Universiteit Leiden

Master Computer Science

Sampling protein sequence space to map global distributions of protein structure.

Name:	Lilli Schuckert
Student ID:	s3658937
Date:	June 30, 2024
Specialisation:	Bioinformatics
1st supervisor:	Dr. E. A. Schultes
2nd supervisor:	Dr. K.J. Wolstencroft

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Proteins are fundamental to cellular function, driven by their unique three-dimensional structures determined through the folding of their linear amino acid sequences. While mutations can significantly impact protein functionality, the Neutral Theory of Molecular Evolution posits that many sequence changes are neutral. This thesis explores the vast protein sequence space by examining the structural consequences along directed mutational trajectories in the SARS-CoV-2 spike protein, central to the virus's infectivity and immune evasion. Utilizing protein structure prediction algorithms like ESMFold, ColabFold, and AlphaFold2, we systematically analyze how amino acid substitutions affect protein structure, particularly focusing on the ACE2 receptor binding domain. By generating a systematic map of structural variations across diverse sequences, this work elucidates trends that could help to illuminate principles governing protein folding and the structural trends emerging from sequence variations. The findings strongly suggest that different regions of sequence are associated with characteristic structural properties as reflected in compactness and residue-residue interactions, and that there are specific regions of sequence space where these properties may undergo sharp transitions. The concept of analyzing mutational trajectories can be enhanced to explore further evolutionary pathways, predict potential viral adaptations, and inform the design of more effective therapeutics and vaccines against SARS-CoV-2 and other emerging pathogens. Future research should expand on these insights to explore intermediate states and apply techniques of evolutionary programming to further unravel protein folding complexities.

Ackowledgements

I want to express my sincere gratitude to some amazing individuals who have supported me throughout my master's program and the completion of this thesis. First and foremost, I thank Dr. Erik Schultes and Dr. Katy Wolstencroft for their supervision and guidance throughout my thesis and the introductory bioinformatics research project. Your expertise, insightful feedback, and encouragement have been instrumental in shaping this thesis.

Thank you, Erik, for your daily mentorship and guidance. I have always felt your appreciation for my work on the project, and I hope we continue to collaborate on more projects in the future. I am grateful for everything you taught me about research this past year and for giving me the confidence to follow up on this master's with a PhD project. Your immense knowledge and expertise are truly inspiring, and I am honored to have had the opportunity to learn from you and work alongside you on this project.

I am also thankful to the individuals I have had the pleasure of working with over the last year. I especially want to thank Max van den Boom for the close collaboration on this project, for all your help, and for being a great unofficial IT support. Thank you to Thomas Hankemeier, Michael van Vliet, Amy Harms, Guus Ruiter, and Alida Kindt from the LACDR; thank you to the project partners from Utrecht, including Berend Jan Bosch, Suki Lee, Daniel Hurdiss, and Bob Coelingh; thank you to Bryan Cardenas Guevara from SURF for all your help and expertise on structure prediction algorithms and their installation. I further want to thank the SURF service desk team for their fast and reliable help. Thank you to Ander Astudillo, Kees Burger, Tobias Kuhn, Barbara Magagna, Barend Mons, and Jacintha Schultes, and to all those who have played a role in this project, whether through direct collaboration, thoughtful discussions, or critical feedback.

Next and most importantly, I want to thank my family for their unconditional love and support and my dad for proofreading part of this thesis. Without your constant encouragement, love, and support, I would not have been able to follow this master's program abroad. I am further thankful for my babysitting families here in Leiden, who have made me feel more at home. I also thank my roommates, friends, and fellow students for their emotional support throughout the long coworking and library days and for all the fun moments outside of university. Thank you, Elisa, for always helping and supporting me - even from far away. And lastly, thank you, Inés. Words won't justify my gratitude for you, your support, and our friendship.

Declaration of originality

I declare to the Universiteit Leiden that I have written this master thesis independently and without using any sources or aids other than those indicated. This thesis is free of plagiarism. All statements taken verbatim or in content from other writings have been marked as such. I understand that any false claim with respect to this thesis will result in disciplinary action in accordance with university and program regulations and that any false claim will be reported to the Board of Examiners. This thesis has not been submitted in the same or similar form to any other university.

Verklaring van originaliteit

Ik verklaar aan de Universiteit Leiden dat ik deze masterscriptie zelfstandig heb geschreven, zonder gebruik te maken van andere bronnen of hulpmiddelen dan aangegeven. Deze scriptie is vrij van plagiaat. Alle uitspraken die woordelijk of inhoudelijk uit andere geschriften zijn overgenomen, zijn als zodanig gemarkeerd. Ik begrijp dat elke valse bewering met betrekking tot deze scriptie zal resulteren in disciplinaire maatregelen in overeenstemming met de universitaire en opleidingsregelingen en dat elke valse bewering zal worden gerapporteerd aan de examencommissie. Deze scriptie is niet in dezelfde of vergelijkbare vorm ingediend bij een andere universiteit.

June 30, 2024

Lilli Schuckert

Contents

1 Introduction **1**
1.1 Research Objectives . . . 3

2 Related Work **7**
2.1 Protein Structure and Function . . . 7
2.2 Protein Sequence Space . . . 9
2.3 Structure Prediction Algorithms . . . 12
2.4 State-of-the-Art SARS-CoV-2 Protein Structure Analysis . . . 14

3 Methodology **17**
3.1 Workflow . . . 17
3.1.1 ACE2 Receptor Binding Domain . . . 19
3.2 Structure Prediction Algorithms . . . 22
3.2.1 ESMFold . . . 23
3.2.2 AlphaFold2 . . . 27
3.2.3 ColabFold . . . 32
3.3 Structural Analysis Metrics . . . 35
3.4 Permutated Sequences . . . 38
3.5 Contact Matrices . . . 39
3.6 Genetic Constraint on ACE2 RBD . . . 40
3.7 Computational Infrastructure . . . 41
3.7.1 SURF . . . 41
3.7.2 Analysis Tools . . . 43
3.8 FAIR Data Management . . . 43
3.8.1 Research Data Lifecycle . . . 44
3.8.2 Creation of Data and Metadata . . . 44
3.8.3 Publication of Research Outputs . . . 49

4 Results **51**
4.1 Structure Prediction Algorithms . . . 51
4.1.1 Sequence Space Trajectories . . . 52
4.1.2 Correlation Analysis . . . 64
4.1.3 Sequential Changes . . . 78
4.1.4 Contact Matrices . . . 87

4.2 Spoke Wheel of Sequence Space 90
4.3 Permutated Sequences 92
4.4 Genetic Constraint on ACE2 RBD 94
4.4.1 Genetic Trajectories 97
4.5 FAIR Digital Twin 102

5 Discussion 107
5.1 Computational Mapping of Sequence Space 107
5.2 Structural Trends Across the Trajectories 108
5.2.1 Phase Transitions 111
5.2.2 Genetic Constraints 114
5.2.3 Structural Predictions in the Context of Protein Folding Principles . . . 115
5.3 Performance of Structure Prediction Algorithms 116
5.4 Ethical Concerns 119
5.5 Future Directions 120

6 Conclusion 125

A Trajectory Plots 135

B Contact Matrix Plots 157

Chapter 1

Introduction

Proteins are essential to life. They play crucial roles within cells, providing physical structure for the cell, transporting molecules, facilitating physiological signals in the organism, and controlling the flux of metabolic pathways [81]. The versatility and precision of protein function are derived from their varied and complex three-dimensional structures. The unique protein structure, is determined by its linear configuration (sequence) of 20 amino acids through a physico-chemical 'folding' process. A protein structure can be described at four levels: primary structure, which is the amino acid sequence; secondary structure, which represents local folding patterns; tertiary structure, which is the overall 3D shape driven in many cases by non-local interactions; and quarternary structure, which is a non-covalent assembly of multiple polypeptide chains. Changes in the structure, even as little as one amino acid substitution, can sometimes significantly impact a protein's function. An example of the large impact small changes in a protein's structure can have is the sickle cell anemia disorder, which is caused by a single amino acid substitution in the hemoglobin protein. The minor structural change of glutamic acid to valine in the beta chain leads to distortion of the typical shape of red blood cells. This causes them to become sickle-shaped, which can lead to various complications associated with the disease [43]. Thus, even minor changes to the amino acid sequence can significantly impact the tertiary structure and biological activity. On the other hand, the Neutral Theory of Molecular Evolution, first proposed by Motoo Kimura in 1983 [33], suggests that many mutations at the molecular level are selectively neutral. That is, these mutations

do not significantly alter the structure or function of proteins. Kimura's theory suggests that such neutral changes can accumulate over successive generations, ultimately contributing to the shaping of protein sequences within the vast sequence space, which is further discussed in [60]. Proteins spontaneously fold into their tertiary structure due to interactions between the amino acid side chains. Not only genetic mutations can cause misfolding but also the cellular environment and insufficient chaperone activity [20]. Understanding protein structures and their folding mechanisms has played a crucial role in the advancements of drug design, protein engineering, and the development of new biomaterials. This understanding has not only been driven by experimental techniques but also by computational efforts.

Despite great advancements in studying proteins, significant challenges remain in understanding their complex structures, functions, and interactions. High-throughput sequencing and computational modeling have enabled researchers to explore a more substantial portion of protein sequence space. However, gaps regarding our knowledge of the functional implications of protein sequence variations have yet to be filled. Data integration, combining information from real-world observations, laboratory experiments, and computational predictions, will help us to gain a more far-reaching understanding of proteins. The COVID-19 pandemic has underlined the importance of studying and understanding viral proteins and their functions. Since its emergence in late 2019, the novel SARS-CoV-2 coronavirus has led to considerable morbidity and mortality worldwide. Key structural and non-structural proteins of SARS-CoV-2 play essential roles in its replication, host cell entry, and evasion of the immune system [84]. Among these proteins, the spike protein (S) is critical for mediating the attachment and fusion of the virus to host cells. Mutations in the spike protein, such as the emergence of variants like Alpha (B.1.1.7), Beta (B.1.351), or Omicron (B.1.1.529), can significantly alter the virus's immune evasion and its transmissibility [11]. Thus, understanding the sequence space and structural changes in SARS-CoV-2 proteins is necessary for predicting the evolution of the virus and developing effective countermeasures. Computational approaches, such as protein structure prediction, can complement experimental studies and help researchers identify potential drug targets, develop vaccines and predict the impact of viral mutations, ultimately benefiting public

health preparedness and response.

1.1 Research Objectives

The diversity of protein sequences that are found in nature is very large, both within an organism and between species. Hence, in practice, proteins are often investigated in isolation, with a focus on their individual structural and functional properties. In this work we shift the focus to the distribution of properties across large numbers of sequences related to each other by continuous variation in their amino acid composition. Hereby, amino acid composition becomes a tunable parameter that is utilized to scan the space of sequence and structure properties that could be used along with other information to infer biological function.

In light of the significant challenges posed by the COVID-19 pandemic, understanding the characteristics of SARS-CoV-2 has become crucial in the global effort against pandemic threads. This work, which is part of the Health-Holland funded STAYAHEAD project at the Leiden Academic Center for Drug Research (LACDR), investigates structural properties of the SARS-CoV-2 virus's spike protein and its variants with the ultimate aim of translating this knowledge in the identification of potential high-risk variants. The project focuses on the molecular structure of the SARS-CoV-2 spike protein [14], because it is crucial for the virus to enter host cells, its structure and variants are of great importance in our understanding of variant emergence and propagation. By delving further into the structural landscape of the spike protein and its variants the project will provide a better understanding of the spike protein's behavior, adaptability, and virulence.

Our strategy in exploring sequence space is as follows: We begin with the known sequence and structure of the ACE2 receptor binding domain (ACE2 RBD) of the SARS-CoV-2 Spike protein (195 amino acids) [77], which plays a key role in the receptor recognition and cell membrane fusion process. We created collections of related sequences by substituting amino acids of the Spike protein at randomly chosen positions for amino acids of the same type.

The amino acid substitutions were made randomly and cumulatively, one by one, until the sequence of the ACE2 epitope converged onto the homopolymer of that amino acid type. This systematic sampling approach was intended to reveal the physical properties of protein folding regardless of any peculiarities of the genetic code. This process was repeated for each of the 20 amino acid types. Throughout this work, we refer to these collections of sequences as directed mutational 'trajectories' in the sequence space.

One of the key challenges in protein structure prediction is the vastness and complexity of the protein sequence space. By utilizing directed mutational trajectories, we can systematically explore this space in a structured and manageable way. Trajectories allow for a comprehensive examination of how different amino acid substitutions affect protein structure and function, revealing general trends and patterns in protein folding in protein sequence space. This method facilitates a comprehensive analysis of the sequence-structure relationship by tracking how gradual modifications in the amino acid composition influence the folding and stability of the protein. Additionally, using trajectories helps to highlight general principles of protein folding and evolution, revealing patterns and trends that might be missed when examining proteins in isolation. By incrementally substituting amino acids and sampling sequences that span a wide range of the sequence space, this approach provides a detailed and informative sample of the entire sequence space. Eventually, this will help to predict the functional impacts of mutations, which is crucial for identifying potentially harmful variants, such as those of the SARS-CoV-2 virus. This collection of sequences was submitted to structure prediction algorithms, namely ESMFold, ColabFold, and AlphaFold2, providing a number of predicted structural features for each sequence. The aim was to reveal general trends in protein structure that span these trajectories. By combining analysis results from the 20 trajectories, a coarse-grained map of the global structure in protein sequence space is created, which sheds light on the boundaries for diverse and complex biological functions.

In this thesis, we sample protein sequence space, aiming to shed light on global distributions of protein structure. The following chapters provide an overview of this field's current state

of knowledge. Moreover, I elaborate on the protein structure prediction algorithms used in this field, namely ESMFold, AlphaFold2 and ColabFold. Finally, we contextualize our findings within the existing field of knowledge.

Chapter 2

Related Work

This chapter will provide an overview of the existing work in the sequence space domain, highlighting the importance of proteins within organisms and elaborating on the concept of protein sequence space. Furthermore, I will dive into state-of-the-art protein structure prediction algorithms. This Chapter will establish the context and significance of the research presented in this thesis, which focuses on mapping the sequence space of the SARS-CoV-2 virus.

2.1 Protein Structure and Function

Proteins are made up of long chains of amino acids, with the specific sequence of amino acids determining the protein's three-dimensional structure [2]. There are 20 different amino acids relevant to biological systems, each having different chemical properties. Amino acids can be grouped into different categories based on the properties of their side chain groups, as seen in Table 2.1. These properties can shed light onto the folding of the protein. Amino acids with nonpolar side chains are hydrophobic for example and tend to cluster more in the interior of folded proteins. In fact, the hydrophobic collapse of these nonpolar residues might be the primary driving force behind protein folding, as the burial of hydrophobic groups in the protein core minimizes unfavorable interactions with the aqueous solvent [16, 17, 22].

The complex relationship between protein shape and function is fundamental to our under-

standing of biological processes at the molecular level in the field of protein biology. Various interactions between amino acid residues drive protein folding, including interactions between hydrogen bonds, ionic interactions, and van der Waals forces. The three-dimensional structures of many biologial proteins are precisely folded into specific configurations that in turn give them their function. The amino acid sequence guides the folding process to attain the most thermodynamically stable and functional conformation. The number of amino acid residues in a protein determines its length, which can vary substantially. Some proteins are quite small, having only a few dozen residues, but others might be thousands of residues long. Longer proteins typically consist of several separate structural and functional modules known as domains. These domains are semi-independent sections that can fold and operate independently and are divided by flexible linker or interdomain regions. The SARS-CoV-2 spike protein, which is 1273 amino acids long consists of two major domains: the S1 domain which is responsible for receptor binding and the S2 domain which is involved in membrane fusion [65]. These domains can then be further classified into subdomains.

Table 2.1: Amino Acid Groups. Classification of amino acids into different groups based on their side chains. The "one-letter" amino-acid code is presented in parentheses.

Groups	Amino Acids
Nonpolar (Hydrophobic)	Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I), Proline (P), Phenylalanine (F), Tryptophan (W), Methionine (M), Cysteine (C)
Polar, Uncharged	Serine (S), Threonine (T), Tyrosine (Y), Asparagine (N), Glutamine (Q)
Acidic	Aspartic acid (D), Glutamic acid (E)
Basic	Lysine (K), Arginine (R), Histidine (H)

Studying the vast landscape of protein structures is essential to solving the puzzles surrounding proteins' diverse functions within living organisms. Understanding proteins' complex folds and spatial configurations helps scientists better understand the principles behind biological processes such as signal transduction, enzymatic catalysis, and molecular recognition. Additionally, elucidating the structure-function relationships holds immense potential for use in biomedical applications, such as protein engineering, drug discovery, and the development of

therapeutics. In light of the COVID-19 pandemic, understanding the tertiary structure of viral proteins is essential to identify epitopes that the immune system targets. This knowledge will then support the development of vaccines specifically targeting these epitopes, effectively training the immune system to recognize the virus upon exposure.

2.2 Protein Sequence Space

Protein sequence space is a theoretical space encompassing all possible sequences of a given length. Scientists have mainly derived information about nucleic acids and proteins by extracting information from naturally existing sequences, which reflect millions of years of evolutionary processes. However, protein sequence space also encompasses sequences that have not yet been instantiated by evolution or synthesized in the laboratory. Newly synthesized sequences of proteins could provide novel insights into relationships between sequence composition and their functional properties. At first, it might seem that exploring a huge amount of arbitrary sequences is too complex to yield scientifically meaningful insights. However, each sequence in sequence space is separated from the others by only a linear number (N) of mutational steps. N hereby refers to the length of the original sequence. This property, where there are an exponential number of sequences, but each separated by no more than a linear number (N) mutational steps, opens the door to a number of tractable and biologically insightful analyses [60].

The concept of sequence space has been shaped over the years by fundamental discoveries and technological advancements. More than 70 years ago, in 1953, the discovery of the double helix structure by Watson and Crick marked the beginning of an era where understanding molecular sequences became central to understanding the genetic basis of life [80]. In 1969, Frank Salisbury proposed that biological complexity arises from the interplay between natural selection and chance, challenging the notion of evolution solely through natural selection. In 1970, Maynard Smith contradicted this idea, suggesting that natural selection can navigate protein space without random exploration but rather using incremental steps akin to transitioning between meaningful words. He uses the example of moving from the word 'WORD'

to the word 'GENE' in 4 steps ($WORD \rightarrow WORE \rightarrow GORE \rightarrow GONE \rightarrow GENE$), which is an analogy to the idea of moving between meaningful sequences in the context of genetic evolution. In this context, 'meaningful' refers to sequences that are functional or have a specific purpose.

Both these arguments paved the way for a more comprehensive understanding of the relationship between natural selection and the complexity of biological molecules. Another significant milestone that undoubtedly advanced the field even further was the completion of the Human Genome Project in 2003. Not only did it lay the foundation for large-scale genomic studies but it also led to a more thorough understanding of genetics. Only six years later in 2009, the concept of sequonomics was formally introduced [60], which entails the study of sequence space as a whole.

Mathematical investigation of sequence space provides an interesting mathematical viewpoint. Let N be the length of the sequence and A be the size of the alphabet. Consider the 20 standard amino acids. By employing basic combinatorial calculations, we can determine the abundance of sequences within the entire sequence space. Additionally, these calculations allow us to observe how altering the size of the alphabet or the length of a sequence can alter the abundance of sequences. The theoretical study of sequence space provides a crucial investigation into the composition and arrangement of sequence space. It provides a mathematical framework and scaling principles that allow one to understand the sequence space's combinatorial richness and clarify how differences in alphabet size and sequence length affect the distribution of biological variety.

Table 2.2 shows some simple calculations on the size of sequence space. Let $A = 20$ be the number of the standard amino acids. Consider a protein of length $N = 100$. This results in a total number of possible sequences equal to $A^N = 20^{100}$ and a dimensionality of $(A-1) \cdot N = 1900$. A study from 2008 suggests that it is not unlikely that all functional protein sequence

Table 2.2: General Scaling Laws for Sequence Space

Description	**Formula**
Total number of possible sequences	A^N
Dimensionality of sequence space	$D = (A-1) \cdot N$
Number of sequences k-steps away from reference sequence	$\frac{N! \cdot (A-1)^k}{(N-k)! \cdot k!}$

space has already been explored. Furthermore, the study proposed that due to structural and functional constraints, the capacity for novel protein structures to evolve is limited [18]. The degree to which protein sequence space has been investigated affects both the possibility of finding new protein functions and our comprehension of molecular evolution. Looking at a protein of length 100, we can calculate that there are 20^{100} possible sequences in this sequences space, which is an astronomically large number. To demonstrate this, we can compare the number of possible sequences to the number of atoms estimated in the known universe (about 10^{80}). Comparing a protein of length 100 to the number of atoms estimated in the known universe we can calculate $ln(10^{80}) \approx 184$ and $ln(20^{100}) \approx 3000$. When $ln(x) = a$, then $x = e^a$. We can then calculate that the number of possible sequences is $e^{ln(y)-ln(x)} \approx e^{116}$ ($\widehat{=}2.3 \cdot 10^{50}$) greater than the number of atoms in the known universe. These numbers increase exponentially with proteins that have more residues. For a protein of length 195, which is the length of the protein we investigate in this thesis, we know that the amount of sequence possibilities is $\approx e^{400}$ ($\widehat{=}5 \cdot 10^{173}$) greater than the number of atoms in the known universe. Despite the suggestion that the overall functional sequence space of this protein may be limited, the evolution of specific sequences, such as in the context of rapidly evolving viruses like SARS-CoV-2, can still lead to novel variants with distinct properties. Exploring the SARS-CoV-2 spike protein's sequence space assist in clarifying the evolutionary mechanisms underlying the formation of novel strains and any possible implications for virulence, transmission, and vaccination effectiveness. By exploring the vast sequence space of the SARS-CoV-2 spike protein, researchers can get better insights into structural properties, such as the stability of conformational changes. These insights are essential for designing therapeutics vaccines targeting SARS-CoV-2.

2.3 Structure Prediction Algorithms

In this thesis we employ protein structure prediction algorithms to explore protein folding across the sequence space of SARS-CoV-2. The development of protein structure prediction algorithms has been a long-standing challenge in computational biology and can be traced back to years of discoveries. The 1950s marked an era when pioneering research on simple machine learning algorithms began to emerge, such as Arthur Samuel's work on a checkers-playing program. However, it was not until the 1980s that the rediscovery of backpropagation led to a renewed interest in machine learning. The rise of recurrent neural networks, support vector machines, and kernel methods marked the 1990s and 2000s. During the late 2000s and early 2010s, several protein structure prediction algorithms paved the way for advancements in the field. These include Rosetta, developed by the Baker laboratory or I-TASSER (Iterative Threading ASSEmbly Refinement), developed by the Zhang laboratory. The development of AlphaFold in 2018 and later AlphaFold2 in 2020 enabled significant breakthroughs in protein structure prediction. Recent significant progress can be traced back to the enhancement of deep learning architectures. These advancements have shown outstanding success in capturing long-range dependencies which ultimately leads to improved predictions of protein structure.

Recent studies have used experimental protein evolution approaches and computational modeling to explore the limits of protein sequence space [13, 9, 69]. Significant progress has been made in developing computational methods for predicting the tertiary structure of proteins from their amino acid sequence [49, 34, 87, 78]. We can classify these prediction algorithms into two main categories: Template-Based Modeling (TBM) methods, which use known protein structures as templates to model the target protein, and Free Modeling (FM) methods, which attempt to predict the structure without using templates [34].

A benchmark to compare different protein structure prediction algorithms is the biennial event 'Critical Assessment of Structure Prediction' (CASP) competition, which evaluates the performance of different protein structure prediction algorithms [51]. The competition differentiates

between TMB methods and FM methods, as well as a hybrid FM/TBM category, which results in a comprehensive assessment of the state-of-the-art field. The research groups that participate predict structures using all structure data available at present to predict newly solved structures that have yet to be released. In 2020, during the 14th edition of CASP, DeepMind's AlphaFold2 achieved a breakthrough in predicting single-chain protein structures, scoring around 90 on the 100-point GDT (Global Distance Test) scale [12]. GDT quantitatively measures how accurately a predicted protein structure aligns with the experimentally determined structure. Scores above 90 are considered to be roughly equivalent to the experimentally determined structure.

Open-access protein databases such as the Protein Data Bank (PDB) or UniProt have been valuable in advancing protein structure prediction algorithms. They provide experimental data and annotated sequences which aids in the development and benchmarking of prediction algorithms. Projects such as OpenFold promote the sharing of data, algorithms and computational resources which will inevitably accelerate scientific progress. OpenFold is a non-profit AI research consortium which aims at developing free and open-source software tools for biology and drug discovery. This also includes accelerating robust and scalable open-source tools for protein structure prediction [1].

Even though there have been great advancements in recent years, challenges remain. Predicting the structure of protein complexes and multi-domain proteins still remain challenging [34, 6]. Researchers are also trying to integrate experimental data with computational methods to further enhance prediction accuracy. Moreover, expanding structural protein databases will further advance structure prediction algorithms. Most structure prediction algorithms predict a single static conformation of a protein but with the integration of physics-based models will allow the prediction of the dynamic nature of proteins, such as multiple conformational states [50]. Moreover, the use of AI in biology has raised ethical concerns regarding data privacy and potential misuse. These concerns can be addressed through the use of a robust governance framework and ethical guidelines [27]. Structure prediction algorithms will continue

to advance, not only due to the rapid growth of protein databases but also due to increases in computing power and advancements in machine learning [34]. This will make them more accessible and affordable to a broader spectrum of users. We will further elaborate on three structure algorithms, namely ESMFold, AlphaFold2, and ColabFold in this thesis, highlighting advantages but also possible limitations.

2.4 State-of-the-Art SARS-CoV-2 Protein Structure Analysis

Several significant efforts have been made to predict and determine the three-dimensional structures of SARS-CoV-2, both computationally and experimentally. In 2020 AlphaFold was used to generate computational predictions of understudied proteins. They acknowledged that these proteins, while not being the main focus of treatments, might help researchers understand how SARS-CoV-2 works [30]. Furthermore, Beshnova *et al.* [7] have made efforts to study the antibody-spike protein binding. Their approach has been validated using existing structures in the Protein Data Bank (PDB). In another large-scale study by Gutnik *et al.* [25] three-dimensional structures for all SARS-CoV-2 proteins were predicted, including those with no experimental structures available. Further research on computationally predicting the three-dimensional structures was done by Ali and Caetano-Anollés [3].

Not only computational efforts have been made over the last years but also experimental efforts. The Jesse Bloom Lab studied the antibody response to various SARS-CoV-2 variants and evolutionary aspects of the virus. They further developed methods to estimate the impact of mutations on the SARS-CoV-2 receptor binding domain [10]. Furthermore, Berend Jan Bosch and colleagues in Utrecht have conducted detailed structural analyses of the SARS-CoV-2 spike protein, focusing on its interactions with host cell receptors such as ACE2 [74, 79].

These structural investigations have made substantial contributions to our understanding of the

function of various SARS-CoV-2 proteins, particularly those involved in viral entry, replication, and immune evasion. The rapid response of the scientific community has evolved from the early pandemic to later stages, showcasing the dynamic nature of structural biology efforts. However, several challenges have been faced in these studies. The virus's fast mutation has made it difficult to keep up with structural changes, and specific proteins have proven especially difficult to research due to their inherent characteristics. Despite these challenges, the integration of computational and experimental methods has provided a more comprehensive understanding of SARS-CoV-2 proteins, informing the development of vaccines, antibody therapies, and small molecule inhibitors.

Chapter 3

Methodology

This chapter provides a comprehensive overview of the methodologies employed throughout the thesis, including the research design and data collection procedures. I will elaborate on the three structure prediction algorithms used in this thesis and present structural metrics that can be derived from their outputs. Additionally, an analysis of missense mutations at the DNA level within the ACE2 RBD will be conducted. This Chapter will also cover the computational infrastructure used in the research. Finally, I will discuss FAIR Data Stewardship and our approach to data storage within this project.

3.1 Workflow

The workflow of this thesis can be seen in Figure 3.1. We begin with the amino acid sequence of the ACE2 binding domain of the SARS-CoV-2 Spike protein, which is 195 amino acids long. To be exact we are using the Wuhan spike sequence residues 333 to 527, extracted from UniProt (UniProt Identifier: P0DTC2) [14]. We created potential variants by substituting amino acids of the sequence at randomly chosen positions for amino acids towards a specific homopolymer. Because we consider the standard 20 amino acids, this will result in variants of 20 different homopolymers. This can be seen in Algorithm 1. In Algorithm 1 we iterate over each amino acid of the 20 standard amino acids and create trajectories by randomly mutating a single amino acid of the input sequence until eventually the homopolymer of that amino acid

is reached. The resulting sequences are formatted as SeqRecords and written to an output file in FASTA format. In total, when considering the sequence length of 195 amino acids, this will result in $195 \cdot (A - 1) = 3705$ possible variants on the trajectories, where A is the number of amino acids, i.e. 20. We further generate variants on the trajectories that are possible when considering the genetic code. Additionally, we create 10 random permutated sequences that have the same amino acid distribution as the ACE2 RBD. These experiments are further described in Sections 3.6 and 3.4, respectively. Both these datatsets will offer additional insights into the distributions of protein structure withing the protein sequence space. An overview can be seen in Table 3.1.

Table 3.1: Overview of Datasets. An overview of the three different datasets and the number of sequences in each dataset.

Dataset	**Description**
Directed Mutational Trajectories	3705 sequences on the 20 mutational trajectories
Genetic trajectories	1135 sequences on the trajectories that are genetically possible
Permutated sequences	10 randomly permutated sequences that have the same amino acid distribution as the ACE2 RBD

Next, these sequences are submitted to a variety of structure prediction algorithms. In this thesis, we employed ESMFold, ALphaFold2, and ColabFold. These algorithms output different structural information about the predicted protein structure, which will be stored on ResearchDrive to give other researchers access to the data upon request. The output is then further analyzed regarding different structural metrics that can be extracted from the output of the structure prediction algorithms. The structural features are analyzed for individual sequences. An overview of the analyses can be seen in Table 3.2.

Taken together, these data allow us to identify general trends in protein structure that span the amino acid trajectories. By mapping the structural features across the explored protein sequence space, we can gain insights into the structural diversity and complexity of proteins. This comprehensive map of protein structure in sequence space sheds light on the boundaries of the complexity of biological functions that can be achieved through protein folding and

Algorithm 1: Variant Sequence Generator

```
Input: ACE2 binding domain
Output: Trajectories for each amino acid
for each amino_acid in amino_acids do
    trajectories_for_amino_acid ← [];
    input_sequence_copy ← copy of ACE2 binding domain;
    sequences_in_trajectory ← [];
    while there exists an amino acid in input_sequence_copy that is not equal to
      amino_acid do
        position_to_mutate ← random integer between 0 and
          length(input_sequence_copy) - 1;
        if amino acid at position_to_mutate in input_sequence_copy is not equal to
          amino_acid then
            replace amino acid at position_to_mutate in input_sequence_copy with
              amino_acid;
            add input_sequence_copy to sequences_in_trajectory;
        end
    end
    add sequences_in_trajectory to trajectories_for_amino_acid;
end
seq_records ← [];
for each sequence in trajectories_for_amino_acid do
    create a SeqRecord with sequence and a unique identifier;
    add the SeqRecord to seq_records;
end
write seq_records to output_filename in FASTA format;
```

structure. To be precise this map can highlight structural features that are most commonly observed and those that are rare or absent. Moreover, it can identify regions of sequence space, where structural diversity is high or limited, implying constraints on function.

3.1.1 ACE2 Receptor Binding Domain

Within the spike protein, there are specific regions known as epitopes that are recognized by the immune system. A schematuc overview of the ACE2 RBD can be seen in Figure 3.2. In this thesis we focus on the ACE2 receptor binding domain (ACE2 RBD) of the spike protein. The ACE2 RBD comprises the epitopes that directly interact with the ACE2 (Angiotensin-Converting Enzyme 2)receptor. The ACE2 receptor plays a key role in the entry of the SARS-CoV-2 virus into host cells [85]. Thus, it is a key functional domain and researchers have been

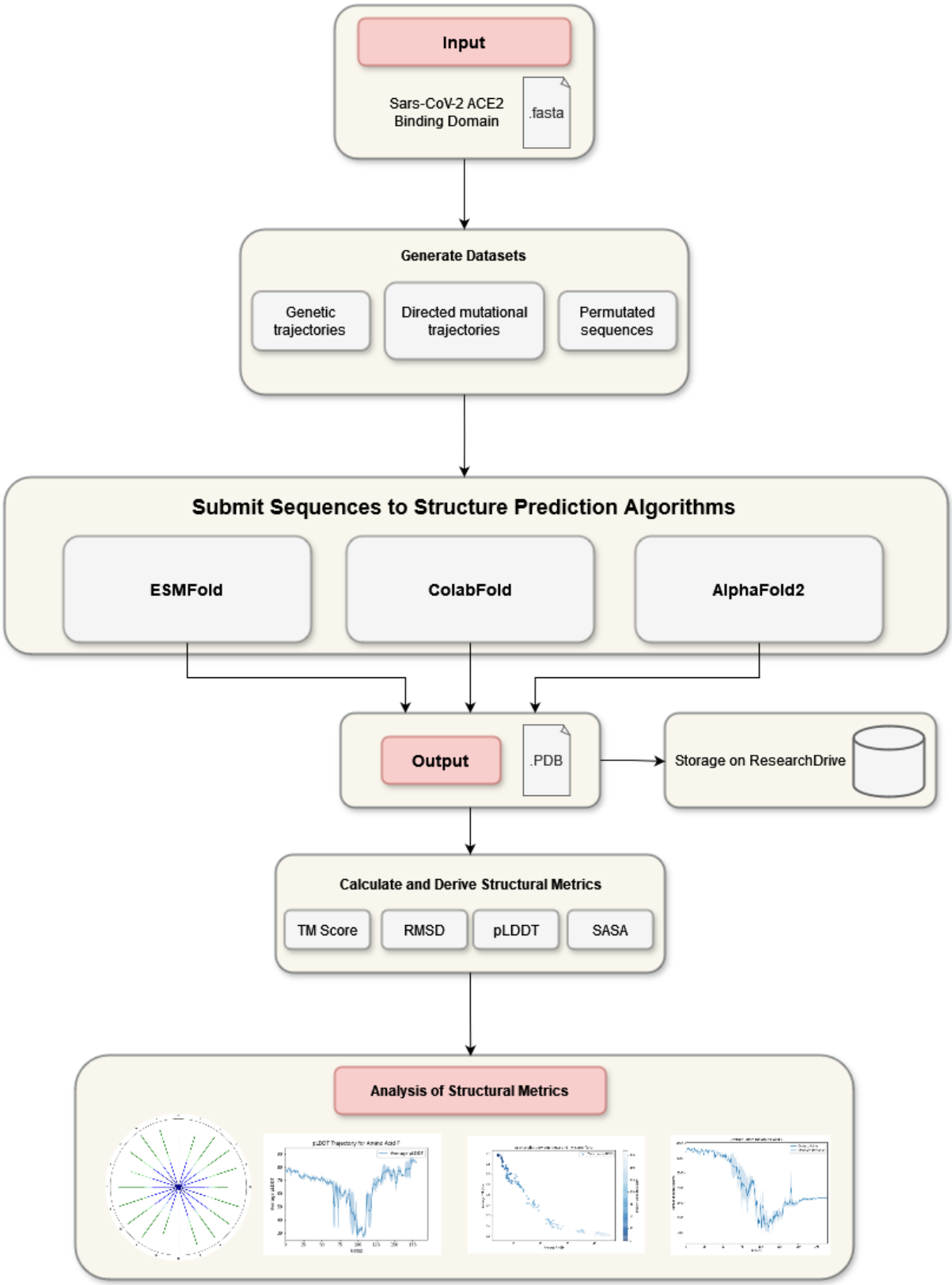


Figure 3.1: Protein Structure Prediction Workflow. We begin with the amino acid sequence of the Sars-Cov-2 ACE2 Binding Domain and the generated variants are then submitted to structure prediciton algorithms whose output is stored on ResearchDrive and then further analyzed regarding different structural metrics.

Table 3.2: Type of Analyses. Overview of the type of analyses we can run using the structural metrics and what insights we can gain from them.

Type of Analysis	Interpretation
Metrics along physical trajectories	Gives insights into protein stability and conformational changes. Additionally, this reveals how structure prediction algorithms perform across the trajectories.
Metrics along genetic trajectories	Gives insights into protein stability and conformational changes when imposing genetic constraints.
Difference between genetic and physical trajectories	Provides insights into structural differences between genetic and physical sequences.
Correlations between metrics along trajectories	Correlations can identify consistent patterns across different trajectories, suggesting common structural or functional themes. Ensuring that metrics correlate well across trajectories acts as a quality control measure, validating the accuracy and reliability of the predictions.
Delta of metrics along trajectories	Identify the rate of change in structural properties, highlighting critical points of transition.
Contact matrix along trajectories	Identify how the overall structure is maintained or disrupted along the trajectories. This can highlight transition points in the protein's three-dimensional structure.

studying it to gain better understanding of the molecular mechanisms of viral entry. This will aid them in vaccine development and therapeutic strategies that are aimed at blocking the viral entry into host cells. Both experimental efforts and computational approaches have been used to characterize these epitopes. The Jesse Bloom Lab [36] has performed experimental efforts on the affinity of mutated proteins on ACE2. The intersection of experimental data and computational approaches will shed more details on the molecular mechanisms.

Given the significance of the ACE2 RBD in viral entry, it is crucial to delve deeper into its properties and behavior. Additionally, focusing on this smaller epitope of the entire spike protein

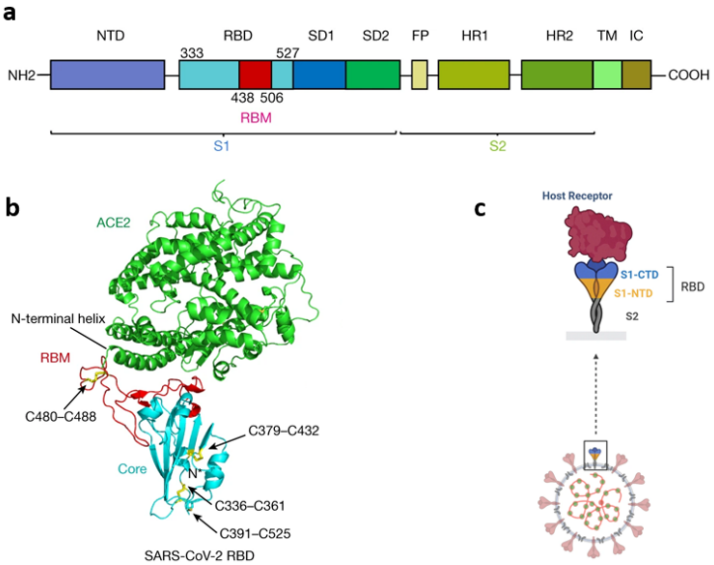


Figure 3.2: Schema of SARS-CoV-2 spike protein and the ACE2 RBD.(a) An overview of the topology of the SARS-CoV-2 spike protein. Adapted from [37]. **(b)** Structure of the ACE2 RBD, with the ACE2 domain shown in green, and the RBD shown in blue. Adapted from [37]. **(c)** A schema of the binding mechanism of the ACE2 RBD. Adapted from [55].

has some advantages. Analyzing a smaller region requires less data storage and computational time compared to the entire spike protein. This can lead to faster turnaround times and more efficient use of computational resources. Moreover, focusing on this specific region of the spike protein helps to tailor the analyses to address specific research questions. By focusing on the ACE2 RBD, we are targeting a key mechanism of viral entry, which can provide valuable insights into virus-host interactions.

3.2 Structure Prediction Algorithms

This section will elaborate on the computational methods for running structure prediction algorithms: ESMFold, AlphaFold2 and ColabFold. Protein structure prediction algorithms can be classified into Template-Based Modeling (TBM) approaches and into Free Modeling (FM) approaches. TBM methods rely on the availability of similar protein structures in databases like the Protein Data Bank [4] because they involve aligning the target protein sequence to a template structure and then building a model based on the structural information. This means that when suitable templates are available, TBM methods are considered to produce

highly accurate models. Two TBM approaches we employ in this thesis are AlphaFold2 and ColabFold, further described in section 3.2.2 and 3.2.3 respectively. A different approach is the FM approach, which attempts to predict protein structures without using templates. These methods can be more challenging but helpful in predicting the structures of proteins without known homologs. One FM approach is ESMFold, further described in section 3.2.1. An overview of the three structure prediction algorithms can be seen in Table 3.3.

Table 3.3: Comparative Overview of Structure Prediction Algorithms. Unique features, strengths, and performance characteristics of ESMFold, ColabFold, and AlphaFold2, highlighting their characteristics and capabilities in protein structure prediction.

