

MScCBBi

MASTER IN
**COMPUTATIONAL BIOLOGY
& BIOINFORMATICS**

Rodrigo António Correia Tavares Bernardino
BSC in Molecular Biology and Genetics

EXPERIMENTS IN THE USE OF $\Delta\Delta G$ PREDICTIONS FOR DE- NOVO IN-SILICO PROTEIN DESIGN

Nov, 2023



EXPERIMENTS IN THE USE OF $\Delta\Delta G$ PREDICTIONS FOR DE-NOVO IN-SILICO PROTEIN DESIGN

RODRIGO ANTÓNIO CORREIA TAVARES BERNARDINO

BSc in Molecular Biology and Genetics

Adviser: Leonardo Vanneschi

Leading Researcher, IMS

Co-adviser: Isabel Rocha

Leading Researcher, ITQB

Examination Committee

Chair: Paula Goncalves

Rapporteur: Diana Lousa

Adviser:

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon

month, year

Experiments in the use of $\Delta\Delta G$ predictions for de-novo in-silico protein design

Copyright © Rodrigo António Correia Tavares Bernardino, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

ACKNOWLEDGEMENTS

This work would not have been possible without the fantastic open source projects of OmegaFold, Pythia and Map-Elites. We also thank Meta and Alex Rives for open sourcing ESM and ESMFold and Steinegger lab for the web access to FoldSeek. I also thank my family and supervisors for their overwhelming support and patience.

ABSTRACT

The biosphere solution to the protein design problem is akin to a planetwide computational machine running a simple stochastic algorithm. The root of its success lies in the open endedness and diversity of its search, willing to extend across all possible niches over time. Nature’s work produces living transient DNA artifacts that are stored in the biosphere, and, at an increasing rate, being read by sequencing technology and transferred to large datasets now hosting millions of raw samples. This data can be leveraged by the use of increasingly large pLM(protein Language models), that have been shown to essentially function as unsupervised protein structure learners. If the structure learned is sufficiently accurate it should by Anfinsen’s dogma reveal the protein’s function and theoretically allow for end-to-end sequence to function pipelines. This in-silico study uses a novel approach by synergistically integrating two pLM models: OmegaFold for single-sequence protein structure prediction, and Pythia for zero-shot mutation effect prediction values in $\Delta\Delta G$. $\Delta\Delta G$ is a measure of the change in energy between the folded and unfolded states $\Delta G_{\text{folding}}$ and the change in $\Delta G_{\text{folding}}$ when a single mutation is present, a determinant factor in protein stability. We employ these models within an evolutionary algorithm framework, creating an evolutionary optimization loop for de-novo protein design with the main focus of evaluating various mutator biases based on $\Delta\Delta G$ values and their efficacy in generating high-confidence secondary structure motifs consisting of 100 amino acids. Our results show that $\Delta\Delta G$ is a sensitive metric for protein stability design and is deeply correlated with backbone secondary structure.

Keywords: $\Delta\Delta G$, OmegaFold, Pythia, pLM

RESUMO

"A solução da biosfera para o problema do design de proteínas é semelhante a uma máquina computacional em escala planetária, executando um algoritmo estocástico simples. A raiz do seu sucesso reside na expansão e diversidade da sua busca, disposta a estender-se por todos os nichos possíveis ao longo do tempo. Este trabalho da natureza produz artefatos de DNA transitórios vivos que são armazenados na biosfera e, a uma taxa crescente, sendo lidos por novas tecnologia de sequenciamento e transferidos para enormes conjuntos de dados que agora hospedam milhões de amostras brutas. Esses dados podem ser aproveitados pelo uso de pLMs (modelos de Linguagem de proteínas) cada vez maiores, que foram demonstrados como essencialmente modelos de "unsupervised learning" de estrutura de proteínas. Se a estrutura aprendida for suficientemente precisa, ela deve revelar a função da proteína pelo dogma de Anfinsen e, teoricamente, permitir "end to end" pipelines de sequência para função. Este estudo in-silico utiliza uma abordagem inovadora, integrando sinergicamente dois protein Language models: Omegafold para previsão de estrutura de proteína de sequência única, e Pythia para previsão de efeitos de mutação zero-shot. Empregamos esses modelos dentro de um framework de algoritmo evolucionário, criando um ciclo de otimização evolucionária para o design de-novo de proteínas. O nosso foco está em avaliar vários tipos de mutadores com base em valores de $\Delta\Delta G$ e sua eficácia na geração de motivos de estrutura secundária de alta confiança consistindo de 100 aminoácidos. Os nossos resultados mostram que $\Delta\Delta G$ é uma métrica útil para o design de proteínas estáveis e está intimamente correlacionado com estrutura secundária. "

Palavras-chave: $\Delta\Delta G$, Omegafold, Pythia, pLm

CONTENTS

List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Sequence to Function	2
1.1.1 Protein folding	3
1.1.2 Proteostasis Network	4
1.1.3 Protein stability	5
1.1.4 Stability design	5
1.1.5 Thermodynamics formulation	6
1.1.6 TM-Score in Protein Modeling	6
1.2 Novelty Search	6
1.2.1 Deception	7
1.2.2 Quality-Diversity	8
1.3 Omegafold - Protein language model specialized in orphan proteins . .	8
1.4 Pythia - Probabilistic Approach Based on Gibbs Free Energy	8
1.4.1 Mathematical Formulation	9
1.4.2 Assumptions	9
2 Aims	10
2.1 Globular proteins	10
2.2 Alpha Helyx and Beta Sheet	11
2.3 Structural stability clues given by chaperon dependent homologs	11
3 Methods	13
3.1 Structure	13
3.2 Initialisation	14
3.2.1 Minimum Dependencies	14
3.3 Omegafold	15

3.4	Prediction and annotation of structural features	15
3.5	Diversity Metric	15
3.6	K-Points Crossover	15
3.7	$\Delta\Delta G$ Mutators	15
3.7.1	Weighted Random	16
3.7.2	Positive Exponential	16
3.7.3	Inverse Exponential	16
3.7.4	Gaussian Centered at 0	16
4	Results	17
4.1	Confidence and structure	17
4.2	Mutational search space	18
4.3	Evaluation of mutation strategies	18
4.4	The problem of objectives	19
4.5	Foldseek Homology Analysis	20
5	Discussion and Future Work	22
5.1	Epistatic Interactions	22
5.2	Binding sites and Functional Activity	22
	Bibliography	24
	Annexes	

LIST OF FIGURES

2.1	Pythia $\Delta\Delta G$ frequency bar plot overlap of chaperon dependent (protein a, e.coli dihydrofolate reductase (DHFR)) and an independent homolog (protein b, mouse DHFR).	11
2.2	evolved β -barrel, pymol [39]	12
3.1	A meta overview.	13
4.1	pLDDT corresponds to the model's prediction of its score on the local Distance Difference Test. It is a measure of local accuracy. Loop is $1-(\alpha+\beta)$. Every point is a final elite protein, the colours indicate what experiment created that protein.	17
4.2	Alpha-Helix and Beta-Sheet percentage versus model confidence	18
4.3	Pythia $\Delta\Delta G$ Maps at Different Stages of Evolutionary Process	18
4.4	The index represents the number of protein structures analysed. The turnover of elite solutions is used as the metric for evolutionary speed.	19
4.5	Average $\Delta\Delta G$ map of all Mutation Strategies	20
4.6	pLDDT objective function optimization.	21
4.7	ESM2 reconstitution of two evolved elite proteins, blue >90 pLDDT.	21

LIST OF TABLES

3.1	Average aminoacid composition percentages.	14
3.2	Dependencies for Computational Biology Pipeline	14

INTRODUCTION

The protein design problem is a complex and multifaceted challenge that resides at the intersection of biology, chemistry, and computer science, even for a small protein of 100 aminoacids the choices 20^{100} outweigh the number of atoms in the universe [2]. The incomprehensibly vast size of the hyperdimensional sequence space puts it in the category of NP-hard problems[3] such as the TSP (Traveling Salesman Problem). However in practice while the aminoacid alphabet is a perfect way to convey the space of all the possible protein structures it doesn't have a structure-centric design and so most of the structures it describes fail to fold due to steric hindrance, vastly exaggerating the possible design space[3].

The current paradigm of Artificial Intelligence technology is advancing extremely fast across multiple disciplines and its profound influence continues to expand [4]. The widespread use of deep unsupervised transformer architecture in biology started with structure modelling and AlphaFold[5]. At the time of AlphaFold's release, approximately 100,000 unique protein structures had been determined through extensive experimental efforts [5]. As of the latest information, the AlphaFold Protein Structure Database contains over 214 million entries. Yet it is still dwarfed by the Evolutionary Scale Modeling (ESM2)[6] applied on the MGnify dataset. The resulting metagenomic atlas now encompasses over 772 million high confidence predicted protein structures. This extensive collection marks a significant expansion, approximately 1000 times larger than the initial size at the time of AlphaFold release in 2021.[7]

Furthermore, the multi-dimensional use of these large scale evolved models [8] only appears to increase, using a technique called diffusion modelling one can turn these models into de-novo generators also known as modelled hallucinations [9]. It works by iteratively learning a function that turns signal to noise, and using its temporal invariance to create de-novo, signal grounded structures from noise [10] An example of this is RFdiffusion, delivering state of the art results in de-novo backbone design by fine tuning the original structure prediction model, RoseTTAFold, on protein structure denoising tasks [11]. Fine tuning these large foundational models [8] can modify their end applications, for example for enzyme classification(EC) tasks.

