Chapter 3

FOLDTUNING TURNS PROTEIN LANGUAGE MODELS INTO STRUCTURE-GUIDED SEQUENCE PROBES

3.1 Introduction

Sampling sequence novelty one fold at a time

In the preceding chapter, we uncovered a pronounced pathology of protein generation with protein language models (PLMs) — a tendency to visit only a limited and unrepresentative collection of structural modes. Generating under conditions meant to prioritize sequence novelty in the search for functional proteins outside the constraints wrought by evolution on Earth only heightens this collapse. In this chapter, we leave this conundrum behind and show how to sample the extremes of permissible sequence novelty while maintaining fidelity to informative guidepost structures, balancing large perturbations to sequence against small perturbations to structure and transforming PLMs into structure-preserving probes of far-from-natural sequence-space. Much stands to be gained by seizing the ability to systematically locate stable, functional proteins that reconstitute known structural motifs but lie in regions of sequence-space with no meaningful similarity to nature. From unlocking expanded repertoires of binding partners, signaling interactions, and substrate scopes for synthetic biology, to revealing key amino-acid sequence rules and constraints undergirding the fundamental biophysics of molecular machines, the as-yet-unrealized design capacity of these outlying zones to tile the functional landscape is tantalizing. Although it has been speculated that such an effort might be accomplished by rationally painting individual interactions and sequence motifs onto a pre-specified backbone template, and accomplished for small folds using physics-based methods, scaling to (1) large folds, (2) full coverage of natural structure-space, (3) truly novel sequences lacking detectable similarity to nature, and (4) sufficiently large variant libraries for functional screening and design rule elucidation — has not come to pass (Dahiyat and Mayo, 1997; Pabo, 1983).[1]

The problem before us then, is to find a search strategy that mines these "döppelganger" proteins from the junk and gibberish that presumably occupies much of the

[1]This explanation deliberately dances around referring to this task by its original name, the "inverse-folding problem," to avoid confusion with the AI-based inverse-folding models for protein sequence design that have increasingly co-opted the term and that we encountered in Chapter 2.

combinatorial vastness of sequence-space. PLMs, with their apparent exploratory capacity — born out in occasional design efforts in the literature and in the immense sequence diversity encountered in Chapter 2 — are natural vehicles for this task (Verkuil et al., 2022). The main obstacle, however, is the structural collapse likewise seen in Chapter 2, the downside of choosing novelty (in sequence) over breadth (in structure). Indeed, this tradeoff between breadth and novelty is considered a general property of LLMs — not just protein models — exemplified by recent theoretical results holding that a language model can sample "in the limit" of all valid texts, beyond yet consistent with its training data, but at the cost of a marked reduction in output diversity (Kleinberg and Mullainathan, 2024). Instead of accepting arbitrary model-induced breadth reduction, we choose the form of this limitation to our advantage. For the protein mimic search problem, we elect to set an anchor structure as a PLM's sole target, directing its exploration-by-generation ability to push the accepted bounds of sequence towards the far-from-natural and sample "in the limit" of meaningful (and functional) sequences encoding a fold of interest, one target at a time.

A novel algorithm for structure-oriented PLM generation "in the limit"

Thus, we envision an approach that takes one target fold at a time — a *family* of related structures, not a single example, so as to retain degrees of freedom for structural and functional innovation — and uses it to force a PLM to push outwards from its familiar training data distribution towards the limit of valid sequences that respect the same underlying biophysical logic and language rules. Deploying such an approach requires three features. The first is a way to decide whether a structure (predicted from a generated sequence) is sufficiently close to the target family to be structure-preserving, versus drifting into structural breakdown and disorder. This can be implemented straightforwardly by feeding generated sequences to a predict-search-assign protocol that links structure prediction models to structural alignment tools to separate valid structural matches from invalid ones. The second is a way to measure whether and how quickly generated sequences are moving in a fruitful direction in sequence-space — towards far-out corners and true novelty — and not towards, say, evolutionarily distant natural analogs. Given that we hope to reach deep into sequence-space, the usual bioinformatic parameters like % identity, bit-score, and E-value are of little help — upon surpassing the bounds of detectable sequence homology to natural proteins they become no longer calculable — evidence, yes, of escaping nature's gravitational pull, but without any sense of by how much. To

quantify sequence novelty in a more informative way, we borrow the concept of semantic change from computational linguistics and natural language processing. Qualitatively, semantic change can be understood as capturing the "displacement in meaning" between two texts that are equally grammatically valid with respect to some governing language; quantitatively it is a distance metric evaluated in a model's high-dimensional latent-space — all sequences can be mapped to embedding vectors, meaning that semantic change can always be computed, irrespective of homology or lack thereof. It should also be noted that when evaluated on real sequences, higher semantic change can correlate with substantive differences in function and binding, e.g. as in antigenic escape (Hie et al., 2021).

Finally, and most crucially, we must link these two arms together. Ready inspiration exists in the machine-learning world in the form of generative adversarial networks (GANs). The traditional picture of a GAN pits two models — a generator and a discriminator — against each other in a so-called counterfeiting game (Goodfellow et al., 2014). The generator's goal is to spit out artificial data (fakes) that go uncaught by the discriminator; the discriminator's goal is to detect all the fakes. Over many rounds of the game, the generator learns to make its output look more like the fakes that got past the discriminator and less like the ones that were stopped. Meanwhile, the discriminator similarly learns from its successes and mistakes to get better at spotting the subtle features separating artificial data from real. In our problem, the PLM becomes the generator, feeding artificial sequences to the predict-search-assign procedure as the discriminator, with the PLM learning to make its later-round outputs more closely resemble those self-generated artificial sequences that "trick" the discriminator by (1) matching the target fold and (2) taking large steps away from nature as tracked by semantic change.

These are the broad strokes of a new algorithm, a PLM-powered engine for making massive-scale sequence perturbations that leap between outlying pockets of structured, sensible proteins populating deep sequence-space. As this method considers a single target fold at any one time and iteratively updates ("finetunes" in LLM parlance) its PLM generator with batches of high-quality synthetic sequence data, we dub it "foldtuning," a portmanteu of "target-*fold*" and "fine*tuning*." We successfully apply foldtuning to 727 targets spanning topologies, functions, and synthetic biology applications, in the process gleaning preliminary insight into features distinguishing easy-to-build targets from more recalcitrant ones. With a battery of *in silico* tests we show that foldtuning preserves the contours of a structural family, maximizes se-

quence novelty as measured by both semantic change and traditional bioinformatics criteria, and proposes thermodynamically-plausible variants with wide-ranging expected functions. Additionally, we unpack how foldtuning samples small structural innovations, expanding the potential scope of downstream engineering campaigns; we also discover shifted amino-acid usage patterns, strongly implying that foldtuned models not only master the signature languages of different protein folds, but experiment with distinct and novel candidate "fold languages" as well. Taken all together, these findings underscore that there is much to be gleaned from deep protein-space in theory and in practice, with foldtuning positioned to drive further exploration and illumination of the sequence-structure map.

3.2 Results & Discussion

Foldtuning: 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. 3.1A. The PLM of choice is first finetuned on natural protein fragments 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 (Madani et al., 2023).[2] 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. 3.1B-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*.[3] 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-

[2] Depending on the target, fragments are drawn either from the custom SCOP-UniRef50 database whose construction was described previously in Section 2.4 or from InterPro entry-associated PDB metadata as described in Section 3.4.

[3] Contrast with a "hard" constraint requiring a small RMSD over the entire backbone, the objective of less-exploratory models like structure→sequence inverse-folding models.

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 suggest 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. 3.1D, Fig. S3.1- S3.2). In this way, 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.

Choosing 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 (out of a collection of 44 target entries; a 43.2% success rate). 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 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. 3.2A). 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. 3.2A). 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. S3.3). 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. 3.2B). Notably, at least up to four rounds, foldtuning does not display any significant tradeoff between structural hit rate and sequence