Algorithm	Unique Features	Strengths	Prediction time per sequence (195 aa)
ESMFold	Language model-based approach	Fast predictions, works well with limited sequence data.	1.6 seconds
ColabFold	Integrates MMseqs2 for faster MSA generation	Efficient and accurate for large-scale predictions.	1.7 minutes
AlphaFold2	Uses attention-based neural networks	High accuracy, especially for proteins with known homologs.	40 minutes

3.2.1 ESMFold

One FM approach to predict protein structures is ESMFold (Evolutionary Scale Modeling Fold) [38]. In a paper published in 2023, Lin *et al.* showed that sequence-to-structure prediction algorithms are notably faster than other employed TBM methods while still maintaining high accuracy. ESMFold (version v1) is available for download on GitHub [21] under an open-source software license known as the MIT License. One limitation of ESMFold is its scalability to longer sequences. The ESMFold documentation mentions that ESMFold requires 16-24GB of GPU memory to run. For this reason, the algorithm can only be run using sequences of 400 amino acids or shorter [76]. However, because we used the sequence of the ACE2 RBD which is 195 amino acids long, this did not pose a problem for our use case. If, however, we were to use the whole Spike protein sequence which is made up of 1273 amino acids, ESMFold would

have not been a suitable tool. To install ESMFold on the national supercomputer Snellius, I followed the README from https://github.com/sara-nl/ESMFold_Snellius. The computational infrastructure is explained in detail in Section 3.7. The setup included loading the required modules like *Python/3.10.4-GCCcore-11.3.0* and *purge* and installing dependencies like *PyTorch*, *torchvision* and the libraries *transformers*, *py3Dmol* and *accelerate*. Additionally, ESMFold requires CUDA-compatible GPUs, which are available on Snellius. To install the necessary modules, CUDA 11.8 and cuDNN 8.6.0.163 modules are required. Moreover, ESMFold requires 16-24GB of GPU memory to run. Figure 3.3 shows the ESMFold model architecture, which is further explained in the next passage.

ESMFold Model Architecture

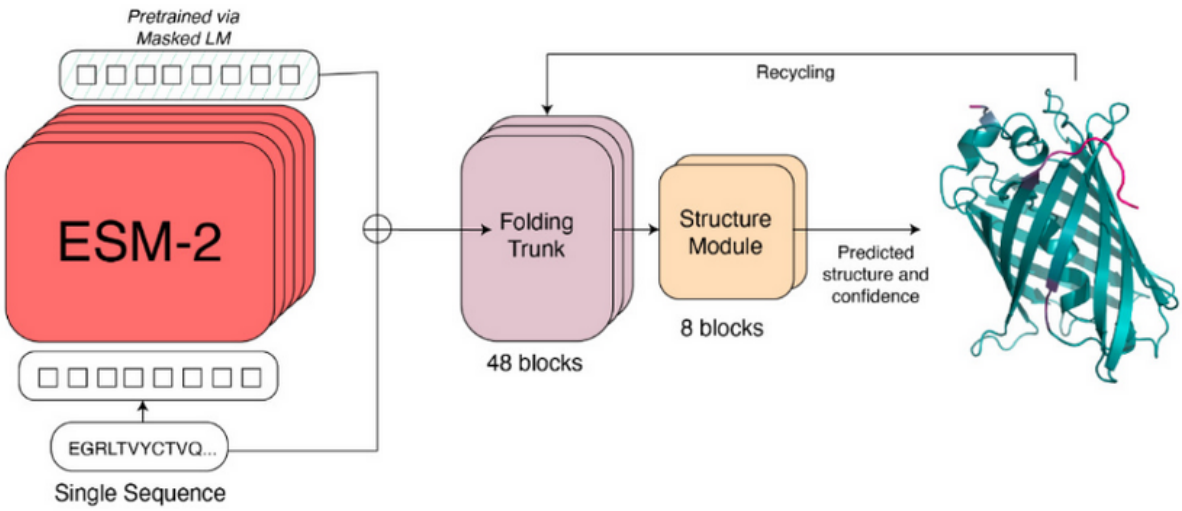


Figure 3.3: ESMFold Model Architecture. The ESMFold architecture is compromised of the ESM-2 language model, a Folding Trunk and a Structure Module. It takes a FASTA file containing a single amino acid sequence as input and outputs the predicted structure. This Figure was adapted from Lin *et al.*[38].

ESMFold requires a FASTA file as input, which consists of the amino acid sequence of the target protein. The first processing step is the **Language Model Processing**. In this step the sequence is fed into the ESM-2 language model. ESM-2 is a transformer-based language model pre-trained on a large amount of protein sequence data. It processes the protein sequence by employing feedforward layers, which are a type of neural network layer where the information

moves in only one direction from the input to the output. They extract important features and representations from the sequence data. The ESM-2 model is trained using a masked language modeling task, which involves randomly masking out amino acids in the protein sequence and then training the model to predict the masked residues by observing the surrounding features. This masked training allows the language model to learn important structural and functional features about proteins without explicit supervision on protein structure. The states and representations obtained from the language model are then passed to the next block.

The **Folding Trunk** is responsible for transforming the latter obtained information into a three-dimensional protein structure. This is achieved through a series of 48 'folding blocks' within the folding head. 'Folding blocks' hereby refer to discrete processing units responsible for specific tasks within the protein folding process. These 'folding blocks' are organized into the folding head. These alternate between updating the sequence representation and a pairwise representation, which encodes the relationships between different parts of the sequence. Ultimately, this ensures the effective translation of sequence information to the three-dimensional structure of a protein.

The output of the Folding Trunk is then passed into a **Structure Module**. The Structure Module comprises eight blocks that perform additional refinement and modifications to the representations received from the Folding Trunk. This transformer Structure Module is based on the architecture proposed in AlphaFold. It is intended to be rotationally and translationally equivariant, ensuring that the predicted structure is unaffected by the input protein sequence's rotations and translation. This ensures that the underlying mechanisms of protein folding are learned without being misled by the arbitrary initial orientation of the input sequence. Lastly, these steps are recycled four times as a default setting to optimize the structural predictions and to ensure that the final atomic-level structure is as accurate as possible. The number of recycling steps can be modified.

Slurm Job Submission

To run ESMFold on Snellius, we employed the SLURM job scheduler. All 3705 sequences were submitted in one job, with each job requesting 1 computational node. A maximum runtime of 5 hours was set, to ensure that a job would have sufficient time to predict all three-dimensional structures. The job was submitted to the GPU partition and the computational node launched 1 task. Each task was allotted 8 CPUs, to enable multi-threaded computation within the allocated resources. These parameters were chosen based on the computational requirements of ESMFold and the available resources in the computing environment. For 3705 sequences, ESMFold took about 1 hour and 40 minutes to complete. Because the mutations in the sequences are at random, three repetitions of the experiment were conducted. Hence, ESMFold was run on a total of 11,115 sequences. This will ensure sufficient data for each point on the trajectories and give an error estimate on the results.

ESMFold Output

ESMFold returns the predicted three-dimensional protein structure file in PDB format. This file contains the structure with the highest predicted local distance difference test (pLDDT) score. From this file, various information about the protein structure can be inferred, such as the x,y,z coordinates for each atom in the protein, as well as information about the amino acid residues. Moreover, ESMFold includes the pLDDT score in the B-factor column for each atom. This score is a measure of the confidence of the predicted three-dimensional structure with values ranging from 0 to 1, 1 being the highest confidence. Additional structural metrics, such as RMSD, TM score and SASA can be calclated from the ouput. These metrics are explained in Section 3.3.

Benchmarking

When comparing the performance results from literature of ESMFold to AlphaFold2, further explained in Section 3.2.2, different prerequisites can be considered. When evaluating ESMFold against AlphaFold2, it becomes apparent that AlphaFold2 achieves higher TM-scores on the

CASP14 dataset [6]. However, when using AlphaFold2 without the use of Multiple Sequence Alignments (MSAs), ESMFold performed better. This is due to their different approaches on predicting the three-dimensional structure of proteins. While AlphaFold2 uses templates in the MSA creation, ESMFold does not. ESMFold is significantly faster for structures that were predicted with high confidence while still reaching good results for RMSD. Predicting all 3705 sequences on Snellius, ESMFold took about 01:40 hours to complete, which means that, on average, predicting a single structure takes about 1.6 seconds. This is significantly faster than AlphaFold2, but it also forfeits accuracy. In total, we performed three independent experiments, thus in total 11,115 protein structures were predicted with ESMFold.

3.2.2 AlphaFold2

While ESMFold is a rapid and still accurate approach for predicting protein structures, AlphaFold2 is much more precise. DeepMind's AlphaFold2, a deep learning-based approach, has been proven to predict protein structures with remarkable accuracy [31]. Following the success of its predecessor, AlphaFold, in the CASP13 competition in 2018, the newer version, AlphaFold2, performed exceptionally in the CASP14 competition, outperforming all other contestants. AlphaFold2 is a TBM approach that predicts the tertiary structure of proteins directly from their amino acid sequence. Its excellent accuracy sets it apart from previous TBM techniques. By leveraging advanced deep learning techniques, AlphaFold2 can predict protein structures at or near experimental resolutions, even for complex structures. These promising outcomes have resulted in the widespread usage of AlphaFold2 around the world, as well as the construction of the AlphaFold Protein Structure Database [78], which gives scientists access to millions of predicted protein structures. AlphaFold2 is available for download on GitHub [24] and published under the Apache License 2.0. On Snellius, the datasets required for AlphaFold2, notably version 2.3.1, were already available in the directory '/projects/2/managed_datasets/AlphaFold'. The genetic databases that AlphaFold2 employs have a total size of 2.62 TB when unzipped, which makes it not feasible to download on a standard computer. They include databases such as BFD [66], MGnify [46], PDB70 [66], PDB [73],

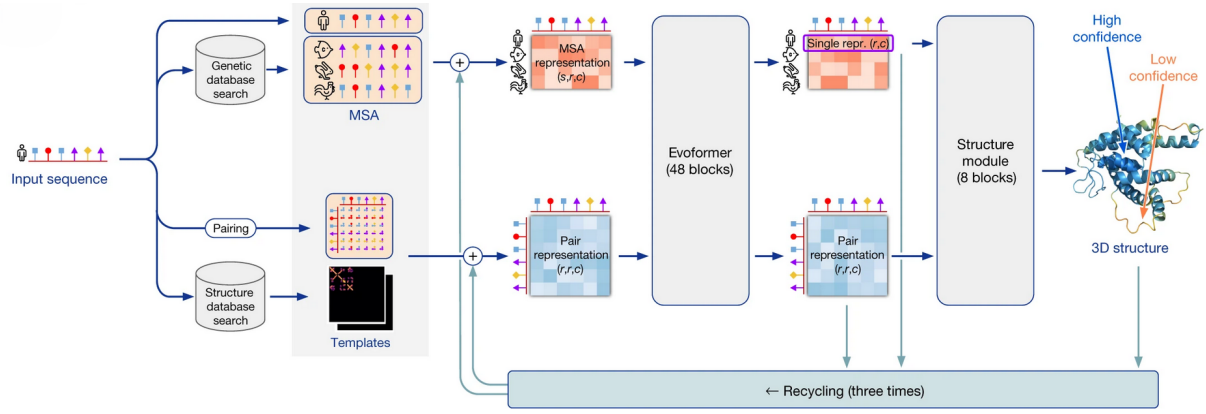


Figure 3.4: AlphaFold2 Model Architecture. The AlphaFold2 model architecture takes a FASTA file containing the amino acid sequence as input. This sequence is then passed through a feature extraction network and a Structure module and after recycling these steps the predicted three dimensional structure is returned. This Figure was adapted from Jumper *et al.*[31].

UniRef30 and Uniref90 [44] and some additional ones if AlphaFold-Multimer is used.

AlphaFold2 performs with such high accuracy because it incorporates advanced neural network architectures and training methods based on evolutionary correlations and homologies to known structures. The use of evolutionary information to put constraints on protein sequences has been deemed successful in other studies as well [63, 41, 29, 54]. Figure 3.4, taken from Jumper *et al.* [31], shows the AlphaFold2 model architecture.

AlphaFold2 Model Architechture

The first step of the AlphaFold2 pipeline is the **input data processing**. AlphaFold2 takes a FASTA file as input, which includes the amino acid sequence and sequence information in the FASTA header. It then performs a genetic database search using JackHMMER [28] and HH-Blits [52]. This database search results in Multiple Sequence Alignments (MSAs), and to avoid biasing the alignment, highly similar or identical sequences are 'de-duplicated' before generating the final MSA. The remaining unique sequences are then arranged into the final MSA by stacking them on each other. Furthermore, a template search is performed. A template is a known three-dimensional protein structure similar to the target protein sequence. Templates are derived from databases such as the Protein Data Bank and serve as starting points for estimating the structure of the target protein sequence. The templates are rated according to

parameters such as the expected number of aligned residues with the target sequence, and the top-ranked templates are used in template-based modeling to construct the predicted structure. Using the MSA from the previous step, a large number of sequences (*default -maxseq 1000000*) in the PDB70 database is searched using HHSearch [68]. The corresponding structural data is then obtained from the mmCIF file in the database. These templates are then further filtered, excluding any that are, for example, identical to the input sequence or that have less than 10% of the primary sequence length. During the training of the model, it is provided with up to 20 templates, and eventually, between 0 and 4 templates are randomly selected from that set. This restriction ensures that the model cannot solely copy the template during training but must learn how to generate the alignment itself. During prediction, the model is then provided with the top 4 templates. The templates are ranked according to their expected number of aligned residues between the template and the query sequence.

The Neural Network Architecture of AlphaFold2 comprises two components: a feature extraction network called **Evoformer** and the **Structure module**. As the name suggests, the feature extraction network processes the input sequence and the MSA generated before to extract meaningful features that will be used in the following stages of the model. This network generates features that capture the evolutionary information present in the MSA, as well as residue embeddings and representations of amino acids. The Evoformer consists of 48 blocks, and each block consists of a neural network that produces an array of the number of sequences and the number of residues (Nseq $\times$ Nres). The resulting rich feature representation enables the model to learn and understand the relationships between different parts of the sequence, which ensures reliable predictions about its structure.

Next, the **Structure module** is where the actual prediction of the protein structure takes place. The Structure module uses a deep neural network with attention mechanisms to predict distances between pairs of amino acids and the protein backbone's angles using the Evoformer features. The attention mechanisms of the Structure module allow for the model to focus on relevant parts of the input data, capture dependencies between amino acids, and learn complex

spatial relationships within the protein. It works by assigning weights to different parts of the input sequence, allowing the model to focus on the most relevant amino acids when making predictions about a particular position. The attention mechanism within the Structure module is one reason why AlphaFold2 is so precise in predicting the three-dimensional structure of a protein.

Finally, AlphaFold's reliability is further enhanced by its use of ensembling techniques. These techniques combine predictions from multiple runs of the model, significantly improving the accuracy and robustness of the final predictions. The model also benefits from the recycling of outputs from previous iterations, which are fed back into the model as additional inputs. This iterative refinement process is a key contributor to the unprecedented accuracy of AlphaFold2, providing a sense of reassurance about the reliability of its predictions.

Slurm Job Submission

To run AlphaFold2 on Snellius, we employed the SLURM job scheduler. Sequences were submitted in batches of 50 sequences, which were then run in parallel using a shell script. In total, we predicted the three-dimensional structure of 263 sequences with AlphaFold2. For each job execution, a total of 6 computational nodes was requested. A maximum runtime of 6 hours was set, ensuring that the job would not exceed the specified duration. Furthermore, the job was submitted to the GPU partition. Each node was configured to have 4 GPUs allocated, enabling parallel processing of GPU-intensive tasks. With each node, 4 tasks were launched simultaneously, optimizing the resource utilization and task distribution. Furthermore, each task was allotted 18 CPUs, enabling efficient multi-threaded computation within the allocated resources. These parameters were chosen based on the computational requirements of AlphaFold2 and the available resources in the computing environment.

AlphaFold2 Output

AlphaFold2 produces various output files that can be used for downstream analyses. Files related to the MSA that were computed during the pre-processing step are saved to a new folder,

which can then also be reused for further predictions. Moreover, *timings.json* provides information on the runtime of each step of the AlphaFold2 pipeline, and *ranking_debug.json* contains information about the quality scores and ranking of the predictive models. Further, AlphaFold2 outputs various PDB files. The raw and unrefined predictions from AlphaFold2 are saved to files *unrelaxed_model_{1,2,3,4,5}_pred_0.pdb*. These are the top 5 models that were generated. The refined predictions that are produced after the structures were run through a force field algorithm to minimize energy are saved in files named *relaxed_model_{1,2,3,4,5}_pred_0.pdb*. Finally, in files *ranked_{0,1,2,3,4}.pdb*, the relaxed versions of the model are stored but now ranked based on their predicted Local Distance Difference Test (pLDDT) score. From this output we can again calculate metrics such as TM score, RMSD and SASA.

Benchmarking

Predicting one sequence with AlphaFold2 on Snellius took about 40 Minutes, which means that without parallelizing the script predicting 3705 sequences would take about 102 days. This is not feasable for high-throughput predictions. However through parallelizing the script, we are able to run 50 sequences in about 4 hours, which means that for one sequence AlphaFold2 took about 5 minutes. This scales to about 13 days for predicting 3705 sequences. Although time wise this would have been feasible, there was not enough computational budget to compute all 3705 sequences. Therefore, we utilized AlphaFold to predict some sequences on the genetic trajectories, totalling to 253 sequences, as well as 10 permutated sequences. AlphaFold2's success was a breakthrough in the field of protein structure prediction. During the CASP competition in 2020 AlphaFold2 achieved accuracy that is comparable to experimental results. A remarkable score was reached for the Global Distance Test (GDT) score, where AlphaFold2 reached a median score of 92.4. This is remarkable because scores above 90 are considered to be equivalent to experimental methods. Moreover, AlphaFold2's predictions only had a median distance of 0.96 $\AA$ to experimental models, with the next-based method having a distance of 2.83 $\AA$ to the experimental model. These promising results made AlphaFold the state-of-the-art method for predicting protein structures. However, because of its computationally expensive nature we utilized another promising method called ColabFold, explained in Section 3.2.3.

3.2.3 ColabFold

While AlphaFold2 is a powerful tool to predict the three-dimensional structures of a protein, it is also computationally very expensive. This can lead to challenges when scaling to large numbers of sequences like we have in this study. A more scalable method to predict protein structures, which is very similar to AlphaFold2, is ColabFold [45]. It allows a user to use the AlphaFold2 protein structure prediction model without having to install the full AlphaFold2 software. It is available on GitHub https://github.com/sokrypton/ColabFold and published under the MIT License. ColabFold is faster than AlphaFold2 because it uses a faster homology search, which can spead up the prediction process immensly. ColabFold has both a web and a command line interface. To install ColabFold on the Snellius supercomputer the following packages had to be installed: *pdbfixer* and *silence_tensorflow*. Moreover, ColabFold (version v1.5.5) was installed from GitHub https://github.com/sokrypton/ColabFold. Because the database and model parameters of AlphaFold2 were already installed on Snellius, some changes were made to the ColabFold scripts. Some additional modules that have to be loaded before running ColabFold, namely *CUDA/11.7.0*, *SciPy-bundle/2022.05-foss-2022a*, *NCCL/2.12.12-GCCcore-11.3.0-CUDA-11.7.0*, *protobuf/3.19.4-GCCcore-11.3.0*, *protobuf-python/3.19.4-GCCcore-11.3.0*, *TensorFlow/2.11.0-foss-2022a-CUDA-11.7.0*, *Biopython/1.79-foss-2022a*, *HH-suite/3.3.0-gompi-2022a*, *HMMER/3.3.2-gompi-2022a*, *Kalign/3.3.5-GCCcore-11.3.0*, *jax/0.3.25-foss-2022a-CUDA-11.7.0*, *OpenMM/8.8.0-foss-2022a* and *Python/3.10.4-GCCcore-11.3.0*.

Because MMseqs2 and the dataset were not installed on Snellius at the time the predictions were run, a server in Korea was pinged. This server, however, has a limit of 1000 searches per day per IP address, which can slow down the prediction process. After completing all structure predictions, MMseqs was also made available on Snellius. Figure 3.5 shows the ColabFold model architecture, which was adapted from Mirdita *et al.*[45].

ColabFold Model Architecture

ColabFold takes a FASTA file containing a sequence as input and sends it to an MMseqs2 server. MMseqs2, which stands for Many-against-Many sequence searching, was developed by

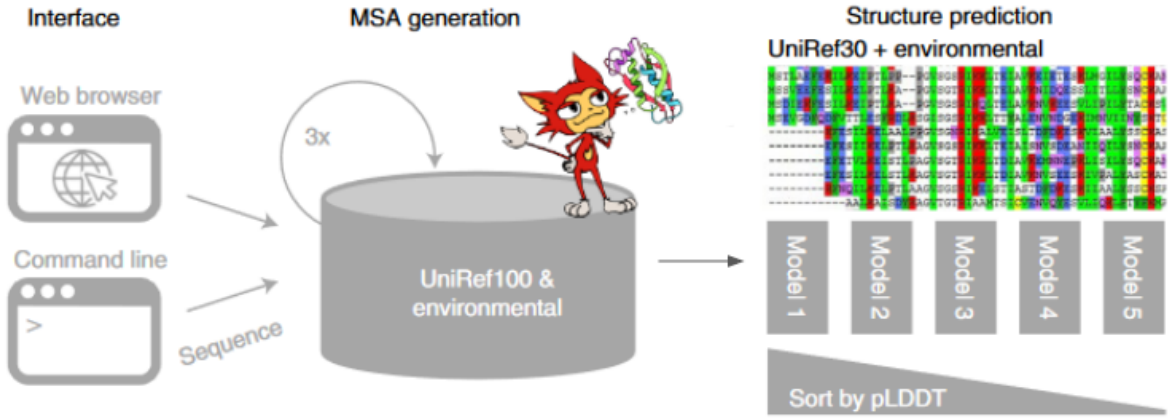


Figure 3.5: ColabFold Model Architecture. ColabFold takes a FASTA sequence as input and generates MSAs by using MMseqs2. The Figure was adapted from Mirdita *et al.*[45].

the Soeding Lab at the Max Planck Institute for Biophysical Chemistry [67]. It provides a fast homology search to build MSAs and to find templates. MMseqs is an open-source and GPL-licensed software and is 40-60 times faster compared to HMMer and HHblits, which are utilized by AlphaFold2. MMseqs2 optimizes the MSA generation by employing an optimized workflow. It first leverages the UniRef30 clustered database, hence searching only 29.3 million sequences instead of the 277.5 million UniRef100 sequences. Moreover, MMseqs2 also uses a new diversity-aware filtering algorithms which is an extension of the HHblits filtering algorithm. The even sampling of sequence space ensures a high quality MSA that captures diversity well. Moreover, ColabFold also combines two of the databses used in AlphaFold2 into a reduced version with less repetition, namely BFD and MGnify.

The MMseqs2 server aligns input sequences against the Uniref100 and PDB70 databases as well as an environmental sequence set. It performs three iterative search rounds to build the final MSA. After the MSA is generated, the input features are provided into the AlphaFold2 models. To improve speed, ColabFold only compiles one instead of five of the AlphaFold2 models. Moreover, weights are reused, which speeds up the process even more. Another criterion that was implemented is an early stop criterium, which would stop the recycling if an amply accurate structure was found.

Slurm Job Submission

To run ColabFold on Snellius, we again employed the SLURM job scheduler. Sequences were submitted in batches of 50 sequences per job. For each job, 1 computational node was requested and a maximum runtime of 3 hours was requested. Additionally, the job was submitted to the GPU partition. Each node was configured to have 1 GPU allocated and with one node 1 task was launched. Each task was allotted 8 CPUs, to again enable efficient multi-threaded computation within the allocated resources. It is important to mention that because the number of sequences that are allowed to be submitted are limited to 1000 per day, that a maximum of 20 jobs each containing 50 sequences could be submitted per day. These parameters were chosen based on the computational requirements of ColabFold and the available resources in the computing environment.

ColabFold Output

ColabFold provides various output files including multiple PDB files that contain the predicted 3D protein structures. They are named sample_id_unrelaxed_rank_001,002,003_model_i.pdb, where rank 001 is the best-ranked structure. Additionally, JSON files containing confidence scores and metrics such as pLDDT and PAE (Predicted Algined Error) are provided. If the MSA was not provided as input they are also provided as an output in A3M format which can be used for subsequent ColabFold runs. Lastly, ColabFold also provides PNG image files, visualizing the pLDDT and PAE score. Again, other metrics such as RMSD, TMscore and SASA can be calculated from the output.

Benchmarking

Predicting sequences with ColabFold on Snellius was significantly faster than with AlphaFold but slower than ESMFold. Predicting the structure of 50 sequences took about 1:25 hours, which means that on average predicting a single protein sequence took about 1.7 minutes. Thus, predicting 3705 sequences takes on average about 104 hours. Mirdita *et al.* [45] found that ColabFold reaches similar accuracy compared to AlphaFold2 while being magnitudes

faster. Moreover, in a preprint from Edmunds, McGuffin, and Genc [19] the authors evaluate various metrics to observed experimental values which show that ColabFold has a strong ranking agreement between predicted and experimental values. However, they found out that the agreement between predicted and experimental values is better for monomers that multimers.

3.3 Structural Analysis Metrics

Various structural metrics can be derived from the output files of the aforementioned structure prediction algorithms. These metrics are distinct measures but their integration can lead to a more comprehensive understanding of protein structure. An overview of the structural metrics can be seen in Table 3.4 and they are further described in this section. We classify them into two groups: comparative metrics, which are calculated by comparing a protein structure to a reference structure, and intrinsic metrics, which are calculated based solely on the predicted structure itself without external reference. These metrics provide insights into different aspects of protein folding, stability, and function, contributing to a deeper analysis of the predicted protein structures.

Table 3.4: Overview of Structural Metrics. Categorization of protein structure analysis metrics into comparative and intrinsic measures, highlighting their distinct roles in assessing structural properties and similarities.

Metric	Description	Unit
	Comparative Metrics	
TM Score	Measures structural similarity between two protein structures.	NA
RMSD	Measures the average distance between the atoms of superimposed molecules.	Å
	Intrinsic Metrics	
pLDDT	Measures a confidence score of the local structure that got predicted for each amino acid	NA
SASA	Measures the accessible area to solvent molecules.	$Å^2$

TM score

The TM (Template Modeling) score measures the structural similarity between two protein structures [88]. The TM score is defined as

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{target}}}\sum_{i=1}^{L_{\text{common}}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right],$$

where L_{target} refers to the length of the protein structure and L_{common} refers to the length of common residues between the target structure and the aligned structure. Moreover, d_i refers to the distance between the ith aligned residues and d_0 is a normalization scale. The TM score ranges between $(0, 1]$ with 1 describing identical structures. The higher the score the more similar two sequences are. It can be assumed that TM scores below 0.5 is considered not similar and any score above 0.9 is considered to be very similar.

RMSD

The root-mean-square-deviation (RMSD) measures the average distance between the atoms of superimposed molecules [39]. Superimposed molecules are structures whose centroids are at the origin, which is done by translating and rotating them. If this step is not done, one might get a misleading RMSD value, which is not the true minimal RMSD. The RMSD is defined as

$$\text{RMSD}(\mathbf{ref}, \mathbf{target}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i^{ref} - w_i^{target}\|^2},$$

where v_i^{ref} refers to the atom coordinates of the reference structure, w_i^{target} refers to the atom coordinates of the translated and rotated target structure and n is the number of atoms that are being compared. The squared difference between the positions of each atom in the two structures is calculated and summed up. The sum is then divided by the number of atoms, and the square root is taken to get the RMSD value. To superimpose the structures, the Biopython module *Superimposer* is used [26]. This module uses the Kabsch algorithm to superimpose the structures [32]. The RMSD is measured in Å, and while there is no universal cutoff value

for what can be considered a good value, generally, values below 2 Å are deemed to be very close. A RMSD of 0 Å means that the structures have identical values, and the higher the value, the more dissimilarity [39]. However, the specific RMSD threshold for defining significant structural changes can vary depending on the analysis context.

pLDDT

The predicted Local Distance Difference Test (pLDDT) value is a confidence score of the local structure that got predicted for each amino acid. The score is based on the local difference test C^α (lDDT-C^α) [40] and is stored in the B-factor column of the PDB file. The lDDT-C^α score is calculated as

$$lDDT - C^\alpha = \frac{1}{4} \sum_{t \in 0.5,1,2,4} \frac{N_d \in t}{N_d},$$

where the t is a threshold measured in Å, N_d is the total number of distances and $N_d \in t$ describes the number of distances that are within the threshold t. These local distances are identified between the C^α pairs and it is determined whether the distance in the predicted structure is within a certain threshold distance of the reference distance.

It ranges from 0 to 100 with higher values indicating a higher confidence in the local structure prediction for an amino acid. pLDDT scores greater than 90 are considered a very high accuracy and pLDDT scores between 70 and 90 indicate a high confidence in the local structure prediction while values below 70 can correspond to a less accurate prediction. Any value below 50 is considered an unreliable prediction. For ESMFold, the pLDDT scores are scaled between 0 and 1.

SASA

Another metric that provides insights into the folding and stability of a protein is the Solvent Accessible Surface Area (SASA). The SASA measures the accessible area to solvent molecules. The SASA was calculated using the *FreeSASA* module in Python, which utilizes the Shrake-Rupley algorithm [64]. This algorithm conceptually divides the protein surface into small sections and rolls a spherical solvent molecule over the surface of these sections. The

SASA of each atom it then calculated by taking the fraction of the solvent molecule surface that does not overlap neighboring atoms. The SASA is calculated as

$$SASA = 100 \cdot \frac{\sum_i \left(\frac{R}{\sqrt{R^2 - Z_i^2}} \right) \cdot \left(\frac{\Delta Z}{2} + \Delta' Z \right) \cdot L_i}{4\pi R^2},$$

where L_i describes the arc length on each section i, Z_i denotes the vertical distance from the sphere's center to section i, ΔZ is the spacing between the sections and R is the radius of the sphere. The total SASA for a protein structure is then calculated by summing up the SASA values of all constituent atoms. A lower SASA value indicates a more stable folded structure while having a higher SASA hints at a more unfolded conformation. This provides insights into the folding and stability of a protein.

3.4 Permutated Sequences

In our study of protein sequence space, we begin with the biological sequence of the ACE2 RBD, from which trajectories are generated through accumulating random mutations. This approach raises the question whether the RBD sequence itself is unusual in its structure and how this influences the resulting sequences composing the trajectories. To gain insight into the ACE2 RBD's structural properties, we generate a set of randomly constructed control sequences as a 'null hypothesis'. To isolate the physical aspects of the molecule from its evolutionary histroy, we employed a permutation method that preserves the sequence length and amino acid composition whole erasing the evolutionary information in the ACE2 RBD. We created 10 random permutations of the original ACE2 RBD sequence. This process effectively separates the physical contributions to protein folding from the biological contributions, helping us calibrate our datasets. To quantify the degree of permutation, we used the Kendall tau coefficient, defined as

$$\tau = \frac{n_c - n_d}{\binom{n}{2}},$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and $\binom{n}{2}$ is the binomial coefficient representing the total number of pairs in the sequence of length n, calculated as $\frac{n(n-1)}{2}$. The Kendall's tau coefficient ranges from -1 to 1, where 1 indicates that the two sequences are identically ordered, -1 indicates that one sequence is in reverse order to the other sequence and 0 indicates that the sequences are randomly ordered to each other. The τ for the 10 random permutations and the structural metrics of the resulting three-dimensional structures can be seen in Section 4.3. Analyzing the permutated sequences allows us to distinguish between the physical and evolutionary contributions to protein structure, offering a more comprehensive understanding of the factors influencing protein folding and stability in our trajectory analysis.

3.5 Contact Matrices

Contact matrices from PDB files can provide valuable insight into the structural properties and the dynamics of biomolecules. These contact matrices represent the residue-residue contacts within a protein structure, offering a representation of the overall folding pattern. The analysis of those matrices along the trajectory can provide valuable insights into various structural features, including potential conformational changes. Contact matrices are usually constructed by analyzing the spatial proximity of residue pairs within a protein structure. A residue-residue contact is defined as such when the distance between two non-hydrogen atoms from different residues falls below a specified cutoff value, usually ranging from 4 to 8 Angstroms. These contacts are presented as binary values, where 1 indicates that a contact between a residue pair is present and a value of 0 represents the absence of a contact. The contact matrix for our sequence of length 195 is represented as a matrix of $N_{seqlength} x N_{seqlength}$ which results in an upper bound of 38,025 contacts, assuming a fully connected contact map.

To analyze the structural properties and dynamics of the protein variants, we calculated contact matrices from the corresponding three-dimensional structures obtained from the structure prediction algorithms. They were generated using a python script that takes a PDB file and

distance cutoff, as input parameters. For each PDB file, the script parses the structure and iterates over all pairs of residues in the structure, and checks if any pair of non-hydrogen atoms between the two residues is within the specified distance cutoff. If a contact is found, the corresponding elements in the matrix are set to `TRUE`. Otherwise, the matrix element is set to `FALSE`. The distance cutoff used for determining residue-residue contacts was set to 8 Angstroms, which is a commonly used value for capturing relevant interactions withing protein structures [83]. The contact matrices were calculated for all predicted structures from ColabFold and ESMFold and saved as CSV files for further analysis. The contact matrices serve as a valuable resource for inferring structural information, such as identifying potential conformational changes. Analyzing these changes in contact patterns along the trajectory towards the homopolymers gives insight into the folding patterns of proteins.

3.6 Genetic Constraint on ACE2 RBD

Our sampling methods of sequence space are based on mathematical relations between sequences. The genetic code offers an additional set of relations to consider. While we initially focused on the general physical properties of protein folding and the distribution of structure in protein sequence space, an intriguing dimension to explore is the impact of single nucleotide mutations on amino acid composition within homopolymeric trajectories. The genetic code was discovered in the early 1960s by Marshall Nirenberg and Johann Matthaei [48] and consists of a series of triplets of nucleotides, called codons, with each codon translating to a specific amino acid or a stop codon. Most amino acids are encoded by more than one codon, however, amino acids such as methionine (M) or tryptophan (W) are only encoded by one codon. The nucleotide sequence of ACE2 RBD was extracted from the NCBI Nucleotide database with accession number *NC_045512.2* [8]. For each codon there are $3^3 = 27$ possible single nucleotide mutations. However, because some single nucleotide mutations do not result in a different amino acid, I only consider the missense mutations that resulted in a different amino acid. We generated mutational trajectories towards the homopolymer of each amino acid, however, only as far out as genetically possible through missense mutations. The three-dimensional structure

of the sequences possible through missense mutation was predicted with ColabFold and the results are shown in Section 4.4.

3.7 Computational Infrastructure

This subsection explains the computational infrastructure that was used to run structure prediction algorithms, as well as the infrastructure used for the analyses of the results. It highlights technical details as well as challenges and issues encountered.

3.7.1 SURF

The Snellius supercomputer, operated by SURF [70], is a high-performance computing (HPC) system that enables advanced scientific research across various domains. It boasts a large number of compute nodes, each equipped with powerful processors capable of executing complex computational tasks in parallel. Moreover, it offers a significant amount of memory per node, allowing for the analysis of large datasets required for computationally intensive tasks. Furthermore, the supercomputer is backed by high-capacity storage systems capable of storing vast amounts of data generated during computations. Another advantage of using the Snellius supercomputer is its optimization for parallel computing, which accelerates the execution of computational workflows. The Snellius supercomputer is typically accessible to researchers affiliated with academic or research institutions, as well as university students through faculty contracts with SURF. Additionally, users have access to dedicated support through the SURF help desk. The parallel computing capabilities of the Snellius supercomputer are particularly relevant to research tasks that involve computationally intensive algorithms such as protein structure prediction algorithms. Moreover, Snellius is equiped with NVIDIA GPUs, which are essentail for running the structure prediction algorithms.

Access to SURF's computing resources was granted through the Leiden Institute for Advanced Computer Science (LIACS). Initially, the predictions were planned to be run on local L40s GPU nodes at LIACS, as this would have been more cost-effective than utilizing SURF's re-

sources. However, due to delays in setting up the nodes caused by the faculty's relocation to a new building, Rob van Nieuwpoort provided access to SURF's Snellius supercomputer. The available computing resources on the Snellius supercomputer are measured in System Billing Units (SBUs), which quantify the amount of compute cores a job utilizes over time. The initial allocation of budgetary units was quantified at 100,000. However, after reevaluation they were increased culminating to 300,000 SBUs. The rates for SBUs can be found in the catalog for SURF services and rates [72].

LIACS's contract with SURF permitted me access to the 'gpu', 'gpu_mig' and 'staging partition'. The 'staging partition' utilizing 'gcn' nodes with 72 cores per node further utilizes a multi-instance GPU (MIG), which allows the partitioning of a serves as a resource hub for managing and transferring data between different storage systems or computing environments. Leveraging powerful 'srv' nodes equipped with Simultaneous Multithreading (SMT) technology, this partition can effectively double its computational throughput, accommodating up to 32 threads per node. Moreover, each node offers 224 GiB of memory capacity. The 'gpu' partition features nodes of type 'gcn', with 72 CPU cores available per node. Furthermore, each node also has a substantial amount of memory available and offers 480 GiB. The 'gpu_mig' partition, also utilizing 'gcn' nodes with 72 CPU cores per node, further utilizes a multi-instance GPU (MIG), which allows the partitioning of a single GPU into multiple instances [71]. To account how many SBUs are used for each job submission, SURF provides a formula to measure the amount of compute cores that are used in a job. Because I was only given access to GPU nodes, the accounting is calculated on a "per GPU" basis. When allocating 1/4th of the GPU node 'gcn', the smallest allocation possible, an accounting weight factor of 128 per GPU is used. This means that for 1 GPU per 1 hour, 128 SBUs are used.

A SSH protocol is used to connect to Snellius. I utilized MobaXterm, which is a SSH client, and FileZilla, which is a File Transfer Protocol client. The host of Snellius is located at *snellius.surf.nl*. Snellius uses a batch system to distribute computational tasks over available nodes. This means that when a job is submitted, it is not started right away, but it enters a queue.

This queue uses a fair share algorithm to ensure users do not overuse their resources. Jobs can have different priorities depending on how much budget a user has already used or on the recourse allocation of a submitted job. Hence, depending on the job size, users sometimes can wait a few days until their job is executed. This can conflict with time-sensitive work.

To make the generated data accessible, it is stored on Research Drive, further explained in Section 3.8.3

3.7.2 Analysis Tools

The generated output by the structure prediction algorithms was then analyzed on a Windows machine with an Intel(R) Core(TM) i5-8265U CPU clocked at 1.60GHz, running at 1800 MHz, with four cores and eight logical processors. Python 3.9. was used for analyses and libraries such as *NumPy*, *Pandas*, *Scipy*, *Bio.PDB* and *Matplotlib* were utilized for extracting statistical metrics and visualizing them.

3.8 FAIR Data Management

The comprehensive collection and accurate data modeling of researched objects are imperative in preserving the scientific record. This ensures the availability and reproducibility of relevant information when such objects are referenced. These practices are key components of research integrity, which encompasses the ethical principles and professional standards that guide responsible conduct in scientific research. A fundamental aspect of effective data management and stewardship is the adherence to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, as defined by Wilkinson *et al.* [82] in 2016. These principles are designed to address the challenges associated with data reuse and maximizing its utility. Proper data curation and management entails the formulation of a robust data management plan and the adoption of FAIR-compliant formatting and storage practices. Furthermore, sustaining the discoverability and accessibility of data, even post-project completion, is essential to fostering FAIR and Equitable Science.

Figure 3.6: Research Data Lifecycle. This Figure shows the research data lifecycle (RDL). It was adapted from https://osf.io/3xjf8 [61]. In this thesis we focus on steps 3 and 4 of the RDL.

3.8.1 Research Data Lifecycle

In this thesis, we will present our FAIRification efforts within the context of the research data lifecycle (RDL), as seen in Figure 3.6. The RDL includes several essential processes that ensure a systematic approach to managing research data from the project proposal to publication. The first stage comprises of a project proposal. Next, a data stewardship plan is created, which also includes a FAIR Implementation Profile. In the third step, the project is carried out, producing research outputs such as data and metadata. The fourth stage is to publish the research outputs in domain-appropriate and reliable repositories. Lastly, research results are published in academic articles. In the following sections, I will focus on parts 3 and 4 of the RDL: the creation of data and metadata and the publication of research outputs.

3.8.2 Creation of Data and Metadata

During the project execution phase, various research outputs are generated. This section will detail the methods and tools used to ensure these outputs adhere to the FAIR principles. To

promote FAIR Data management within the STAYAHEAD project we propose levaraging and combining two approaches: the I-ADOPT framework and nanopublications.

