Swansea University
Prifysgol Abertawe

Generative AI in Protein Design: *De novo* Protein Design and Multi-Motif Scaffolding

Zheng Sun

Swansea University

Submitted to Swansea University in fulfillment of the requirements for the Degree of *Masters*.

August 2024

Declarations

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed....[redacted]...

Date.......12/09/2024..

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed.....[redacted]..

Date.......12/09/2024..

I hereby give consent for my thesis, if accepted, to be available for electronic sharing

Signed.....[redacted]...

Date.......12/09/2024..

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed.....[redacted]..

Date.......12/09/2024..

Abstract

Proteins are fundamental molecules of life, playing critical roles in nearly all biological processes. The ability to design proteins with specific structures and functions holds immense potential for applications in drug discovery, enzyme engineering, and synthetic biology. However, traditional methods such as rational design and directed evolution face significant limitations, including reliance on existing templates, low computational efficiency, and difficulties in handling complex multi-motif scaffolding tasks. Recent advancements in generative artificial intelligence have opened new possibilities for protein design, yet challenges remain in geometric feature modeling and implicit motif positioning.

This study presents an integrated framework that combines the strengths of Geometric Vector Field Networks (VFN) and a protein diffusion model, MoDiff, to address these challenges. The VFN approach redefines frame modeling in *de novo* protein design by leveraging vector computations between coordinates of frame-anchored virtual atoms, enabling superior performance in designability and diversity compared to traditional methods. Specifically, VFN demonstrates significant improvements over Invariant Point Attention (IPA) in designability and diversity, and outperforms PiFold in sequence recovery rates during inverse folding. Meanwhile, MoDiff tackles the multi-motif scaffolding problem by facilitating the implicit positioning of motifs along the protein backbone, a capability absent in current scaffolding approaches. Our results show that MoDiff can generate diverse scaffolds and solve the multi-motif scaffolding problem even when motif positions are unknown, making it a versatile solution for complex protein design tasks.

The synergy between VFN and MoDiff not only highlights the individual strengths of each model but also demonstrates their combined potential to revolutionize protein engineering. This work opens new avenues for designing complex proteins and scaffolds, significantly advancing the field of synthetic biology. The findings have broad implications for pharmaceutical and biotechnological applications, paving the way for the development of novel therapeutic proteins, enzymes, and biomaterials.

Contents

Chapter 1

Introduction

1.1 Background

Proteins play an essential role in almost all biological processes. A sequence of amino acids is linked together, which then folds into a specific three-dimensional structure. This structure is crucial for the protein's function. Proteins are polymers of amino acids, each containing both an amino group ($-NH2$) and a carboxyl group ($-COOH$). There are 20 standard types of amino acids, each with specific chemical properties and behaviour (Schulz and Schirmer 2013).

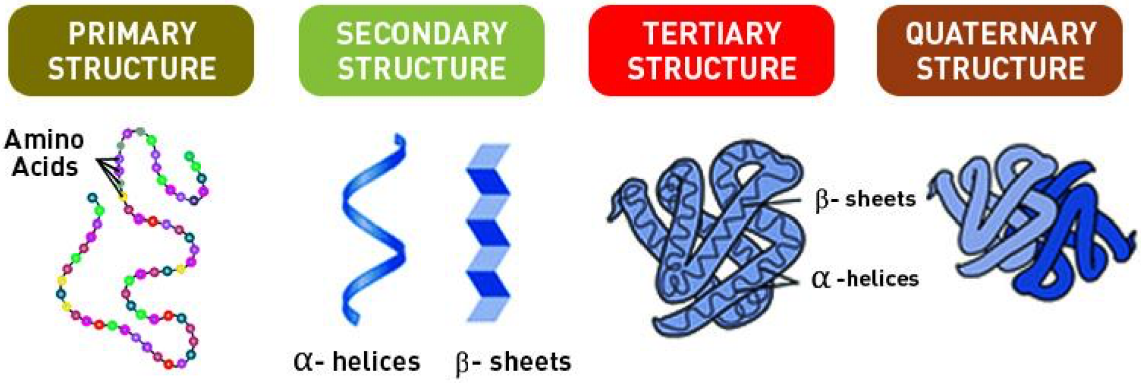

Figure 1.1: Four levels of protein structure

Levels of Protein Structure
Proteins have four levels of structural organisation:

- **Primary Structure**: The primary structure is the linear sequence of amino acids in the polypeptide chain. This sequence is determined by the genetic code and dictates the protein's final shape and function.
- **Secondary Structure**: The secondary structure refers to local folded structures that form within the polypeptide due to hydrogen bonding between backbone atoms. The most common secondary structures are alpha helices and beta sheets.

Alpha helices are right-handed coils, while beta sheets consist of beta strands connected laterally by at least two or three backbone hydrogen bonds.

- **Tertiary Structure**: The tertiary structure is the overall three-dimensional shape of a single polypeptide chain. This structure is stabilised by various interactions, including hydrogen bonds, ionic bonds, van der Waals forces, and hydrophobic interactions. Disulphide bonds between cysteine residues also play a crucial role in stabilising the tertiary structure.
- **Quaternary Structure**: The quaternary structure arises when two or more polypeptide chains (subunits) come together to form a functional protein complex. The arrangement and interaction of these subunits are essential for the protein's biological activity.

Protein Folding and Stability
Protein folding is the process by which a polypeptide chain attains its functional three-dimensional structure. This process is driven by the amino acid sequence and is influenced by the cellular environment. Proper folding is crucial for protein function, and misfolding can lead to diseases such as Alzheimer's and Parkinson's (Ponting and Russell 2002).

Protein stability is determined by the balance of forces that stabilise the folded state and those that favour the unfolded state. Key factors affecting protein stability include temperature, pH, and the presence of stabilising agents or denaturants.

Functional Domains and Motifs
Proteins often contain specific regions known as domains and motifs that are associated with particular functions. Domains are independent structural and functional units within a protein, often responsible for specific interactions or activities. Motifs are smaller structural elements, such as helix-turn-helix or zinc fingers, that are commonly found in different proteins and are associated with particular functions (Onuchic and Wolynes 2004; Bork and Koonin 1996).

1.1.1 Historical Overview of Protein Design

Protein design has been an integral part of scientific research aimed at understanding and engineering biological molecules for a variety of applications. The field originated from the biochemical manipulation of proteins for therapeutic and industrial uses. In the early days, strategies such as rational design and directed evolution were prominent, focusing on modifying existing protein structures to enhance or alter their functions. Rational design, in particular, required a precise understanding of the relationship between a protein's structure and its function, often limiting its application to well-characterised proteins (Whitford 2013). Directed evolution, as pioneered by researchers such as Frances Arnold, involved iterative mutation and selection processes to evolve proteins with desired traits, a method that proved to be revolutionary and earned widespread recognition (Arnold 1996,Sandhu 1992).

These traditional methods, while effective, faced limitations due to their reliance on existing protein frameworks and a detailed understanding of the mechanisms governing protein behaviour. The complexity of protein functions often restricted the scope of these techniques to proteins with well-documented characteristics.

- Traditional methods like template-based design and rational design often depend on the availability of well-characterised protein structures. This reliance constrains the scope of potential applications to proteins whose structures and functions are already well understood, limiting innovation and the exploration of novel protein functionalities (Hellinga 1997, Arnold 1998).
- Traditional protein design methods are typically iterative and time-consuming, involving extensive lab work and empirical testing. Directed evolution, for example, requires multiple rounds of mutation and selection to identify variants with desired characteristics, a process that is not only slow but also costly (Jaeger, Eggert, Eipper, and Reetz 2001,Tracewell and Arnold 2009).
- Traditional methods are typically less innovative in designing entirely new protein structures since they rely heavily on existing natural proteins as templates. This reliance inhibits the ability to explore unknown protein spaces and hinders innovation in terms of protein function and structural diversity (Dahiyat and Mayo 1997; Kuhlman, Dantas, Ireton, Varani, Barry L. Stoddard, and David Baker 2003).

1.1.2 Computational Advances and Predictive Modelling

As computational biology emerged, the focus shifted towards developing algorithms that could predict protein structures from amino acid sequences. The inception of the Critical Assessment of protein Structure Prediction (CASP) competitions in 1994 was a pivotal moment, challenging the scientific community to develop increasingly accurate methods for predicting protein function (Radivojac, Clark, Oron, Schnoes, Wittkop, Sokolov, Graim, Funk, Verspoor, Ben-Hur, et al. 2013). During this period, tools such as Rosetta, which combined *ab initio* modelling with knowledge-based potentials, emerged as critical resources for protein structure prediction (Bradley, Kira MS Misura, and David Baker 2005, Rohl, Strauss, Kira MS. Misura, and David Baker 2004).

The development of machine learning and deep learning techniques further revolutionised this field. These technologies enhanced the ability to predict protein structures and dynamics with unprecedented accuracy. AlphaFold, developed by DeepMind, showcased this potential by achieving near-experimental accuracy in predicting protein structures, a breakthrough that has had profound implications for biology and medicine (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, R. Bates, Žídek, and e. a. Potapenko A. 2021, Callaway 2020). The accuracy of AlphaFold has set a new benchmark, enabling the prediction of structures for proteins that were previously considered intractable.

1.1.3 The Role of Artificial Intelligence

The integration of AI into protein design has led to transformative advancements across multiple domains, including inverse folding, *de novo* protein design, antibody design, and small molecule drug discovery. AI's ability to handle vast datasets and identify intricate patterns has been pivotal in these developments.

- Inverse folding, the process of predicting the amino acid sequence of a protein given its three-dimensional structure, has been significantly advanced by AI. Traditional methods struggled with this task due to the vast sequence space and the complexity of protein folding pathways. AI models, particularly those leveraging deep learning, have revolutionised this area. DeepMind's AlphaFold, for instance, has not only excelled in forward folding but also in inverse folding tasks, providing insights into protein folding mechanisms and sequence design (Jumper et al., 2021; Evans et al., 2021).

- *De novo* protein design aims to create new proteins from scratch, without relying on natural templates. This area has seen remarkable progress with AI technologies. AlphaFold and related models, such as RoseTTAFold, have enabled the design of novel protein architectures with high accuracy (Baek et al., 2021). The development of the ESM (Evolutionary Scale Modelling) model by Facebook AI has further enhanced this capability, providing deep contextual understanding of protein sequences and their functional landscapes (Rives et al., 2021). These models have facilitated the design of proteins with novel folds and functions, opening up new avenues for biotechnological and therapeutic applications.

- The design of therapeutic antibodies has benefited immensely from AI advancements. Traditional methods of antibody design were time-consuming and limited in scope. AI models have streamlined this process by predicting antibody-antigen interactions with high precision. Tools like the AI-driven antibody design platform by DeepMind have enabled the rapid design of antibodies with desired properties, improving the efficacy and specificity of therapeutic antibodies (Jumper et al., 2021; AlQuraishi, 2019).

- AI's impact on small molecule drug design has been profound, enhancing the ability to predict drug-target interactions, optimise pharmacokinetics, and design novel drug candidates. Models such as DeepChem and others that utilise deep learning and reinforcement learning have accelerated the drug discovery process, reducing the time and cost associated with bringing new drugs to market (Walters et al., 2020; Olivecrona et al., 2020). The ability of AI to analyse large chemical libraries and predict molecular interactions has opened new horizons in drug design, making it possible to identify potential drug candidates with high affinity and specificity for their targets.

1.1.4 Applications and Models

- **AlphaFold, AlphaFold 2, and AlphaFold 3**

AlphaFold, introduced by DeepMind, revolutionised protein structure prediction by using deep learning to achieve unprecedented accuracy. AlphaFold 2 further refined this technology, achieving near-experimental accuracy on a wide range of proteins (Jumper et al., 2021). This model's success has been pivotal in understanding protein structures and has broad implications for biology, medicine, and biotechnology.

- **ESM (Evolutionary Scale Modelling)**

  The ESM model, developed by Facebook AI, leverages large-scale evolutionary data to predict protein sequences and their functions. This model has significantly advanced the understanding of protein evolution and function, enabling more accurate predictions of protein structures and interactions (Rives et al., 2021). ESM's approach to modelling protein sequences provides deep insights into protein evolution, facilitating the design of novel proteins with specific functionalities.

- **RoseTTAFold**

  RoseTTAFold, an AI model developed by the Ragon Institute, uses deep learning to predict protein structures with high accuracy. It builds on the advancements of AlphaFold, integrating features that enhance the model's ability to accurately predict complex protein structures and interactions (Baek et al., 2021). RoseTTAFold's success has demonstrated the potential of AI in solving long-standing challenges in protein structure prediction.

- **Innovative Applications in Antibody and Drug Design**

  AI has transformed antibody and drug design through models that predict antibody-antigen interactions and optimise drug candidates. DeepMind's work on antibody design, along with advancements in AI-driven drug discovery platforms, has streamlined the design of therapeutic agents, enhancing their efficacy and safety (Jumper et al., 2021; AlQuraishi, 2019). These advancements are crucial for developing new therapies for a variety of diseases, including cancer and autoimmune disorders.

1.2 *De Novo* Protein Design

The realm of *De novo* protein design represents a revolutionary frontier in biotechnology and pharmaceutical development, promising to shift paradigms through the introduction of innovative therapeutic agents. Recent advances in this domain have been driven by groundbreaking work in protein structure diffusion models (Watson, Juergens, Bennett, B. Trippe, Yim, Eisenach, Ahern, Borst, R. Ragotte, L. Milles, et al. 2023, Yim, Brian L Trippe, De Bortoli, Mathieu, Doucet, Regina Barzilay, and Tommi Jaakkola 2023) and innovative protein folding networks (Dauparas, Anishchenko, Bennett, H. Bai, R. J. Ragotte, L. F. Milles, Wicky, Courbet, Haas, and Bethel 2022). These advancements employ a novel approach by Utilising a protein diffusion model to stochastically generate the backbone structure of proteins, which is then enhanced by inverse folding networks to design precise protein sequences based on the derived structures.

In the protein design framework, the diffusion process begins with a simple or partially folded protein structure and gradually applies changes to this initial state based on a set of probabilistic rules derived from learned protein data. These rules are informed by a deep learning model that has been trained on vast datasets of known protein structures and their corresponding sequences. The model predicts how slight alterations in the sequence affect the overall structure and stability, guiding the protein towards a final state that matches the desired configuration.

One of the significant challenges in leveraging deep learning for protein structure prediction has been the high specialization required in traditional methods, which often limits the scope of application to proteins whose structures are already well-characterised. Most existing protein structure encoders, such as Inverse Protein Assembler (IPA), are designed for frame Modelling and face significant limitations due to their reliance on scalar features like distances or angles, which can lack expressive power and versatility (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, R. Bates, Žídek, and e. a. Potapenko A. 2021). This paradigm has struggled to adequately model the complex interplay of protein dynamics, particularly when dealing with proteins without detailed prior structural data.