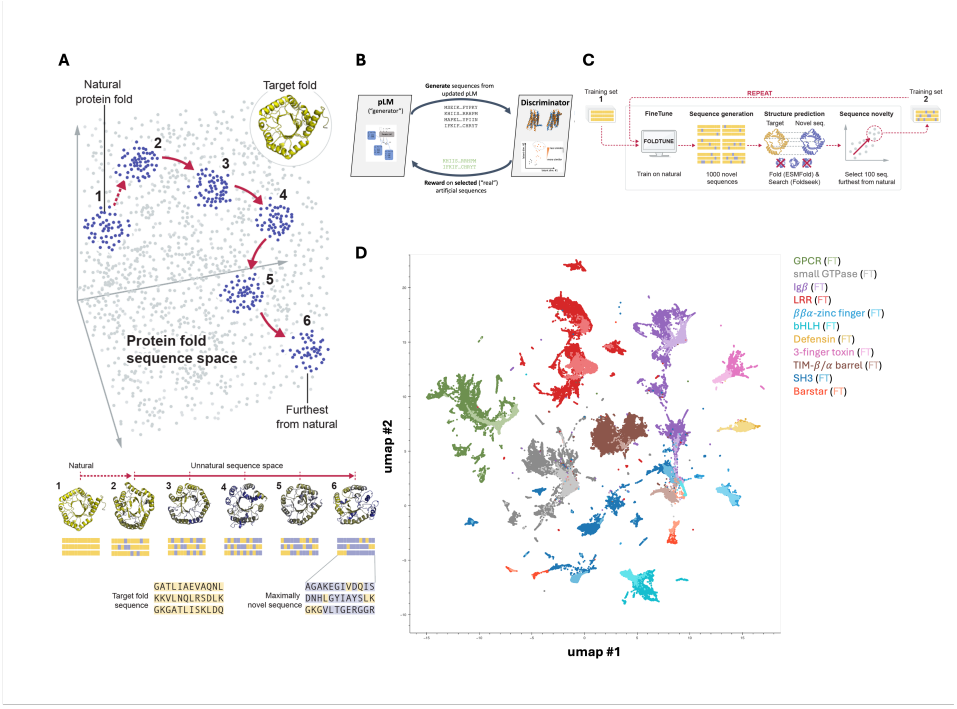


Figure 3.1: **Foldtuning explores far-from-natural sequences encoding alternate versions of natural protein structures.** (**A**) Conceptual overview of foldtuning. Beginning from natural protein sequences coding for a target backbone structure, foldtuning uses a protein language model (pLM)-based strategy to probe outwards in sequence-space, detecting subpopulations that maintain the target backbone while progressively decreasing sequence similarity to the closest natural example. (**B**) Conceptual overview of foldtuning *architecture*, which alternates in a closed-loop between sequence generation and discrimination/selection rounds, roughly analogous to a generative adversarial network. (**C**) Detailed schematic of the foldtuning *workflow*. For a provided backbone target fold, a pLM is initially finetuned (**1**) on examples from structural mining of UniRef50. In each subsequent round of foldtuning, artificial sequences are generated from the current pLM state and filtered for target backbone matching based on ESMFold structure prediction and Foldseek structure-based search (TMalign mode; tmscore cutoff threshold of 0.5); the pLM is then updated by finetuning on those filtered matches that maximize semantic change relative to the natural training examples (**2**). (**D**) 2D UMAP representation of ESM2-650M embeddings of natural (dark) and foldtuned (light) sequence examples for eleven representative target fold classes.

escape rate. In many cases, these metrics can be simultaneously maximized (e.g. TIM β/α barrels, Igβ-like domains); in others, a substantial leap in sequence escape rate — the more critical mark given that sequence novelty at scale is the main goal of foldtuning — can be gained with a minimal drop in structural hit rate (e.g. Ferredoxins, Rossman(2x3)oids) (Fig. 3.2C).

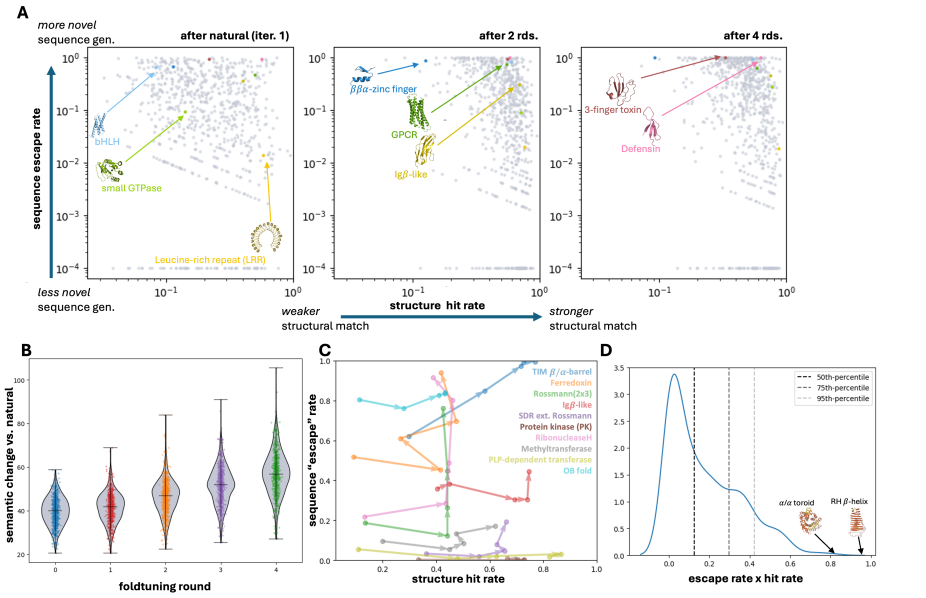


Figure 3.2: **Foldtuned models sample novel sequences for >700 targets.** (A) Sequence escape vs. structural hit rates after natural-only evotuning or two or four rounds of foldtuning for 727 targets. Selected structural/functional targets are highlighted: transcription factors (blue), GPCRs/small GTPases (green), cell surface receptor domains (gold), and small antimicrobial/toxin proteins (red). (B) Semantic change, defined as the minimal L1-norm between the ESM2-650M embeddings of a generated sequence and any sequence in the natural training set, increases with additional rounds of foldtuning. (C) Over up to four rounds of foldtuning, structural hit and sequence escape rates are generally maximized simultaneously without explicit conditioning. (D) Target folds are ranked by sequence "designability," taking the product of structural hit and sequence escape rates as a proxy.

Consistently high structural hit rates, increasing sequence escape rates, and the absence of a tradeoff between the two, together strongly imply that foldtuned models are, as intended, operating as probes that move away from nature and locate patches of viable far-from-natural protein density in sequence-space, all without veering off into regions of garbage. The high structural hit rate / high sequence escape rate regime points to one more interesting feature. 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 β-helix, ribbon-helix-helix (RHH) domain, TIM β/α-barrel, anti-parallel β/α (PT) barrel, and α/α toroid are ranked as the most designable SCOP

motifs, followed by transmembrane β-barrels, Sm-like barrels, defensins, the winged helix domain, and the POU domain (Fig. 3.2D, Table S3.1). 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 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-α 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.[4] 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. S3.4).

Foldtuning explores new sequence rules and populations

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 (Igβ-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_α coordinates vs wild-type (Fig. 3.3A-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. S3.3D). 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. S3.3G). It should also be noted that (1) this apparent sequence identity

[4] As discussed at length in Chapter 2.

barrier for foldtuned immunoglobulins still represents a leap in sequence novelty inaccessible to purely experimental approaches and equivalent to separation over enormous evolutionary timescales, and (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. 3.3C-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 fractions 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.[5] Foldtuning, then, is exploring new semantics at the whole-sequence level; to understand how models reach this point we must consider how foldtuned sequences are assembled from shorter local motifs.

To do so, 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 16,000 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. S3.5- S3.8). 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. S3.5). For $n = 2$, 79.0-94.5% of dipeptide "words" shift (Fig. S3.6). For $n = 3$, 26.5-75.9% of tripeptides shift (Fig. S3.7). 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

[5]Network node counts and total variant counts are not identical due to a necessary preclustering step preceding all-against-all alignment; refer to Section 3.4 for further information.

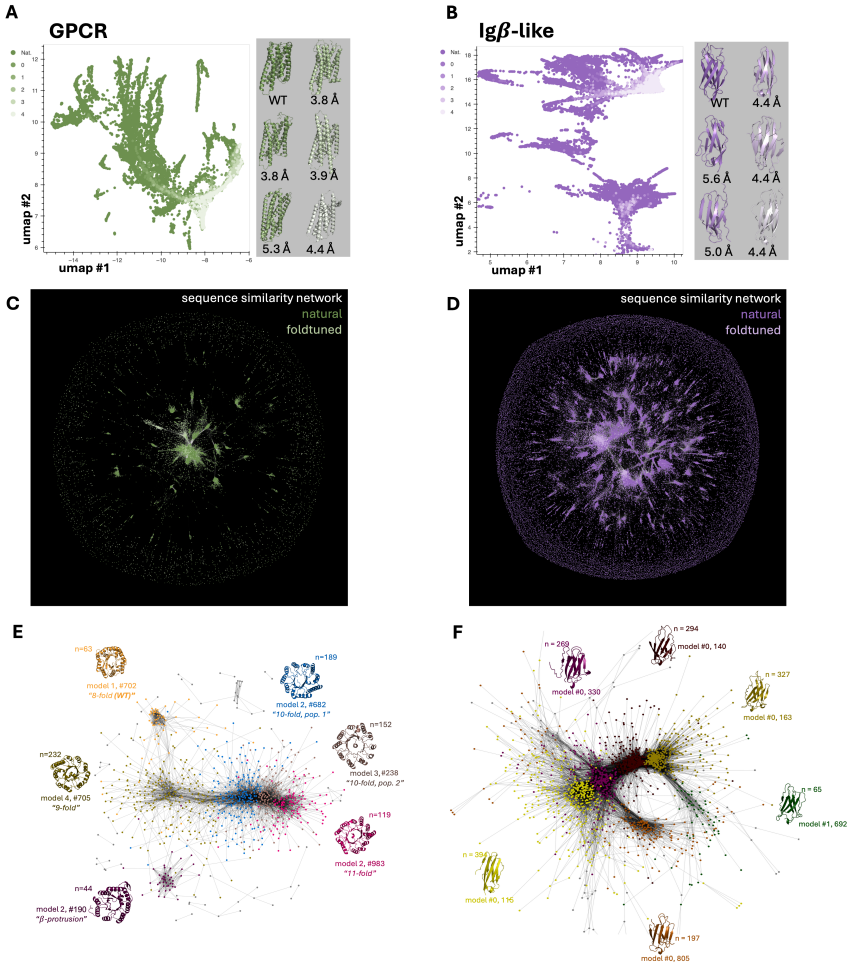


Figure 3.3: **Foldtuning accesses new sequence populations and structural innovations while 'fuzzily' preserving a target backbone.** (**A**) UMAP of round-by-round foldtuning sequence diversification captured by ESM2-650M final-layer hidden states with G-protein coupled receptors (GPCRs, SCOP ID: 2000339) as the target structure. (**B**) Same as (**A**), with Immunoglobulin-like domains (Igβ-like, SCOP ID: 2000051) as the target structure. (**C**) Network representation of similarity between natural (dark green) and foldtuned (light green) GPCR sequences. (**D**) Same as (**C**), with Igβ-like domains as the target structure (natural: dark purple, foldtuned: light purple). (**E**) Network representation of structural similarity between foldtuned TIM β/α barrel (SCOP ID: 2000031) sequences; node coloring reflects Louvain clustering assignments with cluster-representative structures color-coded accordingly. (**F**). Same as (**E**), with Igβ-like domains as the fold class.

