP dimensionality-reduced ESM2-650M embeddings, noting as before that foldtuned versions of ProtGPT2 propose sequences that drift further and further from natural training examples in abstract feature-space; structural fidelity to the targets is preserved as far as high-level shape and connectivity, with the introduction of local plasticity on the order of a few-angstrom root mean square deviation (RMSD) in backbone $C_\alpha$ coordinates vs wild-type (Fig. 3A-B). For GPCRs, foldtuning rapidly converges on generating sequences with no detectable homology against UniRef50, dropping from a median sequence identity of 0.250 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. S3D). Sequence constraints are relaxed more gradually for immunoglobulins, holding at a median sequence identity of 0.336 from evotuning through four foldtuning rounds; the fractional length of the aligning region drops from a median value of 0.695 after evotuning alone to 0.531 after the full four rounds (Fig. S3G). It should also be noted that: (1) this apparent sequence identity barrier for foldtuned immunoglobulins still represents a leap in sequence novelty inaccessible to purely experimental approaches and equivalent to separation over enormous evolutionary timescales; (2) a population of immunoglobulins below the detectable sequence homology threshold persists and expands from 35.9% of valid structure matches (14.5% of all model output) after evotuning to 44.6% of matches (33.3% of all) after four rounds.

All-against-all deep sequence alignment of foldtuned variants (2703 GPCRs, 3035 immunoglobulins) and SCOP-UniRef50 entries (34,327 GPCRs, 150,258 immunoglobulins) reveals that at the sequence level, many foldtuned variants self-cluster into distinct subpopulations infilling regions of sequence-space not sampled by nature (Fig. 3C-D). 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. However, large frac-

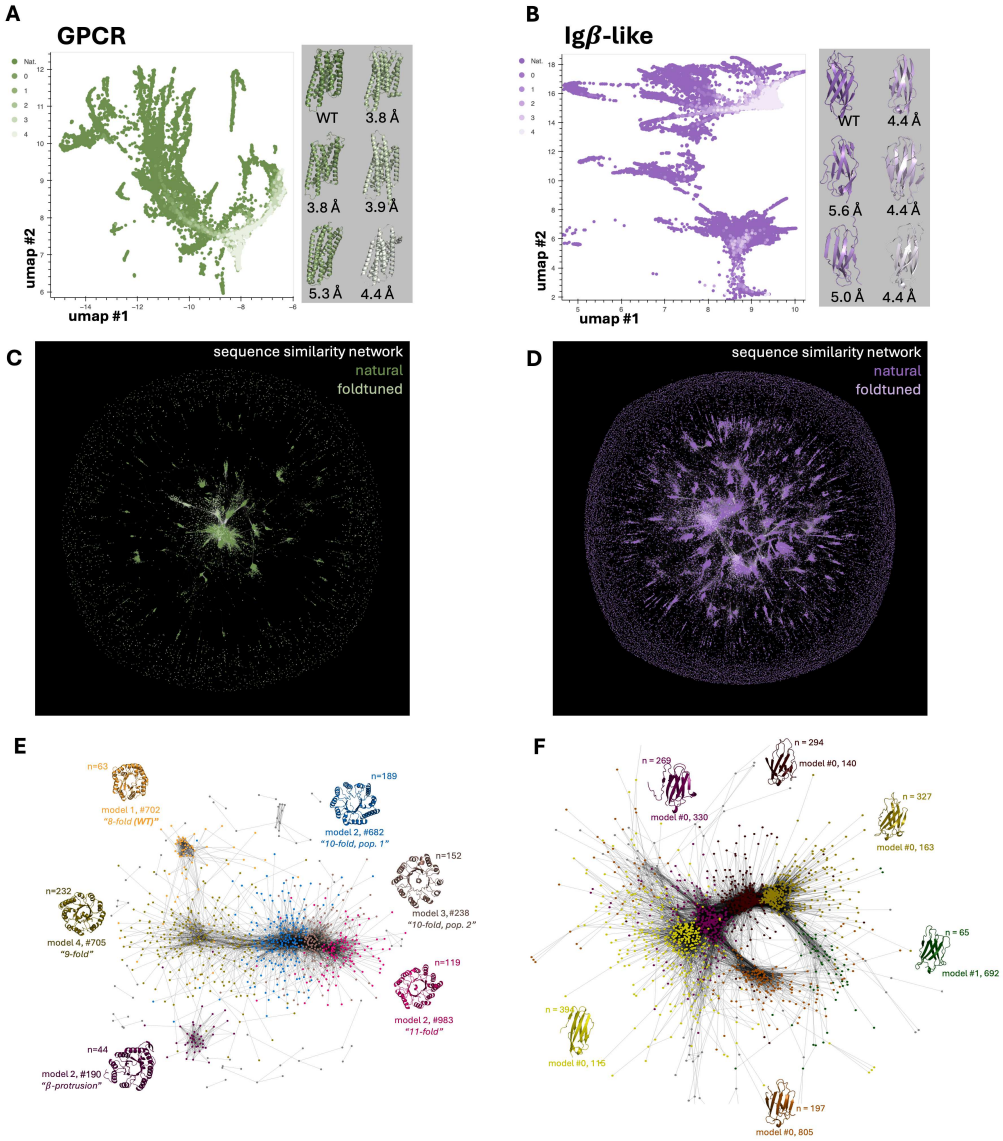



**Figure 3**: **Foldtuning accesses new sequence populations and structural innovations while 'fuzzily' preserving a target backbone.** (**A-B**) UMAP of round-by-round foldtuning sequence diversification captured by ESM2-650M final-layer hidden states for (**A**) G-protein coupled receptors (GPCRs, SCOP: 2000339), (**B**) Immunoglobulin-like domains (Ig$\beta$-like, SCOP: 2000051). (**C-D**) Network representation of similarity between natural (dark) and foldtuned (light) *sequences* for **C** GPCRs (green), (**D**) Ig$\beta$-like domains (purple). (**E-F**) Network representation of *structural* similarity among foldtuned variants, colored by Louvain clustering assignments, for **E** TIM $\beta/\alpha$ barrels (SCOP: 2000031), **F** Ig$\beta$-like domains.

tions of foldtuned variants (332/2323 = 14.3% for GPCRs; 707/2909 = 24.3% for immunoglobulins) are not only dissimilar from natural sequences but from each other, appearing in fold-specific sequence networks as isolated nodes without so much as a homologous snippet to any counterpart real or artificial.

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 an $n$-gram-based ”vocabulary” analysis of foldtuned variants compared to SCOP-UniRef50 examples, splitting 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. Considering the 12 most-abundant natural folds per the SCOP-UniRef50 database, all of which contain $> 50,000$-$250,000$ wild-type examples, we observe noticeable ”vocabulary shifts” – that is, statistically significant upwards or downwards changes in $n$-gram frequency – among foldtuned sequences relative to natural ones for $n$ = 1-4 across all folds analyzed (Fig. S5- S8). For $n = 1$ (equivalent to simple amino-acid composition), 85-100%, or 17 to 20 of the twenty proteinogenic amino acids, shift in usage (Fig. S5). For $n = 2$, 79.0-94.5% of dipeptide ”words” shift (Fig. S6). For $n = 3$, 26.5-75.9% of tripeptides shift (Fig. S7). And for $n = 4$ – a length sufficient as a feature extractor for classifying protein families in past work – as few as 5.7% (Rossmann2x3oid) and as many as 23.3% (PLP-dependent transferases) of ”words” shift in one direction or the other (Fig. S8) (*38*). The substantial vocabulary shift magnitudes (variable from fold to fold) support the contention that foldtuning is stringing new local choices of subsequence motifs into globally perturbed full protein sequences – proposing novel fold-specific sequence languages in lieu of memorizing natural ones. This claim is reinforced by observing that rank-ordered $n$-gram usage by foldtuned models follows the same general distribution as within natural folds – identities of favored and disfavored short motifs change with foldtuning, but semantic breadth is still sampled, forestalling sequence-side compression or collapse (Fig. S9- S12).

## Foldtuning is an implicit innovator of structure and function

We noticed that over the four rounds of foldtuning, without any explicit structural direction to do so, subsets of predicted structures tweak and elaborate on their formal SCOP fold templates, trying out alterations both subtle (e.g. shortening disorded loops, rotating helices) and more substantial (e.g. reversing strand connectivity or altering global symmetry). For instance, the TIM $\beta/\alpha$ -barrel – common to sequentially and functionally diverse enzyme families – undergoes rampant structural exploration in the course of attaining impressive structural hit (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 separates foldtuned TIM barrels into six prominent clusters (Fig. 3E). Only one cluster matches the familiar 8-fold symmetry of the wild-type TIM barrel; a second disrupts that symmetry, ornamenting it with a non-terminal surface $\beta$-hairpin that resembles a natural feature found in predicted structures of cofactor-F420-utilizing bacterial redox proteins. The remaining four clusters correspond to 9-fold, 10-fold (spread across 2 clusters by slight differences in the manner of barrel closure), and 11-fold symmetries, none of which are known to nature based on experimental or predicted structure databases. Applied to foldtuned immunoglobulins, structural clustering differentiates between subpopulations distinguished by the relative orientations of the two $\beta$-sheets in the Ig$\beta$-like sandwich and loop packing (Fig. 3F).

We further evaluate the physical plausibility of sequence-/stucture-pertubed foldtuned proteins *in silico* by scoring their predicted structures with Rosetta to obtain ground-state energy estimates. For eleven target folds of interest, we compute estimated energies for all filtered and validated foldtuned variants and compare to $n \approx 100$ natural training examples (Fig. 4). For all eleven, foldtuned variants sit in the -(1-3) REU/aa regime typically recommended as bounds for distinguishing physically reasonable structures from frustrated ones (*39*). Several targets – $\beta\beta\alpha$-zinc fingers, barstar-like proteins, defensins, GPCRs, small GTPases, helix-loop-helix (HLH) domains – yield energy estimate distributions substantially overlapping those of wild-type examples; others such as immunoglobulin domains, SH3 domains, 3-finger toxins, and TIM barrels hint at possible folding stability penalties relative to WT. For a complementary perspective, variants generated from 55 foldtuned models – targets chosen for potential use in engineering applications as hydrolase and oxidoreductase enzymes, nucleases and base-editors, kinases, proteases, and various scaffolds and

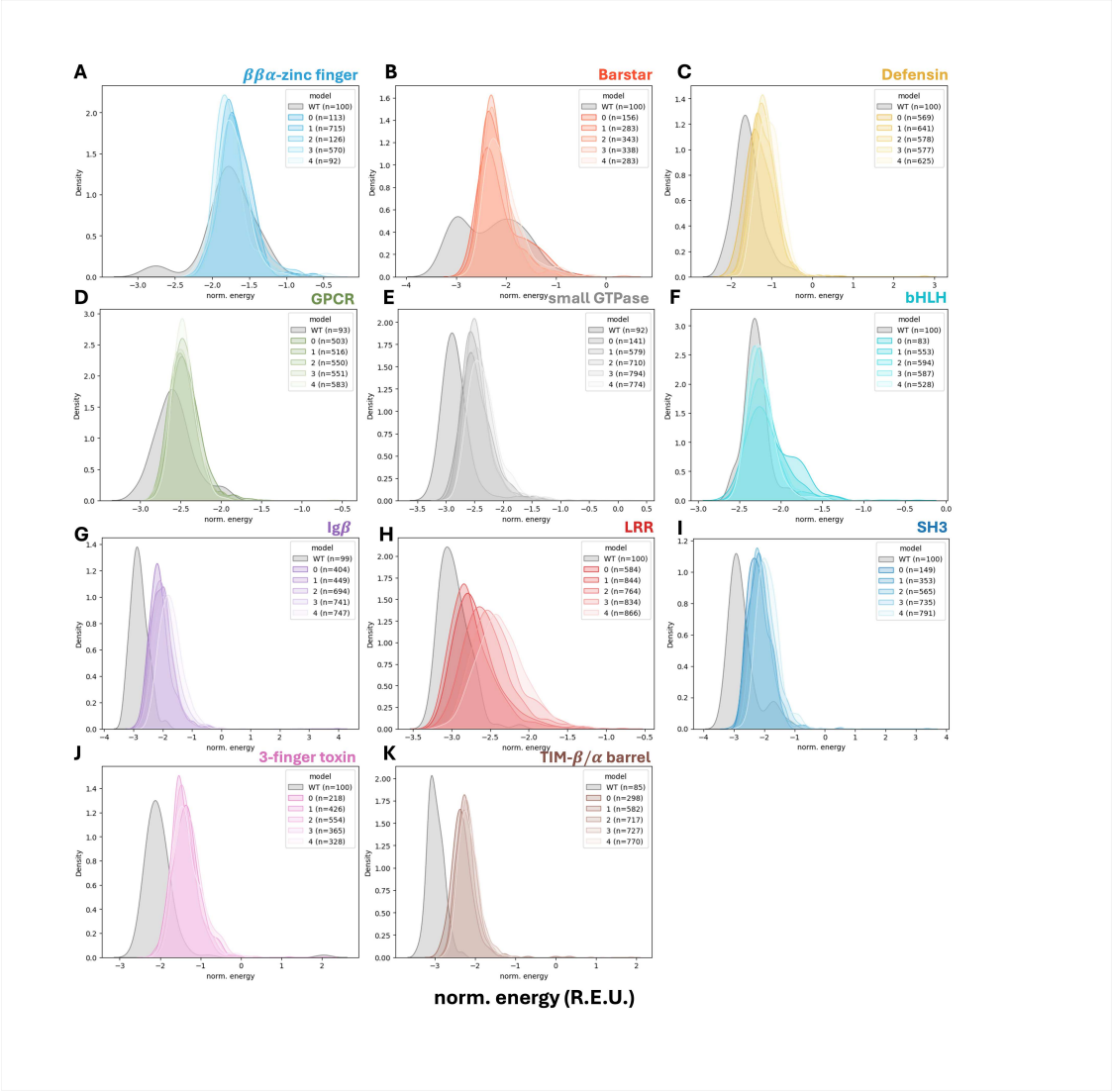



**Figure 4**: **Overlap in estimated energies between foldtuned variants and natural examples.** Histograms of length-normalized (REU / residue) Rosetta energy estimates for foldtuned (colored) and natural (gray) variants. Selected folds – (**A**) $\beta\beta\alpha$-zinc finger. (**B**) Barstar. (**C**) Defensin. (**D**) G-protein coupled receptor (GPCR). (**E**) Small GTPase. (**F**) Basic HLH transcription factor (bHLH). (**G**) Immunoglobulin $\beta$-sandwich (Ig$\beta$). (**H**) Leucine-rich repeat (LRR). (**I**) SH3 domain. (**J**) Three-finger toxin domain (3FTx). (**K**) TIM $\beta/\alpha$ barrel.

mediators of catalysis and protein-protein interactions – were scored with a PLM-based thermostability predictor (Fig. S13) (*40*). Across the board, significant fractions of foldtuned proteins are expected to exhibit melting temperatures > 60°C, restoring some confidence that despite the level of sequence remodeling that occurs, these far-from-natural artificial sequences encode realistic and useful proteins.