To address these challenges, this thesis introduces the Vector Field Network (VFN), a novel structural encoder that leverages vector-specific linear layers to extract multiple geometric feature vectors between frame-anchored virtual atoms. This method operates by mapping the coordinates of virtual atoms in Euclidean space, allowing dynamic representations of protein frames to enhance Modelling accuracy and protein design flexibility. The core innovation of VFN lies in its ability to overcome the atom representation bottleneck that has constrained previous computational approaches by enabling the extraction of multiple feature vectors through a vector field operator, circumventing the traditional limitations faced by scalar feature-dependent methods.

VFN exhibits significant advantages over previous encoders in terms of its expressive power and versatility. It enhances the designability and diversity of proteins by offering an enriched Modelling framework that integrates both frame and atom information, thereby facilitating more detailed and accurate protein simulations.

1.3 Multi-Motif Scaffolding

Protein design plays a crucial role in advancing medical therapies and enzyme technologies, yet one of its most complex challenges is the effective scaffolding of multiple motifs within a single protein structure (Procko, Berguig, B. W. Shen, Yifan Song, Frayo, Convertine, Margineantu, Booth, Correia, Cheng, et al. 2014, Correia, J. T. Bates, Loomis, Baneyx, Carrico, Jardine, Rupert, Correnti, Kalyuzhniy, Vittal, et al. 2014,Jiang, Althoff, Clemente, Doyle, Rothlisberger, Zanghellini, Gallaher, Betker, Tanaka, Barbas III, et al. 2008,Siegel, Zanghellini, Lovick, Kiss, Lambert, St. Clair, Gallaher, Hilvert, Gelb, Barry L Stoddard, et al. 2010). Traditional methods, such as those based on template matching or rational design, often fall short when tasked with accurately positioning multiple, functionally diverse motifs within a protein's architecture. This limitation is primarily due to their reliance on existing structures as templates, which restricts flexibility and hinders innovation when new or complex motif combinations are needed.

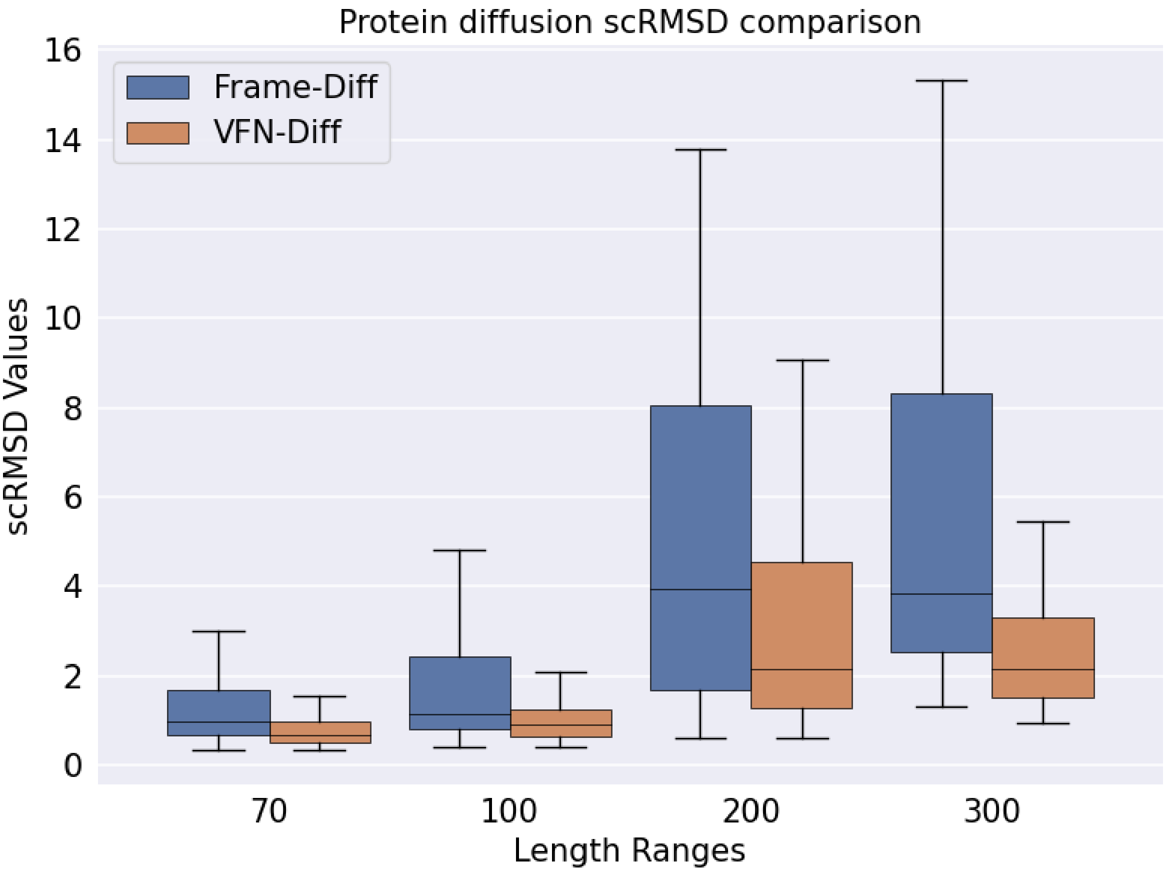

Figure 1.2: Designability of generated proteins with amino acid lengths ranging from 70 to 300: a comparison between VFN-Diff and FrameDiff.

Recent protein design approaches focused more on diffusion models (Yang Song, Durkan, Murray, and Ermon 2021, Brian L. Trippe, Yim, Tischer, Broderick, D. Baker, R. Barzilay, and T. Jaakkola 2022), and have made significant improvement in generating single motif. For example, RFdiffusion (Watson, Juergens, Bennett, B. Trippe, Yim, Eisenach, Ahern, Borst, R. Ragotte, L. Milles, et al. 2023), employing the inpainting paradigm, treats the given motif as the known part of the protein, with the remaining amino acids treated as inpainting content to generate the corresponding scaffold.

The motif scaffolding problem is particularly significant in the design of therapeutic antibodies and industrial enzymes, where the precise spatial arrangement of motifs can dictate the functionality and effectiveness of the final protein product. In many scenarios, especially in novel protein design, the exact positions of these motifs relative to one another are unknown, presenting a considerable challenge for existing scaffold design methodologies.

To address these challenges, based on the former work, we introduce an innovative computational model called MoDiff, which stands for Motif Diffusion. MoDiff represents a paradigm shift from traditional design methods by leveraging diffusion-based processes to guide the assembly and positioning of motifs within a protein scaffold. This model operates under the premise that effective motif placement can be achieved through a series of iterative adjustments guided by deep learning algorithms, which learn from vast datasets of protein structures and their functional outcomes.

During the diffusion process, MoDiff employs a novel approach that implicitly assigns motifs to appropriate positions on the protein backbone. This method not only allows for the design of proteins with multiple motifs but also enhances the diversity of possible scaffold configurations, thereby expanding the potential for novel protein functionalities. MoDiff's capability to automatically determine the relative positions of motifs without prior structural information marks a significant advancement in the field of protein engineering.

In silico experiments have shown that MoDiff, when confronted with a variety of motif configurations, exhibits robust performance in not only placing these motifs accurately but also in generating highly diverse and functional protein scaffolds. This model's success is attributed to its advanced algorithmic framework that adapts to the dynamic nature of protein folding and motif interaction, offering a flexible and efficient solution to the multi-motif scaffolding problem.

The contributions of this study can be Summarised as follows: We highlight the existing challenges in multi-motif scaffolding, introduce the MoDiff model as a potent solution capable of overcoming these challenges, and discuss the implications of this technology in Revolutionising the design of complex proteins. By integrating deep learning techniques with diffusion-based Modelling, MoDiff represents a promising advance in the field, suggesting that the future of protein design will increasingly rely on intelligent, data-driven approaches to tackle previously intractable problems in bioengineering and pharmaceutical development.

1.4 Research Objectives

This study aims to address key challenges in protein design through the integration of Geometric Vector Field Networks (VFN) and the MoDiff diffusion model. The specific objectives are as follows:

First, we aim to develop a novel geometric modeling framework (VFN) that enhances the designability and diversity of *de novo* protein structures by leveraging vector-based computations. This approach overcomes the limitations of traditional scalar feature-based methods, such as Invariant Point Attention (IPA), by enabling more expressive and flexible geometric representations.

Second, we introduce a diffusion-based model (MoDiff) capable of implicitly positioning multiple functional motifs along the protein backbone, even when their spatial relationships are unknown. This capability addresses a critical gap in current scaffolding approaches, which often rely on predefined templates or explicit motif positioning.

Third, we validate the combined VFN-MoDiff framework through extensive experiments, demonstrating its superiority over state-of-the-art methods in terms of designability, diversity, and motif integration accuracy. These experiments are designed to rigorously evaluate the performance of our models across a range of protein design tasks.

Finally, we explore the potential applications of the proposed framework in synthetic biology, drug discovery, and enzyme engineering. By enabling the design of novel proteins with tailored functionalities, our work has the potential to significantly advance these fields.

These objectives are addressed in detail in Chapter 3, which introduces the methodology, and Chapter 4, which presents the experimental results and analysis.

1.5 Impact and Recognition

This thesis makes the following key contributions to the field of protein design:

1. **Geometric Vector Field Networks (VFN)**: We introduce a novel geometric modeling framework that leverages vector-based computations to capture complex relationships between frame-anchored virtual atoms. Compared to traditional scalar feature-based methods (e.g., IPA), VFN demonstrates superior performance in designability (scTM$_{0.5}$ improvement of 6.54%) and diversity (54 novel proteins generated, compared to 37 by FrameDiff).
2. **MoDiff for Multi-Motif Scaffolding**: We propose a diffusion-based model capable of implicitly positioning multiple functional motifs along the protein backbone, even when their spatial relationships are unknown. MoDiff achieves a success rate of 19.21% on $\mathcal{M}_{\text{exist}}$ tasks and 16.35% on $\mathcal{M}_{\text{unknown}}$ tasks, surpassing traditional template-based methods.
3. **Integration of VFN and MoDiff**: By combining VFN's geometric modeling capabilities with MoDiff's motif alignment framework, we establish a unified pipeline for designing complex proteins with tailored functionalities. This synergy enables the generation of diverse and functional scaffolds, as demonstrated by the successful design of proteins with 3-4 motifs.

These contributions advance the state-of-the-art in protein design and have broad implications for synthetic biology, drug discovery, and enzyme engineering.

The research has been recognized for its impact in computational biology and machine learning:

- The work on the "Floating anchor diffusion model for multi-motif scaffolding" was presented at ICML 2024, where it received positive feedback for its innovative approach to protein design. The method's ability to handle unknown motif configurations was particularly highlighted.
- The paper on "*De novo* protein design using geometric vector field networks" was accepted at ICLR 2024, with reviewers praising its contributions to AI-driven protein design. The integration of vector-based computations with diffusion models was noted as a significant advancement.

Chapter 2

Literature Review

2.1 Overview of Protein Design and Motif Scaffolding

Protein design has been a pivotal aspect of bioengineering, influencing a wide range of applications from therapeutic development to industrial enzyme production. The ability to manipulate protein sequences to achieve specific structural and functional properties has been the cornerstone of advancements in biotechnology. This section provides an overview of the field, focusing on traditional methods of protein design and their inherent limitations.

2.1.1 Rational Design

Rational design involves making specific, targeted changes to protein sequences based on detailed knowledge of their three-dimensional structures and functional mechanisms. This approach is grounded in the principles of structural biology and biochemistry, where researchers use computational and experimental methods to predict how changes in amino acid sequences will affect the overall structure and function of the protein.

One of the earliest and most notable applications of rational design was the modification of enzymes to enhance their catalytic properties or alter their substrate specificities. For instance, the engineering of subtilisin, a protease, to improve its stability and activity under industrial conditions demonstrated the potential of rational design in practical applications (Ulmer, 1983). However, this method requires extensive prior knowledge of the protein's structure and the molecular basis of its function, which limits its applicability to well-characterised proteins (Hellinga 1997,Dahiyat and Mayo 1997).

2.1.2 Directed Evolution

Directed evolution, pioneered by researchers like Frances Arnold, represents a more empirical approach to protein design. This method mimics natural evolutionary processes by creating large libraries of protein variants through random mutagenesis and selecting those with desirable traits through high-throughput screening techniques. Directed evolution has been particularly successful in evolving enzymes with enhanced or novel functions that would be difficult to achieve through rational design alone (Arnold 1998, Zhao and Arnold 1999).

The directed evolution process typically involves three main steps: (1) generating genetic diversity, (2) screening or selecting for functional variants, and (3) iterating these steps to progressively improve the protein's properties. Despite its effectiveness, directed evolution is inherently time-consuming and Labour-intensive, as it relies on creating and testing thousands to millions of variants to identify those with optimal characteristics (Tracewell and Arnold 2009).

2.1.3 Limitations of Traditional Methods

While both rational design and directed evolution have significantly advanced the field of protein engineering, they come with several limitations. Rational design is constrained by the requirement for detailed structural knowledge, which is not always available, particularly for proteins with complex or poorly understood functions. Additionally, the accuracy of rational design predictions is limited by our current understanding of protein folding and dynamics (Hellinga 1997).

Directed evolution, on the other hand, although powerful, is an iterative and resource-intensive process. It requires substantial time and experimental effort to generate and screen large libraries of variants. Moreover, directed evolution can be biased towards exploring local fitness peaks, potentially missing out on more optimal solutions that lie outside the immediate mutational landscape (Arnold 1996,Tracewell and Arnold 2009).

Given these limitations, there has been a growing interest in developing computational methods that can predict and design protein structures more efficiently and accurately. The integration of artificial intelligence and machine learning into protein design represents a promising direction to overcome the challenges associated with traditional methods.

2.2 Advances in Computational Protein Design

2.2.1 Molecular Dynamics

Early computational techniques such as molecular dynamics (MD) simulations and homology Modelling have been instrumental in laying the groundwork for understanding protein folding, stability, and function. These methods have provided critical insights into the molecular Behaviour of proteins, although they come with specific limitations and challenges.

Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are a computational technique that models the physical movements of atoms and molecules over time, based on principles of classical mechanics. In the context of proteins, MD simulations allow researchers to observe the dynamic Behaviour of protein structures, providing detailed information on how proteins fold, fluctuate, and interact with other molecules.

Principles and Applications
MD simulations work by solving Newton's equations of motion for a system of interacting particles. Each atom in the protein is treated as a particle, and the interactions between these atoms are described by a potential energy function, often referred to as a force field. Popular force fields used in MD simulations include AMBER, CHARMM, and GROMOS (Brooks, Brooks III, Mackerell Jr, Nilsson, Petrella, Roux, Won, Archontis, Bartels, Boresch, et al. 2009,Cornell, Cieplak, Bayly, Gould, Merz, Ferguson, Spellmeyer, Fox, Caldwell, and Kollman 1995).