"words" shift in one direction or the other (Fig. S3.8) (Islam et al., 2018).[6] In one

[6]For $n = 4$ in particular, shift percentages will chronically *underestimate* the degree to which

sense, the variation in the extent of vocabulary shifts from fold to fold highlights different degrees of attainable sequence relaxation and manipulation. In another, the substantial shift magnitudes 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. S3.9- S3.12).

Foldtuning is an implicit innovator of structure and function

As a PLM-based method, foldtuning only directly interfaces with and generates sequence data. However, over the four rounds of foldtuning, without any explicit structural direction aside from the TMscore-based filtering and validation steps, we notice that 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). The TIM β/α -barrel fold is a particularly sharp example of the latter. The TIM 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. 3.3E). 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 β-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, the same structural clustering procedure picks out six clusters as well; here, the main distinctions between the clusters are relative orientations of the two β-sheets in the Igβ-like

subsequence composition changes across all n-grams, as $20^4 = 160,000$ (# of possible 4-grams) $\approx (f \times 10^3) \times 10^2 = f \times 10^5$ (approximate # of total subsequences of length 4 in a collection of foldtuned variants). A 4-gram that is observed in natural sequences but not in foldtuned ones is not considered to be shifted, resulting in a "zero-deflating" effect on the overall vocabulary shift percentage.

sandwich and loop packing (Fig. 3.3F).

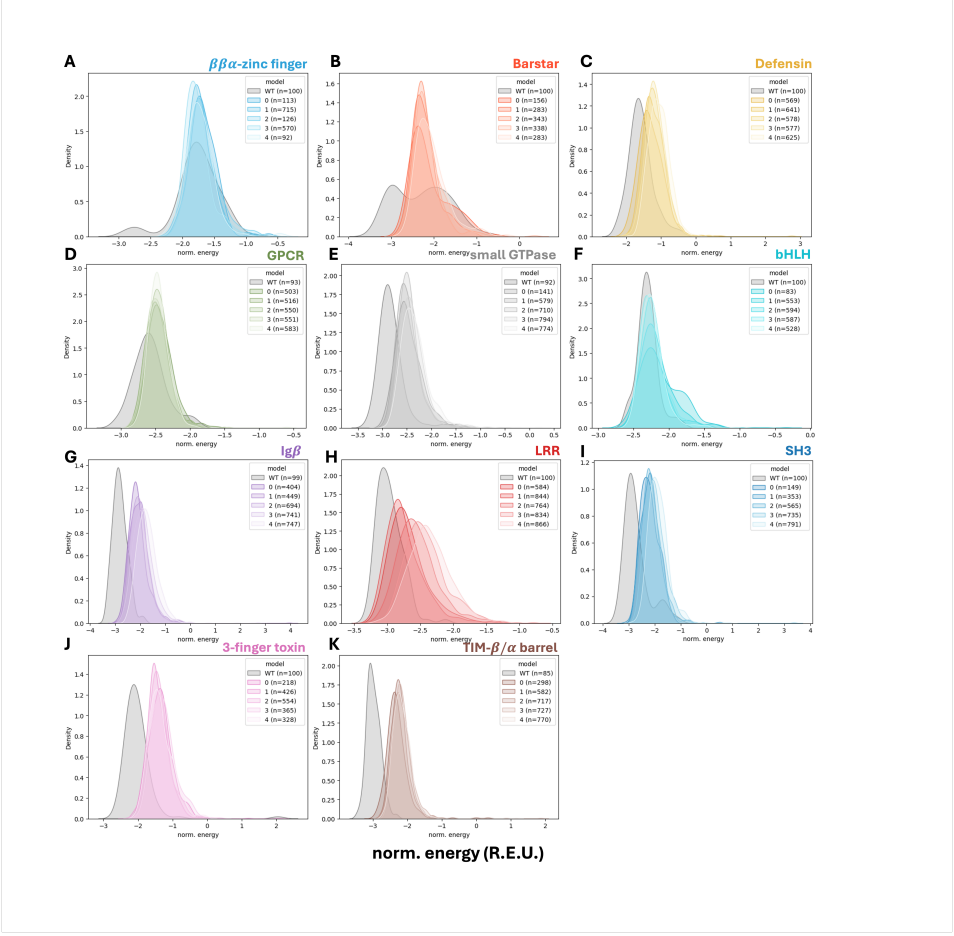


Figure 3.4: **Rosetta energies of foldtuned vs natural variants.** Histograms of length-normalized (REU / residue) Rosetta energy estimates for foldtuned (colored) and natural (gray) variants following standard backbone relaxation. Selected folds: (**A**) $\beta\beta\alpha$-zinc finger. (**B**) Barstar. (**C**) Defensin. (**D**) G-protein coupled receptor (GPCR). (**E**) Small GTPase. (**F**) Basic HLH transcription factor (bHLH). (**G**) Immunoglobulin β-sandwich (Igβ). (**H**) Leucine-rich repeat (LRR). (**I**) SH3 domain. (**J**) Three-finger toxin domain (3FTx). (**K**) TIM β/α barrel.

Given the significant sequence perturbations and shifts in motif usage achieved by foldtuning, not to mention the multiple scales of structural exploration, we evaluate the physical plausibility of 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 (normalized to sequence length and reported in arbitrary **R**osetta **E**nergy **U**nits, or REUs) for all filtered and validated foldtuned variants and compare to $n = 100$ natural training examples (Fig. 3.4). 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 (Alford et al., 2017). However, examined relative to natural counterparts, the picture is not quite as rosy for all foldtuned populations. Several targets — $\beta\beta\alpha$-zinc fingers, barstar-like proteins, defensins, GPCRs, small GTPases, helix-loop-helix (HLH) domains — do produce energy estimate distributions substantially overlapping those of wild-type examples. Other distributions — for immunoglobulin domains, SH3 domains, 3-finger toxins, and TIM barrels — appear shifted towards lower stability compared to natural versions. Additional rounds of synthetic data feedback to foldtuning may also drive a shift towards lower stability — aside from the leucine-rich repeat (LRR) fold, however, the magnitude of this effect is quite small. 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 mediators of catalysis and protein-protein interactions — were scored with a PLM-based thermostability predictor (Fig. S3.13) (Pudžiuvelytė et al., 2024). 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.

Finally, we briefly consider another level of the protein universe that foldtuning only deals with implicitly, the level downstream of structure, namely that of function. As with stability, we want to verify — to the speculative extent that is possible computationally — that foldtuned proteins recapitulate, or perhaps 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 (Yu et al., 2023). For families with established reactivities and mechanisms, top-level EC #s[7] are largely predicted as expected — P450s and nitrite/sulphite reductases are assigned as oxidoreductases, CRISPR Cas1s and α/β hydrolases are assigned as hydrolases, protein kinases are assigned as transferases, and chelatases, albeit less cleanly, are assigned as lyases and ligases, covering their multiple roles in cofactor biosynthesis (Fig. S3.14). Past the top-level, significant fractions of foldtuned enzymes are annotated into categories asociated with evolv-

[7]Top-level EC #s map to functions as follows: Oxidoreductases (1), transferases (2), hydrolases (3), lyases (4), isomerases (5), ligases (6).

ability and and promiscuous activity against a broad spectrum of substrates. For example, nearly one-in-five foldtuned P450s is placed in EC 1.14.14.1, the catch-all "unspecified monooxygenase" category from which xenobiotic-metabolizing enzymes tend to emerge. Similarly, foldtuned versions of CRISPR Cas1 — a metal-dependent non-site-specific DNA-specific endonuclease — are sometimes labeled instead 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 (Taverna and Goldstein, 2002). 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, perhaps presaging utility in designing bespoke signaling networks. Foldtuned versions of common scaffolds, meanwhile, 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. S3.15).

3.3 Conclusion

In the face of an apparent tradeoff between breadth and novelty in protein language models, we developed foldtuning, a PLM-based method that prioritizes the retrieval of novel — but plausible and useful — protein sequences by using individual target folds as structural guideposts to turn breadth reduction into a facilitator of novel sequence generation, speciating a library of several hundred foldtuned PLMs optimized for diverse structural and functional motifs. Foldtuning is an unabashedly "novelty-first" approach to protein design, predicated on the premise that structural mimics and knockoffs of real functional proteins are logical starting points for navigating the hidden order of protein-space and discovering new downstream binding and catalytic properties including ones less suited to *a priori* specification for *de novo* design. Across folds that vary in 2° and 3° structure composition and preferences, foldtuning stays anchored to its target fold family, as captured by high structural hit rates, while departing from natural sequence-space, as captured by high sequence "escape" rates; often, these metrics are maximized simultaneously, indicating that by accepting a self-imposed restriction on breadth, foldtuning frees its base PLM to chase the extremes of allowable sequence manipulation within the structural confines of the target. We attribute this remarkable performance to round-by-round PLM updates on self-generated synthetic sequences, validated as structural matches

and filtered against resemblance to natural versions. In a way, this contradicts recent claims — made in the context of large language models for text, and invoked in relation to PLMs — that such recursive re-training on model-generated data risks complete and unmitigated model collapse to gibberish (Shumailov et al., 2024).[8] Perhaps such a fate is avoided thanks to the careful filtering and validation steps; mayhaps it is a hazard to stay attuned to, that foldtuning stays Pareto efficient with respect to structure matching and sequence escape as the number of cycles is pushed past "evo+four." Either way, foldtuning explores substantial sequence novelty as hoped, powered by relatively small amounts of synthetic data per round; intriguingly, this is also consistent with recent efforts aiming chemical language models at the problem of unearthing previously-undetected small molecules (Qiang et al., 2024; Skinnider et al., 2021).

