Rewriting the sequence and structure rules of deep protein space

Thesis by
Arjuna Michael Subramanian

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended September 15, 2025

ACKNOWLEDGEMENTS

"A man cannot step into the same river twice,
because it is not the same river, and he is not the same man."

Heraclitus might have been miserable dinner company, if his peers are to be believed, but 2,500 years on, he holds a legitimate claim to the title of best metaphor for the PhD process — or any life transition, for that matter. As I prepare to venture out into my next chapter, into a world that is indubitably much changed, I would pause and take a moment to reflect and to recognize many of the people responsible for why I too am not who or what I was six years ago, as a scientist or as a man.

To my advisor, Matt Thomson — my mind drifts, inexorably, to Proverbs 27:17 — *as iron sharpens iron, so one man sharpens another*. Your patience has been endless; your willingness to entertain my most outlandish ideas immense. Day in and day out, you have encouraged my intellectual curiosity and showed me how to de-risk where it would have far easier to dictate "no." In a day and age where none of it is valued enough, you have helped me learn how to recover from setbacks personal and professional, how to handle difficult and unvarnished feedback, and how to not take criticism — no matter how stinging — so personally. You have not merely humored my thoughts of pivoting from science to public service and back again, but actively supported these dalliances with futures outside of academia, temporary as they have been, at least for the moment. I am a more self-assured, more fearless, more resilient scientist and person because of your mentorship, and I can only hope that I am able to foster a fraction of that same growth in my own students down the road.

To my many other mentors in graduate school and on the paths approaching, surrounding, and branching, I am beyond grateful for the time and wisdom that you have shared with me. To Justin Bois — I must have done much right or much wrong to be asked back for three *different* courses in as many years — hopefully the former, as I have never met anyone who cares more about the sum total of student learning, well-being, and preparation for the world beyond the tower walls. From you, I have taken away as much about being an upstanding member of society and pillar of the broader community as I have about how to be an effective and inspiring teacher and leader.

To Martin Wühr — passing through Princeton is always a homecoming for me in part because I feel as welcome in your lab today as on the day I first walked in. Even of your undergraduate students, you demanded real independence and ownership of our science and its ever-evolving story; that I came to grad school feeling — with a fair dose of hubris, it turns out, but methinks that a universal tale — ready to kick the doors down, is a testament to your guidance and high expectations.

To my committee members — Steve Mayo, Richard Murray, Erik Winfree — you have each, in your own way, done much to keep me looking and moving forward with a keen eye for unexpected points of connection and crossover, whilst maintaining focus on the goals dead ahead.

Cato wrote that one of the great virtues worthy of praise in man is to "speak briefly and to the point." Those who know me best know that I often fail — as I am doing once again here — to heed those words. They have lived it doubly if they have ever been forced to share a workspace with me. To the many colleagues and friends who have put up with my chattering and commentary over the years — thank you for the tolerance, the indulgence, and the mental stimulation. Even though, yes, it probably would have been better for all concerned had I just channeled that energy and verbal disgorgement into a podcast launch. You have all touched my time here in profound ways, and I would make the following special, non-exhaustive, remarks.

To Pranav — thank you for lifting my spirits when life events tried their hardest to bring me tumbling down — and, as importantly, for calling me out whenever I tried to wallow in self-pity. "Let's take a walk" and "conference room" are just two phrases that hit differently now.

To Shichen — it often feels like my PhD began anew in year three. Thinking back, that's in large part because you have been a constant presence ever since — to commiserate, to carry on, to carouse, and to contemplate.

To Alec, one man's lunatic is another man's visionary — you are one of the broadest and most creative thinkers I know — case in point, I cannot believe that out of all things, we bonded over moral reframing theory. Attempting to run our own voter study was quite the choice, and I am glad to have tackled it with you, even if I fell down on the job.

To James — my second-favorite alumnus of that unranked school in NYC — we are cut from the same cloth. Who else still reads the physical newspaper and dresses up to go out to listen to Richard Dawkins live on a Saturday night?

To Zach — I am sorry for breaking TRILL at so many inopportune times — but I would argue that I have made up for it with my company and our shared affinity for toxin projects that might have landed us on an FBI watchlist.

To Jerry — your kindness with your time and expertise never went unnoticed, and I could ask for no better academic peer role model for post-PhD endeavours.

There are so many more people who have changed my Caltech experience for the better, without having to handle my distractions in and around the lab. My gratitude goes out to the compatriots with whom I started in 2019 — and who have weathered the same global challenges — in particular, to Tom, Andy, Mike, steadfast friends from the first day of bootcamp (or even earlier). The same goes for the 50+ softballers who have rotated through five iterations of the BMB Bombers. And to the teammates I booted up alongside for Pasadena RFC, brothers joined for eternity by the black rose.

I would be remiss, of course, to not mention the kindred spirits of the Caltech Alpine Club. From climbing, mountaineering, and skiing in earlier times, to ultrarunning and cycling today, so many folks have made the last several years unforgettable on journeys stretching everywhere from local roads and trails to the high peaks of Colorado, Utah, and the Eastern Sierra. Now I am doomed to have to think twice, thrice, quadrice(?), every time I consider venturing back to my flatland roots.

To Elle and Greg — and Poochio...and Woody...and Vinny (RIP, sweet pup) — I am so thankful that a chance encounter led to such deep friendship, in times of adventure and the more mundane. Here's to more Colorado running, climbing, and peak-bagging around the corner!

To Peter and Kate — I am in your debt more than you can possibly know. Although I wish we had come (back) into each other's lives under happier circumstances than my landing on your doorstep with LA literally aflame, my existence is richer, fuller, and more joyful for your friendship and generosity of spirit. And seeing Simba and Regina become friends too — or at least tolerate each other — melts my heart as much today as it did in January.

Speaking of Regina — your contributions to early drafts of this thesis were much appreciated. However, I have taken the liberty of removing such feline turns of phrase as "ni5u3lt," ";eyut3a   ," and the surefire literary masterpiece "   ," better known by its alternate title of "cat standing on the space bar."

To my family — how do I even attempt to put the whole plethora of feelings and emotions — gratitude, admiration, thankfulness, respect, affection, devotion — on a single page? I owe everything that I am, everything that I hope to be, to you. You have instilled in me that scholarship is never complete, that there is grace in admitting where we know not, and to start from the best arguments made from the other side. Sitting here, composing myself to pick up the mantle from the two generations of PhDs before me, it is your faith, your love, that propels me onwards.

ABSTRACT

With a 20-letter alphabet, conceivable protein sequence-space is enormous; sparks of structure and function are vanishingly rare. Despite massive advances in AI-guided protein design, we remain largely ignorant of the sequences and structures that populate the depths of protein space more than a handful of mutations away from what nature has tried. In this work, we leverage the potential of one specific class of AI protein model — the protein language model, or PLM — to internalize the essential features of the protein sequence-structure map while retaining the capacity to explore its extremes. Guided by a "novelty first, fitness next" mentality, we harness this balance towards systematic discovery of new-to-nature sequences and structures throughout deep protein space.

In the first section, we dissect the ability of PLMs to explore natural and novel regimes of sequence and structure during free generation. We find that while these models readily emit novel sequences encoding artificial proteins that appear biophysically feasible *in silico*, they fail to completely or representatively capture the known distribution of natural protein structures. We expose a fundamental tradeoff between the ability of a PLM to generate with sequence novelty or structural coverage but not both simultaneously; prioritizing sampling of far-from-natural sequences triggers a collapse to a handful of simple structural motifs and disordered regions.

Turning this sequence novelty vs. structural breadth tradeoff to our advantage, the second section is devoted to the development of "foldtuning" — a structure-preserving, sequence-remodeling engine for navigating the far corners of sequence-space with PLM-based probes. We successfully scale and deploy foldtuning for $>$ 700 targets, pushing artificial sequences past the point of detectable homology to any real protein documented in nature, discovering novel sequence-level semantics and grammar for mimicking known protein folds, and accessing potential reservoirs of downstream structural and functional innovation. Experimental validation of select targets reveals that foldtuning produces realizable and functional binders in contexts including a toxin/antitoxin system and peptide hormone signaling.

Shifting to focus on structural novelty, the final section introduces two PLM-driven methods for the discovery of new-to-nature structures. We show that with appropriate steering functions, PLMs readily yield well-structured domains (featuring diverse secondary and supersecondary elements) outside the several thousand such families cataloged from among known proteins. Overall, this work makes substantial inroads towards the challenge of locating viable far-from-natural regions of protein density across the *global* sequence-structure map, and revises our notions of the physical constraints on sequence and structure in valid proteins. Moreover, it sets the stage for future assembly of synthetic biological systems composed fully of new-to-nature parts and ultimately for modeling efforts that close the design loop from sequence all the way to complex phenotype.

PUBLISHED CONTENT AND CONTRIBUTIONS

Lourenco, Alec Luiz, Subramanian, Arjuna Michael, Spencer, Ryan K., Miao, Jiapei, Anaya, Michael, Fu, William, Chow, Eric D., and Thomson, Matt. Protein CREATE enables closed-loop design of de novo synthetic protein binders, January 2025. URL `https://www.biorxiv.org/content/10.1101/2024.12.20.629847v2`.

A.M.S. designed protein binders, analyzed data, and participated in the writing of the manuscript. Chapter 4 includes content based on this article.

Subramanian, Arjuna M. and Thomson, Matt. Rapid discovery of new-to-nature protein domains by novelty-first forcing of language models, October 2025. URL `https://www.biorxiv.org/content/10.1101/2025.10.02.679910v1`.

A.M.S. conceived of the project, developed computational methods, performed computational experiments, prepared and analyzed the data, and wrote the manuscript. Chapter 5 is adapted from this article.

Subramanian, Arjuna M., Martinez, Zachary A., Lourenco, Alec L., Yuan, Sonia C., Liu, Shichen, and Thomson, Matt. Unexplored regions of the protein sequence-structure map revealed at scale by a library of foldtuned language models, September 2025a. URL `https://www.biorxiv.org/content/10.1101/2023.12.22.573145v3`.

A.M.S. conceived of the project, developed computational and wet-lab methods, performed computational and wet-lab experiments, prepared and analyzed the data, and wrote the manuscript. Chapters 3-4 collectively update this article.

Subramanian, Arjuna M., Martinez, Zachary A., and Thomson, Matt. Pretrained protein language models choose between sequence novelty and structural completeness, October 2025b. URL `https://www.biorxiv.org/content/10.1101/2025.10.01.679905v1`.

A.M.S. conceived of the project, performed computational experiments, prepared and analyzed data, and wrote the manuscript. Chapter 2 is adapted from this article.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

"A clone is a dead end, a clade is a promise of immortality." — Freeman Dyson

1.1 Design, exploration, and the "concept of a protein-space"

Nature has likely sampled only a fraction of all protein sequences and structures allowed by the laws of biophysics. High-quality genomic and metagenomic databases together contain ~ 10^9 – 10^{10} *unique* protein sequences distributed across the extant tree of life (Jumper et al., 2021; Mirdita et al., 2022). The observed protein catalog presumably reflects selection over multiple timescales for factors from favorable folding thermodynamics and kinetics, to metal center scaffolding and cofactor usage, to essential DNA/RNA-binding and catalytic functions (Alva et al., 2015; Baker, 2000; Dupont et al., 2010; Vyas et al., 2021; Watters et al., 2007). It likely also reflects the fixation of sequence and structure elements that abounded at the dawn of life. After all, statistical analyses of experimentally determined structures and simple physical models that forgo function and reduce proteins to lattices and spin-glasses agree that backbones vary widely in their "designability" — that is, in the number and diversity of sequences that encode one versus another (Bornberg-Bauer, 1997; England and Shakhnovich, 2003; Govindarajan and Goldstein, 1996; Helling et al., 2001; Li et al., 1996; Yue and Dill, 1995). Evolutionary biologists have long theorized as to how natural sequences and structures are related and organized into a single "protein-space," or, acknowledging Anfinsen's dogma that sequence specifies structure, a unified sequence→structure map where a given sequence encodes exactly one structure, and a given structure is the outcome of many such sequences (Anfinsen, 1973; Choi and Kim, 2006). Thinking more ambitiously, John Maynard Smith pondered how what nature *hasn't tried* might fit in, asking whether functional protein sequences might occupy "two or more distinct networks" separated by evolutionarily uncrossable chasms — entire parallel universes or pocket dimensions of protein density unknown and inaccessible to one another (Maynard Smith, 1970).[1]

Suppose that Maynard Smith was on to something. Suppose that nature's sparse

[1] The title of this section recalls Maynard Smith's seminal 1970 rumination on "Natural Selection and the Concept of a Protein Space." It is indispensable reading for any biologist and may be found at ref. (Maynard Smith, 1970).

sampling of proteins neither exhausts the viable solutions of the sequence→structure map nor offers a neat atlas of direct paths to what's missing. How then ought we approach the search problem of finding and enumerating sequences and structures that are physically realistic, biologically fit, and new-to-nature? The conceivable scope of sequence-space encompassing them all is vast and daunting. Combinatorial scaling with the 20 proteinogenic amino acids translates into $\sim 10^{130}$ strings of length 100; the size, roughly, of a small $\sim$ 10 kDa protein domain. The mass of the visible universe — every last speck of it — would allow for making exactly one molecular copy of each of $\sim 10^{75}$ of those. Evolution on Earth has had 4 billion years to work with; under generous assumptions about population size and reproductive capacity, and if somehow, no point mutation were ever stepped on twice, this would be enough to try out "only" $\sim 10^{40}$.

We might hope then to get lucky and land on one of the as many as 1-in-10^{11} functional sparks strewn about sequence-space according to measurements on random sequence libraries (Keefe and Szostak, 2001; Tong et al., 2021). Or, we might get luckier still, and find these sparks anchoring their own neutral nets, dense local pockets of stability and mutational tolerance reminiscent of Maynard Smith-ian parallel networks derived from distant evolutionary seeds (Bornberg-Bauer and Chan, 1999; Fontana et al., 1993).

Reality has proved a harsher mistress. Decades of painstaking protein design work has elucidated rules and heuristics that illuminate only small corners of the global sequence-structure map. For instance, an alphabet of just three amino-acids — glutamine (polar), leucine (hydrophobic), and arginine (charged) — is enough in certain arrangements to confer globularity and cooperative folding (Davidson and Sauer, 1994; Davidson et al., 1995). Simple hydrophobic/polar patterning — the protein version of "like prefers to associate with like" — is sufficient to generate stable α-helical bundle proteins encoded by novel sequences (Hecht et al., 1997; Kamtekar et al., 1993). Similar alphabet reduction strategies have been applied to mimic small β-barrels like the SH3 domain (Riddle et al., 1997). Deep multiple sequence alignments (MSAs) can capture sparse co-evolutionary signals strong enough to generate artificial variants with comparable stability to natural examples (Lockless and Ranganathan, 1999; Socolich et al., 2005; Süel et al., 2003). Meanwhile, taking a structure-first perspective has produced idealized, minimized, and embellished versions of some of the more ubiquitous natural proteins folds inlcuding TIM β/α-barrels, thioredoxins, and Rossmannoids (Huang et al., 2016b; Linsky et al., 2022; Pan et al., 2020). And experimentally-grounded force fields, saintly patience, and lifetimes' worth of CPU-hrs have realized truly *de novo* folds of all-α, all-β and mixed α, β content (Alford et al., 2017; Kim et al., 2023; Kuhlman et al., 2003; Minami et al., 2023; Sakuma et al., 2024).

On the back of these impressive advances, an explosion in AI-driven modeling, and more, protein design may indeed be hitting its prophesied "coming of age" moment (Huang et al., 2016a). But discovering and characterizing sequence and structural novelty in the far-flung, outlying regions of deep protein-space has remained stubbornly scattershot. This is despite the fact that understanding the makeup of the furthest reaches of protein-space promises to:

1. Empirically answer (without evolutionary biases) the biophysicist's quandary of which three-dimensional structures — including any untouched by nature — are the most designable and what makes them that way.

2. Access untapped ground for protein engineering, improving and expanding target properties from thermostability and solubility, to enzyme substrate specificity, to cell-signaling phenotype.

3. Reveal minimal and/or alternate "rulesets" for assembling viable sequences and structures, a goal with applications and implications for function-centric protein engineering and life in extreme and/or primordial environments alike.

Shedding light on any or all of these questions requires new attitudes, new algorithms, and new assays to reliably reach beyond the dots and glimpses historically offered by protein design and reach whole islands, whole continents of structure and function in deep protein space. The time is nigh to update the sequence→structure map from *hic sunt dracones* into a proper Age of Exploration.

1.2 Protein language models are potent agents of exploration

Where Bartolomeu Dias and Vasco da Gama had the caravel, where the architects of the Space Age had Mariners, Pioneers, and Voyagers, we have the protein language model.

Protein language models (PLMs), as follows from the name, are the children of large language models (LLMs) like BERT or GPT-2, developed for human-derived text and transferred to amino-acid "text" as part of the AI-for-proteins gold rush

(Devlin et al., 2019; Radford et al., 2019). PLMs are intriguing vehicles for searching the sequence→structure map thanks to their ability to balance *exploitation* of internalized knowledge about sequence determinants within natural proteins with *exploration* of other sequence rules that they deem physically plausible.[2] Model size (10^7 – 10^{11} parameters), training data volume (10^6 – 10^8 sequences, generally sampled from UniRef clusters introduced in Suzek et al. (2007)), high-level architecture (e.g. autoencoder vs encoder-only vs decoder-only vs encoder-decoder, etc.), and choices of alphabet/vocabulary discretization (e.g. decomposing sequences into individual amino-acids or longer subsequences) can vary significantly (Chen et al., 2023; Ferruz et al., 2022; Heinzinger et al., 2023; Hie et al., 2022; Lin et al., 2023; Madani et al., 2023). However, modern PLMs share a general organizing principle in that they are composed of stacks and layers of individual transformer blocks that pick out informative patterns and correlations across sequence positions (Vaswani et al., 2017). These patterns are notoriously treacherous to interpret, bordering on a Rorschach test — one transfomer might pick up on α-helical content, another a binding pocket, a third an electrostatic gradient, a fourth a sharp map of 3D contacts — with a few hundred more defying easy biochemical or biophysical explanation (Simon and Zou, 2025; Vig et al., 2021).[3] Whatever these transformers are capturing all together, it's enough to succeed at descriptive tasks (variant prediction, structure prediction) and generative tasks (novel fold and enzyme design), seeping gradually into sequence and structure novelty in the latter case (Madani et al., 2021, 2023; Verkuil et al., 2022).

And among the ever-expanding menagerie of AI-based protein models, PLMs stand out for this emergent exploratory capacity that can bridge the levels of information flow from sequence, to structure, to function. Diffusion models can innovate at the level of structure, but do not handle sequence information at all (Watson et al., 2023). Inverse-folding models that consider the flipped structure→sequence problem can diversify sequence with careful hyperparameter selection, but at the cost of enforcing strict backbone constraints that preclude the sorts of small structural innovations and ornamentations that have conferred new and/or expanded functionalities throughout all of natural protein evolution (Dauparas et al., 2022; Hsu et al., 2022; Pan et al., 2020; Tóth-Petróczy and Tawfik, 2014). The global perturbations required to escape

[2] A bold analogy, appealing to the fundamentals of computational linguistics, is that a PLM learns the production rules of an unrestricted (type-0) grammar that separates semantically meaningful valid proteins from meaningless nonfunctional amino-acid strings (Chomsky, 1959).

[3] Some, including the author, view this lack of interpretability as room for attention (pun fully intended) and growth, although appetite in the field to pursue this point has remained tragically low.

the gravitational pull of natural sequence-space are likewise inaccessible to directed evolution – which searches sequence-space locally under strong stability and fitness restrictions – or to machine learning models trained on high-throughput but unavoidably local fitness data collected in deep-mutational scanning experiments (Fahlberg et al., 2023; Romero and Arnold, 2009; Wilson et al., 2020).

1.3 Organization of the thesis

The overarching goal of this thesis is to **harness the emergent exploratory capacity of PLMs to make the aforementioned global perturbations and systematically discover sequence and structure novelty in deep protein space**. We organize this journey as follows.

In **Chapter 2**, we show that the generative capacity of protein language models does not internalize the body of knowledge of sequence and structure in protein biochemistry and biophysics as perfectly as assumed, headlined by a concerning pathology where sampling with sequence novelty and sampling with structural completeness stand at odds.

Subsequently, we bypass these limitations in **Chapters 3 and 4**, developing, applying, and experimentally validating an original algorithm — that we call "foldtuning" — to systematically search for sequence novelty in deep protein space by using known structures as lodestones.

Finally, we find in **Chapter 5** that the PLM-based techniques created to search for sequence novelty can be reformulated and redirected to discover novel protein folds, completing the crossing of the frontiers of natural protein space.

Chapter 2

BASIC LANGUAGE MODELS ARE SKEWED MIRRORS OF THE PROTEIN UNIVERSE

2.1 Introduction

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

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

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

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

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

2.2 Results & Discussion

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

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

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

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

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

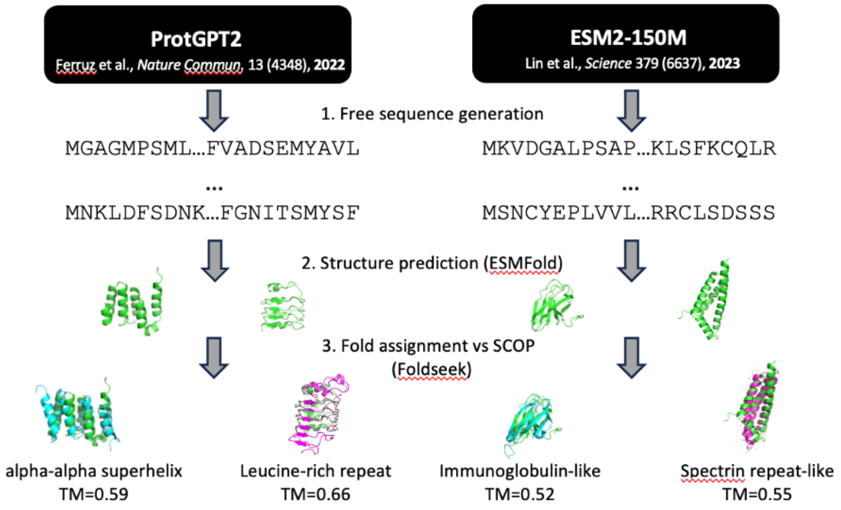


Figure 2.1: **Workflow for structural annotation of pLM-generated sequences.** Schematic overview of pLM generation, structure prediction, and structural search+assignment pipeline for representative models ProtGPT2 and ESM2-150M.

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

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

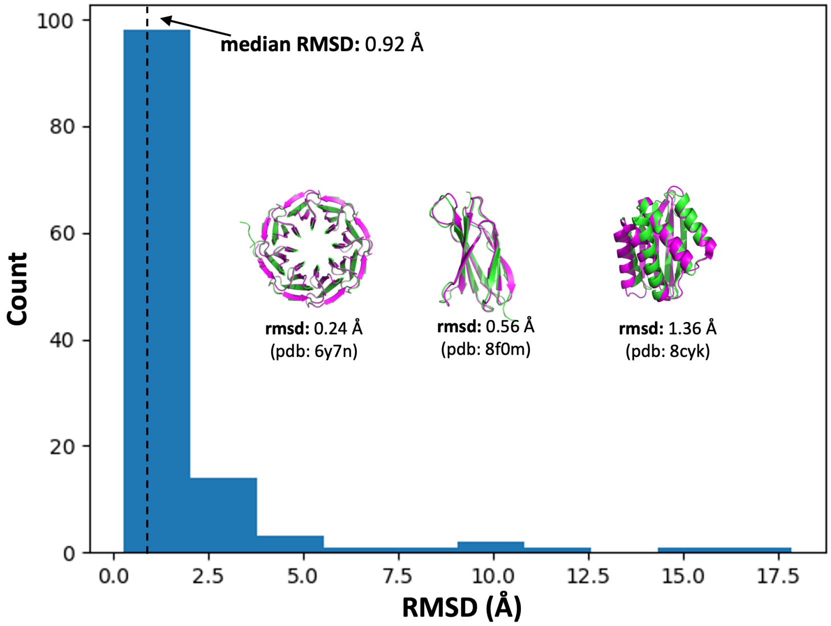


Figure 2.2: **ESMFold achieves single-angstrom structure prediction accuracy on *de novo* designed sequences.** Backbone atom root-mean-square deviation (RMSD; median= 0.92 ± 0.14) for ESMFold predicted structures of $n = 122$ *de novo* designed proteins vs. experimental ground-truth structures, covering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

analysis of model behavior. Bolstering this contention, we assessed the accuracy of ESMFold structural prediction on out-of-distribution samples by evaluating model performance on *de novo* proteins with structures deposited in the Protein Data Bank (PDB) on-or-after the ESMFold training cutoff date of 05-01-2020. Mirroring the training set construction process described in the original ESMFold publication, we filtered out structures with resolution > 9 Å, length ≤ 20aa, rare or ambiguous amino acids (BJOUXZ), or containing > 20% sequence composition of any one amino acid, and clustered remaining sequences at the 40% identity level, obtaining a validation set of $n = 122$ sequences. For each of the 122 sequences, the backbone RMSD was calculated between the ESMFold predicted structure and the ground-truth PDB experimental structure, with a median alignment RMSD of 0.92 ± 0.14 Å and coverage of diverse structure topology classes, indicating sufficient generalization of

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

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

PLM-generated sequences are protein-like

For a first-pass analysis, we consider whether generated sequences and their corresponding (predicted) structures recapitulate the global characteristics of natural proteins. To determine where pLM-generated sequences lie with respect to natural sequence-space, we extract the ESM2-150M final hidden-layer internal representations ("embeddings") of all >400,000 generated sequences and 100,000 diverse natural sequences coding for SCOP fold examples mined from the AlphaFoldDB (Varadi et al., 2022).[4] We reduce dimensionality to 2D using UMAP, and apply a rule-of-thumb that the embeddings of qualitatively similar sequences should co-localize (McInnes et al., 2018). We observe that ProtGPT2-generated sequences separate into two subpopulations, one co-localizing with natural sequences, and a second co-localizing with random sequences (Fig. 2.3A). In contrast, ESM2-150M-generated sequences co-localize most substantially with random sequences. Inverse-folded sequences from ESM-IF1 largely mirror the distribution of natural sequences, implying that they do not represent any significant departure from natural protein-space.