The I-ADOPT (Integrated Annotation and Data Operations Tools) framework [56] is a comprehensive approach to enhancing the FAIRness of biomedical data. The I-ADOPT framework allows us to make the data acquired from various sources machine-actionable. The data stems from three different data sources. We can obtain information about a protein variant from real-world observations, from laboratory-based assessments, and from computational predictions. The framework consists of 6 key components that enable a more seamless data exchange. Firstly, it promotes the usage of controlled vocabularies and ontologies. Another key component is a set of tools for data transformation and integration which promotes interoperability. It also documents data provenance, which includes any changes applied to the data, hence promoting reproducibility. Moreover, data elements are assigned a globally unique identifier to ensure the referencing of data across different platforms. Another important aspect is the robust management of metadata to enhance the discoverability of data. Lastly, I-ADOPT allows for the controlled access and sharing of data.

In addition to the I-ADOPT framework, we promote the use of the nanopublication server [47, 59, 35], which offers a decentralized network for the publication, storage, sharing of nanopublications. Nanopublications are small knowledge graphs which are essentially their own independent publication. A nanopublication consists of three components: the assertion, which is expressed as a set of RDF triples; information about data provenance and information about the publication itself. Nanopublications facilitate the dissemination of individual data points as independent, machine-readable publications, while the I-ADOPT framework ensures consistent annotation, interoperability, and metadata management across diverse data sources. To represent the amino acid sequences of the ACE2 RBD variants in a machine-readable and interoperable format, we utilize the Resource Description Framework (RDF). The core idea is to use a subject-predicate-object structure, known as a triple. The RDF includes information about the amino acid sequence, its composition, and information on what this variant was

derived from. Because the data for each protein variant is now interoperable, it resembles a FAIR Digital Twin, which will further be explained in the following paragraph.

FAIR Digital Twin

To create a digital representation of our physical research entities, we make use of the concept of a FAIR Digital Twin (FDT). Much can be said about a given protein sequence, spanning from the complete amino acid sequence to intricate structural and functional information, epidemiological insights regarding variant prevalence, and experimental data elucidating a variant's infectivity and immune evasion. We advocate using a FDT to ensure machine actionability when storing this information. I-ADOPT and Nanobuplicatons provide a means for us to instantiate a FDT. A schematic representation of a FDT can be seen in Figure 3.7. At the core of this schema lies a sequence variant, which is systematically annotated to a reference sequence. This sequence variant is conceptualized as a nanopublication. The nanopublication framework allows for the sequence variant to be semantically linked to other nanopublications, each delineating specific properties or characteristics of the variant.

A FDT is a virtual representation of a physical asset that adheres to the FAIR principles [58]. This entails providing the Digital Twin with a unique identifier to facilitate discoverability. Additionally, relevant metadata ensures the digital twin is searched and indexed. Data should be accessible through standardized protocols, and metadata should be openly available. The FAIR Digital Twin also uses standardized ontologies to represent its data, facilitating its integration with other data. Lastly, the FAIR Digital Twin must be well-documented with an explicit usage license and data provenance information. Ultimately, adhering to all the principles mentioned above, the digital twin becomes machine-actionable. We propose using computational data, experimental data from the laboratory, real-world observations, and theoretical properties. Leveraging these diverse data properties and storing them in a FAIR-compliant way will make the research data more meaningful. At its core, the proposed model entails a sequence annotated against a reference sequence and different predicates. Each sequence can thereby have 1) a sequence space navigation observation, 2) a real-world observation, 3) an experi-

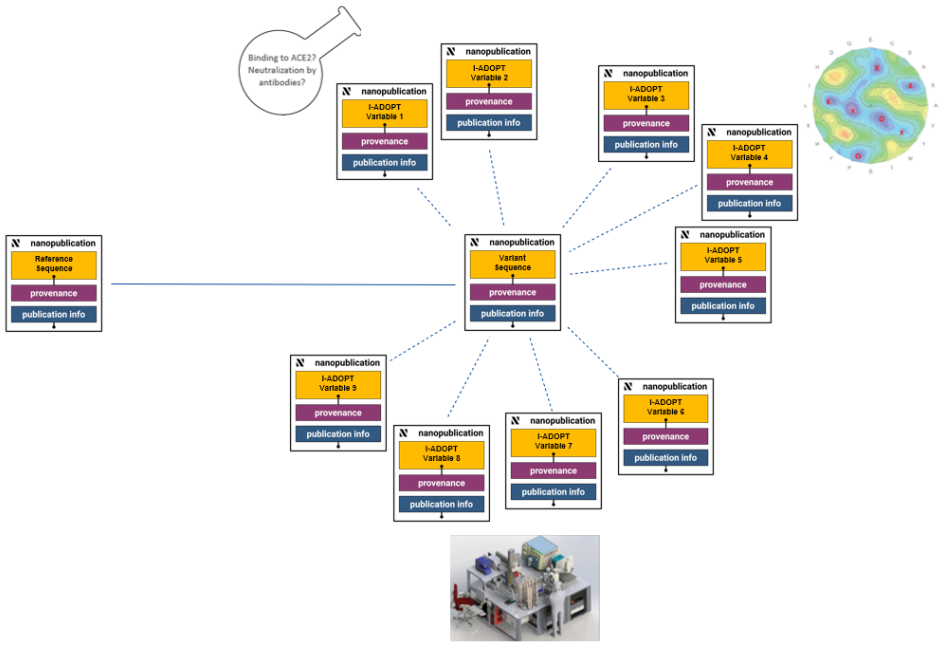


Figure 3.7: Schematic representation of a FAIR Digital Twin. The FDT is instantiated through the use of nanopublications and the I-ADOPT framework. Adapted from https://osf.io/q6r9d [62].

mental observation, and 4) a computational observation. An example model of a FAIR Digital Twin can be seen in Section 4.5.

RDF

Listing 3.1 shows an example RDF of a sequence variant, specifically a sequence variant of step 2 on trajectory A. It defines the namespace for the UniProt core ontology (line 3) and the RDF namespace (line 4). Within the RDF, there are two RDF descriptions defined. The first description defines an RDF resource identified by a URI (line 7) and represents a sequence and a modified sequence. The second description defines an RDF resource identified by a URI (line 17) and represents an amino acid trajectory identifier for the sequence 'A 2'. Within the first RDF description, the label of the sequence is specified as ' A 2' (line 10). The sequence itself is represented by the 'ns1:sequence' property (line 11). The length of the sequence is specified by the 'ns1:sequence_length' property (line 13). Further, we provide a composition vector for the sequence, which represents the amount of each amino acid in the sequence (line 14). Lastly, we state that the sequence is a missense mutation of another sequence identified by the URI

'http://purl.uniprot.org/core/A_1'. Notably, the interrelation between sequences is elucidated through a 'chain', where each subsequent variant on the trajectory is a missense mutation of its predecessor. For instance, the first variant on the trajectory is always a missense mutation of the reference sequence ACE2 RBD. Each following variant refers to the subsequent sequence on the trajectory. Hence, sequence 2 on trajectory A is a missense mutation of sequence 1 on trajectory A. Within the second RDF description, the label of the identifier for the trajectory of the sequence is defined (line 18).

Listing 3.1: RDF of Sequence on Trajectory A 2 steps away from reference sequence

```
1  <?xml version="1.0" encoding="utf-8"?>
2  <rdf:RDF
3     xmlns:ns1="http://purl.uniprot.org/core/"
4     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
5     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
6  >
7    <rdf:Description rdf:about="http://purl.uniprot.org/core/
        A_2">
8      <rdf:type rdf:resource="http://purl.uniprot.org/core/
          Sequence"/>
9      <rdf:type rdf:resource="http://purl.uniprot.org/core/
          Modified_Sequence"/>
10     <rdfs:label>A 2</rdfs:label>
11     <ns1:sequence>TNLCPFGEVFNA...TVCGP</ns1:sequence>
12     <ns1:referred_to_as rdf:resource="http://purl.uniprot.org
          /core/identifier_A_2"/>
13     <ns1:sequence_length rdf:datatype="http://www.w3.org
          /2001/XMLSchema#integer">195</ns1:sequence_length>
14     <ns1:composition_vector>
          14,8,9,6,14,15,1,6,8,13,0,17,11,6,9,14,11,16,2,15</
          ns1:composition_vector>
15     <ns1:is_missense_mutation_of rdf:resource="http://purl.
          uniprot.org/core/A_1"/>
16   </rdf:Description>
17   <rdf:Description rdf:about="http://purl.uniprot.org/core/
        identifier_A_2">
18     <rdfs:label>Amino acid trajectory identifier for A 2</
          rdfs:label>
19   </rdf:Description>
20 </rdf:RDF>
```

3.8.3 Publication of Research Outputs

Upon project completion, the research outputs are published in trusted repositories. This section covers the platforms used for data and code publication. The Open Science Framework (OSF)[23] is used to host and share research data, ensuring it is accessible and citable. Sequence data is uploaded in both FASTA and RDF formats to cater to different research needs. Additional figures from analyses, as well as this thesis manuscript are shared there as well. The project can be accessed through this link https://doi.org/10.17605/OSF.IO/PMJ9X [57]. Furthermore, to facilitate secure data storage of the predicted three-dimensional protein structures and a more seamless collaboration during the current stage of the STAYAHEAD project, we utilize Reseach Drive, which is a cloud-based storage service operated by SURF. Research Drive offers ample storage capacity, enabling us to store and share large volumes of research data generated within this project. It proves to be particularly useful for our collaboration with project partners at Utrecht University, who are responsible for the experimental component of the STAYAHEAD project. Research Drive enables your data to be shared with colleagues or upon request also with people outside the project. Moreover, Research Drive provides a feasible solution when local storage resources on personal computers are insufficient to accommodate the substantial data volumes generated during the project. All code generated during the project is shared on GitHub under https://github.com/slilli/Protein-Sequence-Space. This platform is chosen for its wide adoption in the research community, version control capabilities, and support for open science.

Chapter 4

Results

This chapter presents a summary of the results of the structure prediction algorithms ESMFold, AlphaFold2, and ColabFold. The full results can be found on OSF under https://doi.org/10.17605/OSF.IO/PMJ9X [57]. They are evaluated using a variety of structural metrics. Furthermore, we analyze permutated sequences of the reference sequence and the sequence space of ACE2 RBD when posing genetic constraints. To demonstrate FAIR data management, we include an example of a FAIR Digital Twin.

4.1 Structure Prediction Algorithms

Table 4.1 shows the number of sequences that were predicted with the structure prediction algorithms. For ESMFold, we performed three statistical repetitions of the sequences on the trajectories (3705 sequences), which resulted in 11,115 sequences that were successfully predicted using ESMFold. We performed two statistical repetitions of the trajectories for ColabFold. Notably, ColabFold failed to predict some structures for both repetitions. All these structures were close to the respective homopolymers. For the first repetition, ColabFold failed to predict the three-dimensional structures for the following sequences towards their respective homopolymers: trajectory A sequences 157-183 (27 sequences), trajectory C sequences 161-187 (27 sequences), trajectory G sequences 150-180 (30 sequences), trajectory N sequences 151-178 (28 sequences), and trajectory T sequences 154-184 (30 sequences). For the sec-

ond repetition, ColabFold failed to predict the three-dimensional structures for trajectory A sequences 155-183 (29 sequences), trajectory C sequences 161-187 (27 sequences), trajectory G sequences 154-180 (27 sequences), trajectory N sequences 156-178 (23 sequences), and trajectory T sequences 152-184 (33 sequences). Hence, for repetition 1, ColabFold failed to predict the three-dimensional structure for 142 sequences; for repetition 2, ColabFold was unable to predict the three-dimensional structure for 139 sequences. This occurred because ColabFold employs templates to predict protein structures, and when no suitable templates exist in databases, ColabFold is unable to predict the three-dimensional structures. Additionally, the genetic trajectories (1135 sequences) and the permutated sequences (10) were successfully predicted with ColabFOld. We utilized AlphaFold2 to predict 10 permutated sequences and three genetic trajectories (A, P, and S), totaling 263 sequences. An overview can be seen in Table 4.1.

Table 4.1: Predicted Protein Structures. An overview of the number of sequences predicted with each structure prediction algorithms ESMFold, ColabFold, and AlphaFold2. The experiment column refers to either the directed mutational trajectories ('Trajectories'), the trajectories that are genetically possible through missense mutations ('Genetic trajectories') or the permutated sequences ('Permutated seqs').

ESMFold				
Experiment	**Seqs Predicted**	**Repetitions**	**Predictions Failed**	**Total Structures Predicted**
Trajectories	3705	3	0	11,115
ColabFold				
Experiment	**Seqs Predicted**	**Repetitions**	**Predictions Failed**	**Total Structures Predicted**
Trajectories	3705	2	repetition 1: 144 repetition 2: 139	7,127
Genetic trajectories	1135	1	0	1135
Permutated seqs	10	1	0	10
AlphaFold2				
Experiment	**Seqs Predicted**	**Repetitions**	**Predictions Failed**	**Total Structures Predicted**
Permutated seqs	10	1	0	10
Genetic trajectories	253	1	0	253

4.1.1 Sequence Space Trajectories

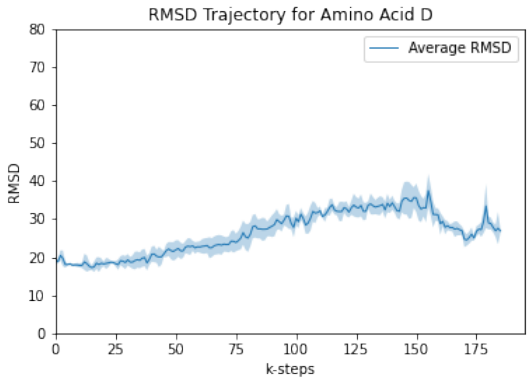
We present representative examples of structural metric analyses performed for each sequence along the trajectories toward the homopolymers of the 20 different amino acids. Analyzing

these metrics along the trajectories provides insights into the distinct structural behaviors and folding mechanisms as we diverge from the reference sequence. The full results can be found in Appendix A and on OSF under https://doi.org/10.17605/OSF.IO/PMJ9X [57].

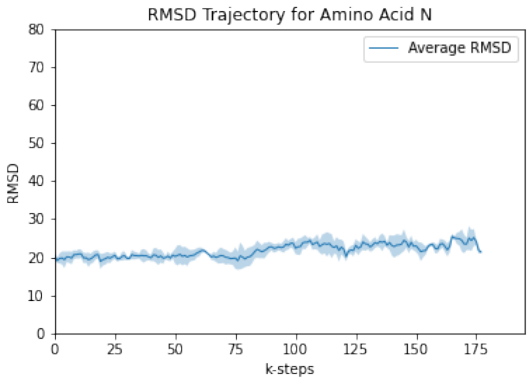
ESMFold

We quantified the structural deviations from the reference sequence conformation along the trajectories using the RMSD. Figure 4.1 shows the RMSD values along the trajectories towards the homopolymers of amino acids D, N, P, and Q. Notably, even close to the reference sequence, the RMSD values were around 18 Å, indicating significant structural divergence from the original conformation. As we diverge further from the reference sequence, the RMSD values increase even more for most trajectories. For the trajectory towards the homopolymer of D (Figure 4.1a), we observed a steady rise in RMSD values, followed by a decrease around step 150. The trajectories towards the homopolymers of amino acids P (Figure 4.1c) and Q (Figure 4.1d) exhibited a more pronounced increase in RMSD values, with a transition point around step 125, where the rate of increase became more rapid. In contrast, the trajectory towards homopolymer N (Figure 4.1b) displayed relatively stagnant RMSD values throughout the optimization process.

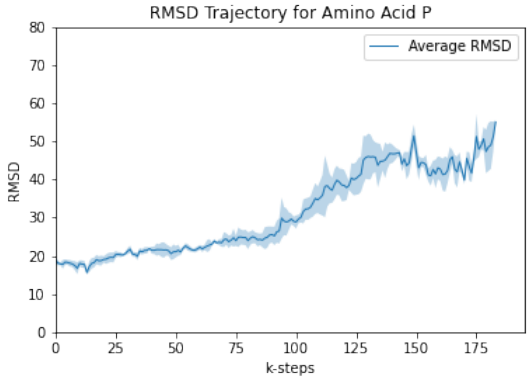
To quantify the structural similarity between the predicted structures from ESMFold and the parent structure, we calculated the TM score along the trajectories toward the homopolymers of all 20 amino acids. Figure 4.2 illustrates the TM scores along the trajectory towards two representative homopolymers, A and P. Notably, both trajectories exhibited consistently low TM scores across all homopolymer trajectories. For the trajectory toward the homopolymer of A (Figure 4.2a), the TM scores remained relatively constant, while a slightly decreasing trend was observed for the trajectory toward the homopolymer of P (Figure 4.2b). The persistently low TM scores suggest that the predicted models deviated significantly from the parent structure regarding topology and fold similarity. This observation highlights that it was challenging for ESM to accurately generate the three-dimensional structures for the variants of the ACE2 RBD.



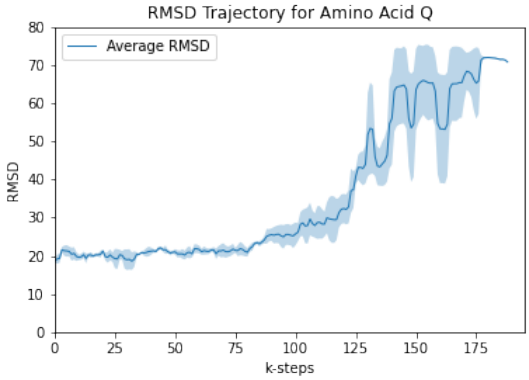
(a) Trajectory D



(b) Trajectory N



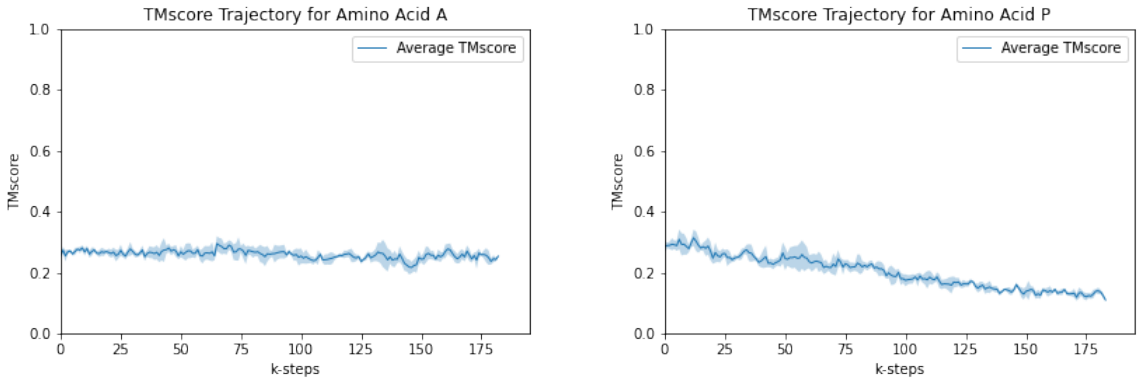
(c) Trajectory P



(d) Trajectory Q

Figure 4.1: RMSD (ESMFold) along Trajectories. RMSD values for trajectories towards homopolymers of amino acids D, N, P, and Q, predicted using ESMFold. Values are averaged over three repetitions with standard deviations shown.

To quantify the structural disorder of atoms in the predicted structures along the trajectories, we analyze the pLDDT along the trajectories towards the 20 homopolymers. Figure 4.3 shows the pLDDT along the trajectories toward four representative homopolymers, namely A, C, E, and S. We observed an initial increase in pLDDT values as the sequences diverged further away from the reference sequence, indicating an increase in atomic displacement and structural disorder. For specific trajectories, such as the trajectory towards the homopolymer E (Figure 4.3c), the increase in pLDDT was more pronounced, suggesting a steeper rise in structural disorder. In contrast, trajectories toward homopolymers C (Figure 4.3b) and S (Figure 4.3d) exhibited a less steep increase and generally lower pLDDT values. Interestingly, we observed a decrease in the pLDDT after step 140 for the trajectories toward the homopolymer

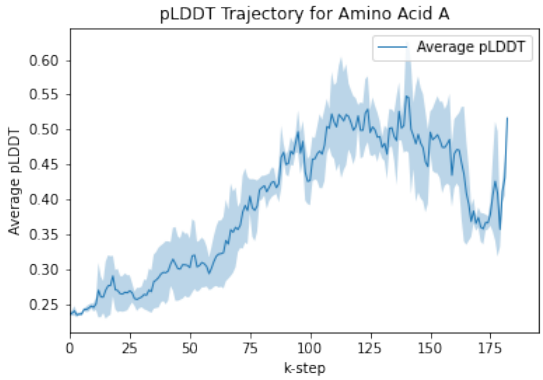


(a) Trajectory A **(b)** Trajectory P

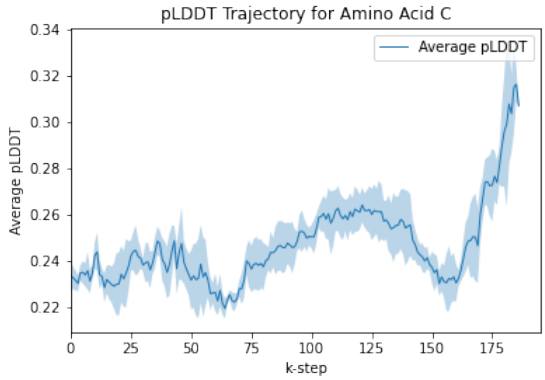
Figure 4.2: TM Score (ESMFold) along Trajectories. TM scores for trajectories towards homopolymers of amino acids A and P, predicted using ESMFold. Values are averaged over three repetitions, with standard deviations shown.

of A, suggesting a potential reduction in structural disorder at later stages of the trajectory. These observations highlight the diverse behavior of structural disorder along the trajectories toward different homopolymers. The varying patterns of pLDDT profiles provide insights into the intrinsic disorder and flexibility of the predicted structures, which can have implications for understanding their conformational dynamics, and potentially immunological features, like anti-body escape.

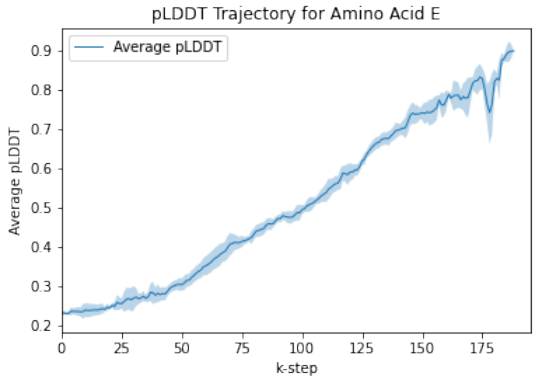
We analyzed the SASA metric to further elucidate the structural changes occurring along the trajectories. Figure 4.4 illustrates the SASA values along the trajectories toward the homopolymers of amino acids C, D, I, and V. We can observe distinct patterns in SASA profiles, suggesting diverse conformational transitions along different trajectories. For certain trajectories, such as those toward the homopolymers of amino acids C (Figure 4.4a) and D (Figure 4.4b), we observed an initial increase in SASA values until a specific point along the trajectory, followed by a subsequent decrease. This transition point occurred at step 90 for trajectory C, after which the SASA values increased again after step 150. For trajectory D, the transition point was observed at step 110. In contrast, the trajectory toward the homopolymer of amino acid V (Figure 4.3d) exhibited a general decrease in SASA values as the trajectory diverged from the reference sequence. The values decreased with some fluctuations until a rapid de-



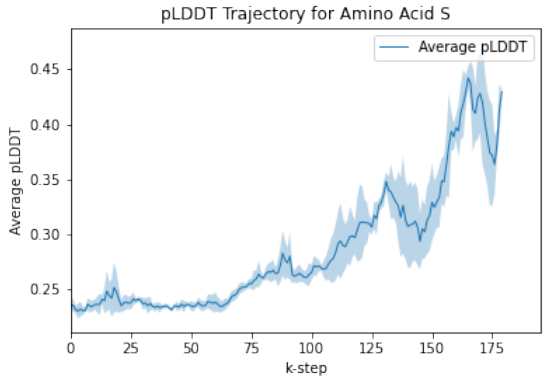
(a) Trajectory A



(b) Trajectory C



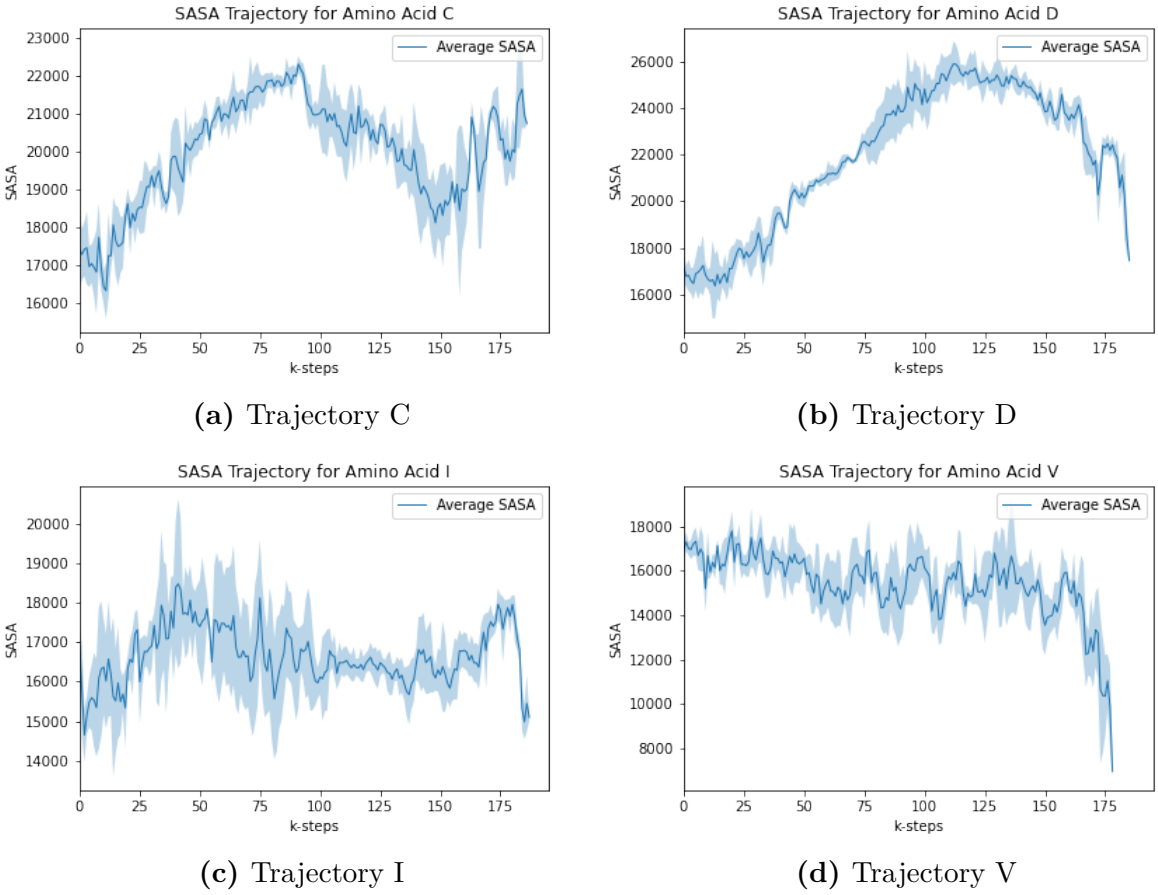
(c) Trajectory E



(d) Trajectory S

Figure 4.3: pLDDT (ESMFold) along Trajectories. pLDDT scores for trajectories towards homopolymers of amino acids A, C, E, and S, predicted using ESMFold. Values are averaged over three repetitions, with standard deviations shown.

crease was observed at approximately step 160. The trajectory toward the homopolymer of amino acid I (Figure 4.4c) displayed a more complex pattern with multiple fluctuations in SASA values. An initial increase was observed, followed by a decrease at step 50. Subsequently, the values increased again until step 150, after which a rapid decrease occurred at step 175. These diverse results suggest that distinct conformational dynamics govern the structural transitions along the trajectories toward different homopolymers, potentially involving changes in solvent exposure and surface area accessibility. The analysis of SASA, in conjunction with other structural metrics, provides valuable insights into the conformational landscape of protein structures.



(a) Trajectory C

(b) Trajectory D

(c) Trajectory I

(d) Trajectory V

Figure 4.4: SASA values (ESMFold) along Trajectories.SASA values for trajectories towards homopolymers of amino acids C, D, I, and V, predicted using ESMFold. Values are averaged over three repetitions, with standard deviations shown.

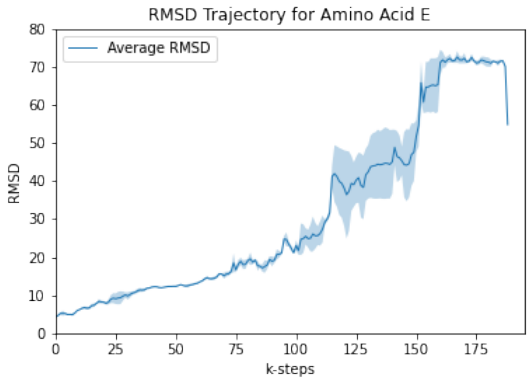
In Section 4.1.2, we integrate the results obtained from the various structural metrics and analyze their correlation. This approach aims to provide a comprehensive understanding of the

change in structural characteristics as we diverge further away from the reference sequence.

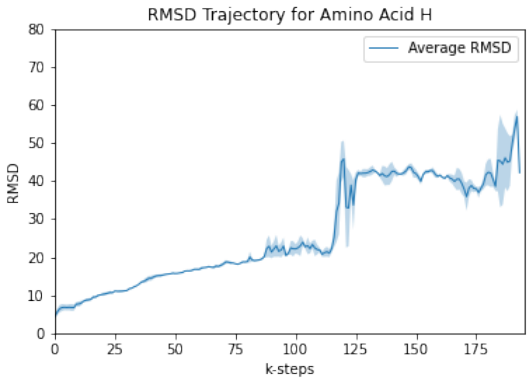
ColabFold

To quantify the structural changes from the reference sequence conformation along the trajectories, we also calculate the RMSD of the predicted models. Figure 4.5 shows the RMSD values on the trajectories towards the homopolymer of amino acids E, H, I, and V. In contrast to the more distinct transition points observed in the TM score, pLDDT, and SASA, the RMSD values gradually increased as the sequences diverged from the reference sequence. While RMSD values near the reference sequence ranged between 3 and 5 Å, indicating relatively minor deviations, they progressively increased, reaching values up to 80 Å closer to the homopolymers. The rate of the RMSD increase varied across trajectories, with some showing a more rapid rise, as exemplified by the trajectory towards the homopolymer of amino acid E (Figure 4.5a). Others, such as the trajectory towards amino acid C (Figure 4.5c), exhibited a more gradual increase, with RMSD values remaining relatively low until around step 110. The maximum RMSD values also differed among the trajectories, with some reaching values up to 80 Å, while others reached their maximum at around 40 Å. It is important to note that while RMSD values provide a quantitative measure of structural deviation, their results should also be integrated with other structural measures to gain a more reliable and comprehensive understanding of structural attributes. The RMSD value complements the following structural metrics by providing a direct measure of the conformational deviations from the parent structure. By integrating the RMSD with metrics such as TM score, pLDDT, and SASA, we better understand the structural transitions and conformational landscapes explored by the protein sequences along the 20 trajectories.

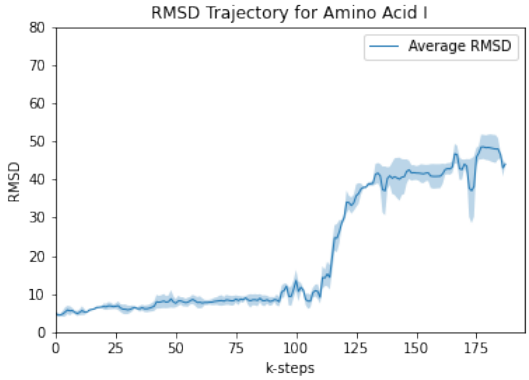
The structural similarity between the predicted models and the parent structure was quantified using the TM score metric along the trajectories toward homopolymeric sequences for all 20 amino acids. Figure 4.6 shows the TM score along the trajectories of 4 different trajectories, namely towards the homopolymer of amino acids D, F, I, and M. For sequences close to the reference sequence, ColabFold predictions exhibited high structural similarity, with TM-scores



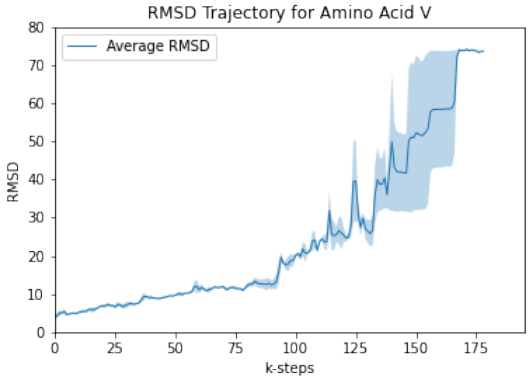
(a) Trajectory E



(b) Trajectory H



(c) Trajectory C

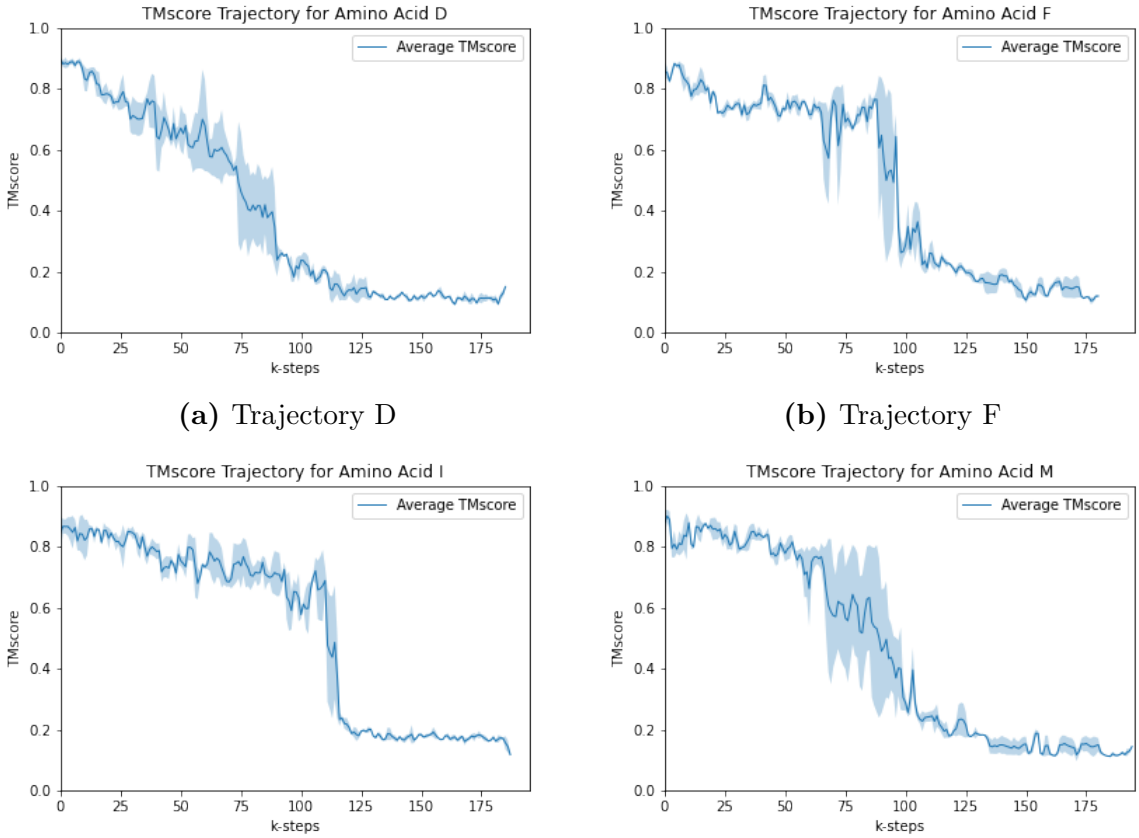


(d) Trajectory V

Figure 4.5: RMSD (ColabFold) along Trajectories. RMSD values for trajectories towards homopolymers of amino acids E, H, I, and V, predicted using ColabFold. Values are averaged over 2 repetitions, with standard deviations shown.

reaching up to 0.9, indicating a strong correspondence with the native fold. However, as the trajectories progressed further into sequence space, a notable decline in TM scores was observed across all 20 trajectories. Mainly, between steps 75 and 100, a more drastic decrease in TM scores was observed, with TM scores approaching values around 0.2. For trajectory I (Figure 4.6c), this change happens a bit further on the trajectory, around step 110. Closer to the reference sequence, the TM score reaches values up to 0.9, considered a very high similarity score. The mentioned transition point suggests a shift in the physiochemical properties of the sequences. It is also observed for all other 14 trajectories not shown here. While the overall trend was consistent, some trajectories had more significant fluctuations and higher standard deviations in TM scores, as seen on trajectories towards amino acid M, seen in figures 4.6d. These fluctuations might arise from the specific sequence compositions or amino

acid properties. After the transition region, the TM scores stabilize around values of 0.2 or lower for most trajectories, indicating a significant divergence from the parent structure. The analysis of TM score trajectories revealed a consistent pattern of higher structural similarity for sequences close to the reference sequence, followed by a significant decrease in values as sequences diverge further into sequence space.



(a) Trajectory D

(b) Trajectory F

(c) Trajectory I

(d) Trajectory M

Figure 4.6: TM Score (ColabFold) along Trajectories. TM Scores for trajectories towards homopolymers of amino acids D, F, I, and M, predicted using ColabFold. Values are averaged over 2 repetitions, with standard deviation shown.

To assess the reliability of the predicted structures along the trajectories, we analyzed the pLDDT scores, which estimate the confidence in the predicted local structure. Figure 4.9 depicts the average pLDDT values for representative trajectories towards the homopolymers of amino acids D, E, F, K, S, and W. In contrast to the trends we observed in TM scores, the pLDDT profiles showed more diverse patterns spanning the trajectories. For sequences that

are about up to 25 steps away from the reference sequence, the average pLDDT values range between 70 and 80, indicating accurate predictions. However, a notable decline in those values was observed as the trajectories progressed, with a transition point occurring between steps 75 and 100. This trend is consistent with the TM score analysis. Beyond this region, the pLDDT value profiles diverged when looking at trajectories for different amino acids. This reflects the varying degrees of confidence in the predicted structures for various amino acid trajectories. For trajectories toward amino acids E (Figure 4.9b), F (Figure 4.9c), and W (Figure 4.9d), we can observe that the pLDDT scores rebounded to values higher than those close to the reference sequence. This suggests increased confidence in predicting structures close to their homopolymers despite their divergence from their parents' three-dimensional structure. Homopolymeric sequences often exhibit simpler and more regular structural patterns that have been extensively studied. This simplicity may make it easier for algorithms such as ColabFold to predict homopolymers' structures accurately. Polylysine sequences are, for example, known to form stable alpha-helical structures. These structural properties arise from the physiochemical properties of specific amino acids. This also includes restricting the conformational space accessible to proteins when converging towards the homopolymer. When visualizing the three-dimensional structure of the homopolymer of K, we see a stable alpha-helical structure. This can be seen in Figure 4.7. In contrast, we see a less organized helical structure when visualizing the homopolymer of P. The structures were visualized with the 3D Viewer of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank [53].

Figure 4.7: Three-dimensional Structure of Homopolymeric Lysine sequence. This sequence shows the homopolymeric sequence of Lysine (K), which is 187 steps away from the reference sequence. We can observe a long helical structure.