Despite — as is inherent in the definition of a PLM — only interfacing directly with *sequence* data — foldtuned models explore novelty at the levels of sequence, structure, and function. As far as sequence is concerned, foldtuned models display distinct preferences for amino-acid and short subsequence usage that depart from natural examples. To put the magnitude of these changes in subsequence usage in perspective — the differences in cognate "word" selection here are so pronounced as to be roughly on par with the pairwise lexical distances[9] between Spanish, French, and Portuguese — emphasizing that foldtuning is penetrating so far into far-from-natural sequence-space as to generate proteins reflecting newly-accessed underlying rules of language. Continuing down a level in information flow to structure, foldtuning samples elaborations, ornamentations, minimizations, and re-symmetrizations on and of its target fold — structurally plastic modifications that might translate to perturbed binding surfaces, active/allosteric sites, and so forth — remaining in the same neighborhood as the target fold, but strictly verboten to *de novo* backbone design specification. One gets the sense that as far as sequence and structure considered jointly, foldtuning behaves as an evolution-esque

[8] For an evocative example of such a catastrophe, consider the following: An image generation model is trained on real photographs of pet dogs. The generator's output might start, innocently enough, by over-emphasizing the most common household breeds: golden retrievers and German shepherds. Then when updated on its own output, it would produce photorealistic images of goldens only. And when trained yet further on this newest output, it would offer up a collection of golden grotesques sprouting extra legs, leathery tails, or Cyclopian eyes — more likely to grace the walls of a Cubist gallery than the rescue or shelter nearest you.

[9] "Lexical distance" or lexical similarity is analogous to 1 - (vocabulary shift) as defined above; for human languages it is generally calculated based on a standard word list of $\sim 100-225$ cognates (Swadesh, 1955). Here we compare lexical distance to vocabulary shift computed over $160,000$ 4-grams.

novelty generator, unencumbered by fitness, tooling about parts of protein-space removed from nature's little sliver. And in light of variable stability predictions for all these foldtuning-generated variants, a synergistic angle for exploiting the relative advantages of foldtuning and inverse-folding might be to feed some of the novel structures emitted by foldtuning to a state-of-the-art inverse-folding model as templates, in hopes of bolstering stability without fully reverting to natural sequence patterns. Sliding the last level down to function, according to functional predictors, foldtuning whets the appetite for everything from evolvable enzymes poised to perform new-to-nature chemistry, to modular domain parts for signaling pathways and programmable gene-editors, further validating the synthetic biology potential of coarsely structure-mimicking, sequence-perturbing döppelganger proteins. Of course, the monstrous caveat to all of the light foldtuning casts on hitherto unseen parts of the sequence→structure→function map is that all of the findings in this chapter have been steadfastly *in silico*; the true burden of utility for bioengineering and synthetic biology is experimental, and the subject of the next chapter.

3.4 Methods

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

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 in Section 2.4 (Andreeva et al., 2020). 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 (Blum et al., 2025).

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 in Section 2.4) 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 (Blum et al., 2025).

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

Sampling from finetuned ProtGPT2 models followed the same general procedures, hyperparameters, and processing steps as for sampling from the base pretrained ProtGPT2 model as described in Section 2.4, with the following differences: (1) in each road of foldtuning, 1000 sequences were generated from the appropriate finetuned model; (2) termination was after $0.4 \times M$ tokens, where M is the median length of SCOP-UniRef50 natural sequences for target fold f, or the first STOP token, whichever occurred first; and (3) generated sequences were force-truncated to a maximum length of Maa. Inference batch size on a single NVIDIA A100-80G GPU ranged from 125-500 sequences depending on target sequence length.

Structure Prediction and Assignment

All structures were predicted with default ESMFold inference parameters as in Lin et al. (2023). 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 to either (1) SCOP fold labels via Foldseek structure-based search against a custom database comprised of the $n = 36,900$

superfamily-level representative structures in SCOP v2, or to (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

$$z_k^{(i)} = \min_j \|x_k^{(i)} - x_{train}^{(j)}\|_1 \tag{3.1}$$

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 and the top 100 combined as 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 3.1, Figure S3.1, and Figure S3.2, 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 $k_{ij} \propto \{\text{seq. iden. between } s_i, s_j\}$.

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 Alford et al. (2017). 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 (Pudžiuvelytė et al., 2024).

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 (Yu et al., 2023). 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), and 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 (Benjamini and Yekutieli, 2001).[10]

Model Availability

A streamlined implementation of foldtuning is now distributed in trill (v1.8.3 and later; https://pypi.org/project/trill-proteins/) (Martinez et al., 2023).

[10]This is a conservative handling of false discovery for the problem at hand; testing for n-gram usage change is indubitably positively correlated between individual "words" as the relative overuse or underuse of any one word affects the available "lexical density" shared by all the remaining words. While the standard Binyami-Hochberg correction of rejecting all null hypotheses H_i for $i = 1, 2, ...k$ where k is the largest integer s.t. $p_k < (k/m)\alpha$ holds and is applied in this case, a resampling-based approach as in (Yekutieli and Benjamini, 1999) might be a preferable choice that does not sacrifice statistical power to the same extent.

3.5 Supplemental Material

Supplemental Figures

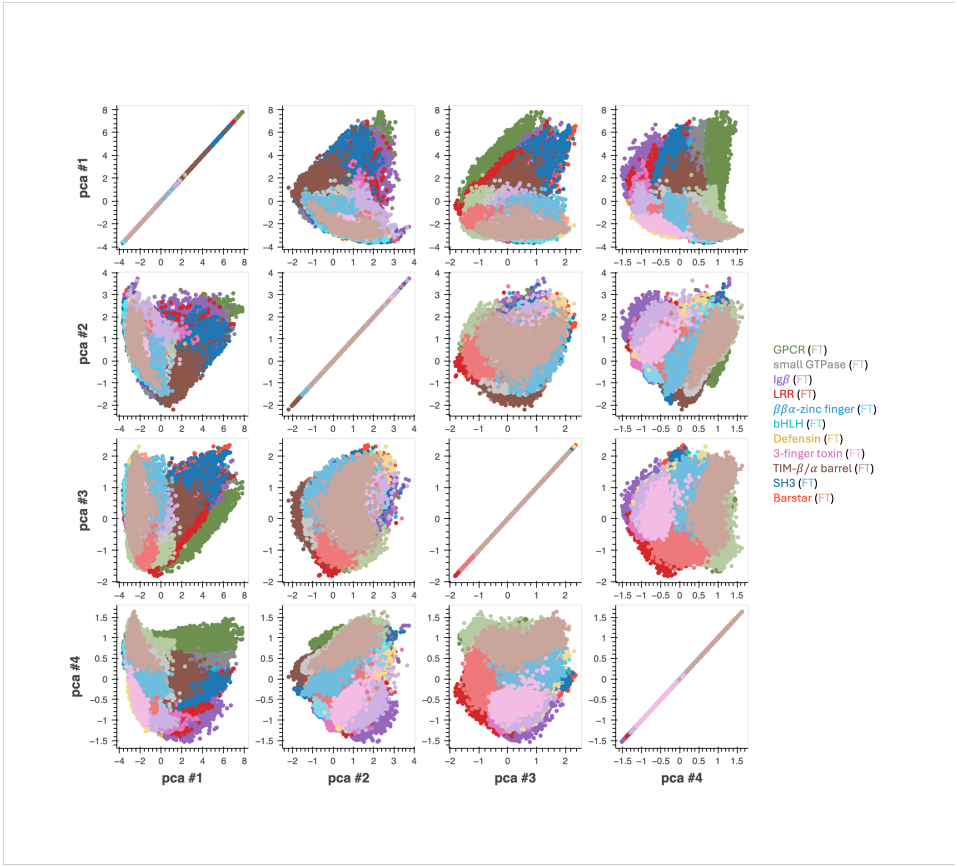


Figure S3.1: **Principal component analysis (PCA) of natural and foldtuned ESM2-650M embeddings.** Pairwise plots of top four principal components (fractional variance: 0.386, 0.103, 0.047, 0.039, respectively) of ESM2-650M embeddings of natural (SCOP-UniRef50) and foldtuning-generated proteins for 11 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-β/α barrels, SH3 domains, and barstar-like domains.