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

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

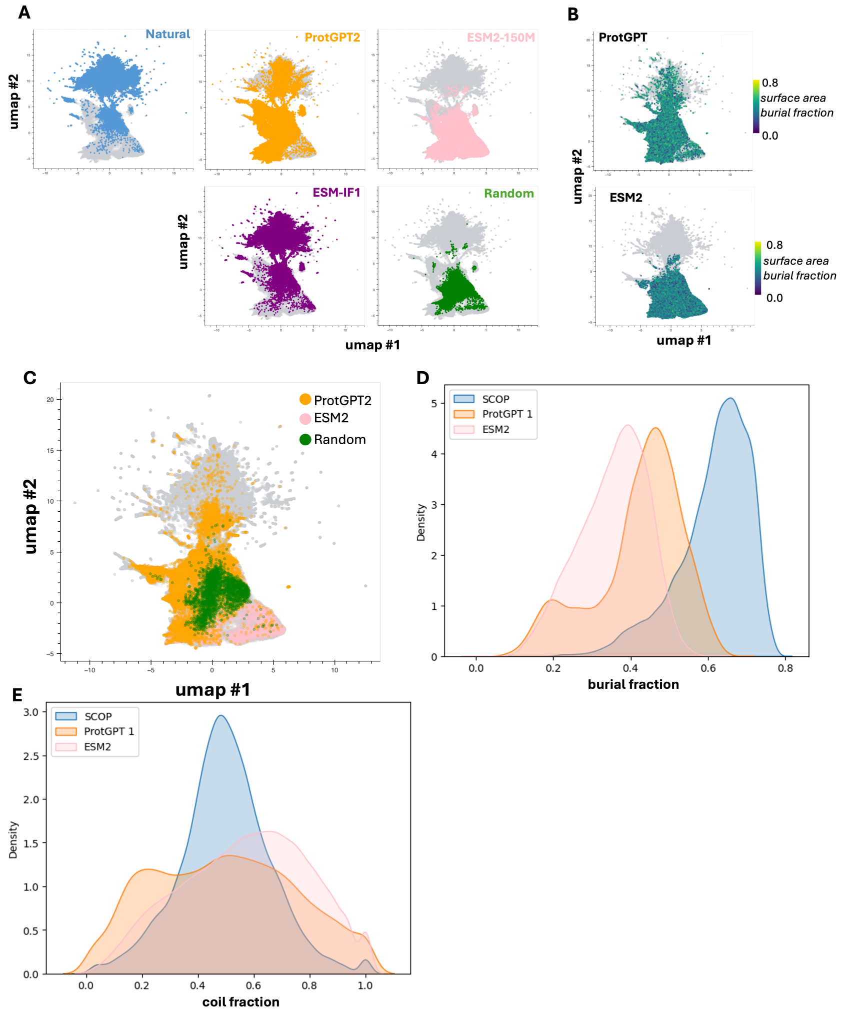


Figure 2.3: **PLM-generated sequences reflect the basic properties of compact, globular proteins.** (**A**) Dimension-reduced UMAP representation of ESM2-150M embeddings of natural, pLM-generated, inverse-folded, and random control sequences. (**B**) UMAP representation of pLM-generated sequences, colored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

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

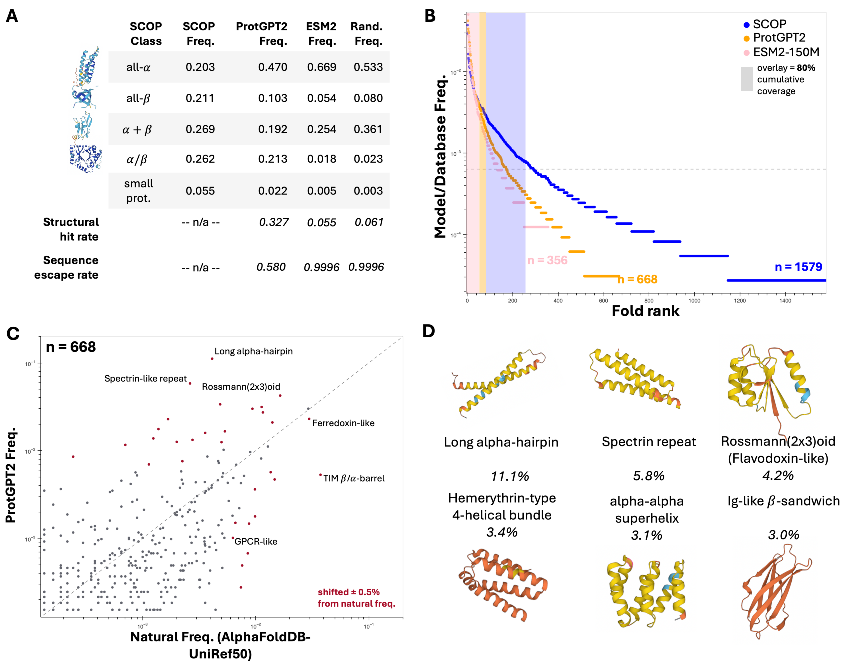
PLM-generated structures do not follow the natural distribution

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

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

Prioritizing sequence novelty shrinks accessible structure-space

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



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

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

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

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

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

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

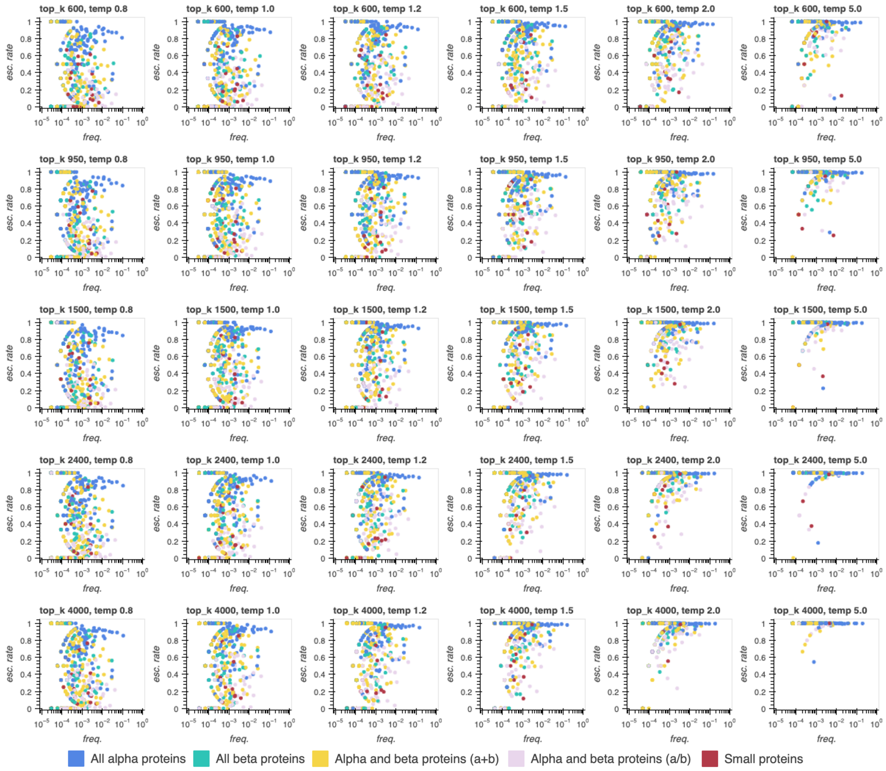


Figure 2.5: **Sequence escape rates increase across most folds as sampling temperature increases, at the cost of a shift towards all-α topologies.** Sequence escape rates for all assigned SCOP folds generated from ProtGPT2 within batches of 100k sequences for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) x several top_k values (number of highest-probability tokens considered in sampling out of 50,256 total; 600, 950, 1500, 2400, 4000.

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

2.3 Conclusion

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

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

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

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

2.4 Methods

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

Sequence Generation from Protein Language Models

For the model comparison experiment, sequences (n = 100000) were sampled from ProtGPT2 by L-to-R next-token prediction with the default best-performing hyperparameters from Ferruz et al. (2022); sampling temperature 1, top_k 950, top_p 1.0, repetition penalty 1.2. The termination condition was set following the 40th token or the first STOP token occurring prior to the 40th token; sequences longer than 100aa were truncated to 100aa as the maximal length. Sequences containing rare or ambiguous amino acids (B, J, O, U, X, or Z) were filtered out as invalid, leaving 99,982 sequences. Sequences were sampled from ESM2-150M (n = 148500), from L-to-R with next-token prediction with Gibbs sampling, with a default sampling temperature of 1, no repetition penalty, and allowing for sampling from the full token distribution. The termination condition was set following the 100th amino-acid or the first STOP token occurring prior to the 100th amino-acid. Truncation and filtering were applied as for ProtGPT2.

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

Sequence Generation from Control Models

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

The inverse-folding control set was constructed by generating three sequences from ESM-IF1 with each of the 36,900 representative structures in the SCOP database as a backbone template, for $n = 110700$ sequences in total. Pre- and post-processing for rare ligands in templates and rare/ambiguous amino acids in outputs, respectively, reduced inverse-folding output to 104,591 sequences. Default hyperparameters for sampling were taken as in Hsu et al. (2022).

Structure Prediction and Assignment

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

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

Sensitive Sequence Search and Novelty Characterization

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

Construction of the SCOP-UniRef50 Sequence-Structure Database

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

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

Basic Chemical Property Calculations

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

2.5 Supplemental Material

Supplemental Figures

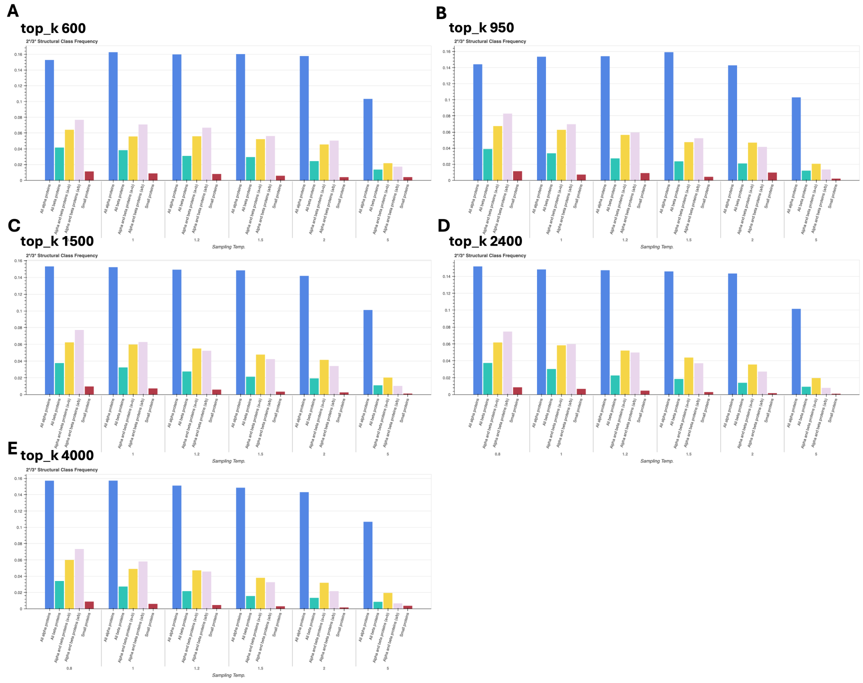


Figure S2.1: **Structure hit rates from base ProtGPT2 decrease as sampling temperature and top_k increase.** Structure hit rates from batches of 100k sequences generated from ProtGPT2 for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) and top_k values (number of highest-probability tokens considered in sampling out of 50,256 total) — (**A**) 600, (**B**) 950, (**C**) 1500, (**D**) 2400, (**E**) 4000; broken down by protein global topology class (α, β, $\alpha + \beta$, α / β, or "small / minimal 2° structure").

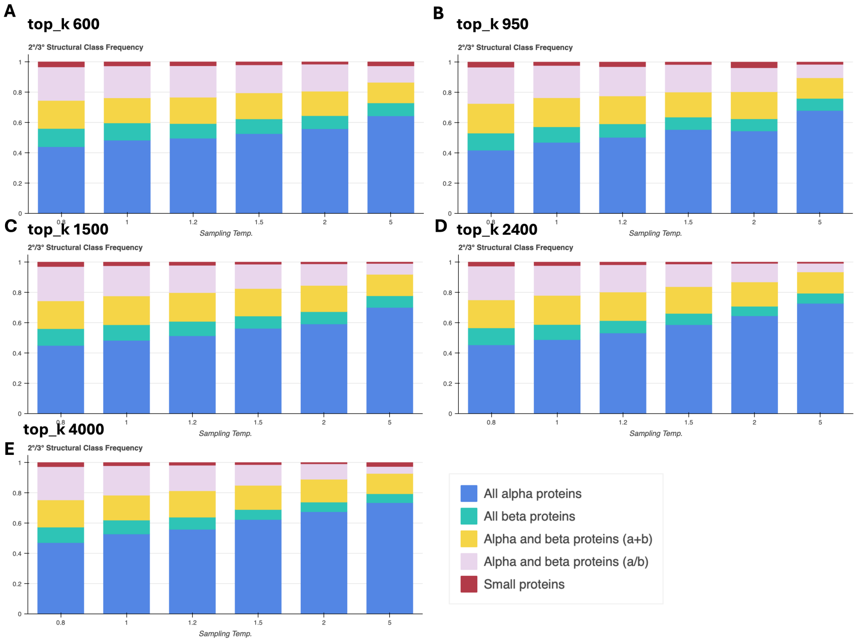


Figure S2.2: **Generated fold distributions shift towards all-α proteins and away from α/β proteins as sampling temperature increases.** Frequency of each protein global topology class (α, β, $\alpha+\beta$, α/β, or "small / minimal 2° structure") among all structure hits within batches of 100k sequences generated from ProtGPT2 for several sampling temperatures (0.8, 1, 1.2, 1.5, 2, 5) and top_k values (number of highest-probability tokens considered in sampling out of 50,256 total) — (**A**) 600, (**B**) 950, (**C**) 1500, (**D**) 2400, (**E**) 4000.

Supplemental Tables

Table S2.1: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k 600.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.095	0.033	0.810
Spectrin repeat-like	α	0.050	0.017	0.871
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.048	0.017	0.146
Immunoglobulin-like beta-sandwich	β	0.036	0.012	0.510
alpha-alpha superhelix	α	0.033	0.011	0.386
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.112	0.038	0.874
Spectrin repeat-like	α	0.056	0.019	0.907
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.014	0.252
Immunoglobulin-like beta-sandwich	β	0.034	0.012	0.596
Hemerythrin-type up-and-down 4-helical bundle	α	0.032	0.011	0.903
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.122	0.039	0.908
Spectrin repeat-like	α	0.063	0.020	0.929
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.014	0.330
Hemerythrin-type up-and-down 4-helical bundle	α	0.036	0.012	0.937
Immunoglobulin-like beta-sandwich	β	0.030	0.010	0.696
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.130	0.040	0.943
Spectrin repeat-like	α	0.068	0.021	0.950
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.013	0.424
Hemerythrin-type up-and-down 4-helical bundle	α	0.040	0.012	0.958
Immunoglobulin/albumin-binding domain-like	α	0.033	0.010	0.955
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.141	0.040	0.969
Spectrin repeat-like	α	0.075	0.021	0.968
Hemerythrin-type up-and-down 4-helical bundle	α	0.044	0.013	0.978
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.012	0.500
Immunoglobulin/albumin-binding domain-like	α	0.032	0.009	0.966
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.154	0.025	0.989
Spectrin repeat-like	α	0.084	0.014	0.989
Hemerythrin-type up-and-down 4-helical bundle	α	0.060	0.010	0.984
alpha-alpha superhelix	α	0.040	0.006	0.905
Immunoglobulin/albumin-binding domain-like	α	0.033	0.005	0.987

Table S2.2: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 950.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.101	0.035	0.841
Spectrin repeat-like	α	0.050	0.017	0.872
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.048	0.016	0.177
Immunoglobulin-like beta-sandwich	β	0.032	0.011	0.534
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.031	0.011	0.342
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.111	0.036	0.893
Spectrin repeat-like	α	0.058	0.019	0.918
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.014	0.273
Hemerythrin-type up-and-down 4-helical bundle	α	0.034	0.011	0.926
alpha-alpha superhelix	α	0.031	0.010	0.571
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.126	0.039	0.930
Spectrin repeat-like	α	0.065	0.020	0.946
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.041	0.013	0.345
Hemerythrin-type up-and-down 4-helical bundle	α	0.038	0.012	0.948
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.030	0.009	0.530
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.136	0.039	0.943
Spectrin repeat-like	α	0.069	0.020	0.969
Hemerythrin-type up-and-down 4-helical bundle	α	0.046	0.013	0.960
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.012	0.491
Immunoglobulin/albumin-binding domain-like	α	0.030	0.009	0.960
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.149	0.039	0.978
Spectrin repeat-like	α	0.076	0.020	0.984
Hemerythrin-type up-and-down 4-helical bundle	α	0.045	0.012	0.976
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.010	0.596
Immunoglobulin/albumin-binding domain-like	α	0.035	0.009	0.974
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.178	0.027	0.991
Spectrin repeat-like	α	0.090	0.014	0.996
Hemerythrin-type up-and-down 4-helical bundle	α	0.064	0.010	0.989
Immunoglobulin/albumin-binding domain-like	α	0.038	0.006	0.986
alpha-alpha superhelix	α	0.035	0.005	0.934

Table S2.3: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 1500.**

temp: 0.8				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.101	0.035	0.847
Spectrin repeat-like	α	0.052	0.018	0.878
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.045	0.015	0.210
Immunoglobulin-like beta-sandwich	β	0.033	0.011	0.555
alpha-alpha superhelix	α	0.031	0.010	0.426
temp: 1				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.119	0.037	0.895
Spectrin repeat-like	α	0.059	0.019	0.918
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.013	0.304
Hemerythrin-type up-and-down 4-helical bundle	α	0.035	0.011	0.930
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.029	0.009	0.472
temp: 1.2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.129	0.038	0.930
Spectrin repeat-like	α	0.067	0.019	0.956
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.012	0.425
Hemerythrin-type up-and-down 4-helical bundle	α	0.040	0.012	0.951
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.029	0.008	0.528
temp: 1.5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.145	0.038	0.963
Spectrin repeat-like	α	0.077	0.020	0.984
Hemerythrin-type up-and-down 4-helical bundle	α	0.047	0.012	0.976
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.011	0.566
Immunoglobulin/albumin-binding domain-like	α	0.033	0.009	0.968
temp: 2				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.158	0.038	0.988
Spectrin repeat-like	α	0.078	0.019	0.989
Hemerythrin-type up-and-down 4-helical bundle	α	0.050	0.012	0.986
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.039	0.009	0.708
Immunoglobulin/albumin-binding domain-like	α	0.034	0.008	0.987
temp: 5				
Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.182	0.026	0.993
Spectrin repeat-like	α	0.094	0.014	0.999
Hemerythrin-type up-and-down 4-helical bundle	α	0.070	0.010	0.994
Ferredoxin-like	$\alpha+\beta$	0.040	0.006	0.984
Immunoglobulin/albumin-binding domain-like	α	0.038	0.005	0.993

Table S2.4: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 2400.**

temp: 0.8

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.106	0.036	0.850
Spectrin repeat-like	α	0.052	0.018	0.894
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.043	0.014	0.210
alpha-alpha superhelix	α	0.032	0.011	0.435
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.031	0.011	0.358

temp: 1

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.125	0.038	0.905
Spectrin repeat-like	α	0.062	0.019	0.939
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.040	0.012	0.353
Hemerythrin-type up-and-down 4-helical bundle	α	0.038	0.011	0.917
Canonical WHD (winged helix domain) fold	$\alpha+\beta$	0.028	0.008	0.446

temp: 1.2

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.138	0.038	0.945
Spectrin repeat-like	α	0.071	0.020	0.959
Hemerythrin-type up-and-down 4-helical bundle	α	0.043	0.012	0.957
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.041	0.011	0.456
Ferredoxin-like	$\alpha+\beta$	0.030	0.008	0.792

temp: 1.5

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.158	0.039	0.976
Spectrin repeat-like	α	0.077	0.019	0.985
Hemerythrin-type up-and-down 4-helical bundle	α	0.052	0.013	0.981
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.038	0.010	0.601
Ferredoxin-like	$\alpha+\beta$	0.033	0.008	0.888

temp: 2

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.167	0.037	0.994
Spectrin repeat-like	α	0.086	0.019	0.992
Hemerythrin-type up-and-down 4-helical bundle	α	0.056	0.012	0.991
Immunoglobulin/albumin-binding domain-like	α	0.041	0.009	0.991
Ferredoxin-like	$\alpha+\beta$	0.036	0.008	0.956

temp: 5

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.190	0.027	0.998
Spectrin repeat-like	α	0.095	0.013	0.998
Hemerythrin-type up-and-down 4-helical bundle	α	0.069	0.010	0.998
Ferredoxin-like	$\alpha+\beta$	0.041	0.006	0.993
Immunoglobulin/albumin-binding domain-like	α	0.041	0.006	0.996

Table S2.5: **Most common SCOP folds generated by base ProtGPT2 at various sampling temperatures with top_k (vocabulary size) 4000.**

temp: 0.8

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.108	0.036	0.855
Spectrin repeat-like	α	0.054	0.018	0.892
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.042	0.014	0.217
alpha-alpha superhelix	α	0.031	0.010	0.448
Hemerythrin-type up-and-down 4-helical bundle	α	0.031	0.010	0.896

temp: 1

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.123	0.037	0.904
Spectrin repeat-like	α	0.065	0.019	0.939
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.039	0.012	0.377
Hemerythrin-type up-and-down 4-helical bundle	α	0.038	0.011	0.930
alpha-alpha superhelix	α	0.028	0.008	0.609

temp: 1.2

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.146	0.039	0.949
Spectrin repeat-like	α	0.071	0.019	0.974
Hemerythrin-type up-and-down 4-helical bundle	α	0.046	0.012	0.967
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.041	0.011	0.544
Ferredoxin-like	$\alpha+\beta$	0.031	0.008	0.812

temp: 1.5

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.161	0.038	0.981
Spectrin repeat-like	α	0.086	0.020	0.991
Hemerythrin-type up-and-down 4-helical bundle	α	0.054	0.013	0.982
Immunoglobulin/albumin-binding domain-like	α	0.039	0.009	0.983
Rossmann(2x3)oid (Flavodoxin-like)	α/β	0.035	0.008	0.699

temp: 2

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.183	0.039	0.997
Spectrin repeat-like	α	0.092	0.019	0.994
Hemerythrin-type up-and-down 4-helical bundle	α	0.062	0.013	0.998
Immunoglobulin/albumin-binding domain-like	α	0.038	0.008	0.995
Ferredoxin-like	$\alpha+\beta$	0.038	0.008	0.970

temp: 5

Fold	Class	Freq.	Abs. Hit Rate	Esc. Rate
Long alpha-hairpin	α	0.196	0.029	0.999
Spectrin repeat-like	α	0.097	0.014	1.000
Hemerythrin-type up-and-down 4-helical bundle	α	0.079	0.011	1.000
Ferredoxin-like	$\alpha+\beta$	0.040	0.006	0.998
Immunoglobulin/albumin-binding domain-like	α	0.038	0.005	1.000

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 sequence 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

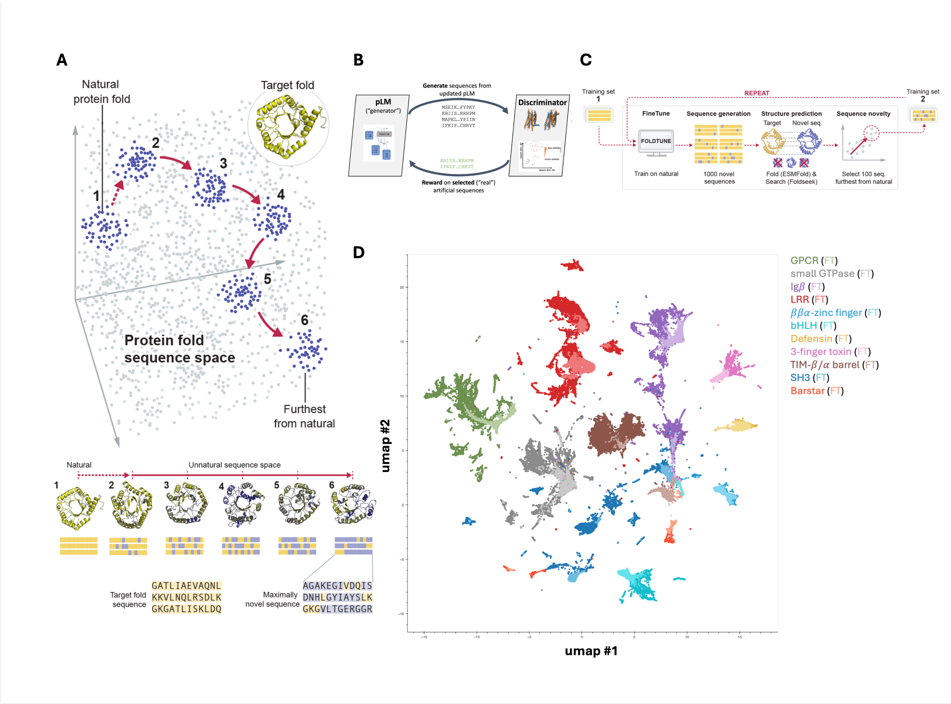


Figure 3.1: **Foldtuning explores far-from-natural sequences encoding alternate versions of natural protein structures.** (**A**) Conceptual overview of foldtuning. Beginning from natural protein sequences coding for a target backbone structure, foldtuning uses a protein language model (pLM)-based strategy 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).

