

LEVERAGING DEEP GENERATIVE MODEL FOR COMPUTATIONAL PROTEIN
DESIGN AND OPTIMIZATION
BY
BOQIAO LAI

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

AT THE
TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO
CHICAGO, ILLINOIS
AUGUST, 2024

THESIS COMMITTEE:

Jinbo Xu(Chair)

Avrim Blum

Aly Azeem Khan

Copyright © 2024 by Boqiao Lai

All Rights Reserved

“I have approximate answers and possible beliefs in different degrees of certainty about different things, but I’m not absolutely sure of anything.”

-Richard Feynman

TABLE OF CONTENTS

LIST OF FIGURES	vi
LIST OF TABLES	xv
ACKNOWLEDGMENTS	xvi
ABSTRACT	xviii
1 BACKGROUND	1
1.1 Proteins	1
1.2 Computational Protein Design	2
1.3 Machine learning	10
2 END-TO-END DEEP STRUCTURE GENERATIVE MODEL FOR PROTEIN DESIGN AND OPTIMIZATION	18
2.1 Motivation	18
2.2 Abstract	19
2.3 Introduction	20
2.4 Literature Review	22
2.4.1 Deep Generative models	22
2.4.2 Protein structure generative models	23
2.4.3 Fixed backbone sequence design	24
2.5 Methods	25
2.5.1 Protein representations	26
2.5.2 Model Architecture	28
2.5.3 Data	31
2.6 Results	32
2.6.1 End-to-End structure reconstruction & generation	32
2.6.2 Ablation study on input features	34
2.6.3 Conformational decoy sampling for robust protein sequence design	35
2.6.4 Unconditional structure inpainting for antibody design	41
2.7 Discussion	43
3 ADAPTIVE <i>DE NOVO</i> PROTEIN DESIGN VIA ITERATIVE SEQUENCE STRUCTURE CO-OPTIMIZATION	47
3.1 Motivation	47
3.2 Introduction	48
3.3 Literature Review	50
3.3.1 Structure based protein design	51
3.3.2 Unconditional structure generation	53
3.3.3 Computational enzyme and small-molecule binder design	54
3.3.4 Protein-protein binder design and motif scaffolding	55

3.3.5	Structure based antibody design	56
3.4	Methods	58
3.4.1	Overall Approach	58
3.4.2	Deep structural generative models	61
3.4.3	in silico validation and sequence design	61
3.4.4	Experimental Validation Methods	62
3.4.5	Iterative optimization algorithms	64
3.4.6	Data	67
3.5	Results	68
3.5.1	Unconditional Structure Generation	69
3.5.2	De novo design of DFHBI activated fluorescent protein	73
3.5.3	de novo protein binder design via motif grounded scaffolding	80
3.5.4	Structure based Antibody design via conditional CDR inpainting	83
3.6	Discussion	88
4	CONCLUDING REMARKS	92
	REFERENCES	94

LIST OF FIGURES

1.1	Illustration of the protein structure	2
1.2	Illustration of in vitro directed evolution adopted from [151] CC-BY 4.0	3
1.3	Illustration of different type of protein design tasks adopted from [43] CC-BY 4.0. a) Fixed backbone sequence design. b) Structure generation via constraints. c) Direct sequence generation from sequence-only models. d) Sequence-Structure co-design models.	4
1.4	Methods for <i>de novo</i> protein backbone generation adopted from [127] CC-BY 4.0. A) Assembly based backbone generation methods, the top shows examples of blue print guided short fragment assembly design. The middle shows an exam- ples of designs leveraging modular structure motifs such as Leucine-rich-motifs. The bottom shows fragment assembly using substructure graphs. B) Rational design methods such as TopoBuilder that leverages expert curated blueprints. C) Symmetric Protein design with repeats and symmetrical fold elements and repeat fragments. D) Family based fold design with family-wide geometry sam- pling. E) Machine learning methods generate <i>de novo</i> protein structures and through various neural network architectures.	5
1.5	Protein diffusion models illustration adopted from [187] with CC-BY-NC 4.0. The top plot shows diffusion process for protein backbone generation, the bottom plot shows the diffusion process for protein ligand docking.	8
1.6	Illustration of graphical representation of protein structures GNNs.	12
1.7	Probabilistic graphical model for Variational Autoencoder	15

2.1	Model Overview. Structures are first distilled into pairwise (inter-residue distance & orientation) and scalar (torsion angles & amino acid sequence) features, Then our model separately encodes the pairwise and scalar features into the latent space. The latent representation is then sampled and decoded by the pair and scalar feature decoder before being fed into a graph-transformer based structure module for coordinate generation. A locally aligned loss is then computed between the reconstructed coordinates and the native input structure.	26
2.2	(A) Examples of reconstructed backbone structures of different folds and sizes. 1WTE (top), 3ZIJ (bottom left), 2JAC (bottom right). (B) LDDT (blue) and TM-score (purple) against native structure vs. the length of the test target. (C) Ramachandran plot of the native (red) and reconstructed(blue) structures. (D) t-SNE embedding of the latent space for the test targets colored by their CATH class annotation.	33
2.3	Outline of robust protein design procedure using conformational decoy library. Starting with a primary backbone structure target, our structure generative model embeds it into a latent space and decodes into conformational decoys to form a structure library. Using a fixed backbone sequence design program on the backbone structure library, one can obtain a preliminary sequence library. To filter and structurally validate the primary sequence library, a structure prediction oracle is used and the validated sequence library is ready for further downstream tasks.	36

2.4 Examples of protein designed from conformational decoys. (A) AF2 folded sequence designed from conformational decoys. Overlay of AF2 predicted structure of decoy designed sequence & native backbone designed sequence (green & blue), native backbone(red) for PDB:1VF6 (top left), PDB:2JUA (top right), PDB:3G67 (bottom). (B) Scatter plots of TM-scores between the best folded conformational decoy designed sequences and sequences designed from the native backbones(bottom). Δ TM-score between decoy designed sequences and fixed backbone designed sequences versus the target size.(top)	37
2.5 (A) Examples of AF2 predicted designed sequences and native sequences. (B) TM-score distributions of decoy designed sequences vs. noisy backbone designed sequences.	38
2.6 A) Example of backbone ensemble sampling vs. sequence design sampling of PDB:1JDI. B) Self-consistent TM score comparison between backbone ensemble sampling and single shot native backbone design(Top). Design success rate across different sc-TM threshold comparison(Bottom). C) pLDDT of designed sequence comparison between conformer sampling vs. single shot native backbone. D) sc-TM distribution vs. sequence identity distribution of the conformer based sequence design(Left). sc-TM distribution vs. sequence identity distribution of the native backbone sequence design(Right)	39

2.7 Examples of antibody structure inpainting. overlay of the CDR-H3 region of PDB:5ILA (bottom left) native structure(green), unmasked region (yellow) and masked inpainting region (red) reconstruction with RMSD = 1.021. overlay of the CDR-H2 region of PDB:6VY2 (bottom right) native structure(green), unmasked region (yellow) and masked inpainting region (red) reconstruction with RMSD = 0.689. overlay of the CDR-H1 region of PDB:7D4G (top) native structure(green), unmasked region (yellow) and masked inpainting region (red) reconstruction with RMSD = 0.782	42
3.1 Structure-based protein design workflow illustration adopted from [125] CC-BY 4.0. Comparison between conventional structure-based protein design workflow vs. DL-based design workflow. For structure generation, the conventional approaches leverages existing structure fragments or functional motifs in the structure database and along with experts' knowledge to build the structure blueprint for subsequent design steps while DL-based approaches uses neural networks trained on the vast structure database that can generate structure templates. For sequence optimization, conventional approaches uses physics based method with energy minimization, DL-based methods use structure conditioned machine learning models to predict the amino acid sequences. For design scoring, conventional methods use energy based simulation such as molecular dynamics simulation or Rosetta energy to select viable candidates. DL-based workflow use <i>in silico</i> structure and property prediction models to evaluate the fitness of design candidates.	52

3.2 Enzyme engineering workflow illustration adopted from [108]. a) Conventional enzyme engineering workflow. Natural protein scaffold with desired structure and function is picked and fitness is optimized via directed evolution. b) Computational <i>de novo</i> enzyme design which starts by selecting or building suitable protein scaffold from scratch via generative models or simulation based filtering, then <i>in silico</i> design and scoring scheme is employed to produce design candidates for downstream validation.	55
3.3 Structure based antibody design workflow illustration adopted from [72] CC-BY 4.0. Workflow of <i>in silico</i> structure based antibody design, the pipeline starts with a antibody framework of choice and the target antigen. A generative model or CDR minding can be used to initialize the seed AB candidates. Computational structure modeling tools and antibody docking tools will then be used to evaluate the designed complex <i>in silico</i> . A scoring function or binding affinity prediction tool will then be applied to filter the design candidates. The best candidates with a given metric will be presented as the resulting design. This process can be iterative and the best candidate can be fed back into the pipeline for further optimization and design.	57
3.4 Overview of iterative design pipeline for adaptive protein design and optimization. The pipeline can be dissected into three modules, the design module which performs structure-sequence co-optimization on the input structure templates. The resulting structurally validated library is then fed into the scoring module which can include various <i>in silico</i> simulation and fitness prediction tools. The resulting functionally validated candidates are then selected as templates for next design iteration or further experimental validation.	60

3.5	Unconditional Generation of monomeric structures. A) Illustration of iterative structure evolution by design iterations. t=0 indicates random initialization of amino acid sequences, t=T indicates design convergence. B) Example of generated <i>de novo</i> structures, helical structures are colored cyan and beta strands are colored purple. C) Conceptual graph of the iterative design framework.	70
3.6	Unconditional Generation of novel protein structures. A) Example pLddT and pTM evolution with design iterations. C) Column 1, <i>de novo</i> generated protein folds. Column 2, initial inter-residue distance map of randomly initialized sequences. Column 3, inter-residue distance map for the <i>de novo</i> generated protein folds. Column 4 overlay of <i>de novo</i> generated protein folds with its closest hit from the PDB. In this case, we found designs from our iterative design framework to exhibit both high <i>in silico</i> folding viability and structural novelty.	71
3.7	De novo design of DFHBI-activated fluorescent β-barrels A) PDB:6CZH previously designed DFHBI-activated β -barrel. B) Distribution of ligand pose RMSD(Left) of the <i>de novo</i> designed β -barrels with starting template 6czh.A and the first design batch BB1.A and the second design batch BB.2. Distribution of computational docking scores from Autodock Vina(Right) of the <i>de novo</i> designed β -barrels with starting template 6czh.A and the first design batch BB1.A and the second design batch BB.2 C) Example overlay of the computationally docked DFHBI with the designed β -barrels. D) Scatter plot of the computationally docked scores vs. the ligand RMSD	74

3.8 **Experimental validation of *de novo* designed DFHBI-activated fluorescent β -barrels** A) Fluorescence microscope image of *de novo* designed protein(Left) and the design template mFAP0. B) Fluorescent emission spectra from the lysates of 20 tested designs and the templates with reference buffer. All of the 20 designs were found with detectable fluorescent emission which represents a 100% design success rate. C) Fluorescent emission spectra of purified designed proteins and their respective original templates of 6czi(up) and 6czh(down). Although we did not observe increased fluorescent peak, we saw slight shift in the emission frequency compared to the design template. D) SDS-PAGE of purified designed protein with its respective design templates. All the designed proteins are observed to be smaller than the reference template. E) Protein thermal stability assay results(top), and protein yield comparison results(bottom). We see improvement of thermal stability on 6czh-based designs and comparable thermal performance in 6czi-based designs. For protein yield, all of the 6czh-based designs exhibited improvement over the design reference and 6czi-V has a significant increase over the reference.

76

3.9 Iterative design workflow for motif grounded protein scaffolding: The design process begins with the identification of the functional motif interest, this can be done by extracting protein-protein binding surfaces or enzyme active sites. Then a randomly initialized scaffold is used as the seed template. The iterative design algorithm is then applied with a motif grounded structure fitness criterion. 80

LIST OF TABLES

2.1	Average structural similarity metrics evaluated on the test targets over different models	34
2.2	Average short(S), medium(M), long(L) contact accuracy evaluated on the 1,120 test targets over different models	34
2.3	Structure inpainting performance on the test monoclonal antibody dataset in CDR-H1(left), CDR-H2(middle), CDR-H3(right) across different models.	42
3.1	Structure inpainting performance on the test monoclonal antibody dataset in CDR-H1(left), CDR-H2(middle), CDR-H3(right) across different models.	86

ACKNOWLEDGMENTS

The completion of this thesis marks the end of a challenging yet rewarding journey, one that would not have been possible without the support and guidance of many individuals.

First and foremost, I extend my sincere gratitude to my advisor, Dr. Jinbo Xu. Your insightful guidance, unwavering support, and intellectual rigor have been instrumental in shaping both this work and my growth as a researcher. Your ability to push me beyond my perceived limits while offering patience and encouragement has been truly invaluable.

To my committee members, Dr. Aly Khan and Dr. Avrim Blum: your expertise, constructive feedback, and challenging questions have significantly elevated the quality of this research. I am deeply appreciative of the time and effort you have invested in my academic development.

I owe a debt of gratitude to my collaborators, Matthew McPartlon and Hugh Yeh. Your contributions, innovative ideas, and our spirited discussions have enriched this work immeasurably. The synergy of our collaboration has been a highlight of this research process.

On a personal note, I want to express my heartfelt thanks to my fiancée, Cindy Zhang. Your unwavering support, understanding, and love have been my anchor throughout this journey. Your belief in me, especially during the most challenging times, has been a source of strength and motivation.

To my parents: your unconditional love, countless sacrifices, and constant encouragement have been the foundation upon which all my achievements stand. Your support has allowed me to pursue my dreams, and for that, I am eternally grateful.

I would be remiss not to mention Hibiki and Cooper, whose feline companionship during long hours of research and writing provided a sense of comfort and routine.

Lastly, I want to acknowledge the friendship and support of Gavin Young, Tom Li, Cathy Zhang, and Vijay Pillai. Your camaraderie, encouragement, and ability to provide perspective have been crucial in maintaining balance throughout this academic pursuit.

To everyone mentioned here, and to those whose names may not appear but whose impact has been felt: your collective support, wisdom, and encouragement have not only made this thesis possible but have also made the journey profoundly meaningful. Thank you.

ABSTRACT

Proteins are the fundamental macromolecules that play diverse and crucial roles in all living matter and have tremendous implications in healthcare, manufacturing, and biotechnology. Their functions are largely determined by the sequences of amino acids that compose them and their unique three-dimensional structures when folded. The recent surge in highly accurate computational protein structure prediction tools has equipped scientists with the means to derive preliminary structural insights without the onerous costs of experimental structure determination. These breakthroughs hold profound promise for building robust and efficient *in silico* protein design systems.

While the prospect of designing *de novo* proteins with precise computational accuracy remains a grand challenge in biochemical engineering, conventional assembly-based and rational design methods often grapple with the expansive design space, resulting in suboptimal design success rates. Despite recently emerged deep learning-based models have shown promise in improving the efficiency of the computational protein design process, a significant gap persists between current design paradigms and their experimental realization. This thesis will investigate the potential of deep generative models in refining protein structure and sequence design methods, aiming to develop frameworks capable of crafting novel protein sequences with predetermined structures or specific functionalities. By harnessing extensive protein databases and cutting-edge neural architectures, this research aims to enhance precision and robustness in current protein design paradigms, potentially paving the way for advancements across various scientific fields.

The thesis is structured into three main Sections. The first section provides a comprehensive background on computational protein design, highlighting the current challenges faced by the scientific community. It then introduces the machine learning techniques that have been developed to address these challenges, focusing on those particularly relevant to our research. This section aims to establish the foundation necessary for understanding the

novel approaches presented in the subsequent sections. The second section introduces a deep structure generative model capable of producing ensembles of high-quality structural variants. We demonstrate how these structure ensembles can facilitate robust and diverse *de novo* protein design pipelines. This section showcases the power of our approach in expanding the possibilities of computational protein design beyond traditional methods. The third section presents an iterative design paradigm that leverages the models described in Section 2. We illustrate the versatility and effectiveness of this paradigm through its application to a diverse array of protein design and engineering tasks. These applications include unconditional *de novo* structure generation, demonstrating our ability to expand current protein fold spaces. We also explore the design and optimization of a small-molecule activated fluorescent protein system, where we show improved *in vitro* fitness and stability in the designed proteins. Furthermore, we present a motif-grounded protein binder design, transforming the important therapeutic target PD1 into distinct *de novo* scaffolds with potential for enhanced stability. Lastly, we demonstrate structure-based *de novo* CDR design for antibody engineering, utilizing an enhanced structure generative model optimized specifically for antibody design.

Through these applications, we demonstrate the broad utility and significant advancements our approach brings to the field of computational protein design. This thesis aims to contribute to the ongoing efforts to expand the capabilities of protein engineering, potentially opening new avenues for therapeutic development and biotechnological applications.

CHAPTER 1

BACKGROUND

1.1 Proteins

Proteins are macromolecules that play vital roles in most biochemical processes. The central dogma of molecular biology established that proteins are the functional endpoint of the genetic code, made up of amino acids that carry out the biochemical interaction among relevant molecules [30]. Proteins perform their function mostly when folded into their three dimensional structures. Figure 1.1 showcased the four levels of proteins structures where the primary structure is the linear amino acid sequence in the protein chain; the secondary structure is the repeating local structure stabilized by hydrogen bonds where the two main types are alpha helices and beta sheets; the tertiary structure is the overall three dimension structure of a single protein chain; the quaternary structure is the global arrangement of multiple proteins chains in a multi-unit protein complex.

The function of a protein is ultimately linked to its structure. Specific structural motifs often correspond to particular functions, and changes in structure (due to mutations or environmental factors) can significantly impact a protein’s function [132]. While often depicted as static structures, proteins are dynamic molecules. They can undergo conformational changes that are often crucial to their function, such as in allosteric regulation or enzyme-substrate interactions[62]. One of the main motivation behind this project is to explore the conformational structure space with deep generative models and how to use these model to improve current protein design paradigms. Understanding protein structure and function is fundamental to many areas of biochemical research, from basic molecular biology to drug design and biotechnology. As our ability to determine and predict protein structures improves, so does our capacity to understand and manipulate biological systems at the molecular level.

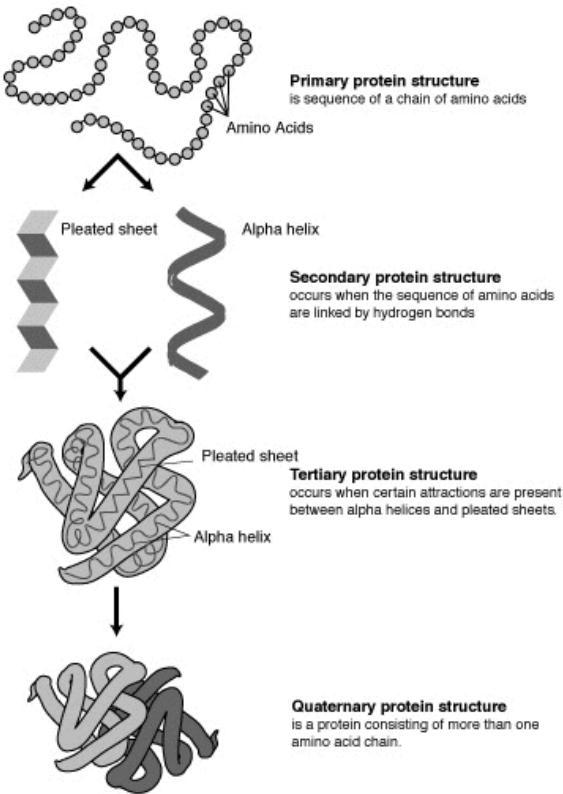


Figure 1.1: Illustration of the protein structure

1.2 Computational Protein Design

Protein molecules are crucial in most biochemical processes, therefore, designing proteins with desired function and structure is of high interest to a wide range of scientists in the scientific community. Thus, the field presents a unique and significant challenge at the intersection of biochemistry, biophysics, and computer science [70]. Thanks to the rapid advances in the deep learning and machine learning field, computational scientists can leverage these techniques to accelerate the field as a whole.

Overall, the fundamental goal of protein design is to determine an amino acid sequence that will fold into a desired structure and perform a specific function. For most of the protein sequences, the combinatorial space of all the possible sequence composition is too vast to explore exhaustively. Conventional approaches to protein design and engineering includes

methods like rational design and directed evolution[147, 10]. Rational design aims to make specific and targeted changes to a protein by chemist to achieve the desired change in protein function. This approach is often labor intensive and low in success rate. Directed evolution designs protein by mimicking natural evolution with artificial selection pressures and does not require detailed understanding of the structure of the protein of interest. While this is a very powerful tool for protein engineering, there are many technical limitations such as limited sequence space exploration and reliance on *in vitro* selection techniques.

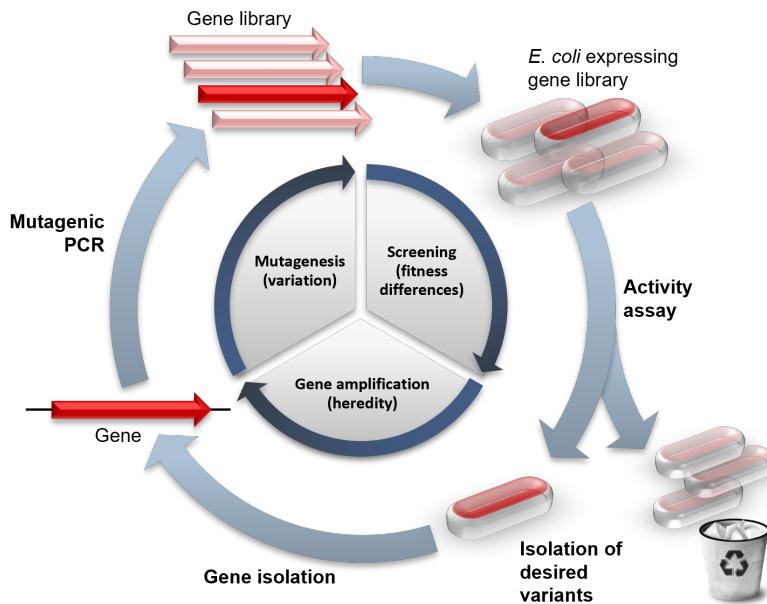


Figure 1.2: Illustration of in vitro directed evolution adopted from [151] CC-BY 4.0

Computational methods have become increasingly important in protein design due to their ability to explore vast sequence spaces efficiently. Central to early computational protein design is the use of energy functions to evaluate the stability and potential functionality of designed sequences [5]. These energy functions typically account for various interactions including van der Waals forces, electrostatic interactions, hydrogen bonding, and solvation effects. However, the folding of a protein is governed by numerous weak interactions, resulting in complex energy landscapes which are still challenging to explore efficiently *in silico*. To reduce the computational complexity, most methods use discrete side-chain conformations

called rotamers. This approach significantly reduces the search space while still capturing the essential features of side-chain packing. Various algorithms are employed to search the sequence space, including Monte Carlo methods, genetic algorithms, dead-end elimination, and integer linear programming. Each of these approaches has its strengths and is suited to different aspects of the design problem[47].

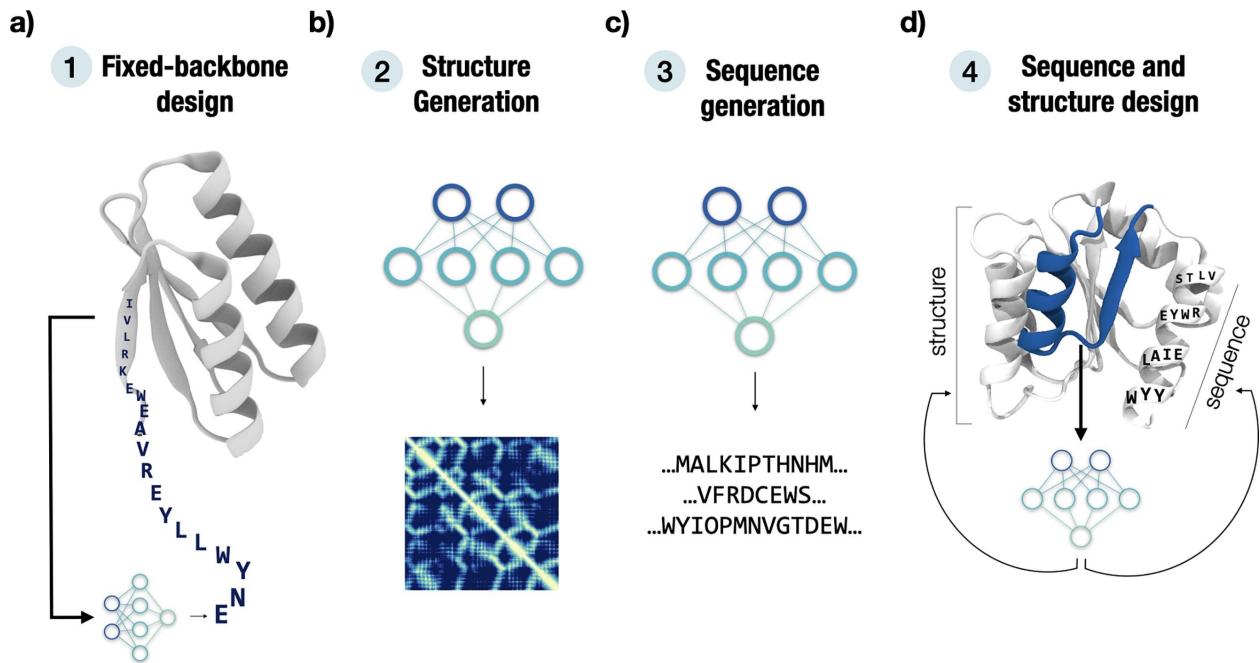


Figure 1.3: Illustration of different type of protein design tasks adopted from [43] CC-BY 4.0. a) Fixed backbone sequence design. b) Structure generation via constraints. c) Direct sequence generation from sequence-only models. d) Sequence-Structure co-design models.

The field of computational protein design has seen significant progress in recent years. For fixed backbone sequence design which seeks to recover the amino acid sequences that conform to the given backbone, methods such as 3D-CNN and geometric graph neural networks has shown promising successes [8, 137, 80, 120, 33, 68] by leveraging the graphical and three dimensional representations of the backbone structures. For structure generation, methods such as the Generative adversarial network(GANs) and diffusion models are widely used to recover the either the two dimensional distance or orientation map [7, 41, 180]. More recent diffusion based methods can also generate the backbone structure directly [165, 177].

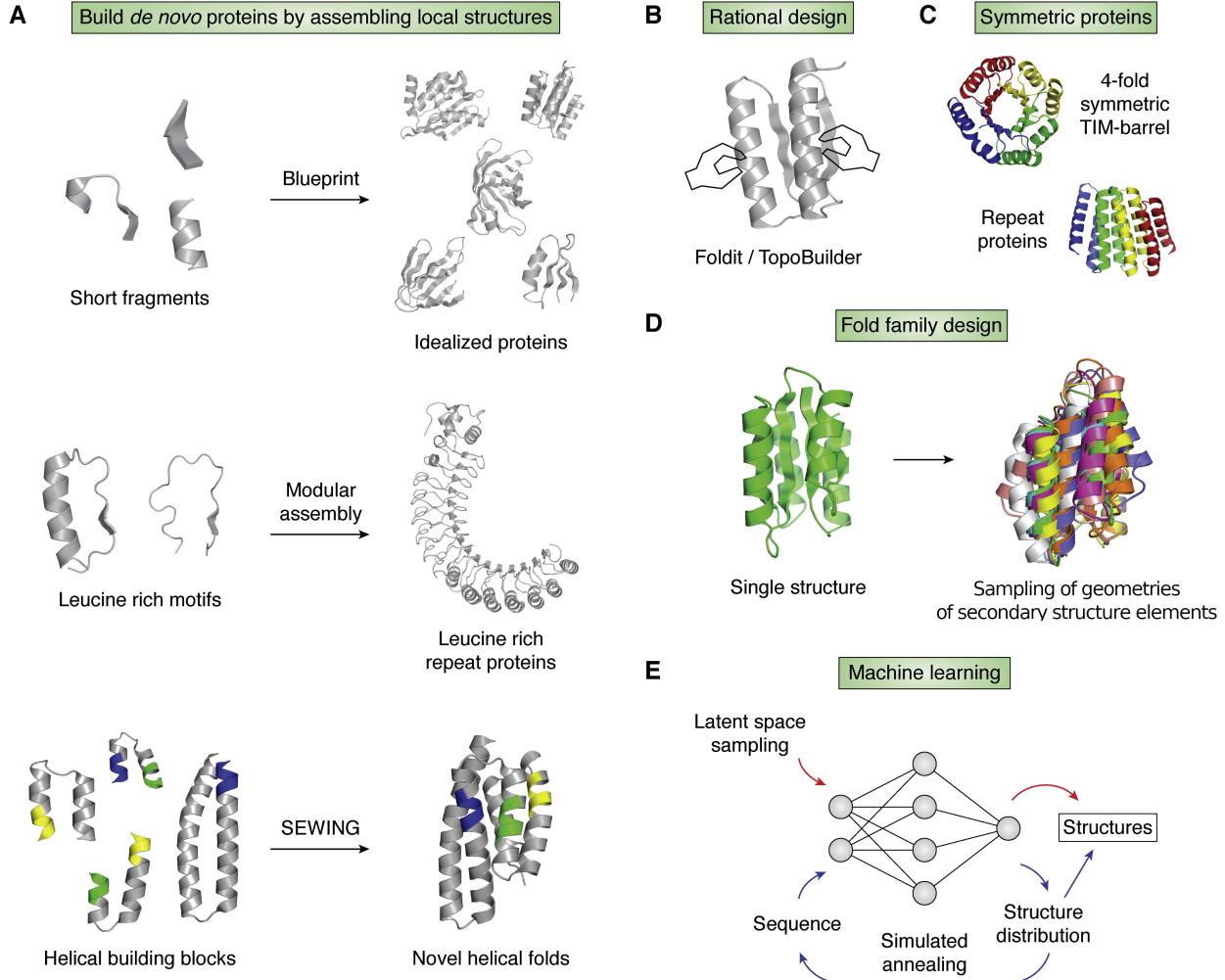


Figure 1.4: Methods for *de novo* protein backbone generation adopted from [127] CC-BY 4.0. A) Assembly based backbone generation methods, the top shows examples of blue print guided short fragment assembly design. The middle shows an examples of designs leveraging modular structure motifs such as Leucine-rich-motifs. The bottom shows fragment assembly using substructure graphs. B) Rational design methods such as TopoBuilder that leverages expert curated blueprints. C) Symmetric Protein design with repeats and symmetrical fold elements and repeat fragments. D) Family based fold design with family-wide geometry sampling. E) Machine learning methods generate *de novo* protein structures and through various neural network architectures.