Moving down one more level of information flow, we consider whether foldtuned proteins recapitulate or even extend the functional capabilities of their parent folds. To this end, foldtuned variants for several SCOP folds corresponding to specific enzyme families or widely-distributed enzyme scaffolds (i.e. catalyzing diverse chemical transformations across nature), were assigned putative **E**nzyme **C**ommission classification numbers (EC #s) with a PLM-based predictor (*41*). For families with established reactivities and mechanisms, top-level EC #s are largely predicted as expected – P450s and nitrite/sulphite reductases are assigned as oxidoreductases, CRISPR Cas1s and $\alpha/\beta$ hydrolases are assigned as hydrolases, protein kinases are assigned as transferases, and chelatases are are assigned as lyases and ligases, covering their multiple roles in cofactor biosynthesis (Fig. S14). Significant fractions of foldtuned enzymes are annotated into categories asociated with evolvability and and promiscuous activity against a broad spectrum of substrates; e.g. nearly one-in-five foldtuned P450s is placed in EC 1.14.14.1, the catch-all ”unspecified monooxygenase” category associated with the emergence of xenometabolism biocatalysts. Similarly, foldtuned versions of CRISPR Cas1 – a metal-dependent non-site-specific DNA-specific endonuclease are alternately labeled as site-specific, as exonucleases, or even as reverse transcriptases – pointing to fertile ground for engineering stable and sequence-specific gene-editing proteins from foldtuned starting points positioned away from the pitfalls of the edge of stability (*42*). Foldtuned protein kinases span serine/threonine kinases (often with unknown or ambiguous specificity), (receptor)-tyrosine kinases, and dual-specificity kinases that can act on serine, threonine, and tyrosine residues, presaging utility in designing bespoke signaling networks. Foldtuned versions of common enzyme scaffolds are typified by consistent annotation coverage spread across the six top-level EC reaction types, suggesting that foldtuning is preserving functional breadth when learning the sequence determinants of nature’s most widely-used and frequently repurposed domains (Fig. S15).

## Foldtuned SH3 domains express stably

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 fo-

cused 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 $\beta$-barrel design difficulty and a lack of orthogonality to natural SH3s (*43*). Applying the standard evo+four foldtuning procedure to ProtGPT2 with SH3s as the target produced 2593 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 0.519 and 0.310 respectively. In contrast to, e.g. deep-mutational scanning libraries, proteins in foldtuned variant libraries – including for SH3s – boast high sequence diversity, featuring low pairwise sequence similarities and unique proteolytic digestion signatures (Fig. S16A-C). This enables direct high-throughput characterization of protein expression and select biophysical properties by mass-spectrometry-based proteomics without the additional complexity and cost of typical yeast-, mRNA-, or cDNA- display methods (Fig. 5A) (*44, 45*). For our SH3 foldtuned library, 1347/2593 (51.9%) variants express at detectable levels in a reconstituted transcription-translation system as measured by untargeted mass-spectrometric profiling (Fig. 5B-C). Using length-normalized signal as a proxy for absolute abundance of expressed proteins, we observe signal intensity spanning $\sim 6$ orders of magnitude, suggesting substantial variance in the intrinsic expressability of foldtuned SH3s; expression level neither correlates with sequence similarity to natural SH3s nor depends on the number of foldtuning cycles performed (Fig. 5B).

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. 5A). Analysis of the native/denatured signal fold-change for an internal control of $N = 91$ *E. coli* proteins originating from the reconstituted transcription-translation system demonstrates that this stability/solubility proxy has a dynamic range spanning up to $\sim 10$ orders of magnitude under the instrument conditions used (Fig. 5D). A total of 361 variants are detected confidently in both the absolute and multiplexed expression assay including a subpopulation of 87 foldtuned SH3s that are both highly abundant in the initial expression assay and displaced away from the denatured fraction in the solubility/aggreggation assay, suggesting expressability,

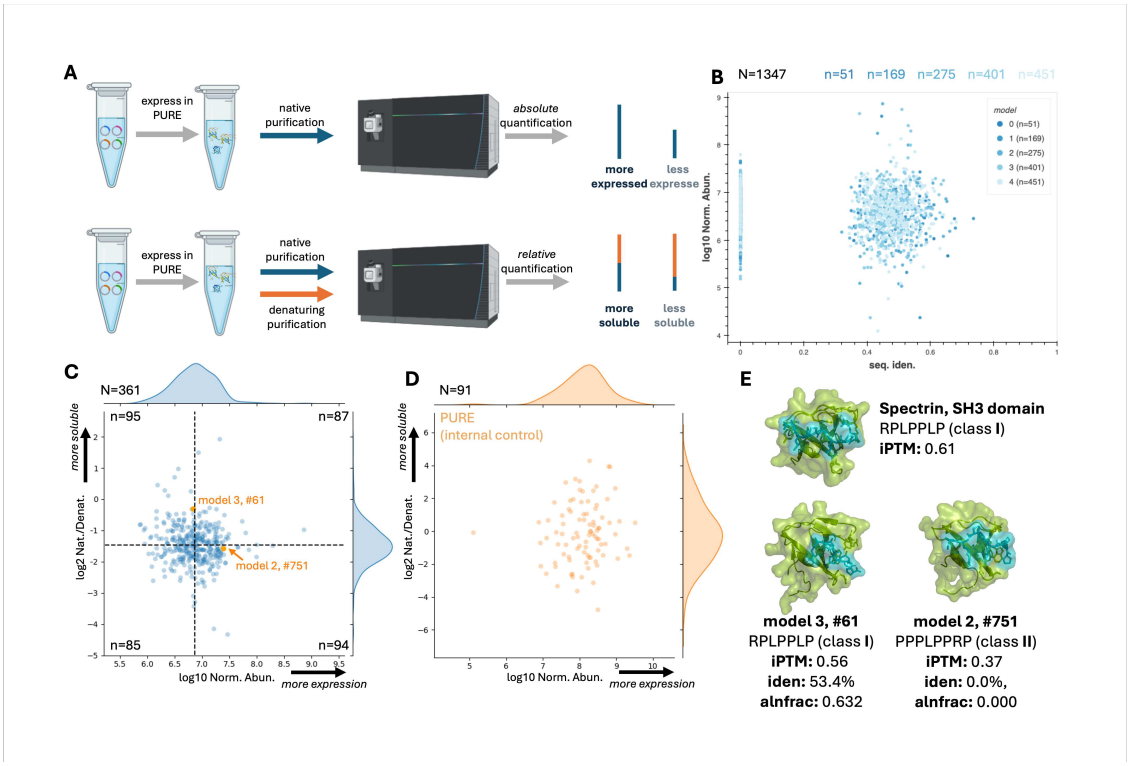



**Figure 5**: **Foldtuned SH3s are expressable and stable.** (**A**). Schematic of mass-spectrometry-based proteomics assays for variant library expression and folding stability. (**B**) SH3 expression assay signal intensity normalized by expected tryptic peptide count vs. sequence identity to closest UniRef50 hit for variants generated after 0-4 rounds of foldtuning. (total $N$ = 1347) (**C**) SH3 folding stability vs. expression assay results for $N$ = 361 variants detected in both screens. Folding stability is measured by relative abundance ratio between native and denaturing purification fractions. Normalized expression is measured as as in (**B**). (**D**) Folding stability vs. expression for internal control set of $N$ = 91 PURExpress components. (**E**) AlphaFold3 predicted structures and iPTM scores for selected SH3 variants (green) bound to a class-I or -II proline-rich peptide (teal), compared to the wildtype *G. gallus* spectrin SH3 domain.

relative folding stability, and low aggregation propensity (Fig. 5C). 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 (*46*). *In silico* screening with AlphaFold3 predicts that certain physically-plausible foldtuned SH3 variants can bind either class I or class II proline-rich ligands in a hydrophobic aromatic-sidechain-rich cleft analogous to the wild-type interface as exemplified by the *G. gallus* spectrin SH3 domain (Fig. 5E). One such example emerging from the AlphaFold-based screen, variant 3_61, is a distant homolog of the

guanine nucleotide exchange factor Vav (involved in cytoskeletal remodeling during lymphocyte development and activation) and is predicted to recognize the canonical class I motif RPLPPLP. Another, variant 2_751, has no detectable sequence homology to any known protein, yet is predicted to recognize the canonical class II motif PPPLPPRP.

To clarify how foldtuned models might be preserving critical structural and functional features in SH3s, including ones responsible for stability or binding of polyproline motifs, we turned to statistical coupling analysis (SCA), applied separately to natural SH3 domains and to the 2593 foldtuned putative SH3s (*15,16,47,48*). Natural and synthetic sectors are composed of non-overlapping sets of core residues; only a single sector position interacts directly with the bound proline-rich motif and it is shared between the natural and synthetic sectors. This suggests that, in line with the promiscuity and diversity of SH3-peptide binding, foldtuning may be preserving a bare-minimum sequence rule for binding few-among-many polyproline-like targets, while trying out a completely different solution for stably packing the SH3 $\beta$-barrel core.

## Foldtuned barstars rescue bacteria from barnase toxicity

We next consider the barstar-like fold (SCOP ID: 2000624). With a single 3-stranded parallel $\beta$-sheet, the barstar-like fold is the simplest $\alpha/\beta$ unit and features a well-studied concerted folding pathway (*49, 50*). Barstar's native function in *B. amyloliquefaciens* is to inhibit, through a high-affinity active-site-occluding non-covalent interaction, the potent broad-spectrum bacterial ribonuclease barnase before its secretion into the surrounding environment. Applying foldtuning to barstar yields 11403 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 0.281 and 0.560 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. 6A) (*51*). 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. 6B). Comparing long-read sequencing counts of variant-coding amplicons, we found that 11 foldtuned barstar variants were significantly enriched ($p < 0.05$; Binyami-Hochberg correction for correlated tests)

relative to uninduced (non-barnase-expressing) control under strong induction of barnase-barstar-variant co-expression, suggesting that the enriched variants are sufficiently functional mimics of barstar so as to mitigate the toxicity of barnase (Fig. 6B). Additionally, enrichment does not correlate with sequence identity relative to wild-type barstars or any natural protein. To this point, 7/11 of survival-enriched foldtuned barstars do not exhibit any detectable homology to natural sequences at the domain or sub-domain level (Fig. 6C).

For mechanistic insight, we obtained AlphaFold3 predicted structures of the survival-enriched variants in complex with barnase. For four foldtuned variants – model 1 #633 (1_633), model 3 #647 (3_647), and model 4 #s 141 (4_141) and 219 (4_219) – these predicted complex structures indicate that barstar mimics are expected to bind barnase analogously to wild-type barstar, inserting an $\alpha$-helix and adjoining loops into the binding pocket, obstructing the RNA hydrolsis active site (Fig. 6D). Detailed examination of predicted binding interfaces reveals that foldtuned barstars are expected to form hydrogen-bonds and salt-bridges with barnase, without steric or electrostatic clashes. Comparison with a published experimental structure of the endogeneous *B. amyloliquefaciens* barnase-barstar complex (pdb: 1BRS) suggests that fewer such contacts are expected with variants than with wild-type barstar, potentially indicating weaker binding, although this different may be an artifact of non-ideal bond geometries that persist due to AlphaFold3's lack of a side-chain or backbone relaxation step (Fig. 6D).

Given the detection of foldtuned barstar mimics with antitoxin-like function and indications that at least some among these mimics may utilize similar structural solutions to wild-type barstar, we continue to ask what sequence and/or structure "rules" foldtuned models themselves have learned. Multiple sequence alignment of wild-type barstar along with the eleven survival-enriched foldtuned variants reveals that in the contiguous nineteen-residue region (columns 38-56) spanning the barnase-binding interface, toxicity-rescuing variants preserve 6-11 (32-58%) of wild-type amino-acid identities (Fig. S18). Clearly, foldtuned models are not simply memorizing the semantics of barnase-binding and scaffolding them into redesigned flanks. We repeat SCA, treating the 1403 foldtuned barstar variants as a synthetic protein family and note a prominent sector mapping onto the barnase-binding interface (Fig. 6E). This suggests that foldtuning has "solved" the barnase-binding problem by distilling the structural-functional "grammar" of barstar into a single rule capturing higher-order sequence dependencies in the critical inserted $\alpha$-helix motif, while

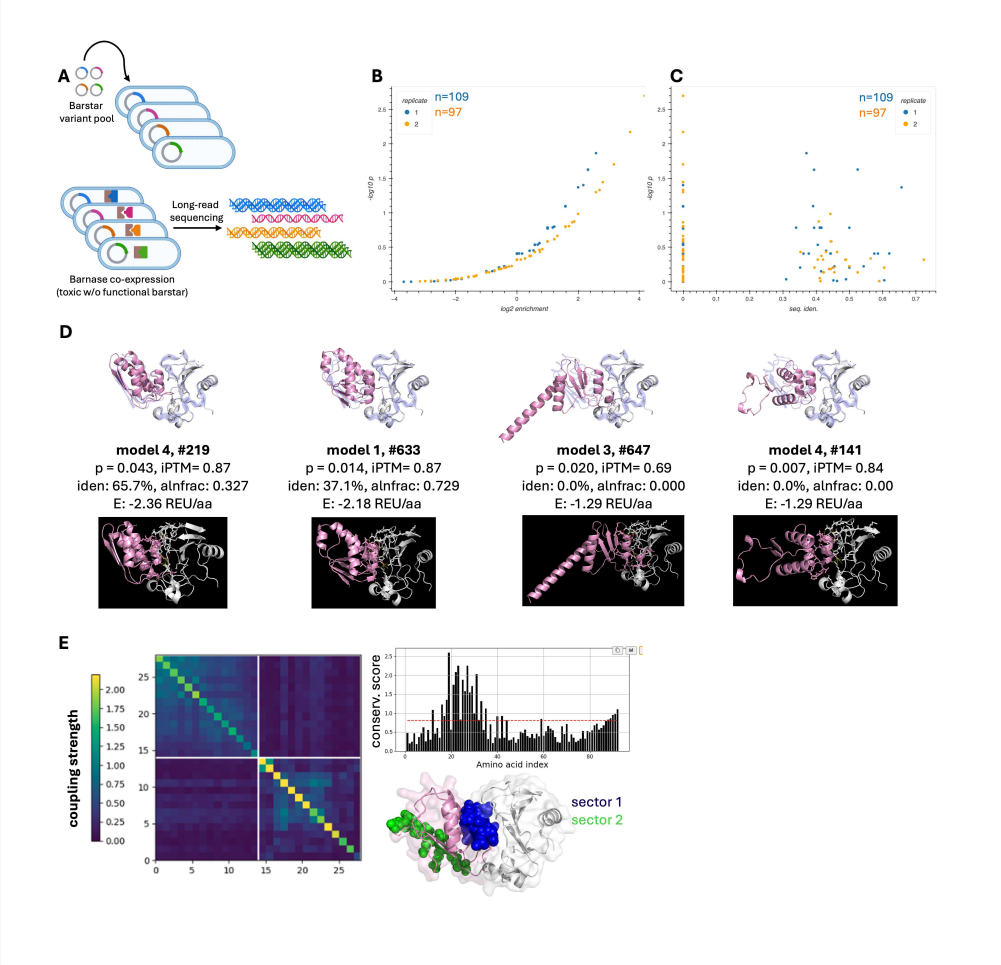



**Figure 6**: **Functional foldtuned barstar variants reflect a minimal structure-function "grammar".** (**A**) Schematic of barnase-inhibition survival assay for barstar variant library stability and function. (**B**) Survival assay p-value rank plot. Enrichment of a given variant is calculated as the ratio of amplicon sequencing reads with and without induction of of barnase co-expression (**C**). Survival assay p-values from (**B**) vs. sequence identity to most-similar UniRef50 hit. (**D**) Top row: AlphaFold3 predicted structures, iPTM scores, and Rosetta energy predictions for selected barstar variants (pink) in complex with barnase (white) with experimental crystal structure of the wild-type *B. aquaforiensis* barnase-barstar complex (pdb: 1BRS) overlaid in blue. Bottom row: Predicted complex structures with putative hydrogen bonds and electrostatic interactions indicated. (**E**) Results of statistical coupling analysis (SCA) on $n$ = 1493 foldtuned barstar sequences. Left: Second-order coupling matrix, blocked into two orthogonally co-evolving sectors. Top right: First-order conservation scores. Bottom right: Visualization of sector positions mapped onto the barstar-barnase complex structure (pdb:2ZA4).

inventing wholly new ways to fill in the more plastic remainders of the barstar fold.