The computational landscape has witnessed a paradigm shift from traditional supervised regression frameworks to Self Supervised Learning (SSL), especially with the advent of masked language modeling (MLM). Masked language modeling is a form of SSL where certain input tokens (like words in a sentence or amino acids in a protein sequence) are masked out, and the model is trained to predict these masked tokens based on their context. This method enables the model to learn rich representations from vast amounts of unlabeled data. Within the realm of biology, pre-trained protein language models (pLM's) have showcased their prowess by extracting biologically relevant representations. Their applications span from protein contact prediction and function prediction to sophisticated tasks such as zero-shot mutation effect prediction, forecasting viral evolution patterns [12] and creating a database of predictions for all possible single amino acid substitutions in the human proteome.[13] These models have now been used to successfully take over the SOTA (state of the art) results in many active areas of protein research. One example is *FoldSeek*, the only method of structural comparison fast enough to handle the current Metagenomic scaled datasets (772 million proteins).

The scale of these models has increased substantially over time, another testament to their increasing use. A study titled “Scaling Up Generative Protein Sequence Models” [8] investigated scaling the models from 8.5×10^7 to 1.2×10^9 parameters, presenting the weights of the models as a proposed benchmark for the effects of scaling. One year later, the current state-of-the-art open-sourced model by Meta, *ESM2*, has 1.5×10^{10} parameters.

Despite all the advances, due to the nature of the technology most of the progress is empirical and there is a wide ongoing effort to understand and characterize these large evolved systems [14]. Protein Language Models are in the forefront of interpretability studies. PLM's have been shown to be unsupervised structure learners, [15] using techniques such as ablation studies and scaling studies, these models have shown to be robust and generalizable and data indicates that they capture the folding structure of proteins, are able to target binding sites, and are able to model increasingly complex biophysical properties with layer depth.[16] [17]

To better understand the scientific advances and obstacles facing in-silico de-novo protein design we need to advance our understanding of them and create new benchmarking across the whole A.I. field.[18]

1.1 Sequence to Function

Structural prediction in molecular biology is grounded in the central dogma, which elucidates the flow of genetic information from DNA to RNA, and ultimately to protein. This process involves the transcription of DNA into messenger RNA (mRNA), which is then translated by ribosomes into proteins. Ribosomes, complex molecular machines comprising ribosomal RNA and proteins, are pivotal in this translation process.

The sequence of these amino acids, determined by the genetic code, directly determines the protein's three-dimensional structure. This structure, in turn, is critical in defining

the protein's function. For example, fibrous proteins like keratin, which forms structural components such as hair, nails, and horns, showcase how protein architecture underpins biological function. Keratin's fibrous nature contributes to the mechanical strength and durability of these structures.

The seminal work of Anfinsen (1973) highlighted that the three-dimensional structure of a protein is intrinsically determined by its amino acid sequence. This principle drives the field's ultimate goal: to predict the three-dimensional structure of proteins from their sequences, thereby inferring their biological functions.

1.1.1 Protein folding

Protein folding is a physical process by which a polypeptide chain folds into its characteristic and functional three-dimensional structure from a random coil. Each protein begins as a linear sequence of amino acids, which then undergoes a transition to a structured state, or native conformation, that is determined by its primary sequence. This folding process is essential for the protein's biological function, as the three-dimensional form dictates how it interacts with other molecules.[19]

Protein stability is fundamentally a thermodynamic concept referring to the propensity of a protein to maintain its structure under various conditions. The stability is determined by the difference in Gibbs free energy between the folded and unfolded states of a protein. A stable protein will have a significantly lower free energy in its folded state, making this state favorable and more populated under physiological conditions. The balance of forces contributing to protein stability includes hydrophobic interactions, hydrogen bonds, Van der Waals forces, and ionic interactions.[20]

Environmental factors such as temperature, pH, and salt concentration can affect these forces, potentially leading to protein denaturation.

Considering the critical role of proteins in biological systems and the fine balance of stability required for their function, it is clear that the precise calibration of protein folding is not merely a biological curiosity, but a vital determinant of health and disease.[21]

Natural proteins are often only marginally stable, with free-energy differences between the native state and the unfolded or misfolded states as low as 5 kcal/mol.[22]

The modest free-energy differential between the native state and less structured forms underpins a protein's robustness to environmental and genetic perturbations. This narrow margin of stability is a key factor in the evolution of proteins, allowing for adaptability while also imposing constraints due to the risk of destabilization. [21]

Diseases, notably, can arise not only from a protein's loss of function but from a reduction in its stability, which diminishes its functional concentration in the cell. This connection is exemplified by certain mutations in the protein p53, where a decrease in stability leads to a significant reduction in the correctly folded, functional protein, contributing to cancer pathogenesis. Furthermore, the tendency of unstable proteins to misfold can trigger the formation of toxic aggregates, implicating protein stability as

a major player in the etiology of numerous pathologies, including neurodegenerative diseases.

1.1.2 Proteostasis Network

As always in biology, nothing exists in a vacuum, and everything is more complex than it first appears. The proteostasis network (PN), is a network of components that collectively ensure the stability and proper folding of proteins within the cell. These components include molecular chaperones that assist in folding, proteasomes that degrade misfolded proteins, autophagy systems that remove larger protein aggregates, and the unfolded protein response which regulates these components. Together, they form a dynamic network that responds to cellular needs and stress signals. The PN is critical not only for the functional maintenance of proteins under normal conditions but also for their efficient folding and stabilization when cells face environmental stressors, such as heat shock or nutrient deprivation.

Experiments on *Escherichia coli* (*E. coli*) have shed light on the proteostasis network's capacity to manage protein folding, with particular emphasis on the roles of chaperones. In these studies, the folding fate of a protein is shown to be an interplay between its own folding energy landscape and the actions of the proteostasis network, including molecular chaperones and degradation enzymes.

For instance, a computational model called FoldEco synthesized experimental data to provide a system-wide description of the proteostasis network in *E. coli*. This model allows for the determination of the probability that a given chaperone system mediates transitions between client protein conformations, offering insights into how different chaperone systems can compensate for each other when one is disabled[23].

In another set of experiments, the folding outcomes of three test proteins biosynthesized in *E. coli* were quantitatively analyzed under various conditions. The results showed that overexpression of the entire proteostasis network benefitted all test proteins, but the effects varied when individual chaperones or the major degradation enzyme Lon were upregulated. This highlighted that the impact of the *E. coli* proteostasis network on protein folding is a collective outcome of the concerted action by multiple systems, including the Hsp70 system (DnaK/DnaJ/GrpE), the Hsp60 system (GroEL/GroES), and Lon.[24]

Further research demonstrated that some protein misfolding intermediates can bypass the chaperone network and remain soluble but function less effectively. This has been observed in the case of Luciferase in *E. coli*, where misfolded states were shown to evade chaperones, not aggregate, resist degradation, and have structurally perturbed residues crucial for function. The likelihood of these misfolded states to interact with the chaperone systems TF, GroEL/GroES, or DnaK was analyzed, revealing that many of these misfolded forms could bypass the *E. coli* chaperone network. This points to the existence of chaperone-independent folding pathways and underscores the complexity of

the proteostasis network.[25]

1.1.3 Protein stability

Proteins are marginally stable due to a balance of energetic factors and functional necessity. The vast number of unfolded and misfolded states compared to the singular folded state imposes a significant entropic cost, which must be offset by the formation of numerous stabilizing molecular contacts within the native structure. Marginal stability allows for the efficient regulation of protein abundance, as proteins that are too stable may hinder cellular turnover and clearance mechanisms. Evolutionarily, proteins have achieved a level of stability that ensures sufficient expression and functionality without incurring additional selection pressure for increased stability, reflecting an equilibrium between genetic drift and functional adequacy.

Certain molecular functions, such as enzyme catalysis or ligand binding, inherently require structural compromises within a protein. Active sites in enzymes often contain clusters of charges critical for catalysis that can destabilize the protein's overall structure. Similarly, binding sites may display hydrophobic patches vital for interaction yet detrimental to stability. Proteins may be larger than necessary for their active functions to offset destabilizing features with stabilizing energy from the rest of the structure. Additionally, some proteins need to fold and unfold, such as those involved in transport or mechanical sensing, where high stability could impede their functional flexibility.

Folding carries a tremendous entropic penalty, estimated to be on the order of 70 kcal/mol for a 100-residue protein [26].

The energy cost of protein folding underscores the significant entropic penalty of assuming a single, ordered structure over a multitude of disordered ones. This cost must be counteracted by stabilizing factors within the protein's structure, notably hydrogen bonds. While an individual backbone hydrogen bond may contribute marginally to the protein's stability, the absence of such a bond would be highly destabilizing, potentially offsetting much of the protein's folding stability, which is generally between -5 and -15 kcal/mol under physiological conditions. Therefore, a network of hydrogen bonds is essential to achieve the net stability of the folded protein[27].