The success of *de novo* design of proteins heavily relies on the ability to generate high-quality designable protein backbone templates[70, 92, 127], various strategies have been successfully employed for backbone structure generation which includes:

Structure variations: Redesigning existing backbone structures for new functions and

simulate the backbone movement through molecular dynamics simulations.[78, 163]. The disadvantage of this approach is apparent such that the process of generating variants with compute intensive methods such as molecular dynamic simulation limited its applicability and stability. In addition, the accuracy and reliability of fast timescale simulation may not be adequate for the design purpose and therefore lead to optimal results.

Family based design: Fold family specific *de novo* protein design such as helical bundles[63, 128, 12]. This approach designs protein with target function by adopting specific folds with desire functional structure motifs such as small molecule binding pockets where rational or computational fine-tunning of the active sites to achieve similar but novel function such as small molecule binders[136], ion transport proteins[82], and protein switches[97]. While this approach is powerful in designing functional variants that are neighboring to existing proteins, however, the flexibility of this approach is limited and structure diversity is often not part of the design objective.

Assembly based design: Assembly based backbone generation strategies leverages existing protein structure databases to find fragments and structure motifs that fits the design blueprint and assemble them into the desired typology. The first successful design of *de novo* protein fold with assembly based methods were done over twenty years ago (Top7)[94]. Other assembly based methods leverages modular structure motifs such as leucine-rich-repeats[129] and helical blocks[76] with substructure graphs to guide the assembly process. Despite the some notable successes, assembly based methods are still challenging due to the constraints presented by the structure motifs[91, 114] and expert construction of viable blueprints[105, 115, 183] render them hard to apply to practical protein design tasks. One of the challenges for assembly based *de novo* protein design is to increase flexibility and expand the usable structure motifs.

Machine learning and deep learning based design: Machine learning based backbone generation models developed recently trained with structures from the PDB are used to

generate *de novo* protein structures. For example, a generative adversarial network(GAN) based method [7] can create protein structures by generating pairwise distance maps and decode with a pre-trained downstream neural networks. Another autoencoder base model[41] focused on immunoglobulins can generate immunoproteins by decoding latent representations. More recently, a model developed can build distance map by iterative refinement and network hallucination[9] and repurposed a structure prediction model[184] to generate three dimensional protein structures.A recent study that utilized this model demonstrated its ability to facilitate the design novel enzymes with competitive enzymatic activity[186]. On the other hand, a slew of diffusion based models [177, 104, 180, 74] emerged for structure biology tasks aimed to address challenges presented by previous methods, specifically the ability to flexibly generate diverse and high-quality protein structures. These models have demonstrated success in designing novel monomeric proteins, protein binders, and enzyme active site scaffolds [177], with some designs experimentally validated. Despite its success, there are still drawbacks from diffusion based structure models which often demand large amount of compute and hard to incorporate physical constraints into the generative process. It is also worth noting that the success of functional protein design still heavily relies on extensive laboratory screening.

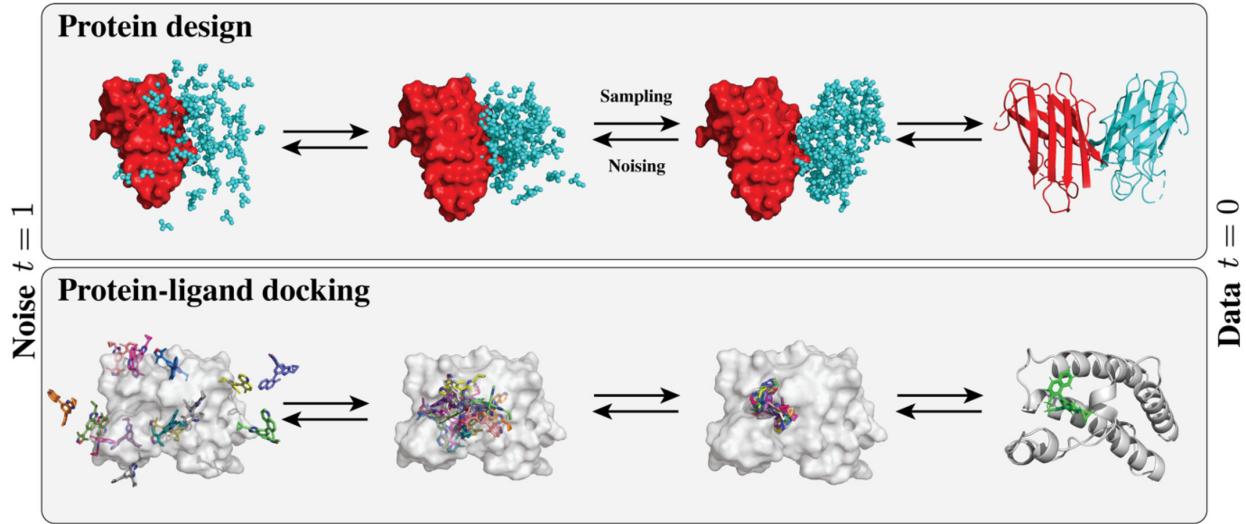


Figure 1.5: Protein diffusion models illustration adopted from [187] with CC-BY-NC 4.0. The top plot shows diffusion process for protein backbone generation, the bottom plot shows the diffusion process for protein ligand docking.

The other crucial aspect of successful computational protein design is sequence optimization also known as fixed-backbone sequence design or the inverse folding problem. The goal of this task is to find the most suitable sequence with a given backbone template such that it will fold and perform desired function in its intended environments[31, 181]. This approach is crucial for both redesigning natural proteins and selecting sequences for *de novo* backbones[70, 124, 136]. We will describe two main class of methods that are most commonly used:

Energy Function-Based Methods: These methods, developed over the past three decades, are based on well-understood physical principles and can provide insights into the physical basis of protein stability and function[31, 181, 111]. They are often computationally efficient for small to medium-sized proteins. However, they have shown relatively low success rates in experimental validation and high sensitivity to target structures[44, 148]. These methods may struggle to capture complex, long-range interactions and are limited by the accuracy of the underlying energy functions. Despite these limitations, they have been the mainstay of computational protein design for many years, with examples including Rosetta Design [98],

Proteus [158], ABACUS [181], and TERM[111].

Deep Learning Methods: More recent approaches use deep learning for sequence design, demonstrating superior performance in both computational tests and wet experiments[73, 33, 102, 8]. These methods can capture complex, non-linear relationships in protein structure and sequence, and have the ability to learn from large datasets of known protein structures. They are potentially more robust to variations in target structures. However, deep learning methods are often difficult to interpret the basis of their predictions. Additionally, they may struggle with novel protein folds not well-represented in training data[102, 120, 52]. Besides, many application specific models are developed for sequence design aimed to enhance performance in domain specific tasks such as antibody design[66, 38]. In this study, we will utilize our structure generative model and the iterative design framework in combination with a wide range of inverse folding models to improve the efficiency and robustness of computational protein design pipelines.

Structure-based *de novo* functional protein design, which involves creating proteins distinct from those exist in the nature based on biophysical principles, has achieved notable successes including the design of novel protein folds[71] and the creation of *de novo* enzymes[144, 186]. Protein redesign and optimization, which involves modifying existing proteins for enhanced stability, altered specificity, or new functions, has also seen substantial advancements [178]. Although significant effort and progress has been made towards computationally designing novel structure and sequence to achieve new functional proteins, current successes in protein design often require extensive manual input from both biochemistry and computational experts, limiting the accessibility of advanced protein design tools to the broader scientific community and hindering their ability to achieve diverse research goals. This thesis will explore and discuss novel methods for developing a more accessible computational framework, aiming to allow easier access of advanced protein design tools and enable their efficient

use for *de novo* protein design across the life science and computational communities.

1.3 Machine learning

Convolutional Neural Network

Convolutional Neural Networks (CNNs) are a type of deep learning architecture initially developed for computer vision and image processing applications [99] by adaptively learning spatial hierarchies of features from input data with specialized layers, namely convolutional layers and pooling layers designed to exploit the 2D structure of image data [100, 46]. CNNs have since become the backbone of numerous state-of-the-art models for image classification, object detection, as well as semantic segmentation[93].

The convolutional layer performs a convolution operation on the input data with learnable kernels. For a 2D input x and a filter w of size $m \times n$, the convolution operation can be expressed as:

$$(x * w)_{i,j} = \sum_{u=0}^{m-1} \sum_{v=0}^{n-1} x_{(i+u,j+v)} \cdot w_{(u,v)} \quad (1.1)$$

Where i and j are the spatial indices of the output feature map. This operation is typically followed by a non-linear activation function, such as the Rectified Linear Unit (ReLU) [55]:

$$f(x) = \max(0, x) \quad (1.2)$$

Pooling layers, often inserted between successive convolutional layers, serve to reduce the spatial dimensions of the feature maps[146] The most common pooling operation is max pooling, which can be defined as:

$$y_{ij} = \max_{(m,n) \in R_{ij}} x_{mn} \quad (1.3)$$

where R_{ij} is a local neighborhood around position (i,j) . The combination of these

operations allows CNNs to learn increasingly abstract representations of the input data as information flows through the network[189].

Besides computer vision and image processing, convolutional neural networks have also been widely successful in the domain of computational biology. In genomic sequence analysis, CNNs are powerful tools for analyzing DNA and RNA protein binding sites and profiling for non-coding variant effects [195, 96]. In protein structure prediction, RaptorX pioneered the use of CNN in protein structure prediction and AphFold later showed drastically improved structure modeling accuracy in CASP13 [176, 149]. In protein-protein interaction prediction, CNNs are used to model inter-protein interaction with sequence data.[60].

Geometric Graph Neural Network

Graph Neural Networks (GNNs) have emerged as a powerful tool for learning on graph-structured data, making them particularly relevant for problems in computational biology and chemistry, including protein design[13, 75]. Unlike conventional neural networks that operate on fixed-size inputs, GNNs can process data of arbitrary size and structure, making them well-suited for three dimensional structural data.

GNNs operate on graph data structures $G = (V, E)$, where V represents a set of nodes and E represents the edges connecting these nodes[196]. In the context of proteins, nodes might represent amino acid residues and their atomic structures, while edges could represent chemical bonds or spatial proximity[48]. GNNs work by iteratively updating node representations based on the features of neighboring nodes and edges. This process, often referred to as message passing, allows the network to capture both local and global structural information. The general form of this update can be expressed as:

$$\begin{aligned} a_v^{(k)} &= \text{AGGREGATE}^{(k)}(h_u^{(k-1)} : u \in N(v)) \\ h_v^{(k)} &= \text{UPDATE}^{(k)}(h_v^{(k-1)}, a_v^{(k)}) \end{aligned} \tag{1.4}$$

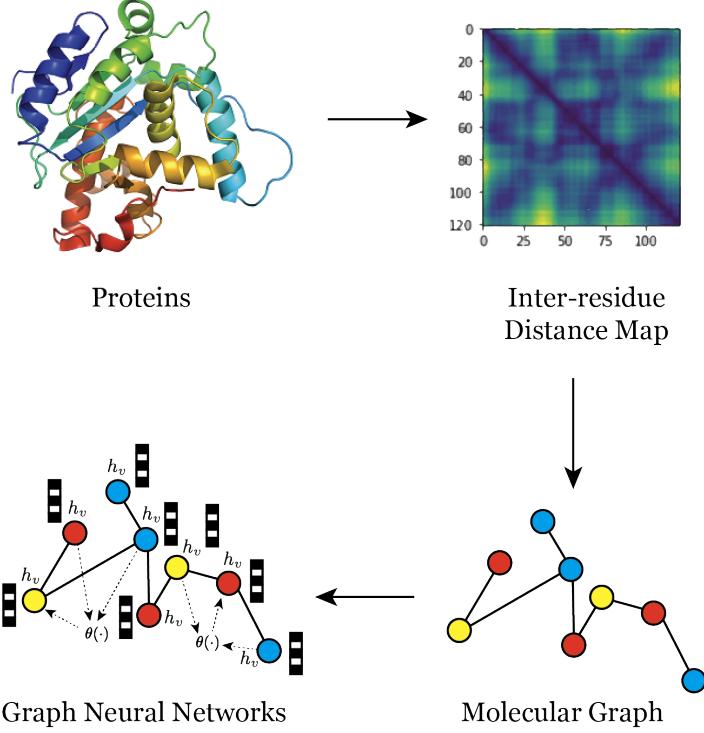


Figure 1.6: Illustration of graphical representation of protein structures GNNs.

where $h_v^{(k)}$ is the feature vector of node v at the k -th iteration, $N(v)$ is the set of neighbors of v , and AGGREGATE and UPDATE are learnable functions. Several variants of GNNs have been developed, each with unique properties. These include Graph Convolutional Networks (GCNs)[89], Graph Attention Networks (GATs)[172], and Message Passing Neural Networks (MPNNs)[53]. For instance, in the case of Graph convolution networks(GCNs) with simple neighbor aggregation:

$$h_v^{(k)} = \sigma \left(W^{(k)} \sum_{u \in N(v)} \frac{1}{|N(v)|} h_u^{(k-1)} + B^{(k)} h_v^{(k-1)} \right) \quad (1.5)$$

where $W^{(k)}$ and $B^{(k)}$ are learnable weight matrices and σ is a non-linear activation function.

In the context of protein design, nodes typically represent amino acids or atoms, while edges represent chemical bonds or spatial proximity. The node features x_v can include amino

acid properties or molecular embedding, while edge features e_{vu} could represent distances or bond types.

Graph Transformer

Graph transformer merges the structural inductive bias of Graph Neural Networks (GNNs) with the powerful attention mechanisms of Transformer models[170]. This combination allows for more expressive and flexible representations of graph-structured data, making them particularly suited for complex tasks in protein design and analysis especially for direct structural generation. The Attention architecture introduced in [170] revolutionized sequence modeling with the self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1.6)$$

where Q , K , and V are query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. To integrate the attention mechanism in graphical settings between node i, j [154]:

$$\begin{aligned} q_{(c,i)}^l &= W_{(c,q)}^l h_i^l + b_{(c,q)}^l \\ k_{(c,j)}^l &= W_{(c,k)}^l h_j^l + b_{(c,k)}^l \\ e_{(c,ij)} &= W_{c,e}^l e_{(i,j)} + b_{c,e}^l \\ \alpha_{(c,ij)}^l &= \frac{< q_{(c,i)}^l, k_{(c,j)}^l + e_{(c,ij)} >}{\sum_{u \in N(i)} < q_{(c,i)}^l, k_{(c,u)}^l + e_{(c,iu)} >} \end{aligned} \quad (1.7)$$

Where $< q, k > = \exp(\frac{q^T k}{\sqrt{d}})$ and d is the size of each head, For the $c - th$ head attention, the node features h_i^l and h_j^l are first transformed into the query vector $q_{(c,i)}^l \in \mathbb{R}^d$ and key vector $k_{(c,j)}^l \in \mathbb{R}^d$ respectively with the trainable weights $W_{(c,q)}^l, W_{(c,k)}^l, b_{(c,q)}^l, b_{(c,k)}^l$. Then the edge features e_{ij} will be encoded and add to the key vector in each layer. After computing

the multi-head attention, message aggregation was performed:

$$\begin{aligned} v_{(c,j)} &= W_{(c,v)}^l h_j^l + b_{(c,v)}^l \\ \hat{h}^{l+1} &= \|_{c=1}^C \sum_{j \in N(i)} \alpha_{(c,ij)}^l [v_{c,j}^l + e_{(c,ij)}] \end{aligned} \quad (1.8)$$

Where $\|$ concatenate for C attention heads. The node feature h_j is transformed into $v_{(c,j)} \in \mathbb{R}^d$. Together with residual connection, and non-linear transformation, the node feature update will be

$$\begin{aligned} r_i^l &= W_r^l h_i^l + b_r^l \\ \beta_i^l &= \text{Sigmoid}(W_g^l [\hat{h}^{l+1}; r_i^l; \hat{h}^{l+1} - r_i^l]) \\ h^{l+1} &= \text{ReLU}(\text{LayerNorm}((1 - \beta_i^l) \hat{h}^{l+1} + \beta_i^l r_i^l)) \end{aligned} \quad (1.9)$$

Graph Transformers represent a powerful tool for learning on protein structures, combining the structural inductive bias of graphs with the expressive power of attention mechanisms. We will use the above formulation of graph transformer in our structure generative model for direct coordinate generation.

Variational Autoencoder

Variational methods have become increasingly important in machine learning and statistical inference. They provide powerful tools for approximating complex probability distributions and learning latent representations of data. This section will provide the background for variational autoencoder used as a framework to model the protein structures.

Variational autoencoders(VAEs) also known as auto-encoding variational bayes, introduced by [88] are a type of deep generative model leverages both the autoencoder architecture and variational inference to learn a compressed latent representation of the data. In many Bayesian models, computing the exact posterior distribution is computationally infeasible. Variational Bayes addresses this by approximating the true posterior with a simpler

distribution, typically from a tractable family. The goal is to find the member of this family that is closest to the true posterior, where "closest" is measured by the Kullback-Leibler (KL) divergence. Let's consider the following simple latent variable graphical model:

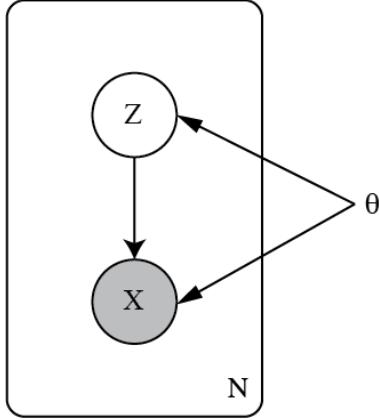


Figure 1.7: Probabilistic graphical model for Variational Autoencoder

Where θ is a set of deterministic parameters; x is our observation which can either be discrete or continuous, z is a continuous latent variable. We can write the joint distribution of the graphical model as $p(x, z; \theta) = p_\theta(z)p_\theta(x|z)$, where $p_\theta(z)$ is the prior of the latent variable and $p_\theta(x|z)$ is the likelihood function of our observation.

For inference on latent variable models, one common challenge is such that the posterior distribution of the latent variable $p_\theta(z|x)$ is often computationally intractable because it requires integration over all possible value of z .

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{\underbrace{\int_z p_\theta(x|z)p_\theta(z)dz}_{intractable}} \quad (1.10)$$

Variational inference approaches this challenge by imposing a tractable variational distribution $q_\phi(z|x)$ in place of the true posterior and minimize the Kullback–Leibler divergence between the target posterior and the variational distribution.

$$\begin{aligned}
KL(q_\phi(z|x) || p_\theta(z|x)) &= \int_z q_\phi(z|x) \log\left(\frac{q_\phi(z|x)}{p_\theta(z|x)}\right) dz \\
&= E_{z \sim q_\phi(z|x)} [\log(q_\phi(z|x)) - \log(p_\theta(z|x))] \\
&= E_{z \sim q_\phi(z|x)} [\log(q_\phi(z|x)) - \log\left(\frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}\right)] \\
&= E_{z \sim q_\phi(z|x)} [\log(q_\phi(z|x))] - (\log(p_\theta(x, z)) - \log(p_\theta(x)))
\end{aligned} \tag{1.11}$$

After rearranging the terms, we can write the log likelihood of x as:

$$\begin{aligned}
\log(p_\theta(x)) &= E_{q_\phi}[\log(p_\theta(x, z))] - E_{q_\phi}[\log(q_\phi(z|x))] + KL(q_\phi(z|x) || p_\phi(z|x)) \\
ELBO &= \log(p_\theta(x)) - \underbrace{KL(q_\phi(z|x) || p_\theta(z|x))}_{\geq 0}
\end{aligned} \tag{1.12}$$

KL divergence is non-negative according to the Jensen's inequality. The log likelihood(evidence) is then lower-bounded by $E_{q_\phi}[\log(p_\theta(x, z))] - E_{q_\phi}[\log(q_\phi(z|x))]$, which is referred to as the Evidence Lower Bound(ELBO). We can estimate the gradient of the ELBO with a naive Monte Carlo estimator, however, it is often challenging due to its high variance[88]. Another important observation is that the evidence is a constant and the KL divergence is non-negative, maximizing the ELBO can be seen as simultaneously minimizing the KL divergence between the variational distribution and the true posterior and maximizing the marginalized data likelihood. We can then rewrite the ELBO as

$$\begin{aligned}
ELBO &= E_{q_\phi}[\log(p_\theta(x, z))] - E_{q_\phi}[\log(q_\phi(z|x))] \\
&= E_{q_\phi}[\log(p_\theta(x|z)p(z))] - E_{q_\phi}[\log(q_\phi(z|x))] \\
&= E_{q_\phi}[\log(p_\theta(x|z)) + \log(p_\theta(z)) - \log(q_\phi(z|x))] \\
&= E_{q_\phi}[\log(p_\theta(x|z))] - KL(q_\phi(z|x) || p_\theta(z))
\end{aligned} \tag{1.13}$$

Here, $q_\phi(z|x), p_\theta(x|z)$ are not defined specifically. Instead of assuming specific distributions, we can take advantage of the fact that we can generate any distribution by mapping a normal distribution with sufficiently sophisticated function[36]. DNNs are used here for

their ability to approximate complicated functions and its differentiability, therefore, we can optimize it with stochastic gradient descent and back propagation. For $q_\phi(z|x)$ we define two neural encoder $Enc_\phi^\mu(x), Enc_\phi^{\sigma^2}(z|x)$ that takes input $x \in \mathcal{R}^n$ into two vectors $\mu, \sigma^2 \in \mathcal{R}^k$ that encodes the mean and variance of the Gaussian distribution of our latent variable z ; this model is also called as a *recognition model*. For $p_\theta(x|z)$ we define a neuro decoder $Dec_\theta(z)$ that takes a latent vector $z \in \mathcal{R}^k$ and output a reconstructed input $\hat{x} \in \mathcal{R}^n$, this is also called as a *generative model*.

$$\begin{aligned} q_\phi(z|x) &= \mathcal{N}(Enc_\phi^\mu(x), diag(Enc_\phi^{\sigma^2}(x))) \\ \hat{x} &= Dec_\theta(z) \quad z \sim q_\phi(z|x) \end{aligned} \tag{1.14}$$

The optimization objective of VAE can then be written as

$$l(x, \theta, \phi) = \underbrace{-E_{q_\phi(z|x)}[\log(p_\theta(x|z))]}_{\text{Reconstruction loss}} + \underbrace{KL(q_\phi(z|x)||p_\theta(z))}_{\text{Regularizer}} \tag{1.15}$$

We use VAE as a crucial component to build the deep generative models for protein structures and downstream protein design tasks.

CHAPTER 2

END-TO-END DEEP STRUCTURE GENERATIVE MODEL FOR PROTEIN DESIGN AND OPTIMIZATION

2.1 Motivation

Computationally designing protein systems has long been a challenging problem due to the complexity of protein structures and interactions among various components within the system. Successful *in silico* protein design demands both accurate modeling of the system of interest and effective structure and sequence generation to accommodate the desired functional properties. Though promising progress has been made towards more accurate design methods [33, 186, 165, 84], the gap between computational design and experimental realization remains significant.

A protein design project often commences with a predefined biochemical objective, such as enhancing enzyme activity, devising binders for specific target molecules, or creating interaction systems involving multiple protein entities. The first step involves the identification of an initial structural template, which can be achieved through two primary methods: a database search aimed at locating existing proteins sharing a similar biochemical function or the *de novo* approach achieved through *in silico* backbone generation targeting specific functional objectives. Following the initial structure template determination, the subsequent phase involves assembling the appropriate amino acid sequence given the predetermined backbone template, also known as inverse protein folding. This process entails the search of amino acid sequences that possess a high likelihood of folding into the designated template structure. Inverse protein folding is often accomplished using deep learning models or energy-based sequence optimization methods. The final phase involves further *in silico* selection pipelines that sift through the generated candidates to identify the most promising

ones for downstream experiments. [92, 183, 186]

There are a couple of main drawbacks to the above design paradigm. 1) Structure templates identified in the first design phase are always assumed to be static, disregarding the inherent dynamics of proteins. Therefore, restricting the inverse folding phase to only a snapshot of the protein system results in a constrained exploration of the potential sequence design landscape. 2) A successfully designed protein system must both be viable in the desired cellular environment and meet functional objectives. Frequently, a single iteration through the outlined design paradigm proves inadequate to accomplish these goals.

To address the abovementioned shortcomings, this thesis project will attempt to develop a structural generative model for variational structure sampling. By doing so, expanding the sampling space available to inverse folding models, a revolutionized adaptive computational design paradigm is proposed. This paradigm integrates structural generative models and advanced system simulation for robust and efficient *in silico* design optimization. Moreover, promising preliminary results suggest that our structural generative model has the potential to serve as a pre-training framework for protein structure, comparable to the protein language models tailored for protein sequences, as discussed earlier. Part of this thesis will be dedicated to the exploration of upscaling the structural generative model for structure embedding and its application to various downstream tasks, including protein function annotation and inverse protein folding.

2.2 Abstract

Designing protein with desirable structure and functional properties is the pinnacle of computational protein design with unlimited potentials in the scientific community from therapeutic development to combating the global climate crisis. However, designing protein macromolecules at scale remains challenging due to hard-to-realize structures and low se-

quence design success rate. Recently, many generative models are proposed for protein design but they come with many limitations. Here, we present a VAE-based universal protein structure generative model that can model proteins in a large fold space and generate high-quality realistic 3-dimensional protein structures. We illustrate how our model can enable robust and efficient protein design pipelines with generated conformational decoys that bridge the gap in designing structure conforming sequences. Specifically, sequences generated from our design pipeline outperform native fixed backbone design in 856 out of the 1,016 tested targets(84.3%) through AF2 validation. We also demonstrate our model’s design capability and structural pre-training potential by structurally inpainting the complementarity-determining regions(CDRs) in a set of monoclonal antibodies and achieving superior performance compared to existing methods.

2.3 Introduction

Computational protein design has been of great interest to the scientific community for decades. Designing protein macromolecules with specific function and structure is a highly sought after technique with broad application to therapeutics, biosensors, and enzyme engineering [59, 183, 109, 190]. However, despite years of effort and advancements, computational protein design still remains a very challenging problem.

Traditionally, the protein design pipeline is often regarded as a two step process where the practitioner first determines the protein backbone structure accommodating the specified structural and biochemical properties, and then designs the amino acid sequence with the given backbone structure. In this regime, the most successful applications rely heavily on template-based fragment sampling and domain-expert specified topologies for backbone determination and energy minimization based sequence design [28, 23]. While this approach is widely adopted, there are apparent drawbacks. For example, the resulting backbone structure from the first step may not be optimal or designable and the designed sequence in the

second step may not readily conform to the desired backbone template. As a result, the success rate for computational protein design remains relatively low [70, 9].

In recent years, progress in machine learning and deep learning research has contributed to significant advances for protein modeling such as mutation effect estimation [45, 121, 140], protein function prediction [95, 54], and structure prediction [84, 11, 176]. These advancements have also played a part in aspects of the computational protein design problem. For fixed backbone sequence design, a series of deep learning methods have emerged to improve conventional energy based approaches [5, 87] by directly incorporating structural information using SE(3)-equivariant frameworks [120, 80, 68]. An array of recent works have studied the use of generative models for structure generation[106, 7, 6], however, these methods often generate topological constraints and rely on downstream tools for 3-dimensional structure determination. One of the major difficulties for direct coordinate generative modelling is properly accounting for the rotation and translation equivariance in the target conformation.

