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 ~ 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.