1.1.4 Stability design

Natural proteins frequently exhibit marginal stability, with minimal free-energy differences between their native and misfolded forms. This marginality can make proteins susceptible to environmental changes or mutations, influencing protein evolution, disease causation, and bioengineering applications. Many disease-causing mutations don't necessarily impact the protein's function but reduce its stability, resulting in inadequate functional protein amounts. Some mutations even cause misfolding, leading to harmful aggregation.[28]

Protein stability, a thermodynamic property, and expressibility, governed by folding trajectory, are intrinsically linked through the free-energy difference between folded and misfolded states. Misfolded protein conformations can act as traps, hindering the correct folding path and potentially leading to aggregation, affecting expressibility. A protein's energy landscape may contain many such misfolded states near the native state's energy level, especially in marginally stable proteins. Stability design aims to widen the energy difference between native and misfolded states, either by reducing the native state's energy or by minimizing high-energy misfolded states. Achieving this not only enhances protein's thermal stability but also its expressibility. [28]

1.1.5 Thermodynamics formulation

The overall free energy of protein stabilization, is the net energy difference between the folded and unfolded states, accounting for both enthalpic (energy from bonds and interactions) and entropic (energy from disorder) contributions. It's defined by the Gibbs free energy equation: $\Delta G = \Delta H - T\Delta S$ where ΔG is the change in free energy, ΔH is the change in enthalpy, T is the temperature, and ΔS is the change in entropy.[29] A stable protein has a negative ΔG , indicating that the folded state is energetically favored over the unfolded state. The overall free energy of stabilization is typically lower (between -5 and -15 kcal/mol for proteins under physiological conditions) than the entropic cost of folding because the enthalpic gains from forming intramolecular interactions like hydrogen bonds, ionic interactions, and van der Waals forces outweigh the entropic cost.

1.1.6 TM-Score in Protein Modeling

The Template Modeling score (TM-score) is a significant metric in bioinformatics, particularly in the domain of protein structure modeling.[30] It serves as a measure of similarity between two protein structures, offering a more nuanced assessment than traditional measures like the root-mean-square deviation (RMSD). The TM-score addresses two core issues prevalent in older metrics such as RMSD.[31] Firstly, it emphasizes smaller distance errors more significantly, making the score more sensitive to global fold similarity rather than local structural variations. [32] Secondly, it incorporates a length-dependent scale, enhancing its accuracy in evaluating the quality of protein structure templates and full-length models. This scale is adaptive to the size of the protein, ensuring a fair comparison across structures of varying sizes.[33] The TM-score ranges from 0 to 1, where a score of 1 indicates a perfect match and a score of ≥ 0.5 suggests that two structures share the same fold.

1.2 Novelty Search

In evolutionary computation one can think of a search space as an expansive, multidimensional landscape dotted with peaks and valleys, each representing a potential solution. In

this terrain, the altitude at any point signifies the solution's fitness or quality. An evolutionary algorithm traversing this landscape should not merely seek the highest peak—akin to the most optimized solution—but aim instead with exploring the breadth of the terrain to uncover a range of high-quality solutions. This exploration parallels a hiker who aims to discover a variety of landscapes, each with distinct characteristics, rather than just ascending to the summit of the tallest mountain[34].

The fundamental strategies of problem-solving in optimization algorithms are encapsulated by two concepts: exploration and exploitation. Exploitation is the strategy of delving deeper into known areas of the search space to optimize solutions, akin to perfecting familiar knowledge. Conversely, exploration is the quest for new information, a strategy that seeks to expand the breadth of the search space and uncover novel solutions. Both strategies are essential and complementary: exploitation leverages existing knowledge to refine solutions, while exploration broadens the knowledge base, potentially revealing innovative solutions that exploitation can then fine-tune. The balance between these strategies is a cornerstone of effective problem-solving across a variety of fields, each one critical to the discovery and optimization of solutions. This general principle underlies many algorithmic approaches, from evolutionary strategies to reinforcement learning, reflecting the universal importance of balancing the known with the unknown in the pursuit of problem resolution.[35].

1.2.1 Deception

Deception in optimization problems refers to situations where evolutionary strategies that prioritize local improvement lead away from the global optimum—essentially, when the path of steepest ascent does not lead to the highest peak. Deceptive problems are those in which the fitness landscape is structured such that the evolutionary pressures guide the algorithms toward local optima that are suboptimal compared to the global solution. In these deceptive landscapes, the local cues that an algorithm uses to determine the direction of search can be misleading. These problems are particularly challenging because they require a more sophisticated approach that can discern when to abandon a seemingly promising path in favor of a less intuitive one that might lead to a better overall solution. The concept of deception underscores the importance of maintaining a balance between exploration and exploitation. While exploitation might drive an algorithm into a local optimum from which it cannot escape, exploration can provide the variability needed to bypass deceptive "traps." Thus, designing algorithms capable of navigating deceptive problems is a critical aspect of research in evolutionary computation, as it mirrors the complexities of real-world problem-solving where the most apparent path forward is not always the most advantageous in the long term. [36].

1.2.2 Quality-Diversity

Quality Diversity (QD) algorithms like MAP-Elites, developed by Jean-Baptiste Mouret, operate by categorizing the search space into distinct niches and populating these niches with elite solutions. The technical process involves discretizing the search space into a grid of cells, each representing a combination of feature dimensions that define a niche. Within each cell, the algorithm seeks to find and maintain the highest-performing individual, thus creating a diverse archive of solutions. MAP-Elites iteratively improves this archive by generating offspring from the elite individuals, evaluating them, and placing them into the appropriate cell if they outperform the current inhabitant. This approach ensures that the algorithm maintains a diverse set of high-quality solutions across the entire range of feature dimensions, facilitating the study of how different solution traits contribute to overall performance in the search space.[37]

1.3 Omegafold - Protein language model specialized in orphan proteins

Omegafold is the first computational method to successfully predict high-resolution protein structure without needing multiple sequence alignments. It was shown to outperform in the specific case of orphan proteins and antibody loops. Omegafold was chosen as the language protein model because of its ability to function with single sequences and the desire to explore de-novo protein structure design. OmegaPLM [38] is a protein language model built on a geformer [38] architecture, trained on an extensive dataset of unaligned and unlabeled single protein sequences. This model generates both single-residue and pairwise-residue embeddings that serve as powerful features for capturing the inherent distribution patterns of amino acid sequences. These embeddings are adept at encapsulating both structural and functional facets of the proteins. It also returns a per-residue pLDDT (predicted Local Distance Difference Test) score, serving as a measure of the quality and reliability of the predicted protein structure.

1.4 Pythia - Probabilistic Approach Based on Gibbs Free Energy

At its core, Pythia employs an attention-based message passing neural network (MPNN), a sophisticated framework that enables the processing of complex biological data. The model operates by independently processing the features of nodes (amino acids) and edges (interactions between amino acids) through an embedding layer, effectively transforming biological data into a neural network-friendly format.

Pythia's approach translates the energy differences into probabilities using thermodynamic principles. It leverages the Gibbs free energy to relate energy, entropy, and probability, providing a physical understanding of the system by tying energy states to the probabilities of amino acid conformations.

1.4.1 Mathematical Formulation

The hypothesis suggests that the energy of a protein in its unfolded state remains largely unchanged with mutations due to a lack of stable, specific interactions between the side chains in the unfolded state. This is mathematically represented as:

$$\Delta\Delta G \approx G_{\text{MUT}} - G_{\text{WT}}$$

The total probability of an amino acid is derived from the sum of probabilities of all its rotameric conformations across the entire rotamer space of all 20 amino acids. In terms of energy, this probability is given by:

$$P = e^{-H/k_B T}$$

Where k_B is the Boltzmann constant, T is the temperature, and H is the enthalpy.

The entropy of each amino acid is related to its microstate (conformation) count as:

$$S = k_B \ln W$$

Substituting this expression for S into the equation for P , we obtain:

$$P = e^{-(H-TS)/k_B T}$$

Given the Gibbs free energy relation $G = H - T \times S$, this simplifies to:

$$P = e^{-G/k_B T}$$

The difference in Gibbs free energy due to mutations is related to the probabilities of the mutant and wild type amino acids as:

$$\Delta\Delta G \propto -\ln \left(\frac{P_{\text{MUT}}}{P_{\text{WT}}} \right)$$

1.4.2 Assumptions

1. The energy of a protein in its unfolded state remains almost invariant with mutations, given the lack of stable specific interactions in this state.
2. A direct relationship exists between the microstates (conformations) of an amino acid and its entropy.
3. The energy states can be effectively translated into probabilities of amino acid conformations, representing their likelihood of occurrence.

This project's aim was to create an optimization loop that could produce novel "interesting" biological artifacts indefinitely, only limited by computation.

In that respect it delivers short, for even if it produces novel biological artifacts their progressive novelty almost plateaus as one starts hitting the physical constraints of both the system and hyperparameters. However it serves as an in-silico evolutionary sandbox experiment, by archiving the proteins at checkpoints through the evolutionary walk we can run experiments and analyse both the consistency of the results of the pLM models and inquire about $\Delta\Delta G$ rational design. All the computing work is done by a single GTX1070 in the span of weeks per experiment. All the code is made public on github and all dependencies are open-licensed.