In the Chapter, we present a versatile VAE-based deep structural generative model that seeks to bridge the gap for robust computational protein design. Our method makes contributions to three fundamental aspects of protein design: First, we directly model protein structure in the 3-dimensional coordinate space which avoids downstream coordinate recovery in constraints based model. Second, Our method is universal such that it can model proteins of arbitrary size and thus exposes our model to the whole fold space while previous generative models are restricted to protein with certain size and can only be trained on a small subset of all available folds for a given model. Third, We address the translation and rotation equivariance in both the input space and the use of a locally aligned coordinate loss proposed by [84]. It is important to note that structure entries in database such as the Protein Data Bank(PDB) [14] often represent a single sample from its conformational

landscape. By conditioning on a given backbone structure, our model is able to generate conformational decoys from the latent space. Combined with fixed-backbone sequence design models and accurate protein structure prediction tools, our model can enable efficient *in silico* design screening.

To evaluate our models from multiple aspects. First, we demonstrate our model’s ability to generate high-quality, realistic protein structure ensembles by comparing the generated three-dimensional coordinates to experimentally determined structures. This comparison allows us to assess the accuracy and realism of our model’s output. Second, we showcase improved efficiency and success rates in conventional design pipelines by producing sequences that recapitulate backbone templates derived from our model. This evaluation highlights the practical applicability of our approach in enhancing existing protein design methodologies. Lastly, we corroborate our model’s design capabilities through the inpainting of backbone coordinates in the complementarity-determining regions (CDRs) of monoclonal antibodies. In this task, we achieve state-of-the-art results, exemplifying our model’s versatility as a structure generator. These evaluations collectively validate our model’s effectiveness in generating, improving, and manipulating protein structures, highlighting its potential as a powerful tool in computational protein design.

2.4 Literature Review

2.4.1 Deep Generative models

. Generative adversarial networks(GAN) [56] and auto-encoding variational Bayes (VAE)[88] are powerful generative frameworks that are used from image synthesis[20, 139] to language modeling[15]. [140, 155, 45] also used VAE for protein sequence design and variant effect prediction. Most of the aforementioned structure generative models also used VAE or

GAN as their generative framework. Recently, Diffusion generative models and flow-based models are emerged in popularity specifically for tasks in protein structure prediction and generation[187, 177, 74, 65]. We build our structural model based on the VAE framework where we first encode the invariant protein structure representations(See Methods) into the latent space then a decoder is used to generate the corresponding three-dimensional coordinates from the latent space. This approach allows our model to generate flexible protein conformations conditioned on the input backbone constraints. Trained on masked input constraints, our model can easily be adopted for backbone inpainting for structural design.

2.4.2 Protein structure generative models

. Existing generative methods for protein structure design focus on generating invariant topological constraints such as inter-residue contacts and distance maps which are then converted into three dimensional structures via downstream coordinate recovery tools such as AMDD[16] or pre-trained structure predictors[184, 11]. While generative models can produce protein structures through 1D and 2D constraints, they often require a second step to recover the Cartesian coordinates. This recovery process can be challenging, especially for 1D representations where small errors in backbone torsion can propagate and lead to unrealistic structures. For 2D representations, methods have been developed to convert distance matrices into 3D structures, but these can be sensitive to systematic noise produced by neural networks. Though these methods have garnered some success, there is no a priori guarantee that the generated constraints are geometrically viable. Consequently, the resulting three dimensional structures are often of low quality or biochemically infeasible. Moreover, it is generally not possible to perform conditioned structure generation on these models[106, 7, 6, 86]. On the other hand, methods that focuses on generating structure ensembles are often fold specific and also relying on generating topological constraints[113, 77].

While [41] proposed the first direct coordinate structure generative model, its application is limited to proteins of a fixed length, and is trained only to recover inter-atom distance and torsion maps. In contrast, our model, addresses rotation and translation equivariance in both the input and output space by distilling invariant representations of protein geometry and by using a equivariant locally aligned coordinate loss function to perform gradient optimization directly on the coordinate space. In this way, our model can directly and flexibly model the three-dimensional structure and generate conformational snsembles. For diffusion based models such as [177, 74, 65], though highly flexible, they are often limited by intensive compute requirement and the quality consistency of the generated structures.

2.4.3 Fixed backbone sequence design

For sequence optimization or often known as the inverse folding problem or fixed backbone sequence design, Traditionally, this problem was addressed by optimizing energy functions through Markov chain Monte Carlo (MCMC) methods, combining physical and statistical potentials [152]. However, these conventional methods often resulted in limited sequence diversity and struggled with designing multi-body interactions crucial for protein function[17, 112].

DL-based sequence design algorithms have emerged to address these limitations. These methods utilize various representations of protein structures, including graphs [161, 73, 33], 2D matrices [24, 123], torsional angles [124], and voxelized volumes [137, 8, 193], to generate sequence probability profiles or full sequences. Some approaches frame the problem as a constraint satisfaction problem [161], while others use MCMC[8] or autoregressive models[73] to generate sequences. DL methods have shown promise in introducing more diverse sequences and capturing multi-body interactions more effectively than conventional approaches. However, it is important to note that while these methods show great potential, their generalizability and reliability have not been extensively validated through experimen-

tal testing[8, 161, 33]. As the field progresses, DL approaches may offer a more flexible and powerful toolkit for protein design, potentially leading to more diverse and functionally optimized sequences.

Although our model does not directly design the sequence alongside the conformation decoys, we demonstrate how our generated conformational ensembles can be utilized and address some of the problems mentioned above and significantly improve the design capability of inverse folding models.

2.5 Methods

In this section we illustrate the overall model design as shown in Figure 2.1. To prepare for input features, each protein structure is distilled into invariant pairwise representations of inter-residue distance and orientations as described in [184] and scalar representations of amino acid sequence and backbone torsion angles. This input is then fed through an encoder network which produces a latent representation of each residue. These representations are reassembled and passed to a decoder module which reconstructs the backbone coordinates.

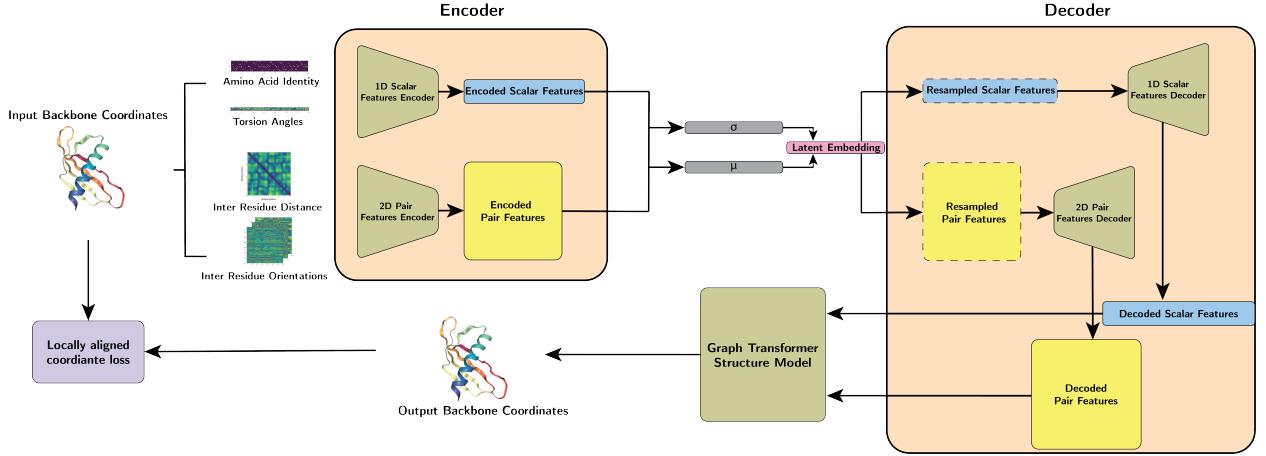


Figure 2.1: Model Overview. Structures are first distilled into pairwise (inter-residue distance & orientation) and scalar (torsion angles & amino acid sequence) features. Then our model separately encodes the pairwise and scalar features into the latent space. The latent representation is then sampled and decoded by the pair and scalar feature decoder before being fed into a graph-transformer based structure module for coordinate generation. A locally aligned loss is then computed between the reconstructed coordinates and the native input structure.

2.5.1 Protein representations

We represent a protein as a complete molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} consists of scalar residue features v_i and \mathcal{E} consists of pairwise features e_{ij} between residues i and j .

Scalar Features

The nodes v_i of our input graph correspond to protein residues $i \in \{1\dots n\}$. The input scalar feature \mathcal{F}_i^{scalar} associated with residue i consists of amino acid identity and dihedral angle encodings:

$$\mathcal{F}_i^{scalar} = \{\mathbf{f}_{AA}(s_i), \mathbf{f}_{dihedral}(\boldsymbol{\theta}_i)\} \quad (2.1)$$

The first, $\mathbf{f}_{AA}(s) \in \mathbb{R}^{20}$, is a one-hot encoding of the amino acid type s using 20 bins for each naturally occurring amino acid. The second, $\mathbf{f}_{dihedral}(\boldsymbol{\theta}) \in \mathbb{R}^6$, is an encoding of the three dihedral angles with Fourier features $\{\sin, \cos\} \circ \{\phi_i, \psi_i, \omega_i\}$, where ϕ_i, ψ_i, ω_i are dihedral angles computed from the coordinates from $C_{i-1}, N_i, C_{\alpha i}, C_i, N_{i+1}$ atoms.

Pairwise Features

for a given pair of residues i and j , we define the edge e_{ij} features as

$$\mathcal{F}_{ij}^{pair} = \{\mathbf{f}_{dist}^{(v_i, v_j)}(\vec{X}_j^{C_\alpha}, \vec{X}_i^{C_\alpha}), \mathbf{f}_{ori}(\theta_{ij})\} \quad (2.2)$$

The first encoding, $\mathbf{f}_{dist}^{(v_i, v_j)}(\vec{X}_j^{C_\alpha}, \vec{X}_i^{C_\alpha}) \in \mathbb{R}^{16}$ is the distance encoding that embeds the inter-residue distance $d_{C_\alpha} = \|\vec{X}_j^{C_\alpha} - \vec{X}_i^{C_\alpha}\|_2$ with 16 Gaussian radial basis functions with centers evenly spaced in $[0, 20]\text{\AA}$ as described in [80]. $\mathbf{F}_{ori}(\theta) \in \mathbb{R}^3$ is the encoding of the angle θ performed in the same manner as the backbone dihedral encoding for residue features. The input angles $\theta_{ij} \in \{\phi_{ij}, \psi_{ij}, \omega_{ij}\}$ are pairwise inter-residue orientations defined in [184]. To produce pairwise orientation information, we impute a unit vector in the direction $C\beta_i - C\alpha_i$ before computing the respective angles. The imputed vector is calculated as in [80] using

$$\sqrt{\frac{1}{3}} \langle \mathbf{n} \times \mathbf{c} \rangle - \sqrt{\frac{2}{3}} \langle \mathbf{n} + \mathbf{c} \rangle$$

where $\langle x \rangle = x/\|x\|_2$, $\mathbf{n} = N_i - C\alpha_i$, and $\mathbf{c} = C_i - C\alpha_i$.

Masking Schemes for structure inpainting and pre-training

To help facilitate our model's ability to inpaint protein structural regions, we employ a masking scheme for a given protein denoted as $Mask(n)$. We implemented three types of masks: **linear** mask where a random residue is selected uniformly at random with $p = 1/L$ where L is the length of the respective protein and the mask will then span 10 residues around the selected one.

spatial mask where a random residue is selected uniformly at random with $p = 1/L$ and all residues within 12\AA are masked.

random mask where each residue will be masked at the rate of $p \sim Uniform(0, 0.5)$ where p is sampled for each individual residue.

For each masked residue, we use zero masks in both the scalar and pairwise features such that $\mathcal{F}_i^{scalar} = \mathbf{0}$ and $\mathcal{F}_{i,:}^{pair} = \mathcal{F}_{:,i}^{pair} = \mathbf{0}$ $\forall i \in Mask(n)$. In the hybrid masking training strategy, 50% of the input proteins are mask with one of the aforementioned masks with equal probability. We also analyzed the effect of various masking probability and strategies, see the Appendix for more details.

2.5.2 Model Architecture

Feature encoders & decoder

The CoordVAE model uses different architectures for its encoder and decoder networks. The encoder network consists of a 1D feature module for the scalar input and a 2D feature module for the pairwise input using dilated convolution network architecture [188].

$$\mathcal{H}_{l+1}^{1D} = \text{LeakyReLU}(\text{1D-InstanceNorm}(\text{1D-Conv}(\mathcal{H}_l^{1D}, \mathcal{K}_l^{1D}))) \quad (2.3)$$

$$\mathcal{H}_{l+1}^{2D} = \text{LeakyReLU}(\text{2D-InstanceNorm}(\text{2D-Conv}(\mathcal{H}_l^{2D}, \mathcal{K}_l^{2D}))) \quad (2.4)$$

Where \mathcal{H}_l^{1D} and \mathcal{H}_l^{2D} denotes the scalar and pairwise representation at layer l respectively and \mathcal{K}_l^{1D} and \mathcal{K}_l^{2D} denotes the kernel size for the scalar and pairwise convolution. Each convolutional block applies a leakyReLU nonlinearity with negative slope parameter set to 0.2, and instance normalization [166].

The decoder network applies a mirrored architecture to the latent scalar and pairwise representations. Finally, the decoded scalar and pairwise features are processed jointly with a graph transformer described in [154] for coordinate generation.

$$\mathcal{H}_{out}^{2D}, \mathcal{H}_{out}^{1D} = \text{GraphTransformer}(\mathcal{H}_{dec}^{1D}, \mathcal{H}_{dec}^{2D}) \quad (2.5)$$

Where \mathcal{H}_{dec}^{1D} and \mathcal{H}_{dec}^{2D} are the scalar and pairwise output of the decoder network respec-

tively. For Coordinate generation, we process the output of the graph transformer module with a dense projection.

$$\vec{X}_i = \text{Linear}(\mathcal{H}_{out}^{1D})_i \quad (2.6)$$

Where \vec{X}_i are the backbone coordinates of residue i and the projection is implemented with as a dense layer without bias. For CNN baseline models, the graph transformer module is removed and coordinates are directly projected from the 1D feature decoder output \mathcal{H}_{dec}^{1D} .

VAE

To implement a VAE framework within our encoder-decoder architecture, we combine the encoder output by a horizontally average-pooling(HAP) the pairwise output feature and concatenate it with the scalar output before applying the mean and variance projection networks.

$$\mathcal{H}_{latent} = \text{HAP}(\mathcal{H}_{enc}^{2D}) \parallel \mathcal{H}_{enc}^{1D} \quad (2.7)$$

Where \mathcal{H}_{enc}^{1D} and \mathcal{H}_{enc}^{2D} are the scalar and pairwise of the encoder output respectively and \parallel denotes concatenation. To produce the mean and variance of the latent variable we use two separate projections on \mathcal{H}_{latent} .

$$\vec{\mu}_i = \text{Linear}^\mu(\mathcal{H}_{latent})_i \quad (2.8)$$

$$\vec{\sigma}_i^2 = \text{Linear}^{\sigma^2}(\mathcal{H}_{latent})_i \quad (2.9)$$

$$\mathcal{Z}_i \sim \mathcal{N}(\vec{\mu}_i, \vec{\sigma}_i^2) \quad (2.10)$$

To produce the input scalar and pairwise features to the decoder networks, we take the outer product of the sampled latent representations.

$$\mathcal{Z}^{1D} = \text{Linear}(Z) \quad (2.11)$$

$$\mathcal{Z}^{2D} = \mathcal{Z}^{1D} \otimes \mathcal{Z}^{1D} \quad (2.12)$$

Where \mathcal{Z}^{1D} and \mathcal{Z}^{2D} are the input to the scalar and pairwise feature decoder respectively. For the architecture, we used [32, 63, 128, 256] number of channels in both the 1D and 2D feature encoder and the inverse for 1D and 2D feature decoder. we used 64 as the latent dimension in our experiments. Models are optimized using Adam with learning rate of $1e^{-3}$

Loss function and objectives

Let \vec{X}_i denote the backbone coordinates of residue v_i . We predict backbone coordinates of $\{N_i, C_{\alpha i}, C_i, O_i\}$ for each residue i . After predicting coordinates $\{N_i, C_{\alpha i}, C_i, O_i\}$, we follow [84], and define the local $C\alpha$ - frame of residue i as the rigid transformation $T_i = (R_i, \vec{t}_i) \in SE(3)$ such that

$$T_i \circ \vec{C}_{\alpha i} = (0, 0, 0, 0) \quad (2.13)$$

$$T_i \circ \vec{N}_i = (\|\vec{C}_{\alpha i} - \vec{N}_i\|_2, 0, 0, 0) \quad (2.14)$$

$$T_i \circ \vec{C}_i = (0, \|\vec{C}_{\alpha i} - \vec{C}_i\|_2, 0, 0) \quad (2.15)$$

$$T_i \circ \vec{O}_i = (0, 0, \|\vec{C}_{\alpha i} - \vec{O}_i\|_2, 0) \quad (2.16)$$

where $T \circ \vec{x} \triangleq R\vec{x} + \vec{t}$. Assuming linear independence between the displacement vectors $\vec{C}_{\alpha i} - \vec{N}_i$ and $\vec{C}_{\alpha i} - \vec{C}_i$, this transformation is unique and well defined. With a single local frame defined from each residue's predicted coordinates, we are able to apply per-residue frame aligned point error (pFAPE) loss against the native coordinates and local frames as

$$\text{pFAPE} \left(T_i, T_i^*, \{\vec{X}_j\}, \{\vec{X}_j^*\}; \theta \right) = \frac{1}{\theta} \sum_j \min \left(\left\| (T_i)^{-1} \circ \vec{X}_j - (T_i^*)^{-1} \circ \vec{X}_j^* \right\|_2, \theta \right) \quad (2.17)$$

where $\theta = 10$ is a threshold determining when the loss value should be clamped, and an asterisk is used to differentiate between native and predicted frames and coordinates. The pFAPE loss is averaged over all frames T_i and all atom types X to produce the final loss \mathcal{L}_{FAPe} . For the VAE loss, in addition to the FAPE reconstruction loss, we also compute

the Kullback–Leibler divergence loss

$$\mathcal{L}_{total} = \mathcal{L}_{FAPE} + \beta D_{KL}(Z||p(Z)) \quad (2.18)$$

Where D_{KL} is the KL Divergence between the latent variable distribution and the prior multivariate Normal distribution.

Fixed backbone sequence design We use the pre-trained fixed backbone design model of [33] to generate sequences from backbone structures. For each test target, we generate 500 conformational decoys and sequences are deranged for each of the decoys independently.

Structtrue prediction oracles We use the AF2[84] implemented by ColabFold[122] as well as ESMFold[107] for structure prediction, for all structures, we use single sequence without MSA and set the number of recycles to three for all sequences. We make structure prediction to each sequence designed from the conformational decoy library and select sequence based on the prediction results.

2.5.3 Data

CATH4.2

We obtained the the CATH4.2 data from [73] which contains 19,752 structures and structurally split with into train/validation/test sets by CATH fold annotations. To ensure sequence and structural independence between the train/test structures, the 40% non-redundant set is used and split is done on the topology/fold level.

Antibody Structures

For antibody structures, we used the structural antibody Database (SAbDab)[39] obtained from [79] which contains 1266, 1564, 2325 structures for CDR-H1, CDR-H2, and CDR-H3 respectively after filtering and splitting, there is no more than 40% sequence identity in the

inpainted regions for each set of structures between the train/test structures. Please refer to [79] for further details.

2.6 Results

In this section, We first demonstrate our model’s ability to generate realistic backbones structures through comparing with a set of experimentally determined and structurally independent structures. Then we show how the conformational decoys generated by our model outperforms native structure backbones for designing structure conforming amino acid sequences. Lastly, we tested out model’s structure design capability by testing CDR inpainting on a set of monoclonal antibodies.

2.6.1 End-to-End structure reconstruction & generation

In this section, we will demonstrate our model’s ability to reconstruct high-quality protein 3-dimensional structures from invariant protein representations. The most crucial capabilities for structure generative models is to construct topologically feasible protein conformations such that the generated backbone structures are viable candidates for downstream design pipelines. For this propose, we presented structural similarity metrics including the local distance difference test (lddt) score[117], TM-score[192], root mean square deviation (RMSD), and the L_1 norm of inter-residue distance between the reconstructed structures and the native input structures to measure how well our models reconstruct the input structures.

Our model is able to consistently construct high-quality structure coordinates of the target proteins up to $L = 500$ residues as shown in Figure 2.2(A & B). To further inspect the geometric feasibility of the reconstructed proteins, we examine the torsion angle distributions and compared it to the native structures as shown in Figure 2.2(C) where we notice the torsion angles from the generated structures are mostly in the feasible regions and matches

up closely to the native torsion angles. To illustrate how our model’s learned latent representation of the input protein contains useful information in its fold space, we plot the t-SNE embedded components colored by its CATH class annotations in Figure 2.2(D). Structures that share the same CATH fold class are clustered in the latent space accordingly.

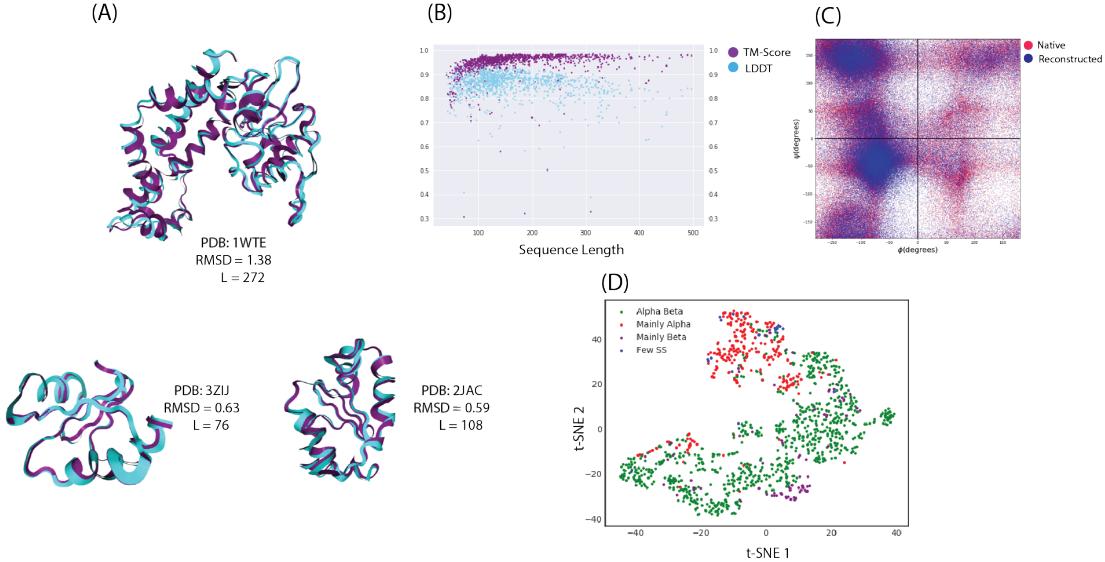


Figure 2.2: (A) Examples of reconstructed backbone structures of different folds and sizes. 1WTE (top), 3ZIJ (bottom left), 2JAC (bottom right). (B) LDDT (blue) and TM-score (purple) against native structure vs. the length of the test target. (C) Ramachandran plot of the native (red) and reconstructed(blue) structures. (D) t-SNE embedding of the latent space for the test targets colored by their CATH class annotation.

In summary, our model demonstrates remarkable consistency in generating high-quality structural decoys across diverse fold classes and protein sizes. The reconstruction accuracy of protein structures not included in the training sets establishes a lower bound for our model’s performance, consistently achieving backbone RMSD values around 1Å. This level of accuracy, combined with the model’s broad applicability, underscores its robustness and flexibility compared to prior methods.

2.6.2 Ablation study on input features

To understand how different input features impact model performance, we compared our model with different features ablated in table 2.1. When both inter-residue distance and orientation are used along with the amino acid sequence, our model achieved average RMSD of 1.122 and TM-score of 0.941 at experimentally comparable resolution. We notice model performance degrades when we remove the inter-residue orientation as input features. This suggests our model relies on spatial orientation information to accurately construct the structure coordinates. It is worth noting that even though our model achieves the best performance with amino acid sequence input, our model has comparable result when operate under sequence free mode for backbone-only decoy generation.

Model	lddt↑	TM↑	RMSD↓	Distance L1↓
CoordVAE CNN	0.472	0.420	9.049	2.666
CoordVAE w/o orientation	0.604	0.649	4.429	1.413
CoodVAE w/o seq	0.788	0.905	1.489	0.699
CoordVAE	0.841	0.941	1.122	0.502

Table 2.1: Average structural similarity metrics evaluated on the test targets over different models

In addition, we also compared the short, medium, long range contact accuracy where the predicted contacts are derived from the inter-residue distance of the generated 3-dimensional coordinates as shown in Table 2.2

We benchmark the contact accuracy across different models, evaluated on 1,120 test targets.

Model	Contact(S)↑	Contact(M)↑	Contact(L)↓
CNN Baseline	0.321	0.3.5	0.193
CoordVAE w/o orientation	0.500	0.436	0.378
CoodVAE w/o seq	0.875	0.833	0.785
CoordVAE	0.885	0.857	0.819

Table 2.2: Average short(S), medium(M), long(L) contact accuracy evaluated on the 1,120 test targets over different models

The results demonstrate a clear progression in performance from the CNN Baseline to the

full CoordVAE model. The CNN Baseline shows the lowest accuracy, with values of 0.321, 0.305, and 0.193 for short, medium, and long-range contacts respectively. The CoordVAE without orientation significantly improves upon this, achieving accuracies of 0.500, 0.436, and 0.378. Further improvement is seen with the CoordVAE without sequence information, which reaches high accuracies of 0.875, 0.833, and 0.785. The full CoordVAE model, incorporating both orientation and sequence information, demonstrates the best performance across all contact ranges, with the highest accuracies of 0.885, 0.857, and 0.819 for short, medium, and long-range contacts respectively. This consistent improvement across all contact ranges, particularly for the challenging long-range contacts, suggests that the full CoordVAE model effectively captures complex structural relationships in proteins despite no loss function was imposed on the contact map. The results highlight the importance of both orientation and sequence information in accurately predicting protein contacts.

2.6.3 Conformational decoy sampling for robust protein sequence design

In this section, we describe how our structural generative model can be used for robust *de novo* protein design. The design pipeline using conformational decoys is outlined in Figure 2.3. A primary design target backbone structure is provided as the design template, such a template can be generated by topology design program such as TopoBuilder[150] or other template based methods. Our structure generative model(CoordVAE) will embed the input backbone into a latent representation then a set of conformational decoys is generated by the structure decoder to form a library. A sequence library is then produced by running a fixed backbone inverse folding program on the decoy library. Each sequence will then be folded and validated by a structure prediction oracle such as AlphaFold[84], and those which pass the structure validation criteria will be used to form a structurally validated sequence library for further filtering and downstream experimental screening. Our structure model

can generate thousands of decoys per minute on a single GPU for fast and efficient library generation.

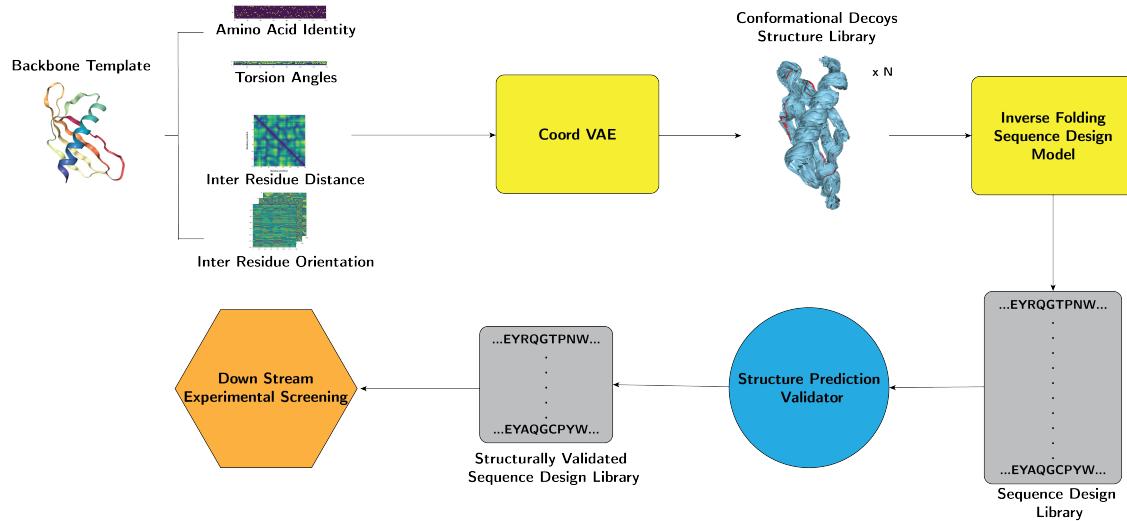


Figure 2.3: Outline of robust protein design procedure using conformational decoy library. Starting with a primary backbone structure target, our structure generative model embeds it into a latent space and decodes into conformational decoys to form a structure library. Using a fixed backbone sequence design program on the backbone structure library, one can obtain a preliminary sequence library. To filter and structurally validate the primary sequence library, a structure prediction oracle is used and the validated sequence library is ready for further downstream tasks.