## Foldtuned insulins are INSR binders

Lastly, we steered foldtuning to design mimics of insulin, a high-value translational target well outside of our initial set of 727 SCOP folds, posing several new challenges for the foldtuning algorithm and workflow to overcome. High-level challenges include that the active form of insulin is deeply conserved across eukaryotes at the sequence level, while additionally sharing a structural neighborhood with related peptide hormones including insulin-like-growth-factor-1 (IGF-1), relaxins, and several insulin-like peptides (ILPs) of unclear function and receptor cross-reactivity (*52*). Practical implementation challenges include the post-translational internal cleavage events required to transform inactive, largely disordered proinsulin into structured, active insulin through excision of the C-peptide, and formation of three disulfide bonds, two of which are interchain staples between the A- and B-peptides. To circumvent this issue and to align with standard expression and characterization processes in industry, we foldtunded ProtGPT2 to generate single-chain insulin variants that are fusions of the A- and B-peptides. Natural training sequences ($n$ = 335, reduced to $n$ = 193 after deduplication clustering) and reference structure fragments were taken from InterPro (entry: IPR004825) and sequences multiply aligned to *H. sapiens* insulin to identify putative C-peptide regions that were removed before clustering and downsampling, leaving single-chain A/B fusion training data. Foldtuning yielded 2889 putative insulin variants with structure hit and sequence escape rates of 0.578 and $7 \times 10^{-4}$ respectively. The atypically low sequence escape hit for foldtuned insulin models (only 2/2889 variants lacking detectable homology to natural proteins), as well as a median 80.0% sequence similarity to the closest natural hit, likely stems from the aforementioned high degree of sequence conservation among natural insulins and from a trade-off in choosing a small InterPro family from which to initiate foldtuning in hopes of prioritizing INSR-specific binders and agonists.

We used the Protein CREATE platform to screen all foldtuned putative insulins for INSR-specific binding as described in (*53*). In brief, variants are displayed on T7 bacteriophage and screened against multiple receptor candidates ligated to magnetic beads, with a sequencing-based readout of amplicon counts before- and after- receptor-bead pulldown, resulting in a vector of

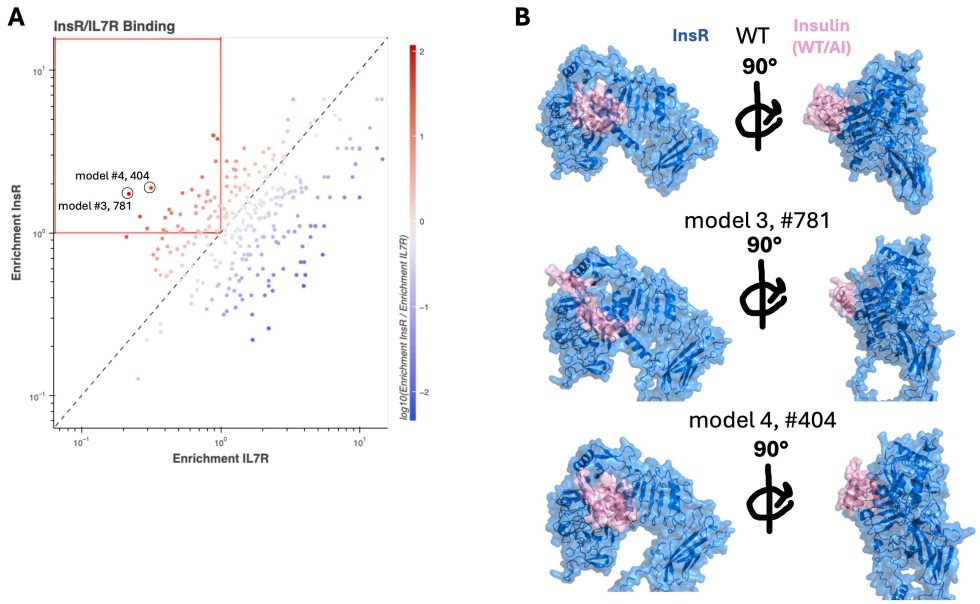



**Figure 7**: **Foldtuned insulin variants bind the native insulin receptor.** (**A**) Relative enrichment plot of foldtuned insulin variant binding to the native INSR receptor (on-target) vs the native IL7R receptor alpha-chain (off-target) using the Protein CREATE platform. (**B**) AlphaFold3 predicted structures of WT *H. sapiens* insulin and enriched foldtuned INSR binders (pink) in complex with the native INSR receptor ectodomain (blue).

enrichment scores for each receptor screened. Here we screen against two receptors, taking enrichment after INSR pulldown as a measure of on-target binding, and enrichment after IL7RA (a type I cytokine receptor) pulldown as a measure of generic off-target binding. For 307 foldtuned variants with sufficient reads to calculate enrichment in both contexts, 41 (13.4% of detected; 1.4% of entire foldtuned pool) are enriched (enrichment > 1) for INSR-binding and de-enriched (enrichment < 1) for IL7RA-binding (Fig. 7A). AlphaFold3 predicts that the varants with the highest relative enrichment scores (3_781, 4_404) bind to the INSR ectodomain with ligand-receptor contacts reminiscent of but not identical to those formed by wild-type insulin, lending support to the emerging paradigm from our investigations of SH3 and barstar that foldtuning retains only those sequence rules minimally necessary for marginal binding while injecting novelty that percolates to perturbed contacts, pockets, and downstream phenotypes (Fig. 7B).

## Conclusion

We introduced foldtuning as a ”novelty first” 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. Experimentally 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.

# 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. *Nature 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. *Nature Reviews Molecular Cell Biology* **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. *Journal of Molecular Evolution* **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. 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. *Nature Communications* **13** (1), 4348 (2022), number: 1 Publisher: Nature Pub-

lishing 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. *Nature Biotechnology* 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. I. J. Goodfellow, *et al.*, Generative Adversarial Nets, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc.), vol. 27 (2014), https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

32. 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.

33. 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.

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

35. 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.

36. G. A. Pavlopoulos, *et al.*, Unraveling the functional dark matter through global metagenomics. *Nature* **622** (7983), 594–602 (2023), number: 7983 Publisher: Nature Pub-

lishing Group, doi:10.1038/s41586-023-06583-7, https://www.nature.com/articles/s41586-023-06583-7.

37. Materials and methods are available as supplementary material.

38. S. M. A. Islam, B. J. Heil, C. M. Kearney, E. J. Baker, Protein classification using modified n-grams and skip-grams. *Bioinformatics* **34** (9), 1481–1487 (2018), doi:10.1093/bioinformatics/btx823, https://doi.org/10.1093/bioinformatics/btx823.

39. R. F. Alford, *et al.*, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13** (6), 3031–3048 (2017), publisher: American Chemical Society, doi:10.1021/acs.jctc.7b00125, https://doi.org/10.1021/acs.jctc.7b00125.

40. I. Pudžiuvelytė, *et al.*, TemStaPro: protein thermostability prediction using sequence representations from protein language models. *Bioinformatics* **40** (4), btae157 (2024), doi:10.1093/bioinformatics/btae157, https://doi.org/10.1093/bioinformatics/btae157.

41. T. Yu, *et al.*, Enzyme function prediction using contrastive learning. *Science* **379** (6639), 1358–1363 (2023), publisher: American Association for the Advancement of Science, doi:10.1126/science.adf2465, https://www.science.org/doi/full/10.1126/science.adf2465.

42. D. M. Taverna, R. A. Goldstein, Why are proteins marginally stable? *Proteins: Structure, Function, and Bioinformatics* **46** (1), 105–109 (2002), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.10016, doi:10.1002/prot.10016, https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10016.

43. 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.

44. G. J. Rocklin, *et al.*, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357** (6347), 168–175 (2017), publisher: American Association for the

Advancement of Science, doi:10.1126/science.aan0693, https://www.science.org/doi/full/10.1126/science.aan0693.

45. 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.

46. B. J. Mayer, SH3 domains: complexity in moderation. *Journal of Cell Science* **114** (7), 1253–1263 (2001), doi:10.1242/jcs.114.7.1253, https://doi.org/10.1242/jcs.114.7.1253.

47. G. M. Süel, S. W. Lockless, M. A. Wall, R. Ranganathan, Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology* **10** (1), 59–69 (2003), number: 1 Publisher: Nature Publishing Group, doi:10.1038/nsb881, https://www.nature.com/articles/nsb881.

48. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138** (4), 774–786 (2009), publisher: Elsevier, doi:10.1016/j.cell.2009.07.038, https://www.cell.com/cell/abstract/S0092-8674(09)00963-5.

49. G. Schreiber, A. M. Buckle, A. R. Fersht, Stability and function: two constraints in the evolution of barstar and other proteins. *Structure* **2** (10), 945–951 (1994), publisher: Elsevier, doi:10.1016/S0969-2126(94)00096-4, https://www.cell.com/structure/abstract/S0969-2126(94)00096-4.

50. 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.

51. R. W. Hartley, [38] - Barnase–Barstar Interaction, in *Methods in Enzymology*, A. W. Nicholson, Ed. (Academic Press), vol. 341 of *Ribonucleases - Part A*, pp. 599–611 (2001), doi:10.1016/S0076-6879(01)41179-7, https://www.sciencedirect.com/science/article/pii/S0076687901411797.

52. I. Claeys, *et al.*, Insulin-related peptides and their conserved signal transduction pathway. *Peptides* **23** (4), 807–816 (2002), doi:10.1016/S0196-9781(01)00666-0, https://www.sciencedirect.com/science/article/pii/S0196978101006660.

53. 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.

54. 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.

55. 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.

56. M. Blum, *et al.*, InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* **53** (D1), D444–D456 (2025), doi:10.1093/nar/gkae1082, https://doi.org/10.1093/nar/gkae1082.

57. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29** (4), 1165–1188 (2001), publisher: Institute of Mathematical Statistics, doi:10.1214/aos/1013699998, https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-4/The-control-of-the-false-discovery-rate-in-multiple-testing/10.1214/aos/1013699998.full.

58. R. C. Edgar, Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* **13** (1), 6968 (2022), publisher: Nature Publishing Group, doi:10.1038/s41467-022-34630-w, https://www.nature.com/articles/s41467-022-34630-w.

59. O. Rivoire, K. A. Reynolds, R. Ranganathan, Evolution-Based Functional Decomposition of Proteins. *PLOS Computational Biology* **12** (6), e1004817 (2016), publisher: Public Library of Science, doi:10.1371/journal.pcbi.1004817, `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004817`.

# 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. S.C.Y. carried out bioinformatics analyses. 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.

# Supplementary Materials for

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

Arjuna M. Subramanian, Zachary A. Martinez, Alec L. Lourenço, Sonia C. Yuan, Shichen Liu, Matthew Thomson*

*Corresponding author. Email: mthomson@caltech.edu

**This PDF file includes:**

Materials and Methods

Figure S1 to S18

Table S1

## Materials and Methods

Except where otherwise specified, all model access and interfacing was via TRILL v1.3.11 (*54*).

### 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 ($n = 36,900$) against the UniRef50 portion (based on the 2021_04 release) included in the July 2022 update to the AlphaFoldDB as first reported in (*55*) and made available as a precompiled FOLDSEEK database in (*34*), filtering for reciprocal hits with fractional query and target coverage > 0.8 and TMscore> 0.5, and clustering the filtered fragments at 100% identity.

### Target Fold Selection for Foldtuning

Out of 1562 folds categorized in SCOP v2, 1474 are present in the SCOP-UniRef50 database whose construction is described above. The top 850 most-abundant of these comprise the intial target set for foldtuning, a cutoff selected in part out of consideration for compute resource constraints and in part to exclude folds with potentially inadequate volumes of natural sequence starting material. As a second target set, we hand-select 44 cytokine, chemokine, and growth factor entries from InterPro, motivated by functional protein engineering applications (*56*).

### Sequence Selection for Evotuning

For the preliminary foldtuning round on SCOP target fold $f$, termed the evotuning round, the base ProtGPT2 model was finetuned for 1-3 epochs on 100 natural sequences selected at random from the subset of sequences in the custom SCOP-UniRef50 database (construction described above) annotated to fold $f$.

For the evotuning round on InterPro target entry $f_{IP}$, 100 natural sequences were selected at random from sequences associated with $f_{IP}$ in INTERPRO v93.0, preliminarily clustered at 100% sequence similarity with MMSEQS2 for deduplication and fragment removal (*56*).

### Finetuning of ProtGPT2

All finetuning of ProtGPT2 was performed with the Adam optimizer using a learning rate of 0.0001, and next-token prediction as the causal language modeling task. For the evotuning round, finetuning proceeded for 1-3 epochs, with the number of epochs for a specific SCOP fold $f$ or InterPro fold $f_{IP}$ 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 $\geq 7\%$ of sequences were assigned to fold $f$ in order to ensure sufficient synthetic data to initiate foldtuning.

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 $f$ or $f_{IP}$ ranked in order of decreasing semantic change as described in the main text and below.

### Sequence Generation from ProtGPT2

Sequences 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. Per round, $n = 1000$ sequences were generated from the appropriate evotuned or foldtuned model. The termination condition was set following $0.4 \times M$ tokens, where $M$ is the median length of SCOP-UniRef50 natural sequences for target fold $f$, or the first STOP token, whichever occurred first; generated sequences were force-truncated to a maximum length of $M$aa. Truncated sequences containing rare or ambiguous amino acids (B, J, O, U, X, or Z) were filtered out as invalid. 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 were predicted with default ESMFold inference parameters as in (*26*). Structures were inferenced in batches of 10-500, depending on sequence length, on single A100-80G GPUs, with compute resource collaboration through Oracle Cloud Infrastructure (OCI).

Predicted structures were annotated either to: (1) SCOP fold labels via Foldseek structure-based search against a custom database comprised of the $n = 36,900$ superfamily-level representa-

tive structures in SCOP v2; (2) InterPro entry labels via FOLDSEEK structure-based search against a custom database comprised of structures compiled from 44 chemokine, cytokine, and growth factor entries in INTERPRO v93.0. Irrespective of target databse, FOLDSEEK was run in accelerated TMalign mode. The consensus SCOP fold or InterPro entry was defined as the fold/entry accounting for the most hits with TMscore > 0.5 and max(query_coverage, target_coverage) > 0.8. In the absence of at least one hit satisfying these criteria, a structure was considered to be un-assignable.

### Sequence Selection for Foldtuning

For each target fold $f, f_{IP}$ and foldtuning round $k = 1, 2, ...N$, the semantic change relative to natural versions was calculated for all generated sequences $\{s_k^{(i)}\}$ structurally assigned to fold $f, f_{IP}$ as

$$z_k^{(i)} = \min_j \|x_k^{(i)} - x_{train}^{(j)}\|_1 \tag{S1}$$

where $s_k^{(i)} \mapsto x_k^{(i)} \in \mathbb{R}^{1280}$ via embedding with ESM2-650M, and the "train" subscript denotes the natural sequences selected from SCOP-UniRef50 or InterPro for the initial foldtuning round. The $\{s_k^{(i)}\}$ were ranked by their corresponding $\{z_k^{(i)}\}$ in descending-order; the top 100 comprising the finetuning sequence data for the $(k+1)$-th round.

### *In Silico* Evaluation of Foldtuned Models & Outputs

**Structural Hit, Sequence Escape, and Designability Rates** 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 $f$. More formally, for a generated sequence $s_i$ and fold $f$, it is $\Pr(s_i \in f)$. 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. Or, formally, $\Pr(s_i \notin \mathbb{N} | s_i \in f)$, where we borrow $\mathbb{N}$ to stand in for the set of all natural/natural-resembling/homologous-to-natural sequences. The "designability" of a fold $f$ was computed as the product of the corresponding structural hit and sequence escape rates, or $d_f = \Pr(s_i \notin \mathbb{N} | s_i \in f) \times \Pr(s_i \in f) = \Pr(s_i \notin \mathbb{N}; s_i \in f)$.

