Chapter 5

PROTEIN LANGUAGE MODELS SPAWN NEW-TO-NATURE STRUCTURES

5.1 Introduction

In the preceding several chapters, we showed how protein language models (PLMs) can be steered with synthetic data to explore far-from-natural regions of protein-space, sampling the extremes of physically-allowable sequence novelty through generation. Through this approach, which we dub "foldtuning," we branched a base pretrained PLM, ProtGPT2, into a library of several hundred models, specialized by protein fold, that favor new "language rules" for assembling amino-acid words and letters into functional proteins. And we introduced experimental evidence, gleaned from a select set of targets, that foldtuned models propose real, buildable, functional proteins, best viewed as reflecting these metamorphosed language rules while respecting underlying fold-specific grammar at the core of what it means to have a protein sequence and not a meaningless string. In the process of exploring novelty of sequence, we touched only glancingly on novelty of structure. In this chapter, we rectify that wrong and prioritize the search for domain-quality structures as-yet-unseen in nature, achieving our aim through two complementary strategies inspired by our preexisting body of work with PLMs.

The beating heart of the pursuit of novel protein structures is a biophysical mystery fusing several questions into one: why might some (or many) compact three-dimensional structures be permitted by the laws of physics and yet apparently unrealized in nature? If casualties of a dearth of possible fold-encoding sequences or a lack of an essential fitness-conferring function, these hypothetical domains should be reachable by *de novo* design, and from first principles at that. Or perhaps flaws in folding thermodynamics or kinetics disfavor or even outright forbid them, and the documented set of structural units is complete after all. Opinions on the structural completeness debate have historically been sharply split (Chitturi et al., 2016; Skolnick et al., 2012; Taylor et al., 2009; Zhang et al., 2006). By the strictest measures only a handful of published *de novo* designed proteins — Top7, five all-α folds, eight α/β folds — qualify as truly new-to-nature (Kuhlman et al., 2003; Minami et al., 2023; Sakuma et al., 2024). And the advent of massive-scale pre-

dicted structure databases, dwarfing the $\sim 200,000$ experimental structures of the Protein Data Bank (PDB) with a rich trove of nearly 1 billion AlphaFold, ESMFold, and ColabFold predictions has only muddied the divide (Kim et al., 2025a,b; Lin et al., 2023; Varadi et al., 2022). With hordes of structural "dark clusters" carved from these databases plus newly collated Pfam families and CATH superfamilies, the PDB looks incomplete indeed (Barrio-Hernandez et al., 2023; Durairaj et al., 2023; Lau et al., 2024; Pavlopoulos et al., 2023). Conversely, if a $\sim$ 1000x inflation of individual structures "only" boosts the number of CATH superfamilies (the finest-grained level in that hierarchy) from 5,841 to 6,573, a 12.5% increase, and the number of topologies/folds (the second-finest-grained) from 1,349 to 2,081, a 54.3% increase, how much can nature really have left by the wayside structurally (Lau et al., 2024)?[1, 2]

On the surface, PLMs may seem an odd choice of tool for unearthing novel structures. "Language" is in the name; they are explicitly sequence models. On one hand, as we have established *ad nauseum*, PLMs may only "see" sequence, yet they implicitly capture the key features of structure and function as well. On the other, we demonstrated in Chapter 2 that left to their own devices, PLMs chop and skew the natural structural ensemble. Even the most vocal proponents of PLMs are prone to treating them as vehicles for infilling nature-adjacent variants into, say, an enzyme class, inducing small structural changes and smaller functional ones, deferring to the bounds of a CATH superfamily rather than breaking out (Madani et al., 2023; Munsamy et al., 2022). And yet, the prospect of novel structure enumeration through PLMs has lingered on the horizon, with examples — sparse ones, but examples nonetheless — reached through free generation and experimentally verified at a preliminary level (Ferruz et al., 2022; Verkuil et al., 2022).

Consequently, we reason that unlocking the full latent capacity of PLMs to access new domain structures requires another embrace of the "novelty first, fitness next" ethos, this time suited for hitting the rare pinpricks of structural novelty. To do so, we build and deploy two distinct fitness-agnostic strategies that enrich PLM output for novel structure generation. The first, inspired by fold recombination events

[1] A clarification — while 1 billion is $\sim$ 5000x 200,000, CATH annotation expansion only considered the $\sim$ 200 million entries in the AlphaFoldDB; 1000x is therefore the relevant inflation factor.

[2] A second clarification — as of CATH v4.4, the 732 novel CATH folds each contain exactly one novel CATH superfamily — the growth at the fold/topology level is hence more representative of the novelty uncovered. Still, we are talking about a new topology discovery rate of $732/214,683,839 \approx 3.4 \times 10^{6}$; or 3-4 per million newly predicted structures.

in real-world protein evolution, is a genetic algorithm; the PLM, ESM2-650M specifically, acts as an oracle favoring sequence plausibility and dense structural contacts. The second revisits foldtuning; instead of chasing sequence-diverging structural matches we select against resemblance to the entire set of CATH domains. Both approaches eschew direct interaction with sequence features. Both employ structural compactness as the primary or sole selective force. And both deliver an abundance of novel folds computationally projected to be stable, foldable, and unmappable to any CATH example, spanning protein topology classes. We contend that despite substantial architectural differences between the two methods, they execute the same overarching tactic of discovery by ignoring natural waypoints, without needing to overtly design against them.

5.2 Results & Discussion

Novel domains emerge from a fold-recombining genetic algorithm

One potential avenue for finding novel protein domains is to start from primitive structural elements and recombine them, evolve them, and put them under selective pressure, all in *in silico*. With a suitable selective force, one that rewards some notion of well-foldedness and/or compactness, stable tertiary folds, alike-to-nature and new-to-nature can both emerge. This approach is a genetic algorithm for domain diversification, loosely inspired by hypotheses for how early enzymes and ancient protein folds may have originated from primoridal polypeptides.[3] As starting material to seed the algorithm, we generate a small library of 800 mini-protein-sized (40aa) fragments *de novo* via PLM-informed replica-exchange Metropolis-Hastings Monte Carlo sampling. Briefly, random amino-acid sequences are evolved in single point mutation steps subject to an energy function that favors greater sequence likelihood and structural contact density, both as inferred by ESM2-650M (full implementation details are provided in Section 5.4). The mini-proteins produced sample a variety of topologies varying in relative α and β content and organization, as well as loop sizes, geometries, and degrees of order (Fig. S5.1). The choice of *de novo* generation is motivated by a desire to mitigate against sequence-side biases in favor of nature that might be introduced by the most straightforward alternative of fragmenting real or experimental structures from published databases. Indeed, while structure-based search with Foldseek ($504/800 = 63.0\%$ hit rate against Al-

[3]The topic of structural and functional emergence and plasticity in polypeptides is far too rich to cover adequately in the context of this chapter. Specific recommended examples include Longo et al. (2020b), Longo et al. (2020a), and Vyas et al. (2021). A highly recommended review, albeit predating the aforementioned studies, is Tóth-Petróczy and Tawfik (2014).

phaFoldDB50) shows that the generated fragments are plausible and representative building blocks, sequence-based search with MMseqs2 (48/800 = 6.0% hit rate against UniRef50) indicates that they are distinct from natural sequences, both as desired.