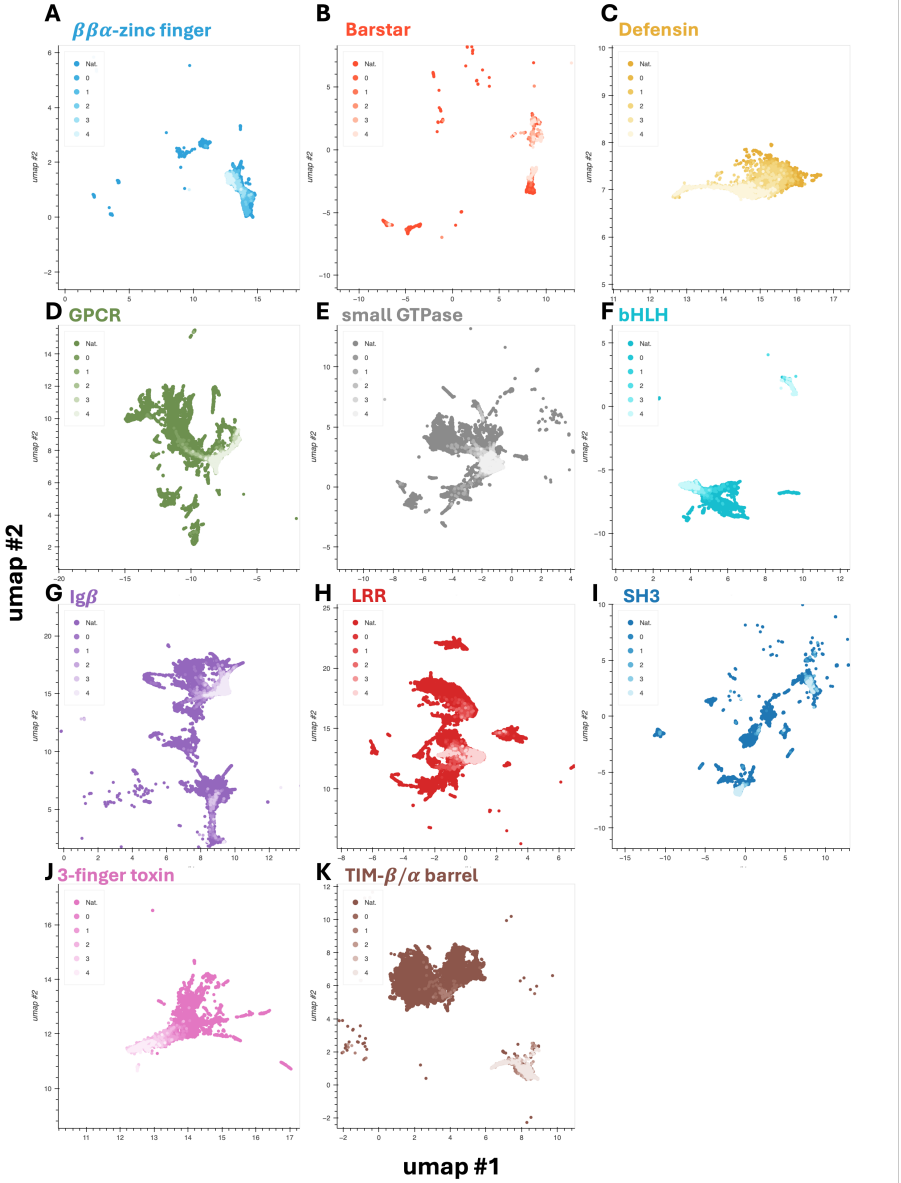


Figure S3.2: **ESM2-650M embeddings capture round-by-round drift of foldtuned sequences from their natural parents.** 2D UMAP representation of ESM2-650M embeddings for eleven representative target fold classes, progressing from natural examples through up to five rounds of foldtuning. Selected folds: (**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 β-sandwich (Igβ). (**H**) Leucine-rich repeat (LRR). (**I**) SH3 domain. (**J**) Three-finger toxin domain (3FTx). (**K**) TIM β/α barrel. Subfigure boundaries are set to the 5th- and 95th- quantiles in each UMAP component.

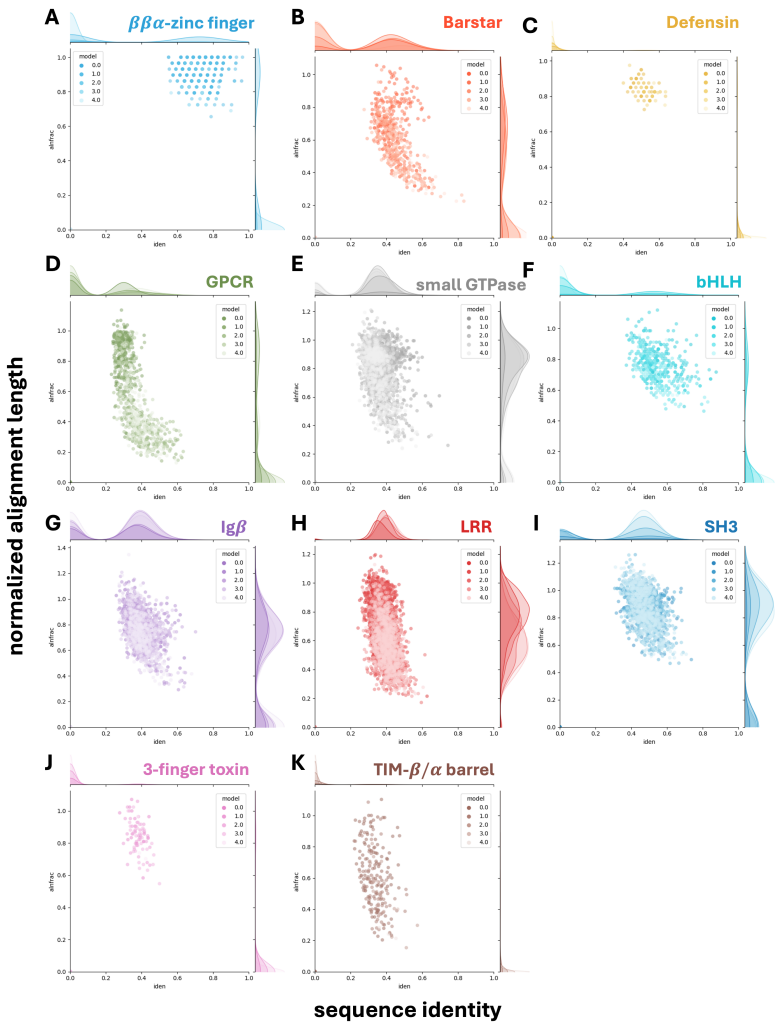


Figure S3.3: **Sequence similarity between foldtuned and natural variants.** Plots of normalized alignment length vs. sequence identity over the aligned region for the closest UniRef50 homolog to each foldtuned variant as identified by ultrasensitive search with MMSeqs2. Selected folds: (**A**) $\beta\beta\alpha$-zinc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

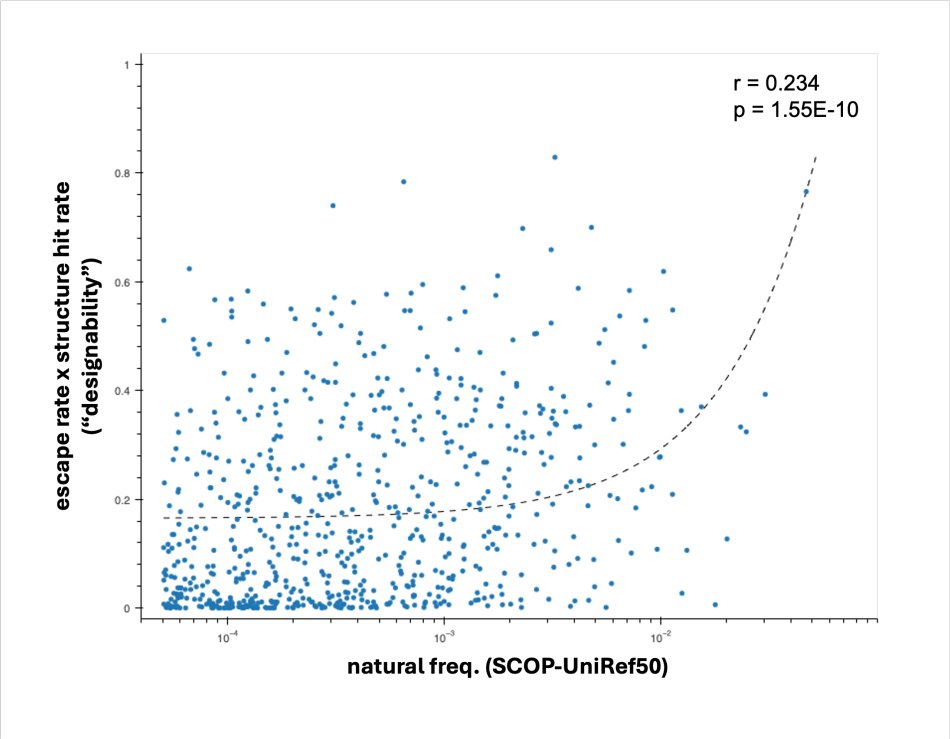


Figure S3.4: **Designability vs natural abundance for $n = 708$ SCOP fold targets.** Designability proxy (structural hit rate × sequence escape rate) across $n = 708$ SCOP fold targets is weakly explained by natural abundance in the custom SCOP-UniRef50 database: linear regression t-test for positive slope; slope= 12.80, $r = 0.234$, $p = 1.55 \times 10^{-10}$.

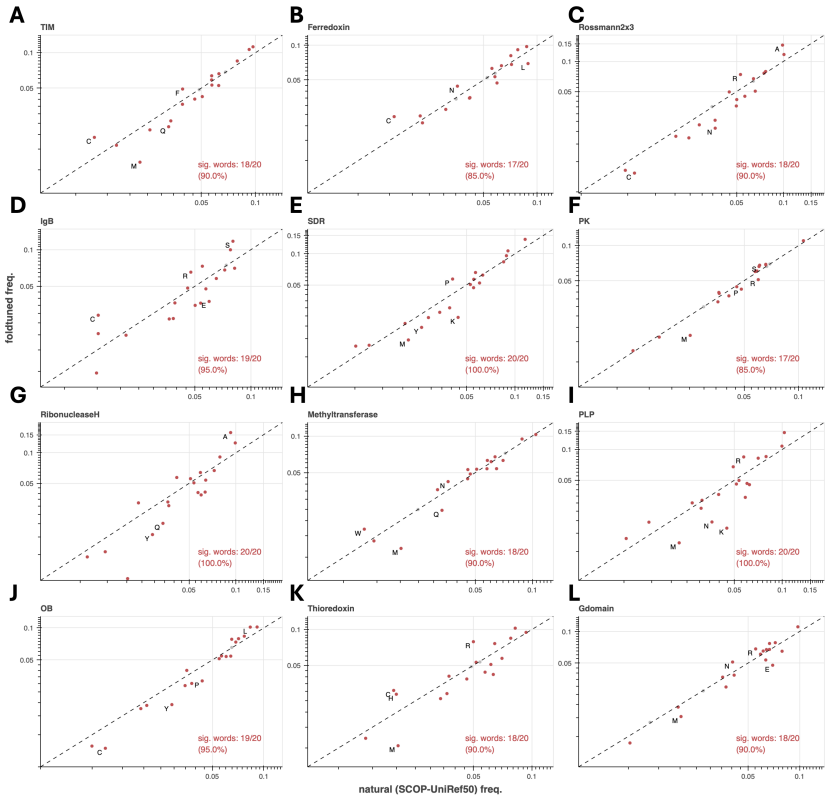


Figure S3.5: **Usage patterns of the 20 canonical amino-acids in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of AAs with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under binyamini-hochberg correction for positively correlated tests). The top-four most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