In contrast, trajectories towards the homopolymer of amino acid D (Figure 4.9a) exhibited more stagnant and low pLDDT values after the transition point, with average scores around

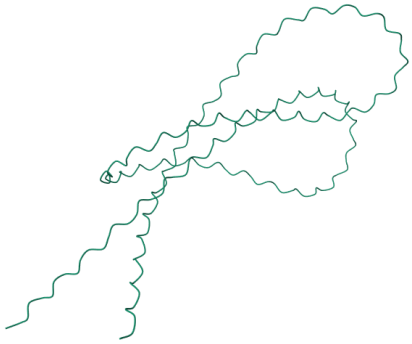
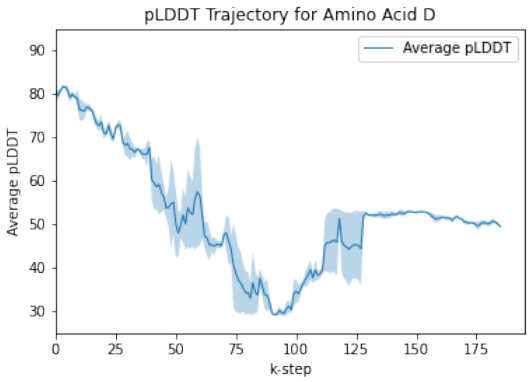


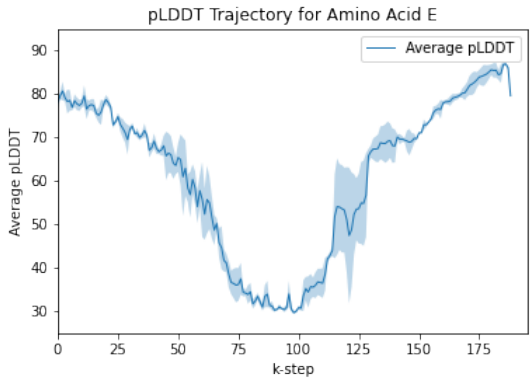
Figure 4.8: Three-dimensional Structure of Homopolymeric Proline sequence. This sequence shows the homopolymeric sequence of Proline (P), which is 184 steps away from the reference sequence

50. These low confidence scores indicate limitations in accurately modeling the structural ensembles for these sequences. Notably, the trajectory towards the homopolymer W sequence (Figure 4.9d) displayed a unique profile, with an initial decline followed by a peak around step 75, a subsequent decline until step 115, and then a sharp increase in pLDDT scores, reaching values close to 90 at the homopolymer. These results reveal a consistent transition point where confidence in predicted structure began to decline, resembling the observed decrease in structural similarity. However, the subsequent trends in pLDDT scores varied across different amino acid trajectories, highlighting the diverse challenges faced by ColabFold in accurately predicting the structural ensembles as the sequences diverged from the reference sequence. These findings undermine the importance of interpreting the pLDDT score alongside other measures to understand the reliability and limitations of protein structure prediction algorithms.

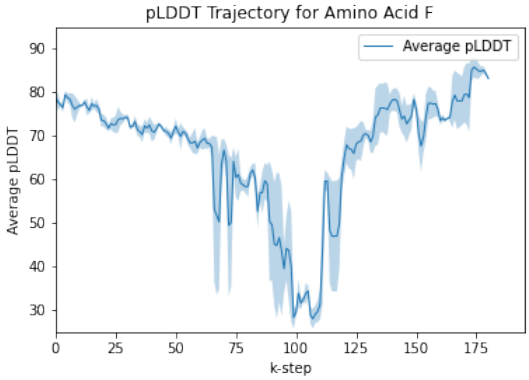
We analyzed the SASA metric to further shed light on the structural changes that occur along the trajectories. Figure 4.10 shows the SASA values for representative trajectories towards the homopolymer of amino acids D, I, R, and W. Consistent with the trends observed in TM score and pLDDT values, a transition point was observed between around steps 75 and 110, where SASA values exhibited a greater increase across all trajectories. This increase in values may suggest a loss of the folded structure, leading to greater residue exposure to the solvent environment. While the overall trend was similar, the maximum SASA values varied considerably



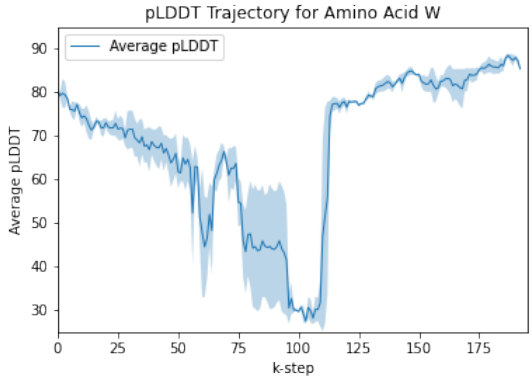
(a) Trajectory D



(b) Trajectory E



(c) Trajectory F

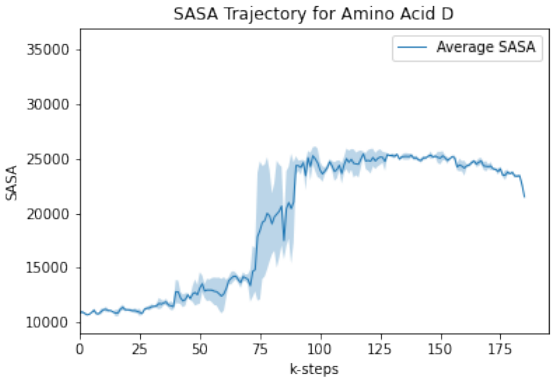


(d) Trajectory W

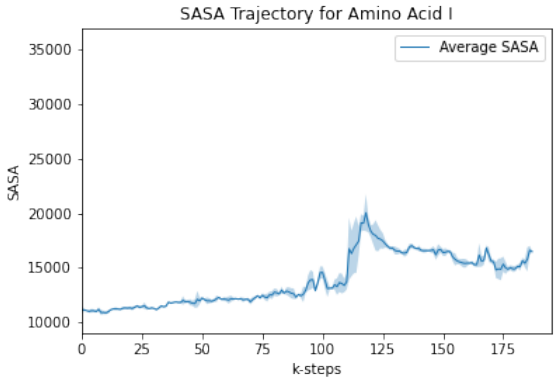
Figure 4.9: pLDDT (ColabFold) along Trajectories. pLDDT values for trajectories towards homopolymers of amino acids D, E, F, and W, predicted using ColabFold. Values are averaged over two repetitions, with standard deviations shown.

among different amino acid trajectories. The trajectory towards the homopolymers of amino acid I (Figure 4.10b) showed relatively lower maximum SASA values compared to the trajectory towards the homopolymer of amino acid R (Figure 4.10c). These differences may arise from their specific sequence properties and ability to adopt distinct structures. This trend along the trajectories can be attributed to several factors. As the protein structure deviates from its parent fold, previously buried regions, including hydrophobic cores, may become more exposed to solvents, contributing to higher SASA values. Furthermore, conformational rearrangements can expose previously buried surface areas, increasing the SASA values. After the transition point, some trajectories show slightly decreased SASA values while others maintain relatively constant values. For some trajectories, the values also steadily increase after the transition point. These divergent behaviors already highlight the diverse structural landscapes in protein

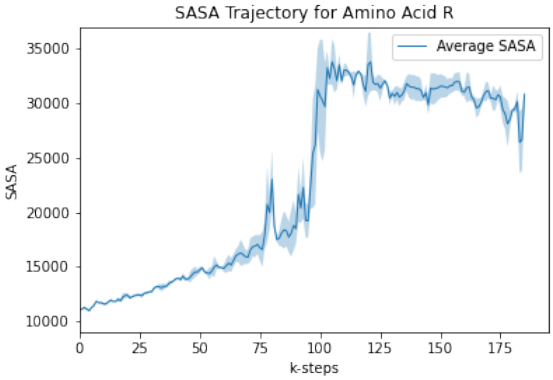
sequence space. However, interpreting SASA values in conjunction with other metrics, such as the RMSD, can provide more comprehensive insights into the conformational changes occurring along the trajectories. Even subtle structural deviations can already result in significant variations in solvent accessibility, undermining the importance of integrating multiple structural metrics for a comprehensive understanding of structural transitions.



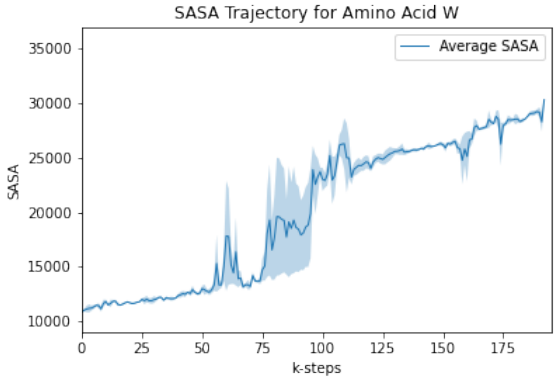
(a) Trajectory D



(b) Trajectory I



(c) Trajectory R



(d) Trajectory W

Figure 4.10: SASA values (ColabFold) along Trajectories. SASA values for trajectories towards homopolymers of amino acids D, I, R, and W. Values are averaged over two repetitions, with standard deviations shown.

4.1.2 Correlation Analysis

To gain further insights into the relationships between different structural metrics and their implications for the predicted models, we performed a correlation analysis by plotting various metrics against each other and then evaluating the resulting trends. From the 4 structural

metrics we can derive 6 possible pairwise correlation analyses, seen in Table 4.2.

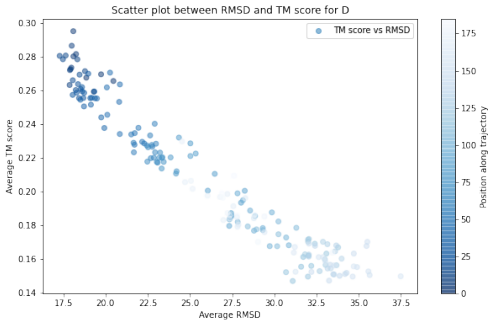
Table 4.2: Correlation Analyses. Overview of the pairwise correlation analyses and their interpretation.

Pairwise Correlation	Interpretation
RMSD vs TM Score	A high TM Score with low RMSD indicates high global and local structural similarity.
RMSD vs pLDDT	High correlation would suggest that pLDDT is a good predictor of structural deviations.
RMSD vs SASA	Strong correlation might indicate that local structural deviations significantly impact surface accessibility.
TM Score vs pLDDT	High correlation between TM Score and pLDDT would suggest that local accuracy predictions align well with structural similarity.
TM Score vs SASA	A strong correlation might suggest that structural changes significantly impact surface accessibility.
pLDDT vs SASA	Strong correlation might suggest that regions with high predicted accuracy tend to be more buried or exposed.

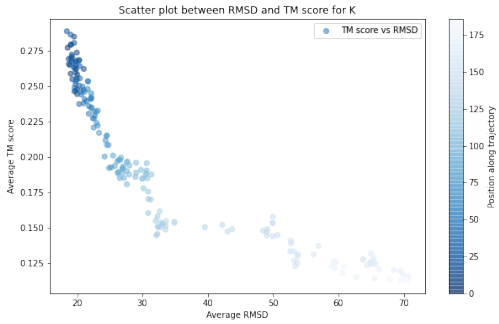
ESMFold

The analysis of the correlation between the RMSD and the TM score revealed a consistent trend across all 20 amino acid trajectories. Generally, as TM scores decreased, RMSD values increased, indicating an inverse relationship between the two metrics. For example, on the trajectory towards the homopolymer D (Figure 4.11a) and the trajectory towards the homopolymer K (Figure 4.11b), a clear inverse correlation was observed. As the TM-score decreased, the RMSD values exhibited a corresponding increase. However, in the trajectory towards the homopolymer K, a less pronounced trend was observed around step 120, where the RMSD value rose steeply despite a relatively smaller decrease in the TM-score. For certain trajectories, such as the one towards the homopolymer V (Figure 4.11c), the trend appeared more scattered, potentially due to the smaller range of TM-score and RMSD values compared to other trajectories. Overall, the analysis demonstrated an inverse correlation between RMSD and TM-score across all 20 amino acid trajectories, with higher RMSD values corresponding

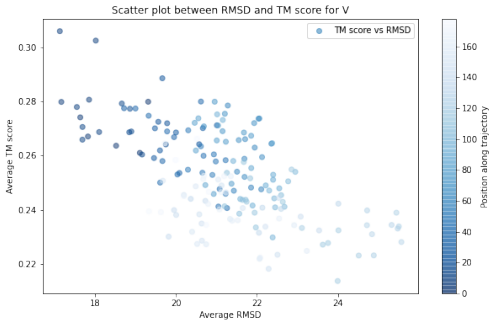
to lower TM-scores, indicating a decrease in structural similarity. However, the strength and linearity of this correlation varied among different trajectories, potentially influenced by the specific amino acid composition.



(a) Trajectory D



(b) Trajectory K

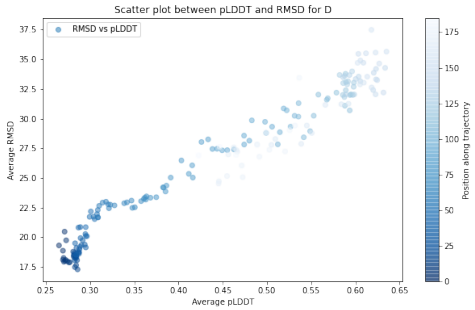


(c) Trajectory V

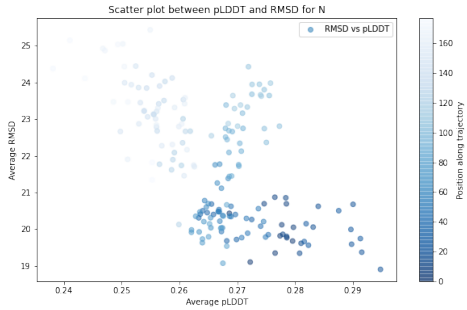
Figure 4.11: RMSD vs TM scores (ESMFold). RMSD values plotted against TM scores for sequences on three different trajectories towards the homopolymers for amino acids D, K, and V. The values are averaged over three repetitions.

Analyzing the correlation between RMSD and pLDDT provides insights into the flexibility and dynamics of protein structures. Figure 4.12 presents the results for three representative trajectories towards the homopolymers of amino acids D, N, and Q. For the trajectory towards the homopolymer of amino acid D (Figure 4.12a), a positive trend is observed until around step 160, indicating that as the pLDDT increases, the RMSD increases as well. After step 160, a negative trend emerges, suggesting that as the pLDDT decreases, the RMSD decreases correspondingly. For the trajectory towards the homopolymer of amino acid Q (Figure 4.12c), a positive trend is observed as we diverge from the reference sequence, implying that higher

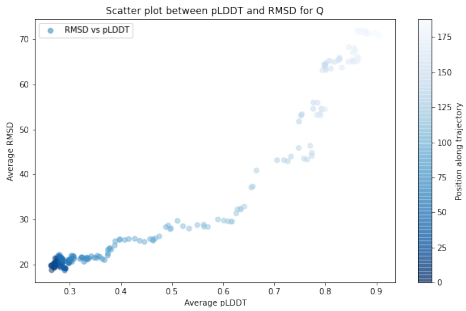
pLDDTs are associated with larger RMSD values. Conversely, for the trajectory towards the homopolymer of amino acid N (Figure 4.12b), the correlation appears more scattered, although the range of RMSD values and pLDDTs is smaller compared to the other trajectories.



(a) Trajectory D



(b) Trajectory N

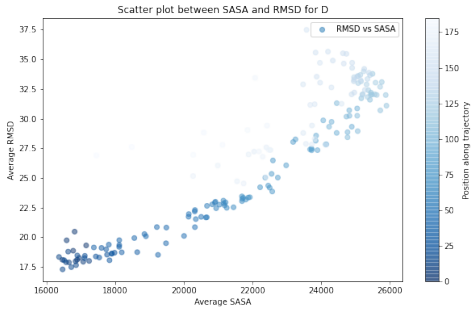


(c) Trajectory Q

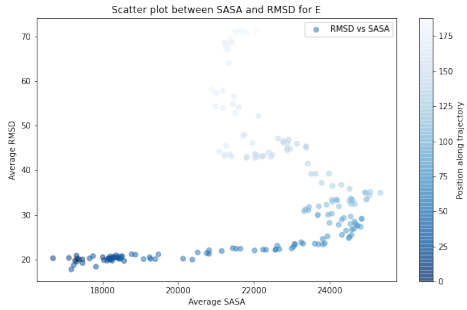
Figure 4.12: RMSD vs pLDDT (ESMFold). RMSD values plotted against the pLDDT values for sequences on three different trajectories towards the homopolymers for amino acids D, N, and Q. The values are averaged over three repetitions.

The correlation between the RMSD and SASA provides insights into the conformational changes and solvent exposure of protein structures on the trajectories. Figure 4.13 shows the results for three representative trajectories towards the homopolymers of amino acids D, E, and I. For the trajectory towards the homopolymer of amino acid D (Figure 4.13a), a positive correlation is observed until around step 160, indicating that as SASA increases, RMSD increases correspondingly. However, after step 160, a negative trend emerges, with both SASA and RMSD values decreasing. A more bifurcated trend is observed in the trajectory towards the homopolymer of amino acid E (Figure 4.13b). Initially, SASA values increase while RMSD values exhibit a slight increase. Subsequently, a point is reached where SASA values decrease,

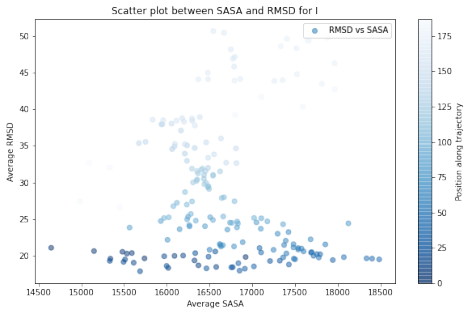
but RMSD values continue to increase. In contrast, the correlation appears more scattered for the trajectory towards the homopolymer of amino acid I (Figure 4.13c). Initially, RMSD values stagnate while SASA values fluctuate. However, as the trajectory progresses, RMSD values rise while SASA values remain within a specific range. The varying trends observed for different amino acids underscore the influence of sequence composition on the structural dynamics and solvent accessibility of the resulting sequences.



(a) Trajectory D



(b) Trajectory E

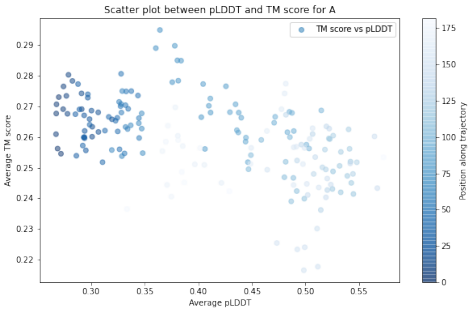


(c) Trajectory I

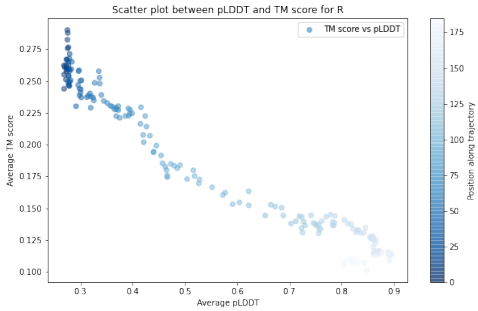
Figure 4.13: RMSD vs SASA (ESMFold). RMSD values plotted against the SASA values for sequences on three different trajectories towards the homopolymers for amino acids D, E, and I. The values are averaged over three repetitions.

We analyzed the correlation between the TM-score and pLDDT for three representative trajectories towards A, R, and Y homopolymers to investigate the relationship between structural similarity and flexibility. We generally observe an inverse trend, indicating that the pLDDT increases as the TM score decreases. A lower TM score coupled with a higher pLDDT may suggest that the algorithm struggled with the flexible part. The trajectory towards the R homopolymer (Figure 4.14b) exhibited a clear inverse relationship between TM-score and pLDDT.

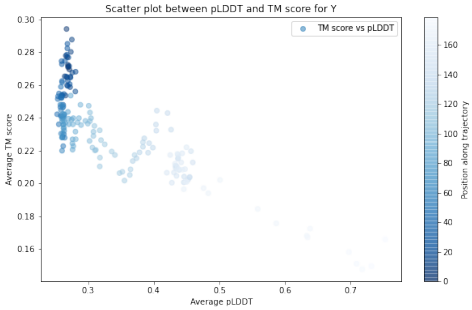
As the sequence diverged further from the parent, the TM-scores decreased while the pLDDTs increased, implying that the algorithm struggled to predict the flexible regions accurately. Figure 4.14a shows a similar trend, albeit with a minor increase in TM scores, leaving the points more scattered. In contrast, for the trajectory towards the Y homopolymer (Figure 4.14c), the pLDDTs remained relatively stable initially, while the TM-scores decreased. After step 120, an increase in pLDDTs was observed, accompanied by fluctuations in the TM-score. These results demonstrate that an inverse correlation between TM-score and pLDDT can indicate challenges in predicting flexible regions, but the relationship is not always straightforward. The metrics should be integrated with other verification methods to assess the predictions comprehensively.



(a) Trajectory A



(b) Trajectory R

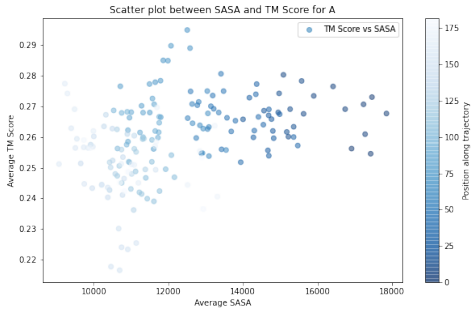


(c) Trajectory Y

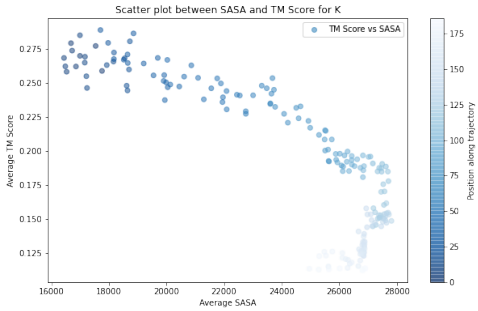
Figure 4.14: TM score vs pLDDT (ESMFold). TM Scores plotted against the pLDDT values for sequences on three different trajectories towards the homopolymers for amino acids A, R, and Y. The values are averaged over three repetitions.

Figure 4.15 shows the correlation between the TM Score and SASA for three representative trajectories towards the homopolymers of amino acids A, K, and R. The observed trends vary across the different trajectories, highlighting the complexity of protein folding dynamics. For

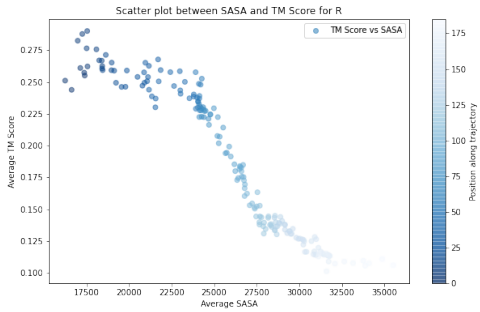
the trajectory towards the homopolymer of R (Figure 4.15c), a clear inverse correlation is evident, indicating that as the TM Score decreases, the SASA values increase. This suggests that structural deviations from the reference conformation are accompanied by an increase in surface exposure. A similar trend is observed for the trajectory towards the homopolymer of K (Figure 4.15b). However, towards the end of this trajectory, a bifurcated trend emerges, where decreasing TM Scores are associated with both increasing and decreasing SASA values. This bifurcation may indicate the presence of alternative structural conformations with varying degrees of surface exposure. In contrast, the trajectory towards the homopolymer of A (Figure 4.15a) exhibits overall decreasing SASA values, while the TM Scores are more scattered.



(a) Trajectory A



(b) Trajectory K

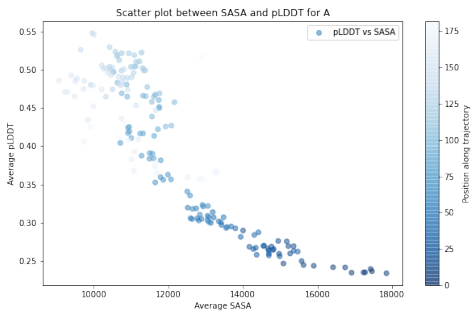


(c) Trajectory R

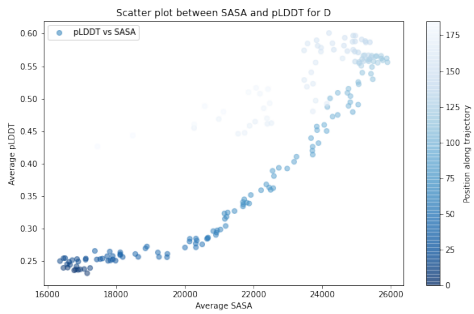
Figure 4.15: TM score vs SASA (ESMFold). TM Scores plotted against the SASA values for sequences on three different trajectories towards the homopolymers for amino acids A, K, and R. The values are averaged over three repetitions.

Figure 4.16 shows the correlation between pLDDT scores and SASA values for three representative trajectories towards the homopolymers of A, D, and E. These trajectories exhibit distinct patterns, reflecting the complex interplay between local structural accuracy and sur-

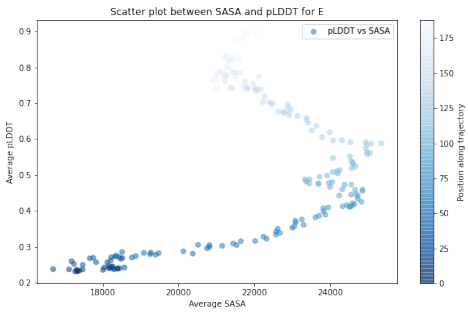
face exposure during protein folding. For the trajectory towards the homopolymer of A (Figure 4.16a), an inverse relationship is observed, wherein decreasing SASA values correspond to increasing pLDDT values. This suggests that as the structure becomes more compact, the local structural predictions become more confident. In contrast, the trajectory towards the homopolymer of D (Figure 4.16b) demonstrates a positive correlation, with increasing pLDDT scores corresponding to increasing SASA values. The trajectory towards homopolymer of D (Figure 4.16c) exhibits a more complex, bifurcated trend. Initially, increasing pLDDT values correlate with increasing SASA values, similar to the trajectory of D. However, approximately midway through the trajectory, this relationship inverts, with increasing pLDDT values corresponding to decreasing SASA values. This bifurcation may suggest a transition point in the folding process, where the structure shifts from favoring solvent exposure to adopting a more compact conformation while maintaining high local structural confidence.



(a) Trajectory A



(b) Trajectory D



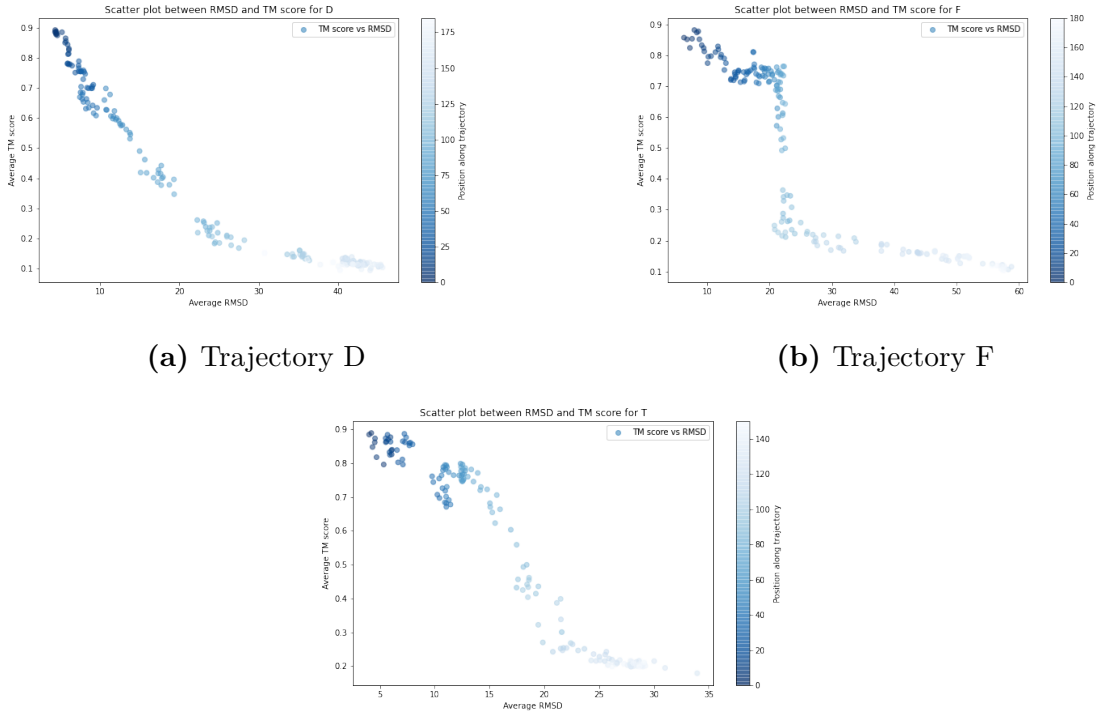
(c) Trajectory E

Figure 4.16: pLDDT vs SASA (ESMFold). pLDDT scores against SASA values for sequences on three different trajectories towards the homopolymers for amino acids A, D, and E. The values are averaged over three repetitions.

ColabFold

Figure 4.17 shows the scatter plots of RMSD against TM score for representative trajectories towards the homopolymers of amino acids D, F, and T. Across all 20 amino acid trajectories, a clear negative correlation was observed between RMSD and TM score, indicating that as the structural similarity (TM score) decreased, the deviation from the reference structure (RMSD) increased. While we observed this negative correlation for all trajectories, the strength of the correlation varied. For example, on the trajectory towards the homopolymer of F (Figure 4.17b), the negative correlation became less strong as TM scores approached values around 0.2, where a more drastic increase in RMSD was observed with only a slight decrease in TM score. This observation implies that ColabFold was more successful in accurately capturing the three-dimensional structures of proteins closer to the reference sequence, with fewer conformational changes or more rigid-body movements while preserving the overall fold of the reference sequence. The observed negative correlation between RMSD and TM score aligns with the expected relationship between these metrics, where higher structural deviations correspond to lower structural similarity. However, the varying strengths of this correlation across different trajectories highlight the diverse structural landscapes of protein sequences as they diverge from the reference sequence. It further gives insights into the accuracy and limitations of the ColabFold predictions.

Figure 4.18 shows the scatter plots of RMSD against pLDDT for representative trajectories towards the homopolymers of amino acids I, R, and T. The scatter plots in Figure 4.18 reveal a consistent trend across all 20 trajectories, albeit differently pronounced for different trajectories. A bifurcation trend is observed for the trajectory towards the homopolymer of I (Figure 4.18a), indicating a change in correlation at some point on the trajectory. We observe that at first, a decrease in pLDDT value correlates with a slight rise in RMSD values, while after some point, this trend changes, and an increase in pLDDT values correlates with an increase in RMSD values. Similarly, for the trajectory towards the homopolymer of R (Figure 4.18b) and T (Figure 4.18c), we can observe a bifurcation trend in the data, although the



(a) Trajectory D

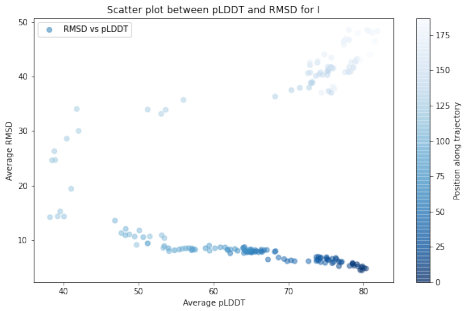
(b) Trajectory F

(c) Trajectory T

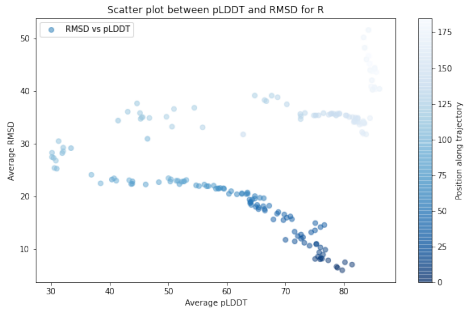
Figure 4.17: RMSD vs TM Scores (ColabFold). RMSD values plotted against the TM scores for sequences on three different trajectories towards the homopolymers for amino acids D, F, and T. The values are averaged over two repetitions.

rise in RMSD for trajectory R after the transition point is not as pronounced compared to trajectory I. For both trajectories R and T, we observe a steadier rise in RMSD values for the first half of the trajectory. These bifurcation trends were observed for all 20 homopolymer trajectories, although with differences in the absolute values. These observations suggest that the relationship between RMSD and pLDDT may be influenced by the structural properties of the target homopolymer. Results highlight that while pLDDT is a valuable local confidence measure, more than a high pLDDT score is needed to guarantee an accurate overall structure prediction.

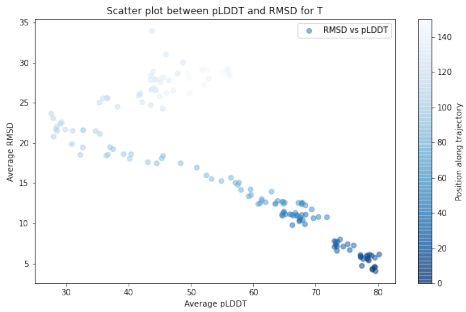
Figure 4.19 shows the relationship between the SASA values and the RMSD for representative trajectories towards homopolymers A, N, and R. The scatter plots in Figure 4.19 reveal a positive trend across all 20 trajectories, indicating that as the SASA values increase, the RMSD also increases. Two distinct clusters of data points can be observed for the trajectory towards



(a) Trajectory I



(b) Trajectory R

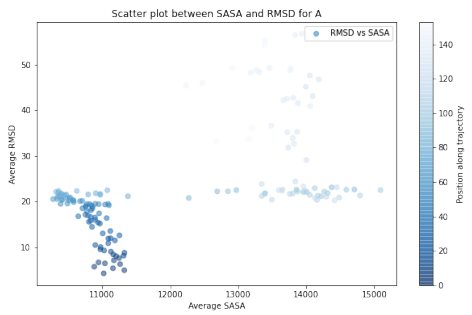


(c) Trajectory T

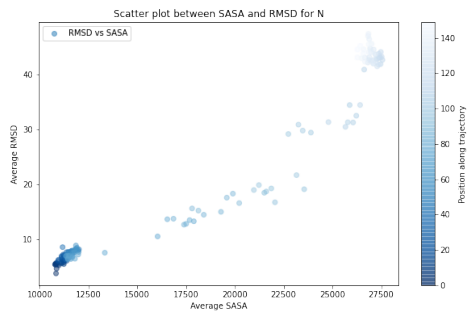
Figure 4.18: RMSD vs pLDDT scores (ColabFold). RMSD values plotted against the pLDDT scores for sequences on three different trajectories towards the homopolymers for amino acids I, R, and T. The values are averaged over two repetitions.

homopolymer N (Figure 4.19b). One cluster exhibits lower RMSD and SASA values, while the other displays higher RMSD and SASA values. Additionally, a few outliers between these clusters exhibit a more positive trend, where higher SASA values correspond to higher RMSD values. The trajectories towards the homopolymer of A (Figure 4.19a) and R (Figure 4.19c also show a positive correlation, albeit less pronounced. The data points are more dispersed; some exhibit similar RMSD values despite higher SASA values. These observations suggest that as the structural deviation from the target homopolymer increases, as indicated by higher RMSD values, the solvent-accessible surface area of the protein also increases. The positive correlation between RMSD and SASA can be attributed to the conformational changes and unfolding events that occur as the protein deviates from the reference sequence's structure, increasing the exposed surface area. Notably, the strength of the correlation can vary among the different homopolymer trajectories, potentially influenced by factors such as amino acid composition and their properties, as well as the degree of deviation from the reference sequence.

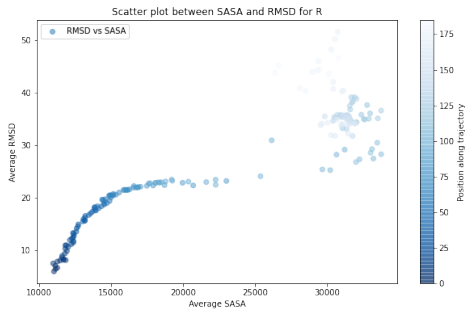
The presence of distinct clusters and more scattered values further highlights the complexity of the relationship between RMSD and SASA.



(a) Trajectory I



(b) Trajectory R

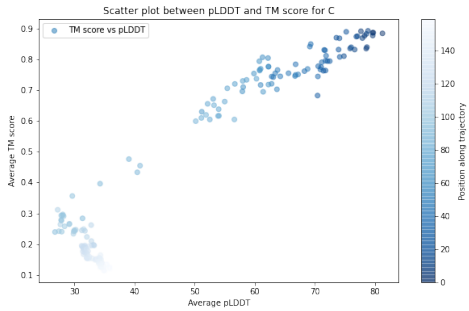


(c) Trajectory T

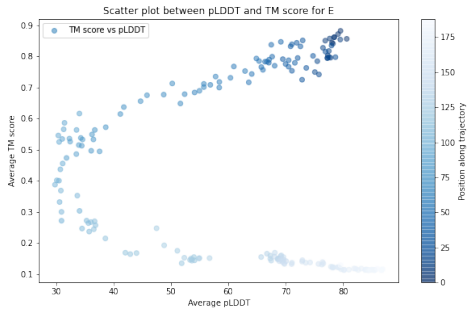
Figure 4.19: RMSD vs SASA scores (ColabFold). RMSD values plotted against the SASA scores for sequences on three different trajectories towards the homopolymers for amino acids A, N, and R. The values are averaged over two repetitions.

To further investigate the relationship between the TM score and the pLDDT score, we analyzed scatter plots for representative trajectories towards the homopolymers of C, E, and W, as seen in Figure (Figure 4.20). For all trajectories, we can observe a bifurcation trend in the data, suggesting that at some point on the trajectory, there is a change in the correlation between pLDDT and RMSD. This trend is pronounced differently for different trajectories. For the trajectory towards the homopolymer of C (Figure 4.20a), we observe this bifurcation trend towards the end of the trajectory, indicating that while at first higher TM scores corresponded to higher pLDDT scores, at some point on the sequence it is the other way around. This transition point happens even earlier for trajectories towards the homopolymers of E (Figure 4.20b) and W 4.20c). As depicted in Figure 4.20b, the first half of the trajectory exhibits a

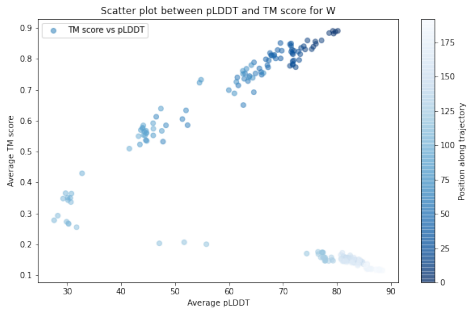
positive trend, indicating that higher TM scores correspond to higher pLDDT scores. However, this trend is the opposite after the transition point, meaning that this region shows increasing pLDDT values with decreasing TM scores. This trend is also observed for the trajectory towards the homopolymer of W 4.20c). These observations can serve as a valuable analysis to assess the reliability and limitations of the confidence metrics.



(a) Trajectory C



(b) Trajectory E

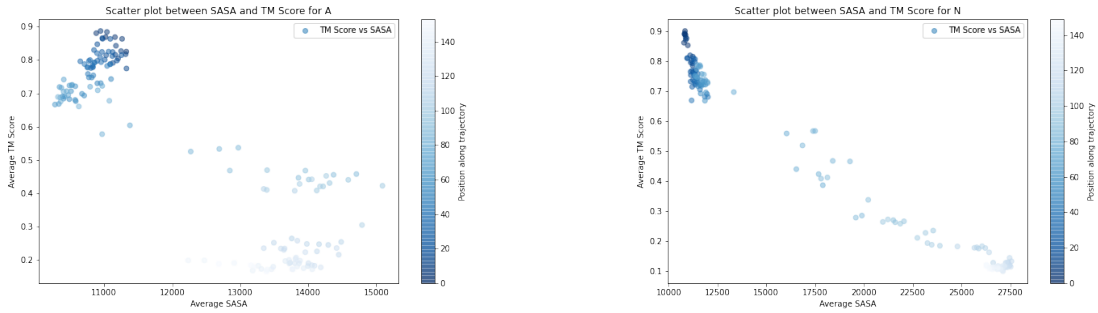


(c) Trajectory W

Figure 4.20: TM Scores vs pLDDT (ColabFold). TM Scores plotted against the pLDDT scores for sequences on three different trajectories towards the homopolymers for amino acids C, E, and W. The values are averaged over two repetitions.