2.1 Globular proteins

Globular proteins, distinguished by their roughly spherical shapes, represent a major class of proteins found abundantly in nature. These proteins showcase a vast array of three-dimensional structures, predominantly comprising α -helices and β -sheets. This intricate folding results in a compact conformation, which is stabilized through a balance of polar and nonpolar interactions. In the globular protein structure, the folding occurs spontaneously and is sustained by the interplay of amino acid side chains.

A key characteristic of these proteins is the spatial arrangement of their amino acid residues. Typically, hydrophobic (water-repelling) side chains of amino acids are tucked away inside the protein's core, avoiding water contact. In contrast, hydrophilic (water-attracting) side chains are positioned on the protein's exterior, exposed to the aqueous environment. This distribution contributes significantly to the high solubility of globular proteins in water.

The diverse structures of globular proteins are intrinsically linked to their wide range of biological functions. These functions include, but are not limited to, molecular binding, catalysis, regulation, transport, immune response, cellular signaling, and more. The structural variety of globular proteins thus mirrors the plethora of roles they play in

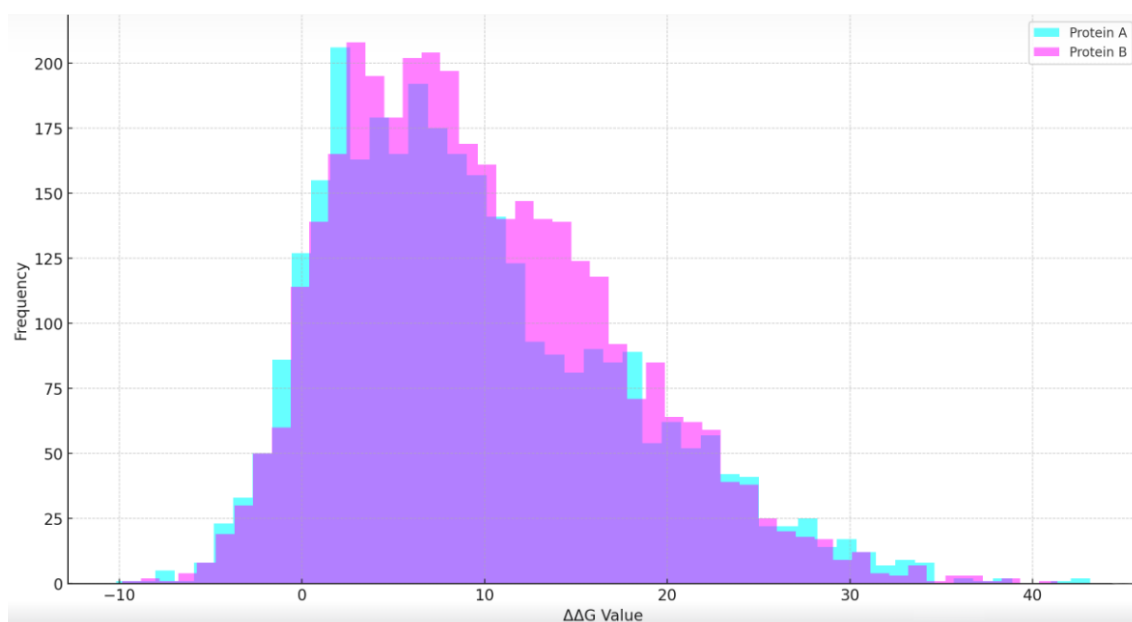


Figure 2.1: Pythia $\Delta\Delta G$ frequency bar plot overlap of chaperon dependent (protein a, e.coli dihydrofolate reductase (DHFR)) and an independent homolog (protein b, mouse DHFR).

biological systems.

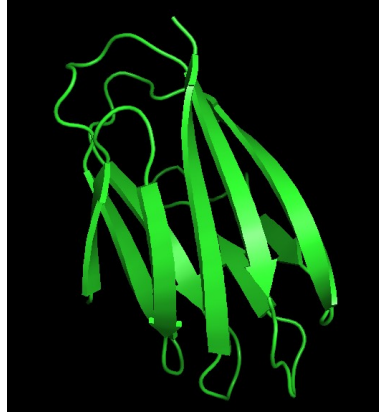
Our focus is on studying protein stability and both fibrous and membranar proteins require meta-structures to properly simulate them, co-factors/co-proteins in the case of fibrous proteins, or the membrane in case of transmembranar proteins. Globular proteins on the other hand, are characterized by their compact and densely folded structures, this allows for a 3-dimensional fitness metric of compactness. By minimising the radius of gyration and the Solvent Accessible Surface Area (SASA) the structures are forced into a globular shape. These metrics were empirically tested to produce the most diverse high-confidence array of globular secondary structures.

2.2 Alpha Helyx and Beta Sheet

In the realm of single small proteins most of the structure is derived by the α helyx and β sheet. However accounting for all possible sequences α are much more prevalent, easier to produce biophysically and come with a harsher $\Delta\Delta G$ penalty to interrupt. On the other hand β sheets are larger less spontaneous motifs but ultimately create extremely stable designs such as the β barrel.(figure 2.2)

2.3 Structural stability clues given by chaperon dependent homologs

Investigations using Pythia to analyze the $\Delta\Delta G$ landscape of *E. coli* proteins and their homologs, particularly those with a high dependency on chaperones for expression,

Figure 2.2: evolved β -barrel, pymol [39]

revealed a possible pattern of a rightward shift in $\Delta\Delta G$ maps (figure 2.1). This shift suggests that as the stability of a protein's structure increases, its $\Delta\Delta G$ transformation map values, reflecting the effect of single amino acid substitutions, also tend to rise.

To conceptualize this in terms of a global energy optimum, we consider a protein sequence as

$$S = \{a_1, a_2, \dots, a_n\}$$

where n is the length of the protein, and each a_i represents an amino acid in the sequence. The global energy optimum, G^* , is defined as a hypothetical state where every single residue transformation is either neutral or disruptive, that is tends to the maximization of values of the global $\Delta\Delta G$ map. In this state, for any amino acid a_i , any possible substitution $a'_i \neq a_i$ results in a positive $\Delta\Delta G$, indicating destabilization.

For each amino acid a_i in S , a set of possible substitutions

$$\{a'_{i1}, a'_{i2}, \dots, a'_{im}\}$$

is defined, where m is the number of possible substitutions (19, considering the 20 standard amino acids excluding the original). The $\Delta\Delta G$ for each substitution is given by

$$\Delta\Delta G(a_i \rightarrow a'_{ij})$$

where $1 \leq j \leq m$. For S to be at G^* , it is required that

$$\Delta\Delta G(a_i \rightarrow a'_{ij}) > 0$$

for all i and j .

However, this theoretical framework oversimplifies the complex nature of amino acid substitutions, the actual $\Delta\Delta G$ landscape is highly variable, influenced by the intricate secondary and tertiary interactions within the protein structure. Despite this an increasing trend in $\Delta\Delta G$ was observed consistently over the archival time period. Subsequent experiments with various $\Delta\Delta G$ -oriented design strategies, including positive, negative, and neutral mutations, were compared against standard random mutation and crossover methods to evaluate their relative effectiveness.

3.1 Structure

Assuming we already have a map of diverse protein structures we can use our mutator model to predict useful mutations, perform them, followed by the use of the sequence to structure model to create new structures that can be studied, placed on the map and repeat the process.(fig 3.1)

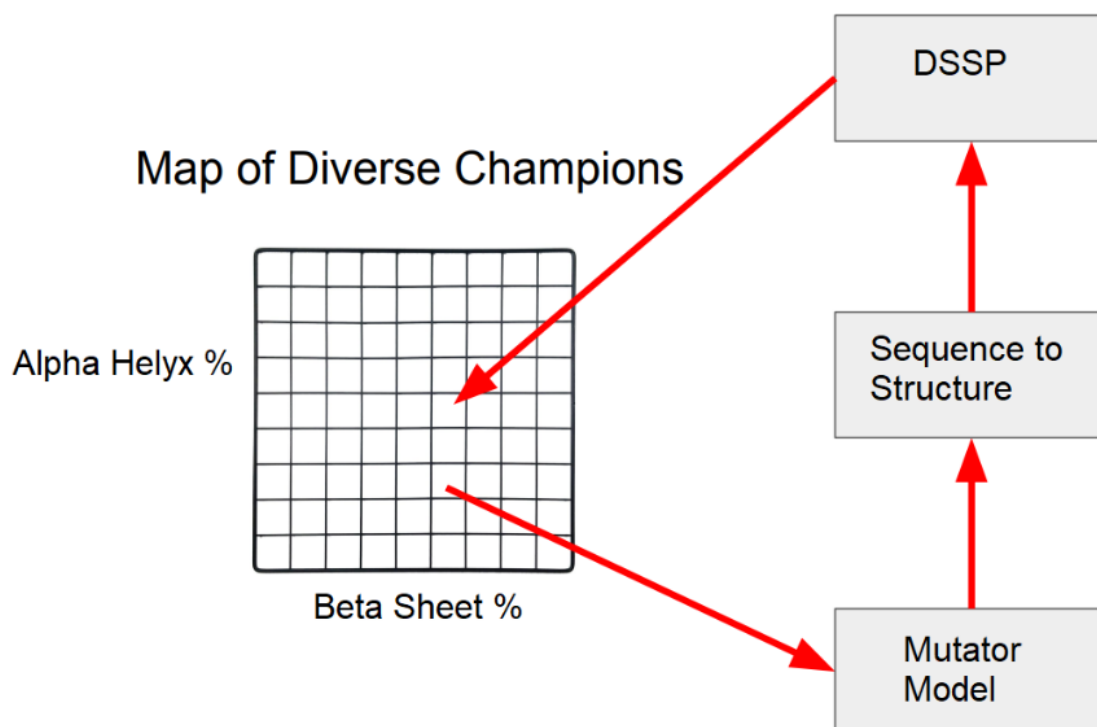


Figure 3.1: A meta overview.