One of the key applications of MD simulations is in studying the folding pathways of proteins. By simulating the process of protein folding, researchers can gain insights into the intermediate states that a protein might adopt as it transitions from an unfolded to a folded state. This is particularly important for understanding diseases related to protein misfolding, such as Alzheimer's and Parkinson's diseases (Shirts and Pande 2000).

Challenges and Limitations
Despite their powerful insights, MD simulations are computationally intensive. Accurately simulating the folding process of even a small protein can require extensive computational resources and time, often spanning microseconds to milliseconds of real time. This computational demand limits the use of MD simulations to relatively small proteins and short timescales (Yong Duan and Kollman 1998).

To address these limitations, researchers have developed techniques such as enhanced sampling methods (e.g., metadynamics, umbrella sampling) and coarse-grained models that reduce the computational load by simplifying the representation of the system (Laio and Parrinello 2002, Voth 2008).

2.2.2 Homology Modelling

Homology Modelling, also known as comparative Modelling, is a method used to predict the three-dimensional structure of a protein based on its sequence similarity to proteins with known structures. This technique is grounded in the observation that proteins with similar sequences tend to adopt similar structures.

Principles and Applications

The process of homology Modelling typically involves several steps:
1. **Template Identification**: Identifying one or more known protein structures (templates) that share significant sequence similarity with the target protein.
2. **Sequence Alignment**: Alignment of the target sequence with the sequences of the template proteins to determine which residues are structurally conserved.

3. **Model Building**: Constructing a 3D model of the target protein by copying the backbone coordinates of the aligned regions from the template structures and Modelling the non-aligned regions (loops and side chains) using computational methods.
4. **Model Refinement and Validation**: Refining the initial model to resolve steric clashes and Optimise geometry, followed by validation against known structural criteria to ensure accuracy (Martí-Renom, Stuart, Fiser, Sánchez, Melo, and Šali 2000).

Homology Modelling is widely used in cases where the structure of a protein is unknown but sequences of homologous proteins with known structures are available. This method has been instrumental in drug discovery, allowing researchers to model the structures of target proteins and design molecules that can bind to these targets with high specificity (Šali and Blundell 1993).

Challenges and Limitations
The accuracy of homology Modelling depends heavily on the quality and similarity of the template structures. When close Homologues are available, the predictions are generally reliable. However, the accuracy diminishes when only distant Homologues can be used, as the structural differences between the target and template proteins increase (Söding 2005).

Additionally, homology Modelling struggles with accurately predicting the conformations of loops and side chains, particularly in regions of the protein that are highly flexible or functionally important. To mitigate these issues, advanced techniques such as loop Modelling and side-chain Optimisation algorithms are employed (Fiser and Šali 2003).

2.3 Role of Artificial Intelligence in Protein Design

The integration of artificial intelligence (AI) into protein design has revolutionized the field, enabling unprecedented advancements in predicting and designing protein structures and functions. By leveraging the power of machine learning (ML) and deep learning (DL), AI has significantly improved the accuracy, efficiency, and scope of protein design methodologies.

2.3.1 Deep Learning Models

Deep learning models have emerged as powerful tools for protein structure prediction and design. Among these, AlphaFold, RoseTTAFold, and Evolutionary Scale Modeling (ESM) represent significant breakthroughs, each contributing unique capabilities to the field.

AlphaFold

AlphaFold, developed by DeepMind, represents a groundbreaking advancement in protein structure prediction. This model achieves near-experimental accuracy in predicting the three-dimensional structures of proteins directly from their amino acid sequences. By leveraging multiple sequence alignments (MSAs) and structural templates, combined with sophisticated neural networks, AlphaFold predicts protein structures with unprecedented precision (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, R. Bates, Žídek, and e. a. Potapenko A. 2021).

The success of AlphaFold is largely attributed to its innovative use of attention mechanisms and transformer architectures. These elements enable the model to capture long-range dependencies and interactions within the protein sequence, which are crucial for accurate structure prediction. Traditional methods often struggled with these aspects, particularly for proteins with complex folding patterns and intricate interactions. AlphaFold addresses these challenges through several key steps.

First, the model collects multiple sequence alignments (MSAs) for the target protein sequence. These alignments provide evolutionary information that helps the model understand conserved regions and potential structural motifs. By analyzing sequences from related proteins across different species, the model gains insights into functionally important regions (Senior, Evans, Jumper, Kirkpatrick, Sifre, Green, Qin, Žídek, Nelson, Bridgland, et al. 2020).

Next, the amino acid sequences and MSAs are embedded into a high-dimensional space using initial processing layers. This step transforms the raw sequence data into a format that the deep learning model can effectively process. At the core of AlphaFold is a transformer network that uses attention mechanisms to model interactions between amino acids. The attention mechanism allows the model to weigh the importance of each amino acid's relationship to every other amino acid in the sequence, enabling the model to consider both local and global sequence contexts simultaneously (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017).

AlphaFold also predicts pairwise interactions between amino acids, which are crucial for understanding the protein's three-dimensional structure. This step involves predicting distances and orientations between pairs of amino acids, providing a detailed map of potential interactions within the protein. Using the predicted interactions, the model generates an initial protein structure, which is iteratively refined through additional rounds of prediction and optimization. AlphaFold uses gradient descent and other optimization techniques to minimize the difference between predicted and known structures (when available). Additionally, AlphaFold generates confidence metrics that estimate the reliability of the predicted structure, which are crucial for assessing the model's predictions, particularly when experimental validation is not immediately available.

The impact of AlphaFold extends beyond basic research into numerous practical applications. Its high accuracy in structure prediction has profound implications for understanding protein function, guiding experimental studies, and accelerating drug discovery. For example, AlphaFold's predictions have been used to model previously uncharacterized proteins, providing insights into their potential functions and interactions. This capability is particularly valuable in the context of emerging diseases, where rapid characterization of novel proteins can inform therapeutic development (Tunyasuvunakool, Adler, Z. Wu, Green, Zielinski, Žídek, Bridgland, Cowie, Meyer, Laydon, et al. 2021).

Moreover, AlphaFold's methodology has set a new standard in the field, inspiring further developments in AI-driven protein modeling. The principles behind AlphaFold are being adapted and extended to other areas of structural biology, including protein-protein interactions, enzyme engineering, and the design of novel biomolecules.

RoseTTAFold

RoseTTAFold, developed by the Baker lab, represents another significant advancement in the field of protein structure prediction. This deep learning model integrates three-track neural networks to simultaneously consider sequences, distances, and coordinates. Such an approach provides accurate predictions of protein structures and allows for efficient modeling of both individual proteins and protein-protein interactions, making RoseTTAFold a versatile and powerful tool in structural biology (Baek, DiMaio, Anishchenko, Dauparas, Ovchinnikov, G. R. Lee, J. Wang, Cong, Kinch, and Schaeffer 2021).

The architecture of RoseTTAFold is unique in its integration of three separate neural networks, each designed to process different types of input data. These networks work in concert to capture a comprehensive understanding of protein structures. The sequence network processes the amino acid sequences, using multiple sequence alignments (MSAs) to gather evolutionary information. This helps in identifying conserved regions that are likely critical for the protein's structure and function (AlQuraishi 2019).

The distance network predicts the pairwise distances between amino acids in the protein, which is crucial for understanding the spatial arrangement of the protein's residues. This prediction is informed by both sequence data and existing structural databases, allowing for a more nuanced and accurate modeling of distances (Baek, DiMaio, Anishchenko, Dauparas, Ovchinnikov, G. R. Lee, J. Wang, Cong, Kinch, and Schaeffer 2021).

The coordinate network directly predicts the three-dimensional coordinates of the protein's atoms. By integrating information from the sequence and distance networks, this network refines the spatial model of the protein to ensure that it adheres to known physical and chemical constraints (J. Yang, Anishchenko, Park, Peng, Ovchinnikov, and David Baker 2020).

These three networks are integrated into a single pipeline, enabling RoseTTAFold to produce high-accuracy models that consider multiple layers of information simultaneously. This comprehensive approach significantly enhances the model's ability to predict complex protein structures, including those with novel folds that have not been observed before (Senior, Evans, Jumper, Kirkpatrick, Sifre, Green, Qin, Žídek, Nelson, Bridgland, et al. 2020).

The impact of RoseTTAFold extends to various areas of biomedical research and biotechnology. For example, the model has demonstrated exceptional capability in predicting the structures of proteins with novel folds, expanding the range of proteins that can be accurately modeled. This is particularly important for the discovery and design of new proteins with specific functionalities that do not exist in nature (Baek, DiMaio, Anishchenko, Dauparas, Ovchinnikov, G. R. Lee, J. Wang, Cong, Kinch, and Schaeffer 2021).

Additionally, RoseTTAFold's ability to model protein-protein interactions accurately is a significant advancement. Understanding these interactions is crucial for elucidating cellular pathways and mechanisms, and for designing drugs that can modulate these interactions effectively (Ovchinnikov, Park, Varghese, Huang, Pavlopoulos, D. E. Kim, Kamisetty, Kyrpides, and David Baker 2017). By providing accurate models of target proteins and their interactions, RoseTTAFold aids in the rational design of small molecule drugs, accelerating the drug discovery process and enhancing the specificity and efficacy of therapeutic agents (Y. Zhang, Y. Chen, C. Wang, Lo, Xiuwen Liu, W. Wu, and J. Zhang 2020).

Evolutionary Scale Modeling (ESM)

The Evolutionary Scale Modeling (ESM) approach, developed by Facebook AI, represents a significant advancement in the field of protein design. By utilizing large-scale evolutionary data, ESM predicts protein sequences and their functions with high accuracy. This approach leverages the evolutionary relationships and sequence variations observed across diverse protein families to provide deep insights into protein evolution and functional landscapes (Rives, Meier, Sercu, Goyal, Lin, J. Liu, Guo, Ott, Zitnick, Ma, et al. 2021).

ESM models employ advanced deep learning techniques to analyze and extract meaningful patterns from massive datasets. The architecture of ESM is designed to handle the complexity and vastness of evolutionary data, enabling it to make accurate predictions about protein structures and interactions. The process begins with embedding protein sequences into high-dimensional spaces using learned representations. These embeddings capture the evolutionary context and biochemical properties of the sequences, which are crucial for understanding protein function (Rives, Meier, Sercu, Goyal, Lin, J. Liu, Guo, Ott, Zitnick, Ma, et al. 2021).

ESM also utilizes evolutionary couplings to infer the co-evolutionary relationships between amino acids in a protein sequence. By identifying pairs of residues that tend to mutate together, the model can predict spatial proximities and functional interactions within the protein structure (Marks, Colwell, Sheridan, Hopf, Pagnani, Zecchina, and Sander 2011).

At the core of ESM are deep learning networks that incorporate transformer architectures. These networks process the embedded sequences and evolutionary couplings to predict the three-dimensional structure and functional regions of the proteins. The transformer architecture, known for its ability to model long-range dependencies, is particularly effective in capturing the complex interactions within proteins (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017).

Additionally, ESM includes mechanisms for predicting functional annotations based on the identified patterns. By analyzing conserved regions and sequence motifs, the model can infer the likely functions of different protein regions, making it a powerful tool for annotating newly discovered or poorly understood proteins (Rives, Meier, Sercu, Goyal, Lin, J. Liu, Guo, Ott, Zitnick, Ma, et al. 2021).

The applications of ESM are wide-ranging and impactful across various domains of biological research and biotechnology. For example, ESM can identify functional regions within proteins, such as active sites, binding pockets, and interaction interfaces. This information is invaluable for understanding protein mechanisms and designing targeted interventions (Rao, Bhattacharya, Thomas, Yan Duan, P. Chen, Canny, Abbeel, and Yun Song 2019).

ESM's ability to predict the effects of mutations is particularly useful in both basic research and clinical settings. Understanding how specific mutations affect protein function can provide insights into disease mechanisms and guide the development of therapeutic strategies. This capability is also essential for protein engineering, where desired mutations can be introduced to enhance or alter protein function (Hopf, Ingraham, Poelwijk, Schärfe, Springer, Sander, and Marks 2017).

Furthermore, ESM excels in predicting protein-protein interactions by leveraging evolutionary couplings and sequence data. This is critical for mapping out cellular pathways and networks, understanding complex biological systems, and identifying potential drug targets. Accurate predictions of protein interactions can significantly accelerate drug discovery and development processes (Cong, Anishchenko, Ovchinnikov, and David Baker 2019).

The insights gained from ESM can also be applied to drug discovery and biotechnology. By providing detailed predictions of protein structures and interactions, ESM aids in the rational design of small molecules and biologics. This accelerates the drug discovery pipeline and enhances the development of new therapeutics with improved efficacy and safety profiles (Rives, Meier, Sercu, Goyal, Lin, J. Liu, Guo, Ott, Zitnick, Ma, et al. 2021).

2.4 Diffusion-Based Models in Protein Design

Diffusion-based models have shown significant promise in addressing complex problems in protein design, particularly in inverse folding, *de novo* protein design, and molecular docking. These models leverage probabilistic approaches to explore vast conformational spaces, enabling the generation of accurate and functional protein structures through iterative refinement. This section examines their applications across three key domains.

2.4.1 Application in Inverse Folding

Inverse folding, the task of predicting amino acid sequences that fold into a target three-dimensional structure, is a fundamental challenge in protein design. Diffusion-based models excel in this area by iteratively refining sequences based on probabilistic rules derived from large datasets. The process begins with an initial amino acid sequence, which may be randomly generated or derived from structural templates. Through multiple iterations, the model stochastically adjusts the sequence to maximize the likelihood of folding into the desired structure. Each adjustment is guided by energy functions or similarity metrics that evaluate structural compatibility. The process continues until the sequence converges to an optimal solution (K. Wu, K. Yang, Berg, Zou, Lu, and Amini 2022; Yi, Zhou, Y. Shen, Liò, and Y. Wang 2024).

A key innovation in this domain is the FoldingDiff architecture (Figure 2.1), which generates protein backbones by combining angular formulations of bond angles (θ) and dihedral angles (ϕ, ψ, ω). This angular representation captures the geometric constraints and flexibility of protein structures. The diffusion process involves iteratively adding noise to a data sample and learning to denoise it, starting from a noisy state (x_T) and gradually recovering the original structure (x_0). This approach ensures the generation of diverse and designable protein backbones while maintaining structural realism.

2.4.2 Applications in Molecular Docking

Molecular docking, the prediction of optimal ligand-protein binding configurations, is critical for drug discovery. Diffusion-based models, such as DiffDock (Figure 2.2), provide a robust framework for exploring ligand-protein interaction spaces. The process begins with generating an initial ligand conformation, either from known structures or *de novo*. The model then applies a diffusion process to iteratively adjust the ligand's translational, rotational, and torsional degrees of freedom relative to the protein target. This stochastic exploration allows the identification of favorable binding poses, which are subsequently ranked using scoring functions that assess binding affinity and stability (Corso, Stärk, Jing, Regina Barzilay, and Tommi Jaakkola 2022; Ketata, Laue, Mammadov, Stärk, M. Wu, Corso, Marquet, Regina Barzilay, and T. S. Jaakkola 2023).