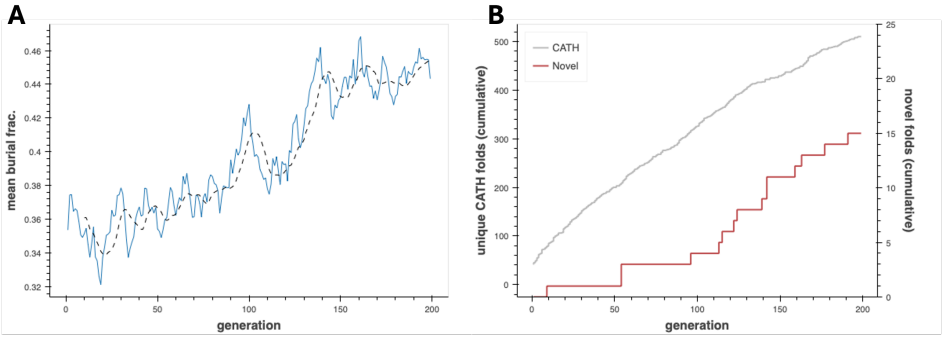

Figure 5.1: **Emergence of novel folds from a PLM-based genetic algorithm.** (**A**) Mean fractional amino-acid surface burial (protein compactness proxy) over 200 generations of the structure discovery genetic algorithm. (**B**) Cumulative counts of unique CATH-annotated folds and putative novel folds detected over 200 generations of the structure discovery genetic algorithm.

A randomly selected subset of 100 mini-protein fragments is carried forward as the initial population for the genetic algorithm, which proceeds for 200 epochs. In each epoch, 20 recombined and mutated fragments are generated and evolved over the same energy landscape as used for the fragment library before being added to the population; stochastic selection with survival rate proportional to burial fraction is performed to reduce the population back to a target constant size of 100.[4, 5] The mean burial fraction increases with time, demonstrating that compact folds become more common and/or folds become more compact on average as the algorithm proceeds (Figure 5.1A). Assigning CATH labels wherever possible with Foldseek-TMalign, natural folds accrue at a roughly constant rate of 2.4 per epoch, while compact (burial fraction > 0.5) yet novel folds emerge sporadically; the first new-to-nature fold (3733B8_R10) appears in epoch 10 with subsequent interfold arrival times as long as 45 and as short as 2 epochs (Figure 5.1A-B). Working off of building blocks

[4]Full implementation details, including construction of the selection function, may be found in Section 5.4.

[5]The only methodological distance of substance between the MHMC sampling process for the fragment library and the recombination algorithm is a switch from multiple chains with replica-exchange in the former case to a single chain with dynamic temperature adjustment in the latter. This choice reduces total run time per epoch by a factor of ~ 5x — a significant speedup when one epoch takes ~ 0.5-1 gpu-hr with a single chain on typical hardware.

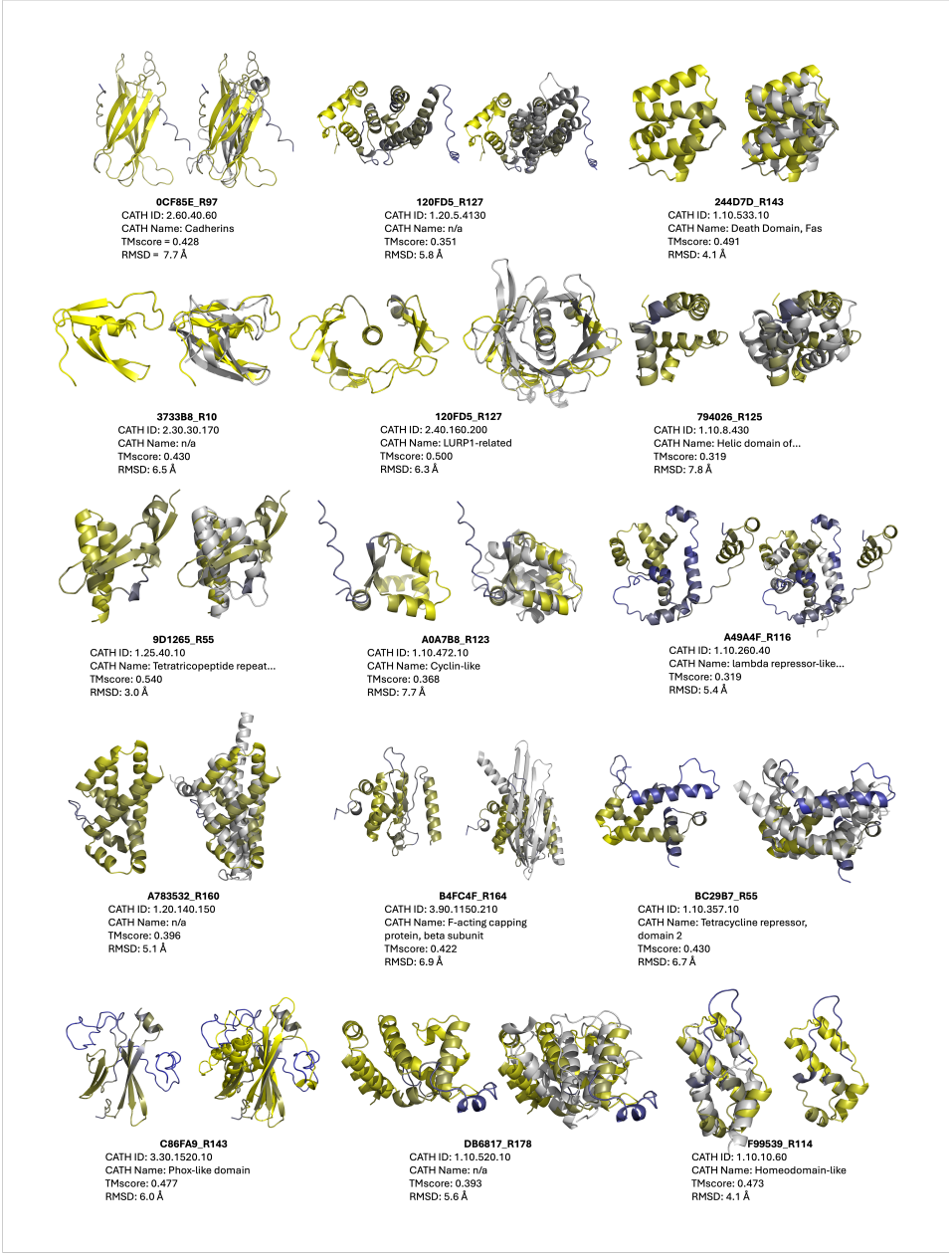

Figure 5.2: **Fifteen novel folds achieved by the structure discovery genetic algorithm.** Within each pair: **left** — putative novel fold (colored by ESMFold pLDDT; yellow=high, blue=low); **right** superimposed with closest CATHDB50 Foldseek hit in TMalign mode, with CATH metadata and global alignment metrics reported below.

that are almost exclusively displaced from nature in sequence but nearby in structure, the algorithm reaches ~500 natural folds and 15 putatively novel ones, suggesting that natural structure-space is far from complete and that additions are surprisingly accessible to design when a backbone is not specified *a priori*.