3.2 Initialisation

The amino acid composition percentages used to create the initial population in this study are derived from the ProtScale tool, which is based on data from the UniProtKB/Swiss-Prot database.[40]

Table 3.1: Average aminoacid composition percentages.

Aminoacid Letter	A	R	N	D	C	Q	E	G	H	I
% Aminoacid Chance	8.25	5.53	4.06	5.45	1.37	3.93	6.75	7.07	2.27	5.96

Aminoacid Letter	L	K	M	F	P	S	T	W	Y	V
% Aminoacid Chance	9.66	5.84	2.42	3.86	4.7	6.56	5.34	1.08	2.92	6.87

3.2.1 Minimum Dependencies

Table 3.2: Dependencies for Computational Biology Pipeline

Tool	Description
NumPy	Numerical computations library
pickle	Object serialization library
BioPython	Bioinformatics tools for computational biology
Pytorch	Machine learning framework
OmegaFold	Single sequence structure model (PyTorch: torch==1.12.0+cu113)
pyDSSP	Secondary structure determination
FreeSASA	Solvent Accessible Surface Area (SASA) calculation
Pythia	$\Delta\Delta G$ single sequence model (PyTorch: torch==1.13.0+cu117)

Algorithm 1 MAP-ELITES ALGORITHM (NOVELFOLD VERSION)

```

1:  $P \leftarrow \emptyset, X \leftarrow \emptyset$       ▶ Initialize population variables: solutions  $X$  and performances  $P$ 
2: Initialize random population
3: Create a feature descriptor map NICHESPACE
4: for genN from 1 to generations do
5:    $x \leftarrow$  random selection from  $X$ 
6:    $x_0 \leftarrow$  random variation of  $x$ 
7:   Simulate  $x_0$  to get its feature descriptor  $b_0$ 
8:   Record the performance  $p_0$  of  $x_0$ 
9:   if  $P(b_0) = \emptyset$  or  $P(b_0) < p_0$  then
10:      $P(b_0) \leftarrow p_0$ 
11:      $X(b_0) \leftarrow x_0$ 
12:   end if
13: end for
14: return  $P$  and  $X$ 

```

The pipeline runs entirely on python and is hosted on [GitHub](#)[41].

3.3 Omegafold

In terms of computational requirements and settings, our OmegaFold implementation was executed on a single NVIDIA GTX 1070 graphics card with number of model cycles 2 (the standard if undefined is 10), this was done to increase the overall computation speed at the expense of pLDDT.

3.4 Prediction and annotation of structural features

Predicted structures of Omegafold were used for annotation of secondary elements with pyDSSP [42]. A proportionate measure of the features was taken by dividing the number of residues with the respective annotations, “H” for α -helices and “B” for β -sheets.

For the prediction of solvent-accessible surface area (SASA), we used the FreeSASA python module [43]. The library implements both Lee and Richards’ [44] and Shrake and Rupley’s approximations [45].

3.5 Diversity Metric

For every residue in the protein, we categorize it as one of three states: α (alpha), β (beta), or loop. Every protein is represented as a distinct combination of these three states, mapped onto a 2D map. The edges of this map, the extreme cases are represented as 100% α , 100% β , or 0% of either; however, biophysical design constraints usually prevent these extremes from occurring. The implementation of the MAP-Elites algorithm ensures that niches are created for every combination of these diversity measures. This approach guarantees a varied population set, permitting vast leaps in structure space through crossover between different secondary structure elements.[37]

3.6 K-Points Crossover

In this method, ‘k’ crossover points are chosen randomly along the length of the parent chromosomes. The segments of the chromosomes between these points are then swapped between the parents to produce new offspring. k is individually rolled per residue at the same time as mutations.

3.7 $\Delta\Delta G$ Mutators

Four different mutator bias strategies are used and compared relative to each other. Using the exponential function positive and negative mutations are highly selected for while neutral and control experiments use gaussian and random weight table respectively.

3.7.1 Weighted Random

The control set, mutations use the initialization weight table (Table 3.1).

3.7.2 Positive Exponential

Prioritizes mutations with higher Pythia values, using a selection probability proportional to e^x , where x is the normalized Pythia value. This method emphasizes the positive tail of the distribution.

3.7.3 Inverse Exponential

Focuses on mutations with lower Pythia values, with a selection probability proportional to e^{-x} . This approach heavily favors the negative tail of the distribution.

3.7.4 Gaussian Centered at 0

Implements a neutral selection based on a Gaussian distribution centered at zero. The probability of selecting a mutation is highest near a Pythia value of 0 and decreases symmetrically for values away from 0, creating a balanced distribution of mutations.

RESULTS

4.1 Confidence and structure

We plotted model confidence (pLDDT) and secondary structure across all experiments to assess if there was clear correlation to the appearance of secondary structure elements with the model's confidence.

We observe a trend where higher loop percentages are correlated with lower pLDDT scores across all mutation types, suggesting that proteins with more loops might be predicted with less confidence, this makes sense due to the dynamic and flexible nature of loop regions which are harder to predict accurately compared to more rigid stable structures like alpha helices and beta sheets.

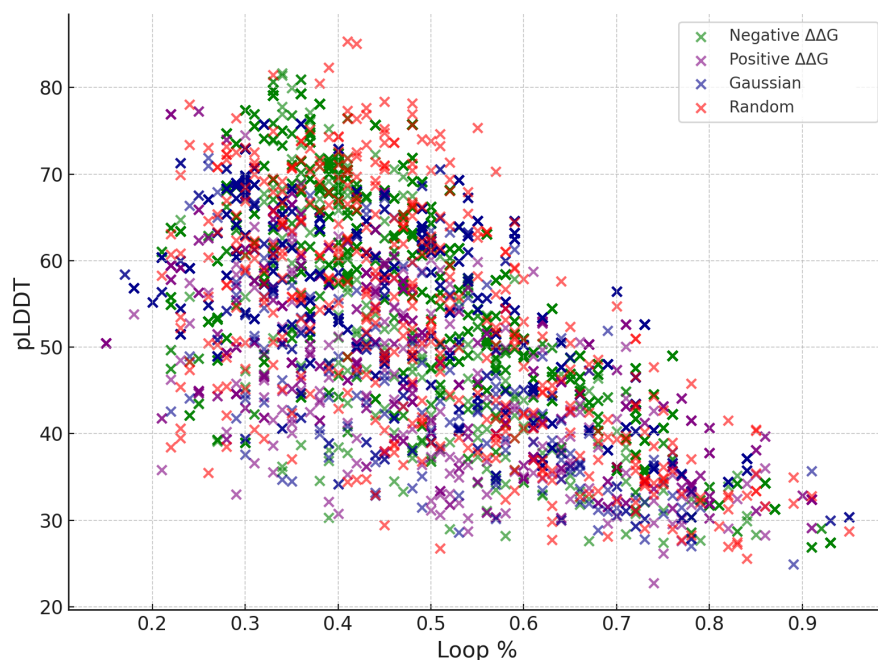


Figure 4.1: pLDDT corresponds to the model's prediction of its score on the local Distance Difference Test. It is a measure of local accuracy. Loop is $1-(\alpha+\beta)$. Every point is a final elite protein, the colours indicate what experiment created that protein.