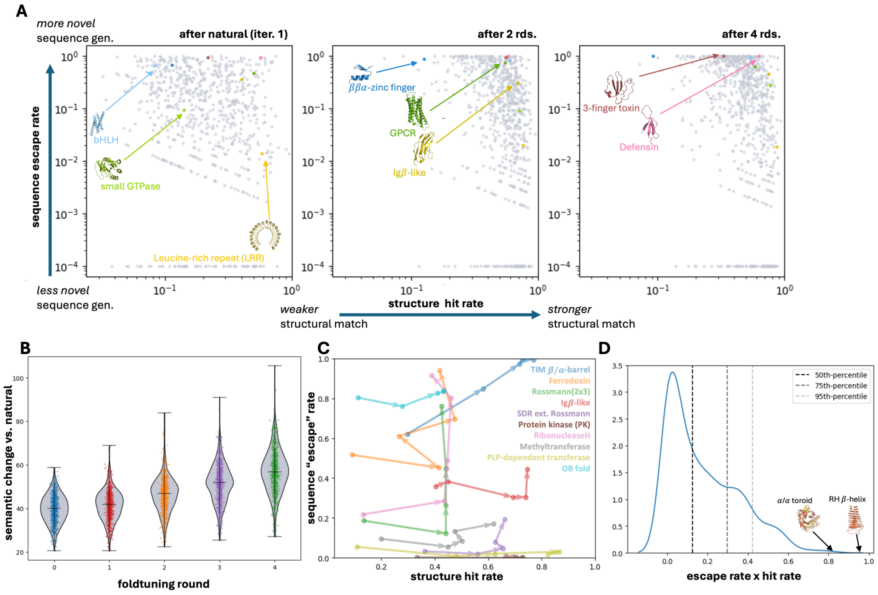


Figure 3.2: **Foldtuned models sample novel sequences for >700 targets.** (A) Sequence escape vs. structural hit rates after natural-only evotuning or two or four rounds of foldtuning for 727 targets. Selected structural/functional targets are highlighted: transcription factors (blue), GPCRs/small GTPases (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.

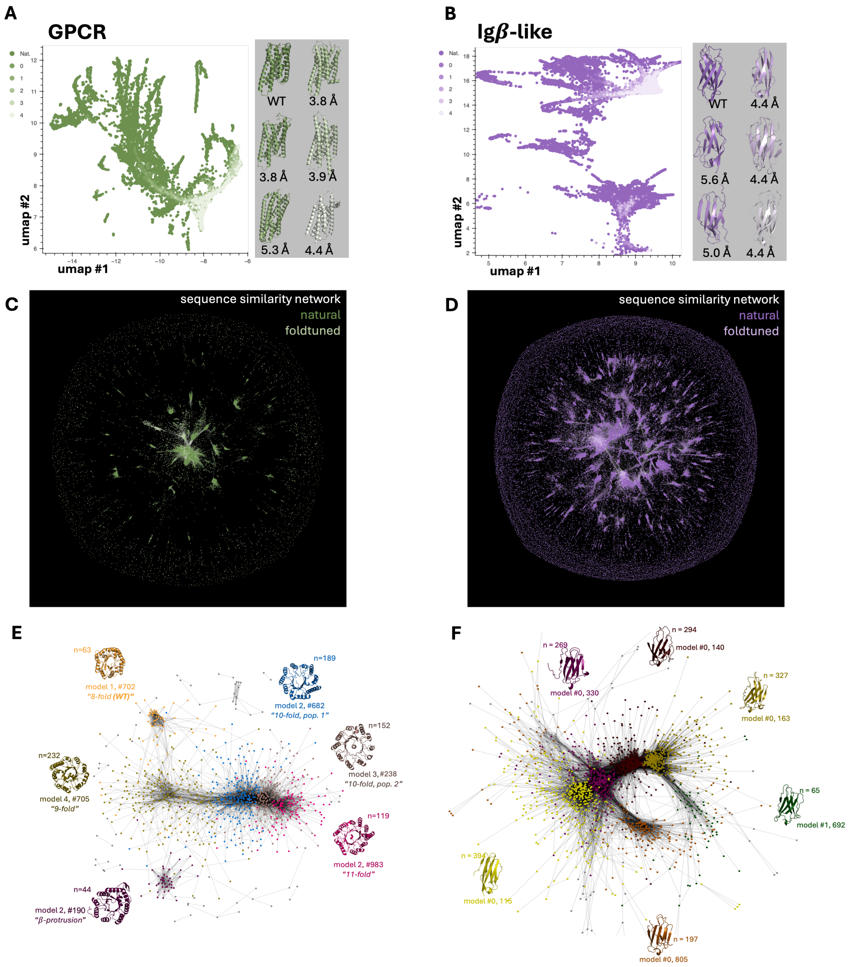


Figure 3.3: **Foldtuning accesses new sequence populations and structural innovations while 'fuzzily' preserving a target backbone.** (**A**) UMAP of round-by-round foldtuning sequence diversification captured by ESM2-650M final-layer hidden states with G-protein coupled receptors (GPCRs, SCOP ID: 2000339) 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).

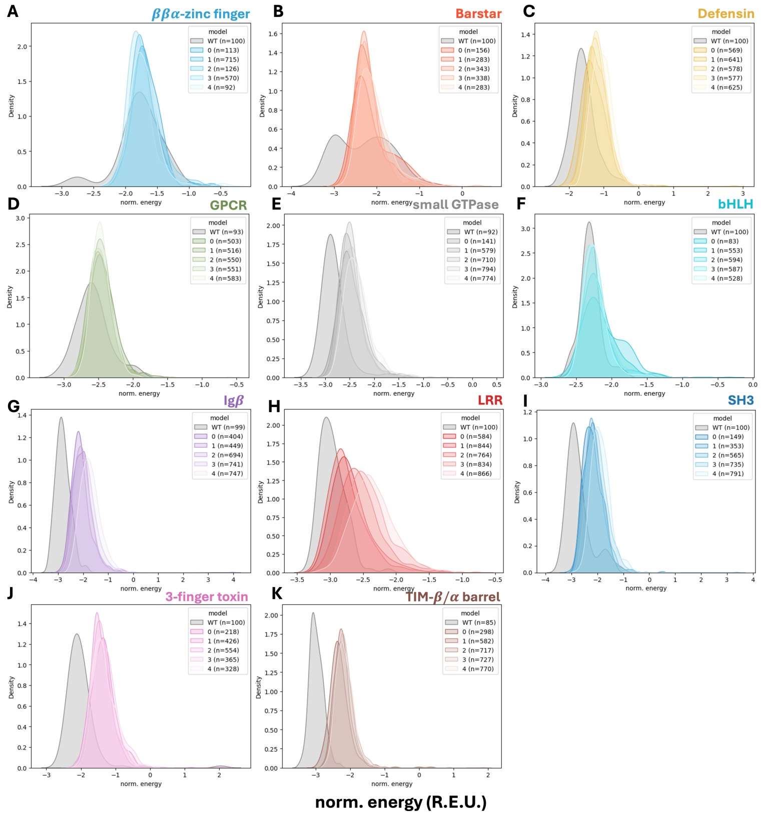


Figure 3.4: **Rosetta energies of foldtuned vs natural variants.** Histograms of length-normalized (REU / residue) Rosetta energy estimates for foldtuned (colored) and natural (gray) variants following standard backbone relaxation. Selected folds: (**A**) $\beta\beta\alpha$-zinc finger. (**B**) Barstar. (**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

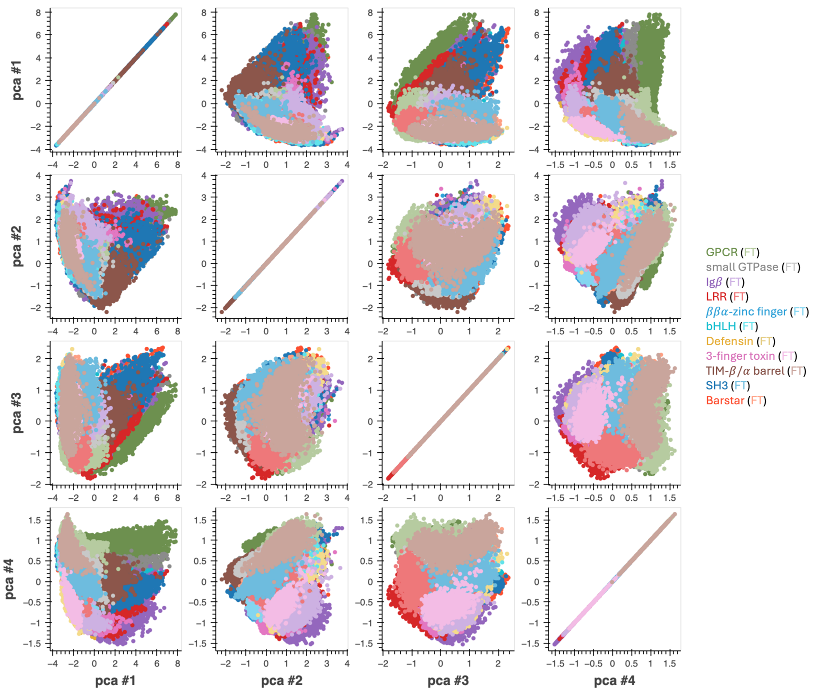


Figure S3.1: **Principal component analysis (PCA) of natural and foldtuned ESM2-650M embeddings.** Pairwise plots of top four principal components (fractional variance: 0.386, 0.103, 0.047, 0.039, respectively) of ESM2-650M embeddings of natural (SCOP-UniRef50) and foldtuning-generated proteins for 11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

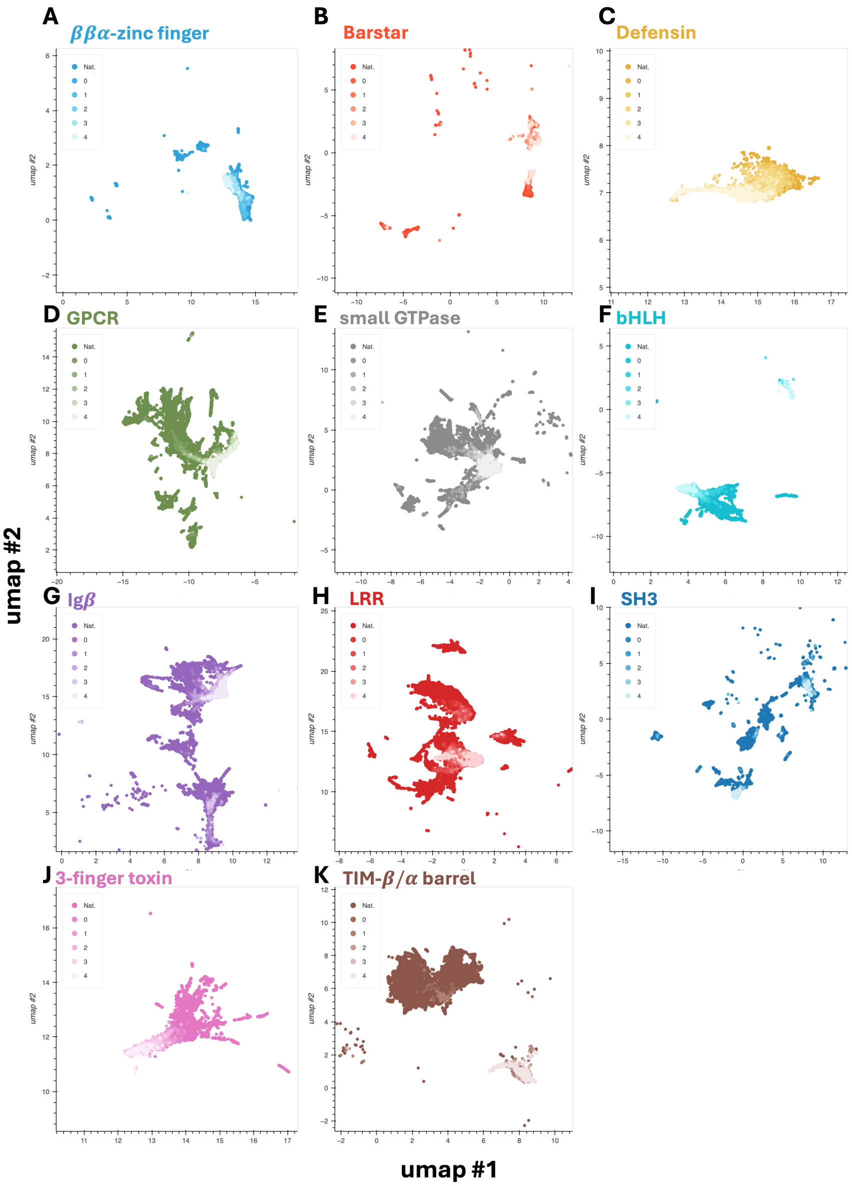


Figure S3.2: **ESM2-650M embeddings capture round-by-round drift of fold-tuned sequences from their natural parents.** 2D UMAP representation of ESM2-650M embeddings for eleven representative target fold classes, progressing from natural examples through up to five rounds of foldtuning. Selected folds: (**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.

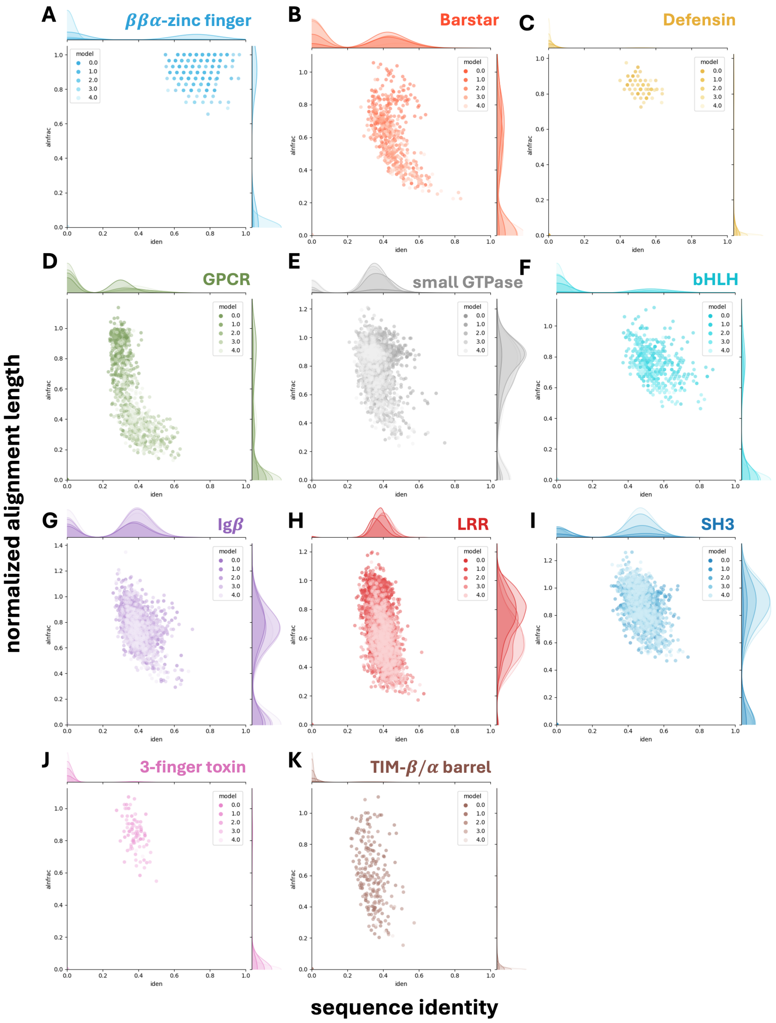


Figure S3.3: **Sequence similarity between foldtuned and natural variants.** Plots of normalized alignment length vs. sequence identity over the aligned region for the closest UniRef50 homolog to each foldtuned variant as identified by ultrasensitive search with MMSeqs2. Selected folds: (**A**) $\beta\beta\alpha$-zinc 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.

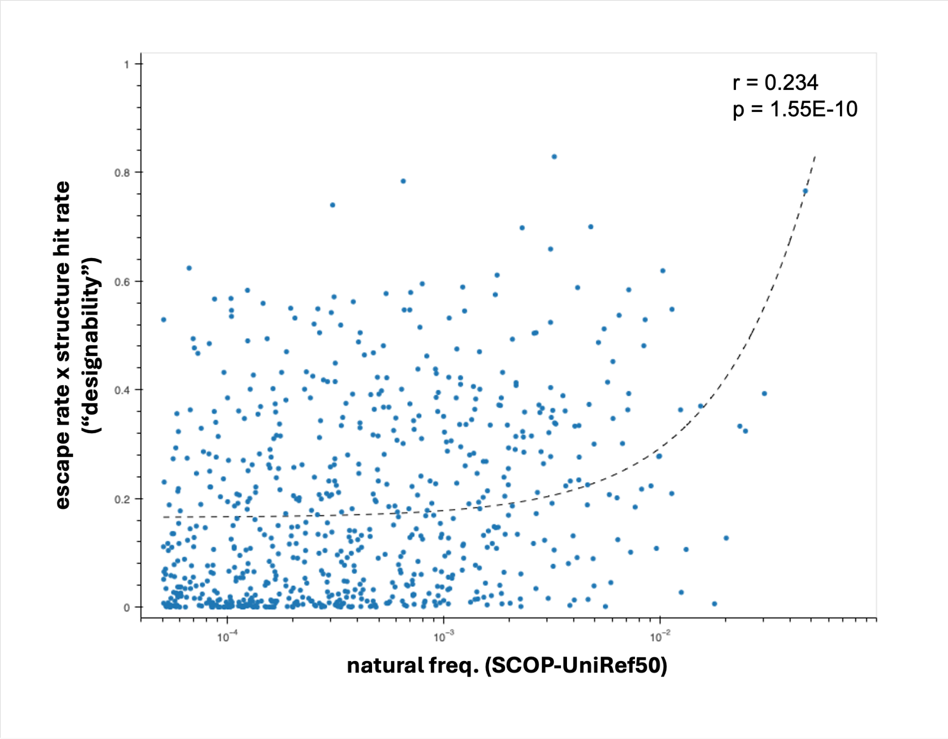


Figure S3.4: **Designability vs natural abundance for $n = 708$ SCOP fold targets.** Designability proxy (structural hit rate × sequence escape rate) across $n = 708$ SCOP fold targets is weakly explained by natural abundance in the custom SCOP-UniRef50 database: linear regression t-test for positive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

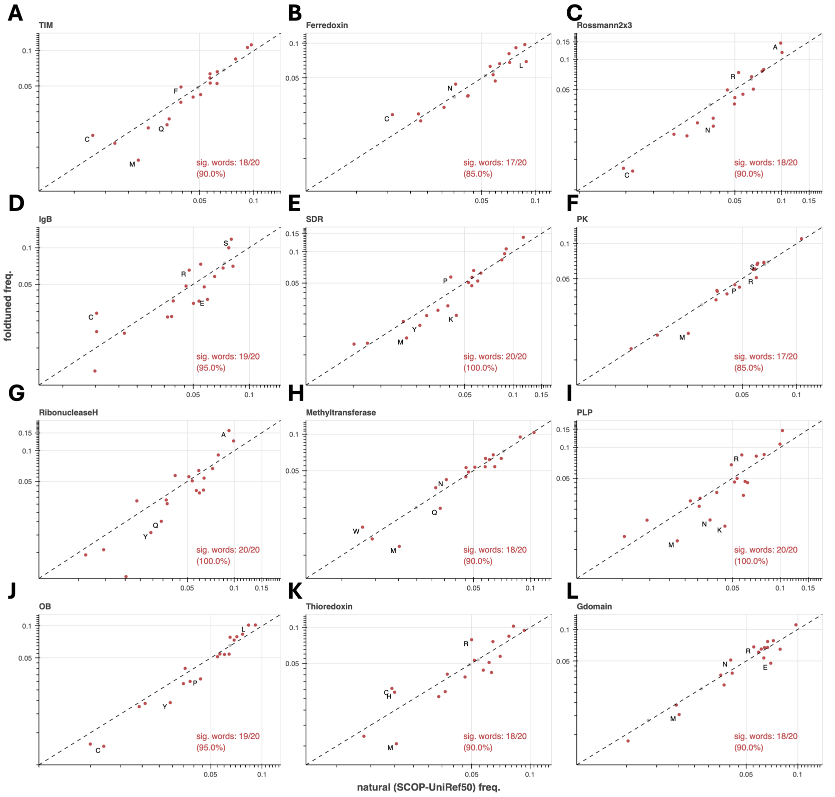


Figure S3.5: **Usage patterns of the 20 canonical amino-acids in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of AAs with a statistically significant usage shift (colored red, vs n = 1000 bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ under 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.

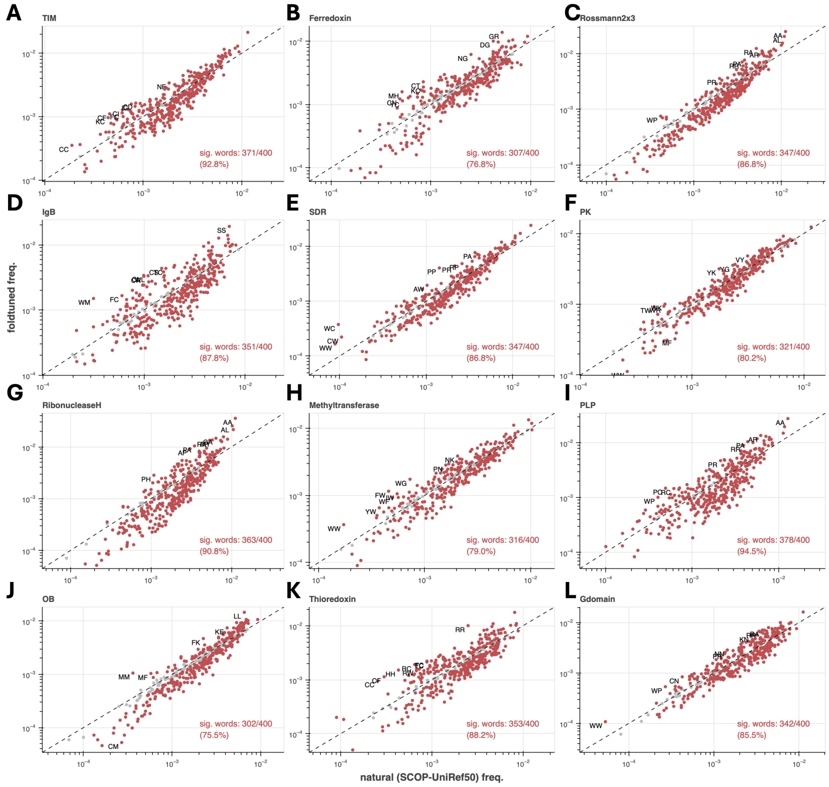


Figure S3.6: **Usage patterns of amino-acid subsequences of length 2 ("2grams", "bigrams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 2grams with a statistically significant usage shift (colored red, vs n = 1000 bootstrapped SCOP-UniRef50 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.

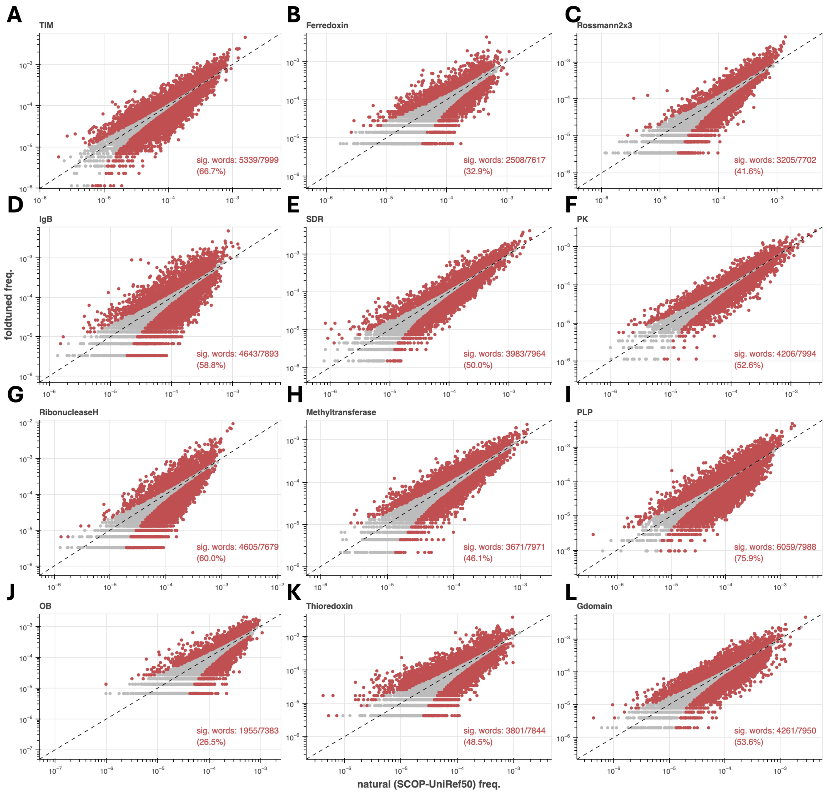


Figure S3.7: **Usage patterns of amino-acid subsequences of length 3 (“3grams”, "trigrams”) in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 3grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 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.

