Deep Learning in Protein Design Post-AlphaFold

Author: Tim Huygelen

Supervisor: Dr Alan Lowe

Advanced Investigative Project in Molecular Biosciences (BIOC0021)

Department of Structural and Molecular Biology

Faculty of Life Sciences

April 18, 2023

Word count: 7481

Table of Contents

Abbreviations

AA	Amino Acid
AF	AlphaFold
AF1	AlphaFold 1
AF2	AlphaFold 2
CASP	Critical Assessment of Structure Prediction
CNN	convolutional neural network
FKIC	Kinematic Closure with Fragments
ML	Machine Learning
MMP	matrix metalloproteinase
MPNN	Message Passing Neural Network
MSA	Multiple Sequence Alignment
MSD	Multi-state design

Acknowledgements

I would like to thank my friends Ado, Anna, César, and Dario for being there for me in times of need. I would also like to thank my supervisor Dr Alan Lowe for his patience and guidance.

Abstract

The protein folding problem, a long-standing challenge in molecular biology, has recently witnessed a breakthrough with the development of AlphaFold. This revolutionary computational tool has achieved remarkable accuracy in predicting protein structures, sparking curiosity about its potential impact on the inverse folding problem - protein design. In this dissertation, we delve into the significant developments in protein design, catalysed by AlphaFold's breakthrough in 2020.

We examine recent progress in machine protein design tools, with an emphasis on the influence of AlphaFold and other deep learning-based computational tools. Our examination encompasses various techniques, including hallucination methods, diffusion models, sequence generation, and physical approaches. We also explore the synergistic relationship between machine learning and traditional computational methods in protein design.

We then analyse the use of traditional and machine learning-based computational tools in protein design through three case studies involving luciferases, metalloproteinase inhibitors, and the design of two-state hinge proteins. We scrutinise the role of AlphaFold 2, and tools inspired by AlphaFold in protein design and emphasise how these algorithms can be harnessed to their maximum potential.

We finally offer a recap of the state of the field, current limitations, and we propose a community-wide protein design competition to accelerate protein design research in the post-AlphaFold era. Additionally, we assess the impact of AlphaFold on protein design and contemplate the future of machine learning in this domain.

Introduction

The Protein Folding Problem

In 1961 Anfinsen postulated that a protein's structure is determined only by its amino acid (AA) sequence (1). Yet even though all the information about a protein's unique fold should be stored in its sequence, it has been particularly difficult to accurately predict a protein's 3D structure from its AA sequence alone. Since the 1960s researchers have been trying to solve the protein folding problem by working on newer and better tools and methodologies. The protein folding problem arguably still has not been solved completely, although recent developments since Google's breakthrough in 2020 have made it so the structures of most known globular proteins can be determined in silico.

Why Do We Want to Solve the Protein Folding Problem?

The ability to correctly determine protein structure from sequence offers two major opportunities: the first is that scientists would not have to rely on expensive and time-consuming experimental pipelines to reveal protein structure by analysing

folded protein, and the second is that scientists would develop an understanding of the inverse relationship, i.e., the ability to know which AA sequence would fold into a desired structure/conformation by determining the sequence of a desired structure (as opposed to determining structure from sequence). The second point would give us the ability to design proteins for several useful functions such as vaccines (2, 3), industrial catalysts, plastic degradation, protein-based logic gates (4), etc. The possibilities for this technology are vast.

CASP

In the 1990s, the Critical Assessment of Structure Prediction (CASP) competition was founded to track progress in solving the problem of protein folding (5). In a biennial competition, research teams compete to predict the protein structures of previously unsolved proteins based on sequence alone. Various metrics, such as atomic distances between predicted and experimentally solved structures, are used to evaluate the quality of the researchers' predictions. Over the years, researchers have achieved better and better results, but until CASP14 in 2020, the methods were still far from experimentally accurate (6).

Pre-AlphaFold

Before AlphaFold entered the field of protein structure prediction, the paradigm for in-silico protein structure determination was based on several computational methods, often used in combination. These include physics-based simulations (algorithms that use a set of constraints and forces defined by our knowledge of the types of interactions and physical forces involved in protein folding), homology modelling (mapping the sequence to proteins with previously determined structures with similar sequences), another method is to look at multiple sequence alignments (MSA), i.e., an alignment of AA sequences related to the protein in question. resembling natural evolution. In this process, ML algorithms continuously undergo "mutations", with those adaptations that better represent the data being favoured by a loss function. This is similar to how advantageous genetic mutations are unconsciously selected across generations, enabling a species to adapt to environmental "patterns".

Deep learning models are a subset of ML models that are more complex, composed of interconnected layers of artificial neurons that modify and transmit information to each other. This

MSA's are then used to look for patterns of AA conservation. AAs that are highly conserved may indicate that they are part of the protein's active site and therefore more likely to be in proximity (7, 8). The total information from these MSAs, which often contain more than 100 sequences, is usually analysed by an algorithm that outputs proportions and patterns contained in the MSA in a visually intuitive format to be read by researchers. MSA data can also be analysed by structure determination packages like Rosetta (9). The Rosetta software has been a major player in the field of de novo protein structure prediction since its inception by the Baker laboratory in 1998 (10). It has been able to gain this prominent position in the field through the continued and successful integration of recent technologies for structure determination (knowledge-based potentials, torsion angle probabilities, homology modelling, etc.). Although some aspects of machine learning (ML) have been integrated into the Rosetta package, e.g., for epitope prediction through RosettaAntibodyDesign (11), it was not until after CASP13 that deep learning was used extensively by Rosetta for protein structure prediction.

What is ML? How does it compare to conventional methods?

Conventional algorithms, including physics-based models, depend on explicit mathematical equations typically grounded in physical laws to depict the behaviour of proteins. These models can be highly accurate but are limited by our understanding of the underlying physics and the very high computational cost of solving these equations for large molecules like proteins (12).

In contrast, ML models learn from datasets without requiring explicit knowledge of the underlying physics and are often considerably less time-consuming to run, though they can be computationally intensive to train (12). ML models identify patterns and relationships in the training data in a manner somewhat

allows the model to discern more complex relationships between input and output (13).

However, if the training data is limited, the ML model may struggle to l generalise well to other proteins (14). A ML-based structure prediction algorithm trained exclusively on naturally occurring proteins may for example have difficulty accurately predicting de novo designed folds.

Another challenge with ML models is their limited interpretability, often referred to as the "black box" problem. This denotes the inherent difficulty in

understanding these algorithms. If a ML model proves more accurate than conventional methods in, for example, protein structure prediction, it may have picked up on patterns that researchers are unaware of. However, due to the black box problem, extracting this information to gain a deeper understanding of the underlying mechanisms behind the model's predictions can be extremely challenging (15).