To elucidate how sequences designed from the conformational decoys compare to sequence designed with native one shot backbone templates, we designed an experiment such that we designed sequences from both a backbone decoy library generated from our structural model and single shot experimentally determined backbones from PDB and we computationally predict the structure of sequences designed from both set and see which set can produce more computationally confident sequences.

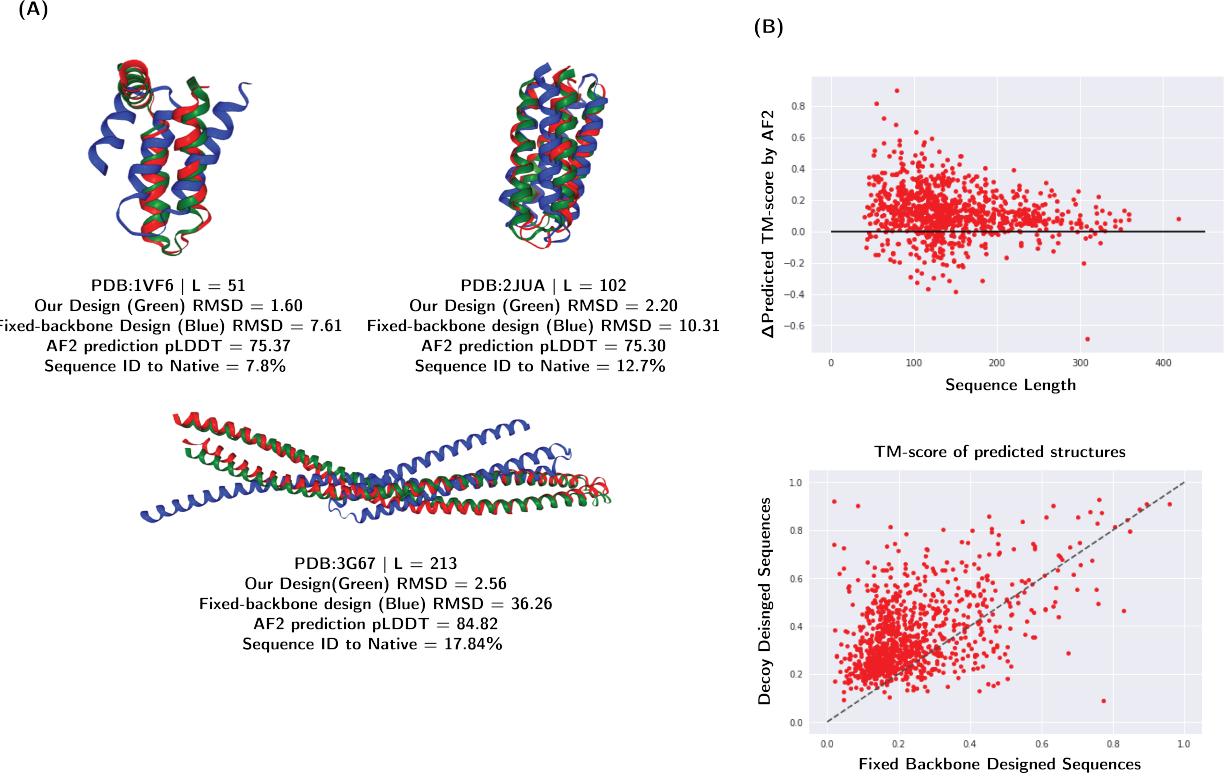
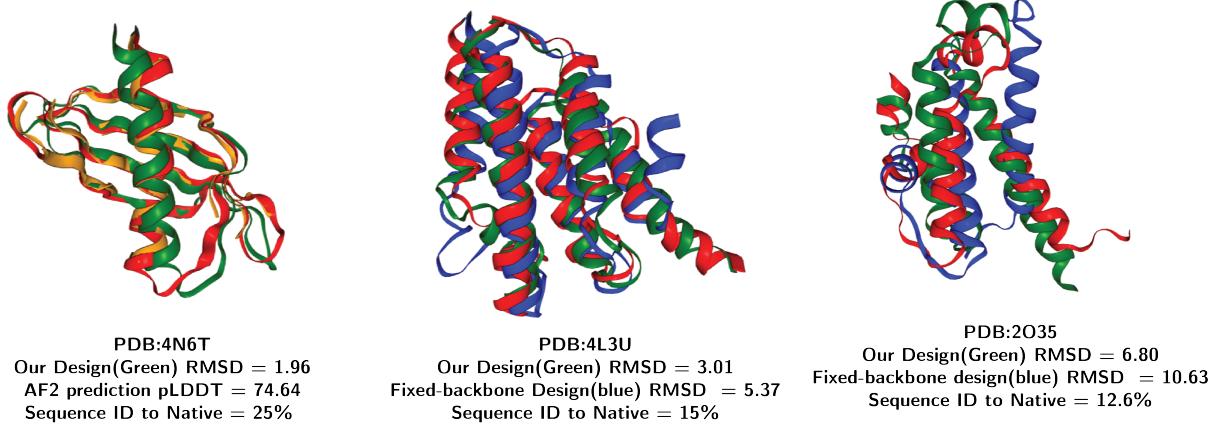


Figure 2.4: Examples of protein designed from conformational decoys. **(A)** AF2 folded sequence designed from conformational decoys. Overlay of AF2 predicted structure of decoy designed sequence & native backbone designed sequence (green & blue), native backbone(red) for PDB:1VF6 (top left), PDB:2JUA (top right), PDB:3G67 (bottom). **(B)** Scatter plots of TM-scores between the best folded conformational decoy designed sequences and sequences designed from the native backbones(bottom). Δ TM-score between decoy designed sequences and fixed backbone designed sequences versus the target size.(top)

We provide examples in Figure 2.4(A) where our designed sequence folded more successfully than the native sequence when MSA and template information is not available. Particularly, as shown for PDB:3G67, the native sequence is folded with $\text{RMSD} = 36.26$ while our best decoy designed sequence folded with $\text{RMSD} = 2.56$ with high confidence. We observed that our model can consistently improve designed quality regardless of the size of the target backbone. In Figure 2.4 (B), we deployed the design pipeline to a set of 211 backbone targets and plotted the TM-score of the predicted structure compared to native sequence(bottom), and sequence designed from the native backbone structure(top). Sequences designed from our pipeline outperform sequences designed from fixed native backbones in

856 out of the 1,016 tested targets(84%).

(A)



(B)

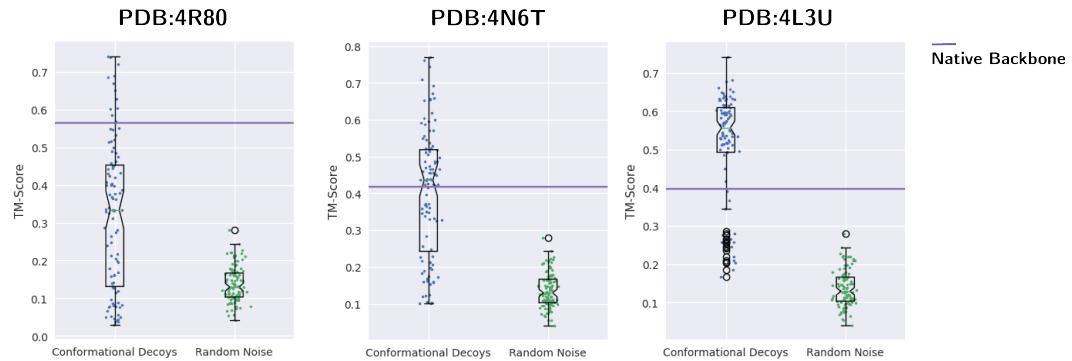


Figure 2.5: (A) Examples of AF2 predicted designed sequences and native sequences. (B) TM-score distributions of decoy designed sequences vs. noisy backbone designed sequences.

To demonstrate our approach can simultaneously drive sequence diversity and design confidence, we showed more examples of conformer based sequence design at Fig 2.5(A). To confirm that our structure generative model does not simply generate noisy versions of the input structure, we compared with sequences designed from backbone coordinates with added random Gaussian noise. As shown in Figure 2.5(B), the results for conformatioal decoys are clearly favorable to those of noisy backbones.

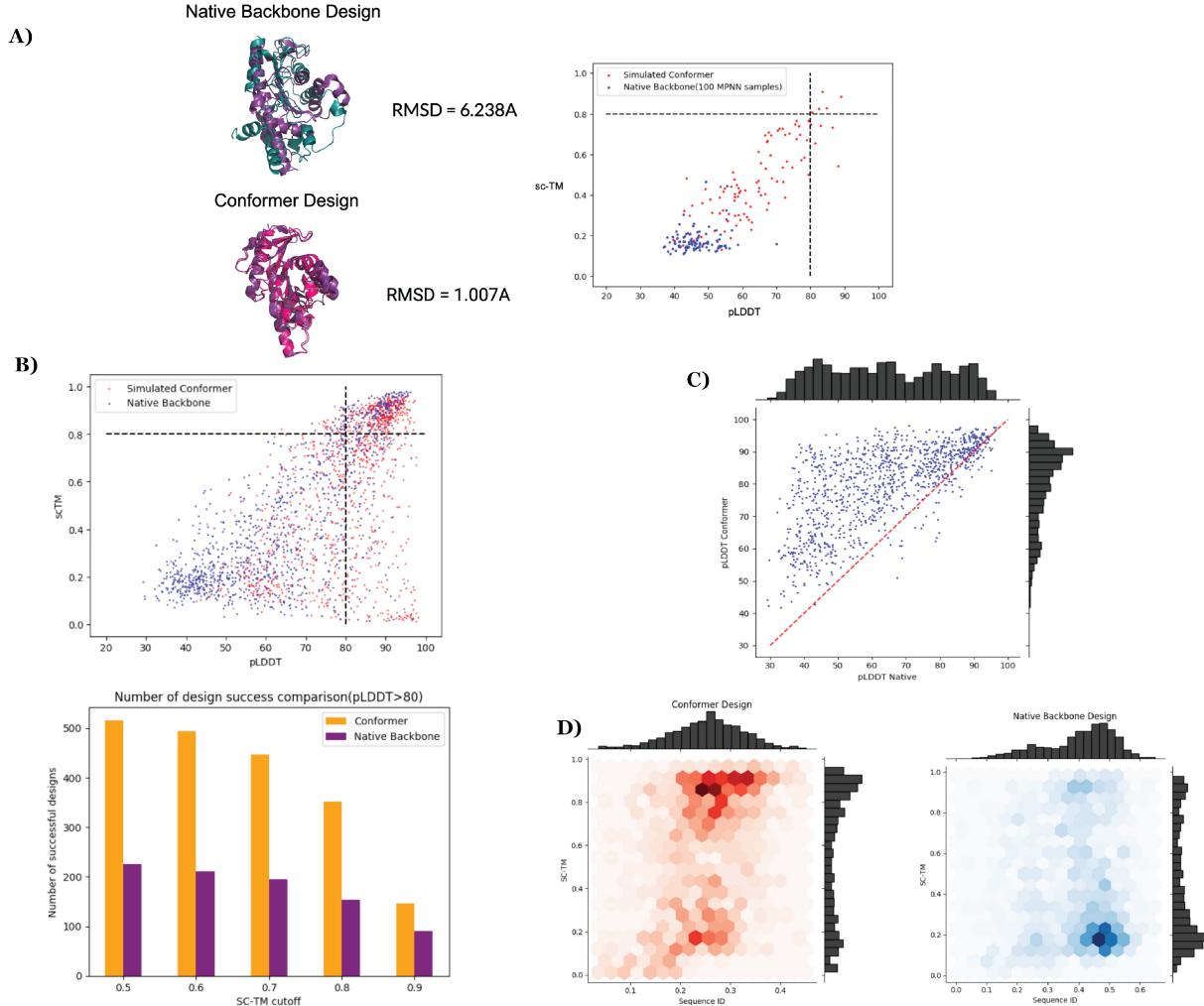


Figure 2.6: A) Example of backbone ensemble sampling vs. sequence design sampling of PDB:1JDI. B) Self-consistent TM score comparison between backbone ensemble sampling and single shot native backbone design(Top). Design success rate across different sc-TM threshold comparison(Bottom). C) pLDDT of designed sequence comparison between conformer sampling vs. single shot native backbone. D) sc-TM distribution vs. sequence identity distribution of the conformer based sequence design(Left). sc-TM distribution vs. sequence identity distribution of the native backbone sequence design(Right)

To further demonstrate our deep structure generative model presents an edge in providing diverse backbone ensembles that can overcome limitation in inverse folding methods relying on single-shot native backbones. Fig. 2.6(A) illustrates an example where proteinMPNN temperature-based sequence sampling with a native backbone failed to generate structure-conforming sequences, while our structure ensemble approach yielded many suc-

cessful designs. The advantages of our method is further evidenced in Fig. 2.6(B), which shows the distribution of sc-TM scores versus pLDDT across all tested targets. Notably, our conformer-based designs (red points) achieved a higher concentration of successful designs in the upper-right quadrant compared to single-shot native backbone templates (blue points). The bar chart below quantifies this advantage, showing that conformer-based designs consistently outperform native backbone designs across various sc-TM cutoffs, with approximately twice the success rate in many cases. Fig. 2.6(C) further supports this conclusion by showing that our conformer-based designs (blue points) consistently achieve higher pLDDT scores compared to native backbone designs across the majority of tested folds.

A particularly important finding is illustrated in Fig. 2.6(D), which reveals the relationship between sequence diversity and design success. Contrary to the conventional wisdom that suggests a trade-off between sequence diversity relative to native proteins and design success rate, our method demonstrates the ability to simultaneously optimize for both. The conformer-based approach (left heatmap) generates much more diverse sequences while increasing design success, as evidenced by the higher density of points in the upper-left region compared to the native backbone design (right heatmap). This is a significant improvement, as it allows for broader exploration of the sequence space without compromising designability.

These results underscore the power of our deep structure generative model in enhancing protein design capabilities. By providing diverse backbone ensembles, our approach not only improves the success rate of inverse folding but also expands the accessible sequence space, potentially leading to the discovery of novel protein designs with unique properties and functions.

In summary, we show that conformational decoy ensembles generated by our model can be used to improve protein design pipeline by providing a structure library for downstream fixed-

backbone sequence design applications therefore significantly increase the available sampling space. Our experiments demonstrate that structure-conforming sequences can be reliably designed from the conformational decoys compared to single backbone targets. With this approach, we present a new path towards robust and efficient protein design.

2.6.4 Unconditional structure inpainting for antibody design

Monoclonal antibodies are important targets for therapeutics development and considerable effort has been dedicated by the community towards computational antibody design[133, 5, 41, 2]. While previous methods mostly focused on sequence design, we adopt our structure generative models to inpaint the complementarity-determining regions (CDRs) for structure based antibody design. CDR grafting is a fundamental technique in antibody engineering which transplants the CDR region from one antibody to another. However, it comes with significant challenges that can impact the success and efficiency of the process due to the delicate nature of antibody binding characteristics. In this section we show how our model can be used for self-consistent structure design with a hybrid mask scheme and provide a direct comparison with existing structure design methods.

To inpaint protein structures, we employed three types of masks; linear, spatial and random and a hybrid masking strategy(see details in the Method section) that impose masks in the input scalar and pair representation and the completed structure will be recovered by the structure generative model. To further test the ability of our model to distill meaningful representation from a larger fold space, we pre-traiend our model on the CATH4.2 dataset and fine-tune the model with a set of monoclonal antibody structures.

Model	Training Data	CDR-H1		CDR-H2		CDR-H3	
		lddt↑	RMSD↓	lddt↑	RMSD↓	lddt↑	RMSD↓
CoordVAE(10AA linear)	CATH4.2	0.587	2.847	0.588	2.864	0.498	4.427
CoordVAE(spatial)		0.515	3.944	0.514	3.914	0.487	4.350
CoordVAE(Random)		0.447	4.216	0.534	4.152	0.450	5.185
CoordVAE(mixed mask)		0.591	2.903	0.575	3.223	0.530	3.889
CoordVAE	SABDAB	0.81	1.55	0.872	1.00	0.809	1.55
CoordVAE(Transfer)		0.88	0.81	0.90	0.85	0.823	1.35
AR-GNN		N/A	2.97	N/A	2.27	N/A	3.63
RefineGNN		N/A	1.18	N/A	0.87	N/A	2.50

Table 2.3: Structure inpainting performance on the test monoclonal antibody dataset in CDR-H1(left), CDR-H2(middle), CDR-H3(right) across different models.

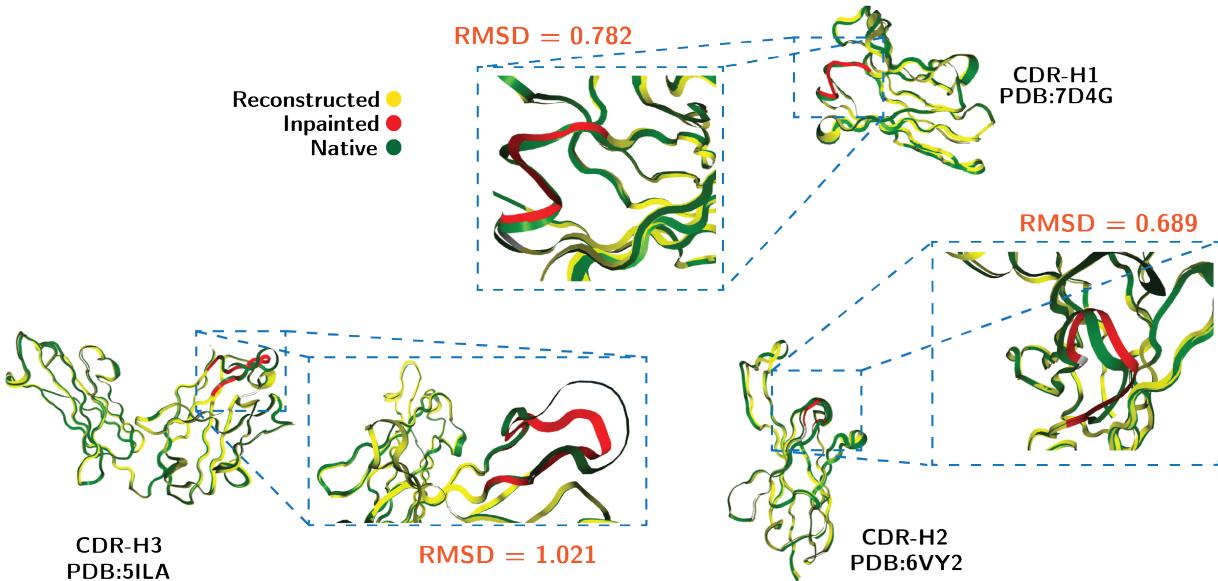


Figure 2.7: Examples of antibody structure inpainting. overlay of the CDR-H3 region of PDB:5ILA (bottom left) native structure(green), unmasked region (yellow) and masked inpainting region (red) reconstruction with RMSD = 1.021. overlay of the CDR-H2 region of PDB:6VY2 (bottom right) native structure(green), unmasked region (yellow) and masked inpainting region (red) reconstruction with RMSD = 0.689. overlay of the CDR-H1 region of PDB:7D4G (top) native structure(green), unmasked region (yellow) and masked inpainting region (red) reconstruction with RMSD = 0.782

We first tested our model’s inpainting performance on three CDRs(H1, H2, H3) with models trained on the CATH4.2 dataset without fine-tuning on antibody structures. Our best model trained on hybrid masking scheme achieved RMSD of 2.903, 3.223, and 3.889 on CDR-H1, CDR-H2, and CDR-H3 respectively as shown on table 2.3. Next, we test models trained on antibody structure with masked CDRs, without CATH4.2 pre-training, our model achieved RMSD of 1.55, 1.00, 1.55 respectively. With model pre-trained on the CATH4.2 dataset and fine-tuned on antibody structure with masked CDRs, our model achieved RMSD of 0.81, 0.85, 1.35 compare to state-of-the-art method RefineGNN[2] with RMSD of 1.18, 0.87, 2.50 respectively. To avoid information leakage, we filtered the CATH4.2 dataset with the antibody test sets for redundancy in pre-training.

To summarize, our structure generative model can be adopted to perform structure inpainting by properly masking the input features. Through our experiment, we show the hybrid masking scheme can improve design performance and our model can adequately inpaint antibody CDRs. With pre-training on large structure databases, our model outperforms state-of-the-art antibody structure design models and we observe that our model has the largest improvement on longer inpainting regions(CDR-H3). In this chapter, we only performed unconditional CDR structure generation in this study with our structure generative model. We will further explore conditional CDR structure generation along with sequence design experiments in next chapter.

2.7 Discussion

This chapter of the thesis presents a novel structural generative model, CoordVAE, for protein design that addresses key limitations in existing computational approaches. The model demonstrates high-quality reconstruction of protein backbone structures, generates diverse conformational ensembles, and enables robust sequence design through a decoy sampling

approach. Additionally, it shows promising capabilities in structure inpainting for antibody design. These results have several important implications for the field of computational protein design and suggest exciting directions for future work.

One of the key innovations of CoordVAE is its ability to directly model protein structures in three-dimensional coordinate space while addressing rotational and translational equivariance. This allows the model to generate realistic protein conformations without relying on intermediate topological constraints or downstream coordinate recovery steps. The high-quality reconstructions achieved across a range of protein sizes and folds, as evidenced by results shown above, demonstrate the model’s ability to capture complex structural relationships. The clustering of latent representations by CATH fold classes further indicates that the model has learned meaningful structural embeddings.

The conformational decoy sampling approach enabled by CoordVAE represents a significant advance for fixed-backbone sequence design. By generating ensembles of backbone conformations, this method expands the search space for compatible sequences beyond what is possible with a single fixed template. The improved performance of sequences designed from decoy ensembles compared to native backbones, as validated by structure prediction, suggests this approach can lead to more designable and stable protein sequences. This could have major implications for *de novo* protein design efforts, potentially increasing success rates and expanding the range of achievable functions.

The model’s success in antibody CDR inpainting, particularly when pre-trained on a diverse protein dataset, highlights its potential for structure-based antibody engineering. The ability to generate plausible CDR conformations while maintaining the overall antibody framework could streamline antibody design processes and potentially lead to improved therapeutic candidates. The superior performance compared to existing methods, especially for longer CDR regions, is particularly promising.

Several aspects of the model design contributed to its strong performance. The use of

both distance and orientation information in the input features proved crucial for accurate coordinate reconstruction. The hybrid masking scheme during training likely enhanced the model’s ability to handle incomplete structural information, as evidenced by its success in inpainting tasks. The effectiveness of transfer learning from a diverse protein dataset to antibody-specific tasks suggests the model captures generalizable principles of protein structure.

While the results are promising, there are limitations and areas for future investigation. The current model focuses on backbone structure generation and does not directly address side chain packing or non-protein components like ligands or cofactors. Extending the model to incorporate these elements could further improve its utility for real-world design challenges. Additionally, while the conformational decoy approach shows clear benefits, further work is needed to understand the optimal sampling strategies and to develop methods for efficiently filtering and selecting the most promising decoys.

In conclusion, this work represents a step forward in computational protein design, introducing a versatile structural generative model that addresses key limitations of existing approaches. The demonstrated capabilities in structure reconstruction, conformational decoy sampling, and structure inpainting open up new possibilities for robust and efficient protein design. As the field continues to advance, integrating such generative models with other computational and experimental techniques promises to accelerate the development of novel proteins for a wide range of applications in biotechnology and medicine.

In the following chapters, I will employ the structural generative model developed and explore *de novo* protein design and protein optimization through a iterative framework and applying the model to challenging real-world protein design problems and validating designs experimentally. Extending the model to handle multi-chain protein complexes and further antibody design. However, other potential directions for future work includes integrating the structural generative model more tightly with sequence design algorithms, potentially in

an end-to-end differentiable framework and exploring the use of the learned structural embeddings for other tasks such as function prediction or protein-protein interaction modeling.

CHAPTER 3

ADAPTIVE *DE NOVO* PROTEIN DESIGN VIA ITERATIVE SEQUENCE STRUCTURE CO-OPTIMIZATION

3.1 Motivation

In the first chapter, we developed a deep structure generative model for protein structure generation. In this chapter we will explore how we can embed our model in a iterative design framework for robust and efficient *de novo* protein design.

Computational protein design has emerged as a powerful tool for creating novel proteins with targeted functional attributes, offering immense potential in fields ranging from therapeutics to industrial enzymes[70, 92, 186]. However, despite significant progress, the gap between computationally designed proteins and those that perform successfully in experimental settings remains substantial[50]. This challenge stems from the complex nature of protein folding and function, which involves a vast design space and intricate relationships between sequence, structure, and activity.

Current approaches to *de novo* protein design pipeline often rely on one-shot design with generated backbone templates[186, 177] which constraints the design space and therefore relatively low success rate. The complexity of protein design often requires satisfying multiple, sometimes competing objectives simultaneously, such as stability, solubility, and specific functional properties[142]. These multi-objective optimization goals are difficult to meet in a single design cycle, motivating the need for an iterative approach.

In this chapter of the thesis, we will outline the proposed iterative algorithm inspired by ideas from directed evolution[10] and evolutionary design algorithms such as genetic algorithm and evolutionary programming[159], detail its integration with our structure generative model, and discuss strategies for incorporating various forms of feedback into the design process. We will also present case studies demonstrating the effectiveness of this adaptive

approach in addressing challenging protein design problems, with the goal of advancing the field towards more reliable and efficient computational protein engineering.

3.2 Introduction

Protein design and engineering have become an indispensable tool in biotechnology, enabling the development of novel enzymes, therapeutics, and biomaterials with enhanced or completely new functions. One of the most successful approaches in this field has been directed evolution, a method that mimics natural evolution in a laboratory setting to optimize protein properties[10]. Directed evolution typically involves iterative cycles of genetic diversity generation followed by screening or selection for desired traits. This approach has led to numerous breakthroughs, including the development of enzymes with improved catalytic efficiency, stability, and even novel functions[126].

Despite its successes, directed evolution faces several challenges. The method is often labor-intensive and time-consuming, requiring the screening of vast libraries of mutants and often the fitness measure is technically sophisticated. Moreover, the random nature of mutations means that fitness advancing variants are rare, and multiple rounds of evolution are typically necessary to achieve significant improvements [10]. The method also struggles with navigating complex fitness landscapes, where beneficial mutations may be separated by fitness valleys that are difficult to traverse through random mutagenesis alone[135].

To address these limitations, researchers attempted to leverage machine learning techniques to accelerate directed evolution. This approach uses computational models to guide the evolution process, potentially reducing the number of variants that need to be experimentally tested and accelerating the discovery of improved proteins[185]. Machine learning algorithms can learn from previous experimental data to predict which mutations are likely to be beneficial, thereby focusing the search on more promising regions of the fitness landscape[118].

While machine learning-assisted directed evolution has shown promise, recent years have seen a surge in interest in purely computational protein design methods, particularly those based on deep generative models[177, 186, 175]. These approaches aim to design proteins *de novo* or to predict beneficial mutations without the need for iterative experimental testing.

However, these methods also face challenges. The predictions made by deep learning models can sometimes be difficult to interpret and often unreliable compared to experimental standard. where it's not always clear why certain designs are predicted to be successful and the generated backbones have limited designability. Additionally, while these models can generate numerous designs *in silico*, experimental validation is still necessary to confirm their function, and the success rate of purely computational designs remains lower than that of evolutionarily refined proteins[70].

Moreever, recent years have witnessed remarkable advancements in deep learning-based approaches for protein structure prediction and molecular simulation. The release of AlphaFold2 by DeepMind in 2020 marked a watershed moment in protein structure prediction, achieving unprecedented accuracy in the Critical Assessment of protein Structure Prediction (CASP) competition[84]. This was quickly followed by other powerful models such as RoseTTAFold[11] and ESMFold[107]. These breakthroughs have dramatically expanded our ability to model protein structures *in silico*. In the field of small molecule docking, deep learning methods have also made significant strides. Models like AtomNet[174] and DeepDock[103] have demonstrated improved accuracy and speed compared to traditional docking algorithms. More recently, end-to-end differentiable docking models like EquiBind[160] and DiffDock[27] have emerged, offering the potential for gradient-based optimization of binding poses. However, while these methods generally performs better in known ligands and computationally predicted protein structures, it still struggles to extend to novel ligand and protein pairs and high resolution local atomic interactions[22]. In practice, We found energy based models such as AutoDock Vina[40] generates more physically feasible

docking poses.