**PCA and UMAP Representations** Mean-pooled embeddings for natural and foldtuned sequences were inferenced with ESM2-650M and dimension-reduced from $\mathbb{R}^{1280}$ to $\mathbb{R}^{100}$ by principal component analysis (PCA) and further to $\mathbb{R}^2$ by Uniform Manifold Approximation and Projection (UMAP). For the eleven chosen folds depicted in Figure 1, Figure S1, and Figure S2, natural sequences were sampled from SCOP-UniRef50 at 5x the number of filtered, validated foldtuned sequences obtained after initial evotuning+four rounds.

**Sequence Similarity Analysis and Clustering** Sequence network analysis was carried out by separately preclustering foldtuned sequences and natural SCOP-UniRef50 sequence fragments assigned to fold $f$ at 50% identity, via MMSEQS2 easy-cluster with default settings and covariance mode 1. Preclustered sequence sets were then merged and searched all-against-all using MMSEQS2 easy-search with maximum e-value $10^{-5}$. 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%. Visualization was with NETWORKX, with node positions calculated according to a force-directed representation with spring constants $k_{ij} \propto \{$ seq. iden. between $s_i, s_j\}$.

**Structural Similarity Analysis and Clustering** 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 reciprocal average TMscore > 0.7, and Louvain clustering was performed with NETWORKX with default parameters to separate the network into fold motif clusters. Isolated nodes were excluded from clustering and visualization.

**Energy Scoring Calculations** Biomolecule energy scores were obtained using the default 'ref2015' energy function and standard relaxation and scoring workflow in ROSETTA v3.11, as described in (*39*). Energy scores are reported in **R**osetta **E**nergy **U**nits (R.E.U.), normalized to sequence length.

### Advanced Chemical Property Prediction and Visualization

Melting temperature bin predictions ($T_m$) for thermostability were obtained for all foldtuned sequences using the $40°C$, $45°C$, $50°C$, $55°C$, $60°C$, and $65°C$ binary classifiers released as part of TemStaPro v0.2.6 (*40*).

Functional enzyme reactivity annotation labels (**E**nzyme **C**ommission #s; EC#s) were inferred for thirty-one classes of foldtuned sequences using the fast "max-separation" mode of CLEAN v1.0.1 (*41*). Where multiple EC#s were inferred for a given sequence, the closest centroid was retained as the best-scoring annotation. The full body of EC# annotations across all scored sequences for a given fold were visualized using KronaTools v2.8.1 with XML customization to maintain a consistent color scheme for top-level EC# classification: oxidoreductases (EC 1; red), transferases (EC 2; yellow), hydrolases (EC 3; green), lyases (EC 4; blue), isomerases (EC 5; purple), ligases (EC 6; pink).

### Sequence *N*-Gram Decomposition and Analysis

*N*-gram vocabulary analysis was carried out with custom code by splitting foldtuned sequences and SCOP-UniRef50 sequence fragments assigned to fold $f$ into subsequences ("words") of length 1,2, 3, or 4 and computing their respective frequency distributions and fold-change for foldtuned variants vs. natural SCOP-UniRef50 sequences. For each fold/word-length pair, $n = 1000$ non-parametric bootstrap replicates were drawn with the SCOP-UniRef50 sequences as the null distribution and significance testing for individual word frequency change performed at significance level $\alpha = 0.05$, applying the Binyamini-Hochberg correction for positively correlated tests (*57*).

### 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 $C_\alpha$ 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 $\leq 0.19$ or $\geq 0.71$ over any subsequence of length 50) and homopolymers (restricted to $< 14$nt). Constant flanks – GACTACAAGGACGACGATGACAAG (5') and GGTTC-

CCACCATCATCACCATCAT (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. T7RNAP 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 NEB-Express 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 $2 \times 25$ 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 per pair, 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.

### Bioinformatics Analysis

Multiple sequence alignments (MSAs) were calculated using MUSCLE v5 via the EMBL-EBI webserver (*58*).

Statistical coupling analysis (SCA) was performed with PYSCA v6.1 and visualizations created with PYMOL v3.1.0 (*59*).

### Binding Mode Prediction and Analysis

Unless specified to the contrary, AlphaFold3 was used for all structure prediction tasks on 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 (PPPLPPRP) proline-rich peptide motif. 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 spectrin SH3 domain from *Gallus gallus* or the barnase-barstar complex from *Bacillus amyloliq-*

*uefaciens* (pdb: 1brs). For insulin, predicted complex structures were computed for foldtuning-generated and/or PLM-sampled putative insulin variants in complex with the monomeric full-length ectodomain of human INSR (insulin receptor).

All predicted structures were visualized with PyMOL v3.1.0. For the barnase-barstar complex, good hydrogen-bonds, acceptable hydrogen-bonds, and electrostatic clashes were inferred and displayed with the PyMOL "show_contacts" third-party plugin. For insulins, hydrophobicity was visualized using the "color_h" third-party plugin and electrostatic potential was calculated and visualized using the APBS Electrostatics plugin.

**High-Throughput Insulin Binding Assay**

A library of 2889 insulin variant amino-acid sequences was constructed by foldtuning on InterPro entry IPR004825, containing 335 natural insulin sequences (reduced to 193 sequences after deduplication clustering at 100% similarity with MMSEQS2) integrated from overlapping entries in the PRINTS, CDD, and PANTHER databases. Foldtuning was executed as otherwise described, with the modification that generated variants were post-processed by aligning to the sequence *H. sapiens* insulin (uniprot: P01308) and removing residues aligning to the C-peptide region that is removed by proteolytic cleavage *in vivo* during the conversion of inactive proinsulin to active insulin, resulting in a library of *single-chain* insulin mimics.

High throughput binding measurements (sequencing read enrichment scores) were obtained using the Protein CREATE platform as described in (*53*) with INSR as the on-target receptor and IL7RA as the off-target decoy receptor.

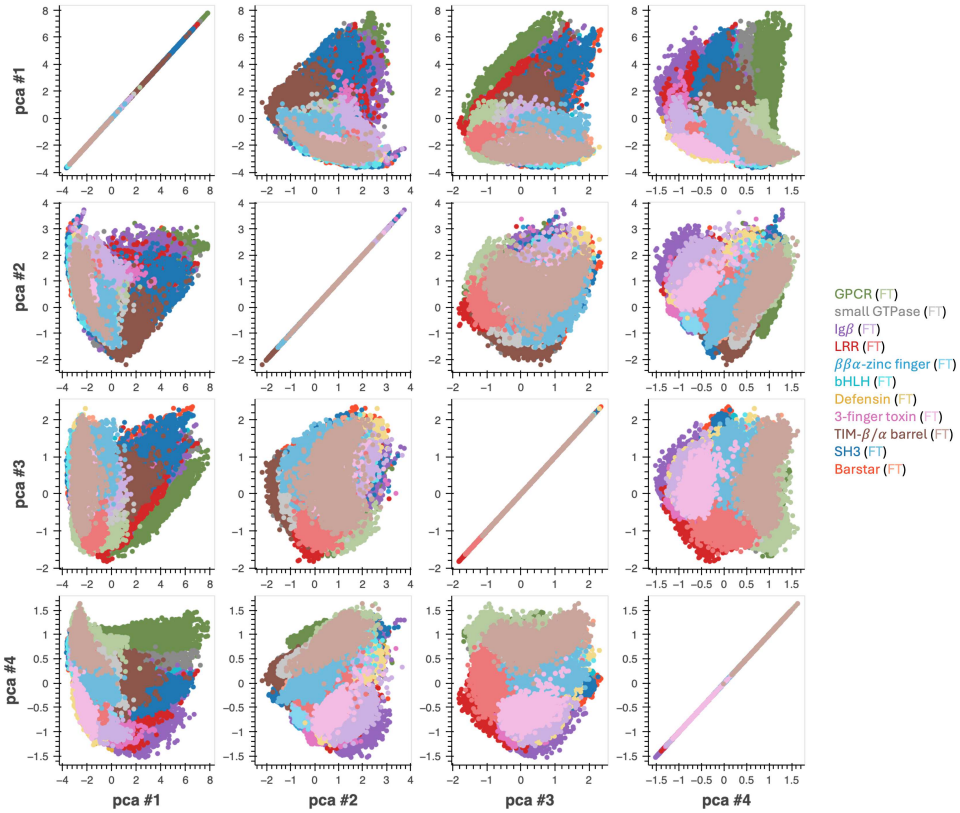



**Figure S1**: **Principal component analysis (PCA) of natural and foldtuned ESM2-650M embeddings.** Pairwise plots of top 4 principal components (fractional variance: 0.386, 0.103, 0.047, 0.039, respectively) of ESM2-650M embeddings of natural (SCOP-UniRef50) and foldtuning-generated proteins for 11 SCOP folds – GPCRs, small GTPases, immunoglobulin-like domains (IgBs), leucine-rich repeat domains (LRRs), $\beta\beta\alpha$-zinc finger transcription factors, bHLH transcription factors, defensins, three-finger toxins (3FTxs), TIM-$\beta/\alpha$ barrels, SH3 domains, and barstar-like domains.

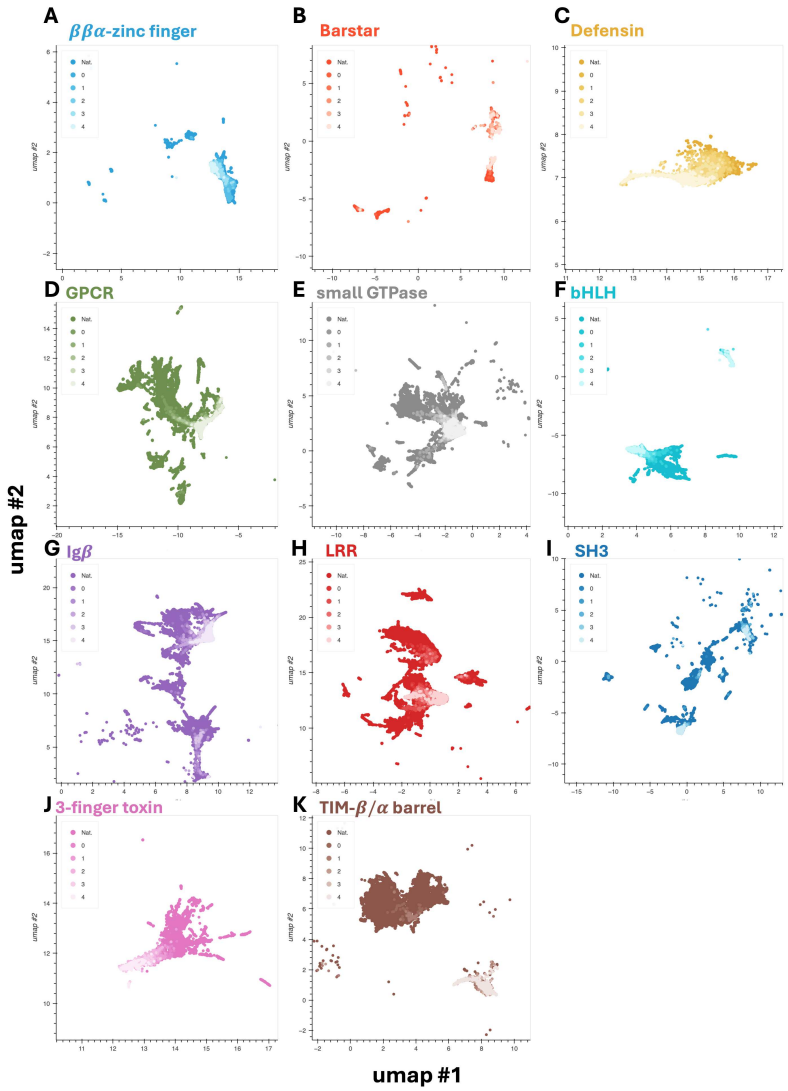



**Figure S2**: **ESM2-650M embeddings capture round-by-round drift of foldtuned sequences from their natural parents.** 2D UMAP representation of ESM2-650M embeddings for eleven representative target fold classes, progressing from natural examples through up to four rounds of foldtuning. Selected folds – (**A**) $\beta\beta\alpha$-zinc finger. (**B**) Barstar. (**C**) Defensin. (**D**) G-protein coupled receptor (GPCR). (**E**) Small GTPase. (**F**) Basic HLH transcription factor (bHLH). (**G**) Immunoglobulin $\beta$-sandwich (Ig$\beta$). (**H**) Leucine-rich repeat (LRR). (**I**) SH3 domain. (**J**) Three-finger toxin domain (3FTx). (**K**) TIM $\beta/\alpha$ barrel. Subfigure boundaries are set to the 5th- and 95th- quantiles in each UMAP component.

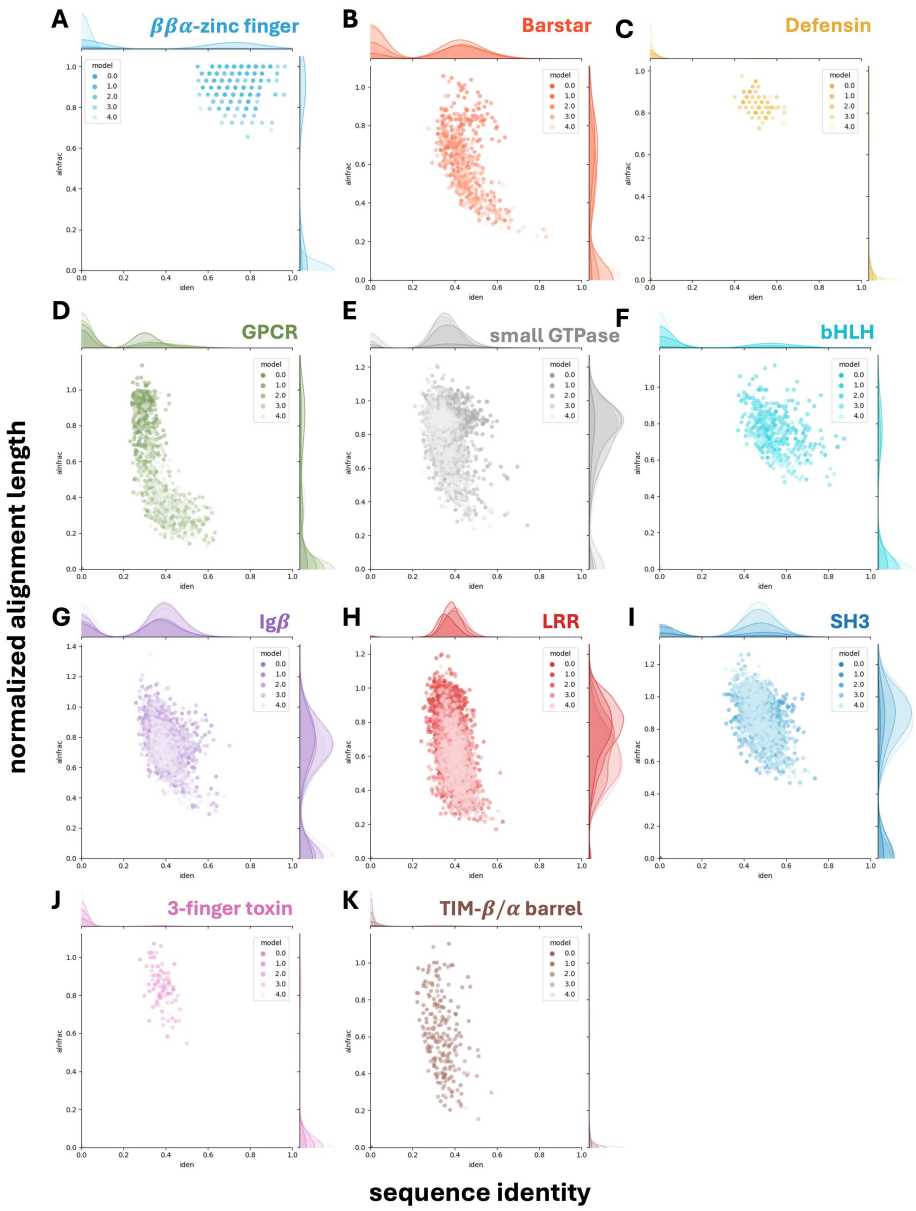



**Figure S3**: **Sequence similarity between foldtuned and natural variants.** Plots of normalized alignment length vs. sequence identity over the aligned region for the closest UniRef50 homolog to each foldtuned variant as identified by ultrasensitive search with MMSeqs2. Selected folds – (**A**) $\beta\beta\alpha$-zinc finger. (**B**) Barstar. (**C**) Defensin. (**D**) G-protein coupled receptor (GPCR). (**E**) Small GTPase. (**F**) Basic HLH transcription factor (bHLH). (**G**) Immunoglobulin $\beta$-sandwich (Ig$\beta$). (**H**) Leucine-rich repeat (LRR). (**I**) SH3 domain. (**J**) Three-finger toxin domain (3FTx). (**K**) TIM $\beta/\alpha$ barrel.