Introducing AlphaFold

In 2018, AlphaFold (AF), a ML-based team at Google's DeepMind focused on protein structure prediction, participated in CASP13 (16). The team developed AlphaFold 1 (AF1), a deep-learning algorithm trained on experimentally determined protein structures that performed significantly better than the runner-up in that year's competition but still strayed from experimental accuracy. Two years later, for the CASP14 competition, AlphaFold 2 (AF2) was released (17). This new ML algorithm was based on a novel type of ML structure (self-attention) and sent shockwaves through the structural biology community by producing results with near experimental accuracy (17).This led to discussion in the community whether the structures predicted by AF can be considered equivalent to experimentally validated structures or not. While currently predictions by AF are not accepted as a proof in themselves to confidently determine a protein's structure, AF can produce a close to accurate result in minutes while experimental methods often take months of work and heaps of funding to determine a protein's structure, if successful at all. AF has directly influenced and accelerated the experimental determination of protein structures by providing structures that can be used as templates to help make sense of complex experimental output e.g., from x-ray crystallography (18).

AlphaFold 1

Previous developments in protein structure prediction in ML and the use of MSAs provided the right foundation for a team like DeepMind's AlphaFold, which put its extensive Deep Learning experience and gigantic stockpile of servers and funds to work. AF1 was based on a ML architecture known as a convolutional neural network (CNN). CNNs are ML algorithms commonly used for tasks involving image processing and recognition, and this was seen as a good fit for inter-residue contact maps (matrices describing pairwise distances between residues), which can be treated similarly to images by CNNs (19), AF1 also relied heavily on MSAs; the algorithm essentially required an MSA of related proteins as input to infer secondary structure. The algorithm was extensively trained by DeepMind using over 100,000 experimentally discovered protein sequences (16). This allowed it to apply general aspects of protein folding and physics as well as to understand how certain patterns in MSAs correlate with the proximity of residues. It is also worth noting that AF1 also uses Rosetta to refine its predicted protein structure produced by the CNN (11). All these contributions were critical to the success of AF1(20).

Attention

Around the same time AF1 entered the CASP13 competition, in 2017 a team at Google working on new ML technologies published a preprint called "Attention Is All You Need" (21). This paper argues that *Attention*, a ML technique that had previously been used for smaller modules within multi-core ML algorithms such as CNNs, could be used as the base structure (then called self-attention) of a new type of ML algorithm they named transformers. These transformers, they argue, are less prone to information loss than CNN and recurrent neural networks (RNNs), while being less computationally intensive and suitable for parallelisation, resulting in faster performance. In the preprint, they describe self-attention as follows: "Self-attention, sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence." (21). This new type of structure for ML algorithms seemed very versatile and promising to many researchers at the time who were intrigued by its unorthodox approach and promising results.

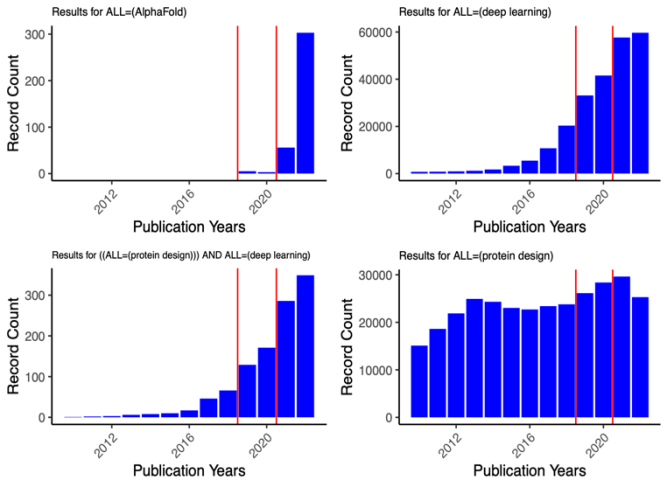

Figure 1. Bar plots illustrating the number of publications matched to specific queries (as shown in plot headings) used as input for Web of Science document search. Number of matches shown by year

of publication. Two red vertical lines are shown to indicate the release of CASP13 and CASP14 results.

AlphaFold 2

When the time came for CASP 14 in 2020, the AF team used this new self-attention technology to develop its new software: AlphaFold 2. AF2 is composed of two Transformer modules, the Evoformer (consisting of 48 blocks) and the Structure module (consisting of 8 blocks). The main inputs to AF2 are the protein sequence, MSA, and existing structural templates. First, the Evoformer looks for specific patterns in the protein's MSA, sequence, and structural information, and gradually builds up an idea of the protein's structure by changing the pair and MSA representations to store this derived information for the next block. At the end, the Evoformer feeds the final MSA- and pair representations into the structure module while also outputting a so-called distogram, a matrix containing information about probabilities of pairwise residue distances. Afterwards the structure module is responsible for reading all the structural information embedded in the two 2D files fed into it by the evoformer and translates them into 3D atom coordinates that form the protein's structure (17, 22). AF 2 still has some major drawbacks. Since AF2 is still so dependent on the evolutionary information from the MSA input, it struggles to model mutations not found in nature or protein sequences that do not have any evolutionarily relevant sequences to fill an MSA with, such as orphan sequences and antibodies. Its predictions also always assume the protein is rigid, making it rather unsuitable for dynamic proteins (23).

Protein Design

Protein design involves creating novel sequences or segments without solely depending on random mutations of an existing sequence. While some argue it only includes designing proteins entirely from scratch (24), the distinction is blurred due to ML's role in training on, but not explicitly reusing, existing sequences. This dissertation adopts the broader definition.

While the challenges are significant when it comes to designing an entirely new protein, the benefits are clear: the ability to deviate from the folds and tertiary structures imposed by evolution and cellular functionality could provide us with a wealth of opportunities for designing industrial, medical, and other necessary applications. D. Baker (24) describes how it is often easier to design a protein *de novo* than to make minor adjustments to naturally occurring proteins, due to the complicated energy landscapes of naturally occurring proteins that can result in major structural changes from a minor mutation. *De novo* proteins can be designed to have steep energy landscapes and a clear energy minimum, making their structure easier to predict.

Designing proteins involves several steps to ensure that the protein performs its intended function effectively. The first step is to think about the problem that needs to be solved or the desired function of the protein. Once this is established, scientists can then conceive the structure and active or functional site necessary for the protein to fulfil its purpose. There are several methods for designing the AA sequence of a protein. One approach is to assemble local structures together. Another is to use modular pieces with overlapping parts in the sequence, a method known as SEWING. Alternatively, scientists can look for structural elements in protein databases or databanks (8). Another method is rationally designing a protein using principles from protein dynamics, physics, and chemistry. To evaluate the designed sequence's structure, scientists use a scoring method and to give a measure of how well it fits the desired structure or function. The structure can be adjusted according to the scoring, and the process can be repeated to optimise the protein's structure. Another option is to use an end-to-end ML method that can design proteins on its own, although it still requires a clear problem or desired function to be defined (25) and for a loss function to be defined according to a scoring mechanism(10).