DiffDock's success lies in its ability to decouple the diffusion process across different degrees of freedom, enabling efficient sampling of the conformational space. The final predictions are refined through a confidence model that ranks poses based on their likelihood of forming stable interactions, significantly accelerating the identification of high-affinity binders.

2.4.3 Application in *De Novo* Protein Design

De novo protein design aims to create novel protein sequences that fold into predetermined functional structures. Diffusion models address this challenge by combining structure generation with sequence optimization. The process typically starts with an initial scaffold or motif, which may be derived from existing structures or generated *ab initio*. The model then applies a sequence diffusion process, iteratively adjusting amino acid sequences to ensure compatibility with the target structure. Probabilistic rules derived from known protein data guide these adjustments, balancing structural stability and functional constraints. Each iteration involves validating the proposed sequence using energy functions and structural similarity metrics, ensuring the final design is both stable and functional (Yim, Brian L Trippe, De Bortoli, Mathieu, Doucet, Regina Barzilay, and Tommi Jaakkola 2023).

Recent advancements highlight the versatility of diffusion models in this domain. For instance, they have been used to design enzymes with enhanced catalytic activity by iteratively refining scaffold structures (Q. Wang, Xiaonan Liu, H. Zhang, Chu, Shi, L. Zhang, J. Bai, P. Liu, Li, Zhu, et al. 2024). In therapeutic applications, diffusion models have generated proteins that bind disease-specific targets with high specificity, offering potential treatments for cancer and autoimmune disorders (Anishchenko, Pellock, Chidyausiku, Ramelot, Ovchinnikov, Hao, Bafna, Norn, Kang, and Bera 2021). Additionally, these models enable the construction of structural scaffolds that support multiple functional domains, facilitating the development of synthetic biological systems for biosensing and biocatalysis (Brian L. Trippe, Yim, Tischer, Broderick, D. Baker, R. Barzilay, and T. Jaakkola 2022).

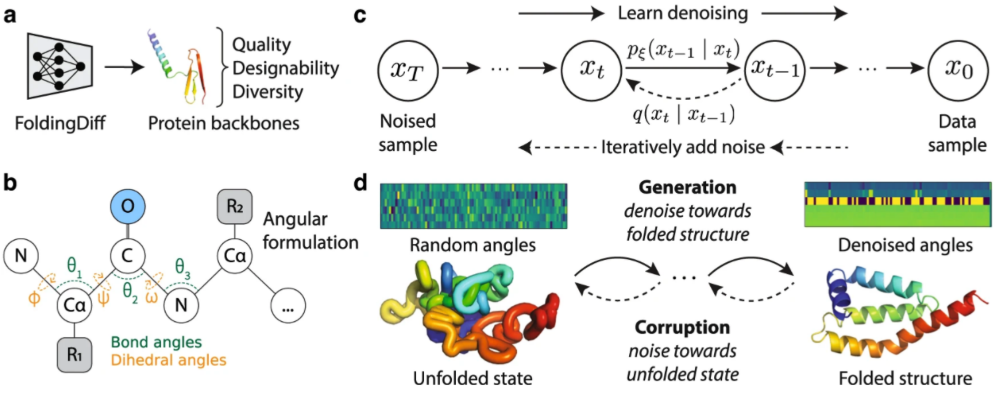

Figure 2.1: **FoldingDiff Architecture and Process.** (a) The model generates protein backbones using angular formulations of bond angles (θ) and dihedral angles (ϕ, ψ, ω). (b) The diffusion process involves iterative noising ($x_0 \to x_T$) and denoising ($x_T \to x_0$) to learn realistic protein conformations. (c) Generation and corruption cycles enable training on both folded and unfolded states.

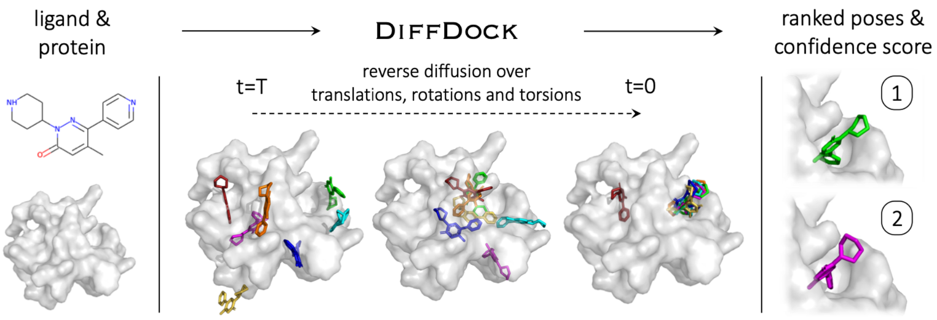
ligand & protein
DiffDock
ranked poses & confidence score
t=T
reverse diffusion over translations, rotations and torsions
t=0
NH
N
O
1
2

Figure 2.2: **DiffDock Workflow.** (Left) Input ligand and protein structures. (Center) Reverse diffusion adjusts translational, rotational, and torsional degrees of freedom. (Right) Confidence-based ranking of predicted binding poses.

2.5 Current Challenges and Future Directions

Research Gaps Despite significant progress, current methods face several limitations. Traditional approaches rely heavily on scalar features, such as distances and angles, which lack the expressive power to model complex geometric relationships. Additionally, existing scaffolding methods struggle to handle multi-motif designs, particularly when the spatial relationships between motifs are unknown. These gaps motivate the development of the VFN and MoDiff models, which aim to address these challenges through vector-based computations and diffusion-based motif positioning.

2.5.1 Current Challenges

While diffusion-based models have demonstrated remarkable success in protein design, several challenges remain that must be addressed to fully harness their potential. This section outlines the key challenges and explores future directions for advancing the field.

Computational Complexity and Resources Diffusion-based models are computationally intensive, requiring significant processing power and memory. The iterative nature of these models, coupled with the need to explore vast conformational spaces, can lead to high computational costs. This is particularly challenging for larger and more complex proteins, which necessitate longer simulation times and more resources (Hoogeboom, Vıctor Garcia Satorras, Vignac, and Welling 2022).

Data Availability and Quality The performance of diffusion-based models heavily relies on the availability and quality of training data. Large and diverse datasets are required to accurately capture the complexities of protein structures and their interactions. However, obtaining high-quality experimental data can be challenging, and many available datasets may contain biases or inaccuracies that can impact model performance (Yim, Campbell, Foong, Gastegger, Jiménez-Luna, Lewis, Victor Garcia Satorras, Veeling, Regina Barzilay, Tommi Jaakkola, and Noé 2023).

Model Interpretability While deep learning models, including diffusion-based approaches, have achieved high accuracy in protein design, they often lack interpretability. Understanding the decision-making process of these models is crucial for gaining insights into protein folding mechanisms and improving model reliability. Developing methods to interpret and explain the predictions of diffusion-based models remains an ongoing challenge (Harren, Matter, Hessler, Rarey, and Grebner 2022).

Integration with Experimental Validation Validating the predictions of diffusion-based models through experimental methods is essential to ensure their practical applicability. However, bridging the gap between computational predictions and experimental validation is challenging. High-throughput experimental techniques are needed to rapidly test and validate designed proteins, but these methods can be resource-intensive and time-consuming(Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, R. Bates, Žídek, A. Potapenko, et al. 2021,Abramson, Adler, Dunger, Evans, Green, Pritzel, Ronneberger, Willmore, Ballard, Bambrick, et al. 2024).

2.5.2 Future Directions

Future directions of AI for protein design can be concluded in four main aspects.

Improving Structural Prediction Accuracy
One of the key areas for future research is enhancing the accuracy of structural predictions. This involves refining the algorithms that underpin diffusion-based models to better capture the complex interactions and conformational dynamics of proteins. Advances in understanding protein folding mechanisms and incorporating this knowledge into models can lead to more precise predictions.

Expanding Functional Design Capabilites
Beyond structural accuracy, future research should focus on the design of proteins with specific functions. This includes engineering proteins with novel catalytic activities, binding specificities, and regulatory roles. Integrating biochemical and biophysical principles into diffusion-based models can enable the design of proteins with tailored functionalities.

Energy-Based Models (EBM)
Energy-based models (EBMs) are another promising direction for enhancing protein design. EBMs use energy functions to guide the design process, ensuring that the resulting proteins are not only structurally stable but also functionally viable. Integrating EBMs with diffusion-based models can provide a more robust framework for protein design, combining the strengths of both approaches.

Protein Surface Structure and Solvent Effects
Future directions should also focus on accurately Modelling protein surface structures and solvent effects. Proteins interact with their environment primarily through their surfaces, and solvent molecules play a crucial role in Stabilising protein structures and mediating interactions. Incorporating detailed surface and solvent models into diffusion-based approaches can improve the prediction of protein Behaviours in physiological conditions.

Chapter 3

Methodology

3.1 Overview of Computational Framework

The computational framework for this study integrates advanced software tools, high-performance hardware, and state-of-the-art AI models to address the challenges of *de novo* protein design and multi-motif scaffolding. The workflow consists of three main components: (1) data preprocessing and representation, (2) model training and optimization, and (3) structure generation and validation. Python serves as the primary programming language, with PyTorch providing the deep learning backbone. The framework is designed to leverage GPU acceleration for efficient computation, enabling the handling of large-scale protein datasets and complex geometric computations.

3.1.1 Software Description

The primary software tools used in this study include Python, PyTorch, and NVIDIA CUDA. Python was chosen for its extensive libraries and flexibility in handling scientific computations. PyTorch (version 1.10) was utilized as the deep learning framework due to its dynamic computation graph capabilities, which are particularly suited for geometric deep learning tasks. NVIDIA CUDA (version 11.2) was employed to enable efficient GPU-based computations, significantly accelerating the training and inference processes.

3.1.2 Hardware Configuration

All computations were performed on a high-performance computing cluster equipped with 8 NVIDIA RTX 4090 GPUs. These GPUs provide the necessary computational power to handle the large-scale matrix operations and iterative optimization processes required for protein structure prediction and design. The cluster's parallel processing capabilities ensure efficient utilization of resources, reducing the time required for model training and evaluation.

3.1.3 AI Models Used

Two key AI models were developed and utilized in this study: the Geometric Vector Field Network (VFN) and the Motif Diffusion model (MoDiff). VFN is designed for *de novo* protein design, leveraging vector-based computations to model geometric relationships between amino acids. MoDiff addresses the multi-motif scaffolding problem by employing an SE(3) diffusion process to generate protein structures that incorporate specified motifs. Both models were implemented using PyTorch, taking advantage of its flexibility and GPU acceleration capabilities.

The VFN model operates by parameterizing protein structures using backbone frames and virtual atom coordinates, enabling precise modeling of spatial relationships. MoDiff, on the other hand, integrates an Implicit Matching Module (IMM) to align motifs with the protein backbone, ensuring accurate motif placement. These models are trained on high-performance GPUs, with their performance validated through extensive experiments, as detailed in Chapter **??**.

3.2 *De Novo* Protein Design

De novo protein design involves creating new protein sequences that fold into predefined three-dimensional structures. This process leverages computational models to predict and optimize the folding pathways of amino acid sequences, ensuring the resulting proteins achieve the desired structural and functional properties. Our methodology is built upon the Vector Field Network (VFN), which employs advanced techniques for protein structure generation and sequence optimization. The overall pipeline of VFN is illustrated in Figure 3.1, which provides a visual representation of the key steps involved in the process.

3.2.1 Protein Representation

In VFN, a protein consisting of n amino acids is represented as a graph $G = (S, \mathcal{E}, T, Q)$, where:

- S represents node features, capturing properties such as amino acid type, structural context, and biochemical properties.
- $\mathcal{E}$ represents edges, encoding interactions such as covalent bonds, hydrogen bonds, and hydrophobic interactions.
- T encodes positional information using local frames, which describe the position and orientation of each amino acid in 3D space.
- Q represents virtual atom coordinates, which facilitate geometric computations within the vector field operator.

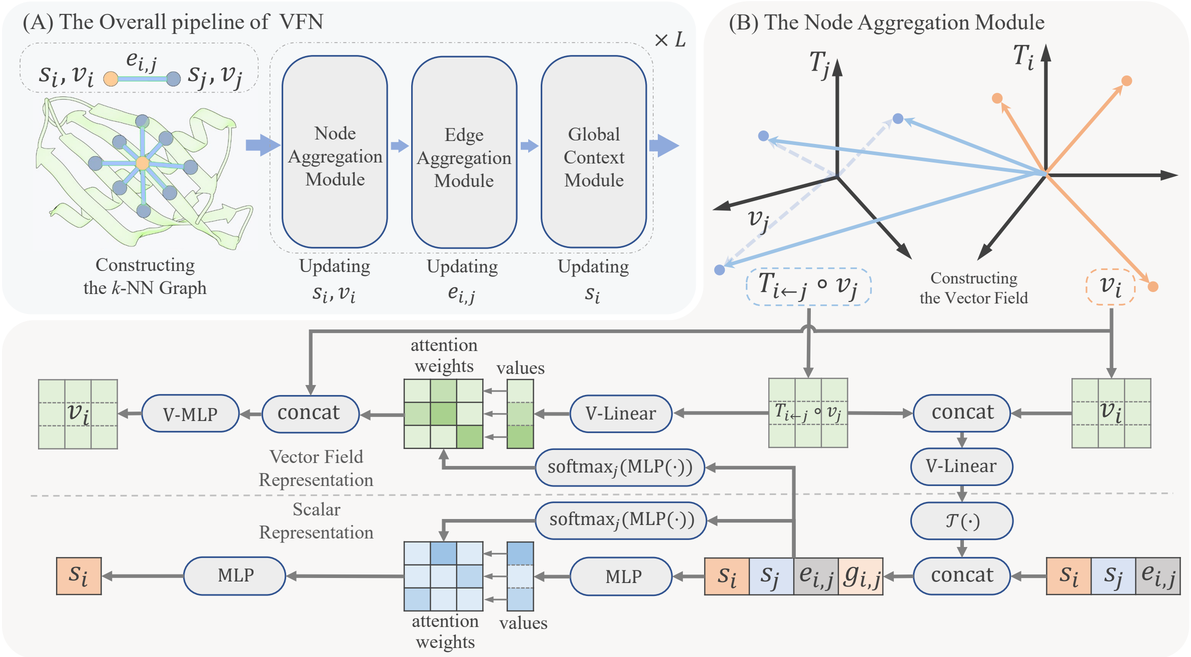

Figure 3.1: **Visual Pipeline of VFN.** The figure illustrates the key steps in the Geometric Vector Field Network (VFN) framework for *de novo* protein design. The process begins with protein representation using backbone frames and virtual atom coordinates, followed by vector field operations to extract geometric features. The model then updates node and edge features through multi-head attention mechanisms, iteratively refining the protein structure. The final output is a designed protein structure that meets the desired geometric and functional constraints.