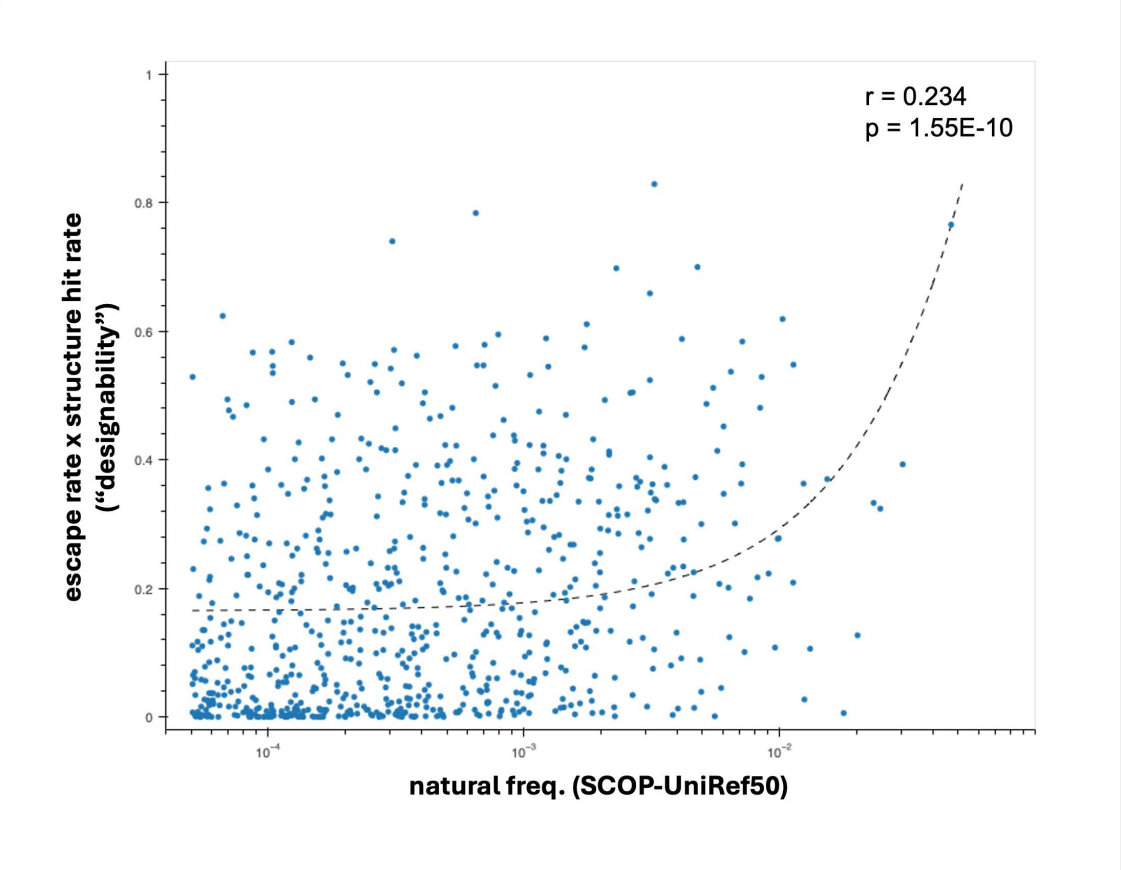



**Figure S4**: **Designability vs natural abundance for** $n = 708$ **SCOP fold targets.** Designability proxy (structural hit rate × sequence escape rate) across $n = 708$ SCOP fold targets is weakly explained by natural abundance in the custom SCOP-UniRef50 database: linear regression $t$-test for positive slope; slope= 12.80, $r = 0.234$, $p = 1.55 \times 10^{-10}$.

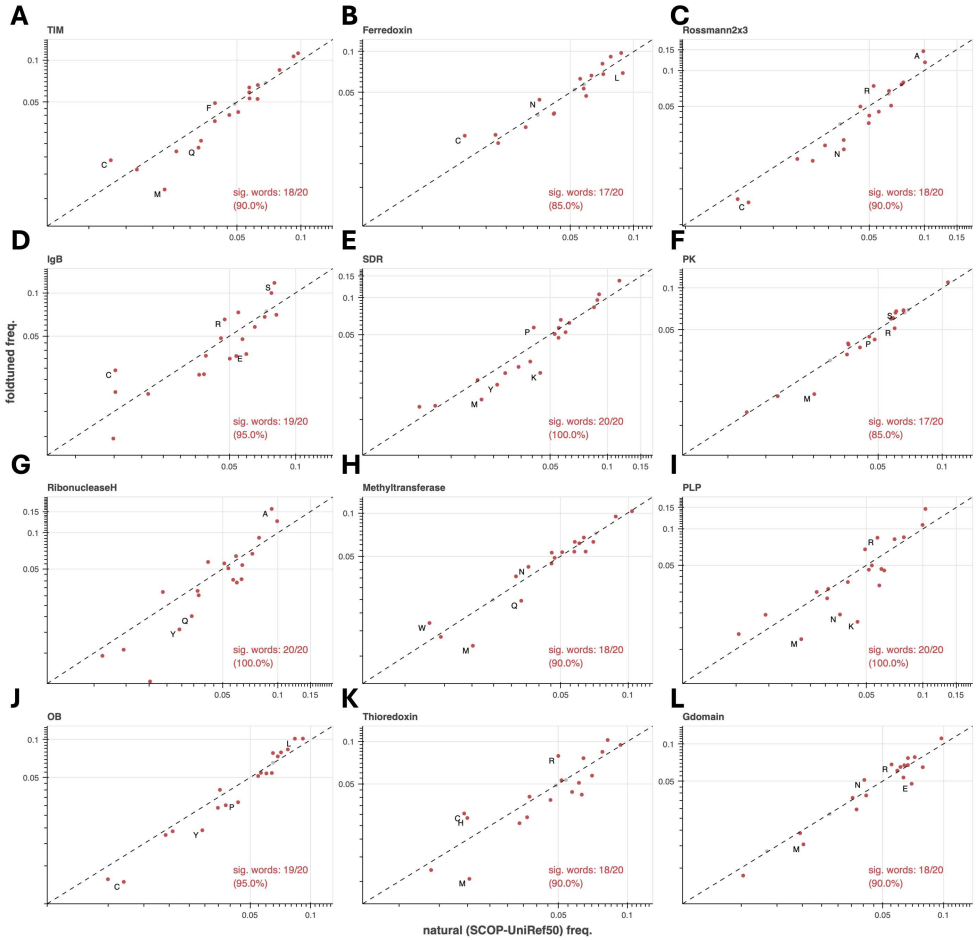



**Figure S5**: **Usage patterns of the 20 canonical amino-acids in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of AAs with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under Binyamini-Hochberg correction for positively correlated tests). The top-4 most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

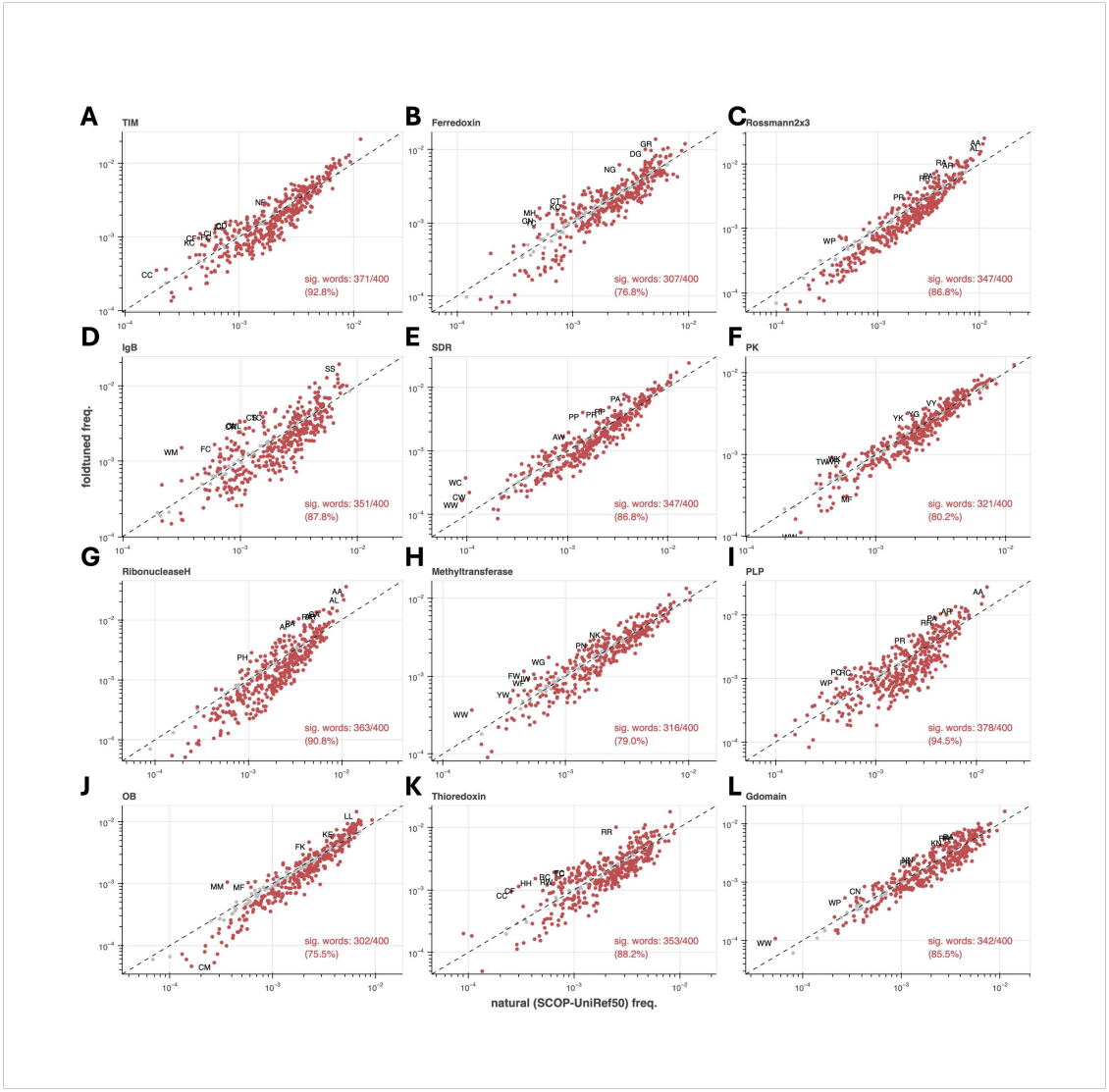



**Figure S6**: **Usage patterns of amino-acid subsequences of length 2 ("2grams", "bigrams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 2grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under Binyamini-Hochberg correction for positively correlated tests). The top-4 most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

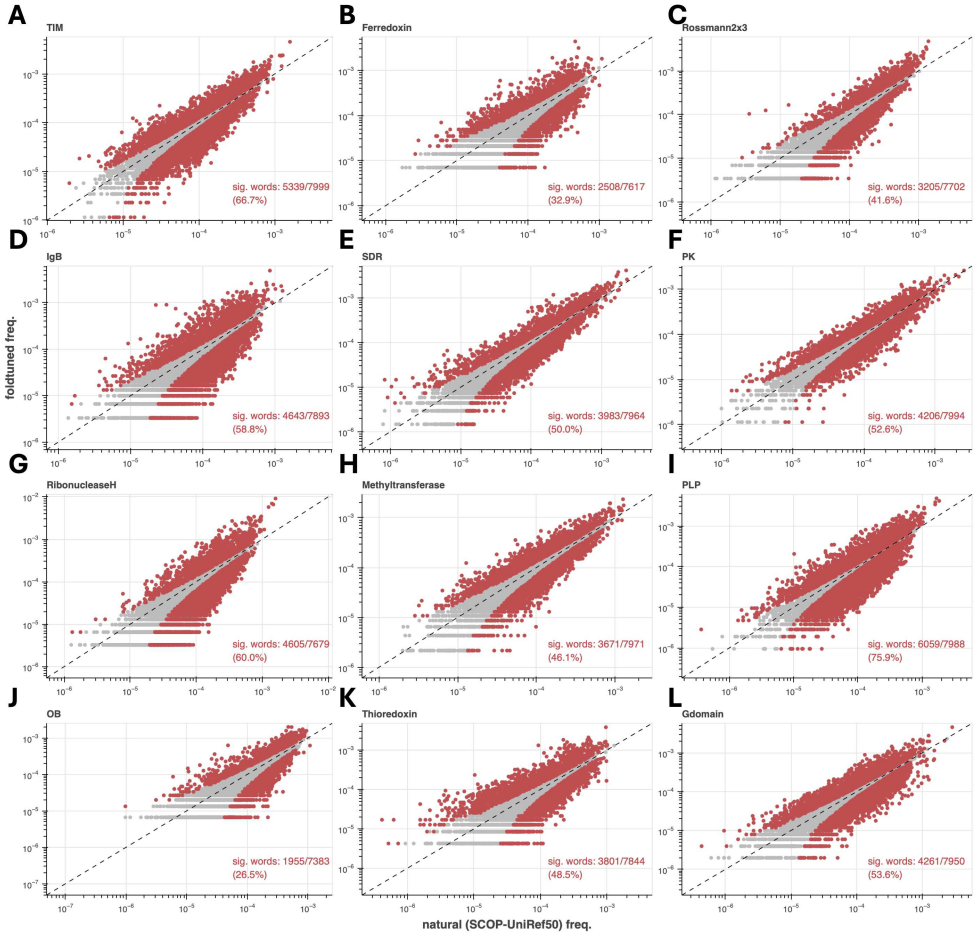



**Figure S7**: **Usage patterns of amino-acid subsequences of length 3 ("3grams", "trigrams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 3grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under Binyamini-Hochberg correction for positively correlated tests). The top-4 most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