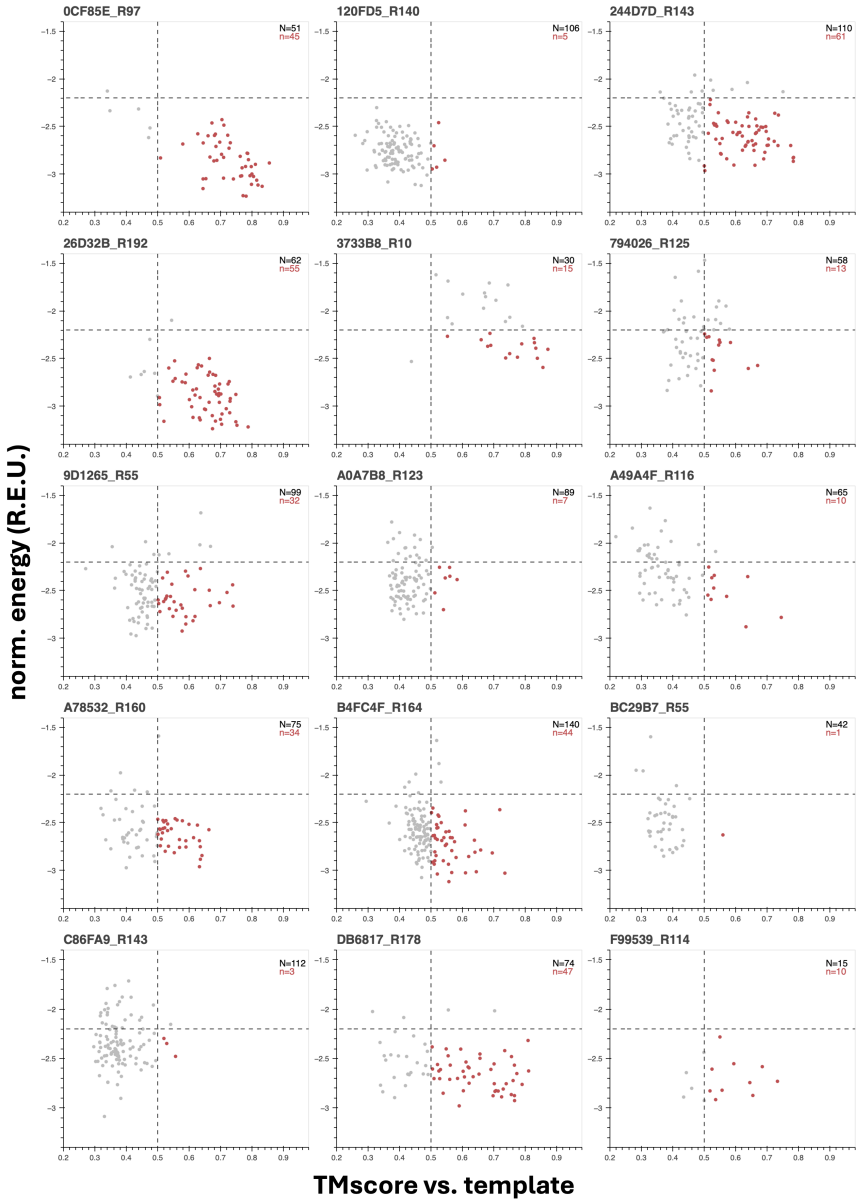

Figure 5.3: **"Inverse-folding landscapes" for fifteen novel folds achieved by the structure discovery genetic algorithm suggest variable stability.** Length-normalized energies (from Rosetta) vs. TM-score (from Foldseek in TMalign mode) for ProteinMPNN-designed sequences inverse-folded off of structure discovery genetic algorithm putative novel folds as templates. Gray dots correspond to all sequences/structures for a given template after clustering 200 initial sequences per template at 60% sequence similarity. Red dots show the subset of inverse-folded seqUences whose ESMFold-predicted structures pass an energy scoring threshold ($\bar{E} < -2.2$ REU/aa) and the standard TM-score global match threshold (TMscore > 0.5).

The 15 new-to-nature domains proposed by the evolutionary algorithm are markedly distinct from their nearest CATH analogs and structurally diverse, visiting three of the four major topology classes — all-α (120FD5_R127, 244D7D_R143, 794026_R125, A49A4F_R116, A783532_R160, B4FC4F_R164, BC2987_R55, DB6817_R178, F99539_R114), all-β (0CF85E_R97, 3733B8_R10, C86FA9_R143), and $\alpha + \beta$ (120FD5_R127, 9D1265_R55, A0A7B8_R123), as categorized by eye (Fig. 5.2, Table S5.1). It is curious that no novel α/β folds occur, given the prominent functional speciation of such domains in nature (Choi and Kim, 2006).

For additional insight into this handful of novel domains and whether they are truly plausible as far as the thermodynamics and kinetics of protein folding, we introduce the "inverse-folding funnel." The inverse-folding funnel is a heuristic inspired by the use of Rosetta *ab initio* structure prediction simulations to explore a protein-folding energy landscape. The traditional result is a plot of estimated energy vs. backbone RMSD to the target for many replicates of the same sequence, with two ideal features: (1) a clear association between lower energy (higher stability, i.e. favorable folding thermodynamics) and smaller RMSD; and (2) an absence of "trapped" subpopulations at moderate-to-high RMSD and local energy minima (presumed metastable states, indicators of poor folding kinetics). A plot satisfying both resembles the prototypical folding funnel of a globular protein spontaneously collapsing to its native-state structure, whereas one failing either or both criteria warns of folding pathologies precluding viable expression let alone function (Dill and Chan, 1997). Analogously, we instead use an inverse-folding model (ProteinMPNN) to generate many sequence-diversified versions expected to encode a given putatively novel domain structure from Figure 5.2 provided as a backbone template. Preclustering by sequence similarity to minimize redundancy, we predict structures with ESMFold, estimate absolute energies with Rosetta, and quantify global alignment between inverse-folded structures and templates as TMscores. For the "inverse-folding" version of the funnel, we look for correlation between lower energy and *higher* TMscore and for a lack of low-TMscore/low-energy states — the former remains a proxy for thermodynamic stability, while the latter rules out both metastability and the possibility that a particular "novel" domain is no more than a noised version of a CATH domain recoverable by the slight re-noising of inverse-folding.