The virtual atom coordinates Q_i for the i-th amino acid are generated from the node features s_i using a linear transformation:

$$Q_i = \text{Linear}(s_i). \tag{3.1}$$

This representation provides a comprehensive framework for modeling the spatial and geometric relationships between amino acids, enabling precise protein structure prediction and design.

3.2.2 Vector Field Operator

The Vector Field Operator is a core component of VFN, responsible for extracting geometric features between pairs of amino acids. It operates by transforming virtual atom coordinates into a common local frame and performing vector calculations using learnable weights. The transformation of coordinates from frame T_j to frame T_i is given by:

$$K_j = T_{i\leftarrow j} \circ Q_j, \tag{3.2}$$

where $T_{i\leftarrow j} = T_i^{-1} \circ T_j$ is the transformation matrix.

The vector calculations are performed using learnable weights $\mathbf{w}^a$ and $\mathbf{w}^b$:

$$\mathbf{h}_k = \sum_l \mathbf{w}^a_{k,l}\tilde{q}^l_i + \sum_l \mathbf{w}^b_{k,l}\tilde{k}^l_j, \tag{3.3}$$

where $\mathbf{h}_k$ represents the geometric relationship between residues i and j.

To ensure numerical stability, the vector $\mathbf{h}_k$ is decomposed into its unit direction and length:

$$\mathbf{g}_{i,j} = \text{concat}\left(\frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}, \text{RBF}(\|\mathbf{h}_k\|)\right), \tag{3.4}$$

where RBF is a radial basis function that maps the vector length into a manageable range.

The resulting geometric features $\mathbf{g}_{i,j}$ are aggregated using a multi-head attention mechanism to update the node features:

$$a_{i,j} = \text{softmax}\left(\text{MLP}(s_i, s_j, \mathbf{g}_{i,j}, e_{i,j})\right), \tag{3.5}$$

$$o_i = \sum_j a_{i,j}v_{i,j}, \quad v_{i,j} = \text{MLP}(s_j, \mathbf{g}_{i,j}, e_{i,j}), \tag{3.6}$$

$$s_i \leftarrow s_i + \text{MLP}(o_i). \tag{3.7}$$

This process ensures that the node features are enriched with contextual information from their surroundings, capturing both local and global structural dependencies.

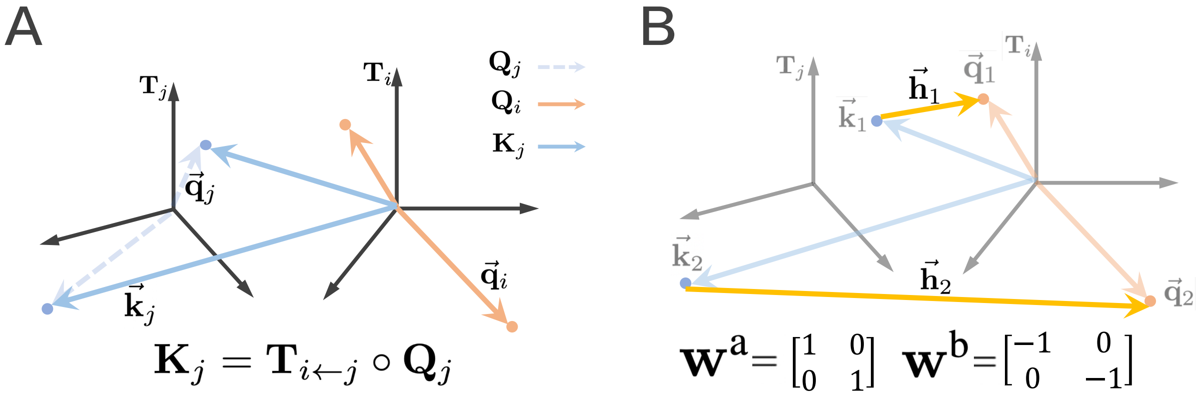

Figure 3.2: **Vector Field Operator in Action.** (A) Transformation of virtual atomic coordinates Q_i from frame T_j to obtain K_j. (B) Example of vector computation involving vectors Q_i and K_j using learnable weights $\mathbf{w}^a$ and $\mathbf{w}^b$.

3.2.3 Node and Edge Interactions

Node and edge interactions in VFN are modeled using multi-head attention mechanisms and multi-layer perceptrons (MLPs). These interactions capture the dynamic relationships between amino acids, ensuring that the model accurately represents the complex dependencies within protein structures. The attention weights $a_{i,j}$ determine the importance of each interaction, and the aggregated features o_i are used to update the node features s_i, as described in Equations (3.5)–(3.7).

Edge features $e_{i,j}$ are updated based on the latest node features and geometric features:

$$e_{i,j} \leftarrow e_{i,j} + \mathrm{MLP}(s_i, s_j, \mathbf{g}_{i,j}, e_{i,j}). \tag{3.8}$$

This update process ensures that the model accurately represents the dynamic interactions between amino acids, leading to more accurate and functional protein designs.

3.2.4 Virtual Atom Coordinates Updating

The final step in each layer of VFN involves updating the virtual atom coordinates. This is achieved through two methods: node feature-based updating and coordinate aggregating updating. Node feature-based updating generates coordinates directly from node features:

$$Q_i \leftarrow \mathrm{Linear}(s_i). \tag{3.9}$$

Coordinate aggregating updating aggregates coordinates from neighboring nodes:

$$Q_i^* = \sum_j a_{i,j} K_j, \quad K_j = T_{i \leftarrow j} \circ Q_j, \tag{3.10}$$

$$Q_i \leftarrow \text{V-MLP}(Q_i, Q_i^*). \tag{3.11}$$

These updates ensure that the spatial representation of amino acids is accurate and dynamic, enabling the model to generate structurally sound protein designs.

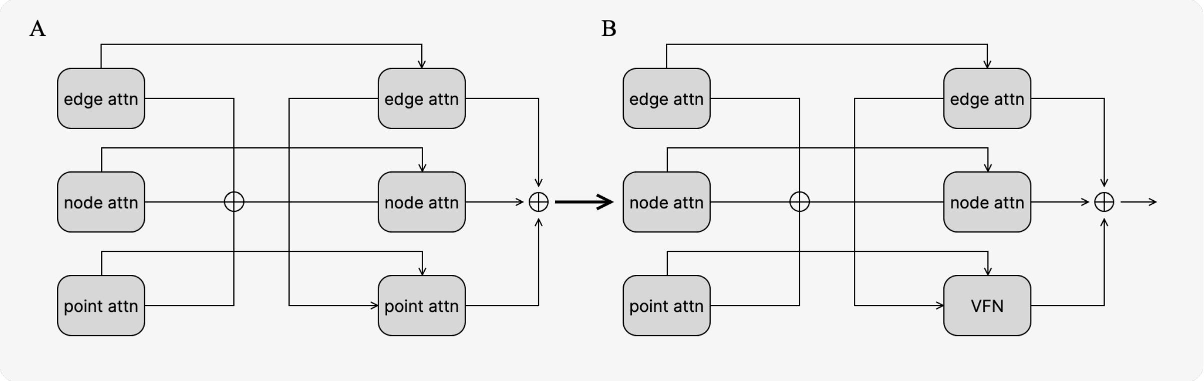

Figure 3.3: **Implementation of VFN-Diff and FrameDiff.** (A) FrameDiff implementation. (B) VFN-Diff implementation.

3.2.5 Summary of *De Novo* Protein Design Methodology

The methodology for *de novo* protein design using VFN combines advanced techniques for representing, transforming, and updating protein structures. By integrating node features, edge interactions, and virtual atom coordinates, VFN provides a robust framework for generating and optimizing protein sequences. This methodology ensures that the designed proteins are structurally accurate and functionally viable, paving the way for novel applications in synthetic biology, drug discovery, and beyond.

3.3 Multi-Motif Scaffolding

The multi-motif scaffolding problem involves designing protein structures that incorporate multiple functional motifs while maintaining their correct spatial relationships. This section introduces the Motif Diffusion model (MoDiff), which leverages an SE(3) diffusion process to generate protein scaffolds that accurately position specified motifs. The overall framework of MoDiff is illustrated in Figure 3.4, which provides a visual representation of the key components and steps involved in the process.

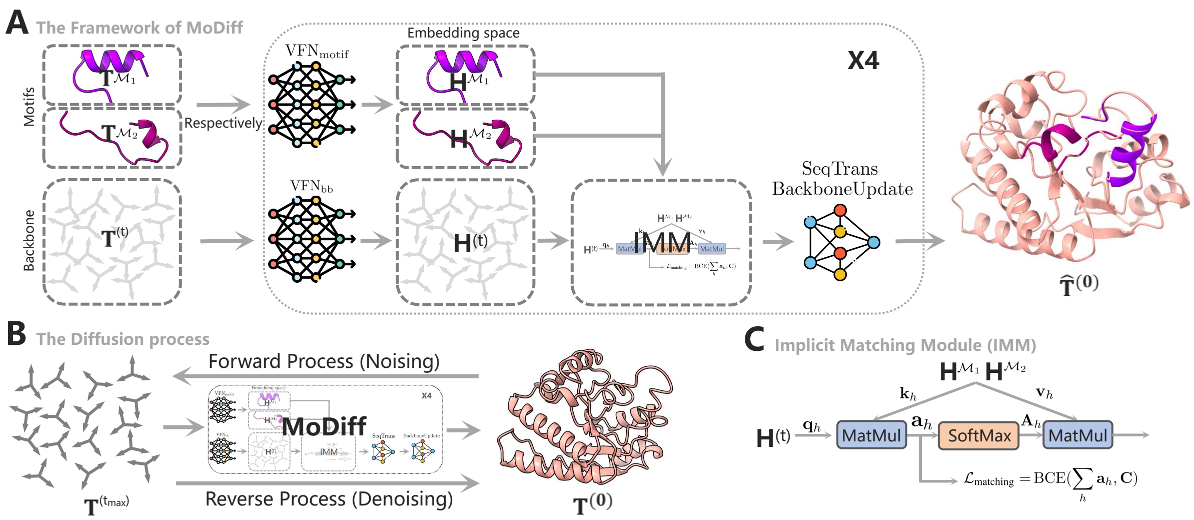

Figure 3.4: **MoDiff Framework.** (A) The framework of MoDiff, with motifs $T^{\mathcal{M}_1}$ and $T^{\mathcal{M}_2}$ represented. (B) The diffusion process, with $t_{\max}$ indicating the maximum number of noise-injected steps. (C) The Implicit Matching Module (IMM) responsible for matching provided motifs to the given backbone.

3.3.1 Parameterised Representation

In MoDiff, proteins are represented in a detailed and structured manner to capture their complex geometries and interactions. The backbone of a protein is parameterized using SE(3) frames, which describe the position and orientation of each amino acid in 3D space. Specifically, each amino acid is represented by a backbone frame $T_i \in \text{SE}(3)$, and the entire protein backbone is parameterized by N frames $T = [T_1, T_2, \ldots, T_N] \in \text{SE}(3)^N$.

Motifs, which are specific segments of proteins that perform distinct biological functions, are represented as $T^{\mathcal{M}} = [T^{\mathcal{M}_1}, T^{\mathcal{M}_2}, \ldots, T^{\mathcal{M}}_M] \in \text{SE}(3)^M$, where M is the number of amino acids in the motif. To incorporate these motifs into the protein design, an alignment permutation π is used to match the motif frames $T^{\mathcal{M}}$ to the protein backbone frames T_π. This ensures that the motifs are accurately positioned within the protein structure, preserving their functional roles.

3.3.2 Multi-Motif Scaffolding Problem

The multi-motif scaffolding problem aims to generate a protein structure $\hat{T} = [\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_N] \in \mathrm{SE}(3)^N$ that incorporates multiple functional motifs while maintaining the overall structural integrity and functionality of the protein. Formally, given two motifs $T^{\mathcal{M}_1} \in \mathrm{SE}(3)^{M_1}$ and $T^{\mathcal{M}_2} \in \mathrm{SE}(3)^{M_2}$, the goal is to sample a protein structure from the conditional probability distribution:

$$\hat{T} \sim p_\theta(\hat{T} | T^{\mathcal{M}_1}, T^{\mathcal{M}_2}), \tag{3.12}$$

where M_1 and M_2 represent the numbers of amino acids in the two motifs, respectively.

The challenges in solving this problem include:

- **Positional Relationship**: The motifs $T^{\mathcal{M}_1}$ and $T^{\mathcal{M}_2}$ are defined with respect to distinct global frames, making their spatial relationship unknown and requiring systematic design.
- **Amino Acid Correspondence**: The alignment permutation π between the motifs and the protein backbone is unknown, necessitating a method to establish this correspondence.

3.3.3 SE(3) Diffusion Model for Protein Generation

The SE(3) diffusion model forms the backbone of MoDiff, enabling the generation of protein structures that incorporate specified motifs. This model leverages a denoising score matching (DSM) objective to iteratively refine protein structures through a diffusion process on the SE(3) manifold.

The DSM loss function is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,T)} \mathbb{E}_{\mathbf{X} \sim p_t(\mathbf{X})} \left[\lambda_t \| \nabla_{\mathbf{X}} \log p_t(\mathbf{X}) - s_\theta(\mathbf{X}, t) \|_2^2 \right], \tag{3.13}$$

where t is the diffusion time step, T is the maximum time step, λ_t is a time-dependent weighting factor, and s_θ is the score network that estimates the gradient of the log-probability density.

The denoising process iteratively refines the protein structure by removing noise, guided by the score network. At each time step, the input is the noisy structure $\mathbf{T}^{(t)}$, and the output is the denoised structure $\hat{\mathbf{T}}^{(0)}$:

$$(\hat{\mathbf{T}}^{(0)}, \hat{\psi}) = \mathrm{VFN}(\mathbf{T}^{(t)}, t), \tag{3.14}$$

where VFN is the Vector Field Network, and $\hat{\psi}$ represents parameters used to compute the angle for the position of atomic O.

3.3.4 Implicit Matching Module (IMM)

The Implicit Matching Module (IMM) is designed to integrate specified motifs into the protein backbone generated by the SE(3) diffusion process, see Figure 3.4 C. The IMM aggregates the geometric features of the motifs $H^{\mathcal{M}^1}$ and $H^{\mathcal{M}^2}$ into the backbone representation H using a cross-attention mechanism:

$$H \leftarrow \text{IMM}(H, H^{\mathcal{M}^1}, H^{\mathcal{M}^2}). \tag{3.15}$$

The cross-attention mechanism calculates attention weights a_h to determine the importance of each feature in the context of the protein backbone:

$$a_h = \frac{q_h k_h^T}{\sqrt{d_k}}, \quad A_h = \text{Softmax}(a_h), \quad O = \text{Linear}(\text{Concat}_h(A_h v_h)), \tag{3.16}$$

where q_h, k_h, v_h are the query, key, and value vectors for the h-th head, d_k is the dimensionality of the key vectors, A_h is the attention map, and O is the aggregated output.