In an effort to harness the full potential of our structure generative model for practical *de novo* protein design and to enhance the robustness and efficiency of current protein engineering pipelines, we have developed an innovative iterative design framework. This framework adaptively optimizes multiple *in silico* fitness objectives within a flexible and modular pipeline. The versatility and power of our proposed framework are demonstrated through its ability to:

- Design *de novo* structures that are distinct from those found in current structural databases, thereby expanding the protein fold space.
- Optimize existing functional proteins with subsequent experimental validation, bridging the gap between computational predictions and real-world performance.
- Engineer scaffolds for functional proteins, resulting in the creation of *de novo* functional proteins with tailored properties.
- Structure based conditional *de novo* CDR design for antibody engineering.

3.3 Literature Review

In this Chapter, we will challenge our iterative design framework that utilizes previously developed structure generative model to an array of real-world design tasks such as unconditional protein structure generation, small-molecule binding . In this section, I will briefly describe those tasks and provide the context for our experiments.

3.3.1 Structure based protein design

Structure-based protein design is a fundamental approach to engineering proteins with specific functions, leveraging the intricate relationship between a protein's three-dimensional structure and its functions. Proteins achieve their functions through the precise spatial organization of amino acids, such as the hydrogen-bonded networks in enzyme active sites that create the ideal chemical environment for catalysis. Understanding this sequence-structure-function relationship is crucial for designing proteins with desired properties[70]. Conventional structure-based design pipeline can often be regarded as a three step process Fig.3.1, the first step is to build the structure blueprint of the protein with desired function and design objectives. The second step is to generate the amino acid sequences optimized for the structure templates that can stably fold and perform the desired function in its intended environments. The last step will be to score the designs to come up with the final design candidates for further in-depth validation and analysis. There have been many success by employing this design principle to design functional proteins in a variety of applications such as enzyme design[85, 90], protein binder design [171, 177, 49, 23], and biosensors[138].

Traditionally, protein design has relied on building 3D structural models to satisfy functional constraints derived from design objectives, using accurate energy models to guide atomic movements in simulated systems[70]. However, the advent of deep learning (DL) algorithms has offered unprecedented opportunities with data-driven methods. These DL approaches offer the potential to enhance the design process by capturing complex patterns in protein structures and sequences more effectively than conventional methods. For more background on methods for *de novo* structure generation and sequence design please refer to the Background section.

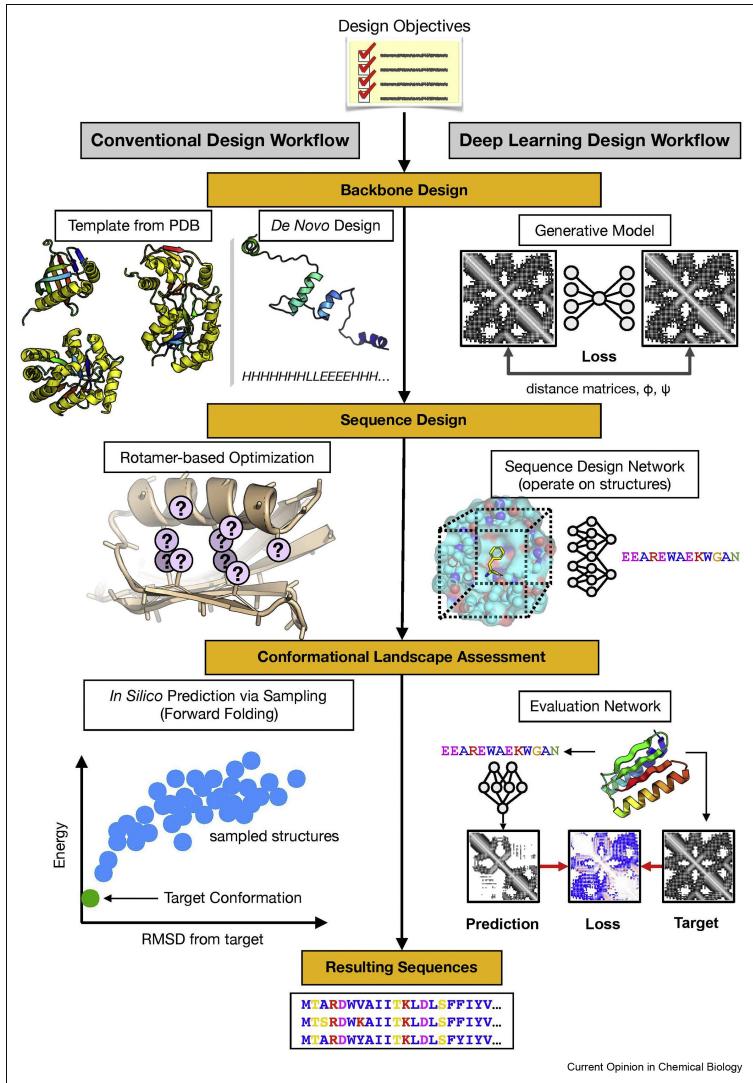


Figure 3.1: Structure-based protein design workflow illustration adopted from [125] CC-BY 4.0. Comparison between conventional structure-based protein design workflow vs. DL-based design workflow. For structure generation, the conventional approaches leverages existing structure fragments or functional motifs in the structure database and along with experts' knowledge to build the structure blueprint for subsequent design steps while DL-based approaches uses neural networks trained on the vast structure database that can generate structure templates. For sequence optimization, conventional approaches uses physics based method with energy minimization, DL-based methods use structure conditioned machine learning models to predict the amino acid sequences. For design scoring, conventional methods use energy based simulation such as molecular dynamics simulation or Rosetta energy to select viable candidates. DL-based workflow use *in silico* structure and property prediction models to evaluate the fitness of design candidates.

3.3.2 Unconditional structure generation

Generating novel protein structures without relying on specific constraints or predetermined templates has gained attention in the field of *de novo* protein design as deep generative models are being more and more popular as structure modeling tools. This approach aims to explore and expand the known protein fold space, potentially leading to the discovery of new structural motifs and functional proteins that may not exist in nature[70]. Traditionally, protein design relied on conventional backbone-design methods, which broke down structure design into hierarchical components of topology and syntax[91, 25]. These methods, while interpretable, were limited in their ability to explore the full space of designable sequences. The advent of deep learning has dramatically changed this landscape, offering new ways to manipulate protein structures in response to functional constraints[179].

Deep generative modeling has emerged as a powerful strategy for efficient sampling from high-dimensional distributions of protein structures [56, 88, 65]. Particularly noteworthy is the rise of diffusion-based generative models, which have shown remarkable success in protein design [8, 177, 101]. These models benefit from an iterative generation mechanism that aligns well with the hierarchical nature of protein structure, breaking down the structure-generation problem into high-level tertiary organization, followed by local secondary structure, and finally chemical detail. The stochastic nature of these generative model allows sampling structures that are previously not being observed in the nature and can potentially harbor novel topology and function. Methods such as GAN, Ginie, and FoldingDiff[7, 104, 180] all showed promising results in generating novel protein folds while [9, 177] experimentally validated numerous novel structures generated by their methods.

3.3.3 Computational enzyme and small-molecule binder design

Computational enzyme and small-molecule binder design have made significant success in recent years, leveraging advances in computational methods, deep learning, and experimental techniques. This field aims to create novel proteins with specific catalytic or binding functions from first principles, offering a complementary approach to traditional protein engineering methods. The general process of computational enzyme design typically involves designing a 'theozyme' - an idealized active-site model that includes a quantum mechanically calculated transition state and key functional groups from amino acid side chains required for transition state stabilization. This theozyme is then docked into structurally characterized proteins to identify suitable scaffolds, followed by redesigning residues in and around the active site to optimize interactions[90, 64].

This approach has led to the successful design of protein catalysts for various model transformations, including the Kemp elimination [144], retro-aldol reactions [78], and Diels-Alder reactions [156]. While initial designs often show low activity, they can be significantly improved through directed evolution [186, 29]. Notably, [186] used a generative model[9] to generate the scaffold with family-wide hallucination, distinct from others which use existing scaffolds. Although the initial success rate of functional design is quite low, this example paved a promising path to emerging *de novo* enzyme design. Despite these advances, challenges remain. Designing highly active enzymes with efficiencies comparable to natural systems is still difficult, and expanding the range of chemistries achievable with *de novo* enzymes remains a key goal [29]. To address this, researchers are exploring hybrid design strategies that combine the strengths of deep learning with fundamental biophysical understanding. These approaches aim to leverage the pattern recognition capabilities of machine learning while incorporating known principles of enzyme catalysis and protein structure.

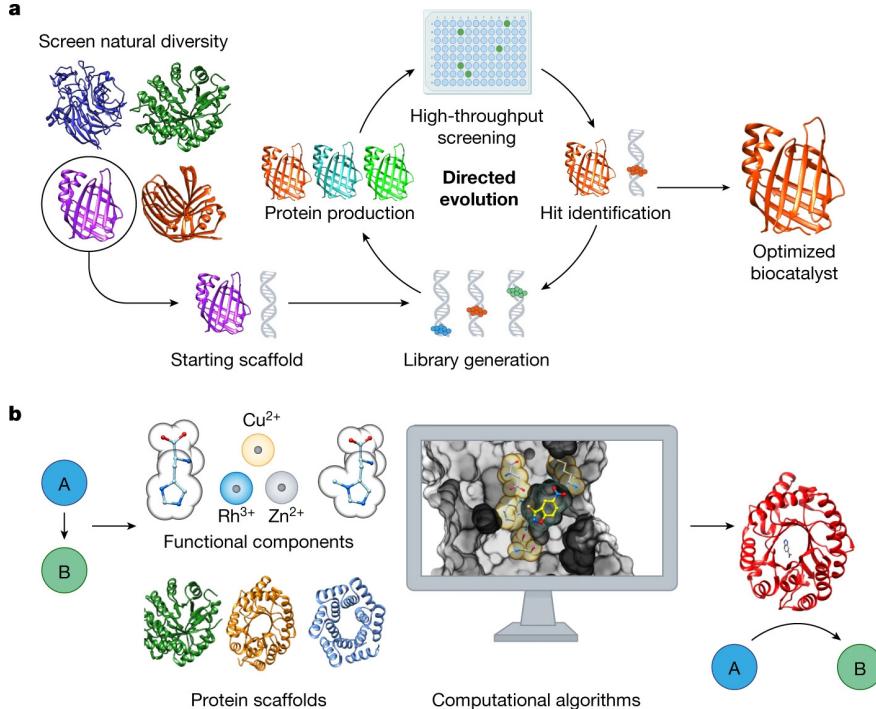


Figure 3.2: Enzyme engineering workflow illustration adopted from [108]. a) Conventional enzyme engineering workflow. Natural protein scaffold with desired structure and function is picked and fitness is optimized via directed evolution. b) Computational *de novo* enzyme design which starts by selecting or building suitable protein scaffold from scratch via generative models or simulation based filtering, then *in silico* design and scoring scheme is employed to produce design candidates for downstream validation.

3.3.4 Protein-protein binder design and motif scaffolding

Functional-motif scaffolding is a critical aspect of computational *de novo* protein design, with applications ranging from designing enzyme active sites to creating high-affinity binders. Recent advancements in deep learning methods, have significantly improved our ability to scaffold protein structural motifs that carry out binding and catalytic functions.

One common challenge of protein engineering is to optimize large and unstable proteins in its natural scaffolds and therefore difficult to apply powerful techniques such as directed evolution on these proteins. Interestingly, in many cases, the critical functional elements, such as active sites or binding interfaces, comprise only a small portion of the protein's over-

all structure. A promising strategy to overcome these limitations involves scaffolding these essential functional motifs into smaller, more stable structures. By transplanting active sites or key functional elements into compact, robust scaffolds, we can potentially create proteins that maintain their function and activity while gaining improved stability, solubility, and expressibility. Most importantly, this approach can increase their availability for experimental optimization. Methods such as [175, 177, 165] showed promising *in silico* results and [177] experimentally validated some of their designs.

3.3.5 *Structure based antibody design*

Structure based antibody design represents the grand challenge in therapeutic protein engineering. This approach aims to overcome the limitations of traditional antibody development techniques, which are often costly, laborious, and may not always produce antibodies that bind to the desired epitope on an antigen[18, 110]. By leveraging computational tools and structural information, structure-based antibody design offers the potential to rapidly create binders for specific targets, whether for combating new diseases or facilitating research.

Recent breakthroughs in computational structure prediction, particularly through deep learning methods, have ushered in a new era for structure-based antibody design[11, 84]. With the increasing availability of accurate protein structures, including those of antibodies and antigens, it is now possible to perform large-scale structural antibody virtual screening. This approach mirrors the successful strategies employed in small molecule drug development[57] and opens up new possibilities for antibody engineering.

Recent advancements in the field have focused on improving various aspects of the structure-

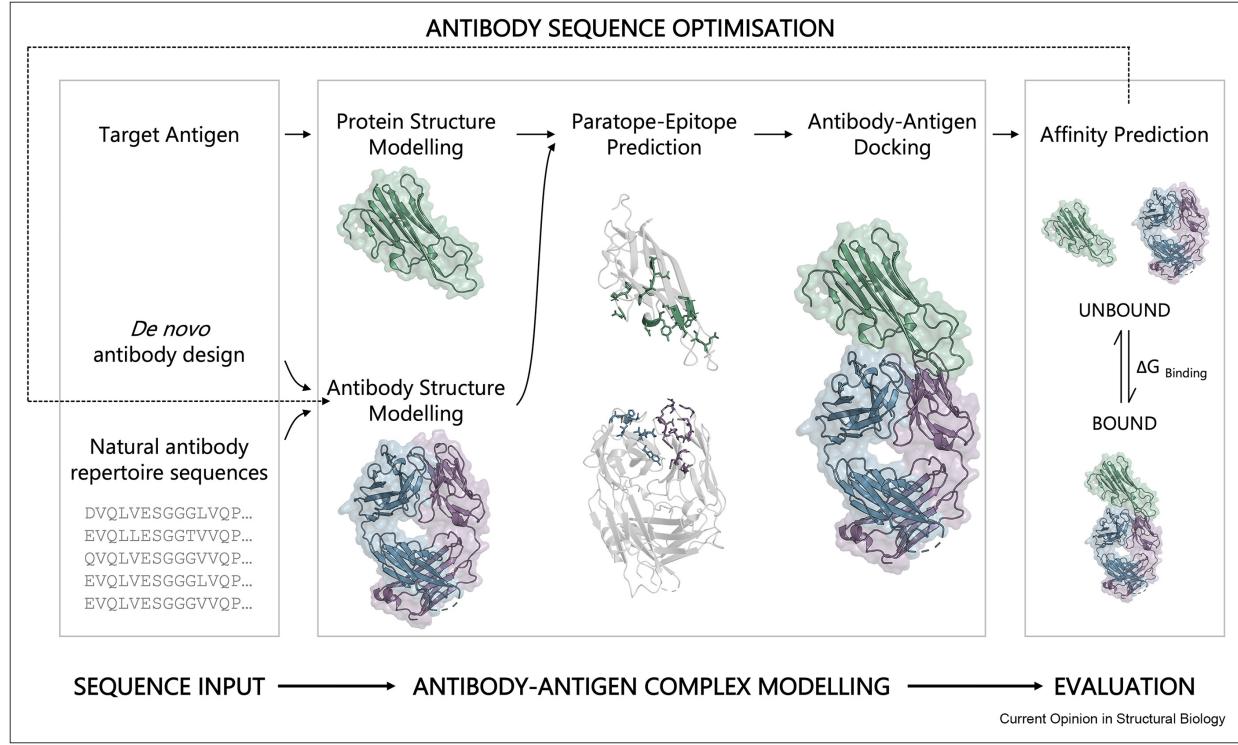


Figure 3.3: Structure based antibody design workflow illustration adopted from [72] CC-BY 4.0. Workflow of *in silico* structure based antibody design, the pipeline starts with a antibody framework of choice and the target antigen. A generative model or CDR minding can be used to initialize the seed AB candidates. Computational structure modeling tools and antibody docking tools will then be used to evaluate the designed complex *in silico*. A scoring function or binding affinity prediction tool will then be applied to filter the design candidates. The best candidates with a given metric will be presented as the resulting design. This process can be iterative and the best candidate can be fed back into the pipeline for further optimization and design.

based design pipeline. Deep learning-based methods specifically developed for antibody structure prediction, such as DeepAb[145] and ABlooper[1], have shown promising results in accurately predicting the structure of complementarity determining regions (CDRs), particularly the challenging CDR-H3 loop. These methods are not only more accurate but also substantially faster than general structure prediction tools, enabling rapid generation of large numbers of antibody structures. Progress has also been made in paratope and epitope prediction, which are crucial steps in assessing binding potential. Methods like PECAN[134], EPMP[34], and PInet[32] have demonstrated improved accuracy in predicting these binding

interfaces, with PInet achieving state-of-the-art performance in epitope prediction. These advancements contribute to more effective antibody design by helping to identify the key residues involved in antigen binding. While challenges remain, particularly in modeling antibody-antigen complexes[42, 4], these developments are paving the way for more effective virtual screening of antibodies against desired antigen targets.

On the other hand, antibody optimized sequence design models[66] that builds up on general inverse folding models[68] have emerged to address the challenge of CDR sequence design with promising *in silico* but lacking experimental validation. While current methods may not yet produce optimal binders directly, they provide a foundation for virtual screening and subsequent *in silico* affinity maturation. The integration of machine learning models to predict the effects of mutations on binding affinity[3] further enhances the potential for computational optimization of antibody-antigen interactions.

3.4 Methods

3.4.1 Overall Approach

In this section, we present a comprehensive overview of our iterative design framework, which integrates a structure generative model with multiple *in silico* structural and functional assessment oracles for design optimization. This versatile pipeline can be tailored to accommodate various design objectives, ensuring its applicability across a wide range of protein design tasks.

The design process initiates with an input structure template, the selection of which is flexible and task-dependent. For instance, when optimizing the structural stability of an existing functional protein, a complete structure template along with active sites can be provided. In scenarios requiring unconditional structure generation, a randomized amino

acid sequence may serve as the starting point. For partial structure scaffolding, the structure of the functional motif forms the initial template. Once a template is selected, it will be processed where it is distilled into one-dimensional and two-dimensional topological features. These features then serve as input for our structure generative model. Subsequently, the model generates a diverse library of structural decoys, which form the template pool for downstream sequence design.

Following structural generation, a fixed-backbone sequence design model[33] is employed to create a library of sequences compatible with the backbone library. These designed sequences are then evaluated using a state-of-the-art structure prediction oracle, such as AlphaFold2[84] or RoseTTAFold[11]. Sequences are filtered based on prediction confidence scores, with only those passing this *in silico* structure validation filter progressing to the next stage.

For designs where specific protein functions are desired, the filtered sequences undergo customized functional oracles such as computational docking and function prediction[40]. This step allows for the evaluation of the designed proteins' potential to perform the intended function. Candidates that successfully pass both the structural and functional filters are then ranked, with the top designs selected as input templates for the subsequent iteration. This iterative process continues until either the specified design objectives are achieved or a predetermined number of iterations is reached. At this point, the pipeline halts and outputs the best candidates from the accumulated design pool.

By integrating structure generation, sequence design, and multi-faceted evaluation within an iterative framework, our approach offers a powerful and flexible tool for protein engineering. It allows for the exploration of vast design spaces while maintaining a focus on both structural integrity and functional requirements. This methodology represents a significant step forward in our ability to design novel proteins with tailored properties and functions.

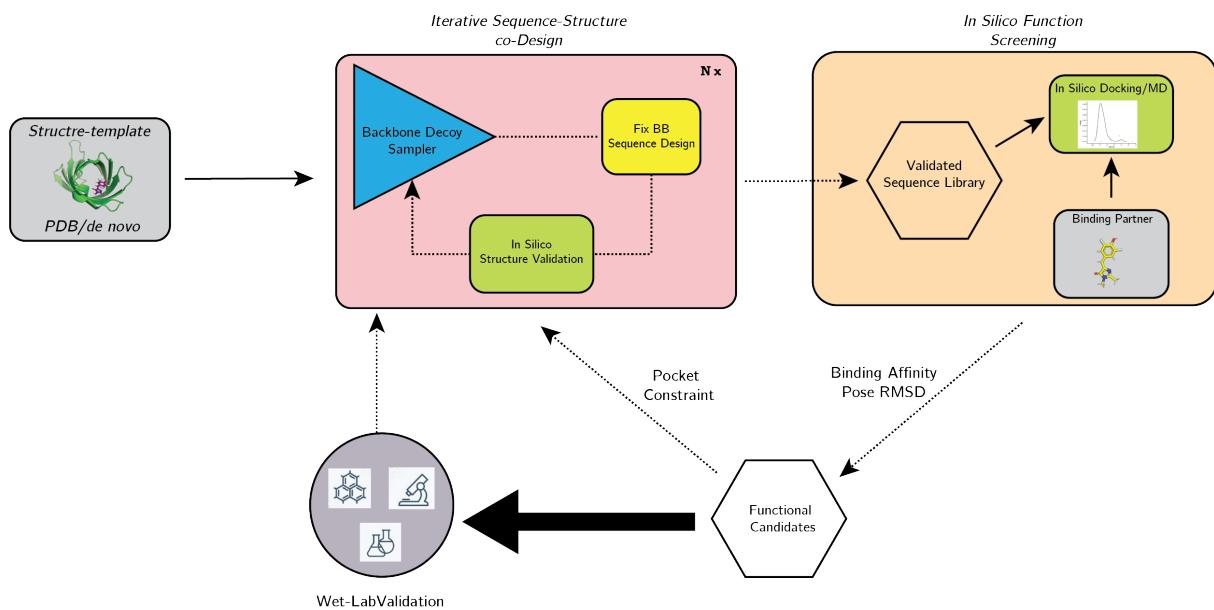


Figure 3.4: Overview of iterative design pipeline for adaptive protein design and optimization. The pipeline can be dissected into three modules, the design module which performs structure-sequence co-optimization on the input structure templates. The resulting structurally validated library is then fed into the scoring module which can include various *in silico* simulation and fitness prediction tools. The resulting functionally validated candidates are then selected as templates for next design iteration or further experimental validation.

3.4.2 Deep structural generative models

For designs presented in this chapter, we use the generative model illustrated in Figure 2.1. To prepare for input features, each protein structure is distilled into invariant pairwise representations of inter-residue distance and orientations as described in [184] and scalar representations of amino acid sequence and backbone torsion angles. This input is then fed through an encoder network which produces a latent representation of each residue. These representations are reassembled and passed to a decoder module which reconstructs the backbone coordinates. For more detail, see the Methods section in Chapter 2.

3.4.3 *in silico* validation and sequence design

Computational structure prediction: For *in silico* structure validation, we employed two protein structure prediction methods: AlphaFold2[84] and ESMFold[107]. To avoid evolutionary bias in *de novo* structure generation, we used AlphaFold2’s single sequence mode, which predicts structures based solely on amino acid sequences without relying on multiple sequence alignments or templates. For optimizing design throughput and partial structure scaffolding where evolutionary bias is preferred, we leveraged ESMFold, which offers faster prediction times and stable motif modeling. The choice between these tools was flexible, adaptable to specific design requirements.

computational docking: For computational molecular docking, we employed AutoDock Vina[40], a widely-used open-source program for protein-ligand docking. AutoDock Vina utilizes a sophisticated scoring function and efficient optimization algorithm to predict the binding modes of small molecules to protein targets. The docking simulations were performed using default parameters, with a search space centered on the predicted binding site of the target protein. The grid box dimensions were set to encompass the entire binding pocket, allowing for comprehensive sampling of possible ligand conformations and positions. Multiple independent docking runs were conducted for each ligand-protein pair to ensure

thorough exploration of the conformational space and to assess the consistency of predicted binding modes.

Fixed backbone sequence design: For fixed backbone sequence design, we employed ProteinMPNN[33], a deep learning-based method for protein sequence design. We primarily utilized the default settings for sequence generation, which have been optimized for a wide range of design tasks. Temperature-based sampling was applied selectively, depending on the specific requirements of each design task. In cases of functional protein design, we implemented constraints to fix the active sites, thereby preserving the critical functional elements of the target protein. For partial structure scaffolding tasks, we extended this approach by fixing not only the functional motif sequence but also its local structure. This strategy ensured the preservation of both the motif structure and the intended function of the designed proteins. These targeted constraints allowed us to explore sequence space effectively while maintaining essential functional and structural features, thus optimizing our protein design process for specific applications. For antibody design, we also employed AntiFold[66] for enhanced antibody specific design performance.

Structure DB search: To assess the novelty of the *de novo* generated structures, we used Foldseek[168] to perform structure search with the PDB[14] and the CATH[157] databases using structure alignment.

3.4.4 Experimental Validation Methods

Plasmid and strain construction

Genes were synthesized and inserted between the BamH I and Not I sites of pET-28b vector by GenecefeBiol (Jiangsu, China), resulting in the plasmids pET-28b. The plasmids were then transformed into E. coli BL21 (DE3) competent cells (TransGen Biotech, Beijing, China) for expression, 6 × His-tag was added to the N-terminus for purification purposes.

Protein expression and purification

The recombinant E. coli BL21 (DE3) cells were cultured in LB medium containing 50 μ g/mL kanamycin at 37 °C with shaking (200 rpm) to OD₆₀₀ = 0.8. Protein expression was induced by adding isopropyl β -D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM, and cell growth was continued overnight at 25 °C. The cells were harvested by centrifugation (10,000 \times g, 10 min, 4 °C).

For purification, the harvested cells were resuspended in buffer A (50 mM Tris, 500 mM NaCl, pH 8.0) and sonicated on ice to lyse the cells using a Scientz JY92-IIN sonicator (Ningbo, China). As the proteins were expressed with 6 \times histidine-tag in their N-terminus, they were then purified using a HiTrap™ Chelating HP column (Cytiva, MA, USA) and equilibrated with buffer A. Unbound proteins were eluted from the column with buffer W1 (50 mM Tris, 500 mM NaCl, 20 mM imidazole, pH 8.0) and buffer W2 (50 mM Tris, 500 mM NaCl, 50 mM imidazole, pH 8.0), respectively. Then, the proteins were eluted from the column with buffer B (50 mM Tris, 500 mM NaCl, 250 mM imidazole, pH 8.0). The purified proteins were further desalted using a Desalting column (Cytiva, MA, USA) with buffer C (20 mM Tris, 150 mM NaCl, pH 8.0).

The protein concentration was measured using the Bradford assay (Thermo Fisher, MA, USA), and the purity was determined using SDS-PAGE.

Fluorescence binding assay

Protein-activated DFHBI fluorescence signals were measured in 96-well plate format (Corning 3650) on a SpectraMax M2 plate reader (Molecular Devices, CA, USA) with $\lambda_{\text{ex}} = 467$ nm, cutoff = 495 nm, and $\lambda_{\text{em}} = 495$ to 595 nm. Binding reactions were performed at 200 μ L total volume in buffer C, containing 25 μ M proteins and 25 μ M DFHBI. DFHBI (Sigma) was suspended in DMSO as instructed to make 100 mM stock.

Differential scanning fluorimetry

The desugbed protein solutions (10 mg/mL) were diluted to 0.2 mg/mL in their respective buffered solutions (PBS at pH 7.4). Protein Thermal Shift Dye (4461146, Applied Bio-

sciences), initially provided at a concentration of 1000x, was diluted to a concentration of 8x using Milli-Q water. The antibody solutions ($12.5 \mu\text{L}$ in each well) were dispensed into 96-well white PCR plates (04729692001, Roche) in triplicate, $2.5 \mu\text{L}$ of the dye solution was added per well, and the solution was mixed by pipetting up and down ten times. The plates were then sealed with foil (04729757001, Roche Diagnostics). Thermal melts were performed using a LightCycler 480 real-time PCR instrument (Roche Diagnostics). The fluorescence (Ex: 558 nm, Em: 610 nm) was measured as the plate was heated from 25 to 99 °C. Many (>50) acquisitions were collected per 1 °C, and the heating rate was 0.6 °C /min. The apparent melting temperatures (T_m) of the GFP mutants were determined by analyzing the first derivative of the fluorescence with respect to temperature. This involved fitting a second order polynomial to the major peak and solving for the temperature at which the maximum occurred.

3.4.5 Iterative optimization algorithms

Genetic algorithms (GAs) represent a powerful class of optimization techniques inspired by the principles of natural selection and evolution[67]. These algorithms simulate the process of natural evolution, including inheritance, mutation, selection, and crossover, to solve complex optimization problems. In the context of computational protein design, GAs have emerged as a valuable tool for exploring the sequence and structural space of proteins that dates back to the 90s[83, 35, 61] for tasks such as optimizing side-chain conformations for protein core packing and functional site design.

One of the key advantages of genetic algorithms in protein design is their ability to efficiently sample large, complex search spaces[173]. This is particularly useful in protein design problems, where the number of possible sequences grows exponentially with protein length. GAs can navigate this vast space by maintaining a population of candidate solutions and evolving them over multiple generations, often leading to innovative and non-obvious

design solutions.