Post-AlphaFold: CASP15

So how do we connect these great advances in solving the protein folding problem to the inverse protein folding problem, namely protein design? And what possibilities does this radical introduction of deep learning into the field of structural biology mean for the scope of future projects in this field now witnessing the strengths of this technology? In 2022 CASP15 was held, and one thing is clear: whether it was the new approaches using language models (26, 27), the groups using aggressive sampling of AF to receive better results (28, 29), or the researchers using deep learning to model RNA structure (30, 31), ML is now a dominant force in structural biology. Transformer-based language models (32), such as ESMFold (33) and OmegaFold (34), use single sequence inputs for faster, more accurate protein structure predictions than single sequence AF2. While less accurate than AF2 with MSA, they are now widely employed by structural biologists (33). Hacks developed by independent researchers to model

protein multimers using AF2 have been incorporated into the main software as AF Multimer (35, 36). All the top ranking protein structure prediction methods at CASP15 were based on AF2 (37). Even Rosetta, the main toolbox for protein structure prediction before AF, has now successfully released a self-attention-based structure prediction algorithm (RoseTTAFold) proving to be competitive with AF2 (38). The impact of AF on protein structure prediction is indisputable; however, evaluating its role in shaping the development of protein design tools warrants additional investigation. In this dissertation, I will explore various notable advancements in ML-based protein design and structure prediction tools, and present three case studies to exemplify the paradigm shift spurred by DeepMind's AF in 2018 and 2020. These case studies will serve to highlight recurring themes within the field, culminating in a comprehensive summary of prospective future developments and opportunities deserving of further exploration.

Advances in protein design tools

Hallucination

Protein backbone hallucination, proposed by Anishchenko et. Al. (39)is a process that involves protein structure prediction algorithms to generate protein backbones. The process begins with inputting a sequence, either pre-existing or randomly generated, into the structure prediction network, which outputs a distance matrix. Then, a random mutation is made to the sequence and compared to the previous one using a loss function, typically the Kullback-Leibler divergence of the distogram with an average background distribution. This process selects proteins with the most distinct structures, and after thousands of iterations, a well-defined backbone emerges. The designs, validated in-vitro, provide unique, monomeric, stable proteins.

Initially performed using trRosetta, a precursor to RoseTTAFold, the method is now often used with algorithms such as AF2, RoseTTAFold, and OmegaFold.

The method was soon improved upon several times (40), including to generate sequences that fold to a specified structure. These designs, which initially lacked in vitro validation, were later proven to be imperfect but able to be rescued by a sequence generation method named ProteinMPNN (41). The same publication also introduced the use of gradient descent, instead of Monte-Carlo, to optimise the loss function, reducing the time required to generate a 120-residue protein from 90 minutes to just 5 minutes (42).

Furthermore, the method has been improved to scaffold existing functional sites, providing a way to incorporate fixed sequence or structural segments and design ideal scaffolds around them (43). To prevent bias, a different structure prediction tool than the one used for hallucination is commonly used for in silico structure validation.

Despite its success, protein backbone hallucination has drawbacks, such as its time-intensive nature and susceptibility to bias, resulting designs have proven to be inconsistent in quality (41).

This method exemplifies a direct contribution of AF to protein backbone design, with structural knowledge being directly harnessed from tools such as AF2 and RoseTTAFold.

Diffusion Models

In recent years, researchers have been focusing on the development of backbone generation models in protein design, with a particular interest in leveraging diffusion-based ML approaches. Commonly employed in image recognition and generation tasks, these diffusion models are trained to remove randomly added noise from geometry encoding matrices of existing proteins.

When presented with a blank matrix, or a partial structure, the models generate an ideal backbone structure, conforming to any given constraints.

A significant breakthrough in this field was the release of the RF *diffusion* model (44) in December 2022, which demonstrated notable improvements over previous methods, such as RosettaDesign.

RF *diffusion* is trained using RoseTTAFold although the authors acknowledge the possible use of other algorithms like AF2 or OmegaFold.

Subsequently, numerous other diffusion models, including Genie (45) and ProteinSGM (46), have emerged, but because of a lack of new protein design papers using them or other direct comparisons it is difficult to evaluate the relative quality of their outputs.

Although diffusion models have not yet been extensively employed in protein design research, the developers of these algorithms suggest that they could be utilised to design proteins with a

range of different properties including ideal backbones for single proteins with multiple active sites (46). Diffusion-based-models also provide us with the opportunity of rapid protein backbone generation, high quality, high throughput results, improving on the previously mentioned hallucination approach (44–46). As such, diffusion-based models hold significant potential for the future of protein design.

Sequence Generation

Protein sequence generation plays a crucial role in understanding and designing protein structures, with ProteinMPNN has quickly emerged as one of the most popular machine-learning-based sequence generation algorithms since its release in 2022 (44). Unlike hallucination-based methods and the mostly “conventional” Rosetta toolbox, ProteinMPNN is a Message Passing Neural Network (MPNN) (41) that offers superior performance in terms of accuracy and computational efficiency. In this approach, proteins are represented as graphs, with nodes corresponding to the protein sequence and edges encapsulating geometric properties such as inter-residue distances and torsion angles.

MPNNs are graph neural networks, they take as input a graph endowed with node and edge features and compute a function that depends on both the features and the graph structure. MPNNs propagate node features by exchanging information between adjacent nodes, allowing information to be iteratively updated and shared throughout the graph.

In the original release paper of ProteinMPNN they experimentally validate its accuracy by giving it the predicted crystal structures of proteins designed with AF deep network hallucination (39, 41). These designed structures have high predicted accuracies but turned out to be highly insoluble. When ProteinMPNN was used to generate sequences that fold to the same coordinates, the median soluble yield increased from 9 mg/l to 247 mg/l. The generated sequences also folded to structures highly similar to the provided coordinates and were highly thermostable (41).

One key advantage of ProteinMPNN is its location-independent sequence generation, which allows researchers to manually fix certain residues in a sequence. The algorithm then constructs the sequence around these fixed points, offering greater control, especially for functional site design.

Researchers have also found their own novel applications of ProteinMPNN such as generating multiple sequences for a specific structure and inferring from residue conservation which residues are most important to the protein’s structure (47). While this method is not fool proof, it can help construct an informed hypothesis to test with more costly experimental methods.

ProteinMPNN is now a major force in sequence design, perhaps being the most powerful sequence generation tool currently available. Even the team at Rosetta working on RF *diffusion* opted to use ProteinMPNN instead of the in-house Rosetta FastDesign to generate sequences for RF *diffusion* generated backbones (44).

Physical methods

Physical methods, such as those contained in Rosetta(48, 49), have long been essential in protein structure prediction and design. Historically, ML has played a role in physical algorithms like Rosetta, primarily by adjusting parameters within defined physical potentials (50). However, it was not until the introduction of AF1 and, more notably, AF2 that we can see an explosion in ML tools in structural biology.

Due to the explainable and adjustable nature of physical methods, they exhibit fewer biases than ML-based methods, and can produce more diverse results (51, 52). However, their requirement for significant computational resources frequently results in homology modelling and MSA analysis being employed for most of the protein structure prediction pipeline, with physical methods used for smaller adjustments(53). Despite these challenges, physical methods remain valuable, particularly in protein dynamics, where they outperform ML models. ML algorithms like AF2 still depend on physical potentials during structure relaxation steps to enhance accuracy (17).

The distinction between physical and ML models isn't clear-cut and may become less distinct over time. Recent publications combine physical and ML methods in protein structure-related algorithms. For example, a study used ML to imitate physics-based potentials for protein dynamics, creating an algorithm that provides similar accuracy with faster runtime (54). This algorithm can also model proteins in varied environments, unlike standard ML models, which only simulate proteins in standard conditions. While its accuracy is limited, this research shows promising progress in integrating physics-based

knowledge into modern ML models in structural biology.