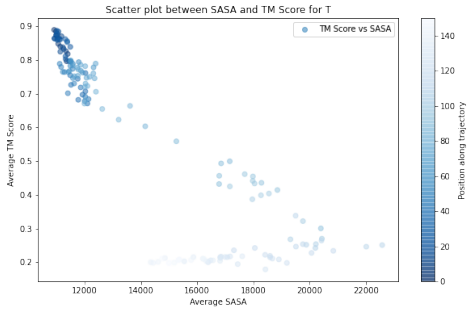
Figure 4.21 shows the correlation between TM Score and SASA values for three representative trajectories towards the homopolymer of A, N, and T. These trajectories exhibit distinct patterns, reflecting the complex relationship between global structural similarity and surface exposure during protein folding. For the trajectory towards the homopolymer of N (Figure 4.21b), an inverse trend is observed, wherein decreasing TM Scores correspond to increasing SASA values. A similar trend is initially observed for the trajectory towards the homopolymer of T (Figure 4.21c). However, towards the end of the trajectory, a notable shift occurs, characterized by decreasing SASA values while TM Scores remain relatively constant. In contrast, the

trajectory towards the homopolymer of A (Figure 4.21a) exhibits a more complex, multi-phase trend. Initially, decreasing TM Scores correlate with decreasing SASA values. Approximately midway through the trajectory, this relationship inverts, with decreasing TM Scores corresponding to increasing SASA values. Interestingly, towards the end of the trajectory, the inverse trend re-emerges. These observations suggest that the folding process may involve multiple phases with varying structural characteristics, even within a single amino acid trajectory.



(a) Trajectory A

(b) Trajectory N

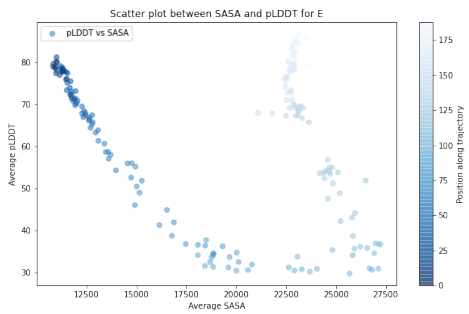


(c) Trajectory T

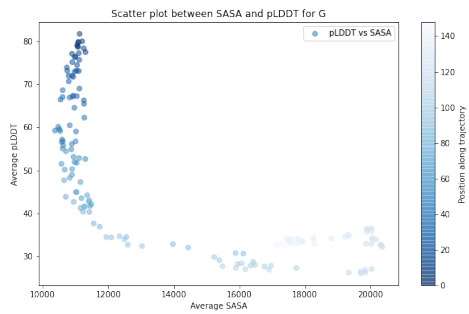
Figure 4.21: TM Scores vs SASA (ColabFold). TM Scores plotted against the SASA values for sequences on three different trajectories towards the homopolymers for amino acids A, N, and T. The values are averaged over 2 repetitions.

Lastly, Figure 4.22 shows the correlation of pLDDT scores and SASA values for three representative trajectories towards the homopolymers of E, G, and T. The trajectory towards homopolymer E (Figure 4.22) demonstrates a bifurcated trend. Initially, there is a decline in pLDDT scores accompanied by an increase in SASA value. However, approximately midway through the trajectory, this trend reverses, with pLDDT values increasing and SASA values slightly decreasing. This bifurcation may indicate a transition point in the folding process. In contrast, the trajectory towards the homopolymer of T (Figure 4.22c) exhibits an less pro-

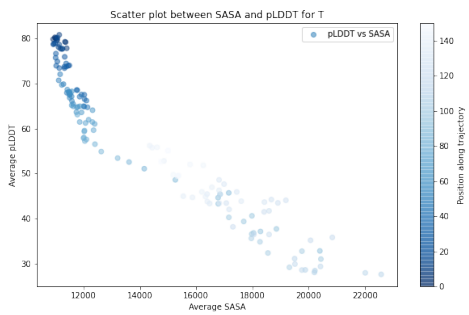
nounced bifurcated trend. Initially, decreasing pLDDT scores correspond to increasing SASA values. However, at a certain point along the trajectory, this relationship inverts, with pLDDT values increasing while SASA values decrease. The trajectory towards the homopolymer of G (Figure 4.22) exhibits relatively constant SASA values while pLDDT values decrease. However, towards the end of the trajectory, a dramatic increase in SASA values is observed while pLDDT values remain constant. These varying results highlight, that high local structural confidence does not always directly correlate with a certain surface exposure.



(a) Trajectory E



(b) Trajectory G



(c) Trajectory T

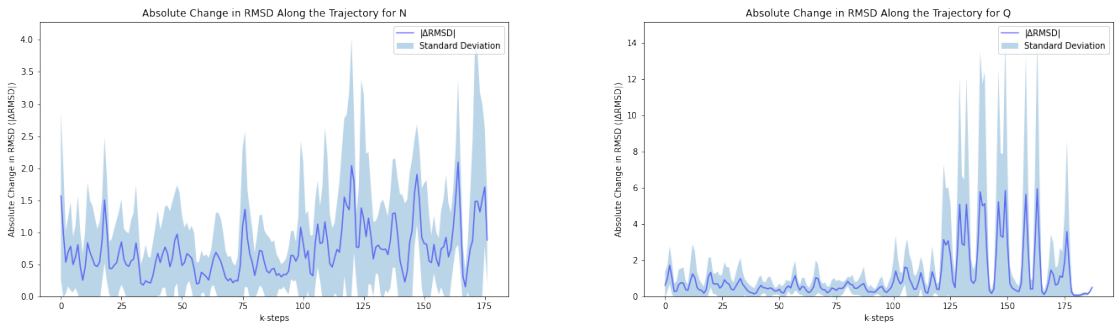
Figure 4.22: pLDDT vs SASA (ColabFold). This figure shows the pLDDT scores plotted against the SASA values for sequences on 3 different trajectories towards the homopolymers for amino acids E, G, and T. The values are averaged over 2 repetitions.

4.1.3 Sequential Changes

To gain further insight into the changes along the trajectories in sequence space, we investigated the change ($\triangle$) in key structural metrics. This analysis aims to identify steps of rapid or abrupt, disproportionately large changes that could give further insights into the protein folding pattern in sequence space.

ESMFold

We analyzed the change in RMSD (△RMSD) along a trajectory. Figure 4.23 shows the △RMSD for two representative trajectories towards the homopolymers of N and Q. For most trajectories we observe larger peaks of △RMSD after step 125 on the trajectories, as seen for the trajectory towards homopolymer Q (Figure 4.23b). In contrast, some trajectories experience overall more change in △RMSD across the entire trajectory, as seen for trajectory towards homopolymer N (Figure 4.23b). This behavior implies that the structural deviations were not concentrated at specific transition points but rather occurred gradually and consistently throughout the trajectory.

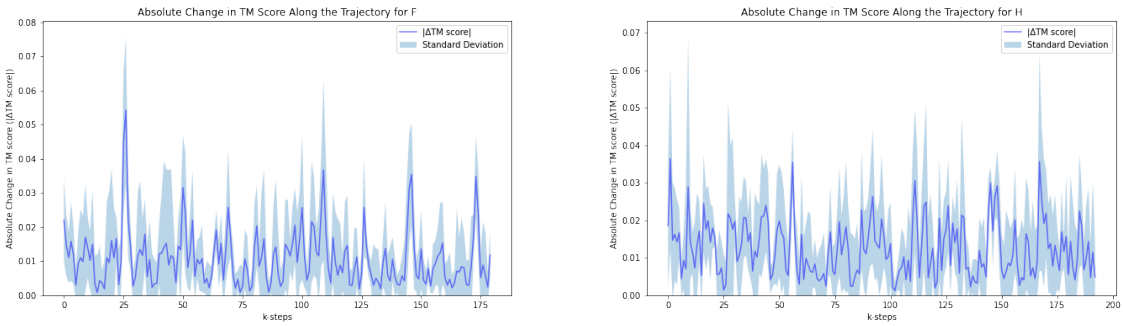


(a) Trajectory N **(b)** Trajectory Q

Figure 4.23: Change in RMSD (ESMFold) along trajectory (△RMSD). Change of RMSD along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids N and Q. The values are averaged over three repetitions, and the standard deviation is plotted.

The structural deviation along the trajectories was quantified by examining the change in TM score (△TM score). Figure 4.24 shows the △TM score towards the homopolymers of F and H. The trajectories exhibited consistent fluctuations in the △TM score, which can be attributed to the relatively low and consistent TM scores calculated along all trajectories. For instance, the trajectory towards the homopolymer F (Figure 4.24a) displayed a peak around step 25, indicating a more substantial structural deviation along the trajectory. However, the overall profile was characterized by frequent fluctuations in the △TM score. Similarly, the trajectory towards the homopolymer H (Figure 4.24b) exhibited a fluctuating pattern in the △TM score.

The observed fluctuations in the △TM score along the trajectories suggest that the structural deviations from the initial conformation were relatively minor and distributed throughout the trajectory paths rather than concentrated at specific transition points.

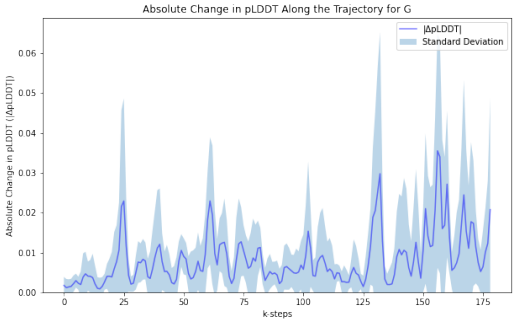


(a) Trajectory F **(b)** Trajectory H

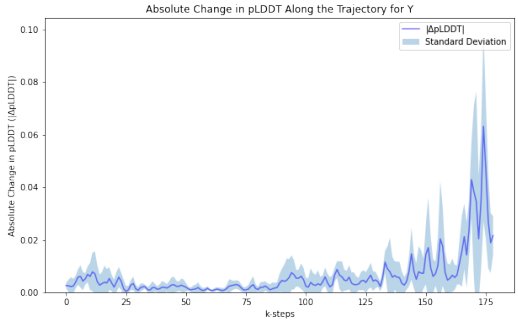
Figure 4.24: Change in TM score (ESMFold) along trajectory (△TM score). Change of TM score along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids F and H. The values are averaged over three repetitions, and the standard deviation is plotted.

The change in confidence of the predictions was analyzed by examining the change in pLDDT along the trajectories. As seen in Figure 4.25, the analysis revealed no clear, uniform trend. While some trajectories exhibited fluctuations and occasional peaks in the △pLDDT, no distinct transition points were observed consistently for all 20 amino acids. For instance, the trajectory towards the homopolymer Y (Figure 4.25b) displayed relatively minor fluctuations until approximately step 160, after which a more substantial △pLDDT was observed. In contrast, the trajectory towards the homopolymer G (Figure 4.25a) exhibited fluctuations along the trajectory with no clear peak.

By examining the change in SASA (△SASA) along the trajectories, we can identify potential points with a more drastic exposure or burial of solvent-accessible regions. Figure 4.26 shows the trends of △SASA towards the homopolymers of F, M, S, and W, suggesting different solvent exposure patterns associated with each homopolymer trajectory. For the trajectory towards the homopolymer F (Figure 4.26a), a significant peak in △SASA was observed around



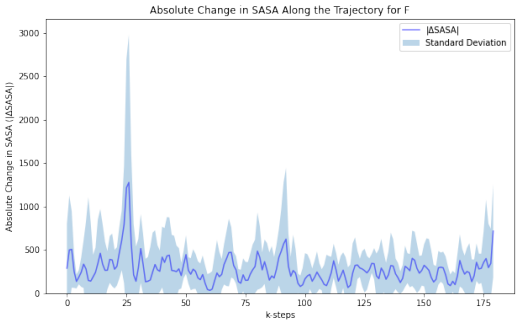
(a) Trajectory G



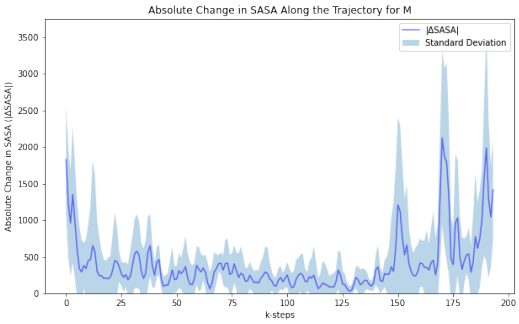
(b) Trajectory Y

Figure 4.25: Change in pLDDT (ESMFold) along trajectory (△pLDDT). This Figure shows the change of pLDDT along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids G and Y. The values are averaged over three repetitions, and the standard deviation is plotted.

step 25, indicating a substantial change in solvent exposure at that point along the trajectory. After this point, the △SASA exhibited minor fluctuations. In contrast, the trajectory towards the homopolymer M (Figure 4.26b) displayed a significant change in △SASA immediately at the beginning of the trajectory, followed by another notable change after step 150, suggesting increased solvent exposure of specific regions as the trajectory approached the homopolymer. These patterns provide insights into the conformational rearrangements and potential exposure or burial of solvent-accessible regions during the folding process, which can have implications for understanding the structural dynamics of the protein.



(a) Trajectory F



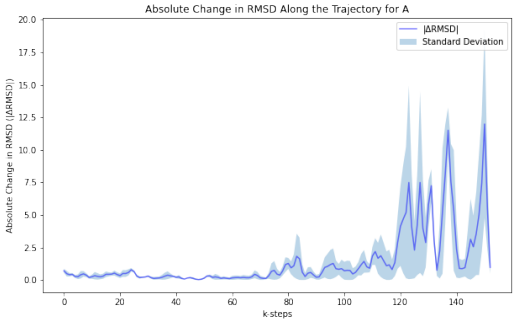
(b) Trajectory M

Figure 4.26: Change in SASA (ESMFold) along trajectory (△SASA). Change of SASA along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids F and M. The values are averaged over three repetitions, and the standard deviation is plotted.

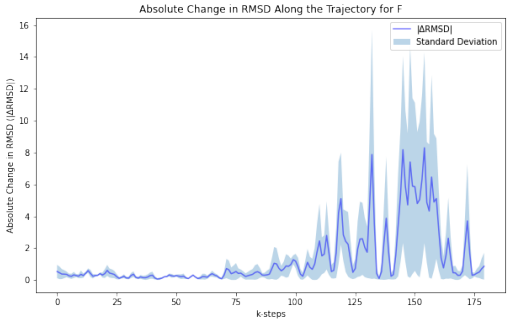
ColabFold

As expected after the analysis above, some trajectories exhibit more drastic changes than others. Analyzing the change in RMSD (△RMSD) along a trajectory can provide valuable insights into protein dynamics and conformational changes. Figure 4.27 shows △RMSD along the trajectories for four representative amino acids, namely A, F, K, S. For the trajectory towards the homopolymer of A (Figure 4.27a), we observed minimal to no change in △RMSD until step 80, after which more fluctuations were evident until a higher magnitude of △RMSD was observed beyond step 120. A similar pattern was observed for the trajectory towards the homopolymer of F (Figure 4.27b). In contrast, the trajectory towards the homopolymer of K (Figure 4.27c) exhibited smaller magnitudes of △RMSD in the steps closer to the reference sequence, followed by a period of higher changes. Subsequently, multiple high peaks of △RMSD were observed, suggesting significant conformational transitions. The trajectory towards the homopolymer of S (Figure 4.27d) shows a distinct pattern, with an initial period of lower △RMSD, followed by a short period of higher peaks and then a subsequent period of lower △RMSD. Notably, the magnitude of △RMSD was generally smaller closer to the reference sequence and experienced more fluctuations as the trajectories neared the homopolymers. This observation suggests that the conformational changes became more pronounced as we diverged from the reference sequence. Like the other structural metrics, the analysis of △RMSD along the trajectories revealed distinct patterns of changes in RMSD. The magnitude and point on the trajectories of these changes varied across the different homopolymer trajectories, highlighting the influence of sequence composition on the folding dynamics and the potential for identifying transitions during the folding process.

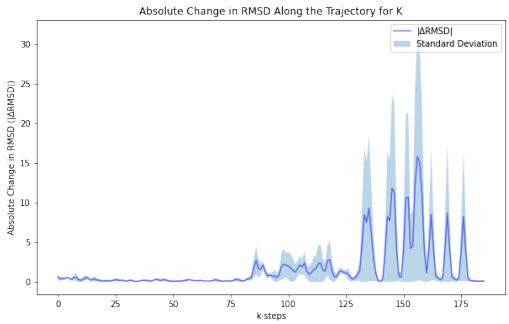
Figure 4.28 depicts the △TM score along the trajectories to homopolymers of F, P, R, and W. The △TM score analysis provides insights into the dynamics of structural similarity during the folding process. For the trajectory towards the homopolymer of F (Figure 4.28a), we observed minor peaks before the highest magnitude of △TM score occurred around step 90, after which the magnitude flattened out. For the trajectory towards homopolymer P (Figure 4.28b), we observe a peak in △TM score around the same step as F, but with more fluctua-



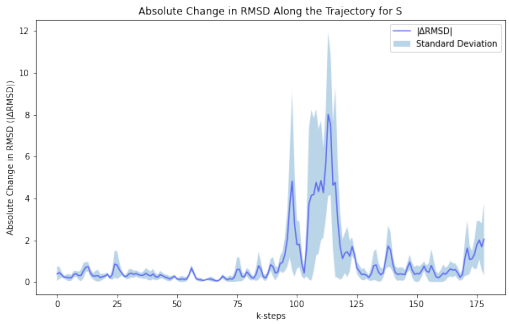
(a) Trajectory A



(b) Trajectory F



(c) Trajectory K

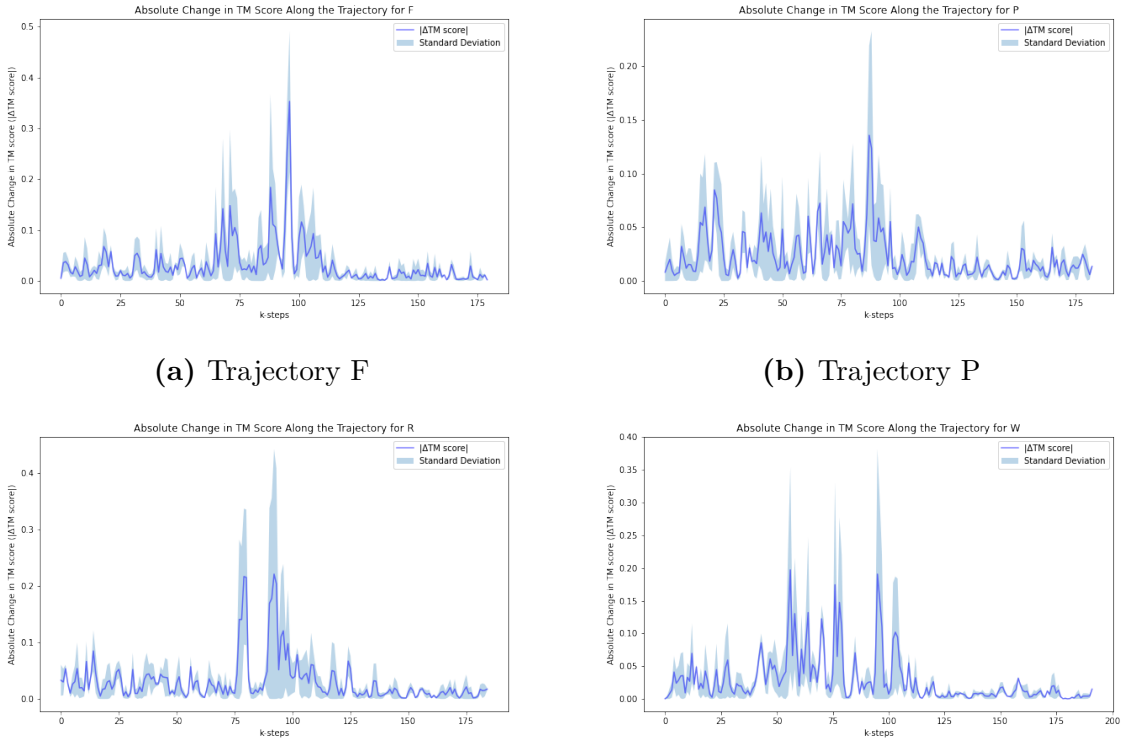


(d) Trajectory S

Figure 4.27: Change in RMSD (ColabFold) along trajectory (△RMSD). Change of RMSD along the trajectories for sequences on four different trajectories towards the homopolymers for amino acids A, F, K, and S. The values are averaged over two repetitions, and the standard deviation is plotted.

tions preceding it. After the peak, the magnitude of △TM score flattened out to some extent, indicating a period of reduced structural change. Looking at the case of the trajectory towards the homopolymer of R (Figure 4.28c), two distinct peaks in △TM score were observed around steps 80 and 90, suggesting two periods of substantial structural rearrangement. Following these peaks, the magnitude of △TM score flattened out again. For the trajectory towards the homopolymer of W (Figure 4.28d), we observed more fluctuations in the △TM score, with three significant peaks at steps 55, 75, and 90. This pattern suggests that there may be multiple periods of substantial structural changes followed by periods of relative stability. The △TM score analysis along the trajectories revealed distinct patterns of structural change, with varying magnitudes and points on the trajectories of occurring peaks across the different trajectories towards various homopolymers. These observations highlight the influence of sequence composition on the folding dynamics and the potential for identifying critical stages

or transitions during the folding process.



(a) Trajectory F

(b) Trajectory P

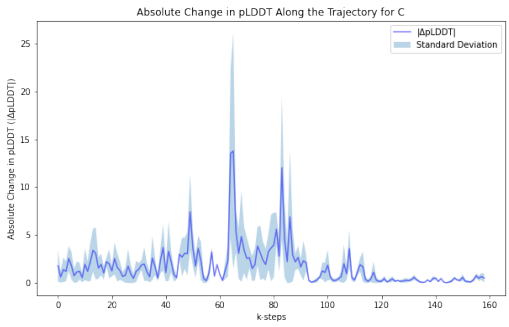
(c) Trajectory R

(d) Trajectory W

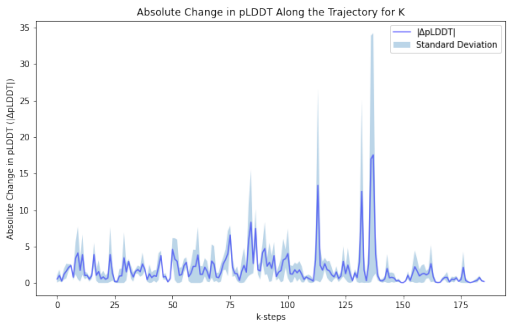
Figure 4.28: Change in TM score (ColabFold) along trajectory (△TM score). Change in TM score along the trajectories for sequences on 4 different trajectories towards the homopolymers for amino acids F, P, R, and W. The values are averaged over two repetitions, and the standard deviation is plotted.

In Figure 4.29, we show △pLDDT along the trajectories to homopolymers of C, K, N, and Y. In Figure 4.29a, we observe the highest magnitude of △pLDDT around step 60 after the magnitude of △pLDDT flattens out. For trajectory K (Figure 4.29b), we observe the highest △pLDDT at around step 135, considerably later than for the trajectory towards the homopolymer of C. For the trajectory towards the homopolymer of N (Figure 4.29c), we observe the highest △pLDDT around steps 70 and 80, and we see very minimal to almost no change after step 110, where the last peak was observed. For the trajectory towards the homopolymer of Y (Figure 4.29d), we see a bit more fluctuation in △pLDDT, but again, we can observe some peaks around step 75. These trends were also observed in the trajectories towards the homopolymers of the other amino acids. Some trajectories exhibit more fluctuation, while others

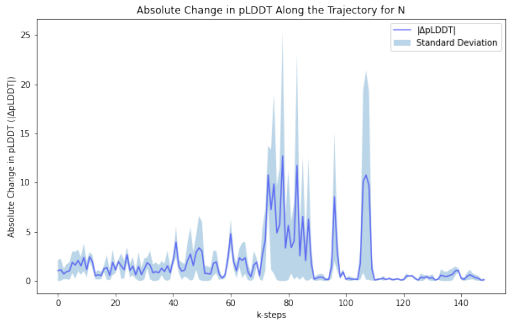
have less rapid changes. These analyses of △pLDDT reveal distinct patterns of change, with periods of more rapid change followed by periods of relative stability. The point on the trajectory and the magnitude of these changes varied across the different homopolymer targets. This may suggest that the specific sequence composition influenced the folding dynamics, but it can also give insight into the accuracy of ColabFold.



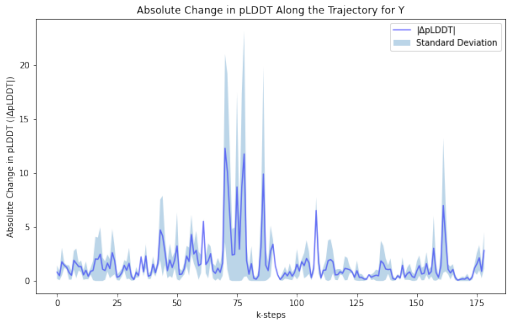
(a) Trajectory C



(b) Trajectory K



(c) Trajectory N

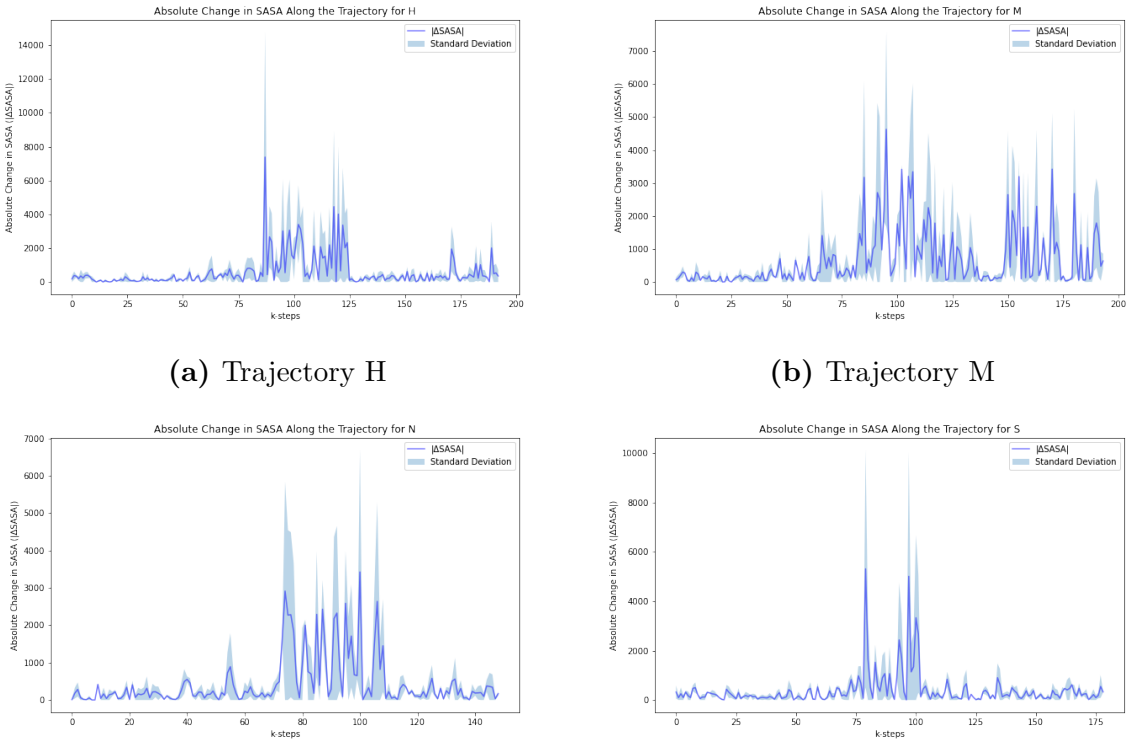


(d) Trajectory Y

Figure 4.29: Change in pLDDT (ColabFold) along trajectory (△pLDDT). Change in pLDDT along the trajectories for sequences on four different trajectories towards the homopolymers for amino acids C, K, N, and Y. The values are averaged over two repetitions, and the standard deviation is plotted.

Analyzing the change in SASA (△SASA) along the trajectories may indicate points where protein regions become exposed or buried to the solvent during folding. Figure 4.30 shows the change in SASA (△SASA) along the trajectories towards the homopolymers of 4 representative trajectories towards the homopolymers of H, M, N and S. For the trajectory towards the homopolymer of H (Figure 4.30a), we observed increased fluctuations in △SASA between steps 85 and 125, preceded and followed by periods of relatively lower fluctuations. A similar

trend was observed for the trajectory towards the homopolymer of N (Figure 4.30c), where the period of higher magnitude △SASA was slightly more extended. In contrast, the trajectory towards the homopolymer of M (Figure 4.30b) exhibited two distinct periods of higher △SASA, separated by a period of lower fluctuations between steps 140 and 150. The trajectory towards the homopolymer of S (Figure 4.30d) displayed less overall fluctuation in △SASA, with two distinct peaks observed around steps 75 and 100. The analysis of △SASA along the trajectories revealed different patterns of solvent-exposed area dynamics, with periods of higher fluctuations possibly indicating the increased burial or exposure of protein regions to the solvent. Again, the point on the trajectory and magnitude of these fluctuations varied across the different homopolymer trajectories.



(a) Trajectory H

(b) Trajectory M

(c) Trajectory N

(d) Trajectory S

Figure 4.30: Change in SASA (ColabFold) along trajectory (△SASA). Change in SASA along the trajectories for sequences on 4 different trajectories towards the homopolymers for amino acids H, M, N, and S. The values are averaged over two repetitions, and the standard deviation is plotted.

4.1.4 Contact Matrices

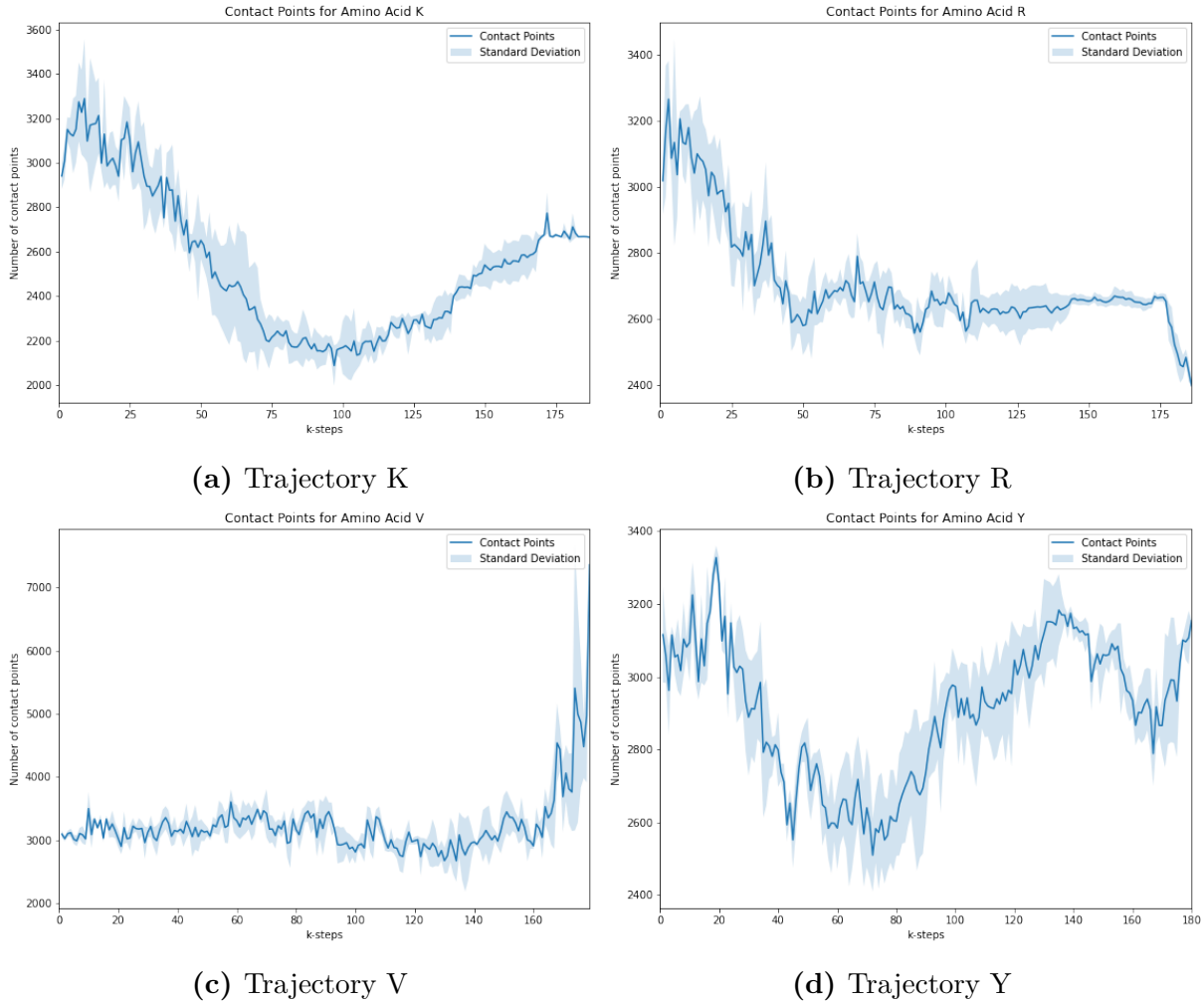
We present the representative examples of residue-residue contact points performed for each sequence along the trajectories toward the homopolymers of the 20 different amino acids. The full results can be found in Appendix B.

ESMFold

The analysis of residue-residue contact points along the trajectories towards homopolymers of different amino acids revealed varying trends, indicating the influence of amino acid properties on the formation of contact points. For the trajectory towards the homopolymer of K (Figure 4.31a) an initial decrease in contact points was observed until approximately step 100, followed by an increase. This trend suggests that the initial folding stages of this positively charged amino acid involve a reduction in residue-residue contact points, potentially due to electrostatic repulsion, before stabilizing interactions lead to an increase in contact points again. For the trajectory towards the homopolymer of amino acid R (Figure 4.31b), an initial decrease in contact points was observed until approximately step 50, before the values stagnate. The trajectory towards the homopolymer of Y (Figure 4.31d) exhibited a more complex pattern, with an initial increase in contact points for the first 20 steps, followed by a steep decline until step 80, and then an increase and subsequent decrease. In contrast, the trajectory towards the homopolymer of V (Figure 4.31c) displayed a relatively stable number of residue-residue contacts until step 160, with no major fluctuations, until a sudden increase after step 160. These findings shed light on the influence amino acid properties may have on the formation and dynamics of protein folding when diverging towards a homopolymer.

ColabFold

The analysis of residue-residue contact points along the mutational trajectories towards homopolymeric sequences reveals intriguing trends. In Figure 4.32 we can observe that as the sequences diverge from the ACE2 RBD, an initial decline in residue-residue contacts is observed, indicating a loss of structural compactness and native interactions.



(a) Trajectory K

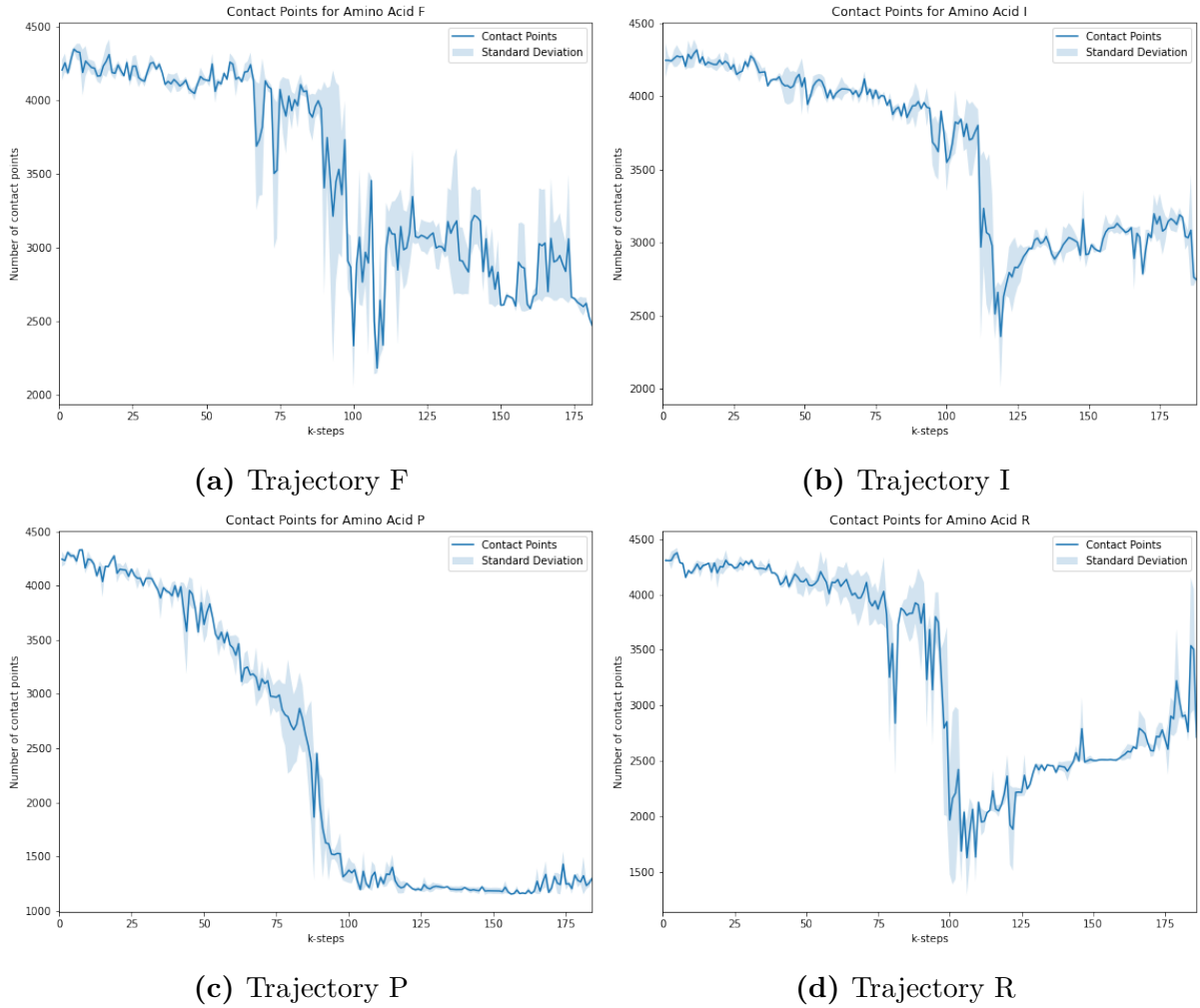
(b) Trajectory R

(c) Trajectory V

(d) Trajectory Y

Figure 4.31: Number of residue-residue contact points along trajectory (ESMFold) . Cchange of residue-residue contact points along the trajectories for sequences on four different trajectories towards the homopolymers for amino acids K, R, V, and Y. The values are averaged over three repetitions, and the standard deviation is plotted.

For the trajectories towards the homopolymers of F (Figure 4.32a) and I (Figure 4.32b), a sharp decrease in contact points is evident around steps 100 and 110, respectively. Notably, these decreases are followed by fluctuations and transient increases in contact points, suggesting the formation of alternative structural arrangements as the sequence diverges from the reference sequence. Similarly, for the trajectory towards the arginine (R) homopolymer, a rapid decrease in contact points is observed around step 100, followed by an increase in values. This highlights the complex nature of the folding landscape and the potential for non-monotonic changes in structural properties. In contrast, the trajectory towards the homopolymer of P exhibits a constant and rapid decrease in residue-residue contacts until approximately step



(a) Trajectory F

(b) Trajectory I

(c) Trajectory P

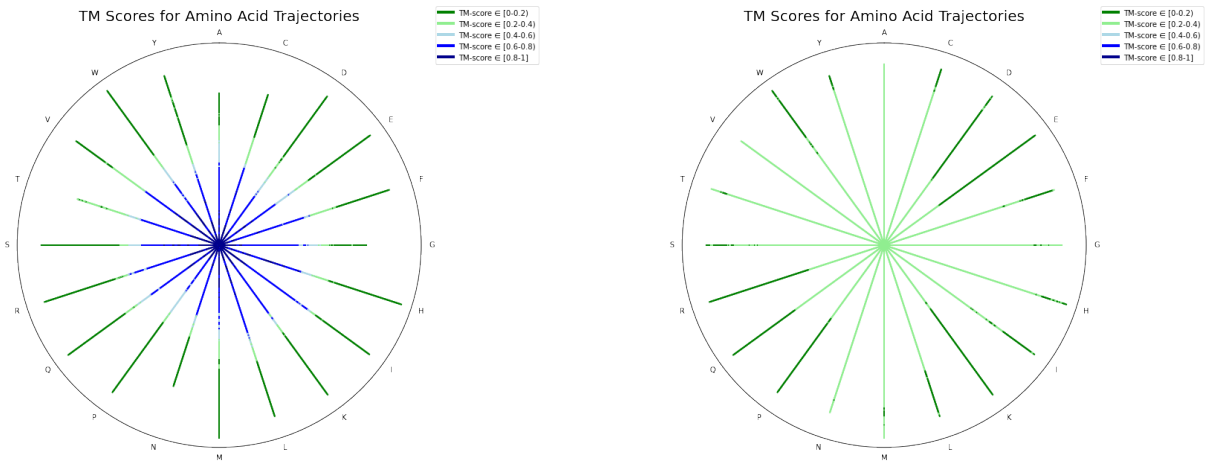
(d) Trajectory R

Figure 4.32: Number of residue-residue contact points along trajectory (ColabFold). Change of residue-residue contact points along the trajectories for sequences on four different trajectories towards the homopolymers for amino acids F, I, P, and R. The values are averaged over two repetitions, and the standard deviation is plotted.