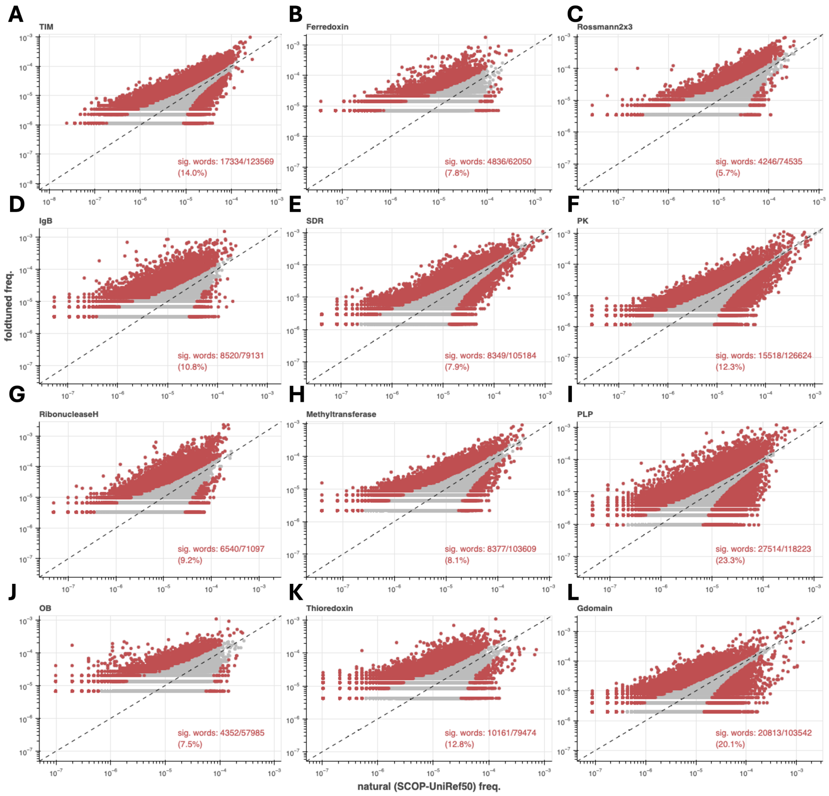


Figure S3.8: **Usage patterns of amino-acid subsequences of length 4 ("4grams") in foldtuned sequences vs. natural sequences for selected folds.** Sig. words denotes the count/fraction of 4grams with a statistically significant usage shift (colored red, vs $n = 1000$ bootstrapped SCOP-UniRef50 replicates, $p < 0.05$ 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.

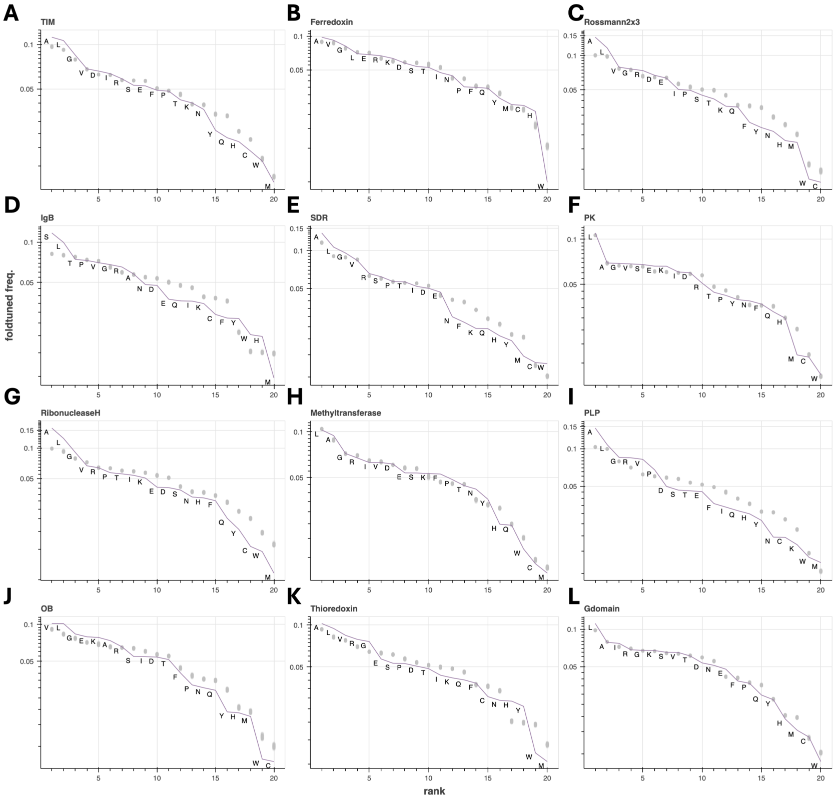


Figure S3.9: **Rank-ordered usage of individual amino-acids for foldtuned sequences (purple, labeled) and natural sequences (**$n = 1000$ **SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**) Igβ-like.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

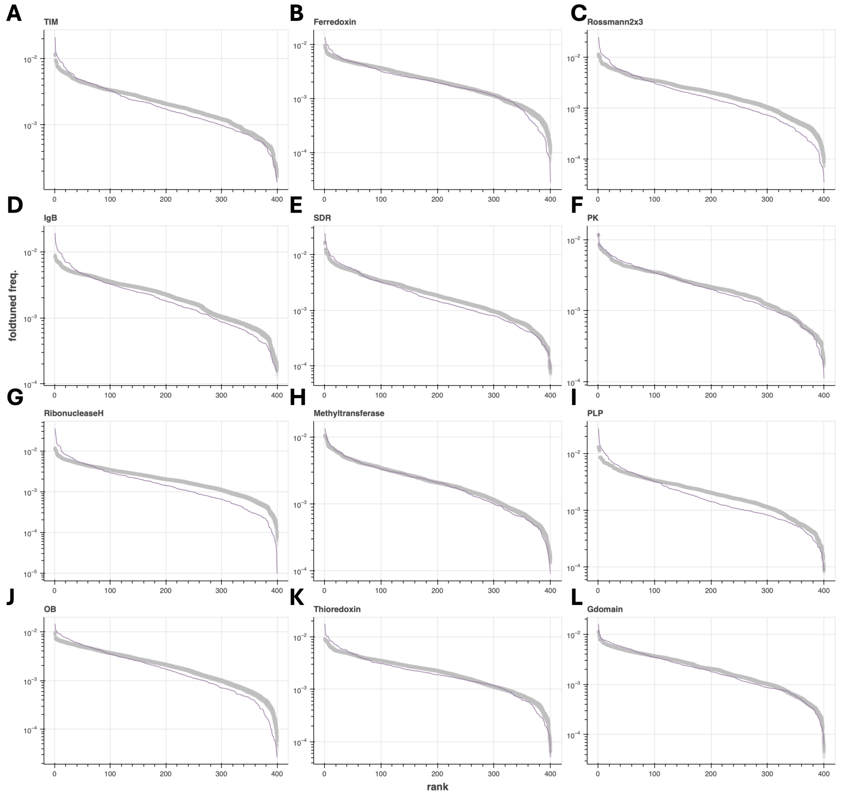


Figure S3.10: **Rank-ordered usage of subsequences of length 2 ("2grams", "bigrams") for foldtuned sequences (purple, labeled) and natural sequences (*n* = 1000 SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid.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

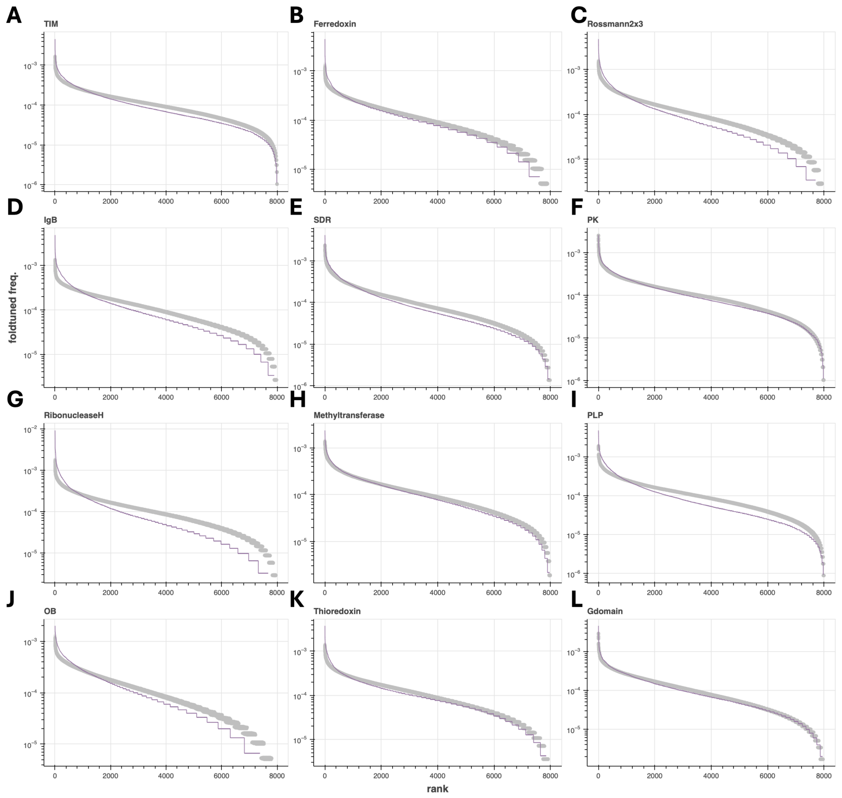


Figure S3.11: **Rank-ordered usage of subsequences of length 3 (“3grams”, “trigrams”) for foldtuned sequences (purple, labeled) and natural sequences (*n* = 1000 SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid.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

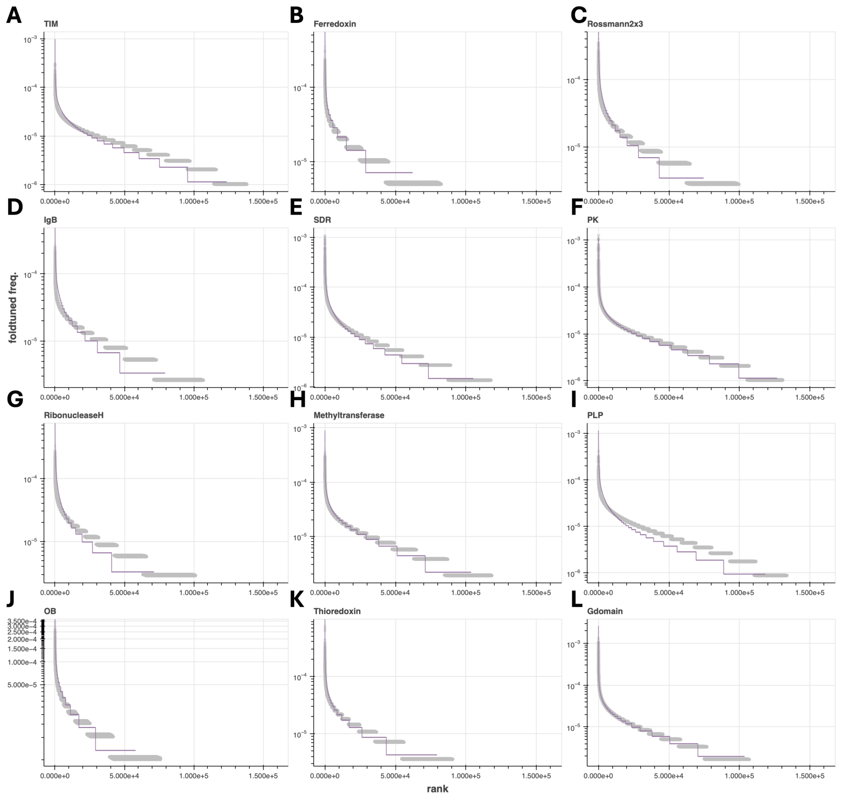


Figure S3.12: **Rank-ordered usage of subsequences of length 4 (“4grams”) for foldtuned sequences (purple, labeled) and natural sequences (*n* = 1000 SCOP-UniRef50 bootstrap samples; gray) for selected folds.** Selected folds: (**A**) TIM β/α barrel. (**B**) Ferredoxin. (**C**) Rossmann2x3oid. (**D**)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

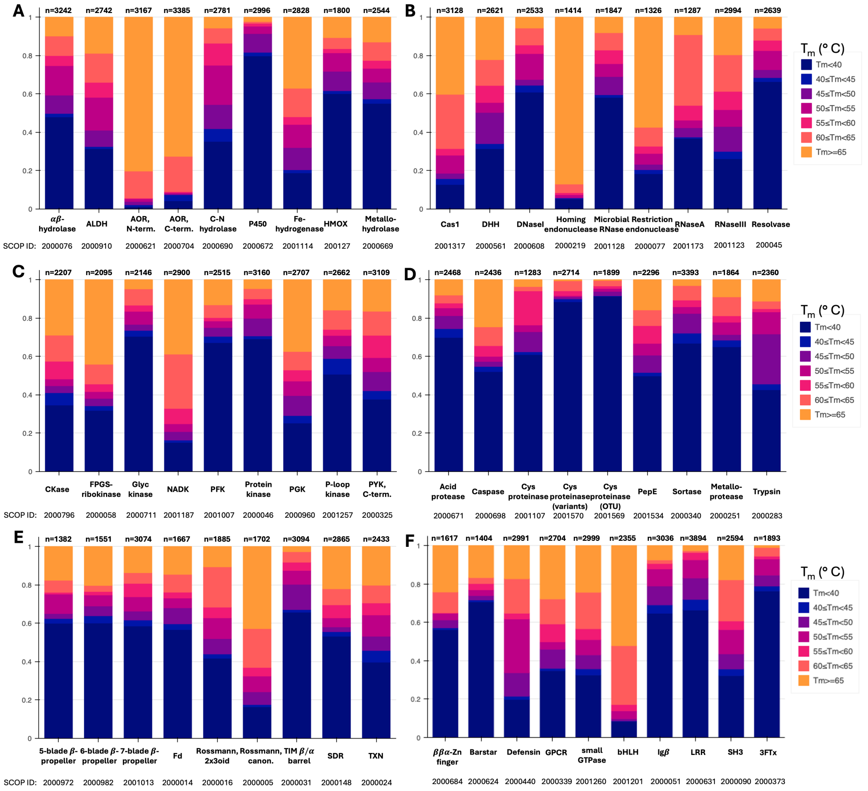
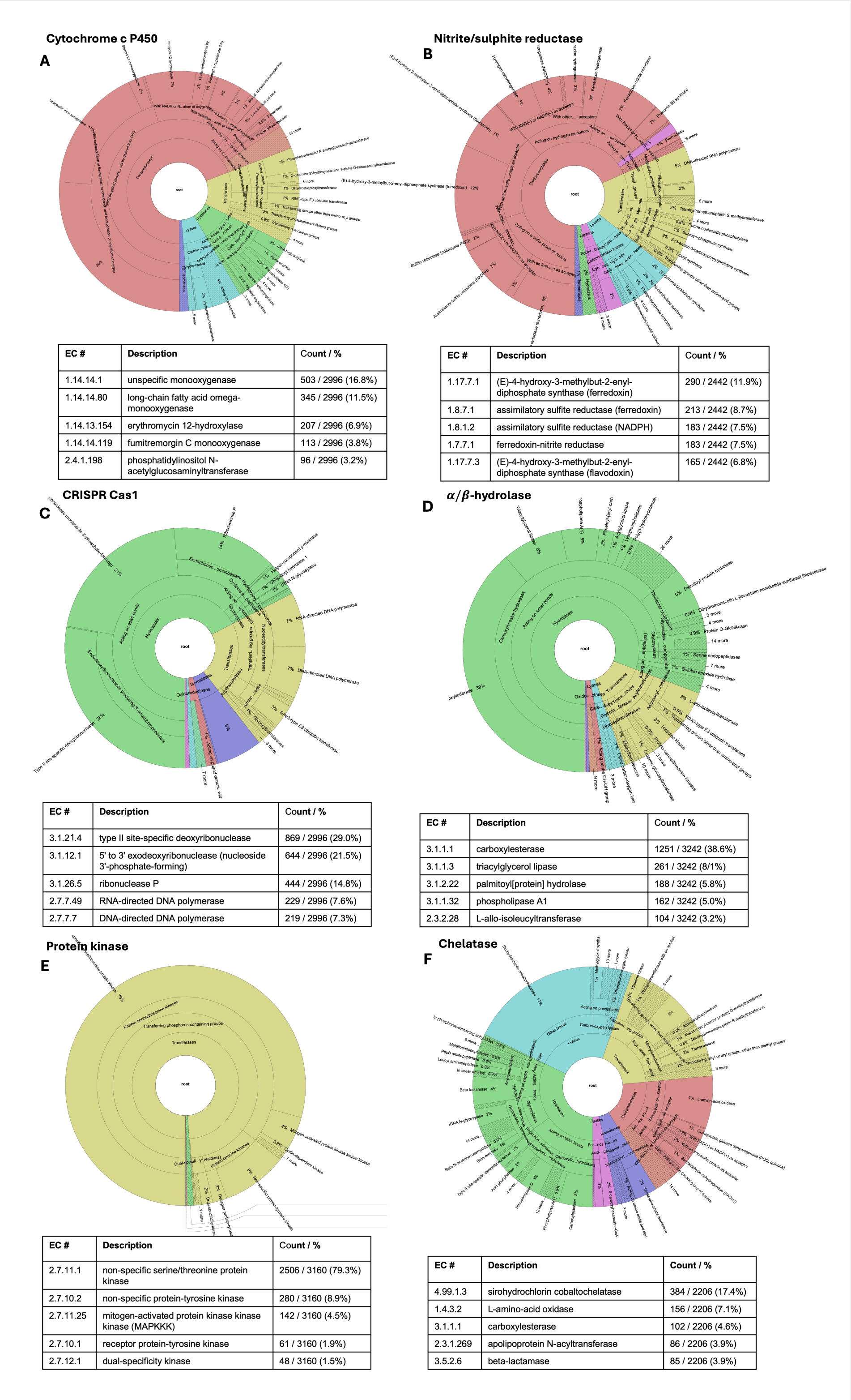


Figure S3.13: **Foldtuned proteins are predicted to exhibit varying degrees of thermostability.** Filtered, validated sequences generated from 55 foldtuned models of interest are expected to exhibit melting temperatures (T_m) ranging from $< 40°C$ to $> 65°C$, as predicted by TemStaPro (Pudžiuvelytė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



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

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

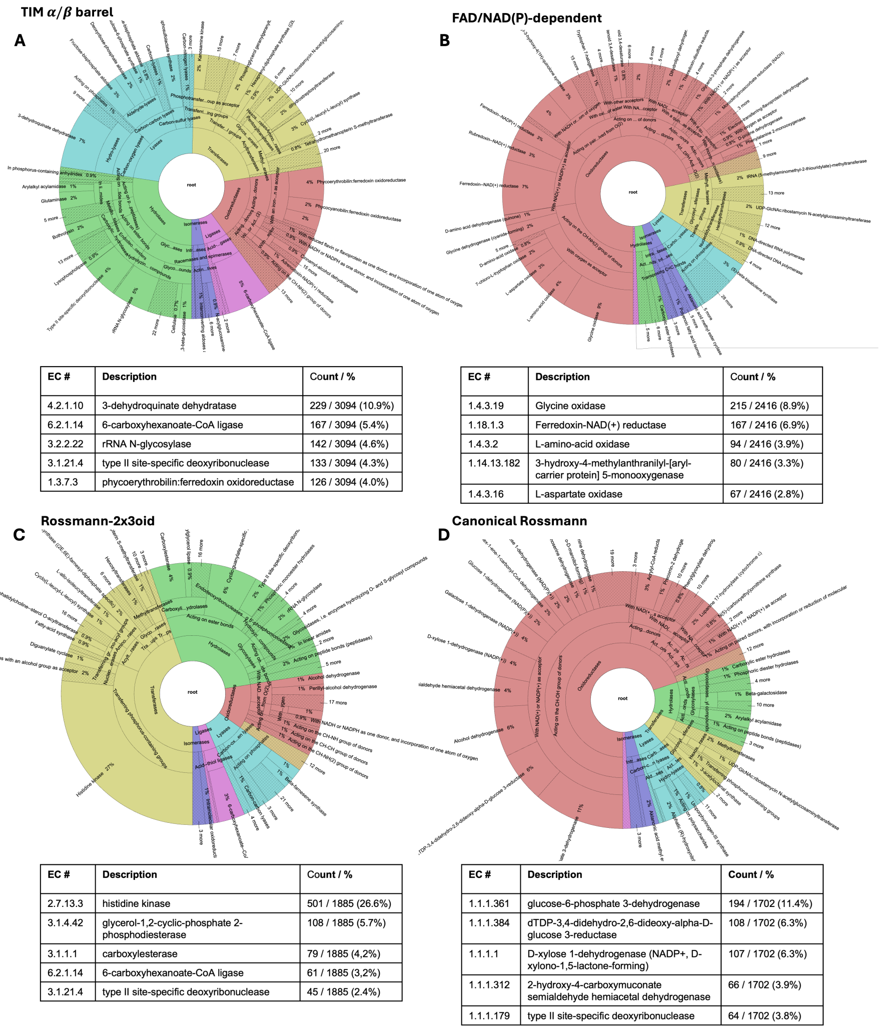
EC #	Description	Count / %
3.1.21.4	type II site-specific deoxyribonuclease	869 / 2996 (29.0%)
3.1.12.1	5' to 3' exodeoxyribonuclease (nucleoside 3'-phosphate-forming)	644 / 2996 (21.5%)
3.1.26.5	ribonuclease P	444 / 2996 (14.8%)
2.7.7.49	RNA-directed DNA polymerase	229 / 2996 (7.6%)
2.7.7.7	DNA-directed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

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

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

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

Figure 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

Chapter 4

FOLDTUNED PROTEINS ARE NOVEL AND FUNCTIONAL

4.1 Introduction

In the preceding chapter, we motivated and introduced "foldtuning" as a promising algorithm for generating far-from-nature and new-to-nature sequences that slot into the broad contours of natural fold families, leveraging structural plasticity in moderation as an extra, evolution-inspired source of structural and functional novelty. We also showed that, in addition to reflecting orthogonal-to-nature "language rules" for "writing" protein sequences, many foldtuned proteins pass a basic computational screen for physical reasonableness by exhibiting predicted-stable folded states.

In this chapter, we report on preliminary *experimental* validation for three foldtuned targets selected for amenability to high-throughput characterization, familiarity in the general field of protein science, and translational relevance for downstream synthetic biology and therapeutic applications. These three targets are as follows: (1) the SH3 domain, a small adaptor domain that mediates protein-protein interactions in receptor-initiated and cytoplasmic signal transduction pathways, often as part of tyrosine kinases (Kurochkina and Guha, 2013; Mayer, 2001); (2) the barstar fold, an antitoxin-like inhibitor of a secreted bacterial ribonuclease, the smallest and simplest of the known α/β folds, and additionally well-studied as a model system for concerted folding pathways and protein-protein interaction energetics (Schreiber and Fersht, 1995; Schreiber et al., 1994); and (3) insulin, the first peptide hormone discovered and characterized, the major regulator of anabolic metabolism in eukaryotes, and whose absence or dysregulation is the causative agent of diabetes (Mayer et al., 2007). Tailoring assays for expression, stability, and binding to the individualized circumstances of the aformentioned three fold targets, we demonstrate that foldtuned proteins are realizable and functional in certain *in vitro* and *in vivo* contexts. Augmenting these experiments with statistical and theoretical analyses of generated sequence architecture, structural features, and physicochemical properties we argue further that foldtuned models learn the minimal structural information required to maintain a core fold and to either (i) preserve existing function or (ii) broaden to novel ones depending on the selective pressure applied.

4.2 Results & Discussion

Foldtuned SH3 domains express stably

Emboldened by the ability of foldtuning to readily propose plausible far-from-natural protein sequences, we sought to validate selected examples experimentally for expression, stability, and function with minimal target-specific platform optimazation. From a roster of small folds ($\leq$ 84aa) for which coding DNA oligo pools could be easily synthesized, we focused first on the SH3-like barrel (SCOP ID: 2000090). The SH3 domain is a notable protein-protein interaction component and regulator of signal transduction, particularly in tyrosine kinase pathways. Engineered SH3 domains have historically been desirable in synthetic biology for roles in designed artificial protein recognition and signaling cascades, but attempts to develop an SH3 "toolkit" have been stymied by difficulties with *de novo* β-barrel design and off-target crosstalk with natural SH3s (Kim et al., 2023). SH3 structural homologs are strewn across functionally diverse superfamilies, including the aforementioned adaptor domains that commonly bind polyproline motifs in protein ligands, chromodomains that recognize histone methylated lysine marks, and large-subunit ribosomal proteins that scaffold rRNA.