Integrating traditional and ML methods

To sum up, protein design is a rapidly evolving field where a variety of computational tools with a wide diversity of architectures are emerging. The best strategies may involve a combination of these methods, as demonstrated by the mutual complementarity of RF diffusion and ProteinMPNN. To facilitate the interoperability of these algorithms, keeping these methods and their parameters universally accessible is crucial. Almost all ML-based methods cited in this article are open source but that does not mean the authors are practicing open science. Way too often the learned parameters of ML models like those of AF2 and RoseTTAFold are not released with the source code (55). This approach discourages open science and prevents the discovery of crucial modifications and workarounds.

Current Use of Protein Design Tools

Traditional physics-based methods and ML models are now widely used in protein design. However, these approaches are often combined to make use of their complementary strengths, as traditional methods offer detailed atomic interaction information but are computationally expensive, while ML efficiently analyses large datasets and identifies complex patterns. Although ML has overtaken some steps in the protein design pipeline, conventional techniques still excel in functional site and dynamic protein design.

To illustrate current usage of traditional and ML methods in protein design research, I will discuss three case studies on designed and engineered functional proteins: Luciferases, Metalloproteinase Inhibitors, and two-state hinge proteins. These case studies highlight diverse use cases of computational protein design tools and their incorporation into creatively designed pipelines, combining ML algorithms with conventional protein design software. My focus will be on the methods employed in the articles in question rather than the broader implications of the designs themselves.

Case Study 1: Luciferase design

In the article titled "*De novo* design of luciferases using deep learning" the authors design novel luciferases that exhibit high activity and specificity for synthetic luciferin substrates(56). To achieve this, the authors implement a family-wide hallucination approach for unconstrained *de novo* design using trRosetta (57), a DL-based precursor to RoseTTAFold, as well as conventional computational design tools such as RosettaDesign and RifDock.

The methodology involved in this research can be described as follows: Initially, the researchers identified the synthetic luciferin DTZ as an ideal substrate, primarily because it does not necessitate cofactors for luminescence. They then docked DTZ into 4,000 small-molecule-binding proteins to analyse binding and identified NTF-2 as the top candidate. The docking process employed RifGen (58) to enumerate rotamer interaction fields (RIFs) surrounding the substrate conformers, significantly reducing the computational time required for modelling the target's interaction energy (58). The complementary RifDock tool was used to dock each conformer and its associated RIF within the central cavity of every scaffold.

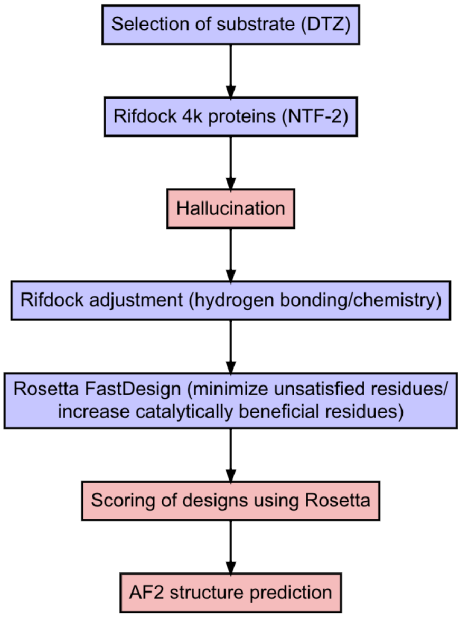

Figure 2. Flow chart of the simplified design pipeline from Case Study 1. Blue indicates steps performed using “conventional” methods, red indicates steps performed using new ML-based methods.

2,000 sequences from the NTF-2 family were then used as input for the deep learning hallucination approach, which involved cutting out sequence sections that seemed suboptimal to the researchers, shortening unnecessary loops, and retaining specific residue pairwise distances to maintain the NTF-2-like consensus fold. Moreover, running the Monte-Carlo deep learning hallucination approach using trRosetta generates

sequences and structures with high predicted accuracy scores and distinct folds.

To further specify the backbone conformation and functionalise the pocket, entire hydrogen bonding networks from native NTF2-like proteins were installed into the designs. On top of this, RifDock was employed to ensure specific hydrogen bond interactions between arginine and the secondary amine in the pyrazine ring of the colenterazine-like substrate. Additionally, RifDock was utilised to score the chemical interactions of the pockets with the substrate and alterations were made accordingly.

Following RifDock, the researchers used Rosetta sequence design, where the score function was reweighted for higher buried `unsat_penalty`. This approach minimised buried unsatisfied residues and increased pre-organised architectures in the core, which are known to be beneficial for a catalytic pocket. Two rounds of Rosetta FastDesign (58, 59) were performed to optimise surrounding residues and enable the redesign of RIF residues. The final set of designs was obtained after filtering by Rosetta ligand-binding interface energy, shape complementarity, contact molecular surface, number of HbondsToResidue, and the presence of N1 hydrogen bond. To assess their design model, single sequence structure prediction using AF2 was performed.

For the identification of h-CTZ as a substrate, the researchers followed similar steps as mentioned above. They utilised ProteinPMNN to redesign sequences for hallucinated NTF-2 scaffolds, and AF2 was employed to predict protein structures. With these scaffolds, the RifDock design strategy and Rosetta were employed to redesign all residues within 4 Å of the ligand. Sequence optimisation was conducted using ProteinPMNN, and AF2 was utilised to obtain predicted 3D protein models. The researchers then experimentally screened luciferase activity and identified two designs (HTZ3-D2 and HTZ3-G4) that exhibited luciferase activity and substrate selectivity to h-CTZ.

The success rate increased significantly in the second round (for CTZ instead of DTZ), likely owing to the knowledge of active-site geometry obtained from the first round and the robustness of the ProteinMPNN sequence design tool.

Case Study 2: Metalloproteinase Inhibitor Design

In the second case study, titled "A Broad Matrix Metalloproteinase Inhibitor with Designed Loop Extension Exhibits Ultra-High Specificity for MMP-14" (60), the authors aimed to redesign a loop extension in an existing metalloproteinase inhibitor to enhance its specificity for the matrix metalloproteinase (MMP) MMP-14. They employed Rosetta to model the protein for visual identification for a design site and to generate a large library of the region to be designed. Structure characterization was performed using creative ML-based approaches and the engineered protein was evaluated for cancer suppression.

They selected N-TIMP2 as the MMP inhibitor to be redesigned and “visually” identified a region suitable for a designed loop insertion to bind distal MMP residues. They computationally modelled different insertion lengths using an unspecified method, an insertion length of 7 residues was “visually determined” to be most suitable for loop design. Using Rosetta Remodel, 2000 loop sequences were generated. The designs were modelled in complex with the MMP-14 target protein using the SciPy Python library and a cluster of sequences visually identified to border two target residues was selected for redesign using Rosetta FastDesign and Rosetta Relax. The resulting sequences were modelled using Rosetta's Kinematic Closure with Fragments (FKIC) tool, both independently and in complex. Seven designs were chosen due to their exhibition of a single low-energy state.