100, after which the values stagnate at around 1000 contact points. This stagnation could indicate that beyond a certain point, further mutations do not substantially alter the overall compactness or contact density of the structures. This behavior may be attributed to the unique physicochemical properties and structural propensities of the respective amino acid types, which influence the folding landscape and the formation of residue-residue interactions. These observations underscore the intricate interplay between sequence and structure in proteins, as well as the potential for amino acid-specific trends in structural properties along the mutational trajectories towards the homopolymers.

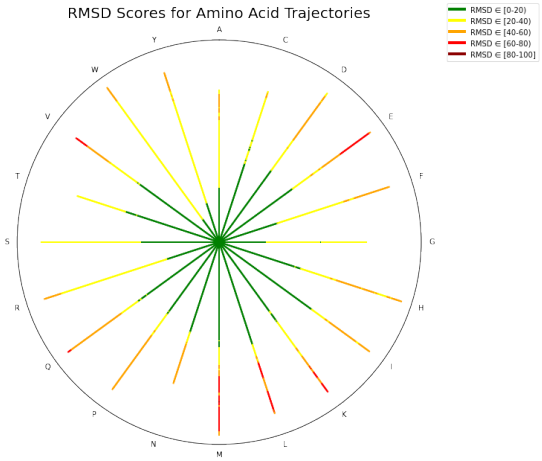
4.2 Spoke Wheel of Sequence Space

The structural metrics computed along the mutational trajectories were visualized using a spoke wheel plot to enable a a convenient and informative comparative analysis of the ESMFold and ColabFold predictions in protein sequence space. This approach offers a concise overview of the structural properties as the sequences evolve from the reference ACE2 RBD towards the homopolymer state. The TM-score, RMSD, pLDDT, and SASA were plotted along the spokes, with the reference sequence at the center of the plot. Figure 4.33 presents the TM-scores along the spokes for both ColabFold (Figure 4.33a) and ESMFold (Figure 4.33b). A notable difference is observed between the TM-score predictions of the two methods, as discussed in Section 4.1.1. The spoke wheel plots provide a concise visualization of these differences and allow for a direct comparison of the structural trends along the trajectories. Similarly, Figure 4.34, Figure 4.35, and Figure 4.36 showcase the RMSD, pLDDT and SASA values, respectively, along the spokes for ColabFold and ESMFold.

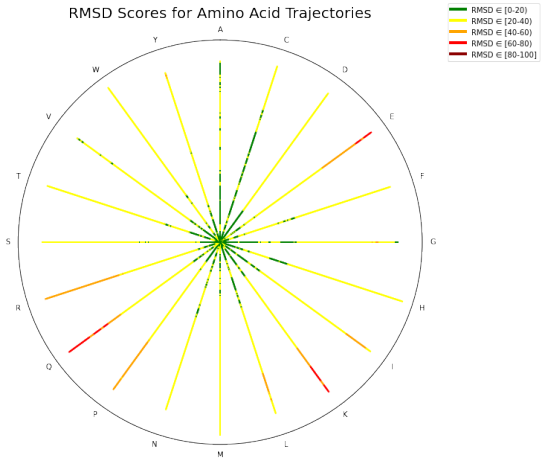


(a) Spoke Wheel TM Score (ColabFold) **(b)** Spoke Wheel TM Score (ESMFold)

Figure 4.33: Spoke Wheel TM Score.

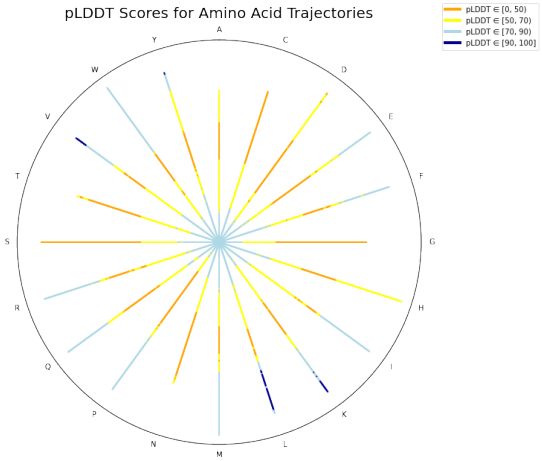


(a) Spoke Wheel RMSD (ColabFold)

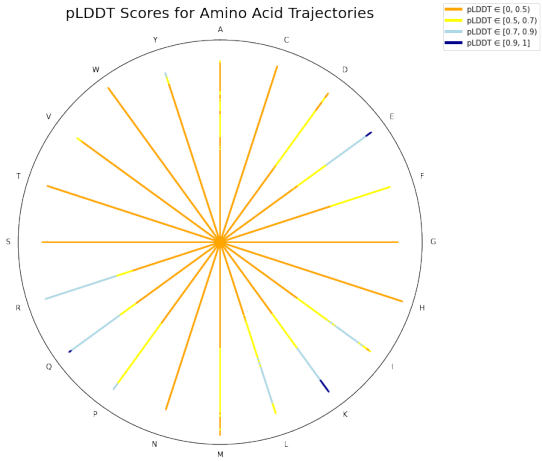


(b) Spoke Wheel RMSD (ESMFold)

Figure 4.34: Spoke Wheel RMSD.

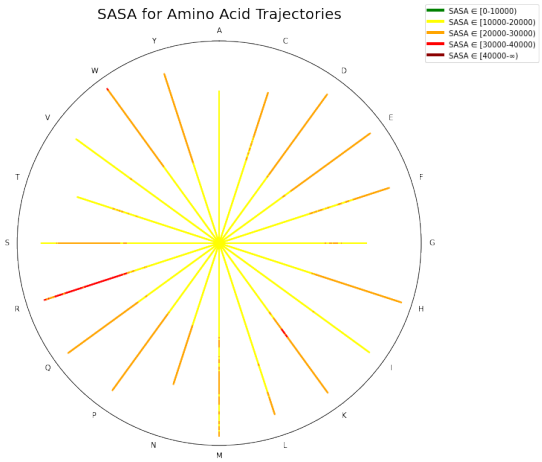


(a) Spoke Wheel pLDDT (ColabFold)

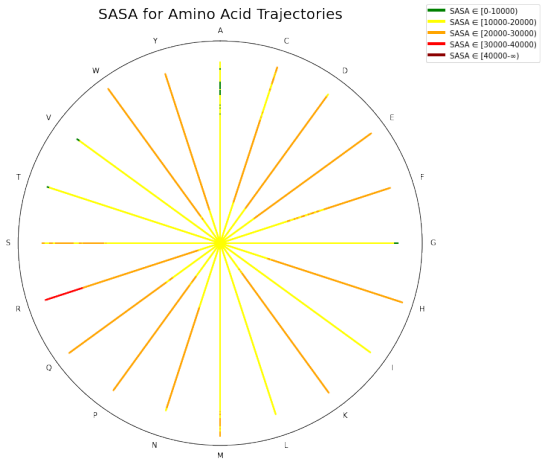


(b) Spoke Wheel pLDDT (ESMFold)

Figure 4.35: Spoke Wheel pLDDT.



(a) Spoke Wheel SASA (ColabFold)

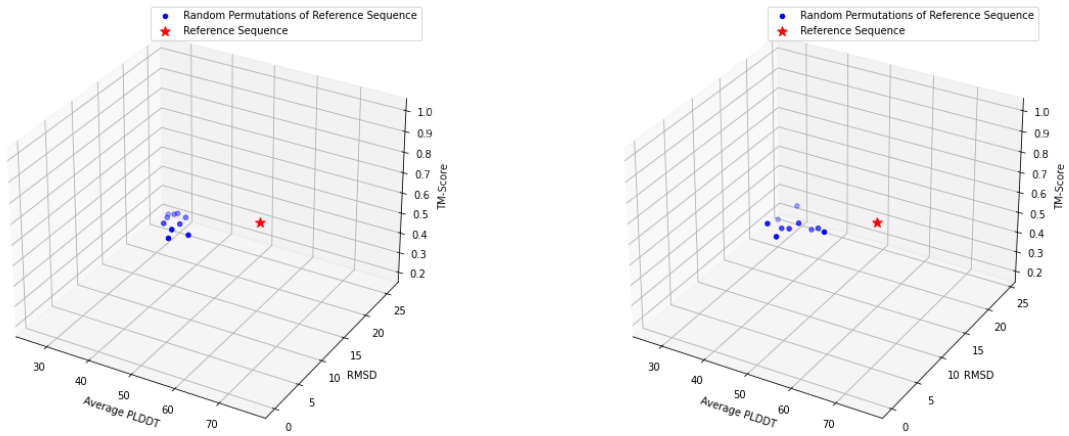


(b) Spoke Wheel SASA (ESMFold)

Figure 4.36: Spoke Wheel SASA.

4.3 Permutated Sequences

To assess the impact of sequence order on structural properties, we analyzed 10 randomly permutated sequences derived from the reference structure. All sequences result in Kendall's tau values close to zero when compared to the reference sequence (Table 4.3). The structural metrics, including the pLDDT, RMSD, and TM-score, were computed for the permutated sequences using the outputs from ColabFold and AlphaFold2. Figure 4.37 presents a three-dimensional scatter plot depicting the relationship between the pLDDT, RMSD, and TM score for the permutated sequences.



(a) ColabFold **(b) AlphaFold**

Figure 4.37: Structural Metrics of Permutated Sequences vs. Reference Sequence. Visualization of the structural metrics pLDDT, RMSD, and TM score of the permutated sequences compared to the reference sequences in a three-dimensional scatter plot.

Notably, the permutated sequences form distinct clusters in the 3D plot, with a more pronounced clustering observed for the ColabFold results (Figure 4.37a) compared to the AlphaFold results (Figure 4.37b). Furthermore, the analysis reveals that the permutated sequences exhibit consistently high RMSD values when compared to the reference structure. This observation is accompanied by low pLDDT scores, indicating a lower confidence in the predicted structural models, and low TM scores, reflecting significant deviations from the reference structure's topology. These results undermine the impact of sequence order on structural

properties.

Table 4.3: Structural and Sequence Metrics for Permutated Sequences. Predicted with ColabFold and AlphaFold.

ColabFold					
Sequence	**Kendall's Tau**	**pLDDT**	**RMSD**	**TM score**	**SASA**
permutated sequence 1	0.01	32.41	21.28	0.26	18151.37
permutated sequence 2	-0.06	28.83	25.65	0.23	22135.41
permutated sequence 3	0.01	31.68	23.64	0.23	19072.56
permutated sequence 4	0.08	26.18	25.86	0.21	26319.98
permutated sequence 5	0.02	28.46	22.92	0.23	18397.25
permutated sequence 6	-0.06	28.16	25.44	0.23	22668.57
permutated sequence 7	-0.02	30.99	25.53	0.23	20665.27
permutated sequence 8	0.11	35.54	22.14	0.23	15140.27
permutated sequence 9	0.06	32.35	20.62	0.23	19179.32
permutated sequence 10	0.04	27.16	24.79	0.22	24605.96
AlphaFold					
Sequence	**Kendall's Tau**	**pLDDT**	**RMSD**	**TM score**	**SASA**
permutated sequence 1	0.01	29.67	24.27	0.26	21091.67
permutated sequence 2	-0.06	40.09	21.21	0.26	15591.03
permutated sequence 3	0.01	31.27	18.76	0.23	17326.41
permutated sequence 4	0.08	33.47	21.48	0.26	16535.96
permutated sequence 5	0.02	28.55	19.20	0.27	19435.90
permutated sequence 6	-0.06	26.68	22.78	0.21	23205.43
permutated sequence 7	-0.02	34.95	22.87	0.20	18616.11
permutated sequence 8	0.11	36.92	22.61	0.23	16447.69
permutated sequence 9	0.06	30.25	20.64	0.23	17711.00
permutated sequence 10	0.04	31.84	20.87	0.23	16599.75

Furthermore, the permutated sequences exhibit varying amount of contact points. The amount of residue-residue contact points for the permutated sequences can be seen in Table 4.4. We observe that the values are mostly lower than for the non permutated sequence close to the reference sequence, with some exceptions such as permuated sequence 8 for ColabFold.

Table 4.4: Amount of residue-residue contact points. Comparison of the amount of residue-residue contact points for the permutated sequences predicted with ColabFold and AlphaFold.

Sequence	ColabFold	AlphaFold
permutated sequence 1	3037	2859
permutated sequence 2	2221	3798
permutated sequence 3	2694	3790
permutated sequence 4	1491	3773
permutated sequence 5	2986	3157
permutated sequence 6	2127	2552
permutated sequence 7	2424	3433
permutated sequence 8	3503	3930
permutated sequence 9	2853	3567
permutated sequence 10	1865	3798

4.4 Genetic Constraint on ACE2 RBD

Thus far we have sampled protein sequence space by systematically scanning along informative (if arbitrary) trajectories. This approach is meant to yield insight into the structure treating proteins as a physical "material". However, these arbitrary trajectories may not in all cases reflect the actual possibilities permitted by the underlying genetic sequence encoding the protein. In the context of genetic variation and its impact on protein structure, we also examined single nucleotide substitutions on DNA level and their impact on amino acid alterations. Particular attention was thus paid to the genetic code redundancy, as not all single nucleotide substitutions result in an amino acid modification. Each codon of the 195 codons can undergo 3^3 single nucleotide mutations, as there are 4 nucleotides (A, G, C, and T), and each nucleotide can be substituted by the remaining 3. Although the range of mutational variation is broad and complex, we focused our analysis on missense mutations, which are nucleotide substitutions that lead to an amino acid alteration. We observed that certain amino acids are more susceptible to mutations than others. This variation in the number of possible missense mutations can among other reasons be attributed to the genetic code, where some amino acids are coded by multiple codons, whereas others, such as W are only coded by one codon.

Figure 4.38 shows the distribution of amino acid alterations resulting from single nucleotide substitutions across all codons.

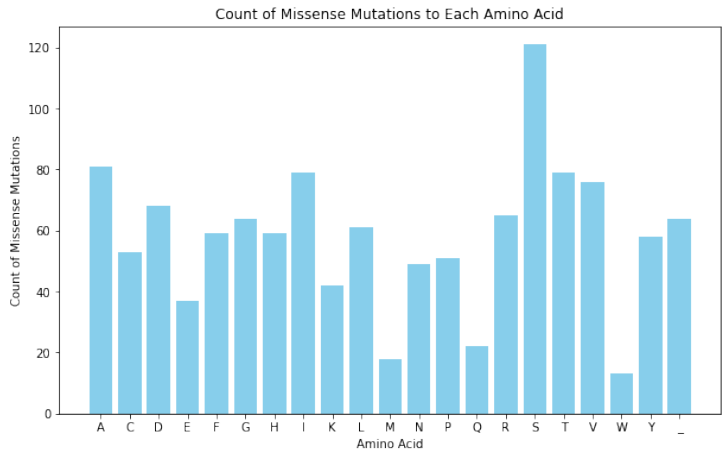


Figure 4.38: Possible Missense Mutations to Each Amino Acid. The absolute number of possible mutations to the standard 20 amino acids and the stop codon (here denoted as _) when considering missense mutations. Notably, the majority of the codons can possibly mutate to amino acid S and the least amount of codons can mutate to amino acid W.

Notably, amino acids such as R, L, and S exhibit a higher propensity to be mutated to, while others, like W and M, are less likely to be a result of a missense mutation. As we are interested in the dynamics of protein sequence space, it is another interesting property to analyze to which extent we can explore the sequence space when considering the constraints imposed by the genetic code. Figure 4.38 illustrates the distribution of amino acid alterations resulting from missense mutations across all codons. To quantify the sequence space traversal, we analyzed the number of steps (amino acid substitutions) that can be achieved from the reference sequence, given the genetic code redundancy and the initial amino acid composition. This spoke wheel can be seen in Figure 4.39 and Table 4.5 further presents the exact counts of the number of steps that can be taken into the sequence space for each of the 20 amino acids. This analysis is particularly relevant because a significant portion of the theoretical sequence space does not overlap with the genetically possible sequence space. Understanding these delineations in our theoretical sequence space adds another dimension to our analysis where we aim to unravel the properties of protein folding as we deviate further away from the

reference sequence.

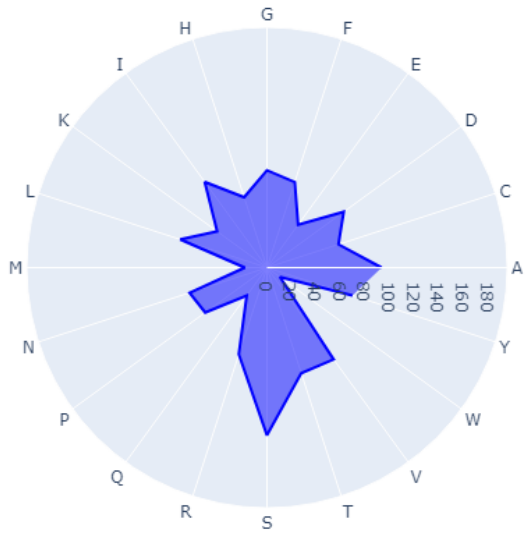


Figure 4.39: Spoke Wheel of Sequence Space. Spoke wheel representation of the sequence space, where each spoke corresponds to a trajectory towards a homopolymer sequence. The blue part is the genetically possible sequence space when considering missense mutations and the genetic code.

Table 4.5: Maximum Occurrence of Amino Acids. Maximum occurrence of an amino acid in a sequence when taking genetic code redundancy into account. This includes the amino acids that were present in the reference sequence as well as all other possible mutations that, for each amino acid, can occur.

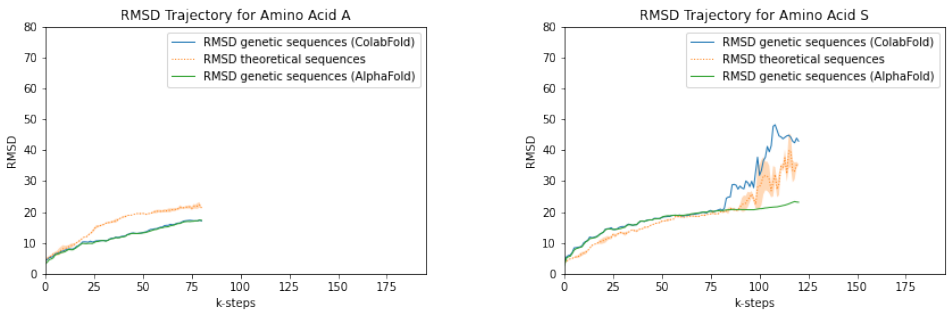
Amino Acid	Count
A	93
C	61
D	77
E	43
F	73
G	79
H	60
I	86
K	50
L	74
M	18
N	66
P	62
Q	28
R	74
S	**136**
T	90
V	92
W	**13**
Y	73

4.4.1 Genetic Trajectories

We performed one repetition of structure prediction with ColabFold for the genetically possible trajectories through missense mutations. The 20 trajectories were analyzed using the structural metrics and methods employed with the theoretical trajectories. To visualize differences between the genetically possible sequences and the theoretical sequences, we plotted them together until the step on the genetically possible trajectory. Even though all sequences are generally theoretical, we will refer in this section to the sequences from our directed mutational trajectories as 'theoretical sequences' and to the genetically possible sequences as 'genetic sequences'.

Trajectory Metrics

The RSMD along the genetic trajectories revealed that for some trajectories there are substantial differences in RMSD between the theoretical sequences and the genetic sequences. Figure 4.40 shows the results for two representative trajectories towards the homopolymer of A and S. Notably, the genetic trajectory towards the homopolymer of A exhibits lower RMSD values compared to the theoretical sequences (Figure 4.40a). In contrast, the trajectory towards the homopolymer of S displays the opposite trend, with higher RMSD values for the genetically possible sequences that were predicted with ColabFold and higher RMSD values for the genetically possible sequences that were predicted with AlphaFold2 (Figure 4.40b).

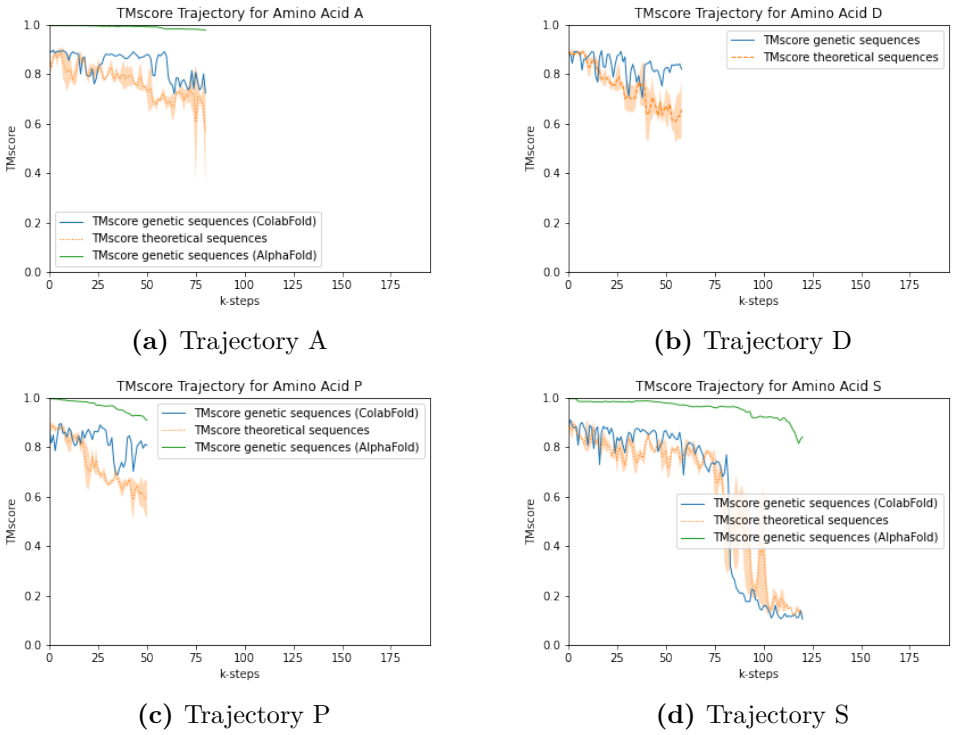


(a) Trajectory A **(b)** Trajectory S

Figure 4.40: RMSD along Trajectories. RMSD along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids A and S. The RMSD values are plotted for the theoretical and genetic sequences until the step on the genetically possible trajectory. The genetic sequences were predicted with ColabFold (blue) and AlphaFold2 (green).

Similar trends in differences can be observed when comparing the TM score of the genetic trajectories to the theoretical trajectories. The TM scores along the genetically possible trajectories towards the homopolymers of amino acids D and P are shown in Figure 4.41. While most of the 20 trajectories exhibit similar trends with some slight variation, we can observe a more notable difference between the genetic and theoretical sequences for the trajectories towards the homopolymer of A (Figure 4.41a) and P (Figure 4.41c). Specifically, the genetically possible sequences generally have higher TM score values compared to the theoretical sequences. Notably, the TM scores are even higher for AlphaFold2 predictions compared to ColabFold

predictions. Interestingly, the trajectory towards the homopolymer of S (Figure 4.41d) shows a more distinct pattern. Initially, the TM score values of the genetic sequences (ColabFold) are slightly higher, albeit within the error bars of the theoretical sequence, until approximately step 85. At this point on the trajectory, a significant decrease in TM score is observed, which is also present in the theoretical sequences, although there it occurs less abruptly on the trajectory. The TM scores of the genetic sequences on the trajectory towards homopolymer S that were predicted with AlphaFold2 were even higher and had a more notable decline at steps 100 and 125.



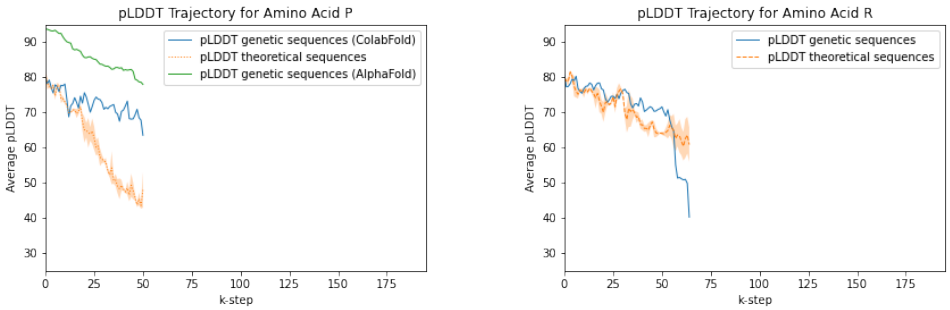
(a) Trajectory A (b) Trajectory D

(c) Trajectory P (d) Trajectory S

Figure 4.41: TM score along Trajectories. TM score along the trajectories for sequences on four different trajectories towards the homopolymer of amino acids A,D, P, and S. The TM score values are plotted for the theoretical and genetic sequences until the step on the genetically possible trajectory. The genetic trajectories towards the homopolymer of amino acids A, P, and S were predicted with ColabFold (blue) and AlphaFold2 (green).

Figure 4.42 shows the pLDDT along the genetically possible trajectory towards the homopolymer of amino acids P and R. Most trajectories exhibited similar values for both the theoretical and genetic sequences. However, for the trajectories shown in Figure 4.42 we can observe a

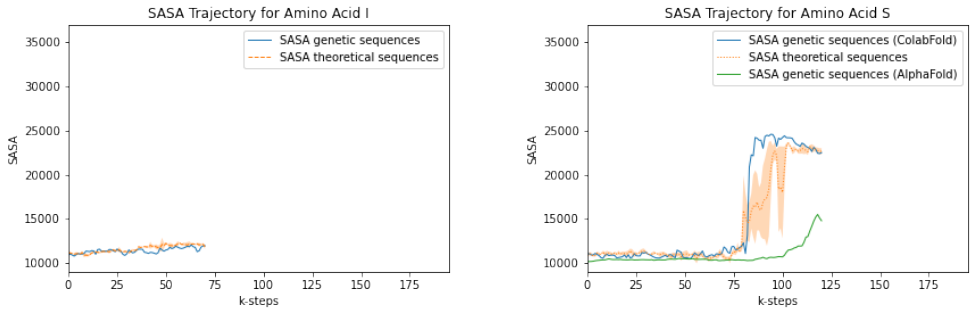
more substantial difference. For trajectories towards the homopolymer of amino acid P (Figure 4.42a) the theoretical sequences exhibited lower pLDDT scores compared to the genetically possible sequences. Interestingly, the trajectory towards the homopolymer of R (Figure 4.42b) showed a more drastic decline in pLDDT scores for the genetically possible sequences around step 50. In contrast, the theoretical sequences exhibited a more gradual decline in pLDDT along this trajectory. Again, we observe higher pLDDT values for the sequences predicted with AlphaFold2, indicating higher confidence in these predictions.



(a) Trajectory P **(b)** Trajectory R

Figure 4.42: pLDDT along Trajectories. pLDDT along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids P and R. The pLDDT values are plotted for the theoretical and genetic sequences until the step on the genetically possible trajectory. The genetic trajectory towards the homopolymer of amino acid P was predicted with ColabFold (blue) and AlphaFold2 (green).

Figure 4.43 shows the SASA values along the genetically possible trajectory towards the homopolymers of amino acids I and S. While no significant differences were observed between the genetic and theoretical sequences for I, the trajectory towards the serine homopolymer (Figure 4.43b) exhibited a more distinct pattern. For the genetic sequences predicted with ColabFold along the trajectory towards the serine homopolymer, we observed a more pronounced increase in SASA values commencing at approximately step 85, compared to the theoretical sequences. In contrast, the genetic trajectory towards the serine homopolymer predicted with AlphaFold2 demonstrated lower SASA values overall. Additionally, a more notable increase in SASA values was observed around step 110 for the AlphaFold2 predictions.

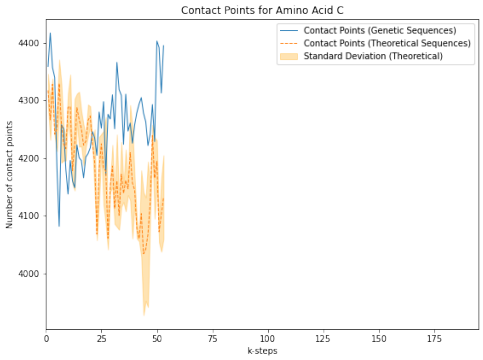


(a) Trajectory I **(b)** Trajectory S

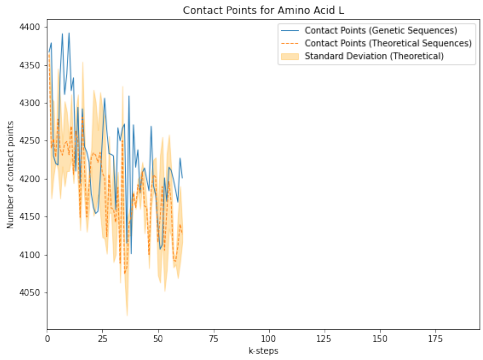
Figure 4.43: SASA along Trajectories. SASA along the trajectories for sequences on two different trajectories towards the homopolymers for amino acids I and S. The SASA values are plotted for the theoretical and genetic sequences until the step on the genetically possible trajectory.

Contact Matrices

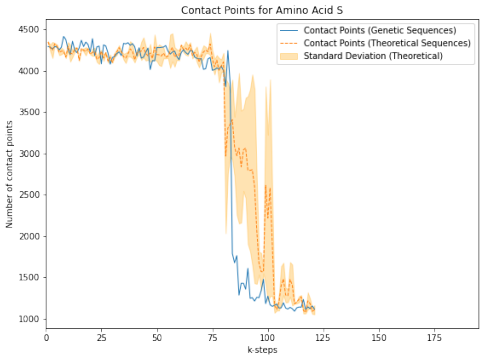
To investigate the structural implications of genetic constraints, we compared the residue-residue contact points between the genetic and theoretical sequences along trajectories towards the homopolymers of amino acids C, L, S, and P (Figure 4.44). For certain trajectories, such as the one towards the homopolymer of L (Figure 4.44b), the number of residue-residue contact points for the genetic sequences remained within the error margins of the theoretical sequences. However, for other trajectories, more notable differences were observed, as seen for the trajectories towards the homopolymers of C (Figure 4.44a) and P (Figure 4.44d). Interestingly, for the trajectory towards the homopolymer of S (Figure 4.44c), the genetic sequences initially exhibited similar residue-residue contact point values compared to the theoretical sequences. However, around step 85 of the trajectory, a sharper decline in the number of contact points was observed for the genetic sequences.



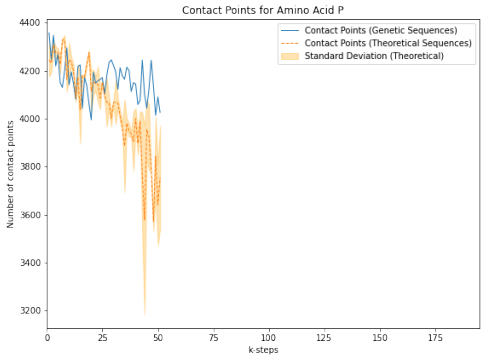
(a) Trajectory C



(b) Trajectory L



(c) Trajectory S



(d) Trajectory P

Figure 4.44: Number of residue-residue contact points along trajectory. Change of residue-residue contact points along the trajectories for sequences on 4 different trajectories towards the homopolymers for amino acids C, L, S, and P. The residue-residue contact points are plotted for the theoretical and genetic sequences until the step on the genetically possible trajectory.

4.5 FAIR Digital Twin

In this section, we show an example model of a FDT that can accommodate the data generated in this project. At the core of the FDT is a protein sequence, specifically a variant of the ACE2 RBD of SARS-CoV-2. The FDT model is depicted in Figures 4.45 and 4.46 and the full model can be accessed under https://doi.org/10.17605/OSF.IO/PMJ9X [57]. Each sequence can have four distinct observations, highlighted in red: 1) a computational observation, 2) an experimental observation, 3) a real-world observation, and 4) an observation of the sequence space navigation. These observations are further characterized by various attributes, shown in yellow.

Computational Observation

The computational observation, for instance, can be described using attributes such as funding, location, method, and date/time. The funding attribute is composed of HealthHolland, as they fund the STAYAHEAD project. The location attribute is described as SURF, where the computational predictions are performed. The date and time attributes record the specific moment when the computational prediction was performed. Additionally, since multiple structure prediction algorithms are employed, the method attribute describes algorithms like ESMFold, AlphaFold2, and ColabFold. These structure prediction algorithms can have different structural metrics, hence, a separate structural metrics attribute is needed. This attribute includes metrics such as RMSD, TM score, SASA, and pLDDT.

Experimental Observation

The experimental observation can be described using attributes such as funding, location, date/time, method, and patient information. The funding attribute is again composed of HealthHolland. The location attribute is described as Utrecht, where the experimental validations are performed in the lab. This can also include information about who performed the experiments. The method attribute can be described as in vitro models, to test the antibody escape of predicted high-risk variants. The date/time attribute then describes the date and time these experiments were performed.

Real-World Observation

Moreover, some sequences can also have a real-world observation, which describes sequences that have been observed in nature already. This includes attributes such as method, date/time, location, patient, and funding. The method could for example be described as the sample collection through a nose swap. The date/time attribute describes the date and time this variant was detected. The location attribute describes the location where the variant was detected, for example at Schipol airport, and can be described using the OGC.org description. Moreover, we can include a patient attribute which has a personal information attribute. This

can further be described using the age, sex, and ethnicity of the patient. Moreover, we have a funding attribute that describes the funding for collecting patient samples.

Sequence Space Navigation Observation

Lastly, each variant can have a sequence space navigation observation which includes attributes such as information content, run length, periodicity, and information about the variants' reference sequence. The information content can be described with entropy, the dimensionality of sequence space around that variant, and the number of possible variants that can be formed through mutations. The run-length attribute refers to the maximum length of consecutive repeated residues in the sequence variant. The periodicity attribute describes whether the sequence variant exhibits any periodic patterns or repeating motifs.

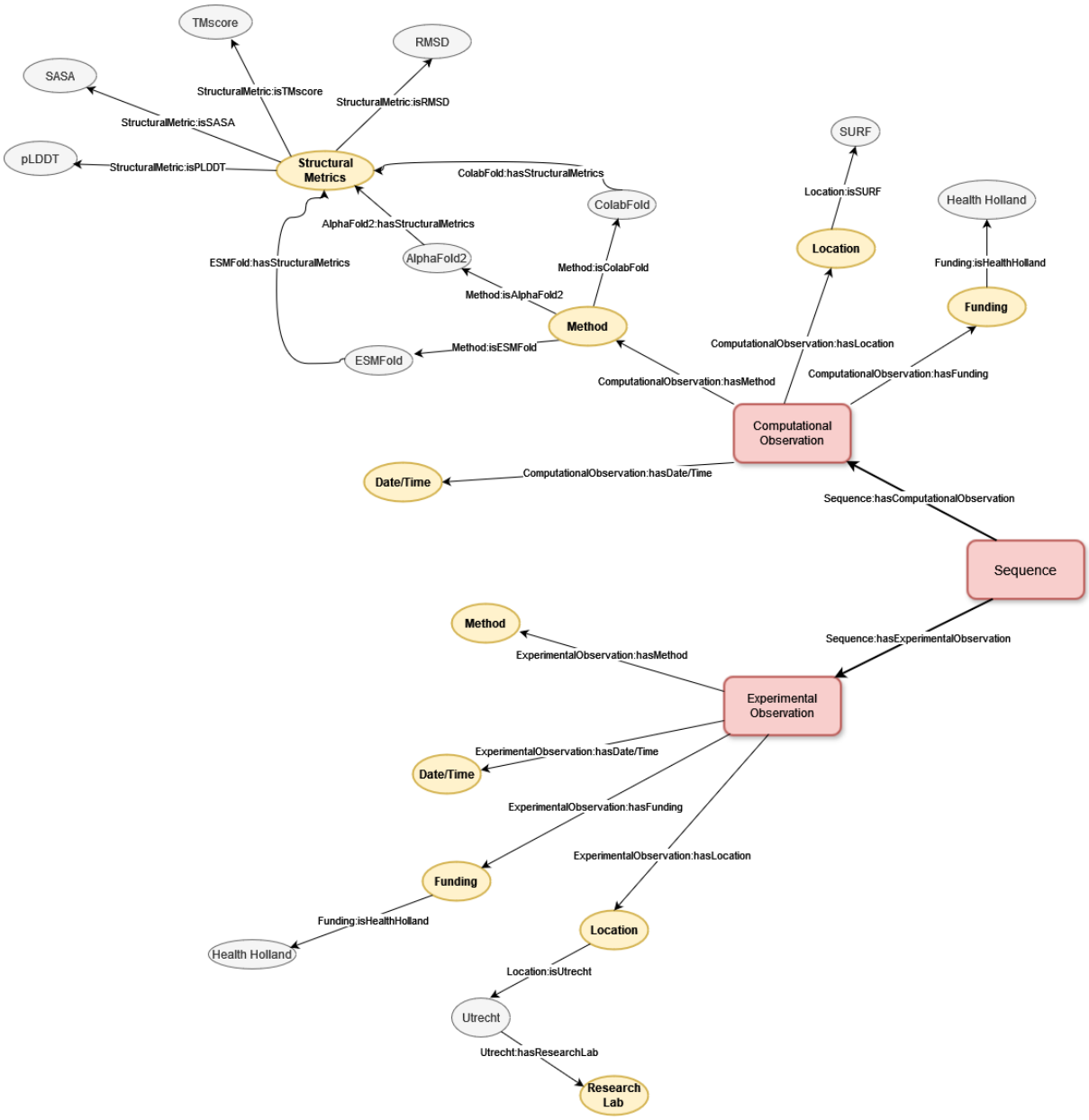


Figure 4.45: Computational and Experimental Observation of the FDT.

Figure 4.46: Theoretical and Sequence Space Navigation Observation of the FDT.

Chapter 5

Discussion

In this chapter, we will discuss the implications of our findings in the context of existing literature and theories in the field of protein sequence space. We will analyze how mapping the sequence space through mutational trajectories can enhance our understanding of critical determinants of protein folding. We will further analyze the structural trends across the trajectories towards the homopolymers and will elaborate on the differences between the different mutational trajectories. Finally, we will elaborate on ethical concerns and outline future research directions and potential for extending this approach.

5.1 Computational Mapping of Sequence Space

Mapping the protein sequence space through the directed mutational trajectories provides invaluable insights into the boundaries and constraints governing protein folding and function. This approach allows us to systematically explore robust properties of the vast sequence space, shedding light onto the distribution structural properties of folded protein sequences. By mapping the entire sequence space, we might identify sparse regions or voids between functional clusters, revealing the constraints imposed by protein structure and folding requirements not only on the trajectories but also for the entire rest of the sequence space. While mapping the entire sequence space, i.e., predicting the three-dimensional structure for every possible sequence, for proteins of significant length may be computationally intractable, directed mu-

tational trajectories offer a practical starting point for exploring the sequence space. Even the sequence space of a protein of length 195 has 3705 dimensions and 20^{195} possible sequences (see Table 2.2), calling for a systematic exploration. Computational techniques, such as machine learning models or molecular dynamics simulations, can predict the structural and functional consequences of mutations along these trajectories. However, experimental validation is crucial for confirming computational predictions and refining our understanding of sequence-structure-function relationships. In summary, although our current data is limited, the initial structural trends across the trajectories observed in the folding properties of sequences as they diverge from the reference sequence demonstrate the potential of this approach.

5.2 Structural Trends Across the Trajectories

Through the analysis of structural trends across the trajectories, we gained insights into the impact that accumulating amino acid substitutions have on protein folding, unraveling the structural landscape as sequences progressively diverged from the original ACE2 RBD sequence. The analysis of residue-residue contact points using ColabFold unveiled compelling patterns as we moved away from the reference sequence. The initial decline in contact points signified a gradual loss of structural compactness and native interactions, indicating an unfolding process as mutations accumulated. Subsequent fluctuations and increases in contact points hinted at the emergence of alternative structural conformations, underscoring the complexity of the protein folding landscape as we diverge further from the ACE2 RBD sequence. The diverse patterns in contact matrices between the different amino acid trajectories may be attributed to other factors. First, amino acids have unique structural properties and preferences for specific conformations or interactions. Structure prediction algorithms account for these amino acid-specific characteristics to predict proteins' folding patterns accurately.