For 8 of the 15 putatively novel domains,[6] this procedure evinces a convincing funnel

[6]Specifically, folds: 0CF85E_R97, 244D7D_R143, 26D32B_R192, 3733B8_R10,

with the aforementioned essential characteristics, bolstering confidence that these are realizable new-to-nature structures (Fig. 5.3). Other faux folding landscapes point to problem spots; for example for 9D1265_R55 multiple equivalent energy minima are observed, while for BC29B7_R55 a single minimum is centered around a TMscore well less than 0.5, as if inverse-folding reliably converges to a more stable neighbor in structure-space (Fig. 5.3). Other landscapes, far from being funnel-shaped, are almost flat, as in the case of F99539_R114, implying that some novel domain candidates may lack a true native state. As general guidelines for stable and robust structures, we additionally set rough threshold values of < -2.2 REU/aa and TMscore > 0.5 for inverse-folded variants to clear and note that even for those domains that do exhibit funnel-like folding landscapes many variants can fail one or both, reiterating the importance of the re-noising step for recovering more-plausible adjacent structures (natural or novel) from novel domain candidates. Despite the ample evidence that not all potentially novel folds brought forth are in fact novel, or, when they are, not created equal as far as folding dynamics and stability, fold recombination and evolution from artificial fragments inculcates a strong belief that natural structure-space does not enumerate all that can be afforded by protein biophysics.

Structure-first foldtuning enriches for domains with new-to-nature structures

In an orthogonal approach, we considered whether foldtuning could be transformed from a sequence-perturbing, fold-preserving method for novel sequence discovery into a fold-perturbing, sequence-insensitive method for novel structure discovery. To estimate the latent capacity of our go-to PLM, ProtGPT2, to generate previously unseen structural motifs off-the-shelf without additional training, we revisited the hyperparameter scan experiment from Chapter 2. The ~3 million predicted structures obtained across thirty (top_k, temperature) pairs were downsampled by 10x and re-annotated with CATH domain labels wherever possible, running Foldseek in accelerated TMalign mode with the precompiled CATHDB50 database as the target. Compactness/globularity was estimated for all predicted structures using fractional burial of total amino-acid surface area relative to the disordered polypeptide chain as a proxy metric sufficient for ranking and coarse binning. Aggregated results are reported in Table 5.1. As thresholds for putative novel structures, we look for predicted structures with a fractional burial > 0.5 and no assignable CATH domain label; occurrence rates range from 0.11% for top_k 1500 and temperature 0.8 to

A49A4F_R116, A78532_R160, B4FC4F_R164, DB6817_R178.

0.41% for top_k 4000 and temperature 5.0. In general, increasing either hyperparameter corresponds to an increase in this novelty rate, but the trend is imperfect. In contrast to the compression of SCOP fold uniqueness reported with increasing top_k and temperature in Chapter 2, the number of unique CATH domains detected *increases* slightly in this context. When we move up rung to the CATH topology/fold level (i.e. CAT), however, we see the same general structural diversity collapse as with SCOP. This implies that increasing top_k and/or temperature to favor textual diversity does somewhat emphasize structural novelty, but this comes in the form of finer-grained structure perturbations and at the expense of the larger supersecondary rearrangements that we hope to see as evidence of satisfyingly novel *folds*. Adding in the fact that the fraction of compact proteins (burial fraction > 0.5) consistently drops by roughly 2x as temperature goes from 0.8 to 5.0, we fix sampling hyperparameters at top_k 950 and temperature 1.5, striking a balance between compactness, CATH non-assignability, and structure perturbation *magnitude* as we move forward to what we refer to as "structure-first" foldtuning.

Structure-first foldtuning (described fully in Section 5.4) mirrors the architecture of the original "sequence-first" foldtuning developed in Chapter 3, with crucial differences on the discrimination/selection side. In brief, in each of five foldtuning rounds, 10,000 sequences are generated out of the current (k-th) model and filtered based on predicted structures to enforce compactness (burial fraction > 0.5) and CATH non-assignability (no Foldseek-TMalign hit in CATHDB50 with TMscore > 0.5). Filtered sequence-structure pairs are ranked in order of descending burial fraction, with the 100 most-compact becoming the training set used to finetune the $(k + 1)$-th model. Given the absence of a specific target fold, there is no need for an initial evotuning round. Over 5 rounds, structure-first foldtuning progressively enriches for sequence-structure pairs meeting the compactness/non-assignable novelty criteria, from 111/10,000 (11.1%) after one round to 269/10,000 (26.9%) after five (Table 5.2). Neither burial fraction nor the number of unique CATH domains is observed to change significantly at the population level, with a concomitant drop in the CATH assignability rate (across all sequences/structures), a further indication that while a non-globular sub-population persists, all of the growth in structural diversity is diverted to putatively novel domains.

Structure-first foldtuning proposes 1018 novel domains in total over five rounds.[7]

[7] As an aside, note that structure-first foldtuning brings along sequence novelty for free, without any explicit design consideration on the sequence side. Only 10/1018 sequences encoding the putative novel domains — $\approx$ 0.1% — exhibit detectable sequence similarity to any natural protein