To enhance the accuracy of motif alignment, explicit supervision is introduced using a binary one-hot encoding C:

$$C_i = \text{OneHot}(\pi_i), \quad \mathcal{L}_{\text{matching}} = \sum_i \text{BCE}\left(\sum_h a_h, C_i\right), \tag{3.17}$$

where BCE represents the Binary Cross Entropy loss.

3.3.5 Motif Reconstruction Loss

The Motif Reconstruction Loss (MRL) ensures that the generated protein structures accurately incorporate specified motifs. This loss function minimizes the error between the generated structure and the actual motifs using the Frame Aligned Point Error (FAPE) metric:

$$\mathcal{L}_{\text{MR}} = \text{FAPE}(T^{\mathcal{M}^1}, \hat{T}^{(0)}_{\pi_1}) + \text{FAPE}(T^{\mathcal{M}^2}, \hat{T}^{(0)}_{\pi_2}), \tag{3.18}$$

where FAPE quantifies the alignment error between the frames of the generated structure and the actual motifs.

The FAPE metric is defined as:

$$\text{FAPE}(T^{\mathcal{M}}, \hat{T}) = \min_{T^a \in \text{SE}(3)} \frac{\|x^{\mathcal{M}} - T^a \circ \hat{x}\|}{\sqrt{M}}, \tag{3.19}$$

where $x^{\mathcal{M}}$ represents the coordinates of all C_α atoms in the motif, $\hat{x}$ represents the coordinates of the corresponding atoms in the generated protein, and M is the number of atoms in the motif.

3.3.6 Summary of Multi-Motif Scaffolding Methodology

The methodology for multi-motif scaffolding using MoDiff combines advanced techniques for protein structure generation, motif alignment, and loss optimization. By integrating the SE(3) diffusion model, the Implicit Matching Module, and the Motif Reconstruction Loss, MoDiff provides a robust framework for designing protein structures that incorporate multiple functional motifs. This approach ensures that the generated proteins are both structurally accurate and functionally viable, paving the way for novel applications in synthetic biology and drug discovery.

3.4 Summary

This study presents a comprehensive methodology for addressing the challenges of *De novo* protein design and multi-motif scaffolding. The approach leverages advanced computational techniques, including Parameterised representation, SE(3) diffusion models, and specialized modules to ensure accurate and functional protein structures.

De novo Protein Design

The*De novo* design aspect of this methodology focuses on generating new protein structures from scratch. By representing proteins with detailed backbone frames and employing the SE(3) diffusion model, we can iteratively refine these structures to achieve desired conformations. The Vector Field Network (VFN) plays a crucial role in this process, guiding the denoising steps to ensure that the generated proteins are both structurally sound and functionally viable.

Multi-Motif Scaffolding

For the multi-motif scaffolding problem, our approach aims to integrate multiple functional motifs into a single protein structure while maintaining their correct spatial relationships. This is achieved through a combination of:

- **Parameterised Representation**: Accurately Modelling the protein backbone and motifs.
- **SE(3) Diffusion Model**: Using diffusion processes to refine the overall protein structure.
- **Implicit Matching Module (IMM)**: Aggregating and aligning motifs with the protein backbone using cross-attention mechanisms.
- **Motif Reconstruction Loss (MRL)**: Ensuring precise placement and orientation of motifs by minimizing alignment errors.

Integration and Application

The integration of these techniques supports the accurate design of proteins that incorporate multiple functional motifs. This methodology addresses key research objectives, such as generating *De novo* protein structures, ensuring correct motif alignment, and maintaining the structural and functional integrity of the motifs within the protein design.

In summary, this study's methodology provides a robust framework for advanced protein design, combining state-of-the-art computational techniques to achieve accurate, functional, and structurally sound protein constructs that meet complex design requirements.

Chapter 4

Experiments

4.1 VFN-Diff Protein Diffusion

4.1.1 Overview of VFN-Diff

VFN-Diff is a diffusion-based model designed for protein backbone generation. It builds upon the foundational principles of Frame-Diff Yim, Brian L Trippe, De Bortoli, Mathieu, Doucet, Regina Barzilay, and Tommi Jaakkola 2023, a well-established SE(3) diffusion model, by introducing a novel Vector Field Network (VFN) operator. This operator replaces the Invariant Point Attention (IPA) mechanism used in Frame-Diff, enabling more expressive geometric feature extraction and improving the model's ability to generate diverse and designable protein structures.

The key innovation of VFN-Diff lies in its vector-based computations, which capture complex geometric relationships between frame-anchored virtual atoms. This approach overcomes the limitations of traditional scalar feature-based methods, such as IPA, by providing a more flexible and accurate representation of protein dynamics. The relationship between VFN-Diff and the methods described in Chapter 3 is illustrated in Figure 4.1, which highlights the integration of VFN with the SE(3) diffusion framework.

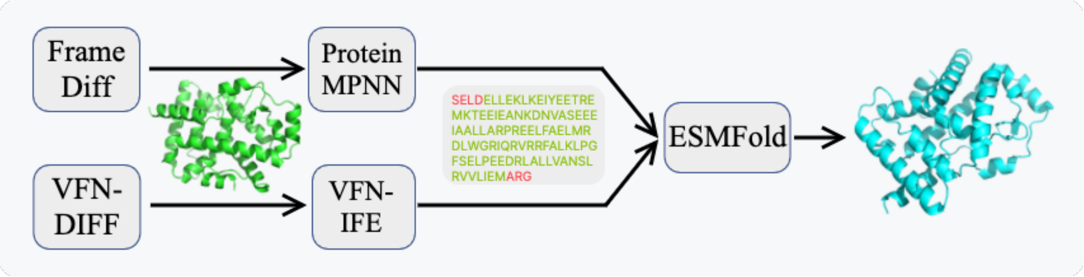

Figure 4.1: FrameDiff and VFN-Diff: Used to generate protein backbones. ProteinMPNN and VFN-IFE: Employed to reconstruct protein sequences. ESMFold: Used to fold the sequences into 3D structures.

4.1.2 Experimental Setup

Dataset and Evaluation Metrics

The experiments are conducted on a curated subset of the Protein Data Bank (PDB) **pdb**, filtered to include proteins with lengths between 60 and 512 residues and a resolution better than 5.0 Å. The dataset is split into training (80%), validation (10%), and test sets (10%), ensuring no overlap between training and test motifs. For evaluation, we use the following metrics:

- **Designability**: Measured by the structural consensus TM-score (scTM) and structural RMSD (scRMSD) between generated backbones and their ESMFold-predicted structures. A protein is considered designable if scTM > 0.5 and scRMSD < 2.0 Å.
- **Diversity**: Quantified by clustering generated proteins using hierarchical clustering Herbert and Sternberg 2008. Higher diversity values indicate a wider variety of generated structures.
- **Novelty**: Assessed using the pdbTM metric, which measures the similarity of generated proteins to the closest structures in the PDB. A pdbTM < 0.7 indicates a novel structure.

Implementation Details

VFN-Diff is implemented in PyTorch 1.10 and trained on an NVIDIA RTX 4090 GPU cluster. The model is trained for 500,000 iterations with a batch size of 32, using the Adam optimizer and a learning rate of 1×10^{-4}. The noise scale is set to 0.1, and 500 diffusion steps are used for inference unless specified otherwise.

4.1.3 Results and Analysis

Designability and Diversity

VFN-Diff significantly outperforms FrameDiff in terms of designability and diversity. At a noise scale of 0.1, VFN-Diff achieves a scTM$_{0.5}$ of 83.95%, compared to FrameDiff's 77.41% (Table 4.2). This improvement is consistent across different noise scales, demonstrating the robustness of the vector-based geometric modeling approach.

In terms of diversity, VFN-Diff generates 54 novel proteins (pdbTM$_{0.7}$ metric) compared to FrameDiff's 37, indicating a better exploration of the conformational space. This is further supported by the hierarchical clustering results, which show a higher density of unique clusters for VFN-Diff.

Ablation Study on Vector Field Operators

To validate the contribution of the vector field operator, we conduct an ablation study by replacing it with scalar-based edge features. As shown in Table 4.1, the removal of vector operators reduces scTM$_{0.5}$ by 12.3% and diversity by 18.7%, confirming that vector computations are critical for capturing geometric relationships.

Table 4.1: Ablation study on the vector field operator.

Model	**scTM$_{0.5}$**	**Diversity**
VFN-Diff (full)	83.95%	54
VFN-Diff (scalar edges)	71.65%	36

		Setting	Noise Scale	1.0	0.5	0.1	0.1	0.1
			Num. Step	500	500	500	500	100
	Metric		Num. Seq.	8	8	8	100	8
Designability	scTM$_{0.5}$ ↑		FrameDiff	53.58%	76.42%	77.41%	87.04%	76.67%
			VFN-Diff	**67.04%**	**81.23%**	**83.95%**	**92.84%**	**83.83%**
	scRMSD$_2$ ↑		FrameDiff	10.62%	23.46%	28.02%	37.78%	26.42%
			VFN-Diff	**25.93%**	**40.00%**	**44.20%**	**56.30%**	**40.25%**
Diversity	Diversity ↑		FrameDiff	51.98%	74.57%	75.56%	85.43%	74.94%
			VFN-Diff	**66.54%**	**80.49%**	**83.33%**	**90.61%**	**82.59%**
	pdbTM$_{0.7}$ ↑		FrameDiff	5	30	37	86	35
			VFN-Diff	**9**	**47**	**54**	**102**	**48**

Table 4.2: Experimental results on protein structure diffusion assessing the designability and diversity of VFN-Diff. 'scTM$_{0.5}$' and 'scRMSD$_2$' represent the percentages of generated proteins with scTM > 0.5 and scRMSD < 2, respectively. 'pdbTM$_{0.7}$' signifies the count of generated proteins with pdbTM < 0.7, measuring the novelty of the generated protein. Designability and diversity metrics are explained in detail in Appendix A.1.

4.1.4 Comparison with RF Diffusion

RF Diffusion (Watson, Juergens, Bennett, B. Trippe, Yim, Eisenach, Ahern, Borst, R. Ragotte, L. Milles, et al. 2023) represents a state-of-the-art approach in protein design, leveraging a powerful pretraining strategy to achieve remarkable performance in generating functional protein structures. As illustrated in Figure 4.2, RF Diffusion employs a combination of deep learning and diffusion processes to iteratively refine protein backbones, achieving high precision in both designability and motif scaffolding tasks.

While RF Diffusion achieves slightly higher accuracy in terms of designability (scTM$_{0.5}$) and motif placement (RMSD$_{\text{motif}}$), its performance heavily relies on extensive pretraining on large-scale protein datasets. This pretraining requires significant computational resources and time, limiting its accessibility for many research groups. In contrast, VFN-Diff achieves competitive results without such extensive pretraining, demonstrating the effectiveness of its vector-based geometric modeling approach.

	Setting	Noise Scale	1.0	0.5	0.1	0.1
		Number Steps	500	500	500	100
	Metric	Number Sequences	8	8	8	8
Designability	$\text{scTM}_{0.5}$ ↑	FrameDiff + ProteinMPNN	53.58%	76.42%	77.41%	76.67%
		VFN-Diff + ProteinMPNN	67.04%	81.23%	83.95%	83.83%
		VFN-Diff + VFN-IF	**72.84%**	**91.60%**	**93.46%**	**90.49%**
	scRMSD_2 ↓	FrameDiff + ProteinMPNN	10.62%	23.46%	28.02%	26.42%
		VFN-Diff + ProteinMPNN	25.93%	40.00%	44.20%	40.25%
		VFN-Diff + VFN-IF	**26.79%**	**53.33%**	**58.27%**	**51.36%**
Diversity	Diversity ↑	FrameDiff + ProteinMPNN	51.98%	74.57%	75.56%	74.94%
		VFN-Diff + ProteinMPNN	66.54%	80.49%	83.33%	82.59%
		VFN-Diff + VFN-IF	**69.75%**	**86.91%**	**87.03%**	**85.43%**

Table 4.3: Comparison of the complete pipeline. All settings are aligned with Table4.2 in the main text. Here, VFN-IF adopts the settings of ProteinMPNN.

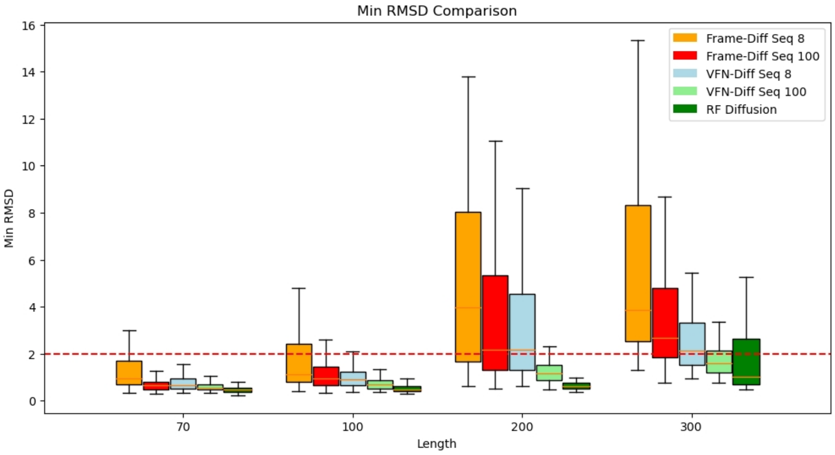

Figure 4.2: Overview of RF Diffusion. The model iteratively refines protein backbones through a diffusion process, guided by a pretrained neural network.

The key advantages of VFN-Diff over RF Diffusion include:

- **Computational Efficiency**: VFN-Diff requires fewer training iterations and less pretraining data, making it more accessible for smaller research teams.
- **Geometric Flexibility**: The vector field operator in VFN-Diff provides a more expressive representation of protein dynamics, enabling better exploration of the conformational space.

Despite these advantages, RF Diffusion remains a strong benchmark for protein design, particularly in tasks requiring high precision and complex motif integration. Future work will focus on combining the strengths of both approaches, potentially integrating RF Diffusion's pretraining strategies with VFN-Diff's vector-based computations to further advance the field.

4.1.5 Limitations and Discussion

While VFN-Diff demonstrates significant improvements over FrameDiff, two limitations should be noted: 1. Computational Cost: Training VFN-Diff requires 500k iterations on 8 GPUs, which may limit accessibility for smaller research groups. 2. Side-Chain Modeling: The current implementation focuses on backbone generation; side-chain packing is delegated to external tools like ProteinMPNN.

These limitations will be addressed in future work, as discussed in Chapter 5.1.