Moreover, the energy landscape of protein folding is also influenced by the interactions between amino acids, which can vary based on their properties. Amino acids with distinct physiochemical features will navigate different energy landscapes during folding, leading to diverse contact

patterns. For example, the stable alpha-helical structure exhibited by polylysine sequences (K) is influenced by the physiochemical properties of lysine and its proneness to form specific secondary structures. In contrast, the less organized helical structure observed for the proline (P) homopolymer reflects the unique structural challenges posed by the proline's rigid conformation. This is reflected in the number of residue-residue points. Additionally, the training data used by structure prediction algorithms may only partially capture the diverse structural behaviors associated with all amino acids. This can be attributed to the limited structural data on databases like the Protein Data Bank.

Considering these factors, it becomes evident that the inherent differences in amino acid properties and their impact on the folding process contribute to the observed variations in contact matrices. However, we can observe a 'transition' point in residue-residue contact points on most trajectories. These transition points were observable around halfway on the trajectory towards the homopolymer. Proteins have specific tertiary and quaternary structures that are stabilized by various interactions, including residue-residue contacts. When mutating towards a homopolymer, introducing similar residues might disrupt existing stabilizing interactions (e.g., hydrogen bonds, hydrophobic interactions, electrostatic interactions) and lead to the formation of new interactions characteristic of the homopolymer. The transition point could represent an intermediate state in the folding/unfolding process. Proteins often fold through a series of intermediates, and the transition might reflect a shift from one dominant intermediate structure to another as the residue composition changes. The transition point might indicate a change in the overall stability of the protein. Initially, the protein might retain its native structure, but as mutations accumulate, the native structure might destabilize, leading to a new, possibly more stable structure characteristic of the homopolymer. In the context of the protein's energy landscape, the transition point could correspond to a crossing from one energy basin (associated with the native structure) to another (associated with the homopolymer). This crossing is a critical point where the protein shifts from being more stable in one conformation to another.

We also investigated different structural metrics to gain insight into the folding patterns of

proteins. We can differentiate between intrinsic physiochemical metrics of a protein structure (pLDDT and SASA) and comparative metrics that assess structural similarity against a reference structure (TM score and RMSD). The observed increase in pLDDT values predicted by ESMFold indicates that as sequences deviated from the reference sequence, there is higher confidence in predicting local accuracy. However, for ColabFold, we observe more bifurcated trends, where initially, the values decrease and then increase again after a transition period. This trend can stem from the different training methods of these structure prediction algorithms. Variations in the patterns of structural disorder among different trajectories not only underscored the diverse behavior of the predicted models but also offered insights into the intrinsic flexibility and adaptability of proteins in response to sequence perturbations. Looking at the trends across the trajectories for the SASA values, we observe a more evident trend for the ColabFold output, where we observe an increase in SASA as we diverge from the reference sequence. Specifically, around steps 75 to 100, we can observe a more gradual increase across all trajectories. Some trajectories naturally have lower SASA values, such as V or L, compared to other trajectories that exhibit higher SASA values, such as K and H. This might be attributed to their side chains (see Table 2.1). Trajectories towards homopolymers of V or L, which are amino acids with hydrophobic side chains, tend to have lower SASA values due to their tendency to form more tightly packed cores.

In contrast, trajectories towards homopolymers of amino acids like K or H that have charged or polar side chains often show higher SASA values due to their tendency to be more solvent-exposed. This trend was captured more by the ColabFold outputs, while for the ESMFold outputs, we observed varying trends. For some trajectories, the SASA values decrease over time, while others exhibit similar patterns to the ColabFold results.

We observe an increase in RMSD in both algorithms outputs as we diverge from the parent sequence. For ColabFold, we can observe a more substantial rise in RMSD around steps 75 to 100 on the trajectories, indicating a transition point. Notably, the RMSD for sequences predicted with ColabFold is lower than those predicted with ESMFold. Even though for some trajectories

predicted with ESMFold, we observe an increase in RMSD, this increase is sometimes relatively mild. Even sequences close to the reference sequence already exhibit high RMSD values close to 20 Å, while for ColabFold, they are around 3 Å. ESMFold predictions result in very low TM scores overall (around 0.2), while ColabFold initially results in higher TM scores (around 0.9) and then decreases. This decrease is happening gradually; however, again, around steps 75-100, we can see a sharper decline, indicating a transition point.

Even though both comparative metrics and intrinsic metrics have some relation to each other, we can argue that the TM-score might be independent of metrics like the pLDDT score. While the TM-score indicates structural differences relative to a reference sequence, the pLDDT score provides a confidence measure for the prediction itself. This distinction is crucial, as a high pLDDT score does not necessarily guarantee a high TM-score, especially in cases where the predicted structure diverges significantly from known templates. We can observe this for a lot of the trajectories, where the pLDDT scores increase again as the sequences are closer to the homopolymers, while the TM score continues to decrease. Therefore, while both metrics are valuable, they serve different purposes and can provide complementary insights into the quality and reliability of protein structure predictions.

5.2.1 Phase Transitions

Phase transitions in general are abrupt changes in the properties of a system that occur due to changes in external conditions such as temperature or the composition. These phenomena are observed in various systems, ranging from condensed matter physics to biology. In protein sequence space we can define phase transitions as the phenomenon where small changes in the amino acid can lead to drastic changes in the protein's structural and functional properties. These phase transitions will eventually provide an outline of the sequence space, where we can determine transitions between clusters of functional and nonfunctional sequences. We know that the protein sequence space is a vast, high-dimensional space that encompasses all possible combinations of amino acid sequences for a given protein length. Each point in this

space represents a unique protein sequence and the properties of the proteins are determined by its position in this space. These properties include not only its structure but also the functionality of the protein. Thus, within this sequence space, regions exist where proteins share similar structural and functional characteristics. These regions are separated by phase transitions, where small changes in the sequence can lead to a significant change in the protein's properties. We can compare the principles of phase transitions in the protein sequence space to known principles from polymer physics and statistical mechanics [42]. Proteins are essentially heteropolymers. Their behavior is controlled by the interplay between various interactions, such as van der Waals forces, electrostatic interactions, hydrogen bonding, and more. These interactions are influenced by the specific arrangement of amino acids in the sequence, as well as other environmental conditions. Small changes in the sequence can alter the balance of these interactions, leading to changes within the protein.

In this thesis, we aimed to create an initial map of the sequence space of the ACE2 RBD, by systematically exploring this space through directed mutational trajectories. The spoke wheel diagrams presented in Section 4.2 show the transitions of structural metrics along these trajectories. Although these results only encompass a very small fraction of the entire sequence space, they reveal significant trends in protein folding behavior as sequences diverge from the reference sequence. This analysis is valuable not only for evaluating the performance of structure prediction algorithms but also for understanding how the actual structure diverges from the ACE2 RBD. The properties of sequences that lie in the interstitial space between the spokes remain unknown. Given the computational constraints of this project, it was not feasible to investigate all possible sequences. However, national supercomputing resources, such as those available at the Oak Ridge National Laboratory, could provide the necessary computational power to explore these intermediate sequences comprehensively.

The observed transition points around steps 75 and 100 in various structural metrics, such as residue-residue contacts, RMSD, and TM-score, indicate a phase transition-like behavior in the folding landscape as sequences deviate from the reference ACE2 RBD sequence. The initial

gradual changes in structural metrics could represent destabilization of the native structure, followed by a more sudden transition to a new structural state characteristic of the homopolymer sequence [15]. Peaks can be seen between steps 75 and 100 and at various points throughout the trajectory. This was clear when examining the change in structural metrics between steps k and k+1 on the trajectory. These peaks may indicate transient structures that the protein explores. The relatively small standard deviation, given that we only had 2 replicates for ColabFold and 3 replicates for ESMFold, is a powerful indicator that these properties per step and over the trajectory are robust.

The bifurcated trends observed in some structural metrics, particularly for the ColabFold predictions, could also indicate phase transition-like behavior. The initial decrease followed by an increase in specific metrics may reflect a transition from the native structure to an intermediate state and subsequently to the final folded state of the homopolymer sequence. Such bifurcated patterns are often observed in physical systems undergoing phase transitions, where the system exhibits distinct behaviors in different regimes separated by a transition point. Overall, the observed transition points, higher peaks, and bifurcated trends in the structural metrics suggest that the folding landscape of proteins exhibits phase transition-like characteristics as sequences diverge from the reference sequence. These findings highlight the complex interplay between sequence variations, structural properties, and the energy landscape of protein folding, providing insights into the fundamental principles governing this intricate process. Identifying and characterizing phase transitions in such a high-dimensional space is incredibly complex and challenging. Most theoretical frameworks and models for phase transitions are developed for low-dimensional systems, and their applicability to high-dimensional spaces is limited. Visualizing and interpreting phase transitions in such a high-dimensional space is exceptionally challenging. Conventional techniques for visualizing phase transitions in low-dimensional spaces become inadequate, and new methods for dimensionality reduction and data representation would be required. However, intelligent sampling techniques combined with clustering algorithms may identify regions or clusters of sequences that exhibit similar structural or functional properties. Phase transitions might be observable between these clusters or regions in the

reduced dimensionality space.

5.2.2 Genetic Constraints

The exploration of the theoretical sequence space, which encompasses all possible amino acid sequences, offers valuable insights into the structural changes in proteins. While missense mutations are a common mutational process, they are limited in their ability to access the entire theoretical sequence space. Other mutational mechanisms, such as insertions, deletions, duplications, and recombination, can also contribute to the exploration of a broader region of the vast sequence space. In our analysis, we saw that through missense mutations a large part of the sequence space can not be accessed. While some sequences within the theoretical space may never occur in nature due to biophysical constraints, their examination can still offer valuable insights into the fundamental principles governing protein folding and stability. Furthermore, investigating the overlap between functional sequence clusters and the regions of sequence space that can be explored through mutations at the DNA level could reveal constraints in the evolutionary process. Hence, the analysis of the theoretical sequence space, while challenging due to its vast size, holds promise for elucidating the potential for structural changes, the constraints on evolutionary trajectories, and the fundamental principles governing protein folding and stability.

Analyzing the genetically possible sequences revealed that their structural conformation differs from the theoretical sequences. This was not the case for all trajectories, but some exhibited significant differences. This difference, especially for intrinsic metrics such as pLDDT or SASA, indicates that the evolutionary constraints imposed by the genetic code and cellular machinery may introduce biases or constraints that shape the accessible conformational landscape. This divergence suggests that the theoretical sequence space may not fully capture the nuances of the biologically relevant protein folding dynamics.

5.2.3 Structural Predictions in the Context of Protein Folding Principles

The structural trends observed along the mutational trajectories can be contextualized within our existing knowledge of protein folding, stability, and sequence-structure relationships. The unique properties of each amino acid, such as size, charge, polarity, and ability to form specific interactions, influence the energy landscape and guide the folding process. Amino acids with small, nonpolar side chains like A are expected to maintain a relatively compact structure. In contrast, the rigid cyclic structure of P is well-documented to introduce kinks and disrupt regular secondary structures like α-helices. G, for example, is the smallest amino acid, with just a hydrogen atom as its side chain. This lack of a bulky side group allows glycine to fit into tight spaces. On the other hand, aromatic amino acids with bulky side chains, including F, Y, and W, are known to maintain structural stability through stacking interactions. The permutated sequences, which preserve amino acid composition while erasing evolutionary information, provide a valuable control to isolate the physical contributions to protein folding. Comparing the structural metrics of the permutated sequences to the mutational trajectories can reveal the extent to which amino acid properties versus the specific sequence order drive the observed trends.

ColabFold captures fundamental protein folding principles to a significant extent. For instance, the trajectories towards the homopolymers of G and V exhibit lower SASA values compared to those towards the homopolymers of H or W. This is because H and W have larger, aromatic side chains more likely to be exposed to the solvent, resulting in higher SASA values. Conversely, the homopolymer of I is expected to show a slower decline in TM-score than amino acids with smaller or more polar side chains. The hydrophobic interactions stabilized by the branched isoleucine side chains help maintain a more native-like structure as the sequence diverges from the reference. As illustrated in the Appendix in Figure A.8, higher TM scores are observed along the trajectory until a later point compared to other trajectories. Additionally, for most trajectories, the confidence in predicting sequences closer to the homopolymer increases again.

This can be attributed to the repeating sequence of the same amino acid, which leads to more regular and predictable structural patterns, simplifying the identification and modeling of local interactions. However, it is also notable that ColabFold sometimes fails to predict structures closer to the homopolymer, indicating a lack of suitable training data. In contrast, ESMFold struggles more with capturing these protein folding principles. The data generally show more fluctuations, and no clear trends are observed with ColabFold. Additionally, ESMFold predictions typically yield lower TM scores, making interpreting the results more challenging.

The observation that permutated sequences exhibit different structural characteristics than the original sequence can be partially attributed to the loss or alteration of evolutionary information. Sequence arrangement plays a crucial role in determining protein structure. Permutations can change these contexts, potentially erasing evolutionary adaptions specific to certain sequence arrangements. Overall, the structural variations predicted along the mutational trajectories largely align with our understanding of protein folding and stability derived from experimental studies and computational models. However, some deviations may highlight limitations in the training data or algorithms used by the structure prediction methods.

5.3 Performance of Structure Prediction Algorithms

In this thesis, we employed different algorithms to compare their ability to accurately predict structures across mutational trajectories. The ability to accurately predict protein structures across mutational trajectories is crucial for understanding the impact of sequence variations on folding and function. Different structure prediction algorithms have strengths and limitations in capturing these structural changes. In this thesis, we employed three different structure prediction algorithms: ESMFold, ColabFold, and AlphaFold2. ESMFold is a free modeling approach that can capture long-range interactions within protein sequences. It is fast and computationally cheap compared to other algorithms. However, we found that the predictions were less accurate than those of ColabFold. This was evident when looking at the RMSD and TM scores. Another disadvantage of ESMFold is its inability to predict structures of proteins that

are longer than 400 amino acid residues. While the accuracy might decrease for sequences diverging significantly from the training data, ESMFold does not rely on templates from the Protein Data Bank to predict the three-dimensional structure of sequences and will, therefore, also be able to predict structures with no similar structure found in nature.

ColabFold is a deep learning-based approach with impressive performance predicting structures for various proteins. It has shown that structural rearrangements are captured accurately for sequences close to the reference sequence, which can be attributed to its incorporation of evolutionary information in the prediction process. While it is computationally more expensive than ESMFold, it is still a scalable structure prediction algorithm for large-scale studies. Like ESMFold, ColabFold's accuracy decreased as the sequences diverged further from the reference sequence. AlphaFold2 has revolutionized the area of protein structure prediction. It has proven at CASP competitions that it can even capture complex structural features with accuracy that is believed to be identical to experimental validation. However, AlphaFold2 is computationally very expensive. For a single sequence of an ACE2 RBD variant, AlphaFold took 40 minutes to predict the three-dimensional structure, making it less suitable if not enough computational resources are available.

When scaling the number of sequences according to their prediction time in Table 3.3, we can estimate that, assuming all resources are available, ESMFold can predict 54,000 sequences per day, ColabFold can predict 847 sequences per day, and AlphaFold2 can predict 36 sequences per day. The process's parallelization will naturally accelerate the predictions; however, this scaling alone highlights the differences in the efficiency of various protein structure prediction algorithms. Moreover, additional resources are required to enable parallelization since multiple GPUs would be utilized. This computational bottleneck was observed during our utilization of the Snellius supercomputer, where we encountered significant queuing times of up to 3 days for a batch of 50 sequences when employing AlphaFold2. Since ColabFold and ESMFold are less computationally expensive and require fewer GPUs, these issues were encountered less frequently. Consequently, the availability of resources, both on supercomputers and through

project funding, significantly influences the number of three-dimensional protein structures that can be predicted using the chosen algorithm.

As sequences diverge from the reference sequence, capturing structural variations becomes increasingly tricky. ColabFold has demonstrated superior performance in capturing structural rearrangements and conformational changes compared to ESMFold. Its incorporation of evolutionary information and physical principles allows it to better model the impact of mutations on protein folding. However, it is essential to note that even these advanced algorithms may need help with highly divergent sequences or sequences that deviate significantly from the training data. In such cases, the predictions become less reliable, and additional experimental validation or refinement may be necessary. ColabFold, for example, failed to predict some of the structures close to the homopolymers. This is one limitation of the template-based modeling approaches because they ultimately rely on fitting templates from data banks. Hence, our ability to explore sequence space is limited by our current understanding of existing, biologically evolved proteins. As sequences approach homopolymers, less suitable templates are available in protein data banks. This introduces evolutionary bias, indicating that large portions of theoretical sequence space remain unrepresented in our structural databases. Increasing the diversity of protein structures in databases by experimentally determining the structures of designed proteins can help improve the performance of template-based methods.

In addition to the algorithms employed in this thesis, there are other approaches for protein structure prediction, such as Rosetta and I-TASSER. Rosetta combines physics-based energy functions and sampling methods to predict protein structures, while I-TASSER (Iterative Threading ASSEmbly Refinement) combines multiple threading programs, structural assembly simulations, and optimization techniques. These algorithms have their strengths and limitations, and the choice of algorithm may depend on the specific requirements of the study, such as computational resources, accuracy requirements, and the availability of homologous structures. Since Rosetta employs physical constraints, it would be a suitable structure prediction tool to investigate the physical properties of protein folding.

In summary, while ESMFold provides efficient and fast predictions, ColabFold has demonstrated superior performance in capturing structural variations across mutational trajectories. However, all algorithms struggle with highly divergent sequences or those deviating significantly from the training data. Integrating multiple approaches, experimental validation, and continued algorithm development will be crucial for improving the accuracy of structure predictions across diverse mutational landscapes. While the patterns observed across the trajectories can be attributed to the sequences' biophysical features, they can also be attributed to the limitations of the algorithms. The differences between the algorithms emphasize each algorithm's unique strengths and weaknesses emphasizing the importance of algorithmic design in improving the accuracy and dependability of protein structure predictions. As a result, the observed structural patterns may not only represent the intrinsic properties of the sequences but also the specific biases and limits of the used prediction algorithms. The concept of combining multiple weather prediction systems or models to form a more robust ensemble prediction system can be applied to protein structure prediction algorithms as well. Just like weather forecasting benefits from considering an ensemble of models to account for uncertainties and leverage the strengths of different approaches, integrating multiple structure prediction algorithms can improve the accuracy and robustness of protein structure predictions.

5.4 Ethical Concerns

Although the advancements in protein structure prediction algorithms are promising for medical and scientific advancements, people have raised concerns about ethical considerations. First are the potential risks that are posed when misused for, for example, bio-weapon development. These algorithms can potentially be misused to design novel pathogens or to enhance the virulence of existing ones. Looking at the history of biological warfare, their danger to the population becomes apparent [5]. The dual dilemma of, on the one hand, having a powerful tool to enhance the field of medicine and science in general, but on the other hand, the danger of misuse necessitates strict oversight and regulation. This calls for the evolvement of inter-

national organizations, such as the Biological Weapons Convention (BWC), to address these emerging threats. Another concern that is raised is privacy and data governance concerns. When sensitive data is mishandled, this could lead to breaches of privacy. Hence, employing robust data governance frameworks that can protect individual privacy while enabling scientific advancements is of utmost importance. There have to be policies that regulate these data governance concerns to regulate data access, patient consent, and data ownership. The FAIR principles, for example, promote a foundation for ensuring a robust data governance framework. Additionally, inaccurate protein structure predictions can have profound implications in critical settings such as drug design. The developers of structure prediction algorithms are responsible for ensuring high accuracy and otherwise transparently stating the limitations of their algorithms. This includes rigorous validation of their models, continuously updating them, and providing users with manuals on accurately interpreting their results. In science, we still have an underrepresentation of diverse genetic backgrounds in data. This could lead to biases and inaccuracies in predictions for underrepresented populations. Thus, a more inclusive data collection that includes a wide range of genetic, environmental, and demographic factors will improve the robustness and fairness of the algorithms. The rapid advancements in protein structure prediction algorithms offer immense potential for scientific and medical breakthroughs. However, they also raise significant ethical, social, and governance challenges. Addressing these challenges requires a multifaceted approach involving continuous regulation, equitable access, moral responsibility, and inclusive data practices. By proactively addressing these issues, we can utilize the full potential of these technologies while mitigating risks and ensuring that the benefits are distributed equitably across all regions and populations.

5.5 Future Directions

The work presented in this thesis opens up various directions for further researching protein sequence spaces and the folding principles of proteins. Despite the interesting insights made, several open questions remain and require further investigation. We propose the following avenues for future research:

1. **Functional Implications:** Further research is needed to correlate structural changes with functional consequences. This could involve experimental validation of predicted structures to assess their biological activity and relevance. Computational techniques, such as machine learning models or molecular dynamics simulations, can be employed to predict the structural and functional consequences of mutations along these trajectories. One computational method that can be employed to predict the functional effects of sequence variations is AlphaPulldown [86]. AlphaPulldown leverages AlphaFold's high-accuracy structural predictions to assess the binding affinity and overall functionality of protein complexes. It is particularly useful in the context of protein-protein interactions. However, experimental validation is crucial for validating the computational predictions and for refining our understanding of the sequence-structure-function relationships. The integration of computational and experimental approaches can provide a more comprehensive picture of protein sequence space, facilitating the exploration of new functional sequences. This approach holds great promise for advancing our fundamental understanding of protein folding and function.
2. **Genetic Constraint on ACE2 RBD:** We have examined the effects of single nucleotide substitutions on amino acid mutations. However, the genetic landscape of viral evolution encompasses a broader spectrum of modifications, including insertions and deletions (indels). Incorporating these additional forms of genetic variation into our analysis would be an extension of this experiment. This would allow for a more comprehensive exploration of the genetically accessible sequence space and its physical constraints.
3. **Genetic Algorithms:** Genetic algorithms (GAs) offer a promising approach to further explore the protein sequence space. GAs can efficiently search for optimal sequences by mimicking the process of natural selection, allowing for the identification of sequences with desired properties. It will be insightful to take the permutated sequences and use a GA that will optimize the thermostability so that the permutated sequences reach the same or similar fold to the reference sequence. Furthermore, incorporating GAs with current structure prediction tools can enhance the exploration process by iteratively mutating and selecting sequences based on their predicted structural stability and functional

relevance. This approach can be particularly useful for identifying rare or highly optimized sequences that may not be easily accessible through traditional methods.

4. **Expanding the Datasets:** Expanding the current datasets to more sequences will shed more light on the folding mechanisms within sequence space. One way to expand the mutational trajectories of the ACE2 RBD is by randomly mutating more positions at each step on the trajectory, which will result in not only the 20 trajectories but also at each step on the trajectories (in total there are 3705 sequences on the trajectories) we propose to have ten more sequences that are mutated. This will result in a dataset of 37,050 more sequences that venture out more into the space surrounding the trajectories. Furthermore, since we now only focus on single-nucleotide mutations, another possible exploration would be to include deletions or insertions and investigate their effect on the folding mechanisms of proteins. We could further include a broader range of proteins with varying lengths, complexities, and functions to capture the diversity within sequence space and to compare this behavior to the sequence space of ACE2 RBD. Expanding this exploration to include post-translational modifications and their impact on protein structure could offer deeper insights into functional sequence clusters.
5. **Intermediate States and Phase Transitions:** The identification and characterization of intermediate states and phase transitions are crucial for a comprehensive understanding of protein folding. While this will be challenging due to the high dimensionality of the sequence space, it provides invaluable insights into the folding mechanisms and energy landscapes. By expanding the datasets, more data on three-dimensional structures will be available in the sequence space, which might provide clearer insights into intermediate states and phase transitions.
6. **Using other Structure Prediction Algorithms:** Using different structure prediction algorithms and combining their strength will enhance our understanding of sequence-structure relationship in protein sequence space. Expanding the structure prediction algorithms selection will reveal different insights. Sequence-based structure prediction methods, such as RoseTTAFold, can indirectly investigate the effects of environmental

conditions on protein folding and stability by predicting structures for different mutant sequences. These sequence changes can mimic the effects of environmental conditions on folding by altering charge distributions, hydrophobicity, etc.

7. **Enhancing Computational Throughput:** Larger datasets require more computational resources, especially for computationally expensive Algorithms like AlphaFold2. Utilizing high-throughput clusters with more resources will result in a faster and more reliable method to apply computationally expensive structure prediction algorithms and will enable us to investigate larger parts of the sequence space.
8. **Improvement of Protein Structure Prediction Algorithms:** The field of protein structure prediction algorithms is rapidly developing and evolving. This will undoubtedly revolutionize our understanding of the protein universe. Competitions like the Critical Assessment of Protein Structure Prediction (CASP) provide a competitive environment where research groups strive to present their state-of-the-art algorithms. The upcoming CASP16, scheduled for early December 2024, will serve as a platform for evaluating the latest developments in the field. In May of 2024, Deepmind announced the development of AlphaFold 3.0, which is now able to predict protein-molecule complexes [75]. Although AlphaFold 3.0 is still restricted to non-commercial use, it holds great promise for future developments. However, for these algorithms to achieve broader adoption, they must address scalability challenges and improve usability for a wider spectrum of users.
9. **Exploring Trajectories Between Sequences of Interest:** A particularly promising application of the approach presented in this thesis lies in investigating trajectories between specific sequences of interest. This method could be especially valuable in the context of viral evolution, as seen by the SARS-CoV-2 pandemic. While the entirety of sequence space is astronomically large, we can efficiently navigate the paths between two sequences in a maximum of N steps, where N is the length of the sequences. This method could help identify critical mutations that significantly alter protein structure or function, and might also identify potential variants of concern before they emerge in the population. This approach further allows for the comparison of mutational trajecto-

ries between different viral strains or proteins, potentially revealing common patterns or divergent evolutionary strategies.

Chapter 6

Conclusion

The exploration of protein sequence space and folding principles presented in this thesis has illuminated the intricate relationship between sequence variations and structural outcomes. By analyzing mutational trajectories and examining structural trends across diverse sequences, we have gained valuable insights into the behavior of protein folding as mutations accumulate. These findings enhance our understanding of the principles governing protein structure and provide a roadmap for further exploration and discovery in this field. By mapping protein sequence space through directed mutational trajectories, we have systematically explored the vast landscape of possible sequences, shedding light on the distribution of structural properties within this space. This approach has allowed us to identify transition points on the trajectories that govern protein folding, offering a deeper understanding of the complexities inherent in protein structure prediction.

The analysis of structural trends across mutational trajectories has revealed the impact of accumulating amino acid substitutions on protein folding. By observing changes in residue-residue contact points and structural compactness as sequences diverge from a reference sequence, we have gained insights into the unfolding process and the emergence of alternative structural conformations. These observations underscore the complexity of protein folding landscapes and highlight the diverse structural patterns that arise as sequences deviate from their original state. In addition to the scientific advancements made, we consider FAIR Data management to

be an integral component of our research methodology, ensuring that our findings are Findable, Accessible, Interoperable, and Reusable for the broader scientific community. Addressing these challenges requires the promotion of inclusive data collection and analysis methods, moral responsibility in research practices as well as equitable access to the algorithms.

Looking ahead, future research directions should focus on exploring intermediate states, phase transitions, and the application of genetic algorithms to unravel the complexities of protein folding mechanisms and energy landscapes further. Furthermore, this method can be applied to any other 'trajectory' in the sequence space, such as a mutational trajectory from one SARS-CoV-2 variant of concern to another. By delving deeper into these areas, we can expand our knowledge of the evolution of proteins and unlock new possibilities for scientific and medical breakthroughs.

In conclusion, this thesis is a stepping stone for ongoing research in protein sequence space and folding principles. By embracing the insights gained from this study and addressing the associated challenges, we can pave the way for a future where advanced protein structure prediction algorithms are integral to pandemic preparedness efforts, benefiting society as a whole.

Bibliography

[1] Gustaf Ahdritz et al. "OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization". In: *bioRxiv* (2022). DOI: 10.1101/2022.11.20.517210. eprint: https://www.biorxiv.org/content/early/2022/11/22/2022.11.20.517210.full.pdf. URL: https://www.biorxiv.org/content/10.1101/2022.11.20.517210.

[2] Bruce Alberts et al. *Molecular Biology of the Cell*. Garland Science, 2002.

[3] Muhammad Asif Ali and Gustavo Caetano-Anollés. "AlphaFold2 Reveals Structural Patterns of Seasonal Haplotype Diversification in SARS-CoV-2 Spike Protein Variants". In: *Biology* 13.3 (2024), p. 134.

[4] Protein Data Bank. "Protein data bank". In: *Nature New Biol* 233.223 (1971), pp. 10–1038.

[5] Vincent Barras and Gilbert Greub. "History of biological warfare and bioterrorism". In: *Clinical Microbiology and infection* 20.6 (2014), pp. 497–502.

[6] Letícia MF Bertoline et al. "Before and after AlphaFold2: An overview of protein structure prediction". In: *Frontiers in Bioinformatics* 3 (2023), p. 1120370.

[7] Daria Beshnova et al. "Computational approach for binding prediction of SARS-CoV-2 with neutralizing antibodies". In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 2212–2222.

[8] National Center for Biotechnology Information (NCBI). *Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome*. Nucleotide Database Record. Accessed: May 14, 2024, 10:06 CET. Bethesda (MD), 2020. URL: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2.

[9] Matteo Bisardi et al. "Modeling sequence-space exploration and emergence of epistatic signals in protein evolution". In: *Molecular biology and evolution* 39.1 (2022), msab321.

[10] Jesse Bloom and lab members. *SARS2-RBD-escape-mutants: designed escape mutants of the RBD of SARS-CoV-2 that contain potential future antigenic mutations.* 2024. URL: https://github.com/jbloomlab/SARS2-RBD-escape-mutants.

[11] Brantly Callaway and Pedro HC Sant'Anna. "Difference-in-differences with multiple time periods". In: *Journal of econometrics* 225.2 (2021), pp. 200–230.

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

[13] Ben E Clifton, Dan Kozome, and Paola Laurino. "Efficient exploration of sequence space by sequence-guided protein engineering and design". In: *Biochemistry* 62.2 (2022), pp. 210–220.

[14] UniProt Consortium. *UniProt: the universal protein knowledgebase in 2021.* UniProt website, https://www.uniprot.org/uniprotkb/P0DTC2/entry. Accessed: February 14, 2024, 17:38 CET. 2021.

[15] Loris Di Cairano et al. "Topological origin of the protein folding transition". In: *Physical Review E* 106.5 (2022), p. 054134.

[16] Ken A Dill. "Dominant forces in protein folding". In: *Biochemistry* 29.31 (1990), pp. 7133–7155.

[17] Ken A Dill and Justin L MacCallum. "The protein-folding problem, 50 years on". In: *science* 338.6110 (2012), pp. 1042–1046.

[18] David TF Dryden, Andrew R Thomson, and John H White. "How much of protein sequence space has been explored by life on Earth?" In: *Journal of The Royal Society Interface* 5.25 (2008), pp. 953–956.

[19] Nicholas S Edmunds, Liam J McGuffin, and Ahmet G Genc. "Benchmarking of AlphaFold2 accuracy self-estimates as empirical quality measures and model ranking indicators and their comparison with independent model quality assessment programs." In: *bioRxiv* (2023), pp. 2023–12.

[20] Nature Education. *Protein Structure*. https://www.nature.com/scitable/topicpage/protein-structure-14122136/. Accessed: April 15, 2024, 18:30 CET.

[21] Facebook Research. *ESM (Evolutionary Scale Models) Repository*. https://github.com/facebookresearch/esm. Accessed: February 14, 2024, 14:30 CET.

[22] Alan Fersht. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. Macmillan, 1999.

[23] Erin D Foster and Ariel Deardorff. "Open science framework (OSF)". In: *Journal of the Medical Library Association: JMLA* 105.2 (2017), p. 203.

[24] Google DeepMind. *AlphaFold GitHub Repository*. https://github.com/google-deepmind/alphafold. Accessed: March 12, 2024, 09:38 CET.

[25] Daria Gutnik et al. "Using AlphaFold predictions in viral research". In: *Current Issues in Molecular Biology* 45.4 (2023), pp. 3705–3732.

[26] Thomas Hamelryck and Bernard Manderick. "PDB parser and structure class for the Biopython molecular biology library". In: *Bioinformatics* 19.17 (2003), pp. 2308–2310.

[27] Bin Huang et al. "Protein structure prediction: challenges, advances, and the shift of research paradigms". In: *Genomics, Proteomics & Bioinformatics* (2023).

[28] L Steven Johnson, Sean R Eddy, and Elon Portugaly. "Hidden Markov model speed heuristic and iterative HMM search procedure". In: *BMC bioinformatics* 11 (2010), pp. 1–8.

[29] David T Jones et al. "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". In: *Bioinformatics* 28.2 (2012), pp. 184–190.

[30] John Jumper et al. *Computational predictions of protein structures associated with COVID-19*. Version 3. DeepMind website. Aug. 2020. URL: https://deepmind.google/discover/blog/computational-predictions-of-protein-structures-associated-with-covid-19/.

[31] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[32] Wolfgang Kabsch. "A solution for the best rotation to relate two sets of vectors". In: *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 32.5 (1976), pp. 922–923.

[33] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[34] Brian Kuhlman and Philip Bradley. "Advances in protein structure prediction and design". In: *Nature reviews molecular cell biology* 20.11 (2019), pp. 681–697.

[35] Tobias Kuhn et al. "Semantic micro-contributions with decentralized nanopublication services". In: *PeerJ Computer Science* 7 (2021), e387.

[36] Jesse Bloom Lab. *jbloomlab GitHub Organization*. https://github.com/jbloomlab. Accessed: March 16, 2024, 12:39 CET.

[37] Jun Lan et al. "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor". In: *nature* 581.7807 (2020), pp. 215–220.

[38] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130.

[39] Vladimir N Maiorov and Gordon M Crippen. "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins". In: *Journal of molecular biology* 235.2 (1994), pp. 625–634.

[40] Valerio Mariani et al. "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests". In: *Bioinformatics* 29.21 (2013), pp. 2722–2728.

[41] Debora S Marks et al. "Protein 3D structure computed from evolutionary sequence variation". In: *PloS one* 6.12 (2011), e28766.

[42] Erik W Martin and Tanja Mittag. "Relationship of sequence and phase separation in protein low-complexity regions". In: *Biochemistry* 57.17 (2018), pp. 2478–2487.

[43] Michigan State University. *Protein Structure and Function*. https://openbooks.lib.msu.edu/isb202/chapter/protein-structure-and-function/. Accessed: April 15, 2024, 19:37 CET.

[44] Milot Mirdita et al. "Uniclust databases of clustered and deeply annotated protein sequences and structures". In: *Nucleic acids research* 45.D1 (2017), pp. D170–D176.

[45] Milot Mirdita et al. "ColabFold: making protein folding accessible to all". In: *Nature methods* 19.6 (2022), pp. 679–682.

[46] Alex L Mitchell et al. "MGnify: the microbiome analysis resource in 2020". In: *Nucleic acids research* 48.D1 (2020), pp. D570–D578.

[47] Barend Mons, Jan Velterop, et al. "Nano-publication in the e-science era". In: *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*. Vol. 523. sn. 2009.

[48] Marshall W Nirenberg and J Heinrich Matthaei. "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides". In: *Proceedings of the National Academy of Sciences* 47.10 (1961), pp. 1588–1602.

[49] Subash C Pakhrin et al. "Deep learning-based advances in protein structure prediction". In: *International journal of molecular sciences* 22.11 (2021), p. 5553.

[50] Chun-Xiang Peng et al. "Recent Advances and Challenges in Protein Structure Prediction". In: *Journal of Chemical Information and Modeling* 64.1 (2023), pp. 76–95.

[51] *Prediction Center for Protein Structure Prediction*. https://predictioncenter.org/. Accessed: April 16, 2024, 13:07 CET.

[52] Michael Remmert et al. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment". In: *Nature methods* 9.2 (2012), pp. 173–175.

[53] Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank. *Mol* 3D Viewer*. https://www.rcsb.org/3d-view. Accessed: 2024-05-26. 2023.

[54] Ambrish Roy, Alper Kucukural, and Yang Zhang. "I-TASSER: a unified platform for automated protein structure and function prediction". In: *Nature protocols* 5.4 (2010), pp. 725–738.

[55] Humna Sajjad et al. "Origin, pathogenesis, diagnosis and treatment options for SARS-CoV-2: a review". In: *Biologia* 76.9 (2021), pp. 2655–2673.

[56] Sirko Schindler et al. "The I-ADOPT Interoperability Framework for FAIRer data descriptions of biodiversity". In: *arXiv preprint arXiv:2107.06547* (2021).

[57] Lilli Schuckert, Erik A Schultes, and Max van den Boom. *Mapping Protein Sequence Space*. June 2024. DOI: https://doi.org/10.17605/OSF.IO/PMJ9X.

[58] Erik Schultes et al. "Fair digital twins for data-intensive research". In: *Frontiers in big data* 5 (2022), p. 883341.

[59] Erik Schultes et al. "The comparative anatomy of nanopublications and FAIR digital objects". In: *Research Ideas and Outcomes* 8 (2022), e94150.

[60] Erik A Schultes, Peter T Hraber, and Thomas H LaBean. "No molecule is an island: molecular evolution and the study of sequence space". In: *Algorithmic Bioprocesses* (2009), pp. 675–704.

[61] Erik A Schultes et al. *Putting FDOs to Work*. Mar. 2024. URL: https://osf.io/zpbwy/.

[62] Erik A Schultes et al. *Step 3: Project Execution*. June 2024. URL: osf.io/r2u3t.

[63] IN Shindyalov, NA Kolchanov, and CHRIS Sander. "Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?" In: *Protein Engineering, Design and Selection* 7.3 (1994), pp. 349–358.

[64] Andrew Shrake and John A Rupley. "Environment and exposure to solvent of protein atoms. Lysozyme and insulin". In: *Journal of molecular biology* 79.2 (1973), pp. 351–371.

[65] Lok Bahadur Shrestha, Nicodemus Tedla, and Rowena A Bull. "Broadly-neutralizing antibodies against emerging SARS-CoV-2 variants". In: *Frontiers in immunology* 12 (2021), p. 752003.

[66] Martin Steinegger and Johannes S"oding. "MMseqs2 cluster and search proteins". In: *Nature methods* 16.12 (2019), pp. 1044–1046.

[67] Martin Steinegger and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature biotechnology* 35.11 (2017), pp. 1026–1028.

[68] Martin Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". In: *BMC bioinformatics* 20 (2019), pp. 1–15.

[69] María Suárez and Alfonso Jaramillo. "Challenges in the computational design of proteins". In: *Journal of the Royal Society Interface* 6.suppl_4 (2009), S477–S491.

[70] *SURF*. https://www.surf.nl/en. Accessed: April 17, 2024, 08:36 CET. 2024.

[71] SURF. *Snellius Partitions and Accounting*. https://servicedesk.surf.nl/wiki/display/WIKI/Snellius+partitions+and+accounting. Accessed: April 17, 2024, 15:39 CET.

[72] *SURF Services and Rates 2024*. https://www.surf.nl/files/2023-08/surf-services-and-rates-2024_version-aug-2023.pdf. Accessed: April 16, 2024, 21:27 CET. Aug. 2023.

[73] Joel L Sussman et al. "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules". In: *Acta Crystallographica Section D: Biological Crystallography* 54.6 (1998), pp. 1078–1084.

[74] Vito Thijssen et al. "A broad-spectrum macrocyclic peptide inhibitor of the SARS-CoV-2 spike protein". In: *Proceedings of the National Academy of Sciences* 120.26 (2023), e2303292120.

[75] Benjamin Thompson et al. "Alphafold 3.0: the AI protein predictor gets an upgrade." In: *Nature* (2024).

[76] UCSF ChimeraX Development Team. *ESMFold Documentation*. https://www.cgl.ucsf.edu/chimerax/docs/user/tools/esmfold.html. Accessed: February 14, 2024, 14:30 CET.

[77] UniProt Consortium. *UniProtKB Entry for Q9BYF1*. https://www.uniprot.org/uniprotkb/Q9BYF1/entry. Accessed: March 06, 2024, 12:34 CET.

[78] Mihaly Varadi et al. "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences". In: *Nucleic Acids Research* 52.D1 (2024), pp. D368–D375.

[79] Alexandra C Walls et al. "Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion". In: *Proceedings of the National Academy of Sciences* 114.42 (2017), pp. 11157–11162.

[80] James D Watson and Francis HC Crick. "The structure of DNA". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 18. Cold Spring Harbor Laboratory Press. 1953, pp. 123–131.

[81] David Whitford. *Proteins: structure and function*. John Wiley & Sons, 2013.

[82] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9.

[83] Pawel P Wozniak and Malgorzata Kotulska. "Characteristics of protein residue-residue contacts and their application in contact prediction". In: *Journal of Molecular Modeling* 20 (2014), pp. 1–17.