Applying the standard evo+four foldtuning procedure to ProtGPT2 with SH3s as the target produced 2593 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 0.519 and 0.310 respectively. In contrast to, e.g. deep-mutational scanning libraries, proteins in foldtuned variant libraries — including for SH3s — boast high sequence diversity, featuring low pairwise sequence similarities and unique proteolytic digestion signatures (Fig. S4.1A-C). This enables direct high-throughput characterization of protein expression and select biophysical properties by mass-spectrometry-based proteomics without the additional complexity and cost of typical yeast-, mRNA-, or cDNA- display methods (Fig. 4.1A) (Rocklin et al., 2017; Tsuboyama et al., 2023). For our SH3 foldtuned library, 1347/2593 (51.9%) variants express at detectable levels in a reconstituted transcription-translation system as measured by untargeted mass-spectrometric profiling (Fig. 4.1B-C). Using length-normalized signal as a proxy for absolute abundance of expressed proteins, we observe signal intensity spanning $\sim$ 6 orders of magnitude, suggesting substantial variance in the intrinsic expressability of foldtuned SH3s and foldtuned designs more broadly; it must be emphasized, however, that these measurements cannot on their own account for confounding factors such as the imbalances in the makeup of the amplified oligo pool encoding the SH3 library. Regardless of this nuance, we see no evidence that expression level correlates with sequence similarity to natural

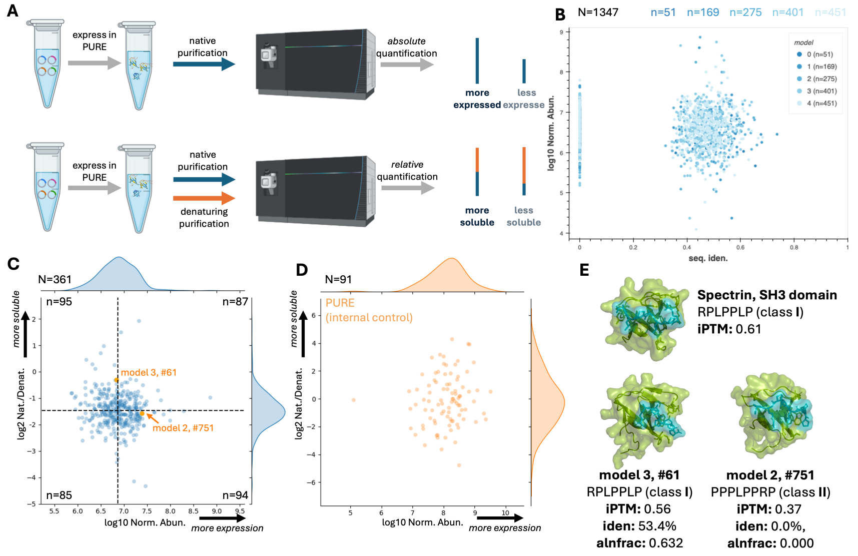


Figure 4.1: **Foldtuning-generated SH3s are expressable and stable.** (**A**). Schematic of mass-spectrometry-based proteomics assays for variant library expression and folding stability. (**B**) SH3 expression assay signal intensity normalized by expected tryptic peptide count vs. sequence identity to most-similar natural hit in UniRef50 for variants generated from models undergoing 0-4 rounds of foldtuning. (total $N = 1347$) (**C**) SH3 folding stability vs. expression assay results for $N = 361$variants detected in both contexts. Folding stability (y-axis) is measured by relative abundance ratio between natural and denaturing purification fractions. Normalized expression (x-axis) is measured as as in (**B**). (**D**) Folding stability vs. expression for PURExpress transcription-translation protein components ($N = 91$) as an internal control. (**E**) AlphaFold3 predicted structures and iPTM scores for selected SH3 variants (green) bound to a classI/II proline-rich peptide (teal), compared to the wildtype *G. gallus* spectrin SH3 domain.

SH3s or that it shifts based on the number of foldtuning cycles performed (Fig. 4.1B).

To rule out cases where high cell-free expression intensity might mask solubility and/or aggregation issues from poor folding stability we compared foldtuned protein recovery under native and denaturing purification conditions via multiplexed proteomics; variants without folding pathologies (e.g. exposed hydrophobic residues, buried and/or electrostatically clashing charged residues) are expected to show equivalent or greater signal in the native fraction relative to the denatured one (Fig. 4.1A). Analysis of the native/denatured signal fold-change for an internal control of $N = 91$ *E. coli* proteins originating from the reconstituted transcription-translation system demonstrates that this stability/solubility proxy has a dynamic range spanning up to ~ 10 orders of magnitude under the instrument conditions used (Fig. 4.1D).[1] Turning back to the foldtuned SH3s, 361 variants are detected confidently in both the absolute and multiplexed expression assay; absolute expression signal varies over ~ 4 orders of magnitude, while the native/denatured signal fold-change varies over ~ 6 orders of magnitude (Fig. 4.1C). Of particular interest is a subpopulation of 87 foldtuned SH3s that are both highly abundant in the initial expression assay and displaced away from the denatured fraction in the solubility/aggreggation assay, suggesting expressability, relative folding stability, and low aggregation propensity (Fig. 4.1C).

With an eye towards designing and optimizing SH3 parts for synthetic signal transduction, we computationally tested the hypothesis that foldtuned SH3 variants with high expressability and relative folding stability might recognize the proline-rich peptide motifs found in the full-length protein binding partners of natural SH3 domains (Mayer, 2001). *In silico* screening with AlphaFold3 predicts that, indeed, certain physically-plausible foldtuned SH3 variants can bind either class I or class II proline-rich ligands in a hydrophobic aromatic-sidechain-rich cleft analogous to the wild-type interface as exemplified by the *G. gallus* spectrin SH3 domain (Fig. 4.1E). Two exemplary foldtuned putative SH3s that emerge in the AlphaFold3-based screen are model 3 #61 (3_61) and model 2 # 751 (2_751). Variant 3_61 is a distant homolog of the guanine nucleotide exchange factor Vav (involved in cytoskeletal remodeling during lymphocyte development and activation) and is predicted to recognize the canonical class I motif RPLPPLP. Variant 2_751 has no detectable sequence homology to any known protein, yet is predicted to recognize the canonical class II motif PPPLPPRP.

To clarify how foldtuned models might be preserving critical structural and functional features in SH3s, including ones responsible for stability or binding of polyproline motifs, we turned to statistical coupling analysis (SCA). Originally developed to identify statistically interacting amino-acids from evolutionary-related sequence data, SCA has historically been applied to natural protein families to infer and extract physically connected "sectors" posited to comprise the minimal sequence information required to specify a fold and/or function (Halabi et al., 2009; Lock-

[1] The exact number and amino-acid sequences of protein components in the particular recombinant transcription-translation system used in this study, PURExpress from New England Biosystems, is proprietary. As a substitute approach, internal control proteins were mapped to an *E. coli* reference proteome.

less and Ranganathan, 1999; Socolich et al., 2005; Süel et al., 2003). Here, we applied SCA separately to natural SH3 domains and to the 2593 foldtuned putative SH3s and extracted sectors from each.[2] For both natural and synthetic SH3s, SCA finds a single small sector, covering eight and five residues in the natural and synthetic cases respectively (Fig. S4.2A-B). Natural and synthetic sectors are composed of non-overlapping sets of core residues; only a single sector position interacts directly with the bound proline-rich motif and it is shared between the natural and synthetic sectors. This suggests that, in line with the promiscuity and diversity of SH3-peptide binding, foldtuning may be preserving a bare-minimum sequence rule for binding few-among-many polyproline-like targets, while trying out a completely different solution for stably packing the SH3 β-barrel core. Ultimately, evaluating this interpretation will necessitate experimental validation of new foldtuning-enabled synthetic "links" in the SH3 connectome, potentially via a high-throughput/high-resolution SH3-peptide all-against-all cross-linking mass spectrometry approach.

Foldtuned barstars rescue bacteria from barnase toxicity

For a target with a more direct experimental readout of not just stability, but also function, we consider the barstar-like fold (SCOP ID: 2000624). With a single three-stranded parallel β-sheet packed against three α-helices, all connected by short loops, the barstar-like fold is an exceedingly simple α/β unit, familiar from foundational studies of protein folding stability (Schreiber and Fersht, 1995; Schreiber et al., 1994). The coding gene for its namesake protein, barstar, was originally identified in *Bacillus amyloliquefaciens* with orthologs distributed across gram-positive bacteria and structural homologs in the DNA double-strand break repair protein Mre11 and ribosomal protein L32e. Leveraging the expanse of the AlphaFoldDB, the custom SCOP-UniRef50 sequence-structure database also detects distant structural homolog barstar-like regions in proteins with putative ATPase and palmitoyltransferase activity, expanding the landscape of sequence motifs to harvest from. Barstar's native function in *B. amyloliquefaciens* is to inhibit, through a high-affinity active-site-occluding non-covalent interaction, the potent broad-spectrum bacterial ribonuclease barnase before its secretion into the surrounding environment.[3] Together, barnase (toxin) and barstar (antitoxin) comprise a toxin-antitoxin

[2] We believe this analysis and an analogous one on barstar later in the chapter to be the first examples of applying SCA to synthetic data to retrieve "pseudo"-evolutionary correlations.

[3] Barnase is so general a ribonuclease that its name is simply a portmanteau of "**ba**cterial" and "**r**ibo**n**ucle**ase**."

system, presenting an opportunity for functional screening of a foldtuned variant library for toxic gene rescue.

We apply the standard evo+four foldtuning approach to barstar, yielding 1403 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 0.281 and 0.560 respectively. Variants were co-expressed with barnase from *B. amyloliquefaciens* under a single *tac* promoter, under strong induction conditions, in a high *lacI E. coli* strain in order to mitigate confounding adaptations to barnase expression (Fig. 4.2A). In the absence of proper barstar expression and function, barnase expression is toxic to *E. coli* (Hartley, 2001). Functional foldtuned variants are expected to rescue host cells from the lethal effects of barnase expression. Comparing long-read sequencing counts of variant-coding amplicons, we found that 11 foldtuned barstar variants were significantly enriched ($p < 0.05$; Binyami-Hochberg correction for correlated tests) relative to uninduced (non-barnase-expressing) control under strong induction of barnase-barstar-variant co-expression, suggesting that the enriched variants are sufficiently functional mimics of barstar so as to mitigate the toxicity of barnase (Fig. 4.2B). Additionally, enrichment does not correlate with sequence identity relative to wild-type barstars or any natural protein. To this point, 7/11 of survival-enriched foldtuned barstars do not exhibit any detectable homology to natural sequences at the domain or sub-domain level (Fig. 4.2C).

For mechanistic insight and hypothesis refinement, we obtained AlphaFold3 predicted structures of the survival-enriched variants in complex with barnase. For four foldtuned variants — model 1 #633 (1_633), model 3 #647 (3_647), and model 4 #s 141 (4_141) and 219 (4_219) — these predicted complex structures indicate that barstar mimics are expected to bind barnase analogously to wild-type barstar, inserting an α-helix and adjoining loops into the binding pocket, obstructing the RNA hydrolsis active site (Fig. 4.2D). Detailed examination of predicted binding interfaces reveals that foldtuned barstars are expected to form hydrogen-bonds and salt-bridges with barnase, without steric or electrostatic clashes. Comparison with a published experimental structure of the endogeneous *B. amyloliquefaciens* barnase-barstar complex (pdb: 1BRS) suggests that fewer such contacts are expected with variants than with wild-type barstar, potentially indicating weaker binding and consequently reduced inhibition of barnase (Fig. 4.2D). It bears noting that this difference may stem at least in part from non-ideal bond geometries that persist due to AlphaFold3's lack of a side-chain or backbone relaxation step; molecular dynamics simulations could prove valuable for discriminating between binding strengths in a more informative manner, as it has in a related design context for IL7RA minibinders (Lourenco et al., 2025).

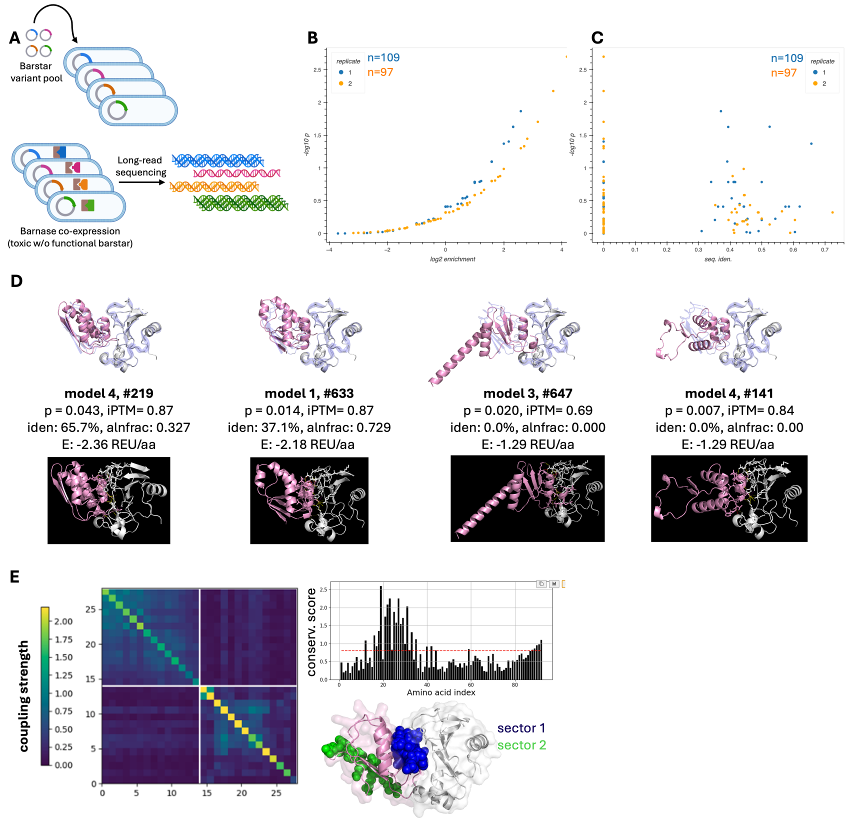


Figure 4.2: **Foldtuning-generated barstar variants are expressable, stable, and functional.** (**A**) Schematic of barnase-inhibition survival assay for barstar variant library stability and function. (**B**) Survival assay p-value rank plot for barstar variants. For a given variant, enrichment is calculated as the ratio of amplicon sequencing reads with and without induction of co-expression of the lethal binding-partner barnase (**C**). Survival assay p-values from (**F**) vs. barstar variant sequence identity to most-similar natural hit in UniRef50. (**D**) Top row: AlphaFold3 predicted structures, iPTM scores, and Rosetta energy predictions for selected barstar variants (pink) in complex with barnase (white). An experimental crystal structure of the wildtype barnase-barstar complex from *B. aquaforiensis* (pdb: 1BRS) is overlaid in blue. Bottom row: Predicted complex structures with putative hydrogen bonds and electrostatic interactions indicated. (**E**) Results of statistical coupling analysis (SCA) on n = 1493 foldtuned barstar sequences. Left: Second-order coupling matrix, blocked into two orthogonally co-evolving sectors. Top right: First-order conservation scores. Bottom right: Visualization of sector positions mapped onto a representative barstar-barnase complex structure (pdb:2ZA4).

Looking beyond the interface, AlphaFold3 predictions also imply that some variants may explore alternate conformations and binding modes in the neighborhood of expected barstar structure and function, exemplified by 4_141, a zero-homology variant that is predicted to have low β-content and rotate 90° relative to the wild-type while maintaining α-helical and loop elements in the barnase binding pocket (Fig. 4.2D). Of note, the two variants predicted to adopt more dramatically different conformations — 4_141 and 3_647 — are also assigned Rosetta energy scores on the higher end of plausible; in the absence of experimental structure determination, these structure-function hypotheses should be taken with a degree of caution.

Given the detection of foldtuned barstar mimics with antitoxin-like function, and circumstantial indications that at least some among these mimics may utilize similar structural solutions to wild-type barstar, a natural question to ask is that of what sequence and/or structure "rules" the foldtuned models themselves have learned? Multiple sequence alignment of wild-type barstar along with the eleven survival-enriched foldtuned variants reveals that in the contiguous nineteen-residue region (columns 38-56) spanning the barnase-binding interface, toxicity-rescuing variants preserve 6-11 (32-58%) of wild-type amino-acid identities (Fig. S4.3). Clearly, foldtuned models are not simply memorizing the semantics of barnase-binding and scaffolding them into redesigned flanks.

For a deeper view of how foldtuned models might be preserving the structural-functional "grammar" of barstar, we return to SCA, this time treating the 1403 foldtuned barstar variants as a synthetic protein family.[4] SCA proposes two sectors; one, at the C-terminus, is most likely an artifact attributable to ESM batch-inference token-padding with residual alanines; the other maps onto the barnase-binding interface (Fig. 4.2E). This suggests that foldtuning has "solved" the barnase-binding problem by decoupling the critical inserted α-helix motif from the rest of the protein, preserving only its most salient sequence features, and inventing wholly new ways to fill in the remainder of the barstar fold. In other words, foldtuning has distilled the structural and functional nature of barstar into a single essential grammar rule.

Foldtuned and PLM-sampled insulins are INSR binders and agonists

Lastly, we steered foldtuning to design mimics of insulin, a high-value translational target well outside of our initial set of 727 SCOP folds, posing several new

[4] Unlike in the SH3 case, we do not conduct the comparative analysis on natural barstars, as the wild-type sequences are so highly conserved across gram-positive bacteria as to swamp out the second-order interaction signatures that SCA relies on.

challenges for the foldtuning algorithm and workflow to overcome. Insulin hardly requires introduction as a protein of interest. It is the preeminent peptide hormone in all eukaryotes; insulin signaling coordinates anabolic metabolism across cells, tissues, and organs (Mayer et al., 2007). Its absence (due to self-reactive destruction of most insulin-producing pancreatic β-cells) is the salient aspect of type I diabetes; its dysregulation (insulin resistance) underlies type II diabetes. Insulin presents multiple challenges to the thematic underpinnings and practical application of foldtuning. To the former aspect, the active form of insulin is deeply conserved across eukaryotes at the sequence level, and shares a structural neighborhood with related peptide hormones including insulin-like-growth-factor-1 (IGF-1), relaxins, and several insulin-like peptides (ILPs) of unclear function; insulin and IGF-1 preferentially bind to and are agonists of their cognate receptor tyrosine kinases, with weak cross-reactivity; relaxins and ILPs cross-react with several GPCRs (Claeys et al., 2002). This suggests that foldtuning, with its emphasis on innovation about a template structure in moderation, may sample a range of specific and promiscuous binding phenotypes as opposed to binders specific to the insulin receptor INSR.

As far as practical implementation obstacles, insulin is a tricky target for foldtuning thanks to the post-translational internal cleavage events required to transform inactive, largely disordered proinsulin into structured, active insulin through excision of the C-peptide (which makes up 31 of the 86 residues in the coding region of the INS gene) and formation of three disulfide bonds (two interchain between the A- and B-peptides; one intrachain within the A-peptide). To circumvent this issue and to align with standard expression and characterization processes in industry, we foldtunded ProtGPT2 to generate single-chain insulin variants that are fusions of the A- and B-peptides. Natural training sequences (n = 335, reduced to n = 193 after deduplication clustering) and reference structure fragments were taken from InterPro entry IPR004825, which ostensibly includes insulin and excludes IGF-1, relaxin, and ILPs, though sequence homology considerations cast some doubt on the robustness of this filtering. IPR004825 sequences were multiply aligned to *H. sapiens* insulin to identify putative C-peptide regions to be removed before clustering and downsampling, leaving single-chain A/B fusion training data. Standard evo+four foldtuning rounds yielded 2889 putative insulin variants with structure hit and sequence escape rates of 0.578 and 7×10^{-4} respectively. The atypically low sequence escape hit for foldtuned insulin models (only 2/2889 variants lacking detectable homology to natural proteins), as well as a median 80.0% sequence similarity to the closest natural hit, likely stems from the aforementioned high degree

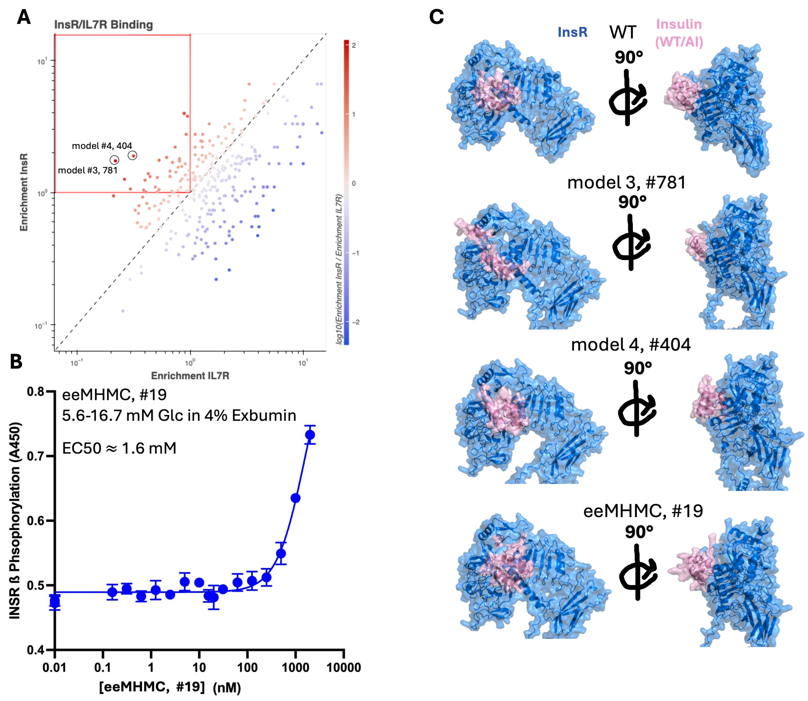


Figure 4.3: **PLM-designed insulin variants bind the endogenous insulin receptor.** (**A**) Relative enrichment plot of foldtuned insulin variant binding to the endogenous INSR receptor (on-target) vs the endogenous IL7R receptor alpha-chain (off-target) using the Protein CREATE platform. (**B**) Sandwich ELISA (A450 spectrophotometric readout) of INSR phosphorylation in response to explore-exploit Metropolis-Hastings Monte Carlo (eeMHMC) designed insulin variant #19 (estimated EC50=1.6 mM). (**C**) AlphaFold3 predicted structures of WT *H. sapiens* insulin and experimentally plausible designed INSR binders (pink) in complex with the native INSR receptor ectodomain (blue).

of sequence conservation among natural insulins and from a tradeoff in choosing a small InterPro family from which to initiate foldtuning in hopes of prioritizing INSR-specific binders and agonists.

We used the Protein CREATE platform to screen all foldtuned putative insulins for INSR-specific binding as described in Lourenco et al. (2025). In brief, variants are displayed on T7 bacteriophage and screened against multiple receptor candidates ligated to magnetic beads, with a sequencing-based readout of amplicon counts before- and after- receptor-bead pulldown, resulting in a vector of enrichment scores for each receptor screened. Here we screen against two receptors, taking enrichment after INSR pulldown as a measure of on-target binding, and enrichment after IL7RA (a type I cytokine receptor) pulldown as a measure of generic off-target binding. For 307 foldtuned variants with sufficient reads to calculate enrichment in both contexts, 41 (13.4% of detected; 1.4% of entire foldtuned pool) are enriched (enrichment > 1) for INSR-binding and de-enriched (enrichment < 1) for IL7RA-binding (Fig. 4.3A). This is fewer than the 61 variants (19.9% of detected; 2.1% of entire pool) with the inverse phenotype of IL7RA-binding enrichment and INSR-binding de-enrichment. The lion's share of variants — 159 (51.9% of detected; 5.5% of entire pool) exhibit a doubly-enriched phenotype in this assay — underscoring, altogether, the inherent difficulty of designing *specific* binders, as well as an opportunity in applying foldtuning as a zero-shot generator of binding *phenotype diversity* suitable for multiple downstream optimizations.

To validate INSR-binding activity, we attempted chemical synthesis of the two foldtuned variants with the highest relative enrichment scores (INSR enrichment / IL7RA enrichment) out of the 41 variants that display the INSR-specific phenotype in Protein CREATE screening — these are model 3, #781 (3_781) and model 4, # 404 (4_404). However, neither variant is observed to refold solubly following synthesis and denaturation, suggesting a failure of proper disulfide bond formation, pointing potentially to a lack of a uniquely stable ground state among multiple cyclization isomers; troubleshooting is ongoing. AlphaFold3 predicts that both priority variants should indeed bind to the INSR ectodomain with ligand-receptor contacts reminiscent of but not identical to those formed by wild-type insulin, lending support to the emerging paradigm from our investigations of SH3 and barstar that foldtuning retains only those sequence rules minimally necessary for marginal binding while injecting novelty that percolates to perturbed contacts, pockets, and downstream phenotypes (Fig. 4.3C).

The challenges presented for foldtuning by insulin's sequence and structure features led us to develop and evaluate a second PLM-based generation strategy in parallel. We call this strategy "**e**xplore-**e**xploit **M**etropolis-**Ha**stings **M**onte **C**arlo", or eeMHMC. We are far from the first to take an MHMC or Markov Chain Monte Carlo (MCMC) approach more generally to the problem of sampling over a sequence landscape from an encoder-only PLM. What is novel about our approach is the energy function that determines the acceptance probability of proposed moves. Where others have written energy functions that consider only the likelihood of a given sequence as inferred by a PLM, sometimes regularized to explicitly favor memorization of natural sequence motifs, we propose a two-term function that balances leveraging what the PLM has internalized about sequence plausibility (the "exploit" term) with an incentive to make large semantic changes (the "explore" term) (Hie et al., 2022; Verkuil et al., 2022).[5] In this way, eeMHMC stands on common ground with the grammar-respecting/semantics-altering behavior of foldtuning. We generate 100 candidate insulin mimics with ESM2-650M-based eeMHMC, seeded by column-wise independent sampling from a deep insulin MSA. Among these 100 variants, sequence identity to human insulin drops as low as 31.4% while still respecting the canonical insulin fold (including disulfide staples) and showing reasonable surface hydrophobicity and electrostatics, according to computational analysis on predicted structures (Fig. S4.4). Outside collaborators attempted to express 20 of the eeMHMC variants under standard industrial conditions; of these, one variant, eeMHMC_19 (56.8% identical to human insulin), was able to be expressed, purified, and refolded. Furthermore, eeMHMC_19 shows agonist activity for INSR according to sandwich ELISA readout of receptor phosphorylation (Fig. 4.3B). Although this activity is weak, with an inferred EC50 of 1.6 mM, ∼ 1000x higher than for wild-type insulin and ∼ 10-50x higher than for wild-type IGF-1, it still represents remarkable progress towards a functional insulin mimetic with substantially reduced resemblance to wild-type.