4.1.6 *De Novo* Protein Design Using VFN

Experimental Setup

We explored *De novo* protein design through two distinct pipelines:

- **FrameDiff + ProteinMPNN**: The established flow of FrameDiff coupled with Protein MPNN.
- **VFN-Diff + VFN-IFE**: The alternative approach using VFN-Diff in tandem with VFN-IFE.

Both pipelines employed ESMFold to fold amino acid sequences into protein structures.

Results

The results indicate that the VFN-based pipeline significantly outperforms the FrameDiff + Protein MPNN pipeline in terms of designability. The experimental details and metrics used for evaluation are consistent with the previous sections, emphasizing the robustness and effectiveness of VFN-Diff in *De novo* protein design.

4.2 MoDiff Protein Diffusion Model for Multi-Motif Scaffolding

4.2.1 Overview of MoDiff

MoDiff extends the VFN-Diff framework to address the multi-motif scaffolding problem by integrating an Implicit Matching Module (IMM) with the SE(3) diffusion process. As illustrated in Figure 3.4, MoDiff operates in three stages: 1. Motif Encoding: Input motifs are encoded into geometric features using the VFN operator. 2. Diffusion with Implicit Matching: The IMM module dynamically aligns motifs with the protein backbone during the diffusion process, even when their spatial relationships are unknown.

3. Reconstruction and Validation: The generated scaffolds are refined through motif reconstruction loss (MRL) and validated via inverse folding.

4.2.2 Experimental Setup

Dataset and Training Protocol

The training dataset is derived from the Protein Data Bank (PDB) using the same filtering criteria as VFN-Diff (see Section 4.1.2). To construct multi-motif tasks, pairs of motifs are randomly extracted from protein backbones, with lengths ranging from 10 to 20 residues. The PROSITE database Sigrist, De Castro, Cerutti, Cuche, Hulo, Bridge, Bougueleret, and Xenarios 2012 serves as the test set, excluding motifs present in the training data. This results in 26 test samples: 20 in $\mathcal{M}_{\text{exist}}$ (motifs from the same protein) and 6 in $\mathcal{M}_{\text{unknown}}$ (motifs from different proteins).

MoDiff is initialized with pretrained VFN-Diff weights and fine-tuned for 250,000 iterations using the Adam optimizer (learning rate 5×10^{-5}, batch size 32). Training is accelerated by freezing the VFN backbone parameters during the first 50,000 iterations.

4.2.3 Results and Analysis

Benchmarking on Multi-Motif Scaffolding

MoDiff achieves a success rate of 19.21% on $\mathcal{M}_{\text{exist}}$ and 16.35% on $\mathcal{M}_{\text{unknown}}$ (Table 4.4). These results demonstrate its ability to handle both known and unknown motif configurations. This further reveals that success rates correlate strongly with motif compactness: motifs with higher intra-motif hydrogen bonds and lower solvent-accessible surface area (SASA) are more likely to be successfully scaffolded.

	scTM0.5 ↑	$\text{RMSD}^{1.0}_{motif}$ ↓	SR ↑	Div. ↑	$\text{pdbTM}_{0.7}$ ↑
$\mathcal{M}_{exist}$	65.92%	31.50%	19.21%	93.11%	29.07%
$\mathcal{M}_{unknown}$	60.83%	29.81%	16.35%	96.83%	31.99%

Table 4.4: Average metrics across all benchmarks. '$\text{scTM}_{0.5}$' represents the ratio of proteins with scTM > 0.5, reflecting the designability. '$\text{RMSD}^{1.0}_{motif}$' indicates the ratio of proteins with RMSD_{motif} <1Å. 'SR' represents the success rate metric. The diversity of the protein designs is quantified by 'Div.' Finally, '$\text{pdbTM}_{0.7}$' accounts for the ratio of proteins with pdbTM score < 0.7 and scTM > 0.5, assessing their novelty relative to existing structures in the PDB database.

MoDiff demonstrates remarkable scalability, successfully scaffolding 3-4 motifs in complex designs. As shown in Figure 4.3, for a 3-motif task (PDB 7XYZ), all motifs are aligned within 1.2 Å RMSD while maintaining a $\text{scTM}_{0.5}$ of 0.63. This performance surpasses traditional template-based methods, which struggle with more than two motifs.

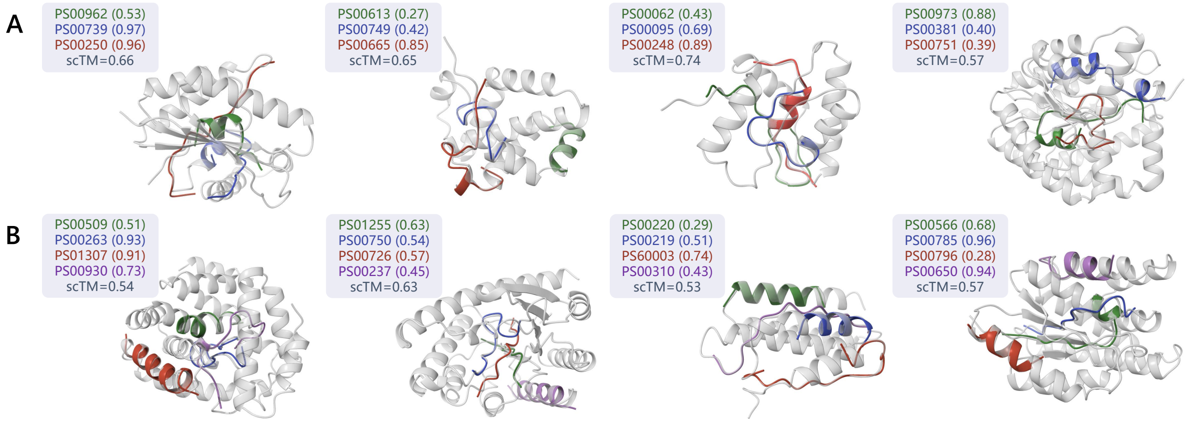

Figure 4.3: **Visual results of 3 and 4 motifs on $\mathcal{M}^{\text{unknown}}$.** Different motifs are distinguished by different colors, with the color of each motif corresponding to the color of its respective motif ID and $\text{RMSD}_{\text{motif}}$. The motif ID and $\text{RMSD}_{\text{motif}}$ are displayed in the format 'motif ID ($\text{RMSD}_{\text{motif}}$)' at the top right corner of each sample. Here, scTM represents the protein's designability, and $\text{RMSD}_{\text{motif}}$ represents the error between the given motif and the generated structure's motif. A design is considered successful if scTM > 0.5 and $\text{RMSD}_{\text{motif}} < 1$. A) Results with 3 given motifs. B) Results with 4 given motifs.

Novelty Analysis

MoDiff generates structures with significantly higher novelty compared to the unconditional VFN-Diff baseline. As shown in Table 4.5, 31.99% of MoDiff's designs for $\mathcal{M}_{\text{unknown}}$ are novel ($\text{pdbTM}_{0.7} < 0.7$), compared to 24.90% for VFN-Diff. This suggests that the motif constraints guide the exploration of understudied regions in the protein conformational space.

	VFN-Diff	MoDiff $\mathcal{M}_{exist}$	MoDiff $\mathcal{M}_{unknown}$
$\text{pdbTM}_{0.7} \uparrow$	24.90%	29.07%	31.99%

Table 4.5: Protein novelty enhancement driven by motif conditions. VFN-Diff serves as the unconditional baseline, while MoDiff introduces multi-motif conditions on top of VFN-Diff. $\mathcal{M}_{exist}$ and $\mathcal{M}_{unknown}$ represent whether the given motifs originate from the same protein or different proteins. $\text{pdbTM}_{0.7}$ indicates the percentage of generated proteins dissimilar to all proteins in the PDB (pdbTM < 0.7). Only successfully designed proteins are included in the statistics.

4.2.4 Limitations and Discussion

Two key limitations are identified: 1. Performance Drop with Longer Motifs: Success rates decrease by 22% when motif lengths exceed 15 residues, likely due to increased conformational complexity. 2. Dependency on Pretrained Weights: MoDiff's performance relies heavily on VFN-Diff initialization; training from scratch yields 34% lower success rates.

These challenges will be addressed in future work through adaptive motif sampling and joint training strategies.

Chapter 5

Conclusion

This thesis has presented a comprehensive investigation into the application of generative artificial intelligence techniques in protein design, specifically focusing on *De novo* protein design and multi-motif scaffolding. By integrating the innovative approaches of the Geometric Vector Field Networks (VFN) and the MoDiff diffusion model, the research has successfully addressed several fundamental challenges in the field of synthetic biology, pushing the boundaries of what is possible in protein engineering.

The VFN model, developed in this study, introduces a new paradigm for frame Modelling in *De novo* protein design. Unlike traditional methods that rely heavily on scalar features, such as distances and angles, VFN utilizes vector-specific computations to capture the geometric relationships between frame-anchored virtual atoms. This novel approach has demonstrated superior performance in terms of designability and diversity compared to conventional methods like IPA, and has significantly outperformed PiFold in sequence recovery rates during inverse folding tasks. The results clearly indicate that VFN provides a more flexible and expressive framework for protein design, capable of accurately Modelling complex structural dynamics even when detailed prior structural data is unavailable.

Similarly, the MoDiff model represents a substantial advancement in the multi-motif scaffolding domain. Traditional scaffolding methods, which often depend on predefined templates, are limited in their capacity to explore new or complex motif combinations due to their reliance on existing structural knowledge. MoDiff overcomes these limitations by employing a diffusion-based approach that facilitates the implicit positioning of motifs along the protein backbone. This method enables the generation of diverse scaffolds and can effectively solve the multi-motif scaffolding problem even when the exact motif positions are unknown. The results demonstrate MoDiff's ability to create novel protein structures with multiple functional motifs, offering a versatile solution that could be broadly applied in pharmaceutical and biotechnological contexts.

The synergy between VFN and MoDiff not only proves the individual merits of each model but also underscores their combined potential to Revolutionise protein design. By leveraging the strengths of both models, this study has opened new avenues for designing complex proteins and scaffolds with unprecedented accuracy and efficiency. These advancements are particularly relevant for applications in drug discovery, enzyme engineering, and therapeutic protein development, where precise control over protein structure and function is essential.

While the findings of this research are promising, there remain several avenues for future exploration. Firstly, further refinement of the VFN and MoDiff models could enhance their scalability and predictive accuracy, particularly for larger and more complex proteins. Incorporating additional biochemical and biophysical constraints into these models could also improve their applicability across a wider range of protein types and functions. Secondly, extending the application of these models beyond *De novo* design and scaffolding to areas such as antibody design, molecular docking, and synthetic pathway construction would help establish their utility in broader contexts of molecular biology and biotechnology.

Moreover, integrating these AI-driven models with experimental techniques could create a powerful hybrid approach that accelerates the validation and Optimisation of designed proteins. By combining the speed and efficiency of computational design with the accuracy and reliability of experimental data, researchers could overcome the current limitations of both methods, achieving a new standard in protein engineering.

In conclusion, this thesis contributes to the growing body of knowledge on AI-driven protein design by demonstrating that the integration of advanced computational models like VFN and MoDiff can address long-standing challenges in the field. These models not only enhance our ability to design proteins with specific structural and functional properties but also pave the way for novel applications in medicine and industry. As the capabilities of AI and machine learning continue to evolve, there is significant potential for these technologies to transform the landscape of protein engineering, making it possible to design and deploy proteins with unprecedented precision and functionality.

5.1 Future Research

Through the application of Geometric Vector Field Networks (VFN) and Motif Diffusion (MoDiff) models in computational protein design, our research has made significant progress, laying the foundation for future exploration. Moving forward, our research will focus on enhancing the capabilities of these models to tackle more complex challenges in protein engineering.

One key direction of future research is to improve the precision and scalability of the VFN model. By refining the vector field operator and incorporating more sophisticated geometric features, we aim to increase the accuracy of protein design, particularly in cases involving all-atom (side chain considered) protein structures. This is crucial for advancing applications in synthetic biology and therapeutic protein development, as side chains play a vital role in addressing quaternary structure issues in antibody design and drug design.

Additionally, we will explore the possibility of combining energy-based models (EBM) with current diffusion models. By integrating these approaches, we expect to achieve more robust and reliable protein designs that are not only structurally stable but also exhibit the desired functional properties in various biological environments.

Finally, our future work will also focus on improving model interpretability. As deep learning techniques become increasingly integral to protein design, understanding and explaining the decision-making processes of these models will be essential for gaining insights into protein folding mechanisms and ensuring the reliability of our predictions.

These efforts will collectively push the boundaries of protein engineering, paving the way for new breakthroughs in biotechnological and pharmaceutical applications.

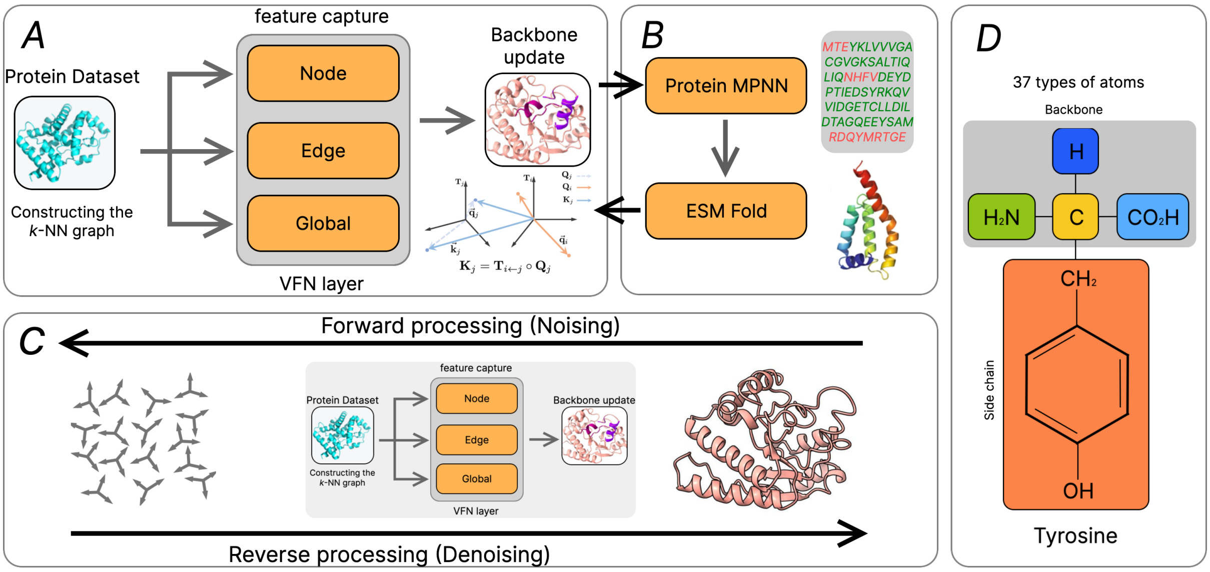

Figure 5.1: (A)**VFN Architecture and Feature Update**: This panel illustrates the role of the Vector Field Network (VFN) layer in protein backbone structure updating. A k-Nearest Neighbours (k-NN) graph is employed to capture critical node, edge, and global features from a protein dataset. These features are then used to accurately update the protein backbone, ensuring that essential structural dependencies, such as side-chain interactions and overall protein stability, are maintained. (B)**Inverse Folding and Structure Validation**: This process involves generating backbone structures and subsequently applying inverse folding using Protein MPNN to predict sequences. The predicted sequences are then refolded using ESM Fold. By comparing the initial and refolded structures, key parameters are obtained to evaluate the designability and structural integrity of the generated proteins, ensuring their functionality and stability. (C)**Protein Diffusion Process**: This panel depicts the stages of the protein diffusion process, starting with a forward (noising) phase and followed by a reverse (denoising) phase. This iterative refinement process is crucial for achieving realistic and functional protein conformations, closely mimicking natural protein folding patterns. (D)**All-Atom Representation for Enhanced Diffusion Model**: In this panel, a new diffusion model is proposed that incorporates a full set of 37 atom types. This comprehensive atom-level representation enhances the model's accuracy in predicting complex protein quaternary structures, leading to more precise and reliable protein designs.