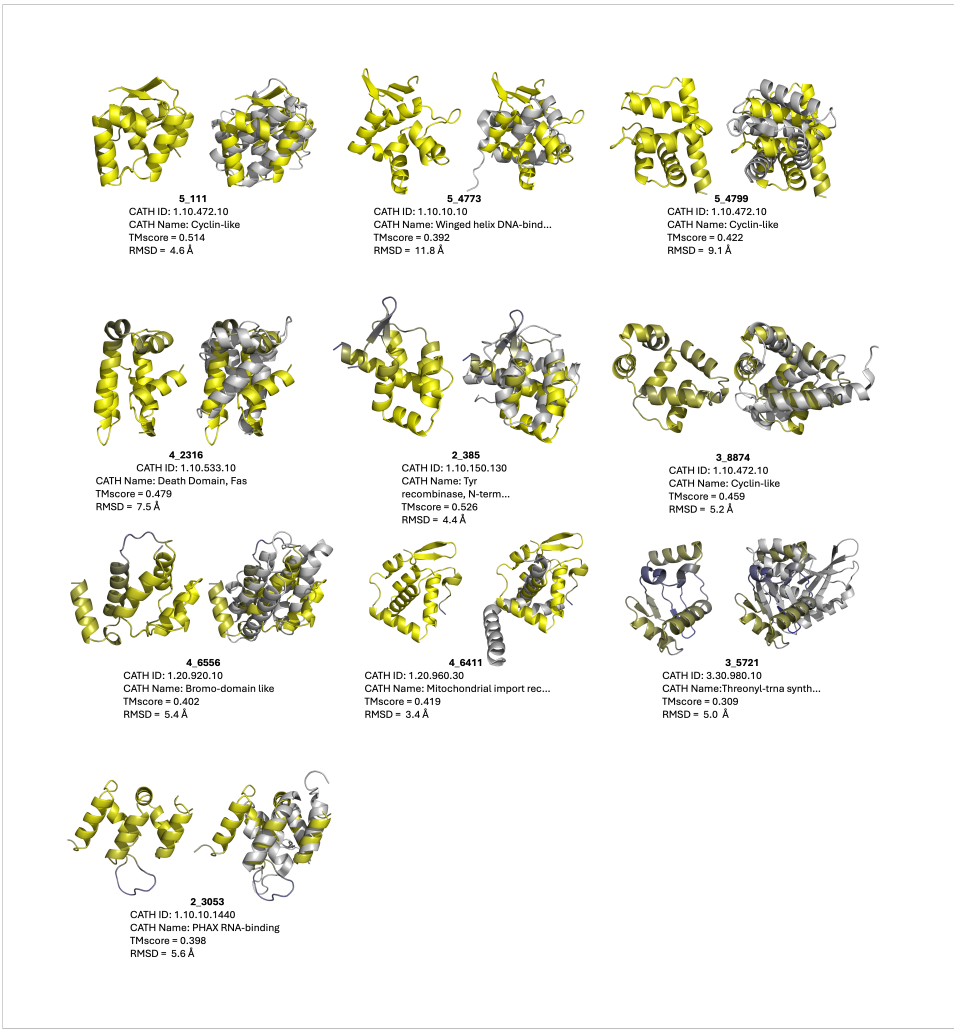

Figure 5.4: **Ten out of 100 novel folds achieved by structure-first foldtuning.** Within each pair: **left** — putative novel fold (colored by ESMFold pLDDT; yellow=high, blue=low); **right** superimposed with closest CATHDB50 Foldseek hit in TMalign mode, with CATH metadata and global alignment metrics reported below.

To accommodate limited computing resources, this set of 1018 is reduced to a set of high-priority templates to 916 by clustering at a TMscore > 0.5 global alignment threshold to group templates that would occupy the same superfamily and/or fold if added to the CATH database. Applying a stricter structural novelty criterion — no Foldseek-TMalign hit with TMscore> 0.5 to any domain in the entire AlphaFoldDB50 database — reduces the priority template set further to 762 members. The final high-priority set is contracted to 100 members after ranking by descending burial fraction and taking the top 100 most-compact. Feeding this

in UniRef50 per MMSEQS2 search.

Table 5.1: **CATH domain coverage, structural compactness, and novel fold discovery rate from base ProtGP2 sampling hyperparameter scan.** CAT(H) folds(superfamilies) detected, CATH hit absence (no hit with TMscore > 0.5), structural compactness (burial fraction > 0.5), and novel fold discovery rate for 30 sampling hyperparameter combinations from varying top_k (vocabulary size: 600, 950, 1500, 2400, 5000) x temperature (0.8, 1.0, 1.2, 1.5, 2.0, 5.0).

Hyperparams		Results				
top_k	temp	# CATH	# CAT	No CATH	Compact	Both
600	0.8	905	401	0.224	0.2376	0.0016
	1.0	935	412	0.2121	0.2298	0.0021
	1.2	975	404	0.2189	0.2176	0.0025
	1.5	988	408	0.2218	0.1983	0.0023
	2.0	977	407	0.2418	0.1839	0.0025
	5.0	967	384	0.3628	0.1039	0.0017
950	0.8	908	402	0.2143	0.2361	0.0018
	1.0	955	416	0.2115	0.2293	0.0018
	1.2	988	430	0.2261	0.1996	0.0031
	1.5	984	**419**	0.2347	**0.1922**	**0.0037**
	2.0	994	421	0.2432	0.1746	0.0036
	5.0	996	394	0.3584	0.1008	0.0029
1500	0.8	954	404	0.2145	0.2313	0.0011
	1.0	964	410	0.2228	0.2113	0.0029
	1.2	994	418	0.2378	0.1908	0.0023
	1.5	1014	415	0.2464	0.1727	0.0028
	2.0	1005	403	0.2612	0.1528	0.0028
	5.0	1017	382	0.3634	0.095	0.0028
2400	0.8	941	406	0.2227	0.2221	0.002
	1.0	970	410	0.2279	0.2045	0.002
	1.2	993	420	0.247	0.1804	0.0029
	1.5	1025	412	0.2572	0.1582	0.0036
	2.0	1055	425	0.2734	0.1417	0.0034
	5.0	1054	396	0.3536	0.0963	0.0033
4000	0.8	962	433	0.2232	0.2303	0.0024
	1.0	1021	440	0.2183	0.2001	0.0026
	1.2	1012	418	0.2521	0.1767	0.0022
	1.5	1076	425	0.2539	0.1519	0.0023
	2.0	1010	380	0.2786	0.1358	0.0027
	5.0	1008	390	0.341	0.1028	0.0041

final set to ProteinMPNN as inverse-folding templates and calculating TM-scores and folded-state energies for the respective outputs yields a set of inverse-folding energy lanscapes as in the preceding section. Predicted structures (with and without