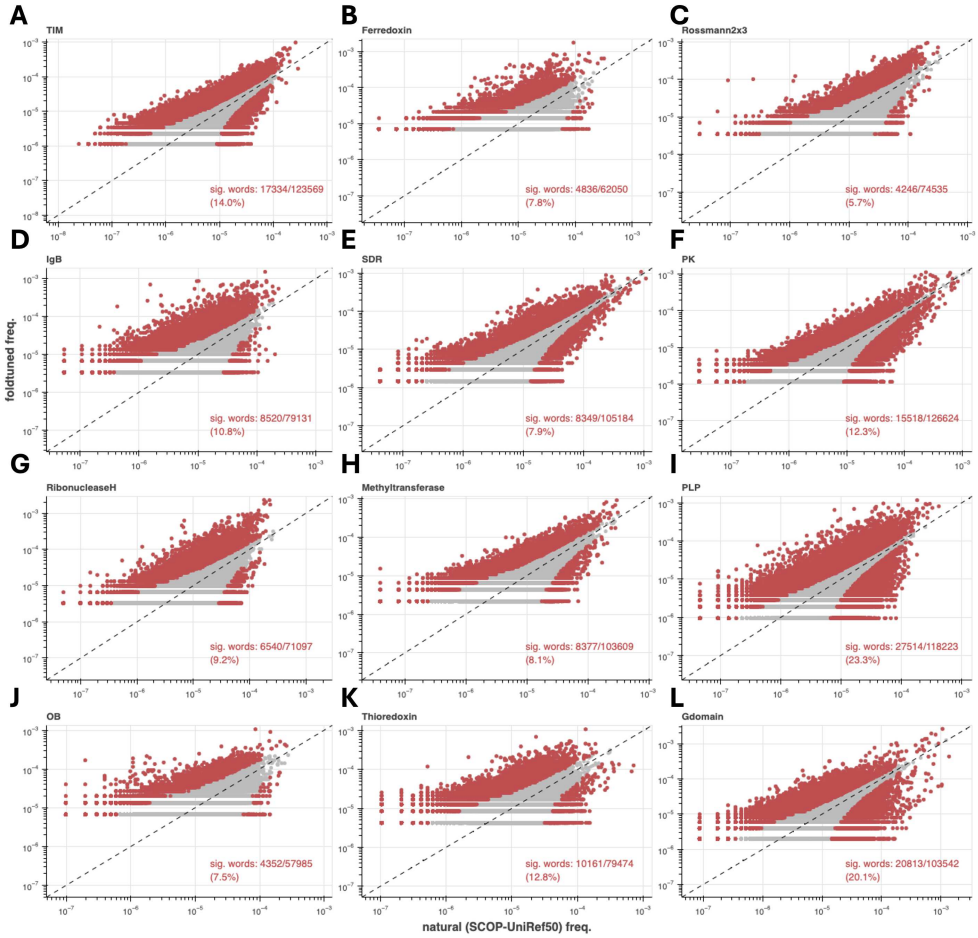



**Figure S8**: **Usage patterns of amino-acid subsequences of length 4 ("4grams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 4grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under Binyamini-Hochberg correction for positively correlated tests). The top-4 most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

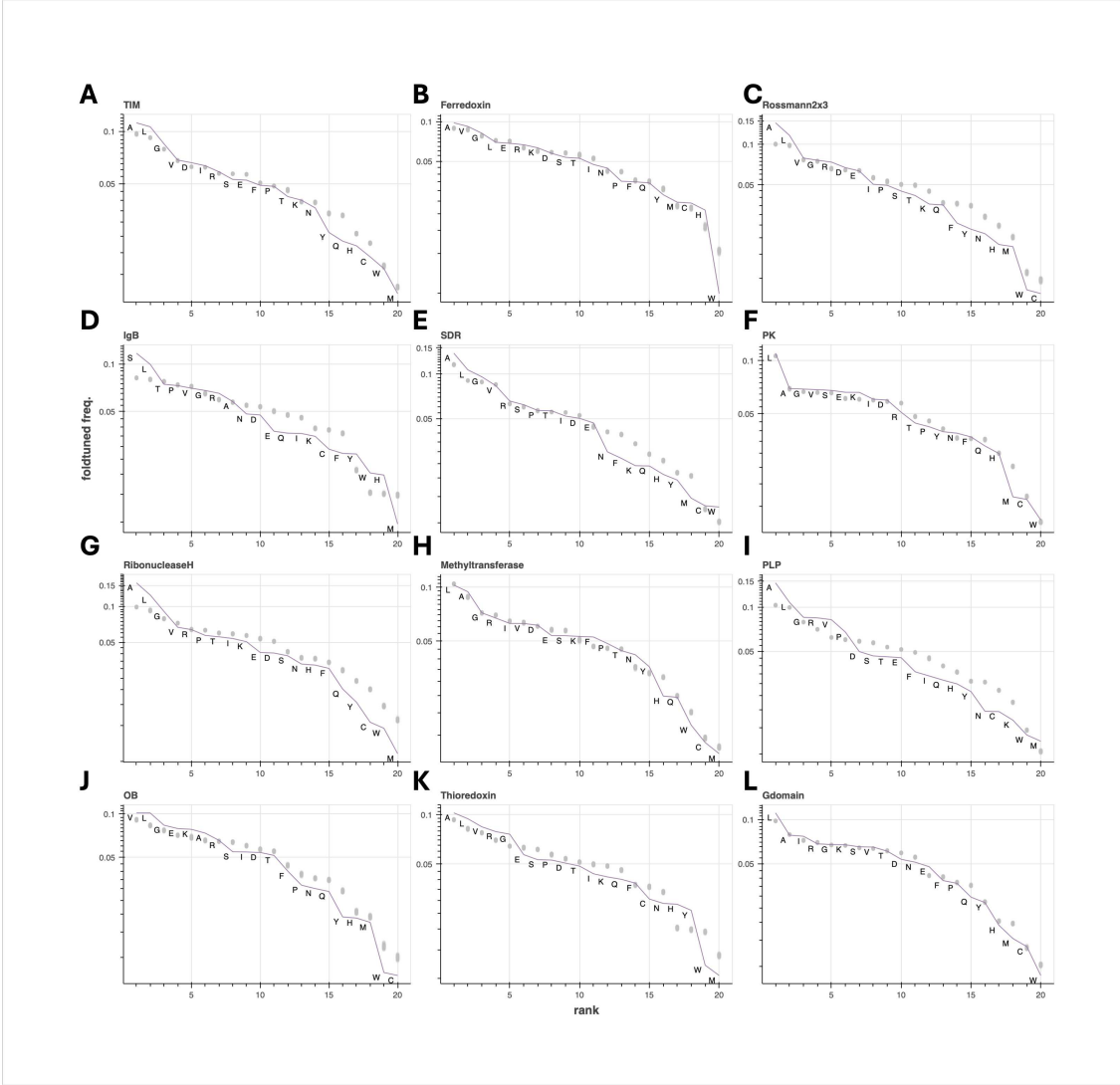



**Figure S9**: **Rank-ordered usage of individual amino-acids for foldtuned sequences (purple, labeled) and natural sequences** ($n = 1000$ **SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

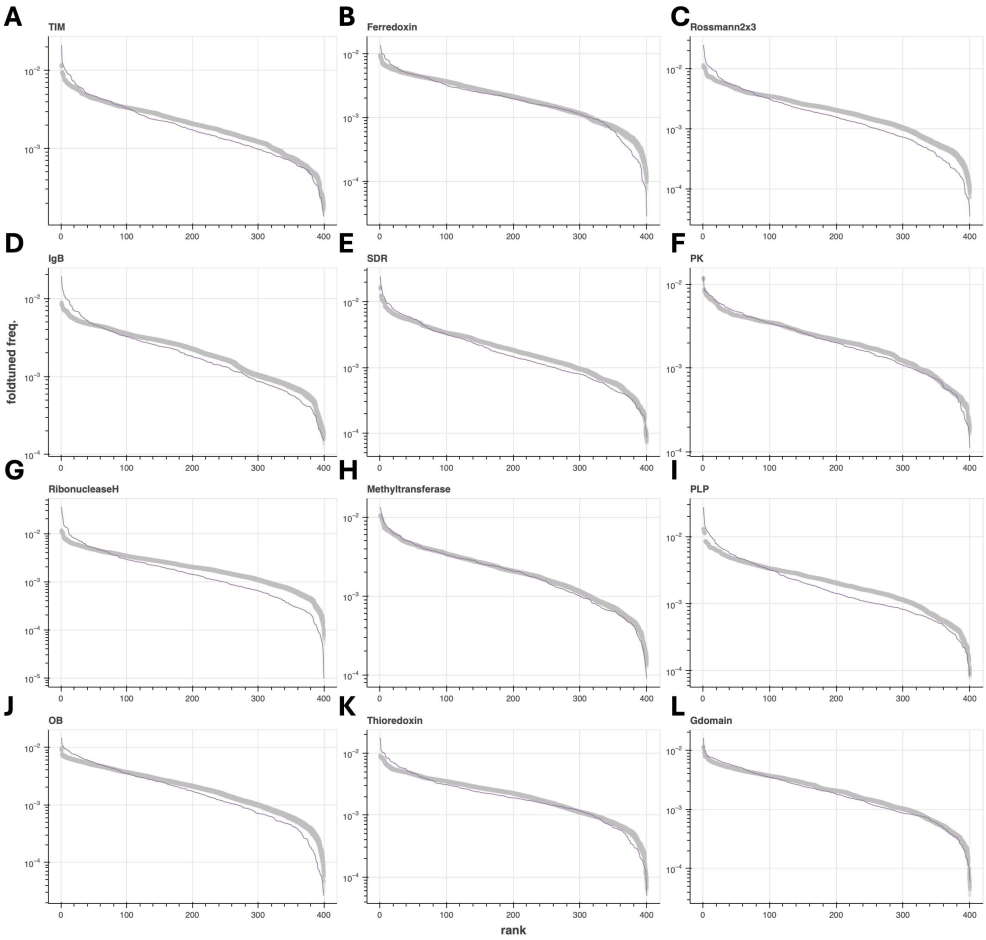



**Figure S10**: **Rank-ordered usage of subsequences of length 2 ("2grams", "bigrams") for foldtuned sequences (purple) and natural sequences** ($n$ = 1000 **SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

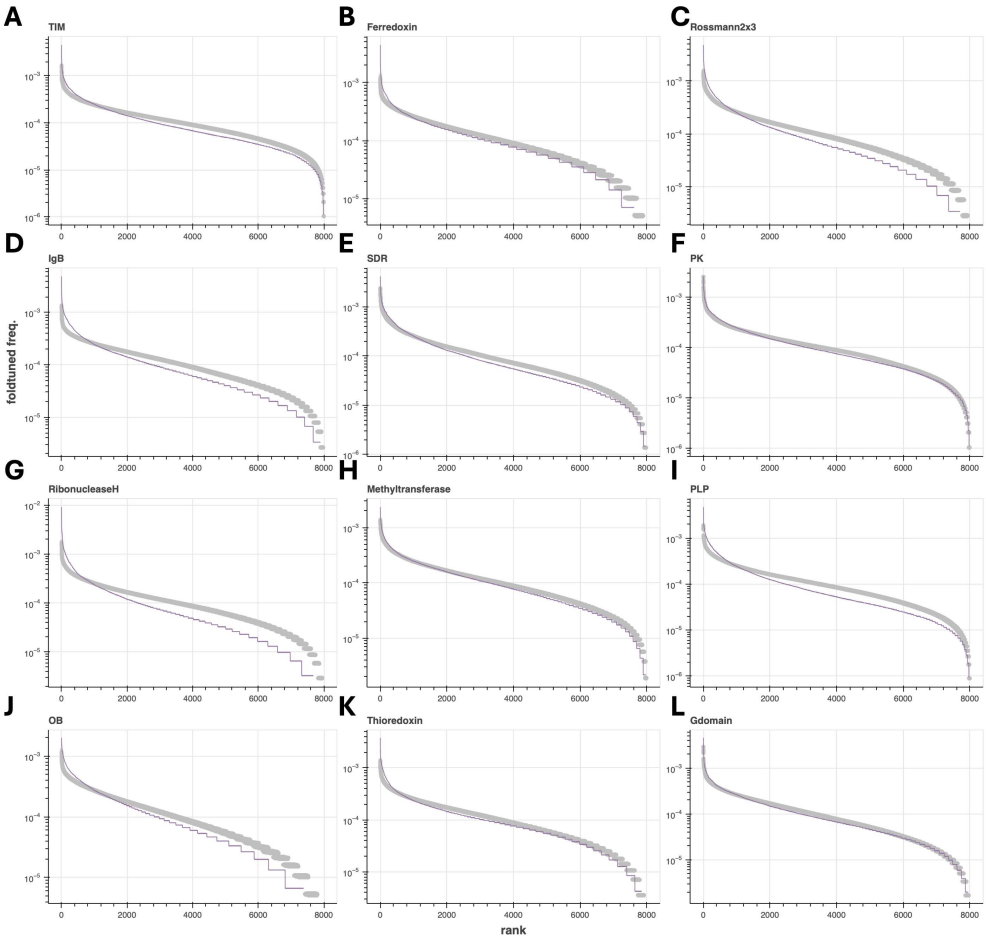



**Figure S11**: **Rank-ordered usage of subsequences of length 3 ("3grams", "trigrams") for foldtuned sequences (purple) and natural sequences ($n$ = 1000 SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

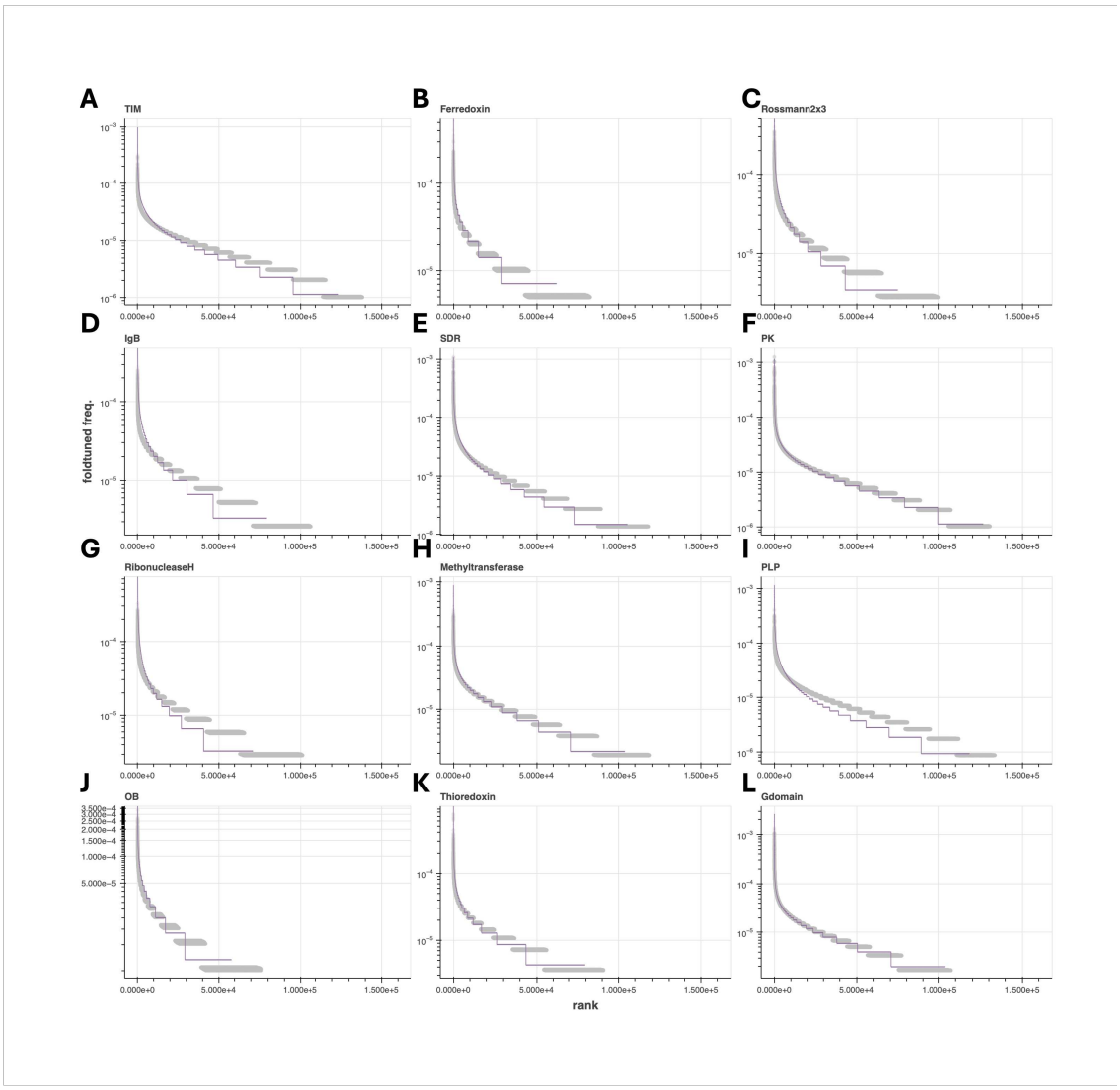



**Figure S12**: **Rank-ordered usage of subsequences of length 4 ("4grams") for foldtuned sequences (purple) and natural sequences** ($n$ = 1000 **SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM $\beta/\alpha$ barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Ig$\beta$-like. (**E**) Short-chain dehydrogenase (SDR). (**F**) Protein kinase (PK). (**G**) Ribonuclease H. (**H**) Methyltransferase. (**I**) PLP-dependent transferase. (**J**) OB fold. (**K**) Thioredoxin. (**L**) small GTPase.