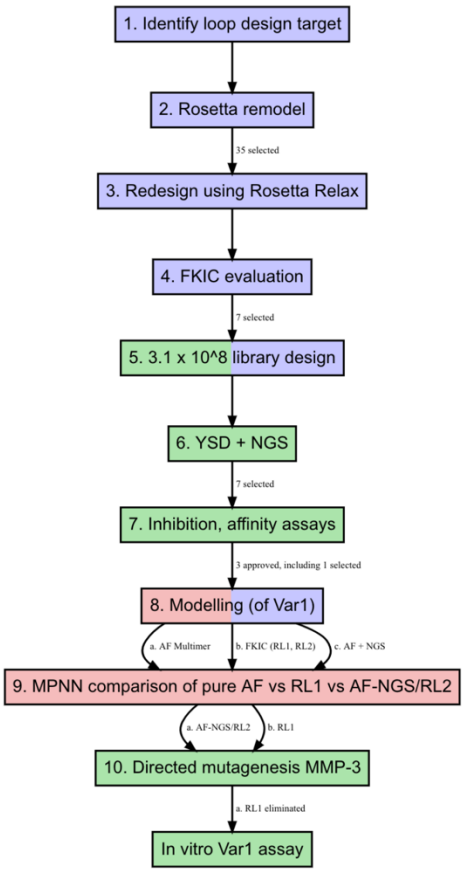

Figure 3. Flow chart of the simplified design pipeline from Case Study 2. Green indicates in vitro steps, blue indicates steps performed using "conventional" methods, red indicates steps performed using new ML-based methods.

The authors kept conserved residues in the 7 sequences intact and the others were randomised to create a library of 3.1 x 10^8 inhibitor variants. These were subsequently evaluated for ligand affinity on yeast surface display (YSD) and were sequenced with next-generation sequencing (NGS). Seven sequences were selected for their high affinity for MMP-14 and low affinity for the off-target MMP-3. In another experiment three of these variants were chosen for their high affinity and specificity for MMP-14, with Var1 performing best.

Var1 and MMP-14 were subsequently modelled using AF Multimer and Rosetta FKIC. Although most of the structure was modelled with high predicted accuracy using AF, the loop region had low predicted accuracy. To overcome this issue, an MSA input was generated with sequences from the same structural cluster as Var1, obtained from post-YSD NGS. This resulted in a substantial increase in predicted accuracy for the loop region, a respectable pLDDT of 85. FKIC modelling produces 2 opposing predicted Var1 structures, named RL1 and RL2.

Different model structures for Var1 were compared using the resulting structures to generate sequences with ProteinMPNN and RosettaFastDesign. Sequences generated from the NGS-aided AF2 approach (NGS-AF) structures were most similar to Var1's sequence. The RL 1 model produced sequences with structures less like those generated from the AF-NGS structure. The RL2 model proved to be better, however, the output sequences' structures were comparable to the NGS-AF model, so the proposed structures were treated as one. The AF model without the aid from NGS performed poorly, with predicted sequences showing a lot of variation, thus AF was discarded. Residue conservation analysis was performed on generated sequences, the authors hypothesised that the conserved residues are more important to its structure.

Directed mutagenesis of three MMP-3 residues was performed to assess binding of Wild Type (WT) and Var1 N-TIMP2 to MMP-3 and see if differential binding of certain residues aligns with the AF-NGS or RL1 (FKIC) models. It was found that the mutations affected Var1's MMP binding in a manner consistent with AF2-NGS, and not RL1. RL1 was thus discarded.

Finally, the researchers assayed Var1's ability to inhibit breast cancer cell invasion in comparison to WT N-TIMP2. They found that Var1 was equally effective to WT N-TIMP2, and more specific. This finding highlights the potential of the redesigned loop extension in the MMP inhibitor to provide enhanced specificity and improved inhibition for MMP-14, which could have implications for developing cancer treatments.

Despite the promising results, the paper's core aspects—generating a library, identifying the highest affinity binder, and evaluating in vitro—relied on basic computational methods and a brute-force approach to create random loop sequences for in vitro evaluation. Although the design chosen post-YSD was not subject to further alterations, the researchers were evidently familiar with sequence design techniques such as ProteinMPNN and Rosetta FastDesign, as these were utilized in the structure determination and evaluation stages. This presents a seemingly missed opportunity to leverage cutting-edge ML or better conventional methods like FastDesign to engineer the potential cancer treatment.

While the strategy of using sequences obtained from NGS appears innovative and ostensibly yields superior results, the article does not provide a clear explanation of how this might lead to

increased accuracy. This is particularly relevant, given the potential for artificially inflated predicted accuracy due to AF2's assumption that the provided MSA reflects evolutionary constraints, when in fact, it has been randomly generated and selected based on ligand affinity and structure. As the template structure remains mostly unchanged with only one loop region modified, this work is more aptly classified as protein engineering rather than protein design.

Case Study 3: Design of Two-State Hinge Proteins Using a ProteinMPNN Tool for Homo-oligomers

In the study titled "Design of Stimulus-Responsive Two-State Hinge Proteins," the authors aimed to design proteins with two distinct conformational states (4), one of which is occupied depending on the presence of an effector peptide, prompting a state switch. The main computational design tools employed are ProteinMPNN and Rosetta FastDesign.

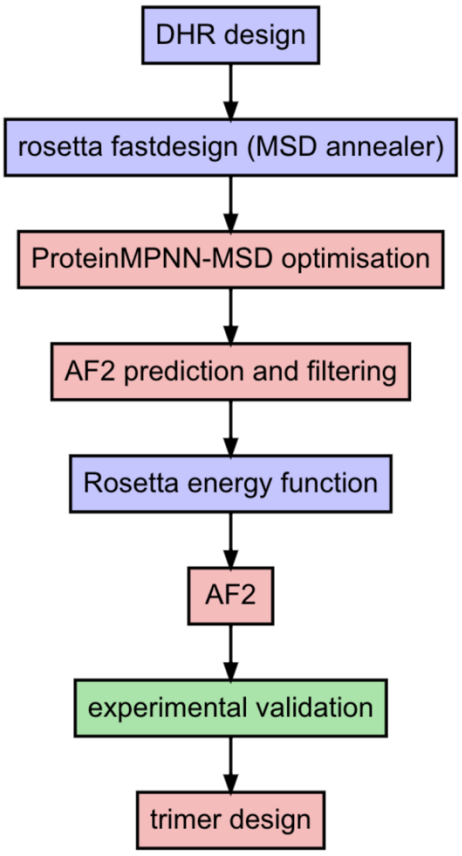

Figure 4. Flow chart of the simplified design pipeline from Case Study 3. Green indicates in vitro steps, blue indicates steps performed using “conventional” methods, red indicates steps performed using new ML-based methods.

The researchers utilised pre-existing Designed Helical Repeat proteins as templates. They repositioned the second conformation's coordinates to create a gap flanked by one hinge domain on each side. Next, a single strand of the repeat protein filled the gap to serve as the initial template for the effector peptide. Any designs that resulted in significant clashes were disregarded. Additionally, the effector peptide was lengthened to increase the interface size between it and state Y. PyRosetta FastDesign was then employed to make minor adjustments to the backbone structure and sequence, optimising affinity between hinge domains and the effector peptide in state Y. Designs that did not form proper contacts between the domains and peptide were discarded. Finally, Rosetta's Blueprint Builder was used to redesign the flexible loop between the hinge domains, tailoring it to state Y.