4.3 Conclusion

Picking up from *in silico* revelation of the potential scope and breath of foldtuning as a synthetic biology design engine, we completed a fastidious series of experiments, supported by mechanistic insight from computation, underscoring that foldtuned proteins are novel, realizable, and functional. Taking the SH3 domain as an initial representative and experimentally tractable fold, we showed that foldtuning clears initial but essential criteria of generating expressable, stable, non-aggregation-prone proteins. Raising the bar to a toxin-antitoxin system where structure and function are closely linked, and where selective pressure disfavors natural exploration, we found that foldtuning maintains function as an apparent byproduct of preserving structural constraints, recapitulating the fold-and-function "grammar" of the barnase-binding interface of barstar with novel semantics. Coming to insulin, an ambitious target with a host of obstacles that could have been erected deliberately to thwart foldtuning — sequence deeply and anciently conserved, structure shared across a bucket of related peptide hormones, significant processing required *in vivo* to go from an

[5] Full mathematical details are reported in Section 4.4.

inactive single-chain propeptide to an active multi-chain species — we still obtained INSR-receptor-specific binder candidates. Moreover, we introduced a second PLM-based generation mechanism, explore-exploit MHMC, which, like foldtuning, perturbs sequence semantics in a stepwise fashion while implicitly retaining structural grammar, and applied it to design and validate a novel INSR agonist.

Beyond their utility as interesting and convenient case studies, the three targets examined in detail here proffer an assortment of pertinent takeaways for the future of foldtuning. The apparent allowance of foldtuning for recognizing variable sequence motifs on binding-partners implies that pools of foldtuned SH3s could contain enough diversity to mine, optimize, and organize sets of orthogonal nature-inspired parts with new-to-nature recognition logic into synthetic signaling cascades or larger fully synthetic connectomes. Such an approach would port over conveniently to other highly modular binding domains such as SH2s and PDZs. Similarly, the insulin-binding results are a powerful reminder of the many degrees of freedom in (ant)agonist design — an information-rich way forward will be to embrace the "novelty first, fitness next" mindset of foldtuning and screen putative binders against whole receptor repertoires to discover brand-new phenotypes in cell signaling space.

4.4 Methods

Oligo Pool Design and Preparation

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

Oligo pools were ordered from Twist Biosciences as ssDNA fragments for sequences $\leq$ 300nt or as dsDNA fragments for sequences $>$ 300bp and PCR-amplified with Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0494S) according to manufacturer instructions. T7RNAP promoter, ribosome binding site, start codon, stop codon, and T7 terminator elements were added in a subsequent PCR-amplification step with the same reagents, and purified, concentrated, and resuspended in ultra-pure water using the Monarch Spin PCR & DNA Cleanup Kit (NEB, T1130S) according to manufacturer instructions.

In vitro Expression Measurements

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

Concentrated purified protein samples were digested in an S-Trap micro spin column (Protifi, USA) according to the manufacturer's instructions and analyzed on Q-Exactive HF mass spectrometer coupled to EASY-nLC 1200. Peptides were separated on an Aurora UHPLC Column (25 cm × 75 µm, 1.7 µm C18, AUR3-25075C18-TS, Ion Opticks) with a flow rate of 0.35 µL/min for a total duration of 1hr and ionized at 2.2 kV in the positive ion mode. Raw data files were searched against the Uniprot Escherichia coli proteome (UP000531813) and foldtuned variant sequences. Searches used the Proteome Discoverer 2.5 software based on the Sequest HT algorithm. Oxidation / +15.995 Da (M), deamidation / +0.984 Da (N), and acetylation / +42.011 Da(N-term) were set as dynamic modifications; carbamidomethylation / +57.021 Da (C) was set as fixed modification. The precursor mass tolerance was set to 10 ppm, whereas fragment mass tolerance was set to 0.05 Da. The maximum false peptide discovery rate was specified as 0.01 using the Percolator Node validated by q-value. Absolute abundance signal intensities were scaled by dividing by the expected peptide count from simulated tryptic digestion.

In vitro Folding Stability Measurements

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

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

Barstar-Barnase Survival Assay

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

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

Bioinformatics Analysis

Multiple sequence alignments (MSAs) were calculated using MUSCLE v5 via the EMBL-EBI webserver (Edgar, 2022).

Statistical coupling analysis (SCA) was performed with PYSCA v6.1 and visualizations created with PYMOL v3.1.0 (Rivoire et al., 2016).

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.

Binding Mode Prediction and Analysis

Unless specified to the contrary, AlphaFold3 was used for all structure prediction tasks involving protein-protein or protein-peptide complexes, via the AlphaFold-Server interface (https://alphafoldserver.com). For the SH3 domain, predicted complex structures were computed for foldtuning-generated putative SH3 variants in the presence of a representative class I (RPLPPLP) or class II (PPPLPPRP) proline-rich peptide motif. For the barstar-like fold, predicted complex structures were computed for foldtuning-generated putative barstar variants in the presence of wild-type barnase from *B. amyloliquefaciens*(uniprot:P00648). Predicted structures were compared to a wild-type reference, either the spectrin SH3 domain from *Gallus gallus* or the barnase-barstar complex from *Bacillus amyloliquefaciens* (pdb: 1brs). For insulin, predicted complex structures were computed for foldtuning-generated and/or PLM-sampled putative insulin variants in complex with the monomeric full-length ectodomain of human INSR (insulin receptor).

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

APBS Electrostatics plugin.

High-Throughput Insulin Binding Assay

A library of 2889 insulin variant amino-acid sequences was constructed by foldtuning on InterPro entry IPR004825, containing 335 natural insulin sequences (reduced to 193 sequences after deduplication clustering at 100% similarity with MMSEQS2) integrated from overlapping entries in the PRINTS, CDD, and PANTHER databases. Foldtuning was executed as described in Section 3.4, with the modification that generated variants were post-processed by aligning to the sequence *H. sapiens* insulin (uniprot: P01308) and removing residues aligning to the C-peptide region that is removed by proteolytic cleavage *in vivo* during the conversion of inactive proinsulin to active insulin, resulting in a library of *single-chain* insulin mimics.

High throughput binding measurements (sequencing read enrichment scores) were obtained using the Protein CREATE platform as described in Lourenco et al. (2025) with INSR as the on-target receptor and IL7RA as the off-target decoy receptor.

Insulin Variant Generation by MHMC Sampling

Additional insulin variants (not screened with Protein CREATE) were generated through Metropolis-Hastings Monte Carlo (MHMC) sampling from an insulin-like sequence landscape with an two-term energy function combining a preference for accepting mutations that increase sequence-likelihood under the ESM2-650M model (the "exploit" term) with a preference for accepting mutations resulting in a large semantic change relative to the current sequence (the "explore" term). An individual sequence s_i of length N has an associated log-likelihood $L_i = \prod_{k=1}^{N} l_k$ where the l_k represent indepedent residue-wise likelihoods, and an ESM2-650M final-layer mean-pooled embedding vector $\mathbf{x_i}$. Semantic change is defined as $S_{i\to j} = ||\mathbf{x_j} - \mathbf{x_i}||_1$ for a pair of sequences s_i, s_j. As semantic change is not defined for individual sequences, it is not possible to define an absolute sequence energy E_i; it is however possible to define $\Delta E_{i\to j}$ for a proposed move from s_i to s_j. Precisely

$$\Delta E_{i\to j} = (\log L_j - \log L_i) + w_s S_{i\to j} \tag{4.1}$$

where $L_i, L_j, S_{i\to j}$ are defined as above and w_s is a coefficient that controls the relative weights assigned to the exploit and explore terms.

The standard Metropolis-Hastings acceptance criterion is used; namely, a proposed move (restricted in this method to a single point mutation, sampled uniformly across sequence positions and amino-acid identities) from s_i to s_j is accepted with probability

$$p_{i \to j} = \min\{1, \exp(\beta \Delta E_{i \to j})\} \tag{4.2}$$

where β refers to the thermodynamic β, the inverse of the sampling temperature T.

In all variant generation runs for this study, MHMC was run for $n = 2000$ steps, $w_s = 0.4$, and adaptive temperature adjustment at 100-step intervals. For each run, the intial sequence s_0 was sampled column-wise from a multiple sequence alignment of 687 wild-type insulin sequences obtained by querying UniProt for all matches to the INS gene, with individual amino-acid sampling probabilities proportional to the amino-acid distribution over each individual column. Sampled gap characters and putative C-peptide regions were removed prior to concatenation into s_0.

A total of 100 independent MHMC replicates were performed. Sequences with fewer than 4 or an odd number of cysteine residues were removed. Sequences were converted to single-chain fusions by inserting a GGGRGG loop in-between the concatenated A-peptide and B-peptide The resulting sequences were binned into four quadrants by sequence identity % to *H. sapiens* insulin and predicted iPTM score in complex with the INSR receptor ectodomain according to AlphaFold-Multimer prediction, with 20 variants across the quadrants forwarded for attempted expression, refolding, and activity characterization.

Expression and Validation of Insulin Mimic Cellular Activity

Twenty insulin mimics designed via "explore-exploit" Metropolis-Hastings Monte Carlo (eeMHMC) sampling from the ESM2-650M model were expressed in *E. coli* and refolding was attempted as previously reported in the literature (Chen et al., 2016; Min et al., 2011). Refolded monomers were isolated by reverse-phase high pressure liquid chromatography (RP-HPLC) and evaluated for insulin receptor agonist activity using a sandwich ELISA INSR-β subunit assay specific for receptor tyrosine residue phosphorylation (Maloney et al., 2003). EC50 values were inferred from a four-parameter logistic regression model fit to 450 nm absorbance vs. variant concentration data.

4.5 Supplemental Material

Supplemental Figures

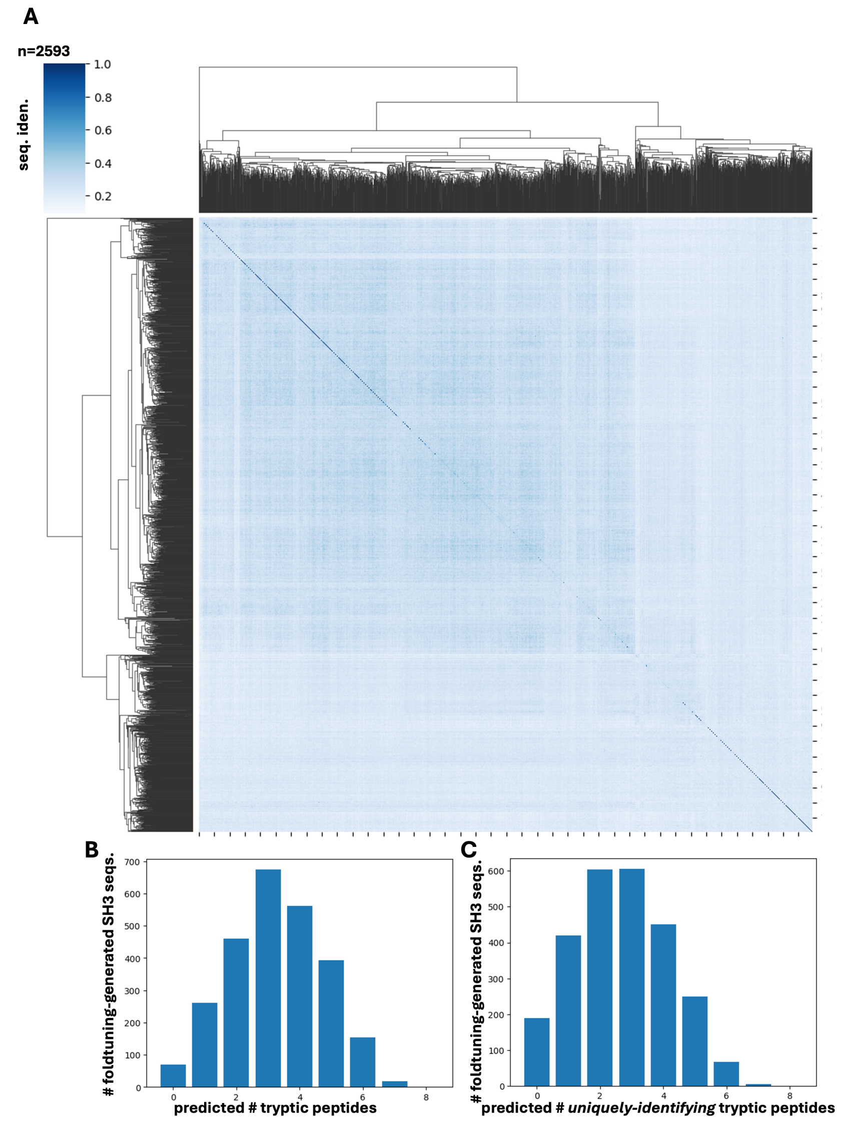


Figure S4.1: **Sequence diversity and detectability of foldtuning-generated SH3 domains.** (**A**) Hierarchically clustered heatmap of pairwise sequence identity between $n = 2593$ SH3 domain candidate sequences generated via foldtuning. (**B**) Expected detectable peptide counts predicted by in silico tryptic digestion. (**C**) Counts of predicted tryptic peptides that map uniquely to single foldtuned SH3 variants.

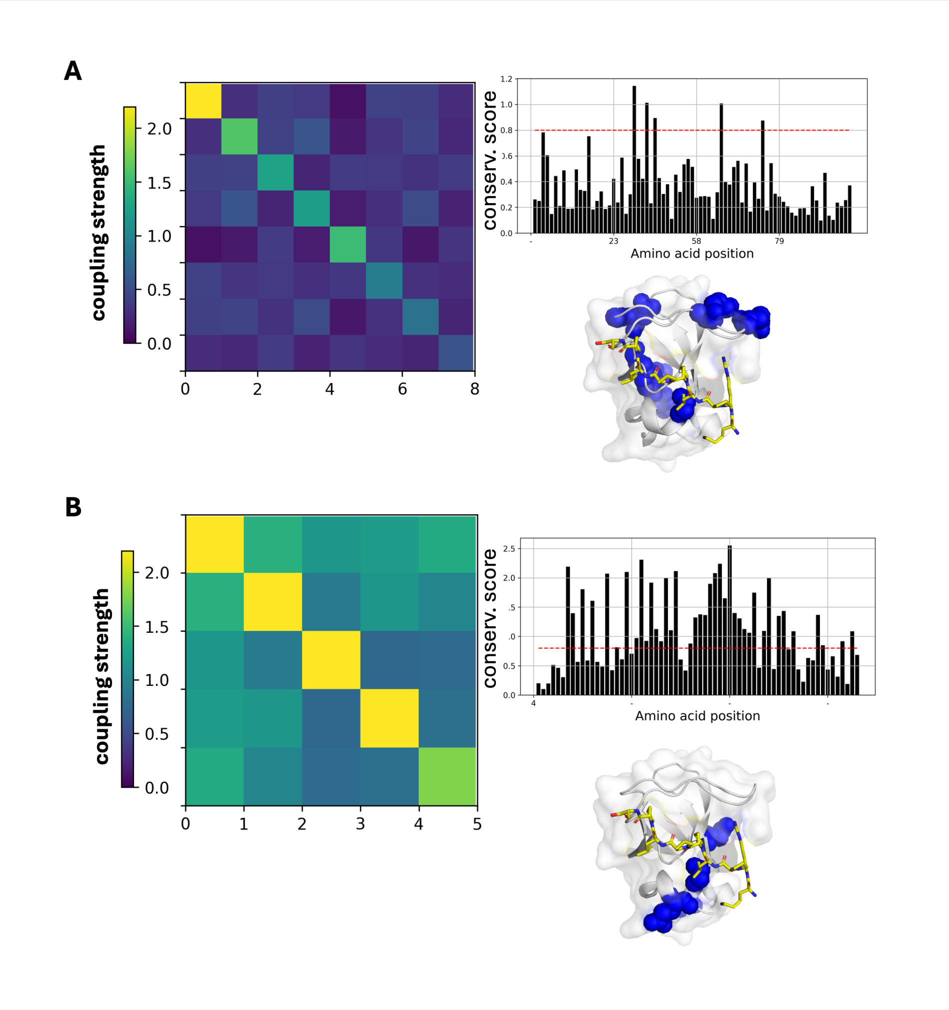
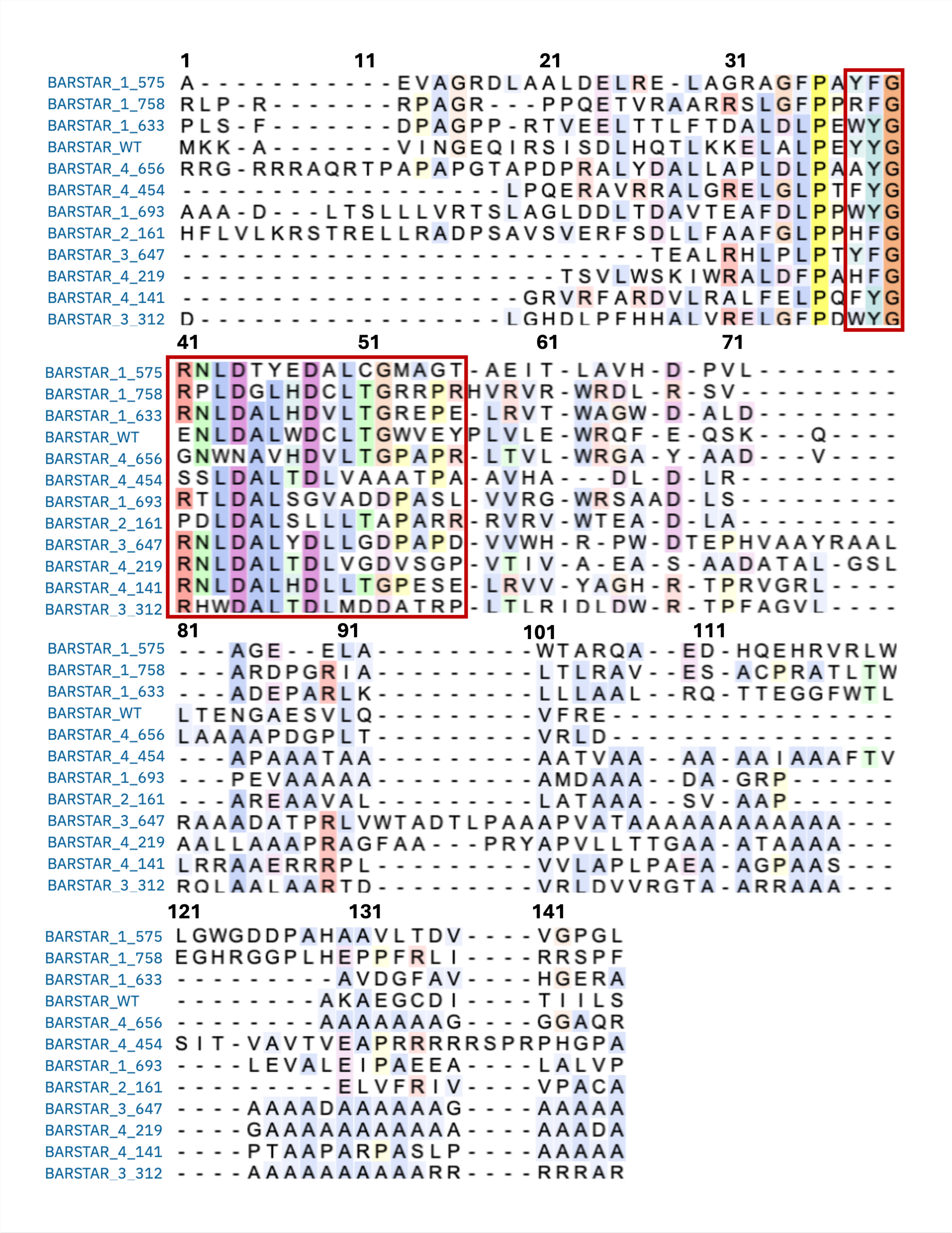


Figure S4.2: **Statistical coupling analysis of natural and synthetic SH3s.** (**A**) Results of statistical coupling analysis (SCA) on $n \approx 2500$ natural SH3 domain sequences. Left: Second-order compressed coupling matrix, blocked into a single statistically interacting sector. Top right: First-order conservation scores. Bottom right: Visualization of sector positions (blue) mapped onto a representative structure of a natural SH3 domain (from PI3K) bound to a proline-rich peptide ligand (pdb: 3I5R). (**B**) Results of statistical coupling analysis (SCA) on $n = 2593$ foldtuned SH3 sequences. Left: Second-order compressed coupling matrix, blocked into a single statistically interacting sector. Top right: First-order conservation scores. Bottom right: Visualization of sector positions (blue) mapped onto a representative structure of a natural SH3 domain (from PI3K) bound to a proline-rich peptide ligand (pdb: 3I5R)



```
             1         11        21        31
BARSTAR_1_575 A-----------EVAGRDLAALDELRE-LAGRAGFPAYFG
BARSTAR_1_758 RLP-R-------RPAGR---PPQETVRAARRSLGFPPRFG
BARSTAR_1_633 PLS-F-------DPAGPP-RTVEELTTLFTDALDLPEWYG
BARSTAR_WT    MKK-A-------VINGEQIRSISDLHQTLKKELALPEYYG
BARSTAR_4_656 RRG-RRRAQRTPAPAPGTAPDPRALYDALLAPLDLPAAYG
BARSTAR_4_454 ------------------LPQERAVRRALGRELGLPTFYG
BARSTAR_1_693 AAA-D---LTSLLLVRTSLAGLDDLTDAVTEAFDLPPWYG
BARSTAR_2_161 HFLVLKRSTRELLRADPSAVSVERFSDLLFAAFGLPPHFG
BARSTAR_3_647 --------------------------TEALRHLPLPTYFG
BARSTAR_4_219 ---------------------TSVLWSKIWRALDFPAHFG
BARSTAR_4_141 -------------------GRVRFARDVLRALFELPQFYG
BARSTAR_3_312 D-----------------LGHDLPFHHALVRELGFPDWYG
             41        51        61        71
BARSTAR_1_575 RNLDTYEDALCGMAGT-AEIT-LAVH-D-PVL--------
BARSTAR_1_758 RPLDGLHDCLTGRRPRHVRVR-WRDL-R-SV---------
BARSTAR_1_633 RNLDALHDVLTGREPE-LRVT-WAGW-D-ALD--------
BARSTAR_WT    ENLDALWDCLTGWVEYPLVLE-WRQF-E-QSK---Q----
BARSTAR_4_656 GNWNAVHDVLTGPAPR-LTVL-WRGA-Y-AAD---V----
BARSTAR_4_454 SSLDALTDLVAAATPA-AVHA---DL-D-LR---------
BARSTAR_1_693 RTLDALSGVADDPASL-VVRG-WRSAAD-LS---------
BARSTAR_2_161 PDLDALSLLLTAPARR-RVRV-WTEA-D-LA---------
BARSTAR_3_647 RNLDALYDLLGDPAPD-VVWH-R-PW-DTEPHVAAYRAAL
BARSTAR_4_219 RNLDALTDLVGDVSGP-VTIV-A-EA-S-AADATAL-GSL
BARSTAR_4_141 RNLDALHDLLTGPESE-LRVV-YAGH-R-TPRVGRL----
BARSTAR_3_312 RHWDALTDLMDDATRP-LTLRIDLDW-R-TPFAGVL----
             81        91        101       111
BARSTAR_1_575 ---AGE--ELA---------WTARQA--ED-HQEHRVRLW
BARSTAR_1_758 ---ARDPGRIA---------LTLRAV--ES-ACPRATLTW
BARSTAR_1_633 ---ADEPARLK---------LLLAAL--RQ-TTEGGFWTL
BARSTAR_WT    LTENGAESVLQ---------VFRE----------------
BARSTAR_4_656 LAAAAPDGPLT---------VRLD----------------
BARSTAR_4_454 ---APAAATAA---------AATVAA--AA-AAIAAAFTV
BARSTAR_1_693 ---PEVAAAAA---------AMDAAA--DA-GRP------
BARSTAR_2_161 ---AREAAVAL---------LATAAA--SV-AAP------
BARSTAR_3_647 RAAADATPRLVWTADTLPAAAPVATAAAAAAAAAAAA---
BARSTAR_4_219 AALLAAAPRAGFAA---PRYAPVLLTTGAA-ATAAAA---
BARSTAR_4_141 LRRAAERRRPL---------VVLAPLPAEA-AGPAAS---
BARSTAR_3_312 RQLAALAARTD---------VRLDVVRGTA-ARRAAA---
             121       131       141
BARSTAR_1_575 LGWGDDPAHAAVLTDV----VGPGL
BARSTAR_1_758 EGHRGGPLHEPPFRLI----RRSPF
BARSTAR_1_633 ---------AVDGFAV----HGERA
BARSTAR_WT    --------AKAEGCDI----TIILS
BARSTAR_4_656 --------AAAAAAAG----GGAQR
BARSTAR_4_454 SIT-VAVTVEAPRRRRRSPRPHGPA
BARSTAR_1_693 ----LEVALEIPAEEA----LALVP
BARSTAR_2_161 ---------ELVFRIV----VPACA
BARSTAR_3_647 ----AAAADAAAAAAG----AAAAA
BARSTAR_4_219 ----GAAAAAAAAAAA----AAADA
BARSTAR_4_141 ----PTAAPARPASLP----AAAAA
BARSTAR_3_312 ----AAAAAAAAAARR----RRRAR
```

Figure S4.3: **Multiple sequence alignment (MSA) of toxicity-rescuing barstar variants.** Muliple sequence alignment (MSA) of the eleven toxicity-rescuing fold-tuned barstar variants and wild-type barstar from *B. aquaforiensis*. Columns corresponding to residue positions making physical contacts (positions 38-56; distance threshold < 4.0 Å) with barnase in a reference crystal structure (PDB: 1BRS) are boxed.