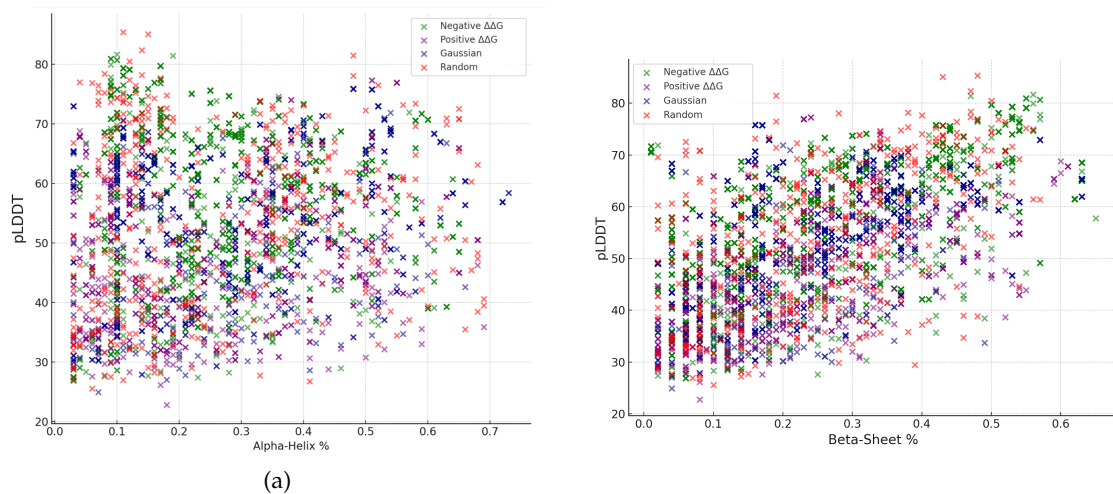


Figure 4.2: Alpha-Helyx and Beta-Sheet percentage versus model confidence

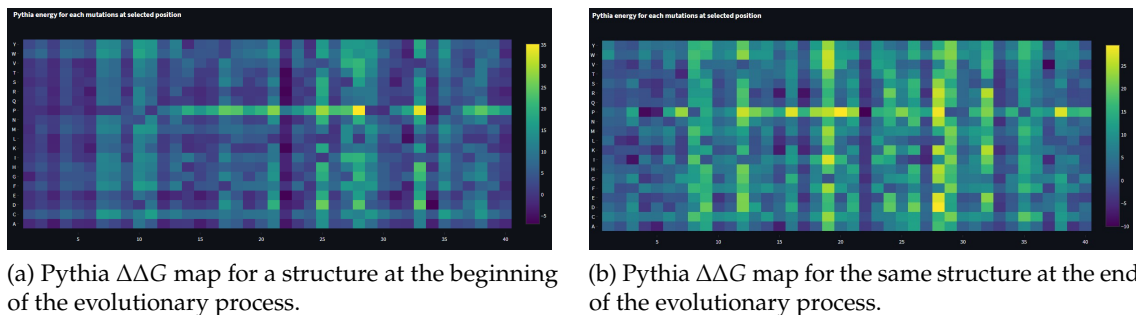


Figure 4.3: Pythia $\Delta\Delta G$ Maps at Different Stages of Evolutionary Process

But strangely we can see that this positive relationship between structure and confidence is mostly due to β sheets. α helyx seems to be both more prevalent and less indicative of model confidence, this indicates that their high prevalence across the hyperspace of possible sequences make their presence alone a poor indicator of an active fold.

4.2 Mutational search space

We can see in the pythia maps (fig 4.4 and fig 4.5) the structure at the start has several dark blue squares (transformations that reduce $\Delta\Delta G$) but after stability optimization a far greater percentage of transformations are now yellow (disruptive >25). The $\Delta\Delta G$ average values attempt to depict the population wide increase in structure stability for all experiments.(fig 4.5)

4.3 Evaluation of mutation strategies

Since the space of niches in the Map-Elites algorithm is limited, the only way to be added to the protein archive is to supersede an existing elite solution in protein fitness in this

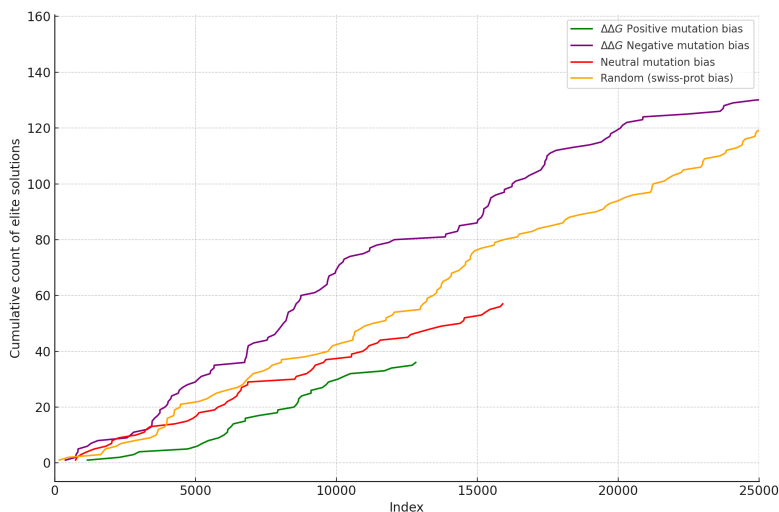


Figure 4.4: The index represents the number of protein structures analysed. The turnover of elite solutions is used as the metric for evolutionary speed.

case, stability. We look at cumulative count of elite solutions as a metric of evolutionary speed.

In the first 1000 generations most mutation strategies yield comparative results on the global trend of rising average $\Delta\Delta G$ values (fig 4.7), however on turnover results (fig 4.6) the positive and neutral strategies lag behind the random and negative bias strategies. The negative bias strategy was given extra computing time, being the the top performing strategy that uses pythia in order to be able to better compare to random (the control only runs omegafold making it substantially faster).

As expected the naive strategies of only mutating across an energetic gradient bias is not an effective global design strategy and eventually should be surpassed by random design however the interesting result is how negative bias seems to better approximate or exceed the random control while positive bias, the exact opposite is the weakest strategy, failing to create stable structures at a comparable rate. Neutral mutations, as expected, sits in the middle of these strategies its neutral effect showcasing the transformative effects of crossover.

4.4 The problem of objectives

Evolutionary algorithms and reinforcement learning share common ground in the ability to produce results that are surprising to their creators [46]. The magical creativity behind these methods is their non-prescriptive approach that leaves no stepping stone unturned[47], often leading to unwanted behavior.

One such behavior is called *Reward Hacking*, a term often used in the context of reinforcement learning. It describes situations where an agent learns to exploit loopholes

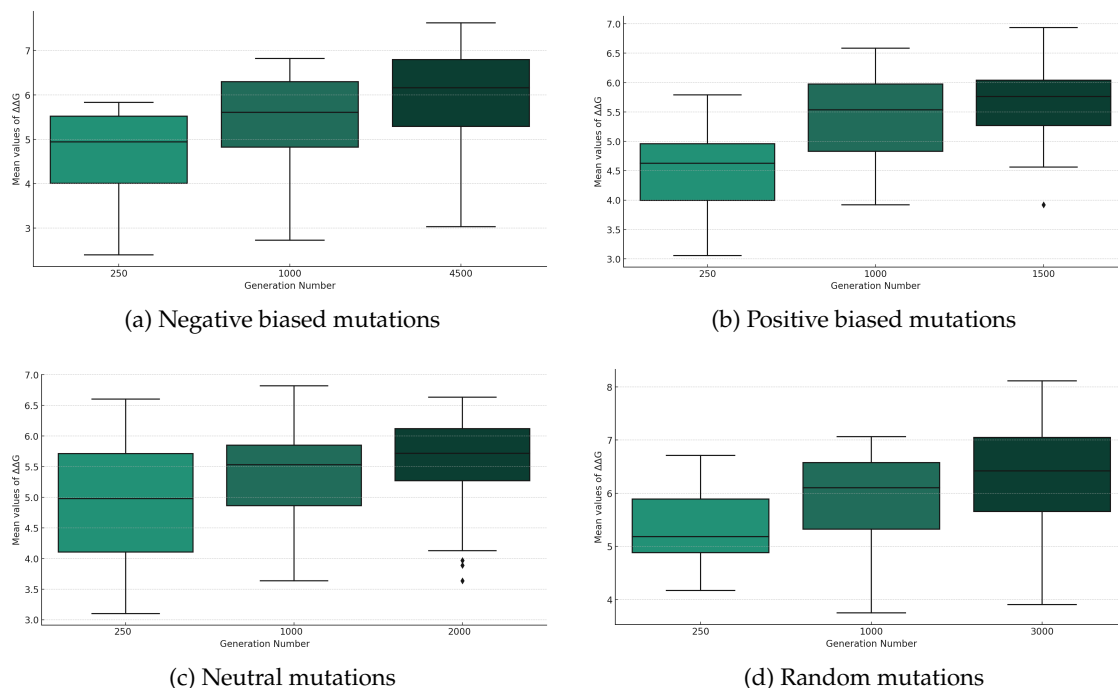


Figure 4.5: Average $\Delta\Delta G$ map of all Mutation Strategies

in the reward function to achieve higher rewards without actually solving the intended task. What might seem like a reasonable metric of performance can quickly turn into an unwanted proxy [48].

A naive protein design objective fitness function is the maximization of pLDDT (predicted Local Distance Difference Test), which is a per-residue confidence score signifying the confidence of the model prediction for each amino acid residue relative to the α -carbon atoms. Maximizing this metric with a simple evolutionary algorithm led to entire populations of near-unbroken α -helices for hundreds of amino acids with near 99% accuracy. Our proposed reason for this phenomenon is twofold:

1. α -Helix is easier to produce; there are a greater number of rotamer conformations that are stable in this motif compared to β -sheet, and it's a self-reinforcing structure that is very stable to continue at any length.
2. The OmegaPLM score simulates the pLDDT prediction signal and was partially trained to predict half-masked structures[38]. This means that a structure with a simple continuous pattern, like an unbroken α -helix, will be preferred over other more deceptive, partially-masked folds.