After experimenting with various multi-state design techniques, the researchers opted for Rosetta's FastDesign to redesign the sequences of the proteins derived from the repeat proteins. Specifically, they used a variant of FastDesign that incorporates a modified, multi-state, symmetric sequence design annealer. It was selected primarily due to its ease of use, computational scalability, and tunability.

The sequences were subsequently refined by employing ML-based ProteinMPNN multi-state design (MSD), utilising a feature designed for homo-oligomer design (41), which links specific residues in multiple sequence (segments), ensuring they remain consistent with each other. ProteinMPNN generated several dozens of possible sequences for each backbone pair.

Sequence structures were then first predicted for the effector-bound state (Y) and subsequently state X using AF2, and filtered according to similarity to reference structures for state Y and X.

After this step, in vitro validation found that hinge proteins tended to be soluble, but many peptides were not. This issue was mitigated by modifying interface residues or truncating effector sequences.

The researchers aimed to design a protein with three possible states: closed conformation (state X), open conformation without peptide, and open conformation with peptide (state Y). To ensure the designed protein functions effectively as a two-state switch, it must prefer state X. However, when the peptide is present, the protein should favour state Y more than both the closed and open states without the peptide. This design ensures that the

peptide's presence dictates whether the open or closed state is more favourable. To achieve this, the researchers utilised the Rosetta Energy Function (48) and predicted solvent-exposed hydrophobicity to filter the generated sequences. Remaining sequence pairs with low Rosetta-predicted binding affinity were also removed.

Using double electron-electron resonance spectroscopy to analyse protein geometries with or without the peptide. It was found that all hinge proteins undergo some form of conformational change when the effector is present. Two pairs the authors deemed most suitable were then selected for X-ray crystallography, X and Y states of both agreed well with the AF2 predicted target structures.

For quantitative investigation of thermodynamics and kinetics of the designs the authors made use of Förster resonance energy transfer (FRET) analysis (61). In FRET, adequate distance between N and C termini is vital. Researchers modified protein sequences near these termini employing ProteinMPNN and stabilised the closed conformation with a disulphide "stapling technique." They added cysteine residues at specific positions. Locations were identified by a sampling method evaluating Rosetta full-atom energies (62) and filtered for correct inter-cysteine distances using AF2, to validate effector peptide binding in state X. Results showed that without the effector present 99.5% of the proteins are correctly in state X.

The researchers also explored whether larger peptides could be used as effectors through RoseTTaFold diffusion-inpainting method, recently succeeded by RF *diffusion,* to add two additional peptide segments to the effector, tripling its size. After filtering these sequences by RoseTTAFold and AF2 predicted local distance difference test (pLDDT), proteins were found to show peptide-dependent conformational changes in vitro.

Analysis/summary of case studies

So how do these case studies highlight ML protein design tools being used in research? There are several themes in the case studies mentioned and discuss the implications of the findings for the broader field of protein design.

In the first case study, a high-activity enzyme was effectively designed by combining an innovative deep learning hallucination method with traditional approaches. This demonstrates the strength of ML methods in reducing the time and effort required for protein design, while also highlighting the continued importance of conventional computational tools and thoughtful integration into the pipeline in question.

The second case study reveals a limitation in the field, as the selection of design tools appeared somewhat arbitrary, indicating a lack of a universal protein design approach. The researchers in this case primarily employed FKIC for protein structure prediction until the designed sequence was finalized, even though language models have demonstrated speed and accuracy. This emphasizes the necessity for researchers to carefully choose the appropriate method for specific tasks instead of relying on a single tool indiscriminately.

In the third case study, the ProteinMPNN homo-oligomer design method was utilised to design multi-state proteins, showcasing the potential of these new ML algorithms to be used in creative ways to further our capabilities in protein design.

These case studies highlight the strong presence post-AF ML models have in design pipelines and how they can be employed to solve an abundance of protein design problems with conventional tools picking up where these tools are lacking and vice-versa. And also the important role of the researcher to curate these pipelines using the variety of tools available to them in a logical manner.

Discussion/Conclusion

Overview of the field and limitations

The field of protein design has experienced rapid growth and innovation in the last 2-3 years, with numerous algorithms continuously being released, seemingly at a faster pace than the release of protein design papers that use these new tools (63). Several tools that were promising when they were released in the last 2 years have already become obsolete due to the release of a better or updated software. Despite these advancements there are several limitations worth noting.

One of the key limitations is the lack of explainability of these DL algorithms. Furthermore, biases in the data can lead to unintended consequences and inaccurate predictions. This is particularly evident in protein dynamics and computational modelling of chemical reactions, where the scarcity of structured data presents significant challenges for ML algorithm training. Research promising to better understand and

prevent these biases should be focused on going forward as our dependence on deep learning is likely to increase. One tangential example of this is research on the strategic incorporation and mixing of experimental and ML generated data as inputs for ML algorithms to prevent large biases while benefitting from large training sets (64).

Open source codes are practically the norm these days in scholarly research But that does not mean that all science is open science, as seen with the non-release ML parameters. This evidences the need to pressure corporations like Google and Meta to release their research openly and freely.

Proposing a New Protein Design Competition: Prospects and Challenges

The Critical Assessment of protein Structure Prediction (CASP) has been highly successful in driving progress in protein structure prediction. This competition has enabled direct comparisons between different pipelines and methods, providing clear objectives for protein design and encouraging groups to compete to prove the efficacy of their tools and pipelines. In the field of molecular biology there are now several similar competitions including for CAFA (65) and CAPRI (66) but while there is a competition for protein design it is based on a specific method requiring the researchers to use a defined set of building block sequence segments. While this competition has been beneficial, I believe it is time for a standardised design competition.

The benefits of these competitions would apply well to research for protein design tools and would foster healthy competition between research groups to achieve more accurate results and to make it easier for researchers designing proteins to pick out accurate and reliable protein design tools to use for their pipelines knowing that they were fairly assessed.

While it could be argued that the design targets of the competition which proteins to design could be an arbitrary process, favouring some tools over others because of their differing applications. And there are cost concerns concerning independent in vitro characterisation of the designed structures.

Potential solutions could involve creating sub-competitions, each with different objectives, such as designing a sequence to fold as close as possible to set coordinates, designing proteins with a specific function, creating new scaffolds for existing functional sites, or redesigning small parts of proteins or functional sites. For the cost concerns there might need to be a pre-selection process to limit the number of protein structures that need to be experimentally determined. This could include validation in silico to remove sequences that are very unlikely to have the desired properties. For other sub-competitions the structure of the protein might not be important and more simple enzyme assays could be conducted instead to measure the catalytic activity of the designs.

Assessing the Impact of AlphaFold on Protein Design

Empirically assessing the impact of AF on protein design is challenging. While deep learning (DL)-based protein design methods have become more popular since AF2's release, DL has gained popularity in general over the years. Nevertheless, the combined growth of DL and protein design has outpaced their individual growth rates over the last two years (as shown in figure 1.).

Most papers cited in this review reference AF, and it is now common practice to run protein structure prediction steps using AF or other DL methods throughout the computational protein design process and before in vitro structure determination.