For unconditional structure generation we used the following design algorithm

Algorithm 1 Unconditional Structure Generation

```

1: procedure UNCONDITIONALSTRUCTGEN( $Seq_{init}, T, N_{decoy}, N_{template}$ )
2:    $Population \leftarrow []$ 
3:    $StructLib_{init} \leftarrow \text{StructurePredictor}(Seq_{init})$ 
4:    $StructLib_0 \leftarrow StructLib_{init}$ 
5:   for  $t \leftarrow 1$  to  $T$  do
6:      $Templatest \leftarrow \text{RankPTM}(StructLib_{t-1}, N_{template})$ 
7:      $DecoyLib_t \leftarrow \text{CoordVAE}(Templatest, N_{decoy})$ 
8:      $SeqLib_t \leftarrow \text{FixedBBDesign}(DecoyLib_t)$ 
9:      $StructLib_t \leftarrow \text{StructurePredictor}(SeqLib_t)$ 
10:     $Population \cup StructLib_t$ 
11:   end for
12:   return RankPopulation( $Population$ )
13: end procedure

```

To begin, a randomly sampled sequence pool Seq_{init} of length L is provided. For each iteration, we predetermine the number of templates we select for each design iteration and the number of decoys to generate for each template. Then we set the number of iteration to be T . In the design pipeline, AlphaFold2 with single sequence mode was used as StructurePrediction and ProteinMPNN was used as the FixedBBDesign algorithm with 1 sequence generated for each decoy. RankPTM is a function that rank a set of prediction structure by their pTM scores.

For functional protein optimization we used the following design algorithm

Algorithm 2 Functional Protein Optimization

```
1: procedure FUNCTIONALPROTOOPT( $Template_{init}$ ,  $T$ ,  $N_{decoy}$ ,  $N_{template}$ ,  $FuncSites$ )
2:    $Population \leftarrow []$ 
3:    $DecoyLib_0 \leftarrow \text{CoordVAE}(Template_{init})$ 
4:    $SqLib_0 \leftarrow \text{FixedBBDesign}(DecoyLib_0, FuncSites)$ 
5:    $StructLib_0 \leftarrow \text{StructurePredictor}(SqLib_0)$ 
6:   for  $t \leftarrow 1$  to  $T$  do
7:      $Templates_t \leftarrow \text{RankFuncRMSD}(Template_{init}, StructLib_{t-1}, FuncSites, N_{template})$ 
8:      $DecoyLib_t \leftarrow \text{CoordVAE}(Templates_t, N_{decoy})$ 
9:      $SqLib_t \leftarrow \text{FixedBBDesign}(DecoyLib_t, FuncSites)$ 
10:     $StructLib_t \leftarrow \text{StructurePredictor}(SqLib_t)$ 
11:     $Population \cup StructLib_t$ 
12:   end for
13:   return RankPopulation( $Population$ )
14: end procedure
```

We initiate our optimization pipeline with a functional protein template $Template_{init}$ of interest as well as the protein functional cites, for each iteration, we predetermine the number of templates we select for each design iteration and the number of decoys to generate for each template. Then we set the number of iteration to be T . In the design pipeline, AlphaFold2 with single sequence mode is used as StructurePrediction and ProteinMPNN is used as the FixedBBDesign algorithm with 1 sequence generated for each decoy and the functional cites were fixed in this step. RankFuncRMSD is a function that rank a set of prediction structure by their the RMSD w.r.t the input template.

For motif grounded protein scaffolding we used the following design algorithm

Algorithm 3 Motif Grounded Protein Scaffolding

```
1: procedure MOTIFGROUNDEDSCAFFOLD( $Template_{init}$ ,  $T$ ,  $N_{decoy}$ ,  $N_{template}$ ,  $MotifSites$ )
2:    $Population \leftarrow []$ 
3:    $SeqLib_0 \leftarrow \text{MotifFill}(Template_{init}, MotifSites, N_{decoy})$ 
4:    $StructLib_0 \leftarrow \text{StructurePredictor}(SeqLib_0)$ 
5:   for  $t \leftarrow 1$  to  $T$  do
6:      $Templates_t \leftarrow \text{RankMotifRMSD}(Template_{init}, StructLib_{t-1}, MotifSites, N_{template})$ 
7:      $DecoyLib_t \leftarrow \text{CoordVAE}(Templates_t, N_{decoy})$ 
8:      $SeqLib_t \leftarrow \text{FixedBBDesign}(DecoyLib_t, MotifSites)$ 
9:      $StructLib_t \leftarrow \text{StructurePredictor}(SeqLib_t)$ 
10:     $Population \cup StructLib_t$ 
11:   end for
12:   return RankPopulation( $Population$ )
13: end procedure
```

We initiate our design pipeline with a structure motif $Template_{init}$ of interest as well as the relative infilling among the segments of the motif, for each iteration, we predetermine the number of templates we select for each design iteration and the number of decoys to generate for each template. Then we set the number of iteration to be T . In the design pipeline, ESMFold is used as StructurePredictor and ProteinMPNN is used as the FixedBBDesign algorithm with 1 sequence generated for each decoy and the motif sequence were fixed in this step. RankMotifRMSD is a function that rank a set of prediction structure by their the RMSD w.r.t the input motif.

3.4.6 Data

DIPS for complex structures

We used the DIPS(Database of Interacting Protein Structures) dataset presented by [164]

to fine tune our structure generative model for binding surface generation. We capped the size of the complex structure to 500 amino acid to account for memory usage. Then we computed the binding surface with a 10A radius inter-chain contact, that is, any residue that is within 10A to any residue of the other chain. During training, the binding interface of one component of the protein complex is randomly masked.

Antibody Structures

For antibody structures, we used the structural antibody Database (SAbDab) obtained from [79] which contains 1266, 1564, 2325 structures for CDR-H1, CDR-H2, and CDR-H3 respectively after filtering and splitting, there is no more than 40% sequence identity in the inpainted regions for each set of structures between the train/test structures. Please refer to [79] for further details.

3.5 Results

Our iterative design framework demonstrates high versatility and efficacy across a spectrum of protein engineering challenges. In this section, we present the results of three distinct applications, each showcasing a different aspect of the framework’s capabilities. First, we test our model’s ability to design *de novo* protein folds by unconditional structure generation, where our approach successfully generates novel protein folds with high *in silico* folding confidence. This not only expands the known protein structure space but also demonstrates the potential of our framework to explore beyond naturally occurring protein architectures. Second, we present a *de novo* design of a small molecule-activated fluorescent protein. This case study illustrates the framework’s ability to engineer proteins with complex, environmentally responsive functions. We provide both *in silico* and experimental validation, offering a comprehensive view of the design process and its outcomes. Lastly, we demonstrate the framework’s ability to perform structure-based protein engineering through a motif-grounded

scaffolding of PD1 (Programmed Cell Death Protein 1). This application highlights the potential of our approach in the field of therapeutic protein design, showcasing how functional motifs can be integrated into novel structural contexts. Together, these results underscore the power and flexibility of our iterative design framework, spanning from fundamental advances in protein structure exploration to the creation of functional proteins with potential real-world applications.

3.5.1 *Unconditional Structure Generation*

The ability to generate novel protein structures is a fundamental challenge in protein engineering, with implications ranging from understanding protein evolution to designing new functional proteins. Traditional approaches to protein design have often been limited by the known protein fold space, typically relying on modifications of existing structures [70]. Our iterative design framework, however, aims to push beyond these boundaries by enabling the generation of entirely new protein folds for a wide range of sizes that do not naturally occur.

In 3.5.A we show an example of structure evolution through our iterative design algorithm, which we found in practice, structures converges within 10 iterations with 5-10 templates chosen at each round of design. Also, we found it advantageous to use the predicted Template Modeling (pTM) score as the primary fitness criterion, rather than relying solely on pLDDT. This strategy helped us avoid a bias towards long helical structures, which computational structure predictors tends to favor for higher pLDDT scores. By incorporating pTM, we were able to generate a more diverse set of structures with varied secondary structure compositions. As illustrated in Fig. 3.5.B, our approach successfully produced designs spanning a wide range of secondary structure elements, including α -helical, β -sheet, and mixed α/β topologies. This diversity in secondary structure composition demonstrates the versatility of our framework and its ability to explore a broad spectrum of protein folds, rather than being limited to particular folds. We began by assessing our framework’s ability to generate

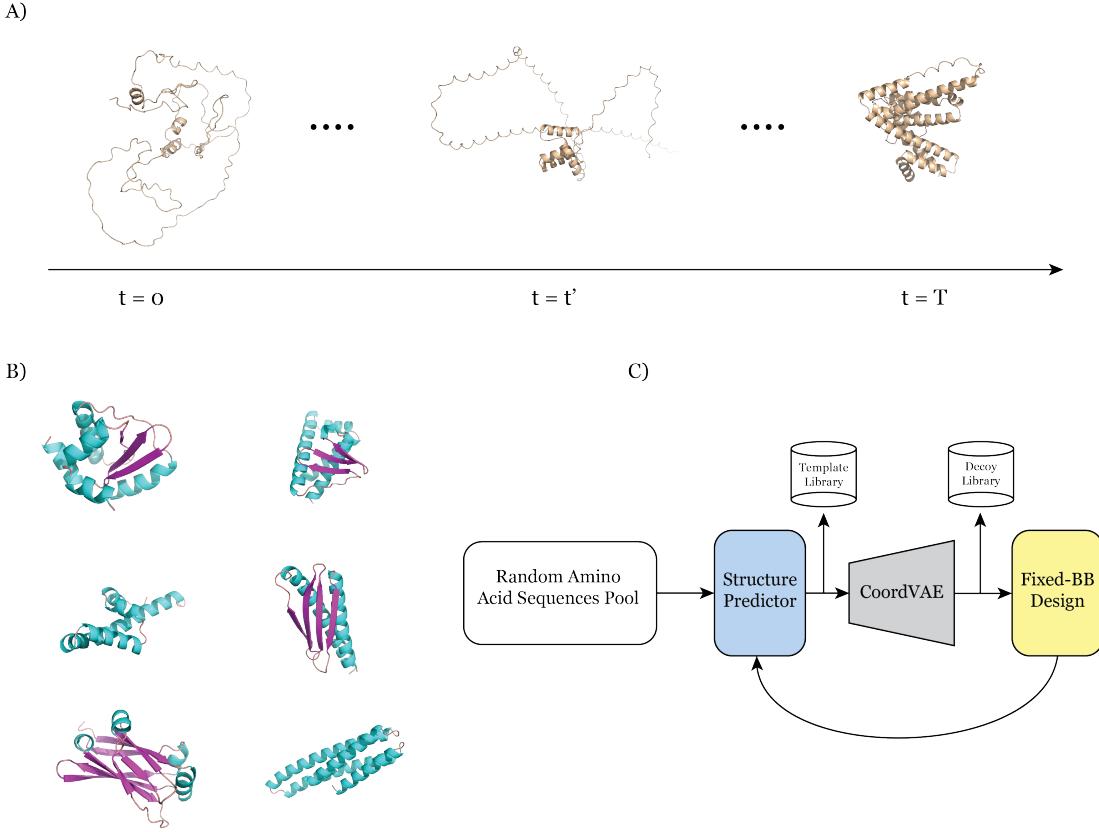


Figure 3.5: Unconditional Generation of monomeric structures. A) Illustration of iterative structure evolution by design iterations. $t=0$ indicates random initialization of amino acid sequences, $t=T$ indicates design convergence. B) Example of generated *de novo* structures, helical structures are colored cyan and beta strands are colored purple. C) Conceptual graph of the iterative design framework.

protein sequences that can be folded *in silico* with high confidence. This is a crucial first step, as it demonstrates the framework's capacity to produce designs that are likely to adopt stable, well-defined structures[9]. We generated a diverse set of protein sequences using our adaptive algorithm, iteratively refining them based on *in silico* folding predictions. To evaluate the quality of our designs, we employed state-of-the-art protein structure prediction tools, including AlphaFold2 [84] and ESMFold[107]. We considered a design successful *in silico* if it achieved a high predicted Local Distance Difference Test (pLDDT) score, which indicates the confidence of the structure prediction [6]. Our framework was able to generate a substantial number of sequences with high pLDDT scores(≥ 90), indicating very high

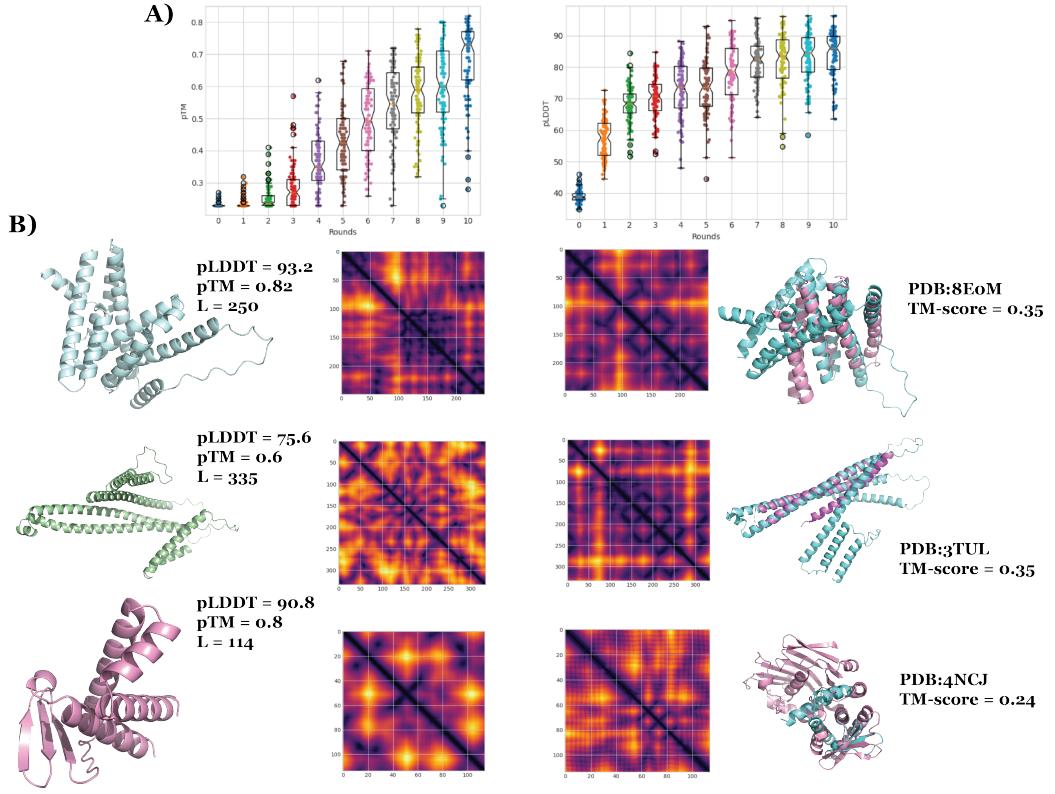


Figure 3.6: **Unconditional Generation of novel protein structures.** A) Example pLddT and pTM evolution with design iterations. C) Column 1, *de novo* generated protein folds. Column 2, initial inter-residue distance map of randomly initialized sequences. Column 3, inter-residue distance map for the *de novo* generated protein folds. Column 4 overlay of *de novo* generated protein folds with its closest hit from the PDB. In this case, we found designs from our iterative design framework to exhibit both high *in silico* folding viability and structural novelty.

confidence in their predicted structures 3.6.A.

Next, we evaluated the novelty of the generated structures. To assess this, we compared our designs against known protein structures in the Protein Data Bank (PDB) using both global and local structural alignment methods [168]. We found that a significant proportion of our high-confidence designs exhibited structural features that were distinct from any known protein fold. This novelty is demonstrated in Fig. 3.6B, which showcases three representative designs of varying lengths and their closest structural matches in the PDB.

For the first design shown in Fig. 3.6B, the closest structural homolog in the PDB (PDB ID: 8EOM) has a TM-score of only 0.35, indicating substantial structural divergence from known folds. We also demonstrated the ability to generate novel folds for longer proteins. The second design in Fig. 3.6B, a protein with 335 amino acids, found its closest structural match in the PDB (PDB ID: 3TUL) with a TM-score of 0.35. We observed that our method could produce diverse and novel folds for shorter proteins as well. The third design in Fig. 3.6B, despite its smaller size, exhibited remarkable novelty. Its closest structural match in the PDB (PDB ID: 4NCJ) had a TM-score of only 0.24. These results collectively demonstrate the capability of our iterative design framework to generate protein structures that are substantially different from any known folds across a range of protein sizes.

Our iterative design framework demonstrated success in generating novel protein structures with high *in silico* folding confidence. We produced a diverse array of protein designs spanning various sizes and secondary structure compositions. Notably, many of these designs exhibited significant structural novelty, with low TM-scores to any existing experimental structures when compared to their closest matches in the Protein Data Bank. This ability to generate stable, novel protein folds across different protein sizes potentially pave new paths for exploring protein structure-function relationships beyond the confines of naturally evolved proteins. In contrast to previous methods such as network hallucination[9] and diffusion models[180, 104], our model is much more compute efficient and flexible. One can use other fitness criterion such as topology preference and the globularity. However, there are several limitations and considerations to keep in mind. The results are based *in silico* on predictions, and experimental validation of these structures is crucial to confirm their stability and folding in real-world conditions. While the designs show novelty compared to known structures, their functional potential remains to be explored. The reliance on computational structure prediction tools like AlphaFold2 and ESMFold means that the method's success is partially dependent on the accuracy and limitations of these prediction tools. The approach

may still have biases or limitations in the types of folds it can generate, which may not be immediately apparent from the *in silico* results. Future work should focus on experimental validation of these designs, exploration of their functional potential, and further refinement of the design algorithm to address any biases or limitations

3.5.2 *De novo* design of DFHBI activated fluorescent protein

Computational design of functional proteins represents a frontier in protein engineering, offering the potential to create tailored molecular tools for a wide range of applications in biotechnology and medicine[70]. This approach not only allows for the optimization of existing protein functions but also enables the creation of entirely new functionalities not found in nature. Despite many recent advances in the field, there is still a gap between the computationally generated design and their experimental success. In this section we demonstrate how to utilize our iterative design framework for *de novo* functional protein design with 100% experimental success rate and improved thermal stability and production yield.

Fluorescent proteins, have been a prime target for computational design due to their immense utility in biological imaging and their relatively well-understood structure-function relationships[131]. While naturally occurring fluorescent proteins like GFP have been widely used, there is a growing demand for proteins with tailored properties, such as specific activation mechanisms or spectral characteristics [143]. Small molecule-activated fluorescent proteins are especially valuable as they allow for temporal control of fluorescence, enabling more precise experimental manipulations [37].

In this study, we focused on further developing and optimizing a DFHBI-activated β -barrel fluorescent protein previously designed by the Baker lab[37]. DFHBI (3,5-difluoro-4-hydroxybenzylidene imidazolinone) is a small molecule that becomes fluorescent when bound in a specific pose, making it an ideal candidate for designing controllable fluorescent sys-

tems3.7.A. Leveraging our iterative design framework, we set out to enhance the properties of this DFHBI-activated fluorescent protein. Our goal was to create a variant with improved fluorescence characteristics, higher stability, and more precise small molecule control, while maintaining the β -barrel scaffold that has proven effective for fluorescent proteins. Building

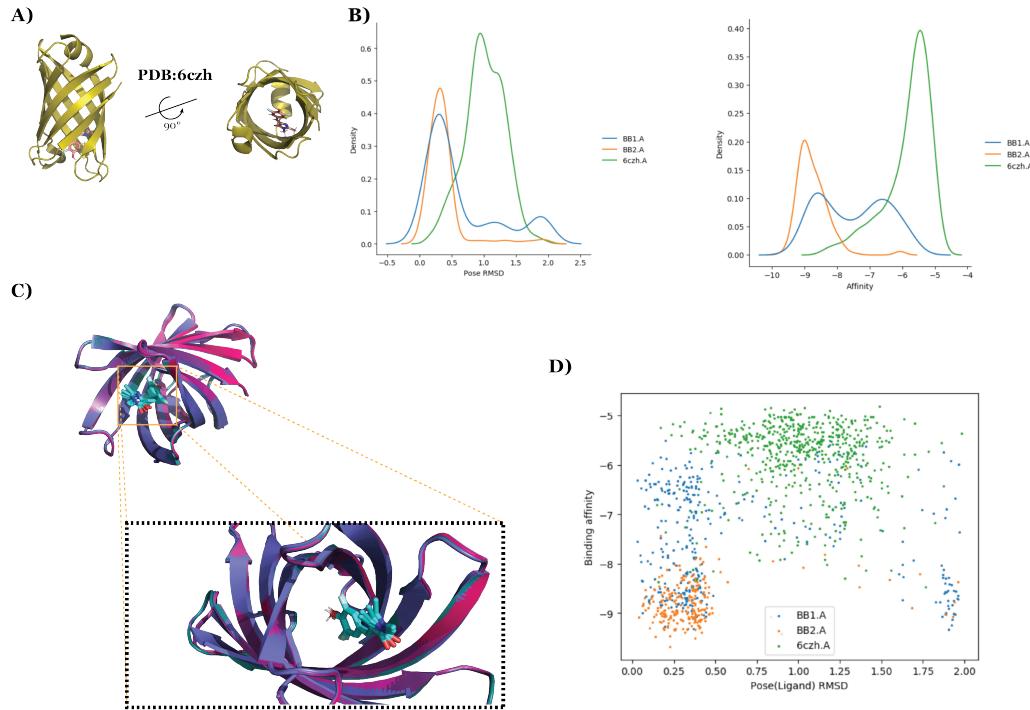


Figure 3.7: De novo design of DFHBI-activated fluorescent β -barrels A) PDB:6CZH previously designed DFHBI-activated β -barrel. B) Distribution of ligand pose RMSD(Left) of the *de novo* designed β -barrels with starting template 6czh.A and the first design batch BB1.A and the second design batch BB.2. Distribution of computational docking scores from Autodock Vina(Right) of the *de novo* designed β -barrels with starting template 6czh.A and the first design batch BB1.A and the second design batch BB.2 C) Example overlay of the computationally docked DFHBI with the designed β -barrels. D) Scatter plot of the computationally docked scores vs. the ligand RMSD

upon the initial DFHBI-activated β -barrel fluorescent protein designed by the Baker lab, we employed our iterative design framework to enhance its properties. The design process began with the generation of structural variants based on the initial design template. We used our structural generative model to generate an backbone library, preserving the DFHBI binding pocket and the surrounding residues that influence binding characteristics. A key

component of our *in silico* evaluation was the use of AutoDock Vina[40] for computational docking simulations. For each generated structure, we performed docking simulations with DFHBI to assess binding affinity and pose. The docking results were crucial in our iterative design algorithm, serving as a primary criterion for selecting promising candidates for further refinement. Candidates were first filtered for *in silico* folding viability and selected using the computational docking metrics. As shown in 3.7.B, the distribution of the first round of design compared to subsequent design shown significant improvement of both the ligand pose position and the computational affinity scores. On 3.7.C we examined the ligand docking positions with the computationally designed proteins. We found stable and consistent functionally active docking simulated candidates. After multiple rounds of iteration and refinement, guided by the docking results and other computational analyses, we identified a promising candidate for further experimental validation. From 3.7.D, we observed the progression of functional metric improvement through design rounds towards lower ligand post RMSD and better simulated binding affinities. For the designs we used two different structure design templates that are previously characterized experimentally(6CZI and 6CZH).

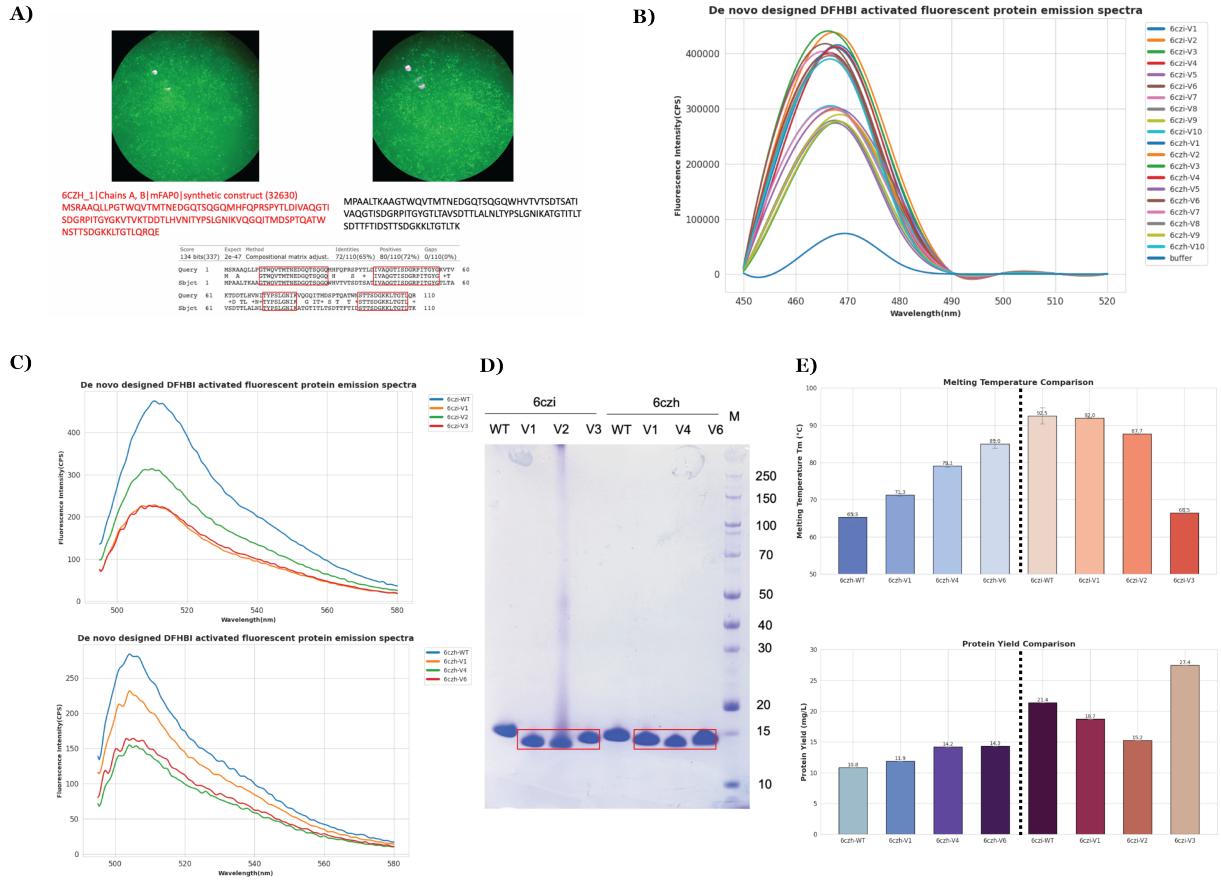


Figure 3.8: Experimental validation of *de novo* designed DFHBI-activated fluorescent β -barrels A) Fluorescence microscope image of *de novo* designed protein(Left) and the design template mFAP0. B) Fluorescent emission spectra from the lysates of 20 tested designs and the templates with reference buffer. All of the 20 designs were found with detectable fluorescent emission which represents a 100% design success rate. C) Fluorescent emission spectra of purified designed proteins and their respective original templates of 6czi(up) and 6czh(down). Although we did not observe increased fluorescent peak, we saw slight shift in the emission frequency compared to the design template. D) SDS-PAGE of purified designed protein with its respective design templates. All the designed proteins are observed to be smaller than the reference template. E) Protein thermal stability assay results(top), and protein yield comparison results(bottom). We see improvement of thermal stability on 6czh-based designs and comparable thermal performance in 6czi-based designs. For protein yield, all of the 6czh-based designs exhibited improvement over the design reference and 6czi-V has a significant increase over the reference.

We selected the top 20 designs from our computational pipeline for initial experimental screening. These designs were expressed in *E. coli*(See Methods for details) using standard protocols. Crude cell lysates were then subjected to a fluorescence assay in the presence and

absence of DFHBI. This initial screen allowed us to rapidly assess the fluorescence activation of our designs and compare their performance to the design templates. Results from this initial screen revealed that 20 out of 20 designs showed detectable DFHBI-activated fluorescence with a 100% success rate(3.8.B). All the tested designs have sequence similarity to the design template greater than 74% with the lowest candidate to have 65% sequence identity to the design template as shown in 3.8.A.

Based on the results of the initial screen, we selected the top 6 designs for more comprehensive evaluation. These proteins were expressed at a larger scale and purified to homogeneity using affinity chromatography followed by size exclusion chromatography. We then perform the fluorescent assay again on the purified proteins and saw very comparable emission peaks to the design templates as seen in 3.8.C. Although we did not see increased fluorescent intensity in any of our designs, the low sequence similarity to the original templates showed a expanded functional sequence space and further wet lab optimizations such as directed evolution can be employed to improve intensity. For thermal stability, we tested the melting temperature T_m of the the *de novo* designed proteins, we found that for 6cjh based design, all 3 purified candidates showed improved thermal stability and the best one reached 85°C which is 26% improvement over the template. In the case of 6czi based design, we achieved very comparable T_m despite the design template already has a high melting temperature of 92°C. For protein expression yield, we achieved increased unit expression yield in both design case as seen in 3.8.E.