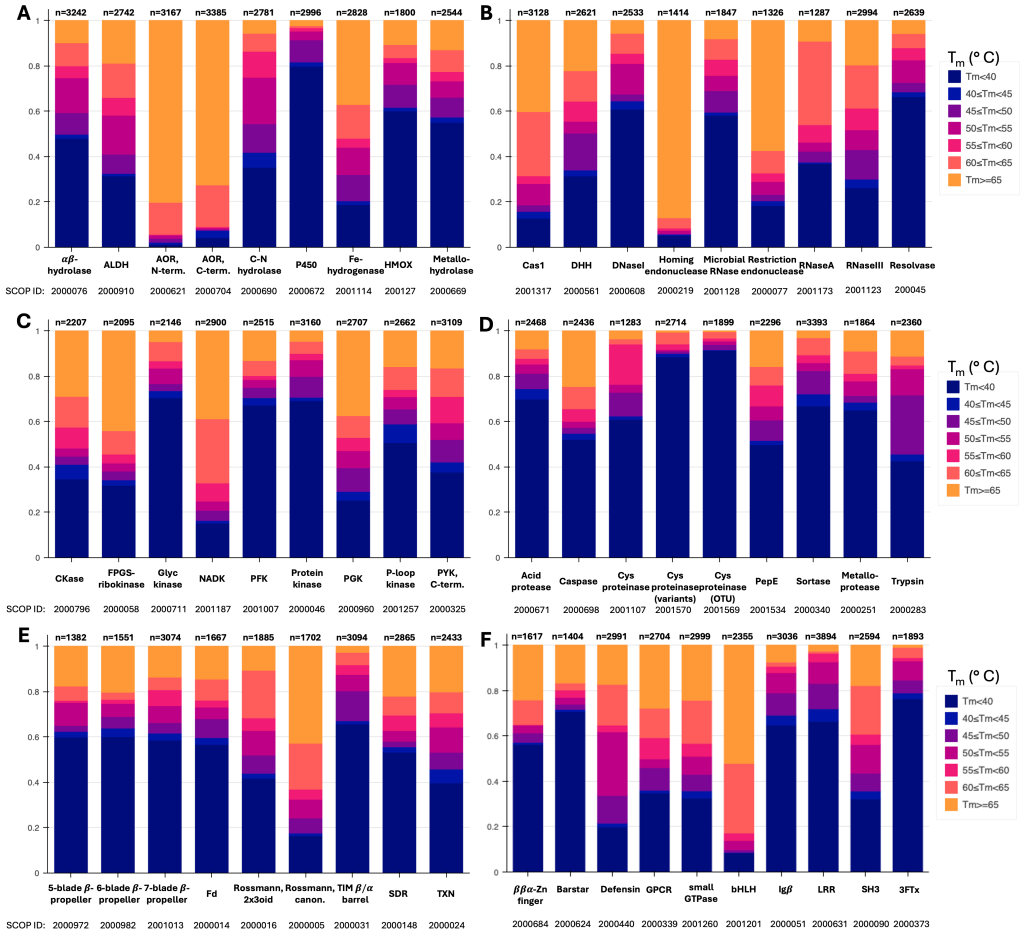



**Figure S13**: **Foldtuned proteins are predicted to exhibit varying degrees of thermostability.** Filtered, validated sequences generated from 55 foldtuned models of interest are expected to exhibit melting temperatures ($T_m$) ranging from $< 40°C$ to $> 65°C$, as predicted by TemStaPro (*40*). Selected models are grouped into: (**A**) Hydrolase and oxidoreductase enzymes. (**B**) Nucleases and other gene-editing-related proteins. (**C**) Kinases. (**D**) Proteases and peptidases. (**E**) Common topologies/scaffolds spanning multiple enzyme families. (**F**) Common synthetic biology "toolkit" parts for cellular engineering applications.

| EC # | Description | Count / % |
|---|---|---|
| 1.14.14.1 | unspecific monooxygenase | 503 / 2996 (16.8%) |
| 1.14.14.80 | long-chain fatty acid omega-monooxygenase | 345 / 2996 (11.5%) |
| 1.14.13.154 | erythromycin 12-hydroxylase | 207 / 2996 (6.9%) |
| 1.14.14.119 | fumitremorgin C monooxygenase | 113 / 2996 (3.8%) |
| 2.4.1.198 | phosphatidylinositol N-acetylglucosaminyltransferase | 96 / 2996 (3.2%) |

| EC # | Description | Count / % |
|---|---|---|
| 1.17.7.1 | (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (ferredoxin) | 290 / 2442 (11.9%) |
| 1.8.7.1 | assimilatory sulfite reductase (ferredoxin) | 213 / 2442 (8.7%) |
| 1.8.1.2 | assimilatory sulfite reductase (NADPH) | 183 / 2442 (7.5%) |
| 1.7.7.1 | ferredoxin-nitrite reductase | 183 / 2442 (7.5%) |
| 1.17.7.3 | (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (flavodoxin) | 165 / 2442 (6.8%) |

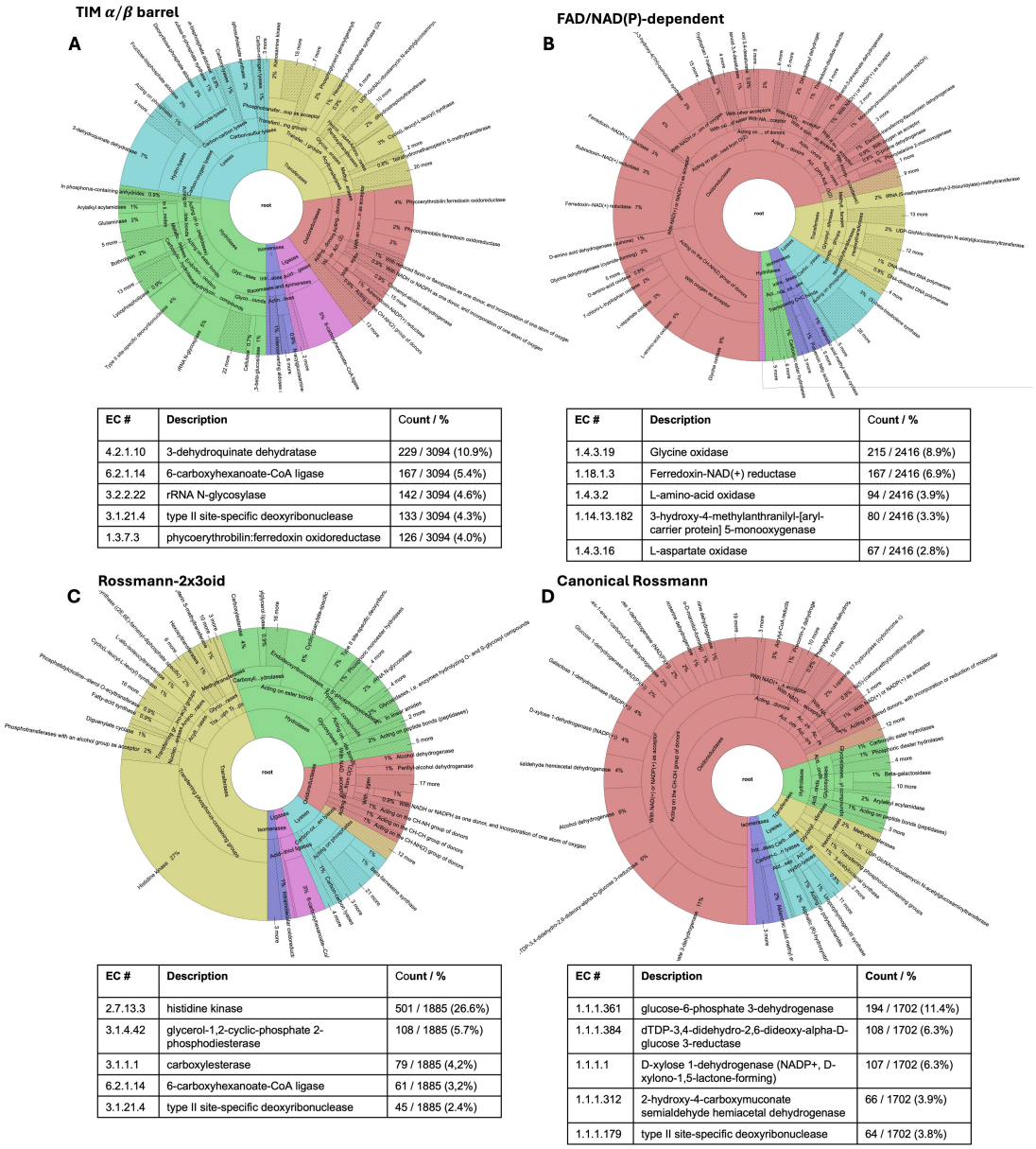
| EC # | Description | Count / % |
|---|---|---|
| 3.1.21.4 | type II site-specific deoxyribonuclease | 869 / 2996 (29.0%) |
| 3.1.12.1 | 5' to 3' exodeoxyribonuclease (nucleoside 3'-phosphate-forming) | 644 / 2996 (21.5%) |
| 3.1.26.5 | ribonuclease P | 444 / 2996 (14.8%) |
| 2.7.7.49 | RNA-directed DNA polymerase | 229 / 2996 (7.6%) |
| 2.7.7.7 | DNA-directed DNA polymerase | 219 / 2996 (7.3%) |

| EC # | Description | Count / % |
|---|---|---|
| 3.1.1.1 | carboxylesterase | 1251 / 3242 (38.6%) |
| 3.1.1.3 | triacylglycerol lipase | 261 / 3242 (8/1%) |
| 3.1.2.22 | palmitoyl[protein] hydrolase | 188 / 3242 (5.8%) |
| 3.1.1.32 | phospholipase A1 | 162 / 3242 (5.0%) |
| 2.3.2.28 | L-allo-isoleucyltransferase | 104 / 3242 (3.2%) |

| EC # | Description | Count / % |
|---|---|---|
| 2.7.11.1 | non-specific serine/threonine protein kinase | 2506 / 3160 (79.3%) |
| 2.7.10.2 | non-specific protein-tyrosine kinase | 280 / 3160 (8.9%) |
| 2.7.11.25 | mitogen-activated protein kinase kinase kinase (MAPKKK) | 142 / 3160 (4.5%) |
| 2.7.10.1 | receptor protein-tyrosine kinase | 61 / 3160 (1.9%) |
| 2.7.12.1 | dual-specificity kinase | 48 / 3160 (1.5%) |

| EC # | Description | Count / % |
|---|---|---|
| 4.99.1.3 | sirohydrochlorin cobaltochelatase | 384 / 2206 (17.4%) |
| 1.4.3.2 | L-amino-acid oxidase | 156 / 2206 (7.1%) |
| 3.1.1.1 | carboxylesterase | 102 / 2206 (4.6%) |
| 2.3.1.269 | apolipoprotein N-acyltransferase | 86 / 2206 (3.9%) |
| 3.5.2.6 | beta-lactamase | 85 / 2206 (3.9%) |

**Figure S14**: **Foldtuned proteins are predicted to mimic or expand parent enzymatic functions.** Wheel plots of predicted Enzyme Commission (EC) numbers (top 5 EC#s per fold tabulated below) for foldtuned variants of select catalytic folds, as annotated by CLEAN (*41*). Sectors are colored by top-level EC #s – oxidoreductases (EC 1; red), transferases (EC 2; yellow), hydrolases (EC 3; green), lyases (EC 4; blue), isomerases (EC 5; purple), ligases (EC 6; pink). Selected folds: (**A**) Cytochrome c P450s. (**B**) Nitrite/sulfite reductases. (**C**) CRISPR Cas1 endonuclease. (**D**) $\alpha/\beta$-hydrolases. (**E**) Protein kinases. (**F**) Chelatases.




| EC # | Description | Count / % |
|---|---|---|
| 4.2.1.10 | 3-dehydroquinate dehydratase | 229 / 3094 (10.9%) |
| 6.2.1.14 | 6-carboxyhexanoate-CoA ligase | 167 / 3094 (5.4%) |
| 3.2.2.22 | rRNA N-glycosylase | 142 / 3094 (4.6%) |
| 3.1.21.4 | type II site-specific deoxyribonuclease | 133 / 3094 (4.3%) |
| 1.3.7.3 | phycoerythrobilin:ferredoxin oxidoreductase | 126 / 3094 (4.0%) |

| EC # | Description | Count / % |
|---|---|---|
| 1.4.3.19 | Glycine oxidase | 215 / 2416 (8.9%) |
| 1.18.1.3 | Ferredoxin-NAD(+) reductase | 167 / 2416 (6.9%) |
| 1.4.3.2 | L-amino-acid oxidase | 94 / 2416 (3.9%) |
| 1.14.13.182 | 3-hydroxy-4-methylanthranilyl-[aryl-carrier protein] 5-monooxygenase | 80 / 2416 (3.3%) |
| 1.4.3.16 | L-aspartate oxidase | 67 / 2416 (2.8%) |

| EC # | Description | Count / % |
|---|---|---|
| 2.7.13.3 | histidine kinase | 501 / 1885 (26.6%) |
| 3.1.4.42 | glycerol-1,2-cyclic-phosphate 2-phosphodiesterase | 108 / 1885 (5.7%) |
| 3.1.1.1 | carboxylesterase | 79 / 1885 (4,2%) |
| 6.2.1.14 | 6-carboxyhexanoate-CoA ligase | 61 / 1885 (3,2%) |
| 3.1.21.4 | type II site-specific deoxyribonuclease | 45 / 1885 (2.4%) |

| EC # | Description | Count / % |
|---|---|---|
| 1.1.1.361 | glucose-6-phosphate 3-dehydrogenase | 194 / 1702 (11.4%) |
| 1.1.1.384 | dTDP-3,4-didehydro-2,6-dideoxy-alpha-D-glucose 3-reductase | 108 / 1702 (6.3%) |
| 1.1.1.1 | D-xylose 1-dehydrogenase (NADP+, D-xylono-1,5-lactone-forming) | 107 / 1702 (6.3%) |
| 1.1.1.312 | 2-hydroxy-4-carboxymuconate semialdehyde hemiacetal dehydrogenase | 66 / 1702 (3.9%) |
| 1.1.1.179 | type II site-specific deoxyribonuclease | 64 / 1702 (3.8%) |

**Figure S15**: **Foldtuned proteins for common enzyme scaffolds are predicted to span wide functional classes.** Wheel plots of predicted Enzyme Commission (EC) numbers (top 5 EC#s per fold tabulated below) for foldtuned variants of select broad-spectrum catalytic folds, as annotated by CLEAN. Sector coloring follows Fig. S14. Selected folds: (**A**) TIM $\beta/\alpha$ barrels. (**B**) FAD/NAD(P)-dependent enzymes. (**C**) Rossmann 2x3oid proteins. (**D**) Canonical Rossmann proteins.