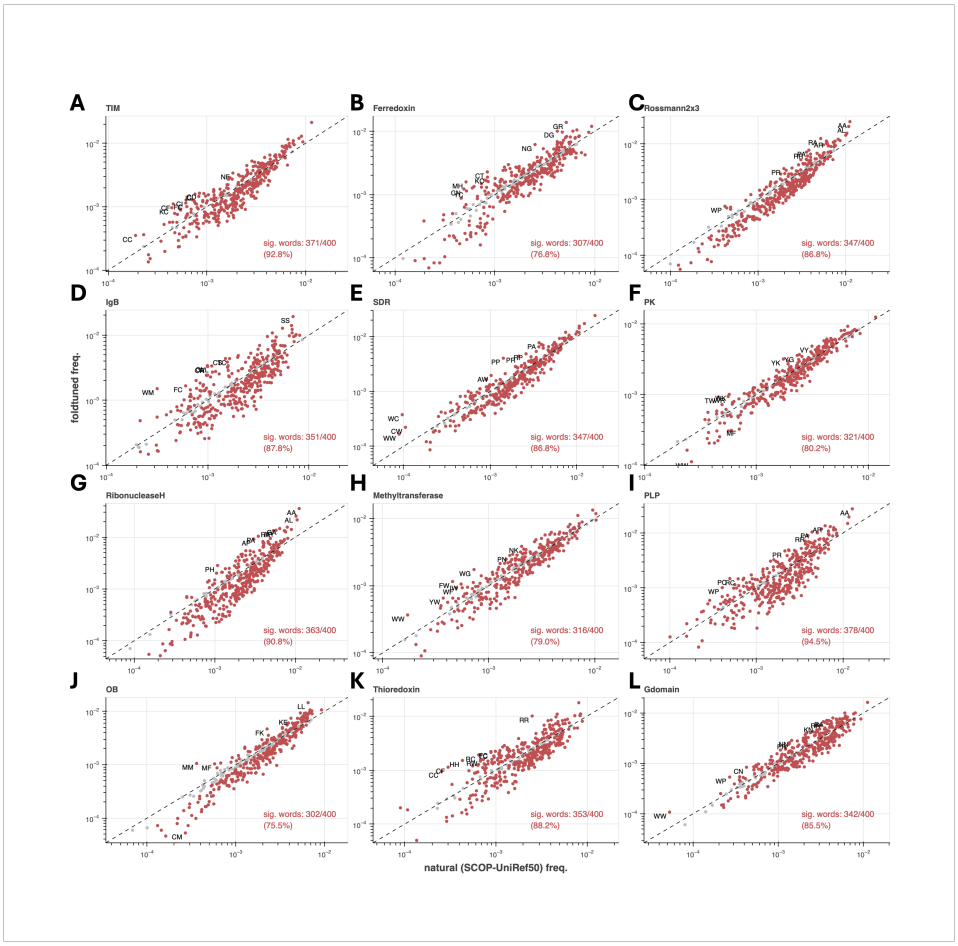


Figure S3.6: **Usage patterns of amino-acid subsequences of length 2 ("2grams", "bigrams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 2grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under binyamini-hochberg correction for positively correlated tests). The top-four most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

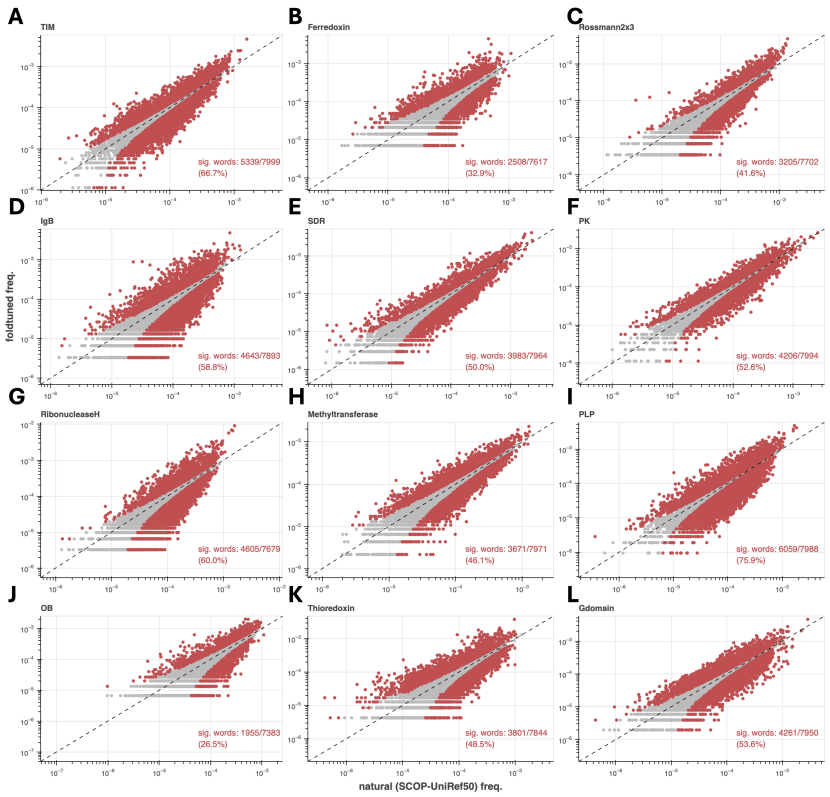


Figure S3.7: **Usage patterns of amino-acid subsequences of length 3 ("3grams", "trigrams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 3grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under binyamini-hochberg correction for positively correlated tests). The top-four most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

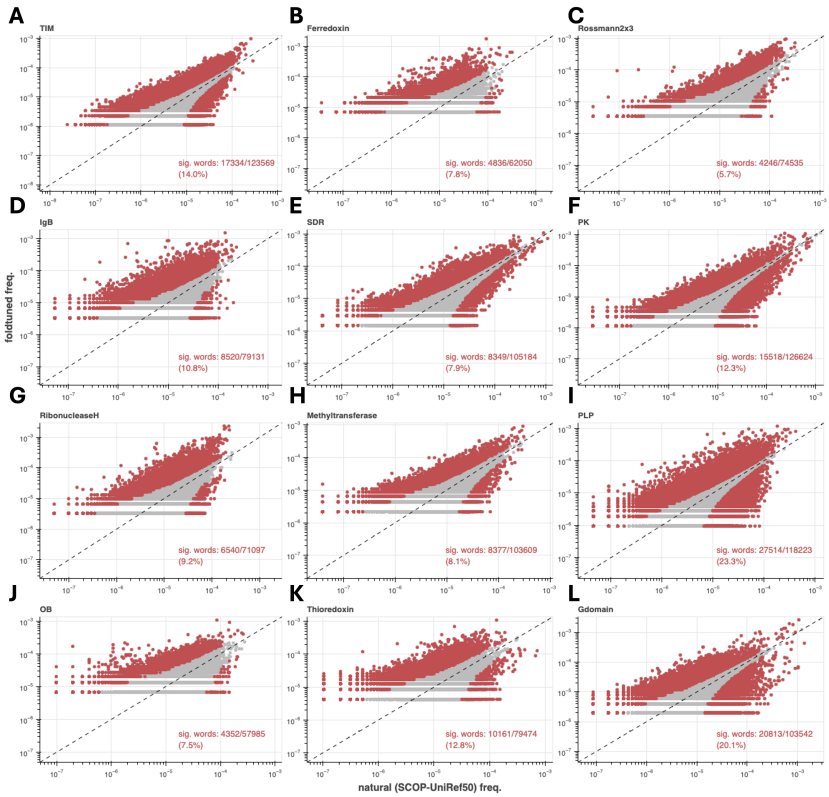


Figure S3.8: **Usage patterns of amino-acid subsequences of length 4 ("4grams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 4grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under binyamini-hochberg correction for positively correlated tests). The top-four most-shifted AAs as ranked by usage fold-change are labeled. Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