Building upon a previously designed DFHBI-activated β -barrel fluorescent protein, we employed our iterative computational framework, incorporating AutoDock Vina[40] for docking simulation as a fitness criterion. Our design approach yielded promising *in silico* design success, as evidenced by computational docking and folding simulations. Remarkably, we achieved a 100% design success rate from 20 computationally designed candidates tested

experimentally. Our designs based on the 6czh template demonstrated improved thermal stability, with the best candidate reaching a melting temperature of 85°C, representing a significant 26% improvement over the original design template. Furthermore, we observed increased protein expression yield for designs based on both 6czh and 6czi templates.

Despite these achievements, our study revealed certain limitations. Most notably, the designs did not achieve increased fluorescence intensity compared to the original templates. This inability to produce higher intensity candidates could be attributed to several factors. The design process may have prioritized properties such as stability or expression yield, potentially at the expense of binding affinity. To enhance structural rigidity and stability, our designs may have inadvertently limited the flexibility of the binding pocket allosterically. Interestingly, we observed a small emission peak shift in the 6czi-based design, indicating a possible change in binding pose induced by our design. However, the general emission peak difference was maintained between the two design templates. We hypothesize that the *de novo* design process has allosteric impacts on the binding pocket, which warrants further investigation.

The differential success in improving thermal stability between the two templates is particularly noteworthy. While we saw significant improvement in thermal stability for the 6czh-based design, the 6czi template already possessed high thermal stability. Our best design based on 6czi was comparable in stability despite low sequence similarity, which is an encouraging result. Additionally, we achieved improved yield from both design templates. This outcome is not entirely surprising, as the original design was optimized for yeast display, whereas our experiments were conducted using *E. coli*. Our optimization pipeline did not incorporate any expression system bias, which may explain the improved yields across different hosts.

Compared to previous design methods, our iterative design framework is fully *in silico* and we leveraged DL based tools in all stages of the design pipeline. Unlike targeted mutation strategies based on predicted single mutational effects, our method enabled us to change upwards of 35% of the total amino acids in the resulting designs. Despite this substantial sequence divergence, we successfully maintained comparable function to the original templates while simultaneously improving other properties such as thermal stability and expression yield. We see this as a promising first step towards practical *in silico* protein design and optimization.

These findings point to several directions for future research. Refining our computational model to better capture the determinants of fluorescence intensity could lead to more comprehensive improvements in protein design. Exploring ways to overcome the apparent trade-offs between different desirable properties, such as stability and fluorescence intensity, would be valuable. The study suggests that our computational approach could be particularly effective for improving proteins that are not already highly optimized, as demonstrated by the successful enhancement of the 6czh-based designs.

To further validate our hypotheses and refine our understanding, more controlled experiments could be conducted. These might include detailed structural studies to examine the binding pocket changes, systematic exploration of the sequence-structure-function relationship in our designs, and comparative studies across different expression systems. Such investigations could provide deeper insights into the complex interaction between protein stability, expression, and function.

3.5.3 *de novo* protein binder design via motif grounded scaffolding

Protein scaffolding is a crucial tool in protein engineering that involves transplanting functional motifs from one protein context to another, often more stable and versatile, protein scaffold[175]. This technique has numerous applications in biotechnology and therapeutics, allowing for the creation of proteins with desired functions in optimized structural contexts. Scaffolding can improve protein stability, enhance expression levels, and even modulate the activity of the transplanted motif. To trial our proposed iterative design framework's ability to embed known functional motif into new protein scaffolds by iteratively co-optimizing structure and sequence around the dedicated motif of interest as seen in 3.9, we performed an excise on computationally scaffold the PD1/PD-L1 complex and validated our design with *in silico* binding and folding experiments.

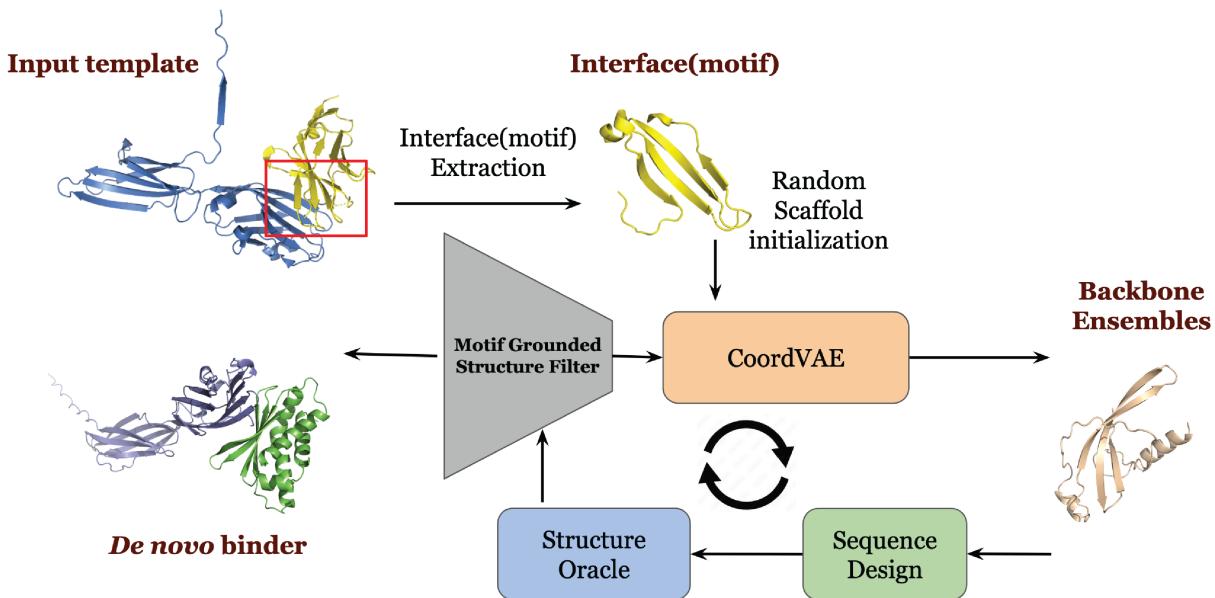


Figure 3.9: Iterative design workflow for motif grounded protein scaffolding: The design process begins with the identification of the functional motif interest, this can be done by extracting protein-protein binding surfaces or enzyme active sites. Then a randomly initialized scaffold is used as the seed template. The iterative design algorithm is then applied with a motif grounded structure fitness criterion.

Programmed Cell Death Protein 1 (PD1) is a crucial immune checkpoint receptor that

plays a significant role in regulating T-cell responses[153]. The interaction between PD1 and its ligands (PD-L1 and PD-L2) is a key target for cancer immunotherapy[162]. However, the use of native PD1 protein or its extracellular domain in therapeutic applications can be challenging due to stability and manufacturing consideration[51]. In this study, we aimed to use our iterative design framework to scaffold the key binding motif of PD1 onto a more stable protein structure. Our goal was to create a novel protein that maintains PD1’s binding specificity and affinity for its ligands while potentially offering improved biophysical properties.

We began by identifying the binding surface residues of PD1 involved in its interaction with PD-L1, based on available structural and mutational data [130]. This binding motif served as the anchor of our scaffolding exercise. We then randomly initialize a scaffold as the starting point then employed our iterative design pipeline to design the new scaffolds for the binding surface using the motif anchored *in silico* folding confidence as the selection criteria(See Methods for details). We performed 20 iterations in our design algorithm for scaffolding, we identified a set of promising candidates with motif centered structure RMSD and *in silico* folding confidence. Our top design, is a 142-residue protein that successfully incorporates the PD1 binding motif into a novel α/β fold in contrast to the native β only fold of PD1. We first computationally predicted the structure of the *de novo* designed protein binder with PD1 binding motif and filtered with pLDDT score for *in silico* folding viability and picked the top candidates for further evaluation. We showed the progression of the pLDDT distribution across all design rounds 3.10.D. We then computationally predicted the structure of the *de novo* designed PD-1 binder with PD-L1 and most of the prediction showed high pLDDT and iPTM score from their AF-multimer models. In the example shown in 3.10.E, the predicted complex showed a iPTM score of 0.87 and the PD1 binding surface has a RMSD of only 0.72A, we also highlighted the polar contacts in the predicted complex model.

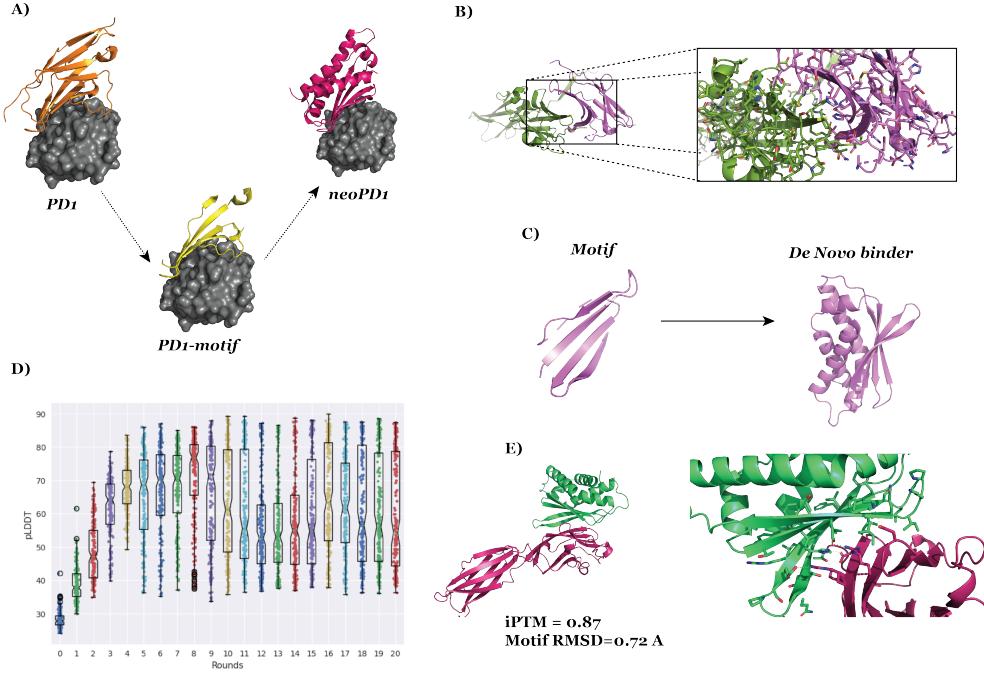


Figure 3.10: Motif grounded protein binder scaffolding for PD1/PD-L1. A) PD1 in complex with PD-L1 at different stage of the design process. PD1/PD-L1(left), PD1-motif/PD-L1(middle), and neoPD1/PD-L1(Right) B) Zoomed in image of the binding surface of the PD1/PD-L1 complex (PDB:5IUS) with polar contacts annotated. C) Starting motif and the *de novo* scaffold binder comparison. D) pLDDT distribution vs. design iterations for the scaffolding process. The structures are mostly converged after 10 iterations, more rounds are conducted to obtain more viable design candidates. E) Predicted complex structure of the design scaffold binder of PD1/PD-L1. Zoomed in image of the binding surface with polar contacts annotated.

In summary, this study employed a novel iterative design pipeline that begins with identifying the binding surface residues of PD1 involved in its interaction with PD-L1. Using this binding motif as an anchor, we randomly initialized a seed scaffold and then iteratively designed new scaffolds around the motif, using *in silico* folding confidence and motif focus structure RMSD as the selection criterion. After 20 iterations, we identified a set of promising candidates, with the top design being a 142-residue protein that successfully incorporates the PD1 binding motif into a novel α/β fold, contrasting with the native β -only fold of PD1.

The results of this study are promising. The designed protein showed high predicted folding stability (as indicated by pLDDT scores) and maintained the critical binding interface. Computational predictions of the designed PD-1 binder with PD-L1 showed high pLDDT and iPTM scores from AF-multimer models, with one example showing an iPTM score of 0.87 and a binding surface RMSD of only 0.72Å compared to the native structure.

Compared to other methods that is capable of scaffolding[177, 175, 165], our method generates the whole protein as a coherent entity without hard fixing any coordinates. we also demonstrates the ability to use flexible fitness selection criterion which previous models can not incorporate. This level of structural redesign is also challenging for traditional protein engineering methods that often rely on more conservative modifications. This *de novo* PD1 design showcases the power of our iterative framework in tackling complex protein engineering challenges, potentially paving the way for new approaches in cancer immunotherapy and beyond.

3.5.4 Structure based Antibody design via conditional CDR inpainting

Antibody engineering is a crucial field in biotechnology and therapeutic development, with applications ranging from cancer immunotherapy to the treatment of autoimmune diseases[110]. The complementarity-determining regions (CDRs) of antibodies are primarily responsible for antigen recognition and binding, making them key targets for design and optimization. Traditional approaches to antibody design often rely on display technologies or in vivo methods, which can be time-consuming and limited in their exploration of sequence space. In recent years, computational approaches have shown promise in accelerating antibody design[26]. However, many of these methods focus on sequence-based design or require extensive prior knowledge of antibody-antigen interactions. Structure-based approaches that can generate novel CDR structures while maintaining the overall antibody architecture are still quite

challenging[58]. Building upon our success with the iterative design framework, we devel-

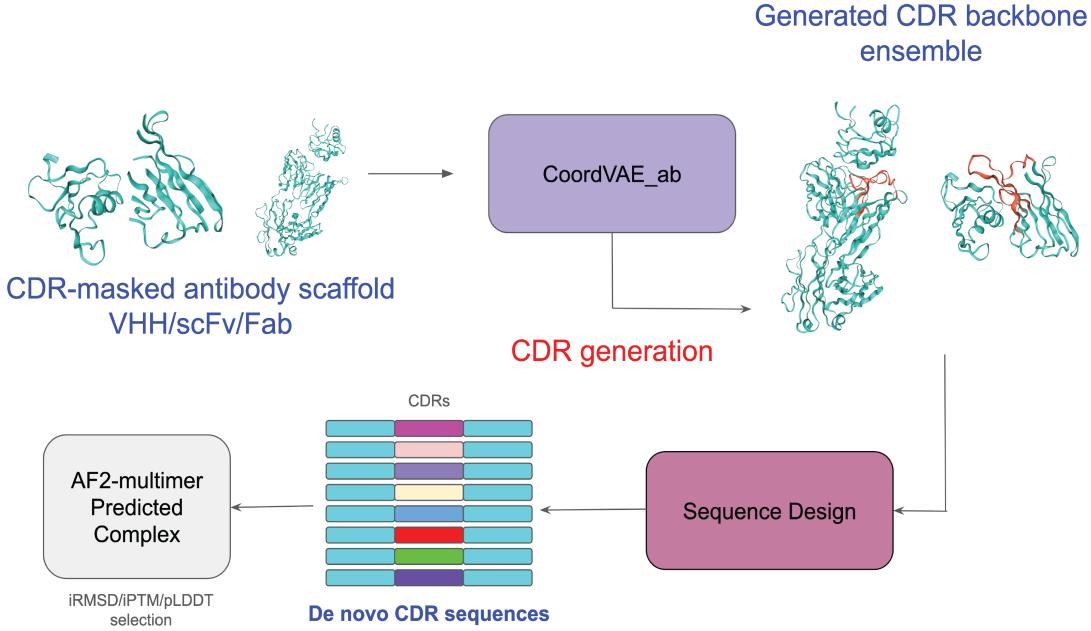


Figure 3.11: Structure-based antibody design via CDR inpainting workflow: The design process begins with a antibody-antigen pair, then the CDRs of interest are masked. CoordVAE-ab will then be used to inpaint the masked CDRs to generate a CDR backbone library. The structure library are then fed to a antibody optimized inverse folding model to recover the CDR sequence. A CDR library is then generated for downstream *in silico* structure prediction tool for selection.

oped a specialized approach for *de novo* CDR design through conditional structure inpainting. This method aims to generate novel CDR structures that are compatible with the antibody framework and optimized for target binding. To adapt our framework for CDR design, we made several key enhancements. First, we expanded our training dataset to include a comprehensive protein-protein interaction (PPI) dataset to capture diverse binding interfaces[164], as well as a curated antibody dataset[39]. We then fine-tuned our CoordVAE model on this expanded dataset with a two-stage approach, where the first stage focused on general PPI interaction and a particular focus on antibody structures and CDR regions in the second stage. The model was modified to perform conditional inpainting, allowing us to generate new CDR structures while maintaining the rest of the antibody framework. Impor-

tantly, we incorporated antigen structure information into the design process to guide the generation of CDRs with potential binding affinity. These enhancements collectively enabled our framework to tackle the specific challenges of structure-based antibody design, leveraging both general protein interaction data and antibody-specific structural information.

Our computational design process (see 3.11) for *de novo* CDR structures begins with an antigen-antibody complex structure, where the CDRs intended for design are masked out of the antibody input. Our enhanced model, $CoordVAE_{ab}$, then generates a diverse ensemble of CDR backbone structures conditioned on this masked antibody structure and the antigen. For selected backbones from this ensemble, we perform inverse folding to obtain a CDR sequence repertoire. These designed sequences undergo rigorous *in silico* folding evaluation to ensure structural integrity and stability. The most promising designs are evaluated for their potential binding interface with the antigen through computational complex structure prediction. This process is iterated, with top-scoring designs serving as seeds for subsequent rounds, allowing for progressive optimization of the CDR structures. This pipeline enables a comprehensive exploration of both CDR structure and sequence space, leveraging our $CoordVAE_{ab}$ model’s ability to generate diverse yet structurally compatible CDRs while maintaining potential antigen interactions.

In Fig3.11 we observed that the two stage fine tuning significantly improved our model’s ability to accurately in paint the CDRs of interest. Our final fine tuned model $CoordVAE_{ab}$, demonstrates significant improvements across all three CDR regions compared to other models. It consistently achieves the highest TM-scores and pLDDT scores, indicating better overall structural similarity and prediction confidence. Notably, it also maintains the lowest RMSD values across all CDR types, suggesting more accurate reconstruction of the original structures. The performance gap is most significant for CDR3, which is often the most consequential CDR and the hardest to design. These results indicate that our model is exceptionally effective at generating accurate and structurally similar CDR regions, which is

crucial for antibody design and engineering.

Model	Training Data	CDR-H1		CDR-H2		CDR-H3	
		lddt↑	RMSD↓	lddt↑	RMSD↓	lddt↑	RMSD↓
CoordVAE(10AA linear)	CATH4.2	0.587	2.847	0.588	2.864	0.498	4.427
CoordVAE(spatial)		0.515	3.944	0.514	3.914	0.487	4.350
CoordVAE(Random)		0.447	4.216	0.534	4.152	0.450	5.185
CoordVAE(mixed mask)		0.591	2.903	0.575	3.223	0.530	3.889
CoordVAE	SabDab	0.81	1.55	0.872	1.00	0.809	1.55
CoordVAE(Transfer)		0.8	0.81	0.90	0.85	0.823	1.35
AR-GNN		N/A	2.97	N/A	2.27	N/A	3.63
RefineGNN		N/A	1.18	N/A	0.87	N/A	2.50
CoordVAE-ab	SabDab + CATH4.2 + DIPS	0.95	0.83	0.97	0.68	0.96	0.74

Table 3.1: Structure inpainting performance on the test monoclonal antibody dataset in CDR-H1(left), CDR-H2(middle), CDR-H3(right) across different models.

To compare the new model with previous models, table3.1 demonstrates that CoordVAE-ab, trained on an extended dataset (SAbDab + CATH4.2 + DIPS) and using a two-stage fine-tuning approach, significantly outperformed all other models across all CDR regions. For CDR-H1, CoordVAE-ab achieved an LDDT score of 0.95 and an RMSD of 0.83, surpassing the next best model (CoordVAE). In CDR-H2, CoordVAE-ab’s performance (LDDT: 0.97, RMSD: 0.68) exceeded the previous best (RefineGNN, RMSD: 0.87). Most notably, for the challenging CDR-H3 region, CoordVAE-ab (LDDT: 0.96, RMSD: 0.74) outperformed the next best model (CoordVAE(Transfer), LDDT: 0.823, RMSD: 1.35). These substantial improvements across all metrics and CDR regions underscore the effectiveness of CoordVAE-ab’s extended dataset and refined training approach in antibody structure inpainting tasks and generate realistic and confident CDR backbone structures.

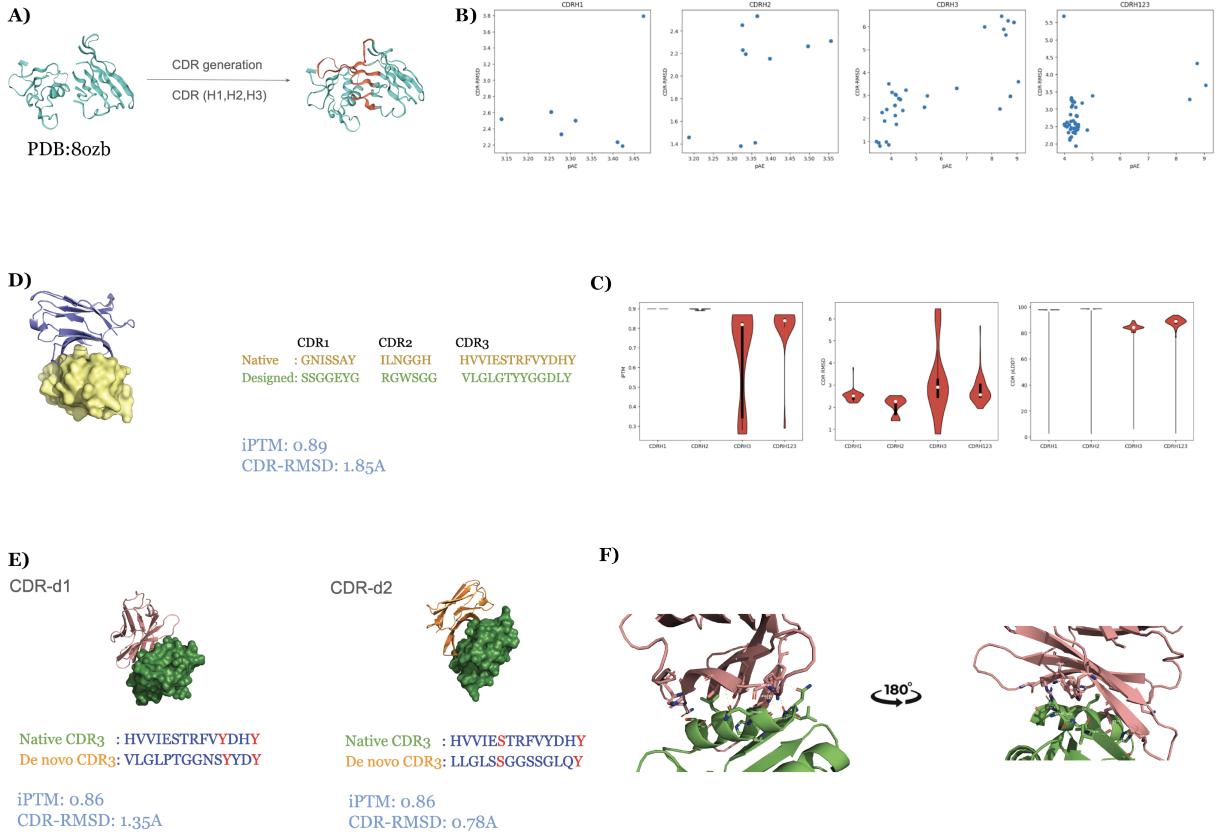


Figure 3.12: Structure based conditional CDR design A) A nanobody-antigen structure for design demonstration PDB:8ozb. B) CDR-RMSD vs. predicted pAE from computational structural models across different design modes. For CDR-H1 and CDR-H2 designs, most of the designed variants converged which is not surprising because the size of the design space is small. For CDR-H3 design, a fair share of designs pass the *in silico* filter (pAE<10 and RMSD<2.5). Surprisingly, we found in the case of full CDR design, the majority of the designs can pass the *in silico* filter. C) Distribution of CDR-RMSD across different design modes. The increased variant on CDR-H3 and full CDR design indicates that in this system the CDR-H3 is the most consequential CDR component. D) Example of full *de novo* CDR design with predicted complex model. E) Example of CDR3 design with computationally predicted complex model. F) Predicted epitope-paratope binding surface of *de novo* designed nanobody-antigen complex.

To evaluate the efficacy of our design framework, we applied it to a case study of nanobody-antigen complexes. We began by selecting a representative nanobody-antigen pair (Fig. 3.12.A) and employed our design pipeline to varying combinations of CDRs. Our assessment of computational structure predictors' reliability in predicting the designed

structures revealed promising results. For designs with predicted Aligned Error (pAE) < 10, the majority of *de novo* CDR-H2 and CDR-H3 designs exhibited < 2 \AA RMSD compared to native CDR structures. In cases of full CDR design, most predicted structures demonstrated RMSD < 2.5 \AA (Fig. 3.12.B).

Examination of the binding influence of different CDRs revealed that CDR-H3 has the most significant impact on predicted binding confidence, showing the highest variance in predicted RMSDs (Fig. 3.12.C). This underscores the critical role of CDR-H3 in antigen recognition and binding. Examples of predicted complex structures for both full CDR design and CDR-H3-only design are illustrated in Fig. 3.12.D and E, respectively. In both scenarios, the resulting *de novo* CDRs yielded high-confidence predicted complex models, demonstrating the robustness of our design approach.

Our design framework shows strong performance in generating *de novo* CDRs for nanobody-antigen complexes. The proposed designs demonstrate low predicted RMSD compared to native structures with highly diverse CDR sequences. Both full CDR and CDR-H3-only designs resulted in high-confidence predicted complexes model *in silico*, with promising results on further design potentials. Compared to previous structure based antibody design methods, our approach offers several advantages. We fully leverage DL-based model in all design stages with both higher efficiency and design capacity. The method also demonstrates the ability to generate diverse yet structurally compatible CDRs while maintaining potential antigen interactions. We also offer high flexibility in the CDR design space by allowing various combination of CDR selections, this approach yield valuable insights into the systems from our computational evaluation.

3.6 Discussion

This chapter presented an innovative iterative design framework for *de novo* protein engineering, demonstrating its versatility and effectiveness across four distinct applications:

unconditional structure generation, functional protein design, motif-grounded scaffolding, and structure-based antibody design. By integrating deep learning-based structure generation with *in silico* evaluation and refinement, our approach showcases the potential for computational methods to accelerate and enhance protein engineering efforts.

Our framework successfully generated novel protein folds with high *in silico* folding confidence across a range of protein sizes. Many of these designs exhibited significant structural novelty when compared to existing structures in the Protein Data Bank, highlighting the framework’s ability to explore beyond naturally occurring protein architectures. This capability opens new avenues for understanding protein structure-function relationships and potentially discovering proteins with novel functions.

For functional protein design, our iterative approach achieved remarkable success in optimizing DFHBI-activated β -barrel fluorescent proteins. The 100% success rate in experimental validation, with all 20 computationally designed candidates showing detectable DFHBI-activated fluorescence, underscores the robustness of our method. Moreover, the enhanced thermal stability and increased production yield compared to the original design reference demonstrate the practical benefits of our approach in creating improved functional proteins for biological applications.

The successful scaffolding of the key binding motif of PD1 onto a novel, potentially more stable protein structure further illustrates the framework’s versatility. Computational predictions suggest that the designed protein maintains binding specificity for PD-L1 while offering a smaller, potentially more stable context. This achievement demonstrates the framework’s ability to tackle complex protein engineering challenges with potential implications for cancer immunotherapy.

Our fourth application, structure-based antibody design via conditional CDR inpainting, represents a significant advancement in computational antibody engineering. The enhanced CoordVAE-ab model, trained on an extended dataset and using a two-stage fine-tuning

approach, demonstrated superior performance in CDR structure prediction across all three CDR regions. This improvement was particularly notable for the challenging CDR-H3 region, which is crucial for antigen recognition and binding. The framework’s ability to generate diverse, high-quality CDR designs while maintaining antigen binding specificity showcases its potential to accelerate and improve antibody design processes.

Despite these promising results, several limitations and areas for future work should be addressed. While we demonstrated experimental success with the fluorescent protein designs, the PD1 scaffolding and antibody design results are based primarily on computational predictions. Future work should include extensive experimental validation of these designs, including binding affinity measurements, structural characterization, and functional assays. The iterative nature of our framework, while powerful, can be computationally intensive. Future efforts should focus on optimizing the pipeline for increased efficiency and scalability, potentially leveraging distributed computing or more efficient sampling methods.

While our framework considers structural stability and function, future iterations could incorporate additional design objectives such as solubility, and immunogenicity. This would require the development and integration of reliable computational predictors for these properties. Our study covered a broad range of applications, but future work should explore the framework’s applicability to an even wider range of protein functions, including enzymatic activity, small molecule binding, and allosteric regulation.

Developing a systematic way to incorporate experimental data into the iterative design process could further improve the success rate and efficiency of the framework. This could involve machine learning models trained on experimental outcomes to guide future design iterations. While state-of-the-art structure prediction tools were used, there is still room for improvement, especially for novel folds and antibody structures. Developing more accurate structure prediction methods tailored for *de novo* designed proteins and antibodies could enhance the overall performance of the framework.

The current study focused primarily on single-domain proteins and antibody CDRs. Extending the framework to design multi-domain proteins, full antibody structures, or protein complexes could open up new possibilities for creating sophisticated molecular machines and therapeutic agents. While we demonstrated success in scaffolding existing functional motifs and designing CDRs, developing methods for *de novo* functional site design within our framework could greatly expand its capabilities. This could enable the creation of entirely new protein functions not found in nature and novel antibody paratopes with enhanced binding properties.