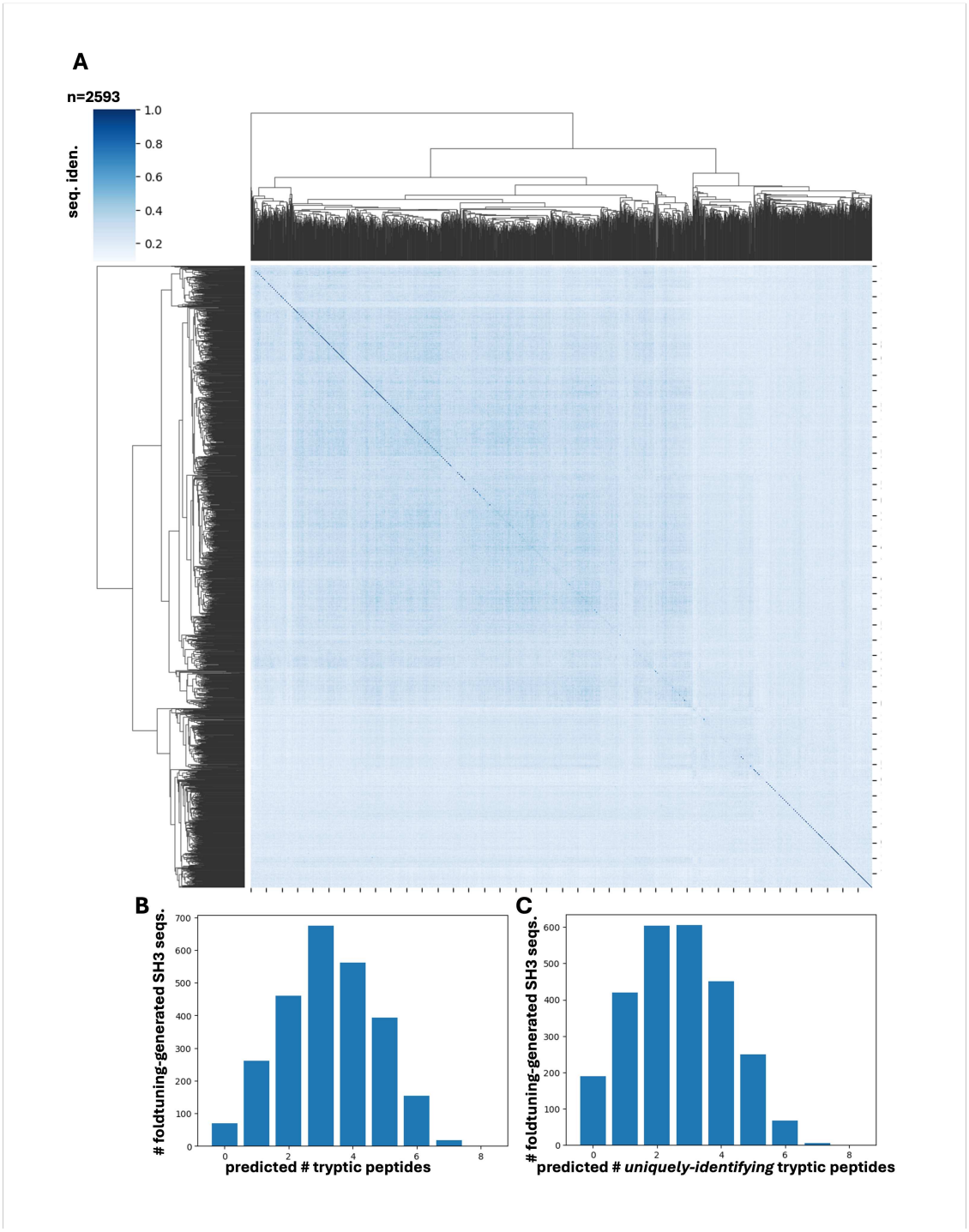



**Figure S16**: **Sequence diversity and detectability of foldtuning-generated SH3 domains.** (**A**) Hierarchically clustered heatmap of pairwise sequence identity between $n = 2593$ SH3 domain candidate sequences generated via foldtuning. (**B**) Expected detectable peptide counts predicted by in silico tryptic digestion. (**C**) Counts of predicted tryptic peptides that map uniquely to single foldtuned SH3 variants.

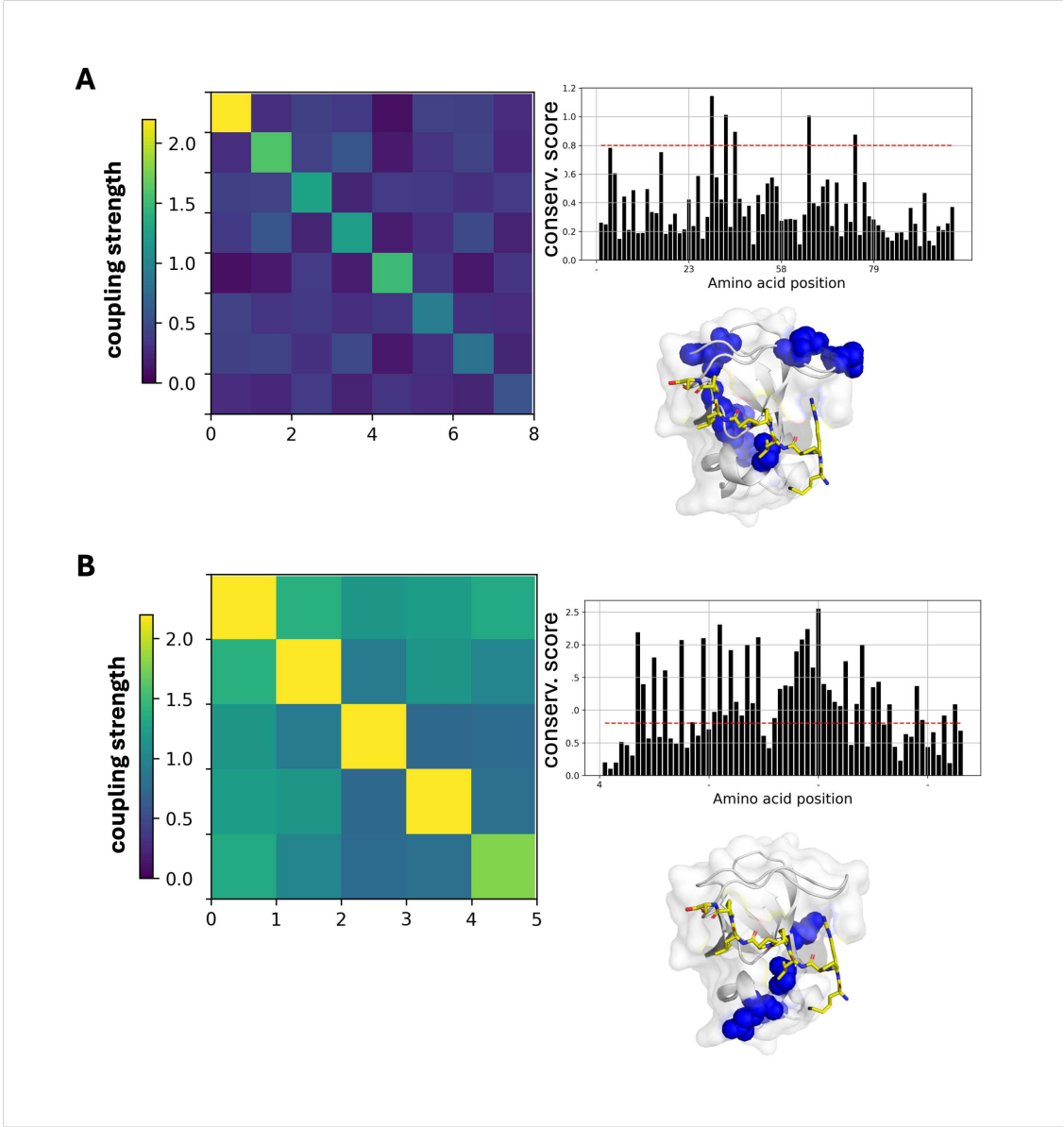



**Figure S17**: **Statistical coupling analysis of natural and synthetic SH3s.** (**A**) Results of statistical coupling analysis (SCA) on $n \approx 2500$ natural SH3 domain sequences. Left: Second-order compressed coupling matrix, blocked into a single statistically interacting sector. Top right: First-order conservation scores. Bottom right: Visualization of sector positions (blue) mapped onto a representative structure of a natural SH3 domain (from PI3K) bound to a proline-rich peptide ligand (pdb: 3I5R). (**B**) Results of statistical coupling analysis (SCA) on $n = 2593$ foldtuned SH3 sequences. Left: Second-order compressed coupling matrix, blocked into a single statistically interacting sector. Top right: First-order conservation scores. Bottom right: Visualization of sector positions (blue) mapped onto a representative structure of a natural SH3 domain (from PI3K) bound to a proline-rich peptide ligand (pdb: 3I5R)

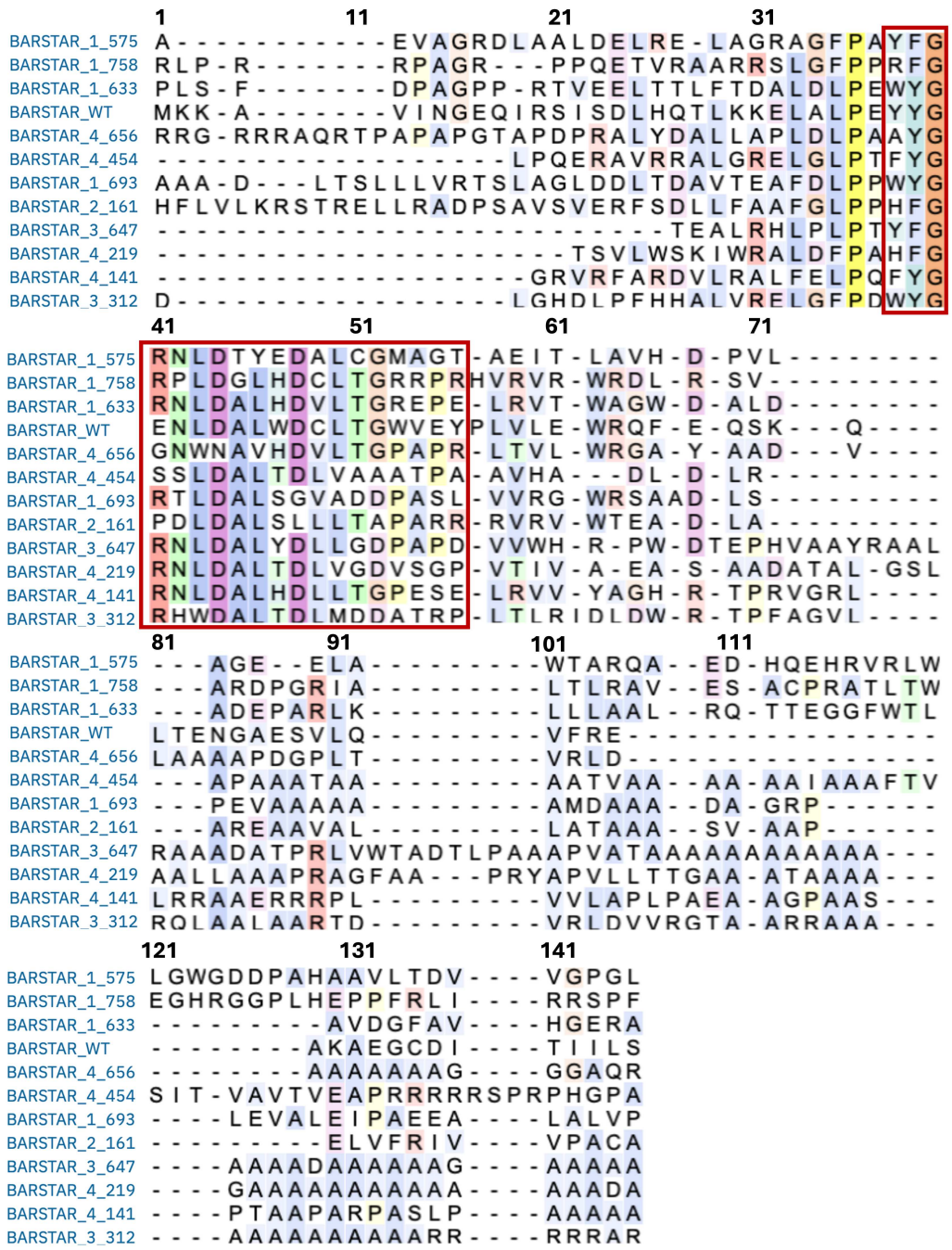



**Figure S18**: **Multiple sequence alignment (MSA) of toxicity-rescuing barstar variants.** Multiple sequence alignment (MSA) of the eleven toxicity-rescuing foldtuned barstar variants and wild-type barstar from *B. aquaforiensis*. Columns corresponding to residue positions making physical contacts (positions 38-56; distance threshold < 4.0 Å) with barnase in a reference crystal structure (PDB: 1BRS) are boxed.

**Table S1**: **Designability of SCOP folds.** Top 2% ($n$ = 14) of succesfully foldtuned SCOP folds ($N = 727$), ranked by designability proxy (structural hit rate × sequence escape rate), with topology class and structural/functional notes.

| SCOP | | | | | | |
|---|---|---|---|---|---|---|
| ID | Fold | Class | Struct. Hit Rate | Seq. Esc. Rate | Design. | Note |
| 2000062 | RH $\beta$-helix | $\beta$ | 0.881 | 0.941 | 0.829 | Periodic |
| 2000239 | Ribbon-helix-helix domain | $\alpha$ | 0.818 | 0.958 | 0.783 | DNA-binding |
| 2000031 | TIM $\beta/\alpha$ barrel | $\alpha/\beta$ | 0.770 | 0.995 | 0.766 | Symmetry (8-fold) |
| 2000920 | Anti-$\|$ $\beta/\alpha$ barrel | $\alpha+\beta$ | 0.743 | 0.996 | 0.740 | Symmetry (5-fold) |
| 2000619 | $\alpha/\alpha$ toroid | $\alpha$ | 0.704 | 0.994 | 0.700 | Periodic |
| 2000193 | Transmembrane $\beta$-barrel | $\beta$ | 0.731 | 0.955 | 0.698 | Symmetry (various) |
| 2000308 | Sm-like fold | $\beta$ | 0.741 | 0.889 | 0.659 | RNA-binding |
| 2000440 | Defensin | n/a | 0.625 | 0.998 | 0.624 | Antimicrobial |
| 2000144 | Winged helix domain | $\alpha+\beta$ | 0.720 | 0.860 | 0.619 | DNA-binding |
| 2000087 | POU domain | $\alpha$ | 0.664 | 0.920 | 0.611 | DNA-binding |
| 2000419 | Pentein $\beta/\alpha$ propeller | $\alpha+\beta$ | 0.624 | 0.954 | 0.595 | Symmetry (5-fold) |
| 2000501 | DNA clamp | $\alpha+\beta$ | 0.658 | 0.895 | 0.589 | DNA-binding |
| 2000114 | Histone fold | $\alpha$ | 0.617 | 0.953 | 0.588 | DNA-binding |
| 2001248 | RecA-like basic | $\alpha/\beta$ | 0.724 | 0.807 | 0.584 | DNA-binding |