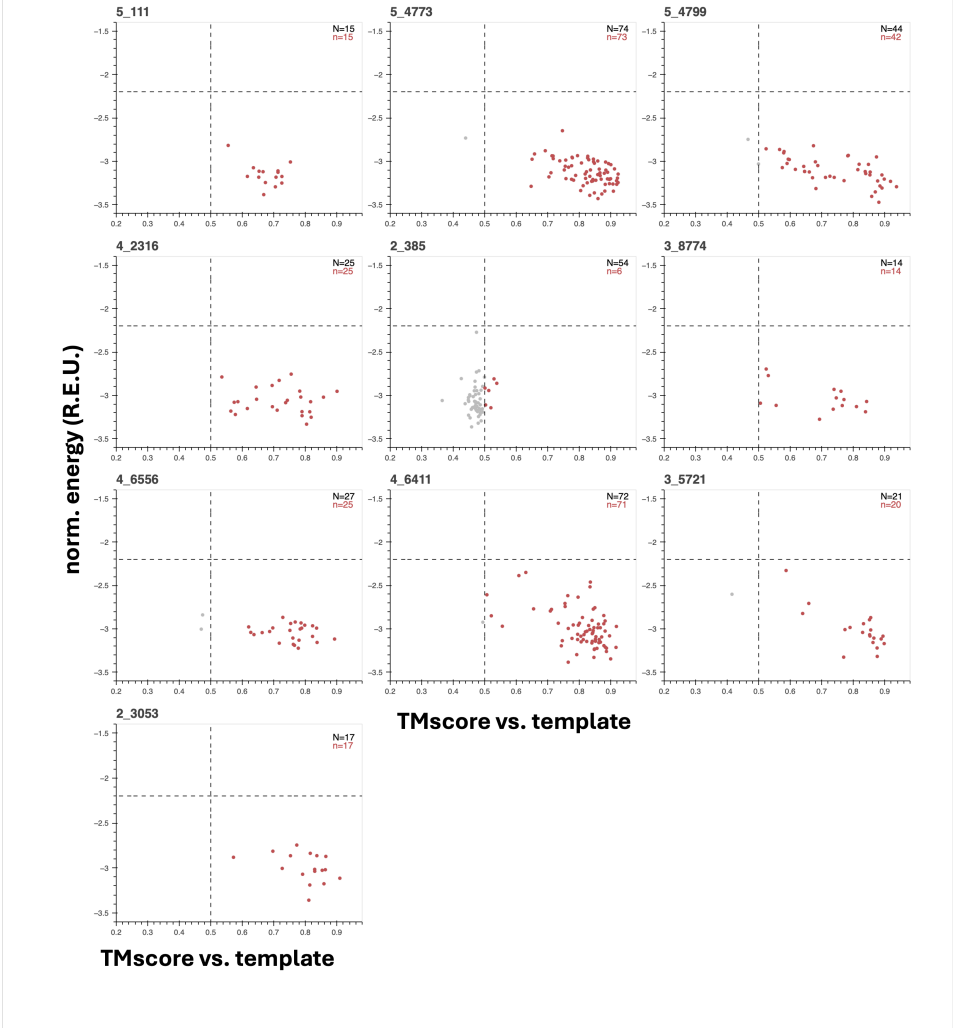

Figure 5.5: **"Inverse-folding landscapes" for ten out of 100 novel folds achieved by structure-first foldtuning imply high stability.** Length-normalized energies (from Rosetta) vs. TM-score (from Foldseek in TMalign mode) for ProteinMPNN-designed sequences inverse-folded off of structure-first foldtuning putative novel folds as templates. Gray dots correspond to all sequences/structures for a given template after clustering 200 initial sequences per template at 60% sequence similarity. Red dots show the subset of inverse-folded seqUences whose ESMFold-predicted structures pass an energy scoring threshold ($\bar{E} < -2.2$ REU/aa) and the standard TM-score global match threshold (TMscore > 0.5).

closest CATH hits) and inverse-folding landscapes for the best 10 templates as ranked by average estimated folded-state energy are shown in Fig. 5.4 and Fig. 5.5 respectively.

Table 5.2: **Emergence of novel and CATH-annotated domains over five rounds of "structure-first" foldtuning.** Number of generated sequences successfully annotated with a CATH domain by Foldseek ("# CATH"), structural hit rate (fraction of generated sequences assigned to *any* CATH label), number of generated sequences assigned as putative novel folds ("# Novel"; burial fraction > 0.5 and no hit with TMscore > 0.5), and mean burial fraction over the course of five rounds of structure-first foldtuning with top_k 950, temperature 1.5, and 10,000 sequences sampled per round.

Round	Mean Burial Frac.	# Novel	# CATH	Struct. Hit Rate
1	0.433	111	1166	0.719
2	0.444	192	1190	0.641
3	0.438	206	1178	0.589
4	0.451	240	1155	0.632
5	0.438	269	1171	0.549

One example, variant 2_385 appears spurious, with a TMscore = 0.526 hit to CATH 1.10.150.130 and an inverse-folding landscape littered with "metastable" analogs with sub-0.5 TMscores upon alignment to the foldtuning-emitted template, suggesting that it is not novel, but a noised version of the natural tyrosine recombinase N-terminal domain (Figs. 5.4- 5.5, Table S5.2). The remaining nine variants, by contrast, impute high stability *in silico*, with strong funnel-esque association between lower-energy folded-states and high TMscore alignments to their putative novel templates and most-if-not-all inverse-folded versions clearing the rough energy targets of < -2.2 REU/aa and TMscore > 0.5 (Fig. 5.5). By eye, TMscore, and RMSD, these nine are clearly distinct from their closest CATH counterparts and, annotating by hand, are distributed across all-α (5_4799, 4_2316, 3_8774, 4_6556, 2_3053), $\alpha+\beta$ (5_4773, 4_6411), and α/β (5_111, 3_5721) topologies (Fig. 5.4). Altogether, this constitutes strong evidence that structure-first foldtuning is able to target novel protein structures with meaningful fitness- and topology-agnostic selection criteria, extracting new-to-nature domains with broad shape diversity from a PLM by steering with synthetic sequences that impart supersecondary structural innovation.

5.3 Conclusion

Expanding our novelty-tinged sights from one-dimensional sequences to three-dimensional structures, we jumped headlong into a long-simmering debate in biophysics and structural biology over the existence and frequency of folded domains

with structures unlike anything found within the bounds of natural protein-space. We conceived and effectuated two radically different methods for probing new-to-nature regions of protein structure-space. These two methods are joined only in that they are both PLM-informed. In one, we endeavored to grow up and fill out fold-space from scratch using an evolutionary algorithm steered by PLM-driven estimates of sequence and structure reasonableness, landing on 8-15 novel folds in the course of tallying 510 natural ones, all collected from 3285 individual sequences/domains. In the second, we revisited foldtuning and flipped the script to enrich for structural novelty, honing in on anywhere between several hundred and one thousand novel folds depending on stringency, close to on par with the 2395 natural ones detected, stemming from a pool of 49,992 individual sequences/domains in total. The rates of fold discovery — roughly 1-in-200 for the evolutionary algorithm and 1-in-50 for structure-first foldtuning — are striking when considering that segmenting and searching the UniRef50 portion of the AlphaFoldDB added new superfamilies to CATH at a rate closer to 3-per-million.

All of these efforts used structure prediction models and structure-based search methods; the difference-maker behind our rapid fold emergence rates appears to come back to our use of PLMs and their capacity to credibly evaluate sequence motifs and now structure motifs that emanate from different generative rules than the operative ones of nature. Yet again, PLMs prove to be the ideal agents of a novelty-first design philosophy. The obvious current limitation of this work is that despite the extra confidence imparted by inverse-folding landscape characterization turning up whispers of folding funnels and reasonable physical driving forces, the ultimate arbiter of whether we have landed on structural novelty must be experimental structure determination. In the interim, however, our findings align squarely with the position that permissible structure-space is much broader than that covered by nature, and that, conjecturing a step further, there may exist numerous fold ensembles sufficient for the essential processes of life, arising or not based on initial conditions and/or population size effects.

5.4 Methods

Fragment Library Assembly for Genetic Algorithm

The initial fragment library for the structure discovery genetic algorithm was assembled via a modification of the eeMHMC method first introduced in Chapter 4 (Section 4.4). The first modification is to the form of the energy function, where the "exploit" term $S_{i \to j}$ is replaced by a term rewarding predicted structural contact

density, so that Eq. 4.1 is replaced by

$$\Delta E_{i\to j} = (\log L_j - \log L_i) + w_c \frac{1}{n^2}(\sum_{kl} C_{j,kl} - \sum_{kl} C_{i,kl}) \tag{5.1}$$

where n is the fixed sequence length and C_i, C_j are binary contact matrices s.t. $C_{i,kl} = 1$ indicates that residues k and l of sequence i are predicted to be in physical contact within < 8 Åin the corresponding three-dimensional structure. Contact matrices are inferred from the contact_prediction head of ESM2-650M, simultaneously with embedding and log-likelihood calculation.