In conclusion, our iterative design framework represents a significant advance in computational protein engineering, demonstrating success across four diverse and challenging design tasks. By addressing the limitations outlined above and expanding its capabilities, this approach has the potential to accelerate the development of novel proteins and antibodies for a wide range of biotechnological and therapeutic applications. Future work should focus on refining the computational methods, expanding the range of designable functions, improving antibody design capabilities, and integrating more closely with experimental validation to realize the full potential of this powerful approach to protein engineering.

CHAPTER 4

CONCLUDING REMARKS

This study has introduced a novel structural generative model for proteins, termed CoordVAE, capable of directly modeling three-dimensional coordinates while addressing rotational and translational equivariance. The model demonstrates high-quality reconstruction of protein backbone structures across a wide range of sizes and folds, and generates diverse conformational ensembles. We also demonstrate both improved sequence diversity and design confidence at the same time using the generated structure ensemble compared to single backbone templates.

Building upon this generative model, we developed an iterative design framework that integrates structure generation, sequence design, and multi-faceted evaluation. This framework enables progressive optimization toward multiple design objectives, mimicking the process of directed evolution *in silico*. The approach iteratively refines both structure and sequence, using computational predictions of stability and function to guide each design cycle.

The versatility and effectiveness of this iterative framework were demonstrated across four challenging applications: *de novo* protein fold generation, functional protein optimization, motif-based scaffolding, and structure-guided antibody engineering. Notable successes include the generation of novel protein folds with high *in silico* folding confidence, the optimization of DFHBI-activated β -barrel fluorescent proteins with 100% experimental success rate and improved properties, the scaffolding of a PD1 binding motif onto a novel structure, and superior performance in antibody CDR structure prediction.

Looking ahead, several promising directions for future work emerge. Expanding experimental validation, particularly for the computationally designed binder scaffolds and antibodies, will be crucial. Incorporating additional design objectives such as solubility and immunogenicity could further improve the practical utility of designed proteins. Exploring applications to multi-domain proteins, protein-protein interfaces, and *de novo* functional site

design represent exciting frontiers for expanding the framework’s capabilities.

Developing systematic ways to incorporate experimental feedback into the iterative process could enhance design success rates and efficiency. Improving the scalability of the pipeline will be important for tackling larger design challenges. Additionally, integrating the framework with recent advances in protein language models and emerging high-throughput experimental techniques could open new avenues for protein engineering.

In conclusion, this work represents a significant advancement in computational protein design, bridging structure generation, sequence design, and multi-faceted *in silico* selection within a flexible iterative framework. By enabling the exploration of vast design spaces while maintaining a focus on both structural integrity and functional requirements, this approach provides a powerful new tool for expanding protein structure and function space. As the field of protein engineering continues to evolve, the integration and refinement of such computational methods promises to accelerate the development of novel proteins for a wide range of biotechnological and therapeutic applications.

REFERENCES

- [1] Brennan Abanades, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. Ablooper: fast accurate antibody cdr loop structure prediction with accuracy estimation. *Bioinformatics*, 38(7):1877–1880, 2022.
- [2] Jared Adolf-Bryfogle, Oleks Kalyuzhnii, Michael Kubitz, Brian D Weitzner, Xiaozhen Hu, Yumiko Adachi, William R Schief, and Roland L Dunbrack Jr. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS computational biology*, 14(4):e1006112, 2018.
- [3] Rahmad Akbar, Philippe A Robert, Cédric R Weber, Michael Widrich, Robert Frank, Milena Pavlović, Lonneke Scheffer, Maria Chernigovskaya, Igor Snapkov, Andrei Slabodkin, et al. In silico proof of principle of machine learning-based antibody design at unconstrained scale. In *MAbs*, volume 14, page 2031482. Taylor & Francis, 2022.
- [4] Mehmet Akdel, Douglas EV Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022.
- [5] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O’Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.
- [6] Namrata Anand and Possu Huang. Generative modeling for protein structures. *Advances in neural information processing systems*, 31, 2018.
- [7] Namrata Anand, Raphael Eguchi, and Po-Ssu Huang. Fully differentiable full-atom protein backbone generation. *bioRxiv*, 2019.
- [8] Namrata Anand, Raphael Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *Nature communications*, 13(1):746, 2022.
- [9] Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.
- [10] Frances H Arnold. Directed evolution: bringing new chemistry to life. *Angewandte Chemie (International Ed. in English)*, 57(16):4143, 2018.

- [11] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.
- [12] Benjamin Basanta, Matthew J Bick, Asim K Bera, Christoffer Norn, Cameron M Chow, Lauren P Carter, Inna Goreshnik, Frank Dimaio, and David Baker. An enumerative algorithm for de novo design of proteins with diverse pocket structures. *Proceedings of the National Academy of Sciences*, 117(36):22135–22145, 2020.
- [13] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vini- cius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- [14] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [15] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.
- [16] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [17] Scott E Boyken, Zibo Chen, Benjamin Groves, Robert A Langan, Gustav Oberdorfer, Alex Ford, Jason M Gilmore, Chunfu Xu, Frank DiMaio, Jose Henrique Pereira, et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science*, 352(6286):680–687, 2016.
- [18] Andrew RM Bradbury, Sachdev Sidhu, Stefan Dübel, and John McCafferty. Beyond natural antibodies: the power of in vitro display technologies. *Nature biotechnology*, 29(3):245–254, 2011.
- [19] Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variants with a deep protein language model. *bioRxiv*, pages 2022–08, 2022.
- [20] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [21] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai- based docking methods fail to generate physically valid poses or generalise to novel sequences. *arXiv preprint arXiv:2308.05777*, 2023.

- [22] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.
- [23] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.
- [24] Sheng Chen, Zhe Sun, Lihua Lin, Zifeng Liu, Xun Liu, Yutian Chong, Yutong Lu, Huiying Zhao, and Yuedong Yang. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. *Journal of chemical information and modeling*, 60(1):391–399, 2019.
- [25] E ChuAlexander, R EguchiRaphael, et al. De novo design of a highly stable ovoid tim barrel: Unlocking pocket shape towards functional design. *BioDesign Research*, 2022.
- [26] Michael F Chungyoun and Jeffrey J Gray. Ai models for protein design are driving antibody engineering. *Current Opinion in Biomedical Engineering*, 28:100473, 2023.
- [27] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.
- [28] A Courbet, J Hansen, Y Hsia, N Bethel, Y-J Park, C Xu, A Moyer, SE Boyken, G Ueda, U Nattermann, et al. Computational design of mechanically coupled axle-rotor protein assemblies. *Science*, 376(6591):383–390, 2022.
- [29] Rebecca Crawshaw, Amy E Crossley, Linus Johannissen, Ashleigh J Burke, Sam Hay, Colin Levy, David Baker, Sarah L Lovelock, and Anthony P Green. Engineering an efficient and enantioselective enzyme for the morita–baylis–hillman reaction. *Nature chemistry*, 14(3):313–320, 2022.
- [30] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.
- [31] Bassil I Dahiyat and Stephen L Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.
- [32] Bowen Dai and Chris Bailey-Kellogg. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, 37(17):2580–2588, 2021.
- [33] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.
- [34] Alice Del Vecchio, Andreea Deac, Pietro Liò, and Petar Veličković. Neural message passing for joint paratope-epitope prediction. *arXiv preprint arXiv:2106.00757*, 2021.

- [35] John R Desjarlais and Tracy M Handel. De novo design of the hydrophobic cores of proteins. *Protein Science*, 4(10):2006–2018, 1995.
- [36] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- [37] Jiayi Dou, Anastassia A Vorobieva, William Sheffler, Lindsey A Doyle, Hahnbeom Park, Matthew J Bick, Binchen Mao, Glenna W Foight, Min Yen Lee, Lauren A Gagnon, et al. De novo design of a fluorescence-activating β -barrel. *Nature*, 561(7724):485–491, 2018.
- [38] Frédéric A Dreyer, Daniel Cutting, Constantin Schneider, Henry Kenlay, and Charlotte M Deane. Inverse folding for antibody sequence design using deep learning. *arXiv preprint arXiv:2310.19513*, 2023.
- [39] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.
- [40] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.
- [41] Raphael R Eguchi, Christian A Choe, and Po-Ssu Huang. Ig-vae: Generative modeling of protein structure by direct 3d coordinate generation. *PLoS computational biology*, 18(6):e1010271, 2022.
- [42] Richard Evans, Michael O’Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.
- [43] Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Gonçarencio, Luca Naef, and Christian Dallago. From sequence to function through structure: deep learning for protein design. *Computational and Structural Biotechnology*, 2022. doi:10.1101/2022.08.31.505981.
- [44] Sarel J Fleishman, Timothy A Whitehead, Damian C Ekiert, Cyrille Dreyfus, Jacob E Corn, Eva-Maria Strauch, Ian A Wilson, and David Baker. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821, 2011.
- [45] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.
- [46] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

- [47] Pablo Gainza, Hunter M Nisonoff, and Bruce R Donald. Algorithms for protein design. *Current opinion in structural biology*, 39:16–26, 2016.
- [48] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.
- [49] Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, 2023.
- [50] Pablo Gainza-Cirauqui and Bruno Emanuel Correia. Computational protein design—the next generation tool to expand synthetic biology applications. *Current opinion in biotechnology*, 52:145–152, 2018.
- [51] Aravindhan Ganesan, Marawan Ahmed, Isobel Okoye, Elena Arutyunova, Dinesh Babu, William L Turnbull, Joydeb Kumar Kundu, Justin Shields, Katharine Cheryl Agopsowicz, Lai Xu, et al. Comprehensive in vitro characterization of pd-l1 small molecule inhibitors. *Scientific reports*, 9(1):12392, 2019.
- [52] Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.
- [53] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [54] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.
- [55] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.
- [56] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [57] Christoph Gorgulla, Andras Boeszoeremenyi, Zi-Fu Wang, Patrick D Fischer, Paul W Coote, Krishna M Padmanabha Das, Yehor S Malets, Dmytro S Radchenko, Yurii S Moroz, David A Scott, et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668, 2020.

- [58] Jordan Graves, Jacob Byerly, Eduardo Priego, Naren Makkapati, S Vince Parish, Brenda Medellin, and Monica Berrondo. A review of deep learning methods for antibodies. *Antibodies*, 9(2):12, 2020.
- [59] Rudolf Griss, Alberto Schena, Luc Reymond, Luc Patiny, Dominique Werner, Christine E Tinberg, David Baker, and Kai Johnsson. Bioluminescent sensor proteins for point-of-care therapeutic drug monitoring. *Nature chemical biology*, 10(7):598–603, 2014.
- [60] Somaye Hashemifar, Behnam Neyshabur, Aly A Khan, and Jinbo Xu. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810, 2018.
- [61] HW Hellinga and FM Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proceedings of the National Academy of Sciences*, 91(13):5803–5807, 1994.
- [62] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.
- [63] R Blake Hill, Daniel P Raleigh, Angela Lombardi, and William F DeGrado. De novo design of helical bundles as models for understanding protein folding and function. *Accounts of chemical research*, 33(11):745–754, 2000.
- [64] Donald Hilvert. Design of protein catalysts. *Annual Review of Biochemistry*, 82(1):447–470, 2013.
- [65] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [66] Magnus Haraldson Høie, Alissa Hummer, Tobias H Olsen, Broncio Aguilar-Sanjuan, Morten Nielsen, and Charlotte M Deane. Antifold: Improved antibody structure-based design using inverse folding. *arXiv preprint arXiv:2405.03370*, 2024.
- [67] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.
- [68] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pages 8946–8970. PMLR, 2022.
- [69] Jing Huang and Alexander D MacKerell Jr. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of computational chemistry*, 34(25):2135–2145, 2013.
- [70] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

- [71] Po-Ssu Huang, Kaspar Feldmeier, Fabio Parmeggiani, D Alejandro Fernandez Velasco, Birte Höcker, and David Baker. De novo design of a four-fold symmetric tim-barrel protein with atomic-level accuracy. *Nature chemical biology*, 12(1):29–34, 2016.
- [72] Alissa M Hummer, Brennan Abanades, and Charlotte M Deane. Advances in computational structure-based antibody design. *Current opinion in structural biology*, 74: 102379, 2022.
- [73] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.
- [74] John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623 (7989):1070–1078, 2023.
- [75] Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023.
- [76] TM Jacobs, B Williams, T Williams, X Xu, A Eletsky, JF Federizon, T Szyperski, and B Kuhlman. Design of structurally distinct proteins using strategies inspired by evolution. *Science*, 352(6286):687–690, 2016.
- [77] Giacomo Janson, Gilberto Valdes-Garcia, Lim Heo, and Michael Feig. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14(1):774, 2023.
- [78] Lin Jiang, Eric A Althoff, Fernando R Clemente, Lindsey Doyle, Daniela Rothlisberger, Alexandre Zanghellini, Jasmine L Gallaher, Jamie L Betker, Fujie Tanaka, Carlos F Barbas III, et al. De novo computational design of retro-aldol enzymes. *science*, 319 (5868):1387–1391, 2008.
- [79] Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.
- [80] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.
- [81] Xiaoyang Jing, Fandi Wu, Xiao Luo, and Jinbo Xu. Raptorgx-single: single-sequence protein structure prediction by integrating protein language models. *bioRxiv*, pages 2023–04, 2023.
- [82] Nathan H Joh, Tuo Wang, Manasi P Bhate, Rudresh Acharya, Yibing Wu, Michael Grabe, Mei Hong, Gevorg Grigoryan, and William F DeGrado. De novo design of a

transmembrane Zn²⁺-transporting four-helix bundle. *Science*, 346(6216):1520–1524, 2014.

- [83] David T Jones. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Science*, 3(4):567–574, 1994.
- [84] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
- [85] J Kaplan and WF DeGrado. De novo design of catalytic proteins. *Proceedings of the National Academy of Sciences*, 101(32):11566–11570, 2004.
- [86] Mostafa Karimi, Shaowen Zhu, Yue Cao, and Yang Shen. De novo protein design for novel folds using guided conditional wasserstein generative adversarial networks. *Journal of chemical information and modeling*, 60(12):5667–5681, 2020.
- [87] Firas Khatib, Seth Cooper, Michael D Tyka, Kefan Xu, Ilya Makedon, Zoran Popović, and David Baker. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953, 2011.
- [88] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [89] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [90] Gert Kiss, Nihan Çelebi-Ölçüm, Rocco Moretti, David Baker, and KN Houk. Computational enzyme design. *Angewandte Chemie International Edition*, 52(22):5700–5725, 2013.
- [91] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B Acton, Gae-tano T Montelione, and David Baker. Principles for designing ideal protein structures. *Nature*, 491(7423):222–227, 2012.
- [92] Tanja Kortemme. De novo protein design—from new structures to programmable functions. *Cell*, 187(3):526–544, 2024.
- [93] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [94] Brian Kuhlman, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *science*, 302(5649):1364–1368, 2003.

- [95] Boqiao Lai and Jinbo Xu. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings in Bioinformatics*, 23(1):bbab502, 2022.
- [96] Boqiao Lai, Sheng Qian, Hanwei Zhang, Siwei Zhang, Alena Kozlova, Jubao Duan, Jinbo Xu, and Xin He. Annotating functional effects of non-coding variants in neuropsychiatric cell types by deep transfer learning. *PLoS Computational Biology*, 18(5):e1010011, 2022.
- [97] Robert A Langan, Scott E Boyken, Andrew H Ng, Jennifer A Samson, Galen Dods, Alexandra M Westbrook, Taylor H Nguyen, Marc J Lajoie, Zibo Chen, Stephanie Berger, et al. De novo design of bioactive protein switches. *Nature*, 572(7768):205–210, 2019.
- [98] A Leaver-Fay, M Tyka, SM Lewis, OF Lange, J Thompson, R Jacak, KW Kaufman, PD Renfrew, CA Smith, W Sheffler, et al. Chapter nineteen-rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Computer Methods*, (Part C):545–574, 2011.
- [99] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [100] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [101] Jin Sub Lee, Jisun Kim, and Philip M Kim. Score-based generative modeling for de novo protein design. *Nature Computational Science*, 3(5):382–392, 2023.
- [102] Zhixiu Li, Yuedong Yang, Eshel Faraggi, Jian Zhan, and Yaoqi Zhou. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2565–2573, 2014.
- [103] Zhirui Liao, Ronghui You, Xiaodi Huang, Xiaojun Yao, Tao Huang, and Shanfeng Zhu. Deepdock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 311–317. IEEE, 2019.
- [104] Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485*, 2023.
- [105] Yu-Ru Lin, Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Amanda F Clouser, Gaetano T Montelione, and David Baker. Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences*, 112(40):E5478–E5485, 2015.

- [106] Zeming Lin, Tom Sercu, Yann LeCun, and Alexander Rives. Deep generative models create new and diverse protein structures. In *Machine Learning for Structural Biology Workshop, NeurIPS*, 2021.
- [107] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- [108] Sarah L Lovelock, Rebecca Crawshaw, Sophie Basler, Colin Levy, David Baker, Donald Hilvert, and Anthony P Green. The road to fully programmable protein catalysis. *Nature*, 606(7912):49–58, 2022.
- [109] Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022.
- [110] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, 27:1–30, 2020.
- [111] Craig O Mackenzie, Jianfu Zhou, and Gevorg Grigoryan. Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences*, 113(47):E7438–E7447, 2016.
- [112] Jack B Maguire, Scott E Boyken, David Baker, and Brian Kuhlman. Rapid sampling of hydrogen bond networks for computational protein design. *Journal of chemical theory and computation*, 14(5):2751–2760, 2018.
- [113] Sanaa Mansoor, Minkyung Baek, Hahnbeom Park, Gyu Rie Lee, and David Baker. Protein ensemble generation through variational autoencoder latent space sampling. *Journal of Chemical Theory and Computation*, 20(7):2689–2695, 2024.
- [114] Enrique Marcos, Benjamin Basanta, Tamuka M Chidyausiku, Yuefeng Tang, Gustav Oberdorfer, Gaohua Liu, GVT Swapna, Rongjin Guan, Daniel-Adriano Silva, Jiayi Dou, et al. Principles for designing proteins with cavities formed by curved β sheets. *Science*, 355(6321):201–206, 2017.
- [115] Enrique Marcos, Tamuka M Chidyausiku, Andrew C McShan, Thomas Evangelidis, Sanrupti Nerli, Lauren Carter, Lucas G Nivón, Audrey Davis, Gustav Oberdorfer, Konstantinos Tripsianes, et al. De novo design of a non-local β -sheet protein with high stability and accuracy. *Nature structural & molecular biology*, 25(11):1028–1034, 2018.
- [116] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. Iddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

- [117] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.
- [118] Stanislav Mazurenko, Zbynek Prokop, and Jiri Damborsky. Machine learning in enzyme engineering. *ACS Catalysis*, 10(2):1210–1223, 2019.
- [119] JA McCammon. Protein dynamics. *Reports on Progress in Physics*, 47(1):1, 1984.
- [120] Matt McPartlon, Ben Lai, and Jinbo Xu. A deep se (3)-equivariant model for learning inverse protein folding. *bioRxiv*, pages 2022–04, 2022.
- [121] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.
- [122] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.
- [123] Christoffer Norn, Basile IM Wicky, David Juergens, Sirui Liu, David Kim, Doug Tischer, Brian Koepnick, Ivan Anishchenko, Foldit Players, David Baker, et al. Protein sequence design by conformational landscape optimization. *Proceedings of the National Academy of Sciences*, 118(11):e2017228118, 2021.
- [124] James O’Connell, Zhixiu Li, Jack Hanson, Rhys Heffernan, James Lyons, Kuldip Palival, Abdollah Dehzangi, Yuedong Yang, and Yaoqi Zhou. Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics*, 86(6):629–633, 2018.
- [125] Sergey Ovchinnikov and Po-Ssu Huang. Structure-based protein design with deep learning. *Current opinion in chemical biology*, 65:136–144, 2021.
- [126] Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7):379–394, 2015.
- [127] Xingjie Pan and Tanja Kortemme. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, 296, 2021.
- [128] Xingjie Pan, Michael C Thompson, Yang Zhang, Lin Liu, James S Fraser, Mark JS Kelly, and Tanja Kortemme. Expanding the space of protein geometries by computational design of de novo fold families. *Science*, 369(6507):1132–1136, 2020.
- [129] Keunwan Park, Betty W Shen, Fabio Parmeggiani, Po-Ssu Huang, Barry L Stoddard, and David Baker. Control of repeat-protein curvature by computational protein design. *Nature structural & molecular biology*, 22(2):167–174, 2015.

- [130] Roberta Pascolutti, Xianqiang Sun, Joseph Kao, Roy L Maute, Aaron M Ring, Gregory R Bowman, and Andrew C Kruse. Structure and dynamics of pd-l1 and an ultra-high-affinity pd-1 receptor mutant. *Structure*, 24(10):1719–1728, 2016.
- [131] Jean-Denis Pédelacq, Stéphanie Cabantous, Timothy Tran, Thomas C Terwilliger, and Geoffrey S Waldo. Engineering and characterization of a superfolder green fluorescent protein. *Nature biotechnology*, 24(1):79–88, 2006.
- [132] Gregory A Petsko and Dagmar Ringe. *Protein structure and function*. New Science Press, 2004.
- [133] Dora Pinto, Young-Jun Park, Martina Beltramello, Alexandra C Walls, M Alejandra Tortorici, Siro Bianchi, Stefano Jaconi, Katja Culap, Fabrizia Zatta, Anna De Marco, et al. Cross-neutralization of sars-cov-2 by a human monoclonal sars-cov antibody. *Nature*, 583(7815):290–295, 2020.
- [134] Srivamshi Pittala and Chris Bailey-Kellogg. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*, 36(13):3996–4003, 2020.
- [135] Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, 2007.
- [136] Nicholas F Polizzi and William F DeGrado. A defined structural unit enables de novo design of small-molecule–binding proteins. *Science*, 369(6508):1227–1233, 2020.
- [137] Yifei Qi and John ZH Zhang. Densecpd: improving the accuracy of neural-network-based computational protein sequence design with densenet. *Journal of chemical information and modeling*, 60(3):1245–1252, 2020.
- [138] Alfredo Quijano-Rubio, Hsien-Wei Yeh, Jooyoung Park, Hansol Lee, Robert A Langan, Scott E Boyken, Marc J Lajoie, Longxing Cao, Cameron M Chow, Marcos C Miranda, et al. De novo design of modular and tunable protein biosensors. *Nature*, 591(7850):482–487, 2021.
- [139] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.
- [140] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.
- [141] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

- [142] Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houlisston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.
- [143] Erik A Rodriguez, Robert E Campbell, John Y Lin, Michael Z Lin, Atsushi Miyawaki, Amy E Palmer, Xiaokun Shu, Jin Zhang, and Roger Y Tsien. The growing and glowing toolbox of fluorescent and photoactive proteins. *Trends in biochemical sciences*, 42(2):111–129, 2017.
- [144] Daniela Röhlisberger, Olga Khersonsky, Andrew M Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L Gallaher, Eric A Althoff, Alexandre Zanghellini, Orly Dym, et al. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, 2008.
- [145] Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature communications*, 14(1):2389, 2023.
- [146] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.
- [147] A Schmid, JS Dordick, B Hauer, Al Kiener, M Wubbolts, and B Witholt. Industrial biocatalysis today and tomorrow. *nature*, 409(6817):258–268, 2001.
- [148] Bettina Schreier, Christian Stumpp, Silke Wiesner, and Birte Höcker. Computational design of ligand binding is not a solved problem. *Proceedings of the National Academy of Sciences*, 106(44):18491–18496, 2009.
- [149] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.
- [150] Fabian Sesterhenn, Che Yang, Jaume Bonet, Johannes T Cramer, Xiaolin Wen, Yimeng Wang, Chi-I Chiang, Luciano A Abriata, Iga Kucharska, Giacomo Castoro, et al. De novo protein design enables the precise induction of rsv-neutralizing antibodies. *Science*, 368(6492):eaay5051, 2020.
- [151] Thomas Shafee. *Evolvability of a viral protease: experimental evolution of catalysis, robustness and specificity*. PhD thesis, University of Cambridge, 2014.
- [152] Maxim V Shapovalov and Roland L Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

- [153] Arlene H Sharpe and Kristen E Pauken. The diverse functions of the pd1 inhibitory pathway. *Nature Reviews Immunology*, 18(3):153–167, 2018.
- [154] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.
- [155] Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):2403, 2021.
- [156] Justin B Siegel, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L St. Clair, Jasmine L Gallaher, Donald Hilvert, Michael H Gelb, Barry L Stoddard, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science*, 329(5989):309–313, 2010.
- [157] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.
- [158] Thomas Simonson, Thomas Gaillard, David Mignon, Marcel Schmidt am Busch, Anne Lopes, Najette Amara, Savvas Polydorides, Audrey Sedano, Karen Druart, and Georgios Archontis. Computational protein design: the proteus software and selected applications. *Journal of computational chemistry*, 34(28):2472–2484, 2013.
- [159] SN Sivanandam, SN Deepa, SN Sivanandam, and SN Deepa. *Genetic algorithms*. Springer, 2008.
- [160] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pages 20503–20521. PMLR, 2022.
- [161] Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M Kim. Fast and flexible protein design using deep graph neural networks. *Cell systems*, 11(4):402–411, 2020.
- [162] Chong Sun, Riccardo Mezzadra, and Ton N Schumacher. Regulation and function of the pd-l1 checkpoint. *Immunity*, 48(3):434–452, 2018.
- [163] Christine E Tinberg, Sagar D Khare, Jiayi Dou, Lindsey Doyle, Jorgen W Nelson, Alberto Schena, Wojciech Jankowski, Charalampos G Kalodimos, Kai Johnsson, Barry L Stoddard, et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, 501(7466):212–216, 2013.

- [164] Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019.
- [165] Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.
- [166] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [167] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.
- [168] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.
- [169] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.
- [170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [171] Susana Vázquez Torres, Philip JY Leung, Preetham Venkatesh, Isaac D Lutz, Fabian Hink, Huu-Hien Huynh, Jessica Becker, Andy Hsien-Wei Yeh, David Juergens, Nathaniel R Bennett, et al. De novo design of high-affinity binders of bioactive helical peptides. *Nature*, 626(7998):435–442, 2024.
- [172] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [173] Christopher A Voigt, Stephen L Mayo, Frances H Arnold, and Zhen-Gang Wang. Computational method to reduce the search space for directed protein evolution. *Proceedings of the National Academy of Sciences*, 98(7):3778–3783, 2001.
- [174] Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

- [175] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.
- [176] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1):e1005324, 2017.
- [177] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, pages 1–3, 2023.
- [178] Hein J Wijma, Robert J Floor, Peter A Jekel, David Baker, Siewert J Marrink, and Dick B Janssen. Computationally designed libraries for rapid enzyme stabilization. *Protein Engineering, Design & Selection*, 27(2):49–58, 2014.
- [179] Adam Winnifrith, Carlos Outeiral, and Brian L Hie. Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology*, 86:102794, 2024.
- [180] Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdar, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024.
- [181] Peng Xiong, Meng Wang, Xiaoqun Zhou, Tongchuan Zhang, Jiahai Zhang, Quan Chen, and Haiyan Liu. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nature communications*, 5(1):5330, 2014.
- [182] Aerin Yang, Kevin M Jude, Ben Lai, Mason Minot, Anna M Kocyla, Caleb R Glassman, Daisuke Nishimiya, Yoon Seok Kim, Sai T Reddy, Aly A Khan, et al. Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science*, 381(6656):eadh1720, 2023.
- [183] Che Yang, Fabian Sesterhenn, Jaume Bonet, Eva A van Aalen, Leo Scheller, Luciano A Abriata, Johannes T Cramer, Xiaolin Wen, Stéphane Rosset, Sandrine Georgeon, et al. Bottom-up de novo design of functional proteins with complex structural features. *Nature Chemical Biology*, 17(4):492–500, 2021.
- [184] Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020.
- [185] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

- [186] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z Zhang, Ivan Anishchenko, et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023.
- [187] Jason Yim, Hannes Stärk, Gabriele Corso, Bowen Jing, Regina Barzilay, and Tommi S Jaakkola. Diffusion models in protein structure and docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 14(2):e1711, 2024.
- [188] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.
- [189] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.
- [190] Jason Z Zhang, Hsien-Wei Yeh, Alexandra C Walls, Basile IM Wicky, Kaitlin R Sprouse, Laura A VanBlargan, Rebecca Treger, Alfredo Quijano-Rubio, Minh N Pham, John C Kraft, et al. Thermodynamically coupled biosensors for detecting neutralizing antibodies against sars-cov-2 variants. *Nature biotechnology*, 40(9):1336–1340, 2022.
- [191] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.
- [192] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.
- [193] Yuan Zhang, Yang Chen, Chenran Wang, Chun-Chao Lo, Xiuwen Liu, Wu Wei, and Jinfeng Zhang. Prodconn-protein design using a convolutional neural network. *Bioophysical Journal*, 118(3):43a–44a, 2020.
- [194] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.
- [195] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.
- [196] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.