AF's usage of transformer architecture has been adopted by many new DL methods, such as ESMFold (67), OmegaFold (34), and RoseTTaFold(38). And the newly sparked interest in Deep Learning in structural biology has presumably directly inspired researchers to consider making use of Deep Learning to create the novel protein design tools I discuss in this dissertation including ProteinMPNN, and several diffusion-based tools, not to speak of the deep network hallucination method which directly requires the inputs of structure determination algorithms like AF2.

The Future of ML in Protein Design

As the field of protein design continues to evolve, diffusion and one-shot protein design methods may become more prevalent. Integration into unified toolboxes like Rosetta can provide a comprehensive resource for researchers.

ML has the potential to play a significant role in protein dynamics, active site design, and other applications, but this will depend on the availability and interpretability of data. Improving the

interpretability of ML protein models will not only enhance their utility but also contribute to a deeper understanding of proteins themselves.

In the long term, the goal should be to achieve a complete understanding of how proteins function. This would reduce the reliance on big data-trained deep learning models and facilitate more targeted, accurate, and efficient protein design.

Bibliography

1. ANFINSEN, C. B., HABER, E., SELA, M., and WHITE, F. H. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A*. **47**, 1309–1314
2. Higgins, M. K. (2021) Can We AlphaFold Our Way Out of the Next Pandemic? *J Mol Biol*. **433**, 167093
3. Delgado-Cunningham, K., López, T., Khatib, F., Arias, C. F., and DuBois, R. M. (2022) Structure of the divergent human astrovirus MLB capsid spike. *Structure*. **30**, 1573-1581.e3
4. Praetorius, F., Leung, P. J. Y., Tessmer, M. H., Broerman, A., Demakis, C., Dishman, A. F., Pillai, A., Idris, A., Juergens, D., Dauparas, J., Li, X., Levine, P. M., Lamb, M., Ballard, R. K., Gerben, S. R., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., Volkman, B. F., Nivala, J., Stoll, S., and Baker, D. Design of stimulus-responsive two-state hinge proteins. *biorxiv.org*. 10.1101/2023.01.27.525968
5. Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*. **23**, ii–iv
6. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021) Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*. **89**, 1607–1617
7. Adhikari, B., Hou, J., and Cheng, J. (2018) Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. *Proteins*. **86**, 84
8. Coluzza, I. (2017) Computational protein design: a review. *J Phys Condens Matter*. 10.1088/1361-648X/AA5C76
9. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011) Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol*. **487**, 545–574
10. Pereira, J. M., Vieira, M., and Santos, S. M. (2021) Step-by-step design of proteins for small molecule interaction: A review on recent milestones. *Protein Science*. **30**, 1502–1520
11. Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., Aprahamian, M., Baker, D., Barlow, K. A., Barth, P., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., Correia, B. E., Coventry, B., Das, R., De Jong, R. M., DiMaio, F., Dsilva, L., Dunbrack, R., Ford, A. S., Frenz, B., Fu, D. Y., Geniesse, C., Goldschmidt, L., Gowthaman, R., Gray, J. J., Gront, D., Guffy, S., Horowitz, S., Huang, P. S., Huber, T., Jacobs, T. M., Jeliazkov, J. R., Johnson, D. K., Kappel, K., Karanicolas, J., Khakzad, H., Khar, K. R., Khare, S. D., Khatib, F., Khramushin, A., King, I. C., Kleffner, R., Koepnick, B., Kortemme, T., Kuenze, G., Kuhlman, B., Kuroda, D., Labonte, J. W., Lai, J. K., Lapidoth, G., Leaver-Fay, A., Lindert, S., Linsky, T., London, N., Lubin, J. H., Lyskov, S., Maguire, J., Malmström, L., Marcos, E., Marcu, O., Marze, N. A., Meiler, J., Moretti, R., Mulligan, V. K., Nerli, S., Norn, C., Ó'Conchúir, S., Ollikainen, N., Ovchinnikov, S., Pacella, M. S., Pan, X., Park, H., Pavlovicz, R. E., Pethe, M., Pierce, B. G., Pilla, K. B., Raveh, B., Renfrew, P. D., Burman, S. S. R., Rubenstein, A., Sauer, M. F., Scheck, A., Schief, W., Schueler-Furman, O., Sedan, Y., Sevy, A. M., Sgourakis, N. G., Shi, L., Siegel, J. B., Silva, D. A., Smith, S., Song, Y., Stein, A., Szegedy, M., Teets, F. D., Thyme, S. B., Wang, R. Y. R., Watkins, A., Zimmerman, L., and Bonneau, R. (2020) Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods 2020 17:7*. **17**, 665–680
12. Tobi, D., and Bahar, I. (2006) Optimal design of protein docking potentials: Efficiency and limitations. *Proteins: Structure, Function, and Bioinformatics*. **62**, 970–981
13. Gershenson CGershenson, C. (2003) Artificial Neural Networks for Beginners. [online] https://arxiv.org/abs/cs/0308031v1 (Accessed April 18, 2023)
14. Hawkins, D. M. (2004) The Problem of Overfitting. *J Chem Inf Comput Sci*. **44**, 1–12

15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*. 10.1145/3236009
16. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020) Improved protein structure prediction using potentials from deep learning. *Nature 2020 577:7792*. **577**, 706–710
17. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*. **596**, 583–589
18. Marcu, Ş. B., Tăbîrcă, S., and Tangney, M. (2022) An Overview of Alphafold's Breakthrough. *Front Artif Intell*. **5**, 112
19. Xia, C., and Shen, H.-B. (2023) Deep Learning Techniques for De novo Protein Structure Prediction . *Machine Learning in Bioinformatics of Protein Sequences*. 10.1142/9789811258589_0001
20. Alquraishi, M. (2019) AlphaFold at CASP13. *Bioinformatics*. **35**, 4862–4865
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017) Attention Is All You Need. *Adv Neural Inf Process Syst*. **2017-December**, 5999–6009
22. Marcu, Ş. B., Tăbîrcă, S., and Tangney, M. (2022) An Overview of Alphafold's Breakthrough. *Front Artif Intell*. **5**, 875587
23. Ruff, K. M., and Pappu, R. V. (2021) AlphaFold and Implications for Intrinsically Disordered Proteins. *J Mol Biol*. **433**, 167208
24. Baker, D. (2019) What has de novo protein design taught us about protein folding and biophysics? *Protein Science*. **28**, 678–683
25. Pan, X., and Kortemme, T. (2021) Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*. **296**, 100558
26. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A. Dos, Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, † Alexander (2022) Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. 10.1101/2022.07.20.500902
27. Barrett, T. D., Villegas-Morcillo, A., Robinson, L., Gaujac, B., Adméte, D., Saquand, E., Beguir, K., and Flajolet, A. (2022) So ManyFolds, So Little Time: Efficient Protein Structure Prediction With pLMs and MSAs. *bioRxiv*. 10.1101/2022.10.15.511553
28. Wallner, B. (2022) AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *bioRxiv*. 10.1101/2022.12.20.521205
29. Wallner, B. (2023) AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *bioRxiv*. 10.1101/2022.12.20.521205
30. Li, Y., Zhang, C., Feng, C., Freddolino, P. L., and Zhang, Y. (2022) Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *bioRxiv*. 10.1101/2022.12.30.522296
31. Baek, M., McHugh, R., Anishchenko, I., Baker, D., and DiMaio, F. (2022) Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. *bioRxiv*. 10.1101/2022.09.09.507333
32. Instadeep, T. D. B., Villegas-Morcillo, A., Instadeep, L. R., Gaujac, B., Adméte, D., Saquand, E., Beguir, K., and Flajolet, A. (2022) So ManyFolds, So Little Time: Efficient Protein Structure Prediction With pLMs and MSAs. *bioRxiv*. 10.1101/2022.10.15.511553
33. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Dos, A., Costa, S., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., and Ai, M. Language models of protein sequences at the scale of evolution enable accurate structure prediction. 10.1101/2022.07.20.500902

34. Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J. (2022) High-resolution de novo structure prediction from primary sequence. *bioRxiv*. 10.1101/2022.07.21.500999
35. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2022) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. 10.1101/2021.10.04.463034
36. Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Science*. **31**, e4379
37. Callaway, E. (2023) After AlphaFold: protein-folding contest seeks next big breakthrough. *Nature*. **613**, 13–14
38. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Christopher Garcia, K., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science (1979)*. **373**, 871–876
39. Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., and Baker, D. (2021) De novo protein design by deep network hallucination. *Nature 2021 600:7889*. **600**, 547–552
40. Frank, C., Khoshouei, A., Stigter, Y. de, Schiewitz, D., Feng, S., Ovchinnikov, S., and Dietz, H. (2023) Efficient and scalable de novo protein design using a relaxed sequence space. *bioRxiv*. 10.1101/2023.02.24.529906
41. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. (2022) Robust deep learning–based protein sequence design using ProteinMPNN. *Science (1979)*. **378**, 49–56
42. Wang, J., Lisanza, S., Juergens, D., Tischer, D., Anishchenko, I., Baek, M., Watson, J. L., Chun, J. H., Milles, L. F., Dauparas, J., Expòsit, M., Yang, W., Saragovi, A., Ovchinnikov, S., and Baker, D. (2021) Deep learning methods for designing proteins scaffolding functional sites. *bioRxiv*. 10.1101/2021.11.10.468128
43. Tischer, D., Lisanza, S., Wang, J., Dong, R., Anishchenko, I., Milles, L. F., Ovchinnikov, S., and Baker, D. (2020) Design of proteins presenting discontinuous functional sites using deep learning. *bioRxiv*. 10.1101/2020.11.29.402743
44. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. De, Mathieu, E., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. (2022) Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*. 10.1101/2022.12.09.519842
45. Lin, Y., and AlQuraishi, M. (2023) Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds. [online] https://arxiv.org/abs/2301.12485v2 (Accessed April 18, 2023)
46. Lee, J. S., Kim, J., Kim, P. M., and Org, P. (2023) ProteinSGM: Score-based generative modeling for de novo protein design. *bioRxiv*. 10.1101/2022.07.13.499967
47. Bonadio, A., Wenig, B. L., Hockla, A., Radisky, E. S., and Shifman, J. M. (2022) A broad matrix metalloproteinase inhibitor with designed loop extension exhibits ultrahigh specificity for MMP-14. *bioRxiv*. 10.1101/2022.12.29.522231
48. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. J. (2017) The Rosetta All-

Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput*. **13**, 3031–3048
49. Das, R., and Baker, D. (2008) Macromolecular Modeling with Rosetta. *https://doi.org/10.1146/annurev.biochem.77.062906.171838*. **77**, 363–382
50. Cheng, J., Tegge, A. N., and Baldi, P. (2008) Machine Learning Methods for Protein Structure Prediction. *IEEE Rev Biomed Eng*. **1**, 41–49
51. Lee, S., Kim, S., Lee, G. R., Kwon, S., Woo, H., Seok, C., and Park, H. (2023) Evaluating GPCR modeling and docking strategies in the era of deep learning-based protein structure prediction. *Comput Struct Biotechnol J*. **21**, 158–167
52. Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2021) A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology 2021 23:1*. **23**, 40–55
53. Kuhlman, B., and Bradley, P. (2019) Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology 2019 20:11*. **20**, 681–697
54. Protein, E., Peslherbe, G., Selvaraj, G., Wang, Y., Ibrahim Omar, S., Keasar, C., Ben-Sasson, A. J., and Haber, E. (2023) Protein Design Using Physics Informed Neural Networks. *Biomolecules 2023, Vol. 13, Page 457*. **13**, 457
55. Jones, D. T., and Thornton, J. M. (2022) The impact of AlphaFold2 one year on. *Nat Methods*. **19**, 15–20
56. Yeh, A., Norn, C., Kipnis, Y., Tischer, D., Nature, S. P.-, and 2023, undefined De novo design of luciferases using deep learning. *nature.com*. [online] https://www.nature.com/articles/s41586-023-05696-3 (Accessed April 18, 2023)
57. Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., Anishchenko, I., Baker, D., and Yang, J. (2021) The trRosetta server for fast and accurate protein structure prediction. *Nature Protocols 2021 16:12*. **16**, 5634–5651
58. Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., Strauch, E. M., DeBouver, N. D., Pires, A., Bera, A. K., Halabiya, S., Hammerson, B., Yang, W., Bernard, S., Stewart, L., Wilson, I. A., Ruohola-Baker, H., Schlessinger, J., Lee, S., Savvides, S. N., Garcia, K. C., and Baker, D. (2022) Design of protein-binding proteins from the target structure alone. *Nature 2022 605:7910*. **605**, 551–560
59. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020) Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A*. **117**, 1496–1503
60. Bonadio, A., Wenig, B. L., Hockla, A., Radisky, E. S., and Shifman, J. M. A broad matrix metalloproteinase inhibitor with designed loop extension exhibits ultrahigh specificity for MMP-14. *biorxiv.org*. 10.1101/2022.12.29.522231
61. Takanishi, C. L., Bykova, E. A., Cheng, W., and Zheng, J. (2006) GFP-based FRET analysis in live cells. *Brain Res*. **1091**, 132–139
62. Fallas, J. A., Ueda, G., Sheffler, W., Nguyen, V., McNamara, D. E., Sankaran, B., Pereira, J. H., Parmeggiani, F., Brunette, T. J., Cascio, D., Yeates, T. R., Zwart, P., and Baker, D. (2016) Computational design of self-assembling cyclic protein homo-oligomers. *Nature Chemistry 2016 9:4*. **9**, 353–360
63. Elofsson, A. (2022) Protein Structure Prediction until CASP15. [online] https://arxiv.org/abs/2212.07702v1 (Accessed April 18, 2023)
64. Moffat, L., Kandathil, S. M., and Jones, D. T. (2022) Design in the DARK: Learning Deep Generative Models for De Novo Protein Design. *bioRxiv*. 10.1101/2022.01.27.478087
65. Gillis, J., and Pavlidis, P. (2013) Characterizing the state of the art in the computational assignment of gene function: Lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinformatics*. **14**, 1–12
66. Janin, J., Henrick, K., Moult, J., Eyck, L. Ten, Sternberg, M. J. E., Vajda, S., Vakser, I., and Wodak, S. J. (2003) CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics*. **52**, 2–9

67. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Dos, A., Costa, S., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., and Ai, M. Language models of protein sequences at the scale of evolution enable accurate structure prediction. 10.1101/2022.07.20.500902