The acceptance probability for a proposed single point mutation move from s_i to s_j remains unchanged from Eq. 4.2, accounting for the change in definition of $\Delta E_{i\to j}$.

The second modification is the use of replica-exchange MHMC (RE-MHMC; RE-eeMHMC for **R**eplica-**E**xchange **e**xplore-**e**xploit **M**etropolis-**H**astings **M**onte **C**arlo. RE-MHMC monitors several chains simultaneously, sampling the same landscape at different temperatures, thereby balancing riskier less-local moves by "hot" chains with more conservation local moves by "cold" chains. Adjacent chains in the temperature array attempt to swap positions on the landscape (and their respective sequences) periodically at a stochastic frequency λ; the proposed swap move between chains i, j is accepted with probability

$$p_{i\leftrightarrow j} = \min\{1, \exp[(E_i - E_j)(\beta_i - \beta_j)]\} \tag{5.2}$$

where as always the $\{\beta_i\}$ refer to thermodynamic β, the inverse of the sampling temperature T.

A total of 800 fragments were generated, running for $n = 5000$ steps, stochastically attempting to swap a uniformly randomly selected pair of adjacent random chains at a rate of $\lambda = 0.01$ swp/step, 5 chains with inverse temperatures $\beta = \{20, 13.\bar{3}, 10, 8, 6.\bar{6}\}$ from "cold" to "hot," and $w_c = 1$. Initial sequences for all chains $\{s_0\}$ were random amino-acid strings of length 40; the coldest chain ($\beta = 20$) sequence at step 5000 was added to the library.

Structure Discovery Genetic Algorithm

The structure discovery genetic algorithm begins by sampling an initial population P_0 of 100 fragments from a fragment library assembled as previously described. For a fixed number of rounds, the k-th round proceeds by:

1. Generating 20 new variants from P_{k-1}. A pair of variants is generated by drawing two sequences uniformly at random from P_{k-1}, performing a crossover operation with the number of crossover points $n_{cross} \sim \text{Poisson}(\lambda = 1.535)$ and the locations of the crossover points uniformly distributed over the sequence length(s), and performing a mutation operation with the number of mutations $n_{mut} \sim \text{Binom}(n_i, \lambda = 0.05)$ and mutation locations and identities uniformly distributed over sequence lengths.

2. Evolving the new (k)-th round variants through eeMHMC with the modified energy function found in Eq. 5.1 for $n = 5000$ steps, with $w_c = 1$, $\beta = 10$, and adaptive temperature adjustment at 100-step intervals.

3. Adding the evolved variants to P_{k-1} to form P_k.

4. Predicting structures and computing the amino-acid surface-area burial fraction for *all* sequences in P_k.

5. Selection for burial fraction, maintaining a target constant population size of 100.

The above procedure repeats up to a desired number of generations (200 in this study). To enforce constant population size while stochastically eliminating sequences from the population, we note that, if the number of surviving sequences after round k is to be $|P_k| = N_{sel}$, then the expectation of N_{sel} must be

$$E[N_{sel}] = N f_{sel} \tag{5.3}$$

where N is the temporary population size after new variants have been added but before any have been removed, and f_{sel} is the fraction of sequences that are to survive. We can additionally write that

$$E[N_{sel}] = \sum_i E[n_i] = \sum_i \Pr(\Theta_i = 1) \tag{5.4}$$

where $\Theta_i \sim \text{Bernoulli}(p_i)$ for some mathematically appropriate p_i, as the survival of a given sequence is independent of the survival probability of all others. We have the choice of the form of p_i and so take $p_i = \exp[-\beta_k(0.8 - \gamma_i)]$, where β_k is a

sampling hyperparameter to be determined and γ_i is the burial fraction of sequence s_i.[8]

$$Nf_{sel} = E[N_{sel}] = \sum_i E[n_i] = \sum_i \exp[-\beta_k(0.8 - \gamma_i)] \tag{5.5}$$

and making the simplifying assumption that the $\{\gamma_i\}$'s are roughly normally distributed, or at least not skewed,[9] we can say

$$Nf_{sel} = E[N_{sel}] = \sum_i E[n_i] \approx N \exp[-\beta_k(0.8 - \bar{\gamma}_i)] \tag{5.6}$$

where $\bar{\gamma}_i$ is the mean of all calculated burial fractions in the temporarily augmented population P_k leaving only algebraic rearrangement to solve for our lone sampling hyperparameter β_k, effectively a selection inverse temperature, as

$$\beta_k = \frac{-\log f_{sel}}{0.8 - \bar{\gamma}_i} \tag{5.7}$$

This completes the material necessary to specify and implement the structure discovery genetic algorithm.

Structure-First Foldtuning

Foldtuning was performed and implemented essentially as described in Chapter 3 and Section 3.4, with the following modifications: (1) generation of 10,000 sequences per round in batches of 250, (2) selection of sequences satisfying structural compactness (amino-acid surface burial fraction > 0.5) and novelty (no CATHDB50 hit with TMscore > 0.5) criteria, and (3) ranking of filtered, validated round n sequences for round $n + 1$ finetuning in descending order of amino-acid surface burial fraction.

Selection of Novel Folds for Computational Characterization

For the genetic algorithm experiment, all fifteen putative novel folds were advanced to the computational validation and characterization. For the foldtuning-based

[8]Note that if $\gamma_i > 0.8$, then this would imply $p_i > 1$. Formally, we ought to say $p_i = \max\{1, \exp[-\beta_k(0.8 - \gamma_i)]\}$, empirically, however, the burial fraction for even exceptionally well-packed and folded protein domains is bounded above by $\gamma_i = 0.8$. See Fig. 2.3D.

[9]This assumption is empirically justified for ESM2-generated sequences, referring again to Fig. 2.3D. It seems reasonable then to extrapolate this claim to sequences being evolved/sampled on a landscape subject to an ESM2-based energy function.

experiment, 1018 putative novel folds were initially cumulatively identified over five rounds of structure-first foldtuning. To remove redundancy, predicted structures of the 1018 were clustered with FOLDSEEK at a similarity threshold of TMscore = 0.5, decreasing the number of templates to 916. Given that the structural diversity of the whole AlphaFoldDB50 runs deeper than that of the CATHDB50 subset, the 916 remaining putative novel folds were searched, again using FOLDSEEK, against the entire AlphaFoldDB50, dropping structures with any single hit with alignment region TMscore > 0.5. This reduced the number of templates to 762. These 762 templates were ranked in order of decreasing surface-area burial fraction and the top 100 carried through for inverse-folding and energy scoring validation. For Fig. 5.4 and Fig. 5.5, only the further top 10 of these top 100, as ranked by lowest (most-stable) mean Rosetta-scored energy over all inverse-folded sequences were are depicted.

Structure Prediction and Assignment

All structures were predicted with default ESMFold inference parameters as in Lin et al. (2023). Predicted structures were annotated to CATH domain labels via FOLDSEEK structure-based search against the prebuilt CATHDB50 database running in accelerated TMalign mode(Lau et al., 2024). The consensus CATH domain was defined as the fold 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.