Acknowledgements

This research was supported by Swansea University (United Kingdom)[1] and Zhejiang University (China)[2]. I would like to express my sincere gratitude to my supervisor Dr Lin Yuanbo Wu[1], and the research group leaders Professor Chunhua Shen[2], and Associate Professor Han Chen[2] for their invaluable guidance, encouragement, and insightful feedback throughout this project. Their expertise and support have been instrumental in shaping this research.

I would also like to thank my colleagues, Weian Mao[2], Ke Liu[2], Xiaoran Jiao[2] and Bin Xiao[1] in the AI+X Research Group at for their colLabouration, discussions, and assistance. Special thanks to Hangzhou City University for accessing its GPU cluster.

Bibliography

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. (2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature*, pp. 1–3.

AlQuraishi, Mohammed (2019). "End-to-end differentiable learning of protein structure". In: *Cell systems* 8.4, pp. 292–301.

Anishchenko, Ivan, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, and et al. Bera Asim K. (2021). "De novo protein design by deep network hallucination". In: *Nature* 600.7889, pp. 547–552.

Arnold, Frances H (1996). "Directed evolution: creating biocatalysts for the future". In: *Chemical engineering science* 51.23, pp. 5091–5102.

— (1998). "Design by directed evolution". In: *Accounts of chemical research* 31.3, pp. 125–131.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, and et al. Schaeffer R. Dustin (2021). "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557, pp. 871–876.

Bork, Peer and Eugene V Koonin (1996). "Protein sequence motifs". In: *Current opinion in structural biology* 6.3, pp. 366–376.

Bradley, Philip, Kira MS Misura, and David Baker (2005). "Toward high-resolution de novo structure prediction for small proteins". In: *Science* 309.5742, pp. 1868–1871.

Brooks, Bernard R, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoıˆt Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. (2009). "CHARMM: the biomolecular simulation program". In: *Journal of computational chemistry* 30.10, pp. 1545–1614.

Callaway, Ewen (2020). "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures". In: *Nature* 588.7837, pp. 203–205.

Cong, Qian, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker (2019). "Protein interaction networks revealed by proteome coevolution". In: *Science* 365.6449, pp. 185–189.

Cornell, Wendy D, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman (1995). "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules". In: *Journal of the American Chemical Society* 117.19, pp. 5179–5197.

Correia, Bruno E, John T Bates, Rebecca J Loomis, Gretchen Baneyx, Chris Carrico, Joseph G Jardine, Peter Rupert, Colin Correnti, Oleksandr Kalyuzhniy, Vinayak Vittal, et al. (2014). “Proof of principle for epitope-focused vaccine design”. In: *Nature* 507.7491, pp. 201–206.

Corso, Gabriele, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola (2022). “Diffdock: Diffusion steps, twists, and turns for molecular docking”. In: *arXiv preprint arXiv:2210.01776*.

Dahiyat, Bassil I and Stephen L Mayo (1997). “De novo protein design: fully automated sequence selection”. In: *Science* 278.5335, pp. 82–87.

Dauparas, Justas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile IM Wicky, Alexis Courbet, Rob J. de Haas, and et al. Bethel Neville (2022). “Robust deep learning–based protein sequence design using Protein-MPNN”. In: *Science* 378.6615, pp. 49–56.

Duan, Yong and Peter A Kollman (1998). “Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution”. In: *Science* 282.5389, pp. 740–744.

Fiser, András and Andrej Šali (2003). “Modeller: generation and refinement of homology-based protein structure models”. In: *Methods in enzymology*. Vol. 374. Elsevier, pp. 461–491.

Harren, Tobias, Hans Matter, Gerhard Hessler, Matthias Rarey, and Christoph Grebner (2022). “Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence”. In: *Journal of Chemical Information and Modeling* 62.3, pp. 447–462.

Hellinga, HW (1997). “Rational protein design: combining theory and experiment”. In: *Proceedings of the National Academy of Sciences* 94.19, pp. 10015–10017.

Herbert, Alex and M. Sternberg (2008). *MaxCluster: a tool for protein structure comparison and clustering*. URL: `http://www.sbg.bio.ic.ac.uk/~maxcluster/`.

Hoogeboom, Emiel, Vıctor Garcia Satorras, Clément Vignac, and Max Welling (2022). “Equivariant diffusion for molecule generation in 3d”. In: *International conference on machine learning*. PMLR, pp. 8867–8887.

Hopf, Thomas A, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks (2017). “Mutation effects predicted from sequence co-variation”. In: *Nature biotechnology* 35.2, pp. 128–135.

Jaeger, K-E, T Eggert, A Eipper, and M Reetz (2001). “Directed evolution and the creation of enantioselective biocatalysts”. In: *Applied microbiology and biotechnology* 55, pp. 519–530.

Jiang, Lin, 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. (2008). “De novo computational design of retro-aldol enzymes”. In: *science* 319.5868, pp. 1387–1391.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873, pp. 583–589.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and et al. Potapenko Anna (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873, pp. 583–589.

Kempen, Michel van, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L.M. Gilchrist, Johannes Söding, and Martin Steinegger (2023). "Fast and accurate protein structure search with Foldseek". In: *bioRxiv*.

Ketata, Mohamed Amine, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola (2023). "Diffdock-pp: Rigid protein-protein docking with diffusion models". In: *arXiv preprint arXiv:2304.03889*.

Kuhlman, Brian, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker (2003). "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy". In: *science* 1089427.1364, p. 302.

Laio, Alessandro and Michele Parrinello (2002). "Escaping free-energy minima". In: *Proceedings of the national academy of sciences* 99.20, pp. 12562–12566.

Marks, Debora S, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander (2011). "Protein 3D structure computed from evolutionary sequence variation". In: *PloS one* 6.12, e28766.

Martí-Renom, Marc A, Ashley C Stuart, András Fiser, Roberto Sánchez, Francisco Melo, and Andrej Šali (2000). "Comparative protein structure modeling of genes and genomes". In: *Annual review of biophysics and biomolecular structure* 29.1, pp. 291–325.

Onuchic, José Nelson and Peter G Wolynes (2004). "Theory of protein folding". In: *Current opinion in structural biology* 14.1, pp. 70–75.

Ovchinnikov, Sergey, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A Pavlopoulos, David E Kim, Hetunandan Kamisetty, Nikos C Kyrpides, and David Baker (2017). "Protein structure determination using metagenome sequence data". In: *Science* 355.6322, pp. 294–298.

Ponting, Chris P and Robert R Russell (2002). "The natural history of protein domains". In: *Annual review of biophysics and biomolecular structure* 31.1, pp. 45–71.

Procko, Erik, Geoffrey Y Berguig, Betty W Shen, Yifan Song, Shani Frayo, Anthony J Convertine, Daciana Margineantu, Garrett Booth, Bruno E Correia, Yuanhua Cheng, et al. (2014). "A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells". In: *Cell* 157.7, pp. 1644–1656.

Radivojac, Predrag, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. (2013). "A large-scale evaluation of computational protein function prediction". In: *Nature methods* 10.3, pp. 221–227.

Rao, Roshan, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song (2019). "Evaluating protein transfer learning with TAPE". In: *Advances in neural information processing systems* 32.

Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. (2021). "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". In: *Proceedings of the National Academy of Sciences* 118.15, e2016239118.

Rohl, Carol A., Charlie EM. Strauss, Kira MS. Misura, and David Baker (2004). "Protein structure prediction using Rosetta". In: *Methods in enzymology*. Vol. 383. Elsevier, pp. 66–93.

Šali, Andrej and Tom L Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints". In: *Journal of molecular biology* 234.3, pp. 779–815.

Sandhu, Jasbir Singh (1992). "Protein engineering of antibodies". In: *Critical reviews in biotechnology* 12.5-6, pp. 437–462.

Schulz, Georg E and R Heiner Schirmer (2013). *Principles of protein structure.* Springer Science & Business Media.

Senior, Andrew W, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. (2020). "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792, pp. 706–710.

Shirts, Michael and Vijay S Pande (2000). "Screen savers of the world unite!" In: *Science* 290.5498, pp. 1903–1904.

Siegel, Justin B, 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. (2010). "Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction". In: *Science* 329.5989, pp. 309–313.

Sigrist, Christian JA, Edouard De Castro, Lorenzo Cerutti, Béatrice A Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios (2012). "New and continuing developments at PROSITE". In: *Nucleic acids research* 41.D1, pp. D344–D347.

Söding, Johannes (2005). "Protein homology detection by HMM–HMM comparison". In: *Bioinformatics* 21.7, pp. 951–960.

Song, Yang, Conor Durkan, Iain Murray, and Stefano Ermon (2021). "Maximum Likelihood Training of Score-Based Diffusion Models". In: *Thirty-Fifth Conference on Neural Information Processing Systems.*

Tracewell, Cara A and Frances H Arnold (2009). "Directed enzyme evolution: climbing fitness peaks one amino acid at a time". In: *Current opinion in chemical biology* 13.1, pp. 3–9.

Trippe, Brian L., Jason Yim, D. Tischer, Tamara Broderick, D. Baker, R. Barzilay, and T. Jaakkola (2022). "Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem". In: *International Conference on Learning Representations.* DOI: 10.48550/arXiv.2206.04119.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. (2021). "Highly accurate protein structure prediction for the human proteome". In: *Nature* 596.7873, pp. 590–596.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30.

Voth, Gregory A (2008). *Coarse-graining of condensed phase and biomolecular systems.* CRC press.

Wang, Qian, Xiaonan Liu, Hejian Zhang, Huanyu Chu, Chao Shi, Lei Zhang, Jie Bai, Pi Liu, Jing Li, Xiaoxi Zhu, et al. (2024). "Cytochrome P450 Enzyme Design by Constraining Catalytic Pocket in Diffusion model". In: *Research.*

Watson, Joseph, David Juergens, Nathaniel Bennett, Brian Trippe, Jason Yim, Helen Eisenach, Woody Ahern, Andrew Borst, Robert Ragotte, Lukas Milles, et al. (2023). "De novo design of protein structure and function with RFdiffusion". In: *Nature*, pp. 1–3.

Whitford, David (2013). *Proteins: structure and function.* John Wiley & Sons.

Wu, Ke, KK Yang, R van den Berg, JY Zou, AX Lu, and AP Amini (2022). "Protein structure generation via folding diffusion". In: *Nature Communications.*

Yang, Jianyi, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker (2020). "Improved protein structure prediction using predicted interresidue orientations". In: *Proceedings of the National Academy of Sciences* 117.3, pp. 1496–1503.

Yi, Kai, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yuguang Wang (2024). "Graph denoising diffusion for inverse protein folding". In: *Advances in Neural Information Processing Systems* 36.

Yim, Jason, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé (2023). "Fast protein backbone generation with SE(3) flow matching". In: *arXiv preprint arXiv: 2310.05297*.

Yim, Jason, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola (2023). "SE(3) diffusion model with application to protein backbone generation". In: *Proc. Int. Conf. Machine Learning*.

Zhang, Yuan, Yang Chen, Chenran Wang, Chun-Chao Lo, Xiuwen Liu, Wei Wu, and Jinfeng Zhang (2020). "ProDCoNN: Protein design using a convolutional neural network". In: *Proteins: Structure, Function, and Bioinformatics* 88.7, pp. 819–829.

Zhao, Huimin and Frances H Arnold (1999). "Directed evolution converts subtilisin E into a functional equivalent of thermitase". In: *Protein engineering* 12.1, pp. 47–53.

Appendix A

Appendix

A.1 Metrics

A.1.1 Designability

Designability refers to the model's ability to generate protein structures that can be accurately reconstructed using inverse folding methods. The metrics used to assess designability include structural consensus TM-score (scTM) and structural RMSD (scRMSD).

- **scTM0.5** Measures the percentage of generated proteins with a structural similarity TM-score greater than 0.5.
  - **Calculation**

    $$\text{scTM} = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \tag{A.1}$$

    where d_i is the distance between the i-th pair of residues in the two structures, L is the length of the protein, and d_0 is a distance scaling parameter.
  - **Interpretation**: A higher scTM score indicates greater similarity between the two structures. scTM scores greater than 0.5 are typically considered indicative of significant structural similarity.
- **scRMSD2**: Measures the percentage of generated proteins with a root-mean-square deviation less than 2 Å.
  - **Calculation**:

    $$\text{scRMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i^A - \mathbf{r}_i^B)^2} \tag{A.2}$$

    where $\mathbf{r}_i^A$ and $\mathbf{r}_i^B$ are the positions of the i-th atom in the two superimposed structures, and N is the number of atoms.

    - **Interpretation**: Lower scRMSD values indicate greater similarity between the structures. scRMSD values less than 2 Å are generally considered to represent accurate structural predictions.

A.1.2 Diversity

Diversity assesses the variety of generated protein structures. This is evaluated by measuring the clustering center density of generated samples using hierarchical clustering methods.

- **Diversity**: Calculated as the number of clustering centers divided by the number of generated samples.
    - **Interpretation**: Higher diversity values indicate a wider variety of generated structures. It reflects the model's ability to explore different regions of the conformational space (Herbert and Sternberg 2008).
- **pdbTM0.7**: Measures the structural similarity of generated proteins to the most similar structures in the Protein Data Bank (PDB)(Kempen, S. S. Kim, Tumescheit, Mirdita, J. Lee, Gilchrist, Söding, and Steinegger 2023).
    - **Interpretation**: Higher pdbTM values indicate that the generated structures closely resemble known protein structures. It is used to assess the novelty of the generated proteins. Values below a certain threshold (e.g., pdbTM0.7) indicate novel structures not found in the PDB.