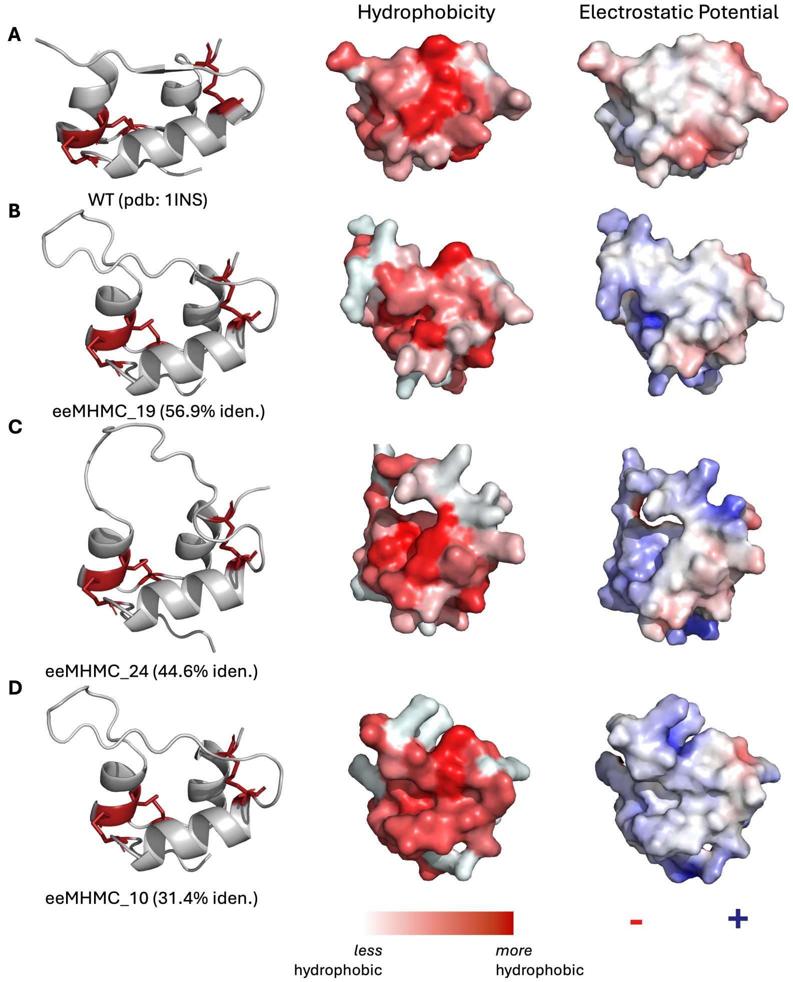


Figure S4.4: **eeMHMC-generated insulin variants respect structural and physicochemical plausibility of the canonical insulin peptide. left**: ESMFold predicted structures with disulfide bonds highlighted in red; **middle**: solvent-accessible surface visualization colored by hydrophobicity (Eisenberg scale); **right**: same as middle, colored by electrostatic potential with positive/negative charge in blue/red respectively. (**A**) wild-type human insulin (pdb: 1INS). (**B**) eeMHMC_19 (56.9% iden.). (**C**) eeMHMC_24 (44.6% iden.). (**D**) eeMHMC_10 (31.4% iden.)

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 predicted 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.

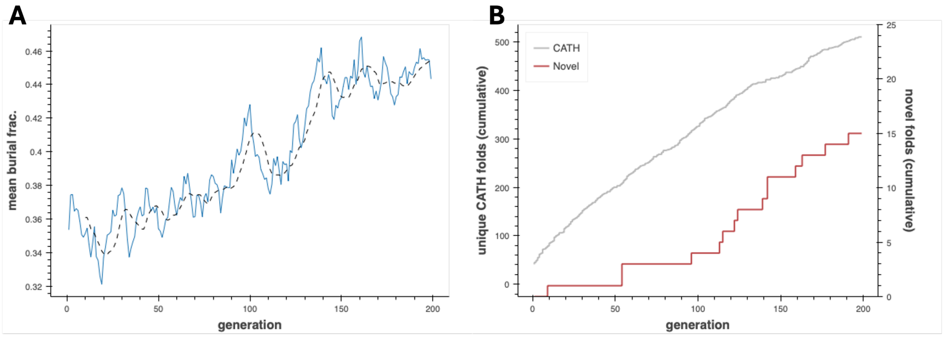


Figure 5.1: **Emergence of novel folds from a PLM-based genetic algorithm.** (**A**) Mean fractional amino-acid surface burial (protein compactness proxy) over 200 generations of the structure discovery genetic algorithm. (**B**) Cumulative counts of unique CATH-annotated folds and putative novel folds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

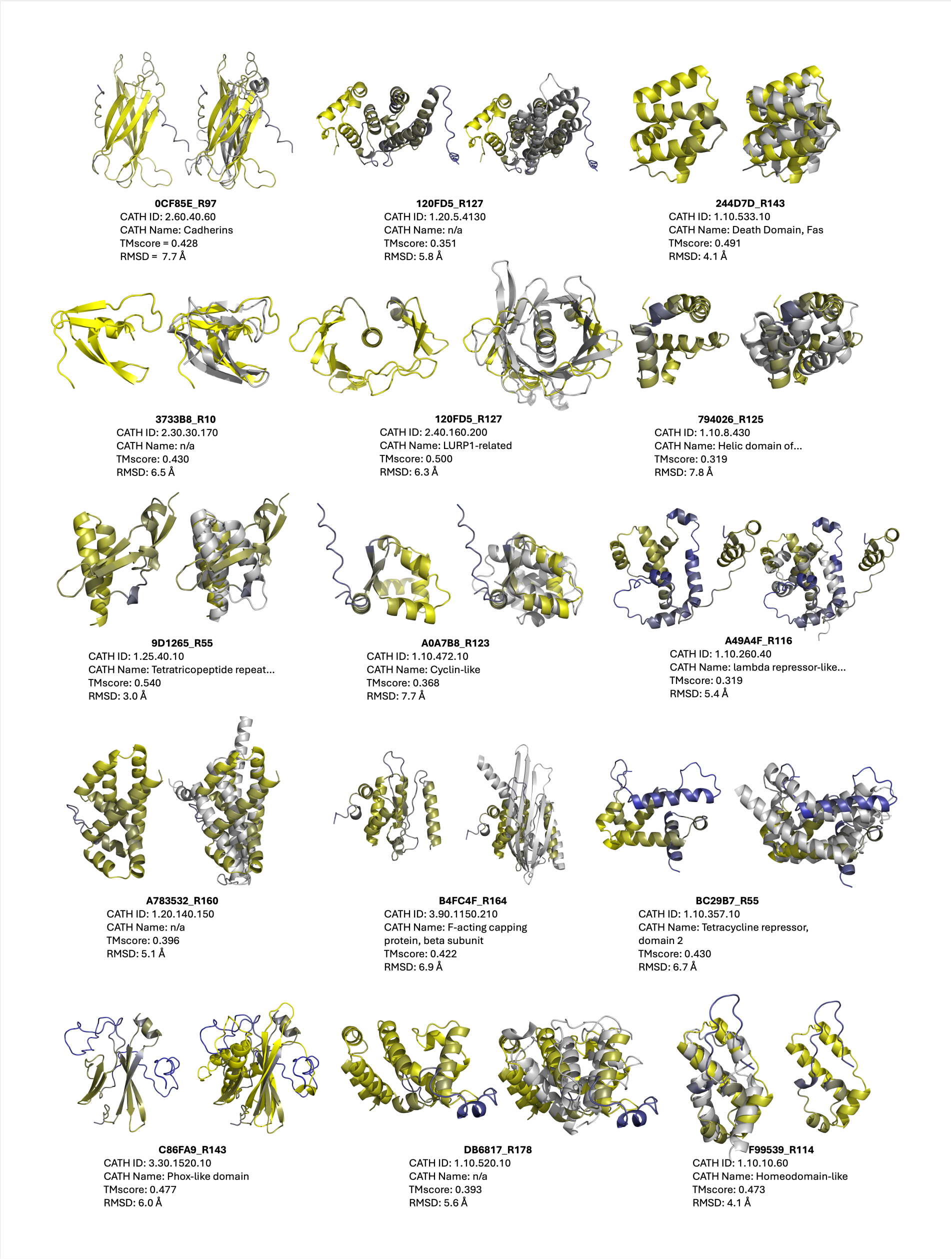


Figure 5.2: **Fifteen novel folds achieved by the structure discovery genetic algorithm.** Within each pair: **left** — putative novel fold (colored by ESMFold pLDDT; yellow=high, blue=low); **right** superimposed with closest CATHDB50 Foldseek hit in TMalign mode, with CATH metadata and global alignment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

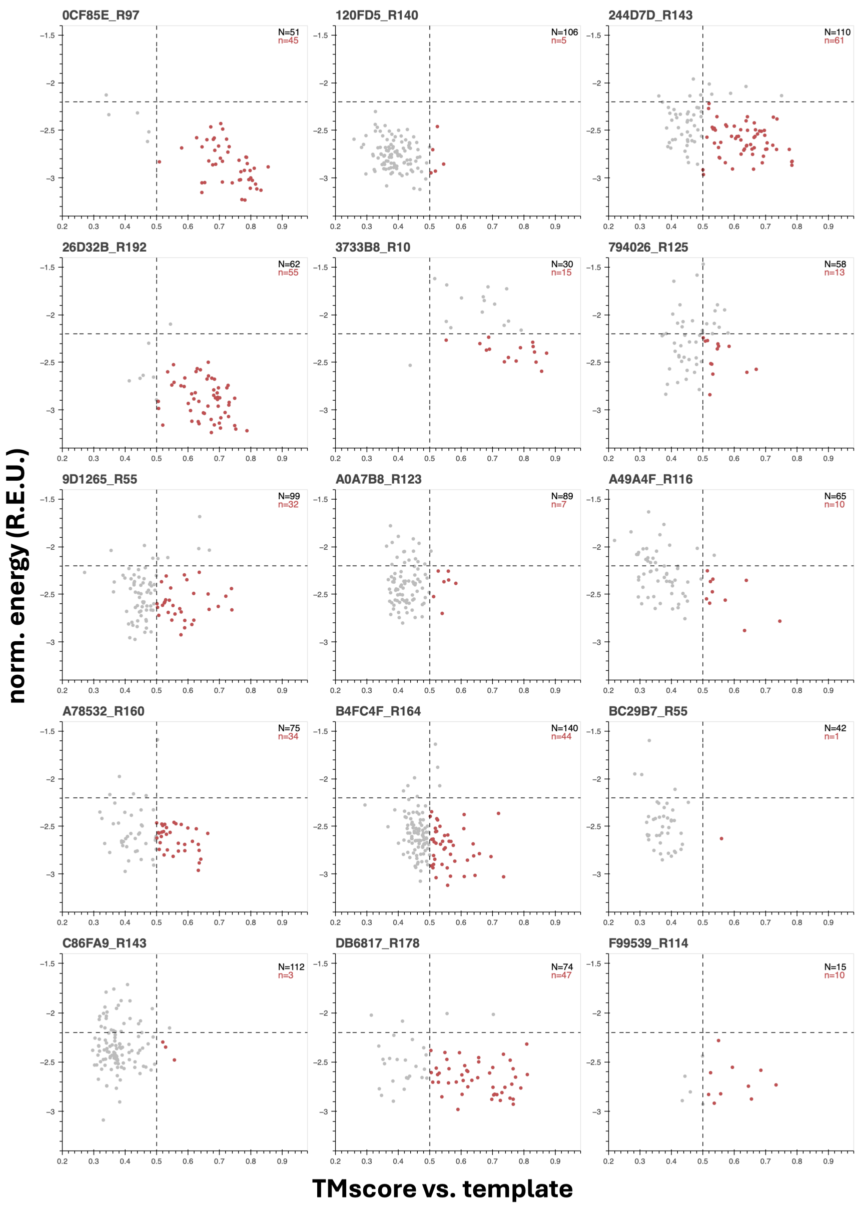


Figure 5.3: **"Inverse-folding landscapes" for fifteen novel folds achieved by the structure discovery genetic algorithm suggest variable stability.** Length-normalized energies (from Rosetta) vs. TM-score (from Foldseek in TMalign mode) for ProteinMPNN-designed sequences inverse-folded off of structure 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 — ≈ 0.1% — exhibit detectable sequence similarity to any natural protein

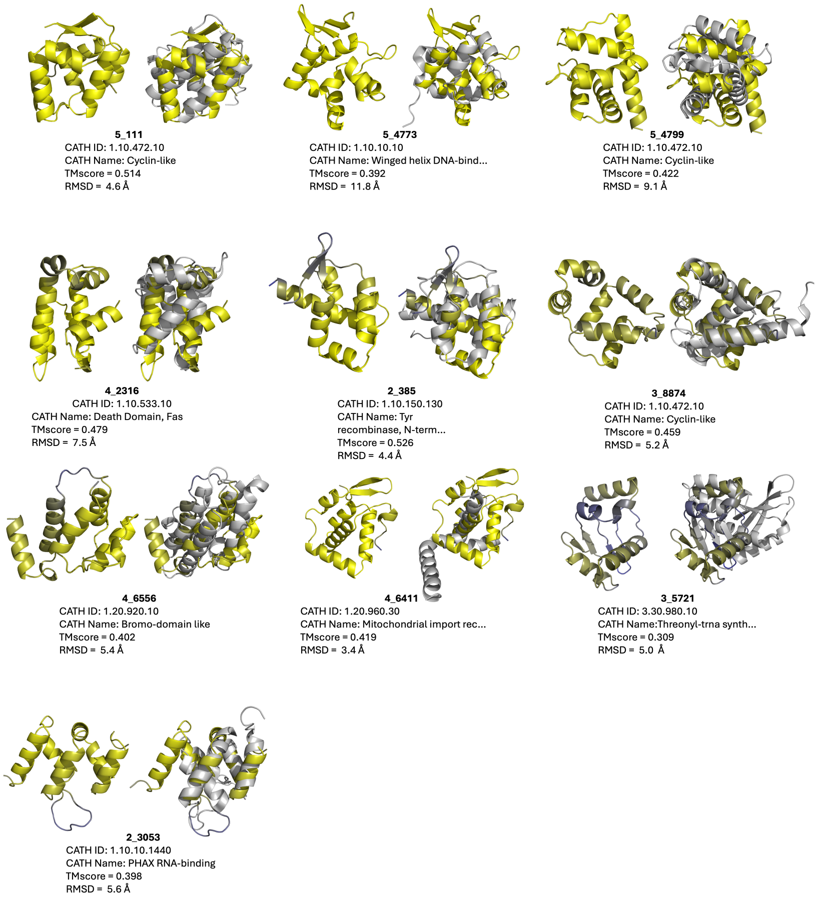


Figure 5.4: **Ten out of 100 novel folds achieved by structure-first foldtuning.** Within each pair: **left** — putative novel fold (colored by ESMFold pLDDT; yellow=high, blue=low); **right** superimposed with closest CATHDB50 Foldseek hit in TMalign mode, with CATH metadata and global alignment metrics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

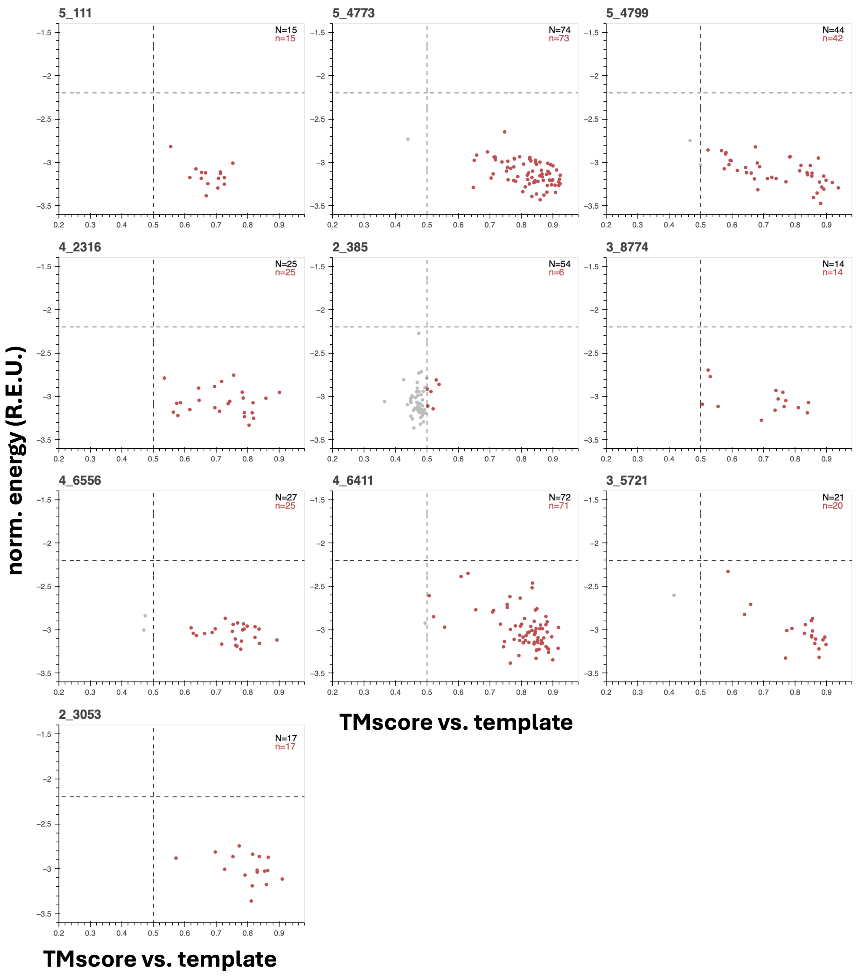


Figure 5.5: **"Inverse-folding landscapes" for ten out of 100 novel folds achieved by structure-first foldtuning imply high stability.** Length-normalized energies (from Rosetta) vs. TM-score (from Foldseek in TMalign mode) for ProteinMPNN-designed sequences inverse-folded off of structure-first foldtuning 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

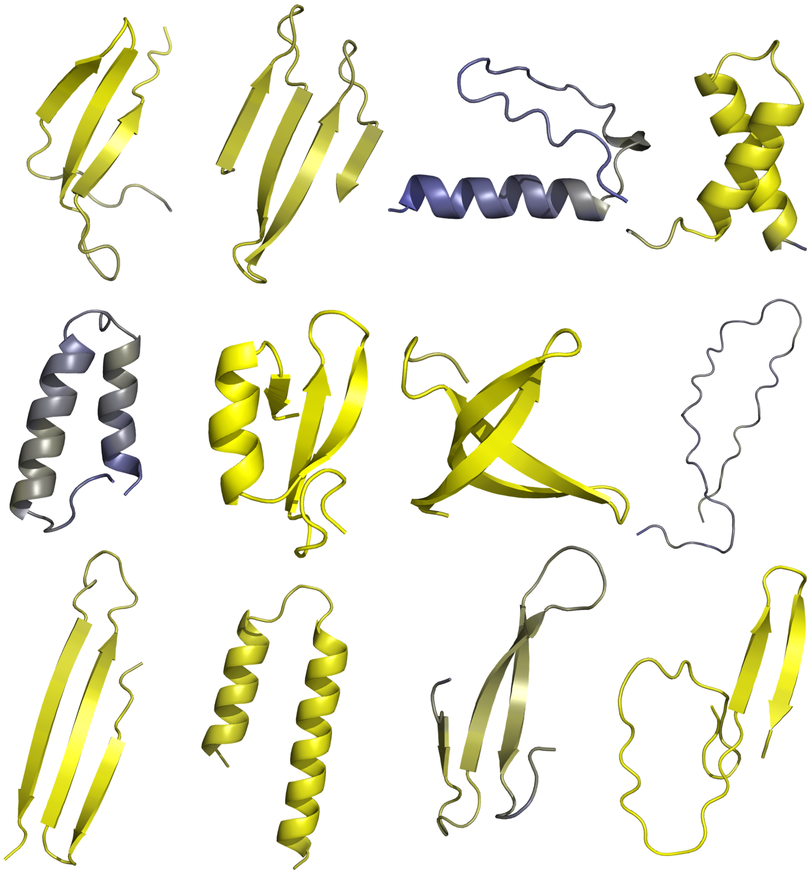


Figure S5.1: **Example structure fragments generated by RE-eeMHMC.** 10 of 800 structure fragments predicted from sequences designed by replica-exchange explore-exploit Metropolis-Hastings Monte Carlo sampling (RE-eeMHMC). Individual structures are colored by ESMFold pLDDT; yellow=high, blue=low.

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

Chapter 6

CONCLUDING REMARKS

6.1 From foldtuning to foundation models

Foldtuning is not an an algorithm or model collection frozen in time. Indeed, a leading strength of foldtuning is its modularity and generalizability to complex design tasks. Much of this modular nature is bestowed by the GAN-like division into generator and discriminator. In principle, any PLM could act as the generator, provided an appropriate procedure for sampling from the model in question. The choice of ProtGPT2 as the initial generative model for foldtuning was motivated by its relative success at proposing novel, reasonable, and representative protein sequences, but other promising approaches — e.g. direct embedding-to-sequence decoders for encoder-only architectures (as recently demonstrated for ESM2) — could be swapped in with minimal implementation burden (Chen et al., 2024). The discriminator side is in some sense endlessly flexible. We began with a structure-based filter to pursue minimal constraints on individual fold families, but foldtuning was designed to be amenable to an arbitrary scoring function — even an ensemble of scoring functions — from predicted stability or optimal pH, to active site preorganization, to molecular-dynamics-derived root mean square fluctuation (RMSF), and model-provided confidence metrics, depending on the exact problem of interest. In that regard, structure-first foldtuning may be considered the first such spinoff of foldtuning, replacing scoring that favors structural matches with an objective that rewards three-dimensional compactness and "anti-matching" so as to prefer structural novelty.

Other discriminator-side changes could confer improvements in compute perfomance and overhead. The total compute burden (cost + time) of foldtuning, as currently implemented, is set by the structure prediction step — even a few seconds of inference time per sequence adds up to four GPU-hrs for four rounds of foldtuning with pre-evotuning.[1] Replacing the time-consuming explicit structure prediction step with conversion to a suitable one-dimensional representation of structure, such as the increasingly utilized 3Di sidechain-aware structural alphabet developed for

[1] Benchmarked on a single NVIDIA A100 GPU with 80GB of memory; the optimal monetary vs time cost tradeoff will depend on the number of foldtuned models required, hardware technical specifications, and highly variable capital and/or hourly cost differences between providers.

Foldseek, could accelerate foldtuning by as much as $\sim$ 10x over current internal benchmarks (Heinzinger et al., 2023; van Kempen et al., 2023).

As a unit, also, foldtuning need not be an end unto itself. With its iterative update structure, foldtuning is architecturally amenable to direct integration[2] of experimental measurements — positive and negative — of generated variants via reinforcement learning (RL). Incorporating RL atop foldtuning is a logical next step for guiding models that have already learned sequence novelty under hidden language rules towards empirical evidence of function, whether for the experimental results presented in this work or for any arbitrary target and/or assay. Beyond improving future batches of generated proteins based on real-world data, we envision two additional related model-side advances of import for AI-guided synthetic biology. First, recent theoretical findings on general LLMs and hands-on application of chemical language models for small-molecule representation and generation have independently pushed back on the axiom that breadth and novelty are incompatible and argued that training on labeled positive and negative examples mitigates the dilemma (Kalavasis et al., 2025; Skinnider, 2024). Consequently, we can imagine incorporating information from negative examples — both those filtered out *in silico* and those deriving from experiment — into a single foldtuning foundation model that achieves full structural coverage (including novel domains) without mode collapse or hallucination. Second, foldtuning can form one end of an end-to-end model linking sequence-level specification of, e.g. a binder with an arbitrary agonism/antagonism profile against cell-surface receptors, to single-cell transcriptomic readout for design of bespoke cell-signaling programs.

6.2 Producing and propagating protein novelty across scales

In this work, foldtuning was restricted to single-domain targets, a biophysically meaningful, and well-annotated level at which to first segment. Generation and optimization of individual domains offers much to be excited about, including cytokine- and chemokine-like binders as in the example posed above, host-defense peptide-mimicking antimicrobials, and biosensor toolkits enabled by fluorescent protein property expansion, all areas of ongoing interest. Complex systems of proteins, on the other hand, are built up in layers of physical organization, compartmentalization, and interaction. As an agent of domain diversification, foldtuning can power the design of full-blown protein *systems* for AI-guided synthetic biology and cellular engineering, such as signaling cascades assembled out of foldtuned kinases, SH2s, and

[2] As opposed to ranking, selection, and updating by finetuning.

SH3s, or gene regulatory networks based on the many flavors of seemingly highly-designable DNA-binding domains, or multi-step pathway catalytic machinery for xenobiotic metabolism.

Lastly, if there is one defining theme of this thesis, it is that PLMs — whether through foldtuning, MHMC sampling, or any other method — are producers and propagators of novelty in sequence, structure, and function. With the right steering and forcing, PLMs readily expand the boundaries of valid protein-space at greater rates than the accumulation and processing of (meta)genomic data can. Putting generative novelty ahead of prespecified phenotype leads to new fold-centered language rules, evolution-esque innovation of structure and function, and mechanistic hints towards the fundamental constraints that dictate the structure→function transition. Ruminating on the virtues of novelty-first methods for looking forward and backwards in time leads to persistent open questions including: What can alternate sequence rules reveal about the primordial emergence of the first proteins? What are the smallest collections of sequences and structures that can sustain the essential functions of a minimal cell and how do they overlap (or not) with what we observe in nature today? How can we leverage the "structure of feature-space" — a PLM's internal navigational charts, as it were — to further accelerate the search for new-to-nature sequences, structures, and functions?

Peering a final time at the sequence→structure map, we have, through several strands of novelty-directed exploration, replaced certain mythical monsters with the outlines of heretofore unknown landmasses; the challenge persists to fully characterize and capitalize on all that these addenda confer.

BIBLIOGRAPHY

Alford, Rebecca F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13 (6):3031–3048, June 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125. Publisher: American Chemical Society.

Alva, Vikram. A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, 4:e09410, December 2015. ISSN 2050-084X. doi: 10.7554/eLife.09410. URL https://doi.org/10.7554/eLife.09410. Publisher: eLife Sciences Publications, Ltd.

Andreeva, Antonina et al. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1):D376–D382, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz1064. URL https://doi.org/10.1093/nar/gkz1064.