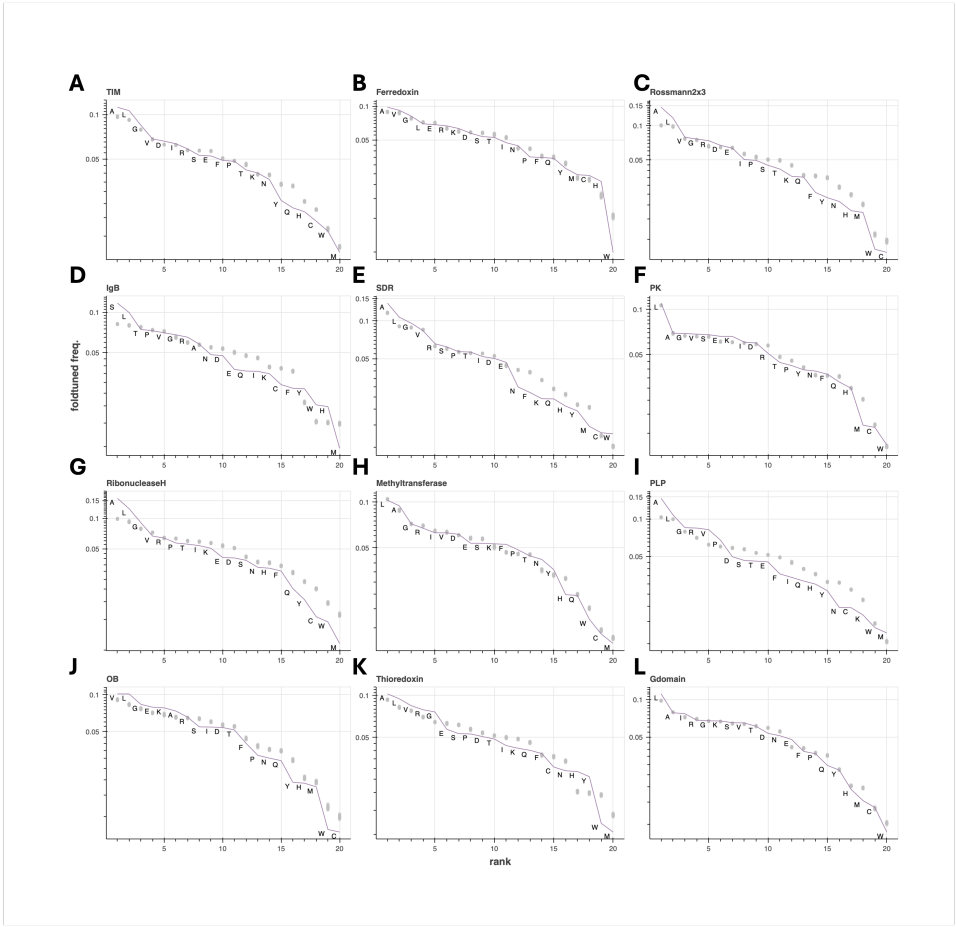


Figure S3.9: **Rank-ordered usage of individual amino-acids for foldtuned sequences (purple, labeled) and natural sequences (**n = 1000 **SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-like. (**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.

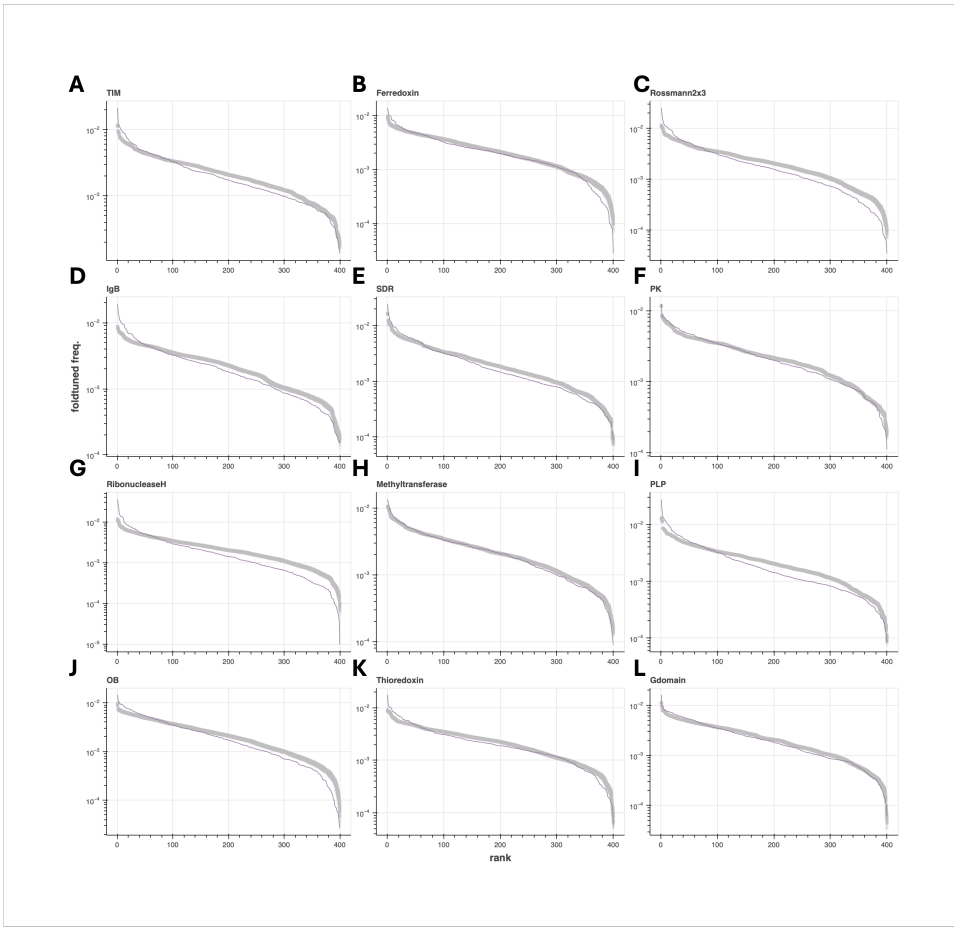


Figure S3.10: **Rank-ordered usage of subsequences of length 2 ("2grams", "bigrams") for foldtuned sequences (purple, labeled) and natural sequences (n = 1000 SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

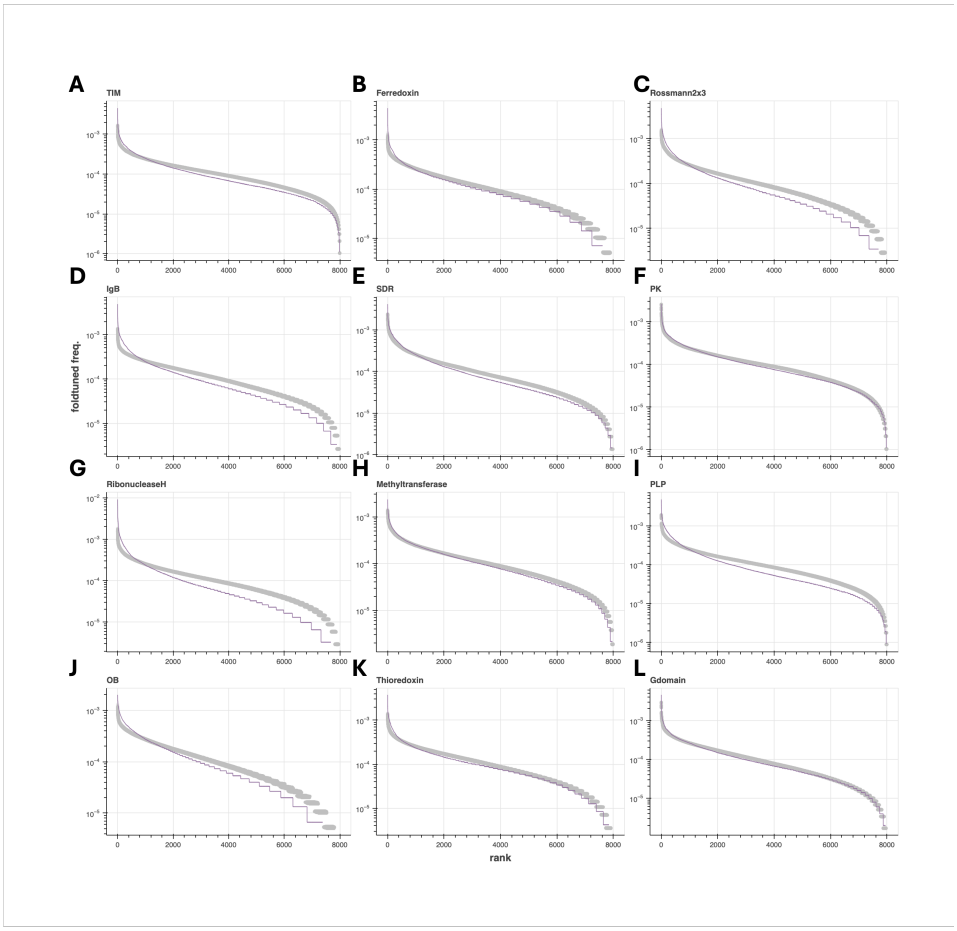


Figure S3.11: **Rank-ordered usage of subsequences of length 3 ("3grams", "trigrams") for foldtuned sequences (purple, labeled) and natural sequences (n = 1000 SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

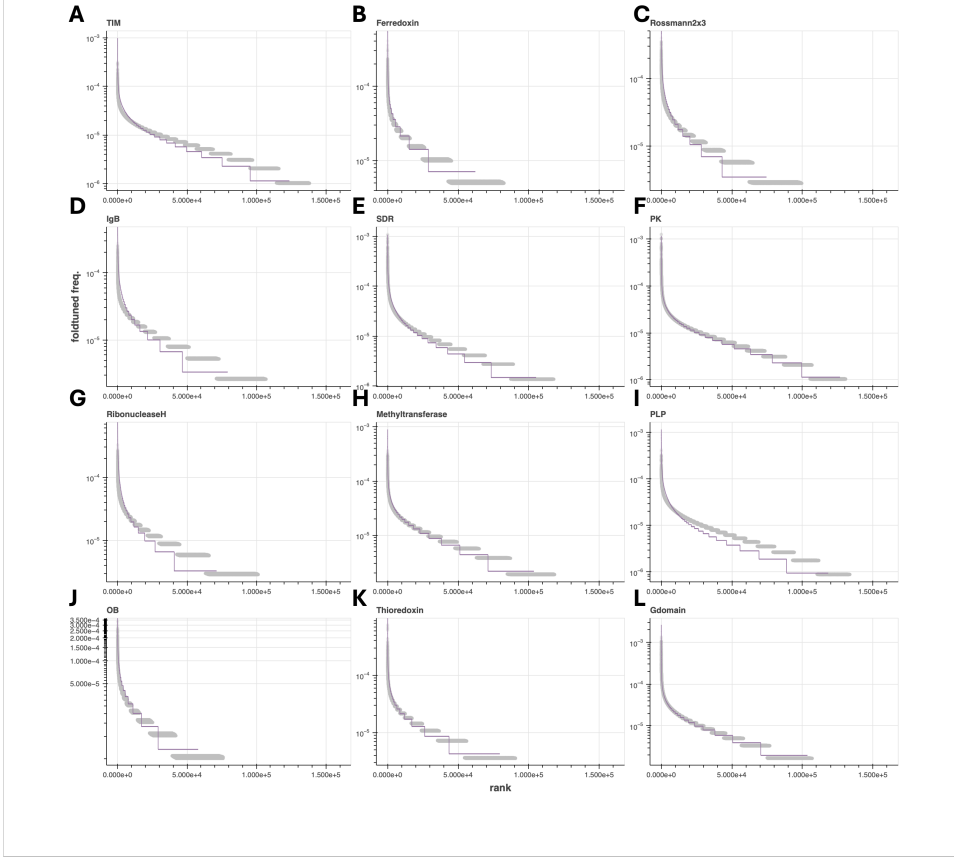


Figure S3.12: **Rank-ordered usage of subsequences of length 4 ("4grams") for foldtuned sequences (purple, labeled) and natural sequences (**n = 1000 **SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-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.