Basic Chemical Property Calculations

Amino-acid surface area burial fraction was calculated using custom code and reference individual amino-acid surface areas (HMS Bionumbers: 103239).

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.

Validation of Inverse-Folding Sequences and Structures

For both the genetic algorithm and foldtuning-based experiments, 200 sequences were generated per structural template with ProteinMPNN, using the vanilla—v_48_020 model, sampling temperature 0.2, backbone noise 0.1 $Å^2$ backbone noise, and forced

omission of the rare/ambiguous amino acids B, J, O, U, X, and Z (Dauparas et al., 2022). Within each batch of 200, sequences were downclustered at 60% sequence identity with MMSEQS2, structures predicted with ESMFold, and queried against the template structure with FOLDSEEK in TMalign mode using the standard TMscore > 0.5 threshold as confirmation of a global match.

5.5 Supplemental Material

Supplemental Figures

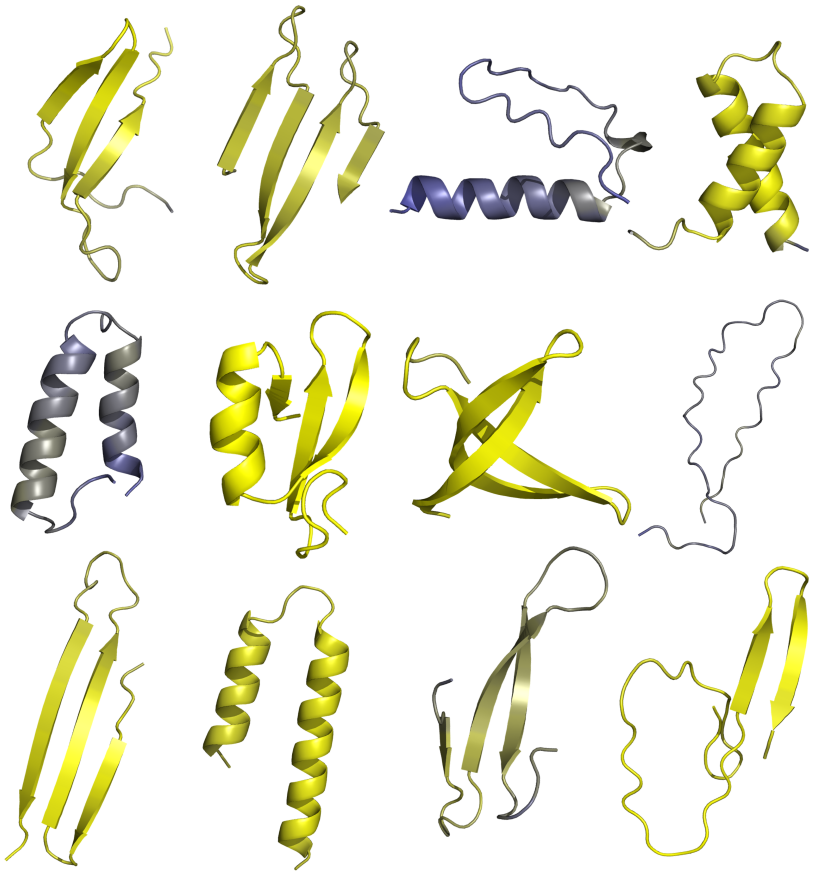

Figure S5.1: **Example structure fragments generated by RE-eeMHMC.** 10 of 800 structure fragments predicted from sequences designed by replica-exchange explore-exploit Metropolis-Hastings Monte Carlo sampling (RE-eeMHMC). Individual structures are colored by ESMFold pLDDT; yellow=high, blue=low.

Supplemental Tables

Table S5.1: **Metadata and structural alignment metrics for closest CATH domain Foldsheek hits to 15 novel folds proposed by the genetic algorithm approach.**

	Closest CATH Hit					
Novel Fold	ID	Name	PDB/AFDB	Pos.	TM	RMSD (Å)
0CF85E_R97	2.60.40.60	Cadherins	0Q7TSF1	383-484	0.428	7.7
120FD5_R140	1.20.5.4130	n/a	A0A0P0YA47	9-126	0.351	5.8
244D7D_R143	1.10.533.10	Death Domain, Fas	Q4QQS0	2-94	0.491	4.1
26D32B_R192	2.40.160.200	LURP1-related	A0A0K3ARQ4	139-303	0.500	6.3
3733B8_R10	2.30.30.170	n/a	Q2FZK7	945-1013	0.430	6.5
794026_R125	1.10.8.430	Helical domain of apop...	Q6Z392	364-452	0.319	7.8
9D1265_R55	1.25.40.10	Tetratricopeptide repeat...	Q9LEX5	342-409	0.540	3.0
A0A7B8_R123	1.10.472.10	Cyclin-like	F4IWI9	175-2664	0.368	7.7
A49A4F_R116	1.10.260.40	λ repressor-like DNA-bind...	1ic8A	87-180	0.319	5.4
A78532_R160	1.20.140.150	n/a	Q7YTM8	1-160	0.396	5.1
B4RC4F_R164	3.90.1150.210	F-actin capping protein...	3aa7B	90-244	0.422	6.9
BC29B7_R55	1.10.357.10	Tet repressor, domain 2	1Z77A	47-200	0.430	6.7
C86FA9_R143	3.30.1520.10	Phox-like domain	Q54S15	808-935	0.477	6.0
DB6817_R173	1.10.520.10	n/a	K7VNV5	33-159	0.393	5.6
F99539_R114	1.10.10.60	Homeodomain-like	1ic8B	203-276	0.473	4.1

Table S5.2: **Metadata and structural alignment metrics for closest CATH domain Foldsheek hits to 10 novel folds proposed by structure-first foldtuning.**

	Closest CATH Hit					
Novel Fold	ID	Name	PDB/AFDB	Pos.	TM	RMSD (Å)
5_111	1.10.472.10	Cyclin-like	I1M2D8	39-142	0.514	4.6
5_4773	1.10.10.10	Winged helix DNA-bind...	Q2FWL6	1-80	0.392	11.8
5_4799	1.10.472.10	Cyclin-like	Q10QA2	94-195	0.422	9.1
4_2316	1.10.533.10	Death Domain, Fas	F8VQ39	371-466	0.479	7.5
2_385	1.10.150.130	Tyr recombinase, N-term...	2keyA	1-112	0.526	4.4
3_8774	1.10.472.10	Cyclin-like	P51946	41-159	0.459	5.2
4_6556	1.20.920.10	Bromodomain-like	A0A1I9LTJ3	289-399	0.402	5.4
4_6411	1.20.960.30	Mitochondrial import rec...	1uujA	2-77	0.419	3.4
3_5721	3.30.980.10	Threonyl-trna synth...	Q9VUJ0	131-293	0.309	5.0
2_3053	1.10.10.1440	PHAX RNA-bind...	2xc7A	1-104	0.398	5.6