4.5 Foldseek Homology Analysis

Utilizing the same cutoff as ESM2[6] only the elite proteins with calculated pLDDT higher than 70 were taken for Foldseek[49] analysis. We saw no significant matches in the

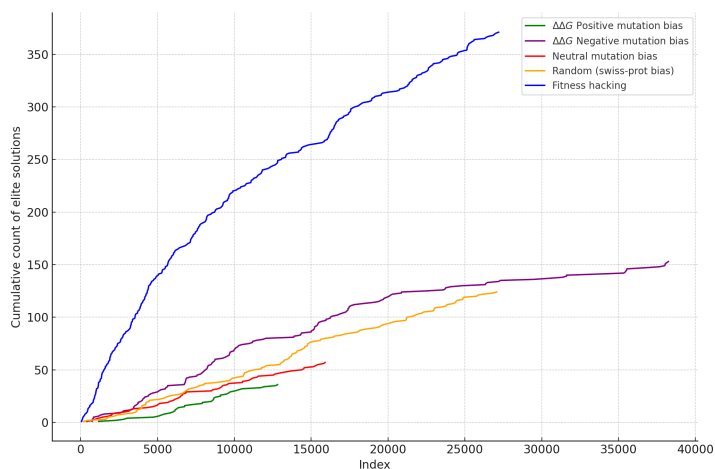


Figure 4.6: pLDDT objective function optimization.

alphafold database (below 0.5 TM-score) but the vast majority of elite proteins(>0.9) have greater than 0.7 TM-score matches on the ESM metagenomic atlas[50] database (FoldSeek version 5-53465f0). This means the solutions hold definitive similarity to several structures calculated by the SOTA metagenomic ESM2[6] model database.



Figure 4.7: ESM2 reconstitution of two evolved elite proteins, blue >90 pLDDT.

DISCUSSION AND FUTURE WORK

The single residue transformation $\Delta\Delta G$ landscape is a useful thermodynamical system correlated with the structural global stability of a protein, we show using two self-supervised models, of structure and mutational prediction, that negative bias mutations tends to produce stabler structures than neutral or positive bias mutations and that this reflects on their comparative future $\Delta\Delta G$ landscape.

5.1 Epistatic Interactions

The $\Delta\Delta G$ map deals with single residue transformations, this might hide very effective pairwise or n^{th} order substitutions. A computational study showed that the key factors that largely account for the variance in a protein's phenotype, as determined by its sequence-function relationship, are primarily the main and pairwise effects, along with a simple form of global nonlinearity[51]. These factors alone account for a median of 96% of the total variance in the measured phenotype and more than 92% in every case studied. Regarding epistasis, the role of higher-order epistatic interactions (involving three or more different amino acid residues) is much less significant than previously thought. While these epistatic interactions among residues were assumed to be pervasive, making protein functions difficult to predict from their sequence, recent findings show that these high-order interactions are not as common. High-order epistasis is negligible for the majority of genotypes, and only a tiny fraction (ranging from 0.02% to 2%) of genotypes are strongly influenced by such interactions. This indicates that the complexity added by high-order epistasis to the sequence-function relationship is limited to a very small subset of genotypes.

5.2 Binding sites and Functional Activity

While this research primarily delved into the stability and structure of proteins, the advancements in binding site identification offer a complementary perspective. In the realm of computational biology, the identification and analysis of protein binding sites

have attained remarkable sophistication, GRASP (Graphical Representation and Analysis of Structural Properties)[52] is a user-friendly accessible tool, built in a fully contained Google Colab setup. It demonstrates state-of-the-art performance in both recall and precision of single sequence binding sites. This method leverages advancements in graph neural networks and machine learning to identify binding sites with higher accuracy and efficiency than before.

Using Proteinfer, a deep network for protein functional inference, we find that a small subset(<10 percent) of proteins evolved show predicted enzymatic activity.[4]

BIBLIOGRAPHY

- [1] J. M. Lourenço. *The NOVAthesis L^AT_EX Template User's Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/main/template.pdf> (cit. on p. i).
- [2] A. J. Faure et al. “The genetic architecture of protein stability”. In: *bioRxiv* (2023). DOI: [10.1101/2023.10.27.564339](https://doi.org/10.1101/2023.10.27.564339). URL: <https://doi.org/10.1101/2023.10.27.564339> (cit. on p. 1).
- [3] N. A. Pierce and E. Winfree. “Protein Design is NP-hard”. In: *Protein Engineering, Design and Selection* 15.10 (2002), pp. 779–782. DOI: [10.1093/protein/15.10.779](https://doi.org/10.1093/protein/15.10.779). URL: <https://doi.org/10.1093/protein/15.10.779> (cit. on p. 1).
- [4] T. Sanderson et al. “ProteinInfer, deep neural networks for protein functional inference”. In: *eLife* 12 (2023), e80942. DOI: [10.7554/eLife.80942](https://doi.org/10.7554/eLife.80942). URL: <https://doi.org/10.7554/eLife.80942> (cit. on pp. 1, 23).
- [5] J. Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596 (2021), pp. 583–589 (cit. on p. 1).
- [6] Z. Lin et al. “Evolutionary-scale prediction of atomic-level protein structure with a language model”. In: *Science* 379.6637 (2023). Earlier versions as preprint: bioRxiv 2022.07.20.500902, pp. 1123–1130. DOI: [10.1126/science.ade2574](https://doi.org/10.1126/science.ade2574). URL: <https://www.science.org/doi/abs/10.1126/science.ade2574> (cit. on pp. 1, 20, 21).
- [7] M. Varadi et al. “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences”. In: *Nucleic Acids Research* gkad1011 (2023). DOI: [10.1093/nar/gkad1011](https://doi.org/10.1093/nar/gkad1011) (cit. on p. 1).
- [8] D. Hesslow et al. *RITA: a Study on Scaling Up Generative Protein Sequence Models*. 2022. arXiv: [2205.05789 \[q-bio.QM\]](https://arxiv.org/abs/2205.05789) (cit. on pp. 1, 2).
- [9] I. Anishchenko et al. “De novo protein design by deep network hallucination”. In: *Nature* 600 (2021), pp. 547–552 (cit. on p. 1).

- [10] J. S. Lee, J. Kim, and P. M. Kim. “Score-based generative modeling for de novo protein design”. In: *Nature Computational Science* 3 (2023), pp. 382–392. DOI: [10.1038/s43588-023-00440-3](https://doi.org/10.1038/s43588-023-00440-3) (cit. on p. 1).
- [11] M. Baek et al. “Accurate prediction of protein structures and interactions using a three-track neural network”. In: *Science* 373 (2021), pp. 871–876 (cit. on p. 1).
- [12] N. N. Thadani et al. “Learning from prepandemic data to forecast viral escape”. In: *Nature* 622 (2023), pp. 818–825. DOI: [10.1038/s41586-023-06617-0](https://doi.org/10.1038/s41586-023-06617-0). URL: <https://www.nature.com/articles/s41586-023-06617-0> (cit. on p. 2).
- [13] A. Jun Cheng. “Accurate proteome-wide missense variant effect prediction with AlphaMissense”. In: *Science* (2023). DOI: [10.1126/science.adg7492](https://doi.org/10.1126/science.adg7492). URL: <https://www.science.org/doi/10.1126/science.adg7492> (cit. on p. 2).
- [14] [Authors]. “Machine behaviour”. In: *Nature* ([Year]). URL: <https://www.nature.com/articles/s41586-019-1138-y> (cit. on p. 2).
- [15] R. Rao et al. “Transformer protein language models are unsupervised structure learners”. In: *bioRxiv* (2020). DOI: [10.1101/2020.12.15.422761](https://doi.org/10.1101/2020.12.15.422761). URL: <https://doi.org/10.1101/2020.12.15.422761> (cit. on p. 2).
- [16] D. Hesslow et al. “RITA: a Study on Scaling Up Generative Protein Sequence Models”. In: *CoRR abs/2205.05789* (2022). DOI: [10.48550/ARXIV.2205.05789](https://doi.org/10.48550/ARXIV.2205.05789). arXiv: [2205.05789](https://arxiv.org/abs/2205.05789). URL: <https://doi.org/10.48550/arXiv.2205.05789> (cit. on p. 2).
- [17] J. Vig et al. “BERTology Meets Biology: Interpreting Attention in Protein Language Models”. In: *arXiv* (2020). URL: <https://arxiv.org/abs/2006.15222> (cit. on p. 2).
- [18] D. Silver. *AlphaFold: Using AI for Scientific Discovery*. https://www.youtube.com/watch?v=Gg-w_n9NJIE. 2023 (cit. on p. 2).
- [19] Biology LibreTexts Contributors. *4.9: Protein Stability - Thermodynamics*. Biology LibreTexts. 2023. URL: https://bio.libretexts.org/Bookshelves/Biochemistry/Section_4%3A_Structure_of_Proteins/4.09%3A_Protein_Stability_-_Thermodynamics (cit. on p. 3).
- [20] PubMed Contributors. “Protein stability and degradation in health and disease”. In: *PubMed* (2023). URL: <https://pubmed.ncbi.nlm.nih.gov/> (cit. on p. 3).
- [21] AAT Bioquest Contributors. *What factors affect protein stability?* AAT Bioquest. Accessed: 2023-11-29. 2023. URL: <https://www.aatbio.com/resources/factors-affecting-protein-stability> (cit. on p. 3).
- [22] A. Goldenzweig and S. J. Fleishman. “Principles of Protein Stability and Their Application in Computational Design”. In: *Annual Review of Biochemistry* 87.1 (2018). PMID: 29401000, pp. 105–129. DOI: [10.1146/annurev-biochem-062917-012102](https://doi.org/10.1146/annurev-biochem-062917-012102). eprint: <https://doi.org/10.1146/annurev-biochem-062917-012102>. URL: <https://doi.org/10.1146/annurev-biochem-062917-012102> (cit. on p. 3).