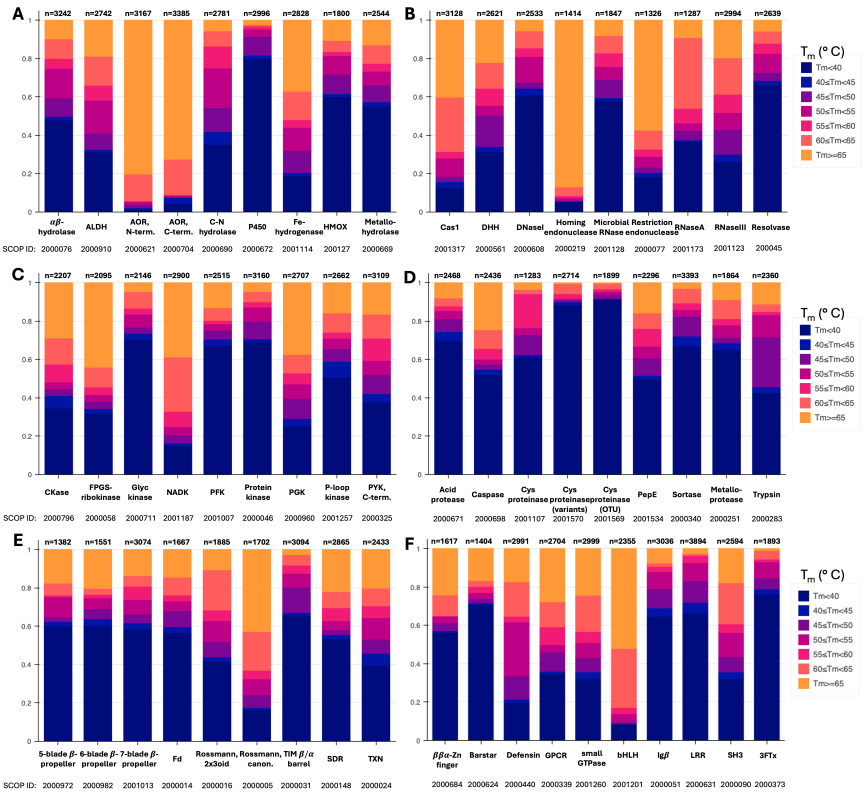
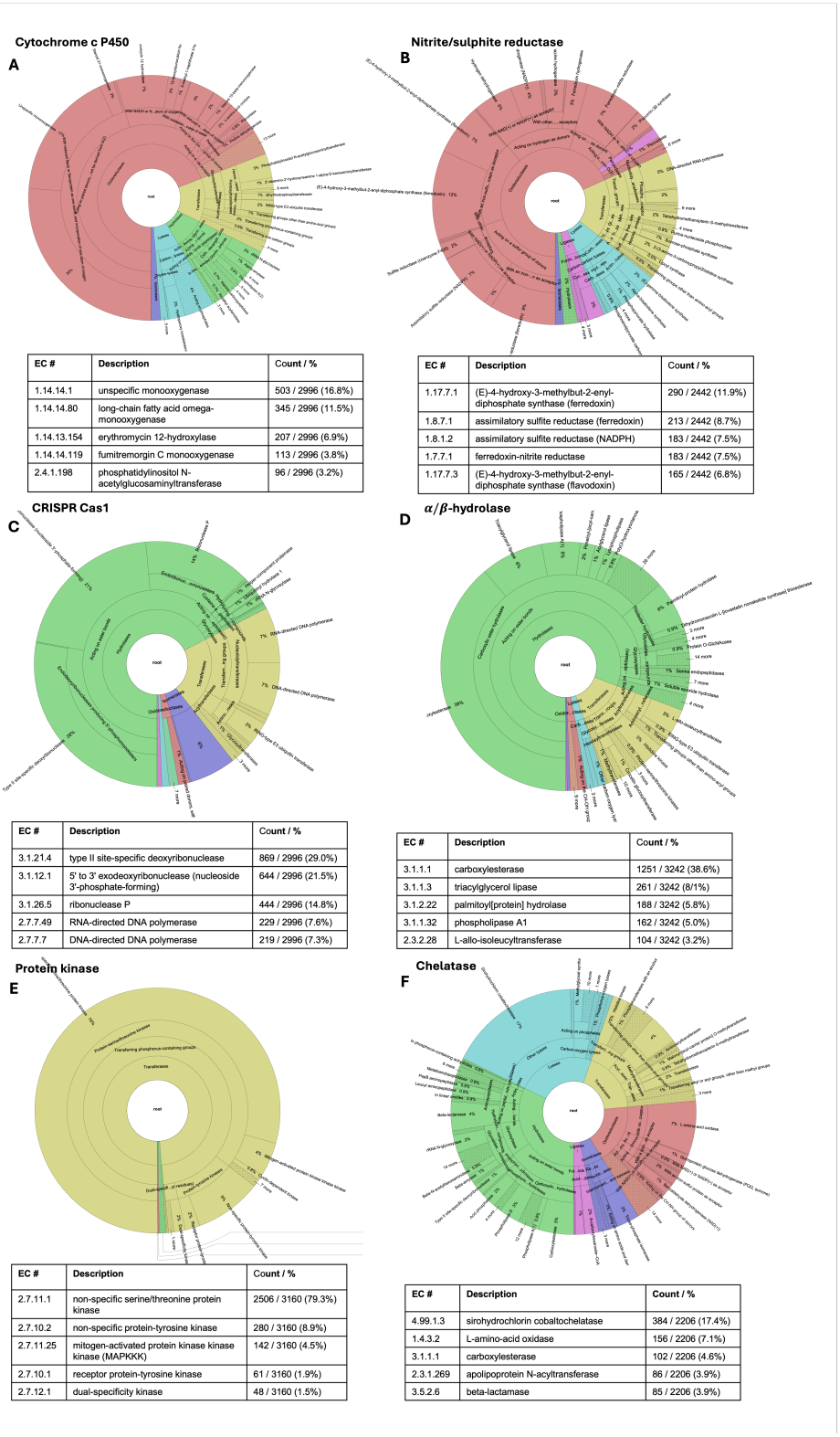


Figure S3.13: **Foldtuned proteins are predicted to exhibit varying degrees of thermostability.** Filtered, validated sequences generated from 55 foldtuned models of interest are expected to exhibit melting temperatures (T_m) ranging from $< 40°C$ to $> 65°C$, as predicted by TemStaPro (Pudžiuvelytė et al., 2024). 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.



Cytochrome c P450

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%)

Nitrite/sulphite reductase

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%)

CRISPR Cas1

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%)

α/β-hydrolase

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%)

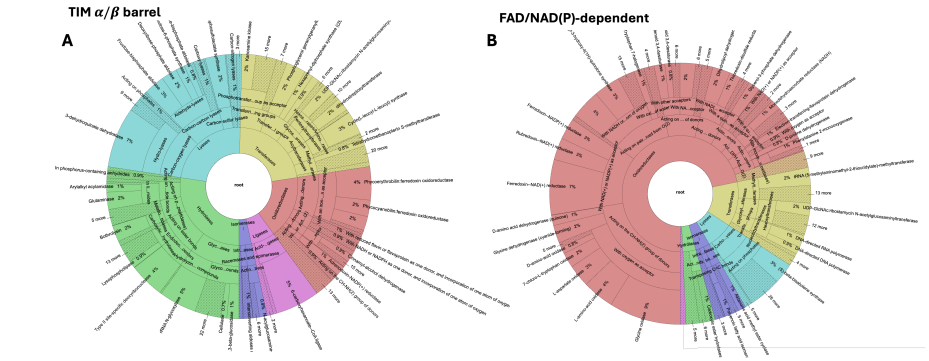
Protein kinase

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%)

Chelatase

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 S3.14: **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 (Yu et al., 2023). 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**) α/β-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 S3.15: **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. S3.14. Selected folds: (**A**) TIM β/α barrels. (**B**) FAD/NAD(P)-dependent enzymes. (**C**) Rossmann 2x3oid proteins. (**D**) Canonical Rossmann proteins.

Supplemental Tables

Table S3.1: **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 β-helix	β	0.881	0.941	0.829	Periodic
2000239	Ribbon-helix-helix domain	α	0.818	0.958	0.783	DNA-binding
2000031	TIM β/α barrel	α/β	0.770	0.995	0.766	Symmetry (8-fold)
2000920	Anti-$\|$ β/α barrel	$\alpha+\beta$	0.743	0.996	0.740	Symmetry (5-fold)
2000619	α/α toroid	α	0.704	0.994	0.700	Periodic
2000193	Transmembrane β-barrel	β	0.731	0.955	0.698	Symmetry (various)
2000308	Sm-like fold	β	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	α	0.664	0.920	0.611	DNA-binding
2000419	Pentein β/α 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	α	0.617	0.953	0.588	DNA-binding
2001248	RecA-like basic	α/β	0.724	0.807	0.584	DNA-binding