Anfinsen, Christian B. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, July 1973. doi: 10.1126/science.181.4096.223. URL https://www.science.org/doi/10.1126/science.181.4096.223. Publisher: American Association for the Advancement of Science.

Baker, David. A surprising simplicity to protein folding. *Nature*, 405(6782): 39–42, May 2000. ISSN 1476-4687. doi: 10.1038/35011000. URL https://www.nature.com/articles/35011000. Number: 6782 Publisher: Nature Publishing Group.

Barrio-Hernandez, Inigo et al. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06510-w. URL https://www.nature.com/articles/s41586-023-06510-w. Number: 7983 Publisher: Nature Publishing Group.

Benjamini, Yoav and Yekutieli, Daniel. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, August 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013699998. URL https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-4/The-control-of-the-false-discovery-rate-in-multiple-testing/10.1214/aos/1013699998.full. Publisher: Institute of Mathematical Statistics.

Blum, Matthias et al. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1082. URL https://doi.org/10.1093/nar/gkae1082.

Bornberg-Bauer, E. How are model protein structures distributed in sequence space? *Biophysical Journal*, 73(5):2393–2403, November 1997. ISSN 00063495. doi: 10.1016/S0006-3495(97)78268-7. URL https://linkinghub.elsevier.com/retrieve/pii/S0006349597782687.

Bornberg-Bauer, Erich and Chan, Hue Sun. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences*, 96(19):10689–10694, September 1999. doi: 10.1073/pnas.96.19.10689. URL https://www.pnas.org/doi/full/10.1073/pnas.96.19.10689. Publisher: Proceedings of the National Academy of Sciences.

Chen, Bo et al. xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. Technical report, bioRxiv, July 2023. URL https://www.biorxiv.org/content/10.1101/2023.07.05.547496v3. Section: New Results Type: article.

Chen, Tianlai et al. PepMLM: Target Sequence-Conditioned Generation of Therapeutic Peptide Binders via Span Masked Language Modeling. *ArXiv*, page arXiv:2310.03842v3, August 2024. ISSN 2331-8422. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10593082/.

Chen, Ying et al. Improving the refolding efficiency for proinsulin aspart inclusion body with optimized buffer compositions. *Protein Expression and Purification*, 122:1–7, June 2016. ISSN 10465928. doi: 10.1016/j.pep.2016.01.015. URL https://linkinghub.elsevier.com/retrieve/pii/S1046592816300158.

Chitturi, Bhadrachalam et al. Compact Structure Patterns in Proteins. *Journal of Molecular Biology*, 428(21):4392–4412, October 2016. ISSN 0022-2836. doi: 10.1016/j.jmb.2016.07.022. URL https://www.sciencedirect.com/science/article/pii/S0022283616302893.

Choi, In-Geol and Kim, Sung-Hou. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences*, 103(38):14056–14061, September 2006. doi: 10.1073/pnas.0606239103. URL https://www.pnas.org/doi/abs/10.1073/pnas.0606239103. Publisher: Proceedings of the National Academy of Sciences.

Chomsky, Noam. On certain formal properties of grammars. *Information and Control*, 2(2):137–167, June 1959. ISSN 0019-9958. doi: 10.1016/S0019-9958(59)90362-6. URL https://www.sciencedirect.com/science/article/pii/S0019995859903626.

Claeys, Ilse et al. Insulin-related peptides and their conserved signal transduction pathway. *Peptides*, 23(4):807–816, April 2002. ISSN 0196-9781. doi: 10.1016/S0196-9781(01)00666-0. URL https://www.sciencedirect.com/science/article/pii/S0196978101006660.

Dahiyat, Bassil I. and Mayo, Stephen L. De Novo Protein Design: Fully Automated Sequence Selection. *Science*, 278(5335):82–87, October 1997. doi: 10.1126/science.278.5335.82. URL https://www.science.org/doi/full/10.1126/science.278.5335.82. Publisher: American Association for the Advancement of Science.

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

Davidson, A R and Sauer, R T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proceedings of the National Academy of Sciences*, 91(6):2146–2150, March 1994. doi: 10.1073/pnas.91.6.2146. URL https://www.pnas.org/doi/abs/10.1073/pnas.91.6.2146. Publisher: Proceedings of the National Academy of Sciences.

Davidson, Alan R. Cooperatively folded proteins in random sequence libraries. *Nature Structural Biology*, 2(10):856–864, October 1995. ISSN 1545-9985. doi: 10.1038/nsb1095-856. URL https://www.nature.com/articles/nsb1095-856. Publisher: Nature Publishing Group.

Devlin, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, Jill, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Dill, Ken A. and Chan, Hue Sun. From Levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10–19, January 1997. ISSN 1545-9985. doi: 10.1038/nsb0197-10. URL https://www.nature.com/articles/nsb0197-10. Publisher: Nature Publishing Group.

Dupont, Christopher L. et al. History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proceedings of the National Academy of Sciences*, 107(23):10567–10572, June 2010. doi: 10.1073/pnas.0912491107. URL https://www.pnas.org/doi/abs/10.1073/pnas.0912491107. Publisher: Proceedings of the National Academy of Sciences.

Durairaj, Janani et al. Uncovering new families and folds in the natural protein universe. *Nature*, 622(7983):646–653, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06622-3. URL https://www.nature.com/articles/s41586-023-06622-3. Publisher: Nature Publishing Group.

Edgar, Robert C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, 13(1): 6968, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34630-w. URL `https://www.nature.com/articles/s41467-022-34630-w`. Publisher: Nature Publishing Group.

Elnaggar, Ahmed et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

England, Jeremy L. and Shakhnovich, Eugene I. Structural Determinant of Protein Designability. *Physical Review Letters*, 90(21):218101, May 2003. doi: 10.1103/PhysRevLett.90.218101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.90.218101`. Publisher: American Physical Society.

Fahlberg, Sarah A. et al. Neural network extrapolation to distant regions of the protein fitness landscape, November 2023. URL `https://www.biorxiv.org/content/10.1101/2023.11.08.566287v1`. Pages: 2023.11.08.566287 Section: New Results.

Ferruz, Noelia and Höcker, Birte. Dreaming ideal protein structures. *Nature Biotechnology*, 40(2):171–172, February 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01196-9. URL `https://www.nature.com/articles/s41587-021-01196-9`. Number: 2 Publisher: Nature Publishing Group.

Ferruz, Noelia. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32007-7. URL `https://www.nature.com/articles/s41467-022-32007-7`. Number: 1 Publisher: Nature Publishing Group.

Fontana, Walter et al. RNA folding and combinatory landscapes. *Physical Review E*, 47(3):2083–2099, March 1993. doi: 10.1103/PhysRevE.47.2083. URL `https://link.aps.org/doi/10.1103/PhysRevE.47.2083`. Publisher: American Physical Society.

Frey, Nathan C. et al. Cramming Protein Language Model Training in 24 GPU Hours, May 2024. URL `https://www.biorxiv.org/content/10.1101/2024.05.14.594108v1`. Pages: 2024.05.14.594108 Section: New Results.

Garcia, Lucas et al. Design of Novel Dehalogenases using Protein Large Language Models, October 2024. URL `https://www.biorxiv.org/content/10.1101/2024.10.28.620469v1`. Pages: 2024.10.28.620469 Section: New Results.

Goodfellow, Ian J. et al. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.,

2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

Govindarajan, S and Goldstein, R A. Why are some proteins structures so common? *Proceedings of the National Academy of Sciences*, 93(8):3341–3345, April 1996. doi: 10.1073/pnas.93.8.3341. URL https://www.pnas.org/doi/abs/10.1073/pnas.93.8.3341. Publisher: Proceedings of the National Academy of Sciences.

Halabi, Najeeb et al. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4):774–786, August 2009. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2009.07.038. URL https://www.cell.com/cell/abstract/S0092-8674(09)00963-5. Publisher: Elsevier.

Hartley, Robert W. [38] - Barnase–Barstar Interaction. In Nicholson, Allen W., editor, *Methods in Enzymology*, volume 341 of *Ribonucleases - Part A*, pages 599–611. Academic Press, January 2001. doi: 10.1016/S0076-6879(01)41179-7. URL https://www.sciencedirect.com/science/article/pii/S0076687901411797.

Hecht, Michael H. et al. De novo heme proteins from designed combinatorial libraries. *Protein Science*, 6(12):2512–2524, 1997. ISSN 1469-896X. doi: 10.1002/pro.5560061204. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.5560061204. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560061204.

Heinzinger, Michael et al. ProstT5: Bilingual Language Model for Protein Sequence and Structure. Technical report, bioRxiv, July 2023. URL https://www.biorxiv.org/content/10.1101/2023.07.23.550085v1. Section: New Results Type: article.

Helling, Robert et al. The designability of protein structures. *Journal of Molecular Graphics and Modelling*, 19(1):157–167, February 2001. ISSN 1093-3263. doi: 10.1016/S1093-3263(00)00137-6. URL https://www.sciencedirect.com/science/article/pii/S1093326300001376.

Hie, Brian et al. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021. doi: 10.1126/science.abd7331. URL https://www.science.org/doi/full/10.1126/science.abd7331. Publisher: American Association for the Advancement of Science.

Hie, Brian L. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285.e6, April 2022. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2022.01.003. URL https://www.cell.com/cell-systems/abstract/S2405-4712(22)00038-2. Publisher: Elsevier.

Hsu, Chloe et al. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8946–8970. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hsu22a.html. ISSN: 2640-3498.

Huang, Po-Ssu. The coming of age of de novo protein design. *Nature*, 537(7620): 320–327, September 2016a. ISSN 1476-4687. doi: 10.1038/nature19946. URL https://www.nature.com/articles/nature19946. Number: 7620 Publisher: Nature Publishing Group.

Huang, Po-Ssu et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chemical Biology*, 12(1):29–34, January 2016b. ISSN 1552-4469. doi: 10.1038/nchembio.1966. URL https://www.nature.com/articles/nchembio.1966. Number: 1 Publisher: Nature Publishing Group.

Hwang, Yunha et al. Genomic language model predicts protein co-regulation and function. *Nature Communications*, 15(1):2880, April 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46947-9. URL https://www.nature.com/articles/s41467-024-46947-9. Publisher: Nature Publishing Group.

Islam, S M Ashiqul et al. Protein classification using modified n-grams and skip-grams. *Bioinformatics*, 34(9):1481–1487, May 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx823. URL https://doi.org/10.1093/bioinformatics/btx823.

Johnson, Sean R. et al. Generating novel protein sequences using Gibbs sampling of masked language models, January 2021. URL https://www.biorxiv.org/content/10.1101/2021.01.26.428322v1. Pages: 2021.01.26.428322 Section: New Results.

Jumper, John et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Kalavasis, Alkis. On the Limits of Language Generation: Trade-Offs between Hallucination and Mode-Collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, pages 1732–1743, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 979-8-4007-1510-5. doi: 10.1145/3717823.3718108. URL https://dl.acm.org/doi/10.1145/3717823.3718108.

Kamtekar, Satwik et al. Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids. *Science*, 262(5140):1680–1685, December 1993. doi: 10.1126/science.8259512. URL https://www.science.org/doi/10.1126/science.8259512. Publisher: American Association for the Advancement of Science.

Keefe, Anthony D. and Szostak, Jack W. Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718, April 2001. ISSN 1476-4687. doi: 10.1038/35070613. URL https://www.nature.com/articles/35070613. Number: 6829 Publisher: Nature Publishing Group.

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

Kim, Rachel Seongeun et al. BFVD—a large repository of predicted viral protein structures. *Nucleic Acids Research*, 53(D1):D340–D347, January 2025a. ISSN 1362-4962. doi: 10.1093/nar/gkae1119. URL https://doi.org/10.1093/nar/gkae1119.

Kim, Woosub et al. Rapid and sensitive protein complex alignment with Foldseek-Multimer. *Nature Methods*, 22(3):469–472, March 2025b. ISSN 1548-7105. doi: 10.1038/s41592-025-02593-7. URL https://www.nature.com/articles/s41592-025-02593-7. Publisher: Nature Publishing Group.

Kleinberg, Jon and Mullainathan, Sendhil. Language Generation in the Limit. *Advances in Neural Information Processing Systems*, 37:66058–66079, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7988e9b3876ad689e921ce05d711442f-Abstract-Conference.html.

Kuhlman, Brian et al. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649):1364–1368, November 2003. doi: 10.1126/science.1089427. URL https://www.science.org/doi/full/10.1126/science.1089427. Publisher: American Association for the Advancement of Science.

Kurochkina, Natalya and Guha, Udayan. SH3 domains: modules of protein–protein interactions. *Biophysical Reviews*, 5(1):29–39, March 2013. ISSN 1867-2469. doi: 10.1007/s12551-012-0081-z. URL https://doi.org/10.1007/s12551-012-0081-z.

Lau, Andy M. et al. Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science*, 386(6721):eadq4946, November 2024. doi: 10.1126/science.adq4946. URL https://www.science.org/doi/10.1126/science.adq4946. Publisher: American Association for the Advancement of Science.

Li, Hao et al. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science*, 273(5275):666–669, August 1996. doi: 10.1126/science.273.5275.666. URL https://www.science.org/doi/abs/10.1126/science.273.5275.666. Publisher: American Association for the Advancement of Science.

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

Linsky, Thomas W. et al. Sampling of structure and sequence space of small protein folds. *Nature Communications*, 13(1):7151, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34937-8. URL https://www.nature.com/articles/s41467-022-34937-8. Number: 1 Publisher: Nature Publishing Group.

Lockless, Steve W. and Ranganathan, Rama. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286(5438):295–299, October 1999. doi: 10.1126/science.286.5438.295. URL https://www.science.org/doi/10.1126/science.286.5438.295. Publisher: American Association for the Advancement of Science.

Longo, Liam M. et al. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proceedings of the National Academy of Sciences*, 117(27):15731–15739, July 2020a. doi: 10.1073/pnas.2001989117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2001989117. Publisher: Proceedings of the National Academy of Sciences.

Longo, Liam M et al. On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *eLife*, 9:e64415, December 2020b. ISSN 2050-084X. doi: 10.7554/eLife.64415. URL https://doi.org/10.7554/eLife.64415. Publisher: eLife Sciences Publications, Ltd.

Lourenco, Alec Luiz et al. Protein CREATE enables closed-loop design of de novo synthetic protein binders, January 2025. URL https://www.biorxiv.org/content/10.1101/2024.12.20.629847v2. Pages: 2024.12.20.629847 Section: New Results.

Madani, Ali et al. Deep neural language modeling enables functional protein generation across families. Technical report, bioRxiv, July 2021. URL https://www.biorxiv.org/content/10.1101/2021.07.18.452833v1. Section: New Results Type: article.

Madani, Ali et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, January 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2. Publisher: Nature Publishing Group.

Maloney, Erin K. et al. An Anti-Insulin-like Growth Factor I Receptor Antibody That Is a Potent Inhibitor of Cancer Cell Proliferation. *Cancer Research*, 63(16): 5073–5083, August 2003. ISSN 0008-5472.

Martinez, Zachary A. TRILL: Orchestrating Modular Deep-Learning Workflows for Democratized, Scalable Protein Analysis and Engineering, October 2023. URL https://www.biorxiv.org/content/10.1101/2023.10.24.563881v1. Pages: 2023.10.24.563881 Section: New Results.

Mayer, Bruce J. SH3 domains: complexity in moderation. *Journal of Cell Science*, 114(7):1253–1263, April 2001. ISSN 0021-9533. doi: 10.1242/jcs.114.7.1253. URL https://doi.org/10.1242/jcs.114.7.1253.

Mayer, John P. Insulin structure and function. *Peptide Science*, 88(5):687–713, 2007. ISSN 1097-0282. doi: 10.1002/bip.20734. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.20734. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.20734.

Maynard Smith, John. Natural Selection and the Concept of a Protein Space. *Nature*, 225(5232):563–564, February 1970. ISSN 1476-4687. doi: 10.1038/225563a0. URL https://www.nature.com/articles/225563a0. Publisher: Nature Publishing Group.

McInnes, Leland. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, February 2018. URL https://arxiv.org/abs/1802.03426v3.

Min, Cheol-Ki et al. Increased expression, folding and enzyme reaction rate of recombinant human insulin by selecting appropriate leader peptide. *Journal of Biotechnology*, 151(4):350–356, February 2011. ISSN 01681656. doi: 10.1016/j.jbiotec.2010.12.023. URL https://linkinghub.elsevier.com/retrieve/pii/S0168165611000265.

Minami, Shintaro et al. Exploration of novel alpha/beta-protein folds through de novo design. *Nature Structural & Molecular Biology*, pages 1–9, July 2023. ISSN 1545-9985. doi: 10.1038/s41594-023-01029-0. URL https://www.nature.com/articles/s41594-023-01029-0. Publisher: Nature Publishing Group.

Mirdita, Milot et al. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://www.nature.com/articles/s41592-022-01488-1. Publisher: Nature Publishing Group.

Munsamy, Geraldene et al. ZymCTRL: a conditional language model for the controllable generation of artificial enzymes. *NeurIPS Machine Learning for Structural Biology Workshop*, December 2022. URL https://www.mlsb.io/papers_2022/ZymCTRL_a_conditional_language_model_for_the_controllable_generation_of_artificial_enzymes.pdf.

Nguyen, Eric et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, November 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/full/10.1126/science.ado9336. Publisher: American Association for the Advancement of Science.

Pabo, Carl. Molecular technology: Designing proteins and peptides. *Nature*, 301 (5897):200–200, January 1983. ISSN 1476-4687. doi: 10.1038/301200a0. URL https://www.nature.com/articles/301200a0. Publisher: Nature Publishing Group.

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

Pavlopoulos, Georgios A. et al. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983):594–602, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06583-7. URL https://www.nature.com/articles/s41586-023-06583-7. Number: 7983 Publisher: Nature Publishing Group.

Pudžiuvelytė, Ieva et al. TemStaPro: protein thermostability prediction using sequence representations from protein language models. *Bioinformatics*, 40(4): btae157, April 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae157. URL https://doi.org/10.1093/bioinformatics/btae157.

Qiang, Hantao et al. Language model-guided anticipation and discovery of unknown metabolites, November 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.13.623458v1. Pages: 2024.11.13.623458 Section: New Results.

Radford, Alec et al. Language Models are Unsupervised Multitask Learners. 2019. URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Riddle, David S. et al. Functional rapidly folding proteins from simplified amino acid sequences. *Nature Structural Biology*, 4(10):805–809, October 1997. ISSN 1545-9985. doi: 10.1038/nsb1097-805. URL https://www.nature.com/articles/nsb1097-805. Publisher: Nature Publishing Group.

Rivoire, Olivier. Evolution-Based Functional Decomposition of Proteins. *PLOS Computational Biology*, 12(6):e1004817, June 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004817. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004817. Publisher: Public Library of Science.

Rocklin, Gabriel J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, July 2017. doi: 10.1126/science.aan0693. URL https://www.science.org/doi/full/10.1126/science.aan0693. Publisher: American Association for the Advancement of Science.

Romero, Philip A. and Arnold, Frances H. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, December 2009. ISSN 1471-0080. doi: 10.1038/nrm2805. URL `https://www.nature.com/articles/nrm2805`. Publisher: Nature Publishing Group.

Roney, James P. and Ovchinnikov, Sergey. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Physical Review Letters*, 129(23):238101, November 2022. doi: 10.1103/PhysRevLett.129.238101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.129.238101`. Publisher: American Physical Society.

Sakuma, Koya et al. Design of complicated all-alpha protein structures. *Nature Structural & Molecular Biology*, pages 1–8, January 2024. ISSN 1545-9985. doi: 10.1038/s41594-023-01147-9. URL `https://www.nature.com/articles/s41594-023-01147-9`. Publisher: Nature Publishing Group.

Schreiber, Gideon and Fersht, Alan R. Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles. *Journal of Molecular Biology*, 248(2):478–486, January 1995. ISSN 0022-2836. doi: 10.1016/S0022-2836(95)80064-6. URL `https://www.sciencedirect.com/science/article/pii/S0022283695800646`.

Schreiber, Gideon. Stability and function: two constraints in the evolution of barstar and other proteins. *Structure*, 2(10):945–951, October 1994. ISSN 0969-2126. doi: 10.1016/S0969-2126(94)00096-4. URL `https://www.cell.com/structure/abstract/S0969-2126(94)00096-4`. Publisher: Elsevier.

Shumailov, Ilia et al. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y. URL `https://www.nature.com/articles/s41586-024-07566-y`. Publisher: Nature Publishing Group.

Simon, Elana and Zou, James. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders, January 2025. URL `https://www.biorxiv.org/content/10.1101/2024.11.14.623630v2`. Pages: 2024.11.14.623630 Section: New Results.

Skinnider, Michael A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nature Machine Intelligence*, 6(4):437–448, April 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00821-x. URL `https://www.nature.com/articles/s42256-024-00821-x`. Publisher: Nature Publishing Group.

Skinnider, Michael A. et al. Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence*, 3(9):759–770, September 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00368-1. URL `https://www.nature.com/articles/s42256-021-00368-1`. Publisher: Nature Publishing Group.

Skolnick, Jeffrey. Further Evidence for the Likely Completeness of the Library of Solved Single Domain Protein Structures. *The Journal of Physical Chemistry B*, 116(23):6654–6664, June 2012. ISSN 1520-6106. doi: 10.1021/jp211052j. URL https://doi.org/10.1021/jp211052j. Publisher: American Chemical Society.

Socolich, Michael et al. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, September 2005. ISSN 1476-4687. doi: 10.1038/nature03991. URL https://www.nature.com/articles/nature03991. Publisher: Nature Publishing Group.

Suzek, Baris E. et al. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)*, 23(10):1282–1288, May 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm098.

Swadesh, Morris. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21(2):121–137, April 1955. ISSN 0020-7071. doi: 10.1086/464321. URL https://www.journals.uchicago.edu/doi/abs/10.1086/464321. Publisher: The University of Chicago Press.

Süel, Gürol M. et al. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1):59–69, January 2003. ISSN 1545-9985. doi: 10.1038/nsb881. URL https://www.nature.com/articles/nsb881. Number: 1 Publisher: Nature Publishing Group.

Taverna, Darin M. and Goldstein, Richard A. Why are proteins marginally stable? *Proteins: Structure, Function, and Bioinformatics*, 46(1):105–109, 2002. ISSN 1097-0134. doi: 10.1002/prot.10016. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.10016.

Taylor, William R. et al. Probing the "Dark Matter" of Protein Fold Space. *Structure*, 17(9):1244–1252, September 2009. ISSN 0969-2126. doi: 10.1016/j.str.2009.07.012. URL https://www.cell.com/structure/abstract/S0969-2126(09)00295-0. Publisher: Elsevier.

Tong, Cher Ling. *De novo* proteins from random sequences through *in vitro* evolution. *Current Opinion in Structural Biology*, 68:129–134, June 2021. ISSN 0959-440X. doi: 10.1016/j.sbi.2020.12.014. URL https://www.sciencedirect.com/science/article/pii/S0959440X21000026.

Tsuboyama, Kotaro et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, pages 1–11, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06328-6. URL https://www.nature.com/articles/s41586-023-06328-6. Publisher: Nature Publishing Group.

Tóth-Petróczy, Ágnes and Tawfik, Dan S. The robustness and innovability of protein folds. *Current Opinion in Structural Biology*, 26:131–138, June 2014. ISSN 0959-440X. doi: 10.1016/j.sbi.2014.06.007. URL `https://www.sciencedirect.com/science/article/pii/S0959440X14000724`.

van Kempen, Michel et al. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, pages 1–4, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL `https://www.nature.com/articles/s41587-023-01773-0`. Publisher: Nature Publishing Group.

Varadi, Mihaly et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL `https://doi.org/10.1093/nar/gkab1061`.

Vaswani, Ashish et al. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

Verkuil, Robert et al. Language models generalize beyond natural proteins. Technical report, bioRxiv, December 2022. URL `https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1`. Section: New Results Type: article.

Vig, Jesse et al. BERTology Meets Biology: Interpreting Attention in Protein Language Models, March 2021. URL `http://arxiv.org/abs/2006.15222`. arXiv:2006.15222 [cs].

Vyas, Pratik et al. Helicase-like functions in phosphate loop containing beta-alpha polypeptides. *Proceedings of the National Academy of Sciences*, 118 (16):e2016131118, April 2021. doi: 10.1073/pnas.2016131118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2016131118`. Publisher: Proceedings of the National Academy of Sciences.

Watson, Joseph L. et al. De novo design of protein structure and function with RFdiffusion. *Nature*, pages 1–3, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL `https://www.nature.com/articles/s41586-023-06415-8`. Publisher: Nature Publishing Group.

Watters, Alexander L. et al. The Highly Cooperative Folding of Small Naturally Occurring Proteins Is Likely the Result of Natural Selection. *Cell*, 128(3):613–624, February 2007. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2006.12.042. URL `https://www.cell.com/cell/abstract/S0092-8674(07)00117-1`. Publisher: Elsevier.

Wilson, Amanda E. Evolutionary Processes and Biophysical Mechanisms: Revisiting Why Evolved Proteins Are Marginally Stable. *Journal of Molecular Evolution*, 88(5):415–417, July 2020. ISSN 1432-1432. doi: 10.1007/s00239-020-09948-y. URL `https://doi.org/10.1007/s00239-020-09948-y`.

Yekutieli, Daniel and Benjamini, Yoav. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1):171–196, December 1999. ISSN 0378-3758. doi: 10.1016/S0378-3758(99)00041-5. URL https://www.sciencedirect.com/science/article/pii/S0378375899000415.

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

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

Zhang, Yang et al. On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences*, 103(8):2605–2610, February 2006. doi: 10.1073/pnas.0509379103. URL https://www.pnas.org/doi/abs/10.1073/pnas.0509379103. Publisher: Proceedings of the National Academy of Sciences.

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