- [23] A. Dickson and C. L. r. Brooks. “Quantifying chaperone-mediated transitions in the proteostasis network of *E. coli*”. In: *PLoS Computational Biology* 9.11 (2013), e1003324. DOI: [10.1371/journal.pcbi.1003324](https://doi.org/10.1371/journal.pcbi.1003324). URL: <https://pubmed.ncbi.nlm.nih.gov/24244134> (cit. on p. 4).
- [24] Y. Cho et al. “Individual and collective contributions of chaperoning and degradation to protein homeostasis in *E. coli*”. In: *Cell Reports* 11 (2 2015), pp. 321–333. DOI: [10.1016/j.celrep.2015.03.018](https://doi.org/10.1016/j.celrep.2015.03.018). URL: <https://pubmed.ncbi.nlm.nih.gov/25843722/> (cit. on p. 4).
- [25] D. A. Nissley et al. “Universal protein misfolding intermediates can bypass the proteostasis network and remain soluble and less functional”. In: *Nature Communications* 13.3081 (2022). DOI: [10.1038/s41467-022-30548-5](https://doi.org/10.1038/s41467-022-30548-5). URL: <https://www.nature.com/articles/s41467-022-30548-5> (cit. on p. 5).
- [26] G. P. Brady and K. A. Sharp. “Entropy in protein folding and in protein-protein interactions”. In: *Current Opinion in Structural Biology* 7.2 (1997), pp. 215–221. DOI: [10.1016/s0959-440x\(97\)80028-0](https://doi.org/10.1016/s0959-440x(97)80028-0). URL: <https://pubmed.ncbi.nlm.nih.gov/9094326/> (cit. on p. 5).
- [27] G. D. Rose. “Reframing the Protein Folding Problem: Entropy as Organizer”. In: *Biochemistry* 60.49 (2021), pp. 3753–3761. DOI: [10.1021/acs.biochem.1c00687](https://doi.org/10.1021/acs.biochem.1c00687). URL: <https://pubmed.ncbi.nlm.nih.gov/34855369/> (cit. on p. 5).
- [28] A. Goldenzweig and S. J. Fleishman. “Principles of Protein Stability and Their Application in Computational Design”. In: *Annual Review of Biochemistry* 87 (2018), pp. 105–129. DOI: [10.1146/annurev-biochem-062917-012102](https://doi.org/10.1146/annurev-biochem-062917-012102). URL: <https://pubmed.ncbi.nlm.nih.gov/29401000/> (cit. on pp. 5, 6).
- [29] J. Smith, H. Van Ness, and M. Abbott. *Introduction to Chemical Engineering Thermodynamics*. 8th. McGraw-Hill Education, 2017 (cit. on p. 6).
- [30] “A new protocol for atomic-level protein structure modeling”. In: NCBI (). Accessed: 2023-11-20. URL: <https://www.ncbi.nlm.nih.gov/> (cit. on p. 6).
- [31] *Template modeling score*. https://en.wikipedia.org/wiki/Template_modeling_score. Accessed: 2023-11-20 (cit. on p. 6).
- [32] “Scoring function for automated assessment of protein structure”. In: PubMed (). Accessed: 2023-11-20. URL: <https://pubmed.ncbi.nlm.nih.gov/> (cit. on p. 6).
- [33] Z. Group. *TM-score: Quantitative assessment of similarity between protein structures*. Accessed: 2023-11-20. URL: <http://zhanggroup.org/TM-score/> (cit. on p. 6).
- [34] D. E. Goldberg and J. H. Holland. “Genetic Algorithms and Machine Learning”. In: *Machine Learning* 3 (1988), pp. 95–99. ISSN: 0885-6125, 1573-0565. URL: <http://hdl.handle.net/2027.42/46947> (cit. on p. 7).

- [35] J. Lehman and K. O. Stanley. "Exploiting Open-Endedness to Solve Problems Through the Search for Novelty". In: *Proceedings of the Eleventh International Conference on Artificial Life (ALIFE XI)*. MIT Press. 2008, pp. 329–336 (cit. on p. 7).
- [36] D. E. Goldberg. "Simple genetic algorithms and the minimal, deceptive problem". In: *Genetic Algorithms and Simulated Annealing*. Ed. by L. Davis. Morgan Kaufmann, 1987, pp. 74–88 (cit. on p. 7).
- [37] J.-B. Mouret and J. Clune. *Illuminating search spaces by mapping elites*. 2015. arXiv: [1504.04909 \[cs.NE\]](#) (cit. on pp. 8, 15).
- [38] R. Wu et al. "High-resolution de novo structure prediction from primary sequence". In: *bioRxiv* (2022). DOI: [10.1101/2022.07.21.500999](#). eprint: <https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999.full.pdf>. URL: <https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999> (cit. on pp. 8, 20).
- [39] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 2.0". PyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC. 2015 (cit. on p. 12).
- [40] E Gasteiger et al. "ExpASY: the proteomics server for in-depth protein knowledge and analysis". In: *Nucleic Acids Res* 31 (2003), pp. 3784–3788 (cit. on p. 14).
- [41] R. Bernardino. *NovelFold: Github*. <https://github.com/rundead3/Novelfold>. Accessed: 2023-11-19. 2023 (cit. on p. 14).
- [42] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers: Original Research on Biomolecules* 22.12 (1983), pp. 2577–2637 (cit. on p. 15).
- [43] S. Mitternacht. "FreeSASA: An open source C library for solvent accessible surface area calculations". In: *F1000Research* 5 (2016), p. 189. DOI: [10.12688/f1000research.7931.1](#). URL: <https://freesasa.github.io> (cit. on p. 15).
- [44] B. Lee and F. M. Richards. "The interpretation of protein structures: estimation of static accessibility". In: *Journal of Molecular Biology* 55.3 (1971), pp. 379–400. DOI: [10.1016/0022-2836\(71\)90324-X](#) (cit. on p. 15).
- [45] A. Shrake and J. Rupley. "Environment and exposure to solvent of protein atoms. Lysozyme and insulin". In: *Journal of Molecular Biology* 79.2 (1973), pp. 351–371. DOI: [10.1016/0022-2836\(73\)90011-9](#) (cit. on p. 15).
- [46] J. Lehman et al. *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*. 2019. arXiv: [1803.03453 \[cs.NE\]](#) (cit. on p. 19).
- [47] B. G. Woolley and K. O. Stanley. *Exploring Promising Stepping Stones by Combining Novelty Search with Interactive Evolution*. 2012. arXiv: [1207.6682 \[cs.NE\]](#) (cit. on p. 19).

- [48] J. Lehman and K. O. Stanley. “Abandoning objectives: evolution through the search for novelty alone”. In: *Evol Comput* 19.2 (2011). Epub 2011 Feb 14, pp. 189–223. DOI: [10.1162/EVCO_a_00025](https://doi.org/10.1162/EVCO_a_00025) (cit. on p. 20).
- [49] M. van Kempen, S. Kim, C. Tumescheit, et al. “Fast and accurate protein structure search with Foldseek”. In: *Nature Biotechnology* 2023 (2023). DOI: [10.1038/s41587-023-01773-0](https://doi.org/10.1038/s41587-023-01773-0). URL: <https://doi.org/10.1038/s41587-023-01773-0> (cit. on p. 20).
- [50] S. Kieser et al. “ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data”. In: *BMC Bioinformatics* 21.257 (2020). DOI: [10.1186/s12859-020-03585-4](https://doi.org/10.1186/s12859-020-03585-4). URL: <https://doi.org/10.1186/s12859-020-03585-4> (cit. on p. 21).
- [51] Y. Park, B. P. Metzger, and J. W. Thornton. “The simplicity of protein sequence-function relationships”. In: *bioRxiv* (2023). DOI: [10.1101/2023.09.02.556057](https://doi.org/10.1101/2023.09.02.556057). URL: <https://www.biorxiv.org/content/10.1101/2023.09.02.556057v1> (cit. on p. 22).
- [52] Z. Smith et al. “Graph Attention Site Prediction (GrASP): Identifying Druggable Binding Sites Using Graph Neural Networks with Attention”. In: *bioRxiv* (2023). DOI: [10.1101/2023.07.25.550565](https://doi.org/10.1101/2023.07.25.550565). URL: <https://www.biorxiv.org/content/10.1101/2023.07.25.550565v1> (cit. on p. 23).



NOVA

UNIVERSIDADE NOVA
DE LISBOA

2023

EXPERIMENTS IN THE USE OF $\Delta\Delta G$ PREDICTIONS FOR DE-NOVO IN-SILICO PROTEIN DESIGN

Rodrigo António Correia Tavares Bernardino

MASTER IN COMPUTATIONAL BIOLOGY & BIOINFORMATICS