[84] Zunyou Wu and Jennifer M McGoogan. "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention". In: *jama* 323.13 (2020), pp. 1239–1242.

[85] Jinsung Yang et al. "Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor". In: *Nature communications* 11.1 (2020), p. 4541.

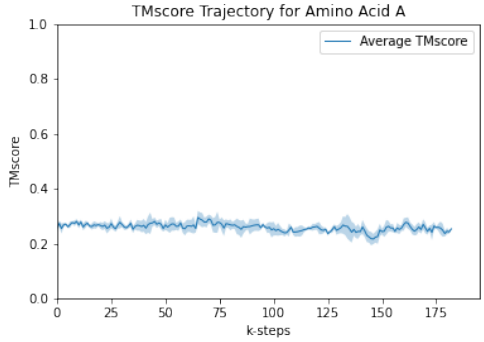
[86] Dingquan Yu et al. "AlphaPulldown—a python package for protein–protein interaction screens using AlphaFold-Multimer". In: *Bioinformatics* 39 (1 2023). DOI: https://doi.org/10.1093/bioinformatics/btac749.

[87] Quan Zhang et al. "Application of the AlphaFold2 Protein Prediction Algorithm Based on Artificial Intelligence". In: *Journal of Theory and Practice of Engineering Science* 4.02 (2024), pp. 58–65.

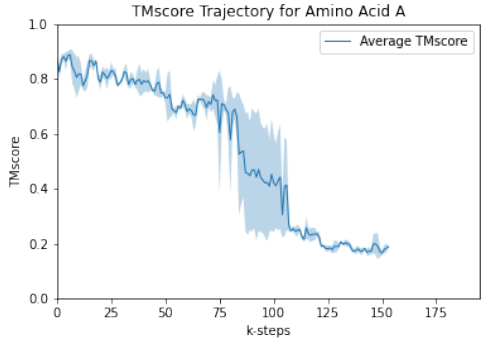
[88] Yang Zhang and Jeffrey Skolnick. "Scoring function for automated assessment of protein structure template quality". In: *Proteins: Structure, Function, and Bioinformatics* 57.4 (2004), pp. 702–710.

Appendix A

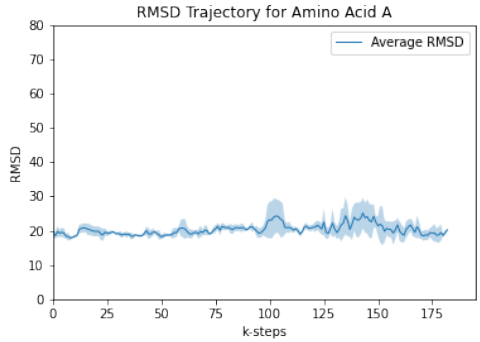
Trajectory Plots



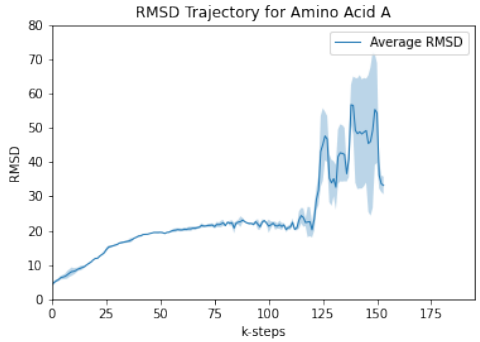
(a) ESMFold TM Score



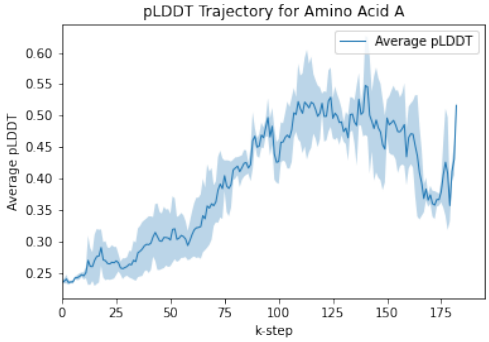
(b) ColabFold TM Score



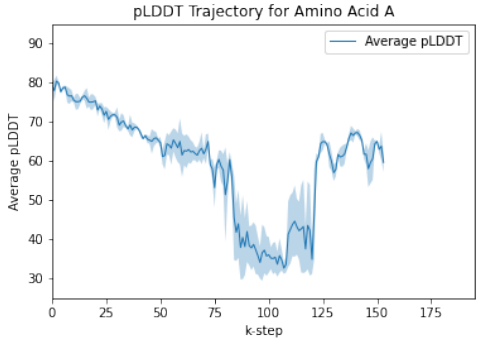
(c) ESMFold RMSD



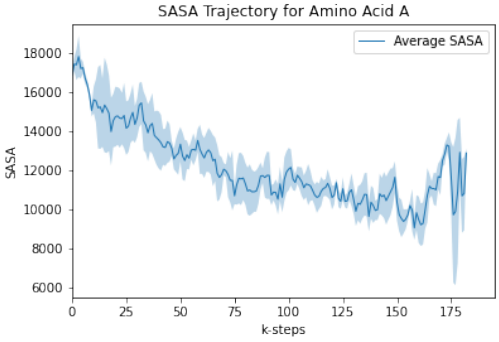
(d) ColabFold RMSD



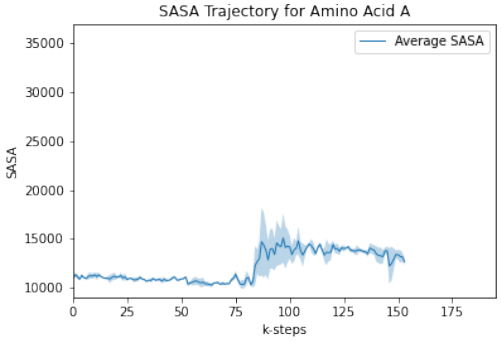
(e) ESMFold pLDDT



(f) ColabFold pLDDT

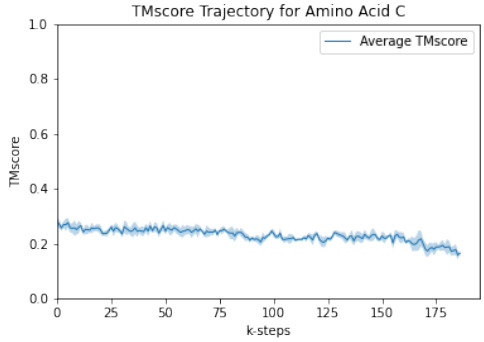


(g) ESMFold SASA

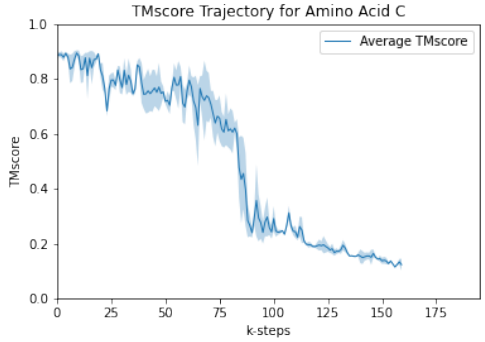


(h) ColabFold SASA

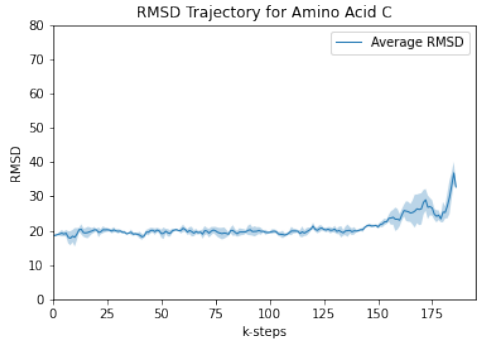
Figure A.1: Trajectory A. Comparison of the structural metrics along trajectory A from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



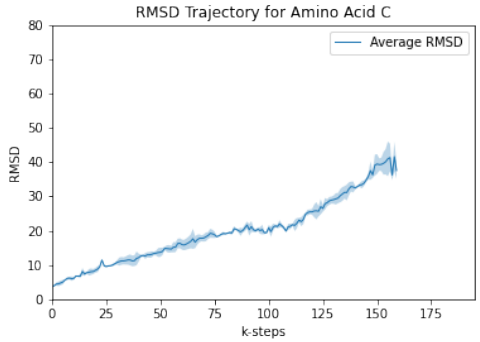
(a) ESMFold TM Score



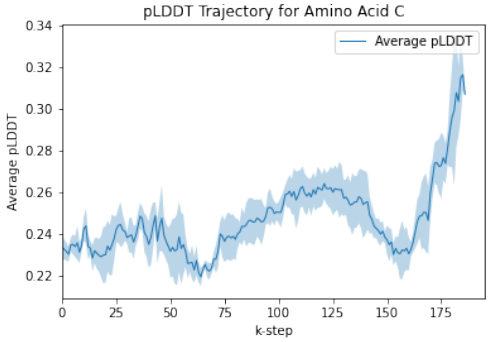
(b) ColabFold TM Score



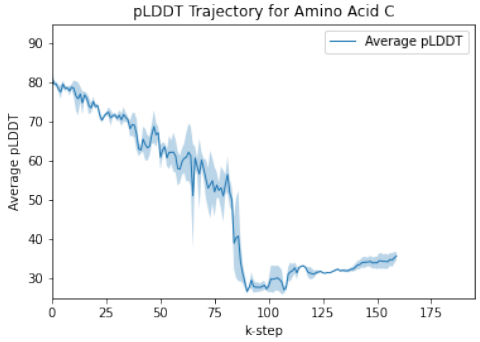
(c) ESMFold RMSD



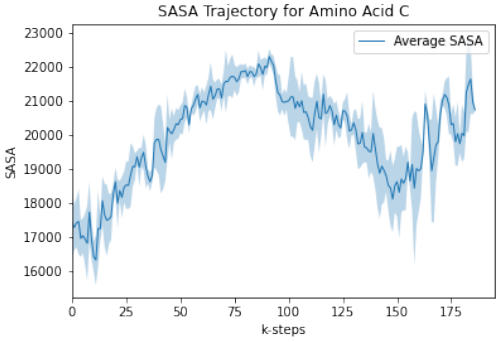
(d) ColabFold RMSD



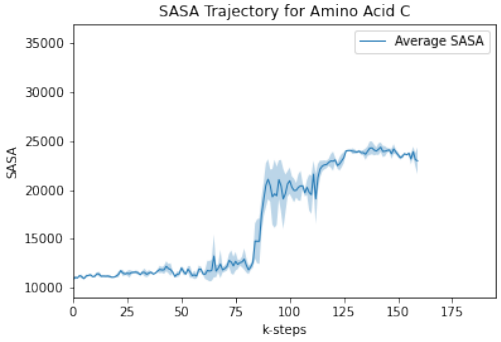
(e) ESMFold pLDDT



(f) ColabFold pLDDT

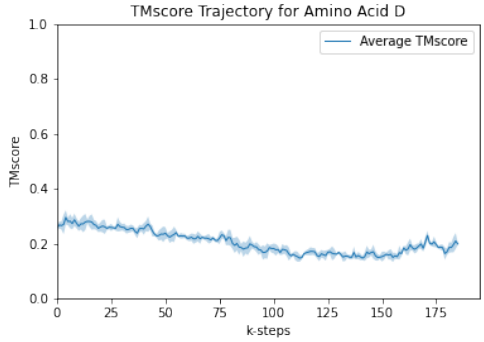


(g) ESMFold SASA

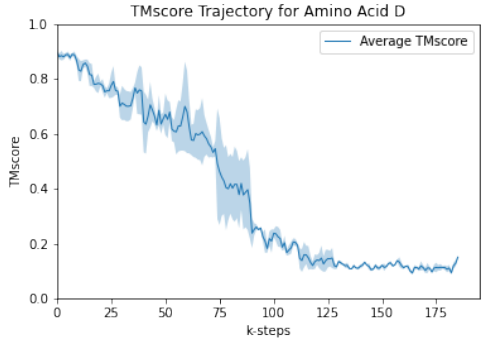


(h) ColabFold SASA

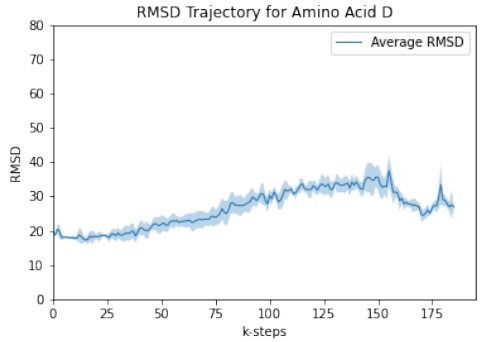
Figure A.2: Trajectory C. Comparison of the structural metrics along trajectory C from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



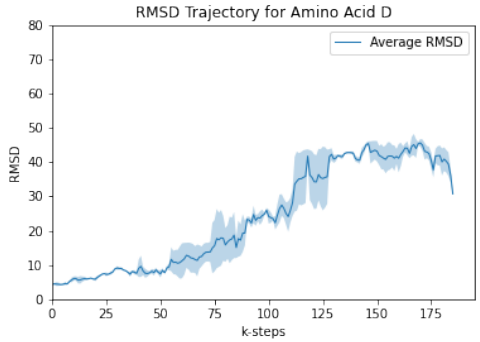
(a) ESMFold TM Score



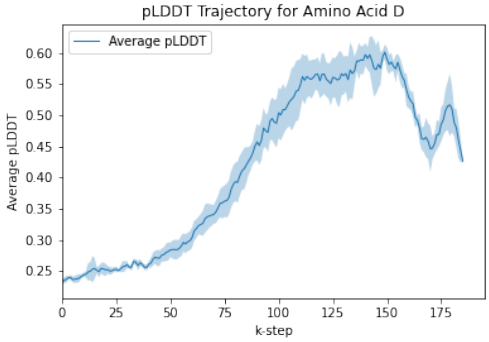
(b) ColabFold TM Score



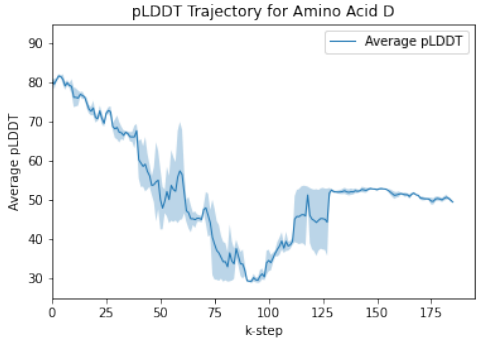
(c) ESMFold RMSD



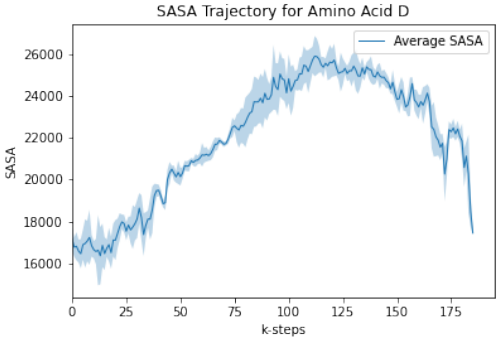
(d) ColabFold RMSD



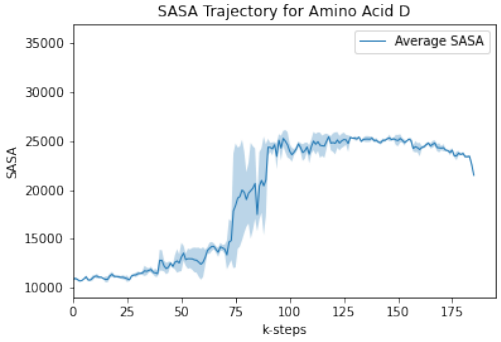
(e) ESMFold pLDDT



(f) ColabFold pLDDT

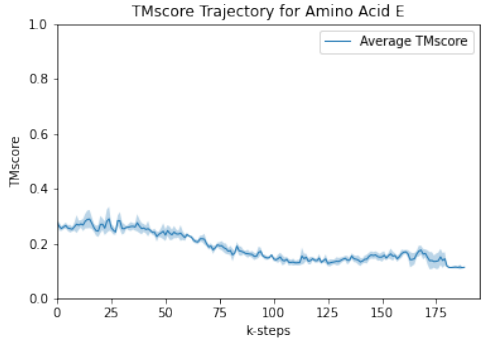


(g) ESMFold SASA

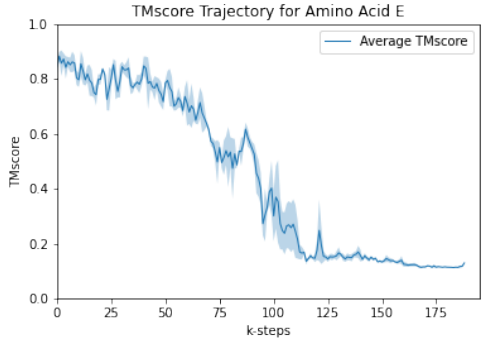


(h) ColabFold SASA

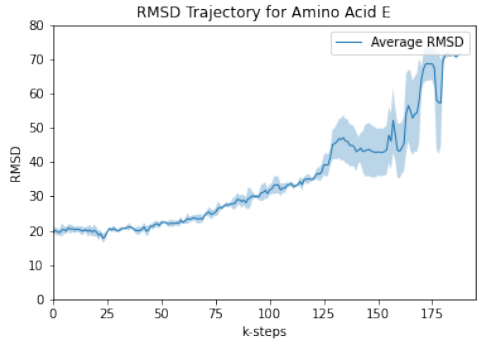
Figure A.3: Trajectory D. Comparison of the structural metrics along trajectory D from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



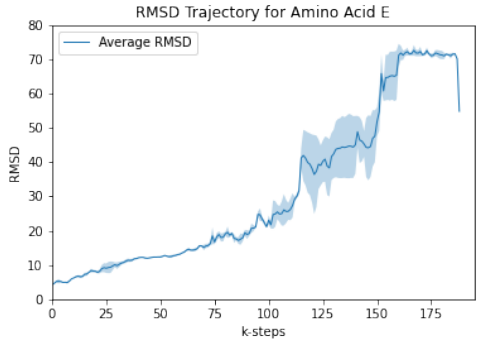
(a) ESMFold TM Score



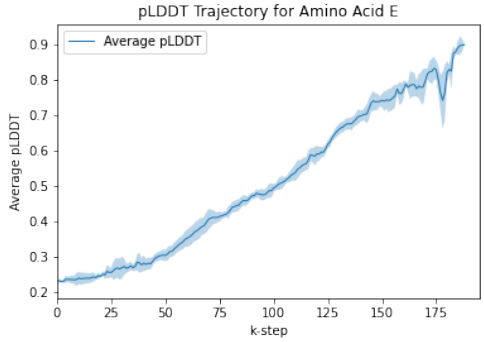
(b) ColabFold TM Score



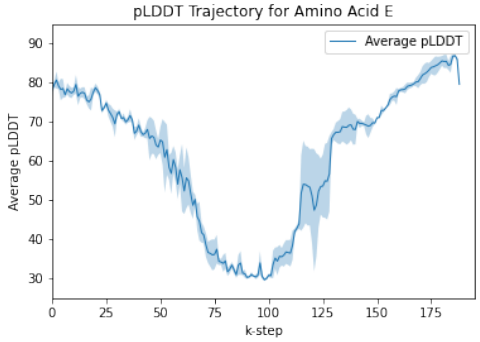
(c) ESMFold RMSD



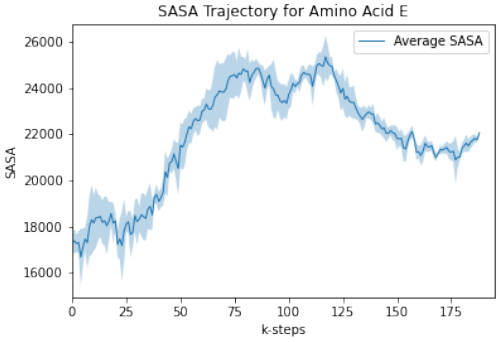
(d) ColabFold RMSD



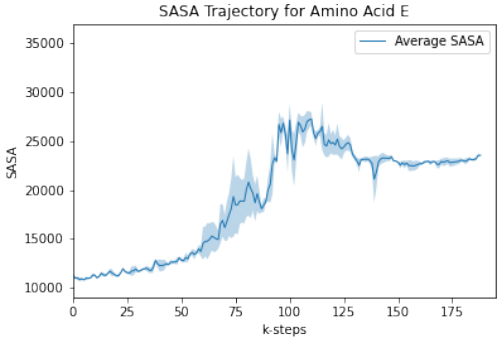
(e) ESMFold pLDDT



(f) ColabFold pLDDT

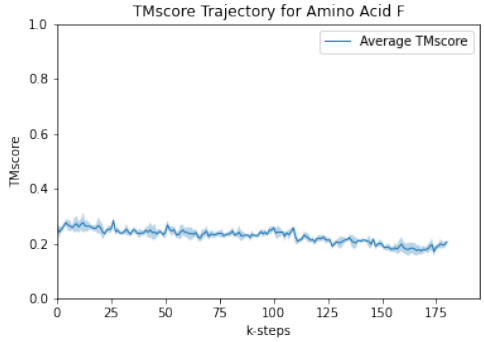


(g) ESMFold SASA

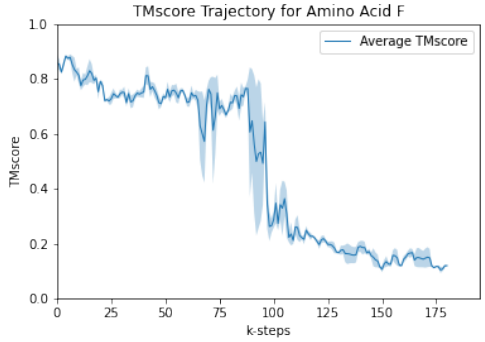


(h) ColabFold SASA

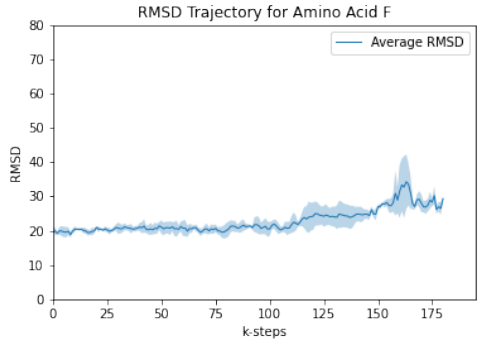
Figure A.4: Trajectory E. Comparison of the structural metrics along trajectory E from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



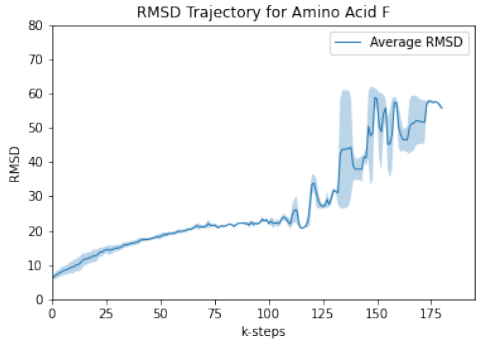
(a) ESMFold TM Score



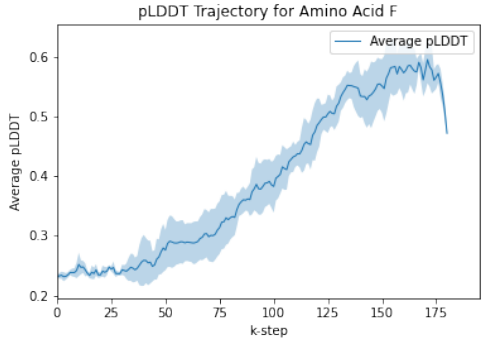
(b) ColabFold TM Score



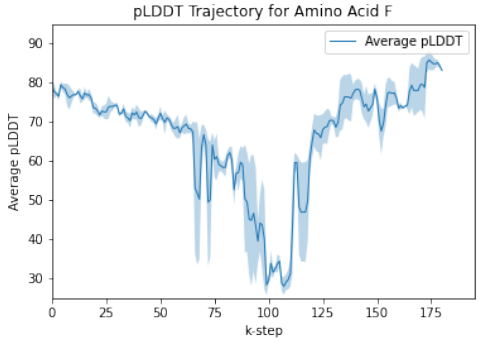
(c) ESMFold RMSD



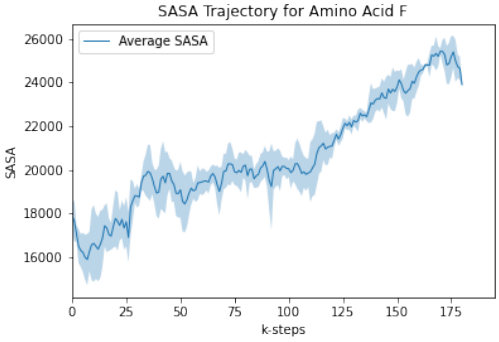
(d) ColabFold RMSD



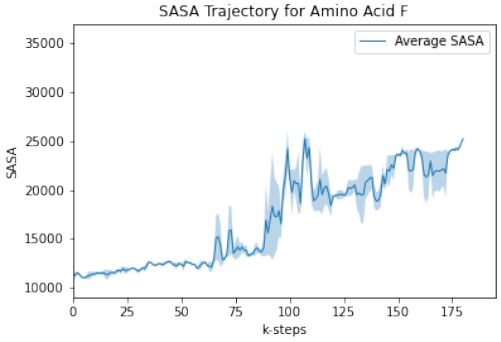
(e) ESMFold pLDDT



(f) ColabFold pLDDT

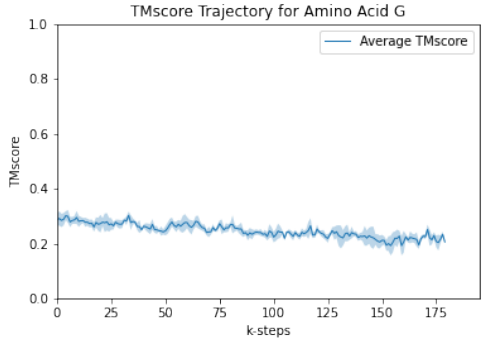


(g) ESMFold SASA

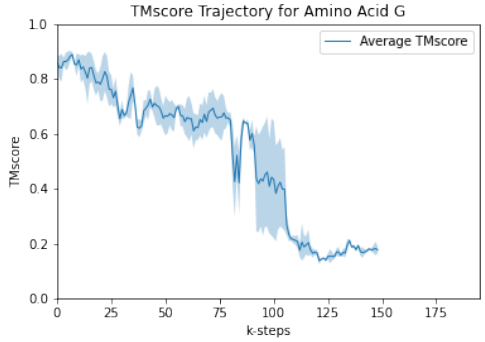


(h) ColabFold SASA

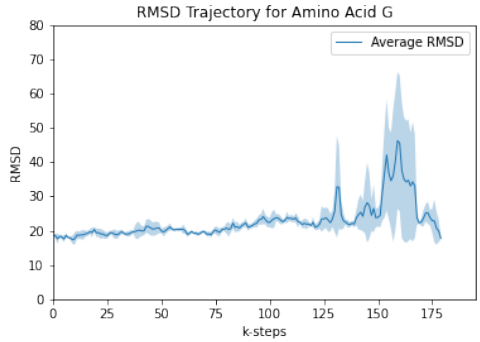
Figure A.5: Trajectory F. Comparison of the structural metrics along trajectory F from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



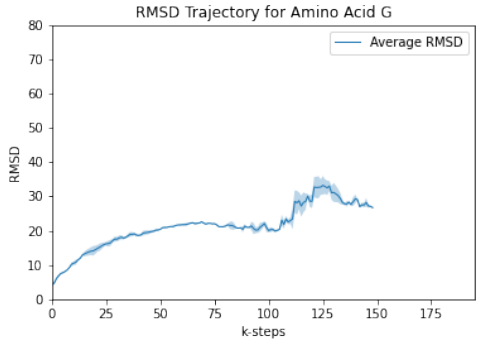
(a) ESMFold TM Score



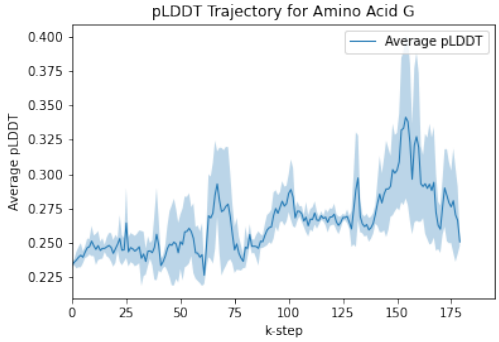
(b) ColabFold TM Score



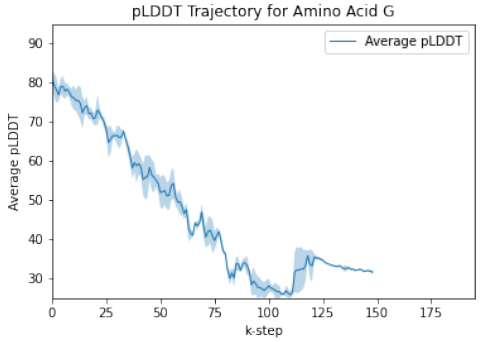
(c) ESMFold RMSD



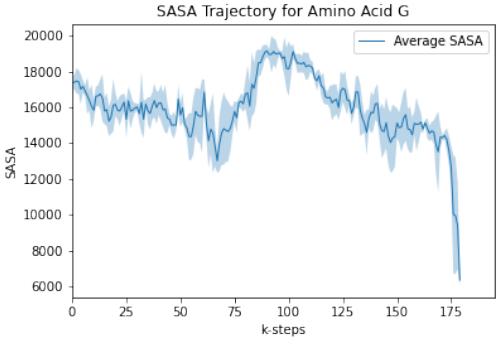
(d) ColabFold RMSD



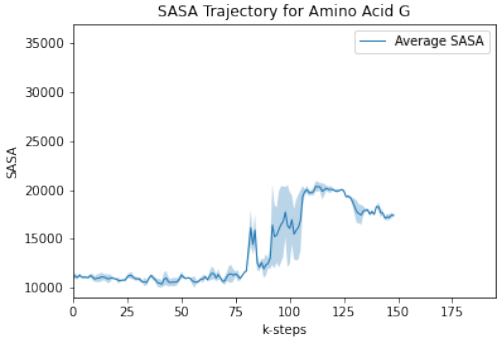
(e) ESMFold pLDDT



(f) ColabFold pLDDT

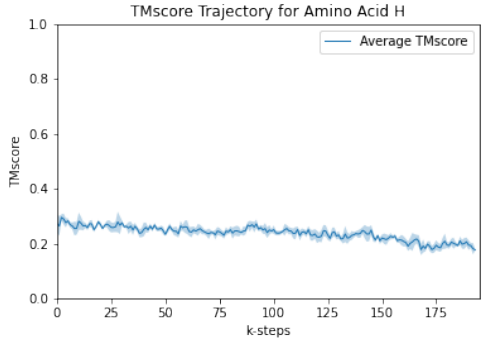


(g) ESMFold SASA

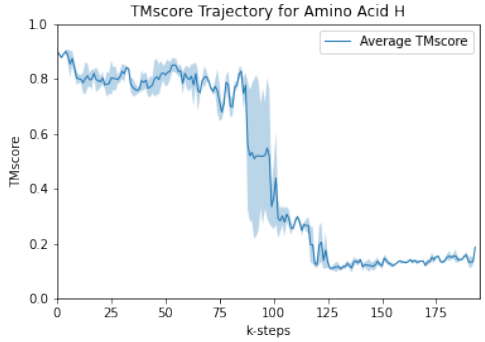


(h) ColabFold SASA

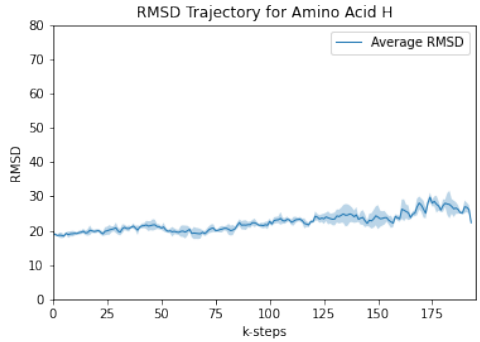
Figure A.6: Trajectory G. Comparison of the structural metrics along trajectory G from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



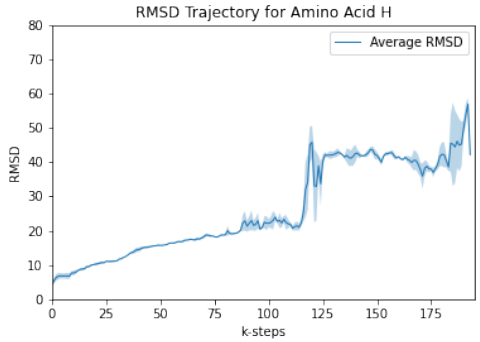
(a) ESMFold TM Score



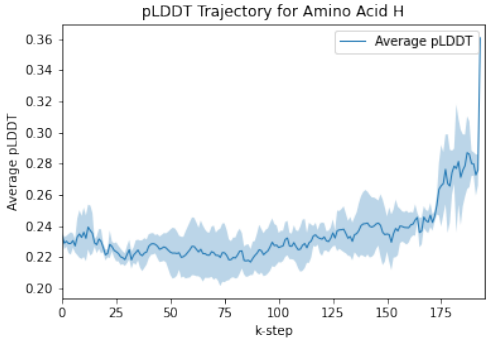
(b) ColabFold TM Score



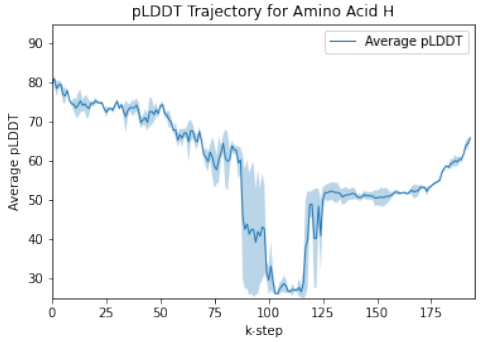
(c) ESMFold RMSD



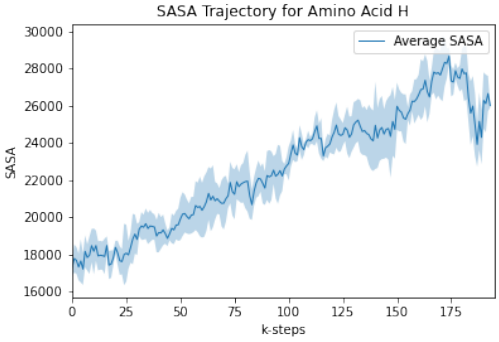
(d) ColabFold RMSD



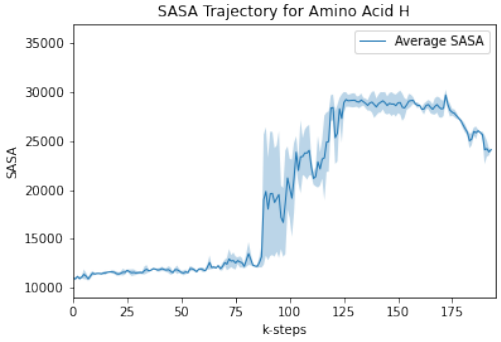
(e) ESMFold pLDDT



(f) ColabFold pLDDT

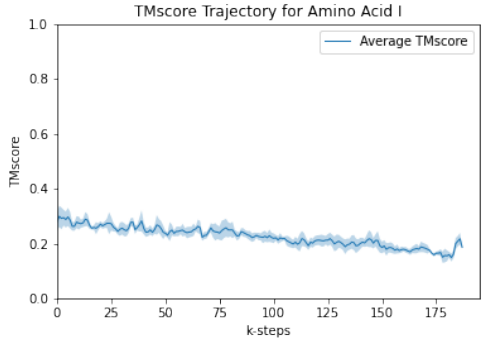


(g) ESMFold SASA

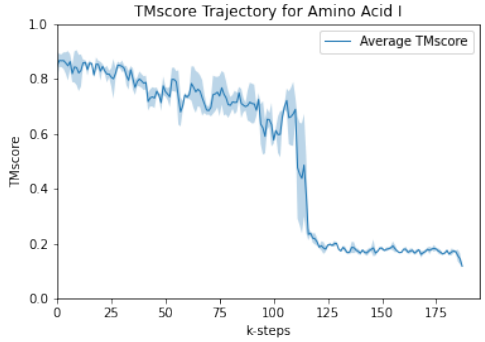


(h) ColabFold SASA

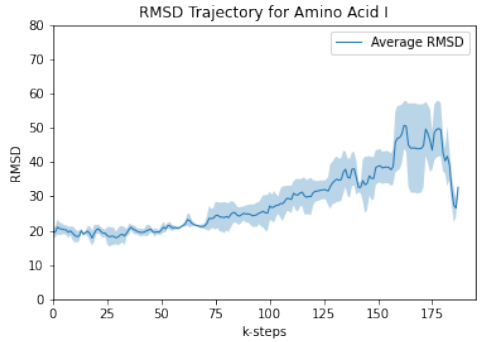
Figure A.7: Trajectory H. Comparison of the structural metrics along trajectory H from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



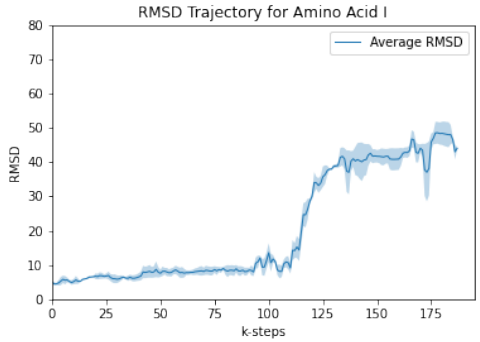
(a) ESMFold TM Score



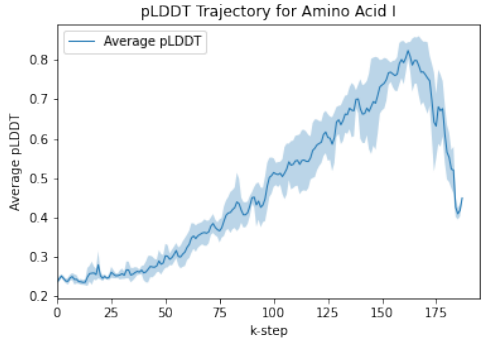
(b) ColabFold TM Score



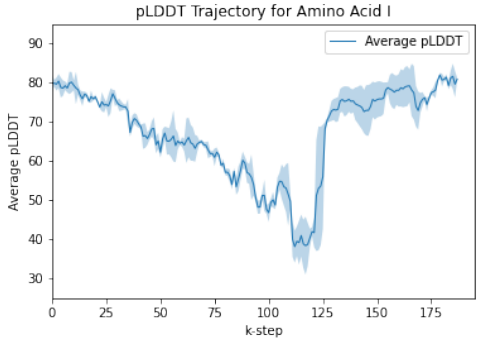
(c) ESMFold RMSD



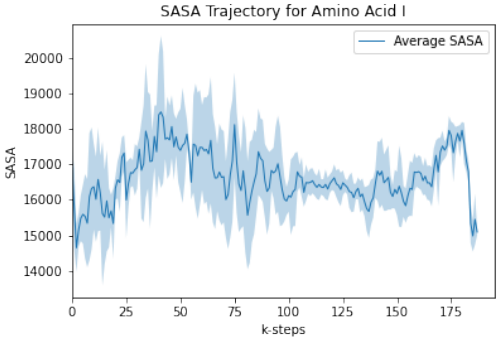
(d) ColabFold RMSD



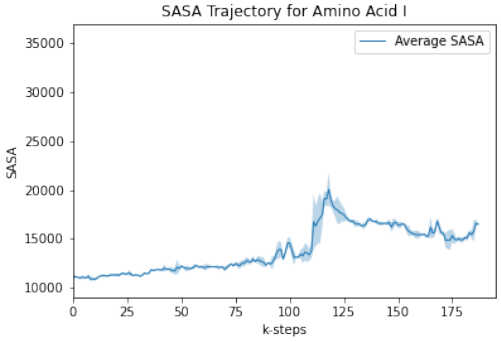
(e) ESMFold pLDDT



(f) ColabFold pLDDT

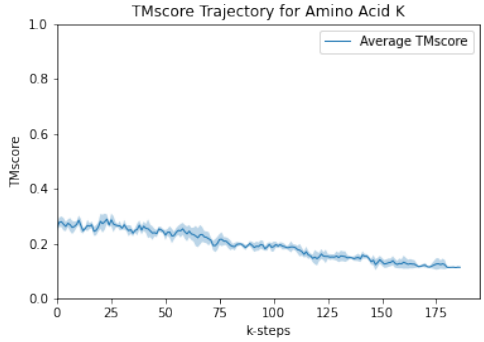


(g) ESMFold SASA

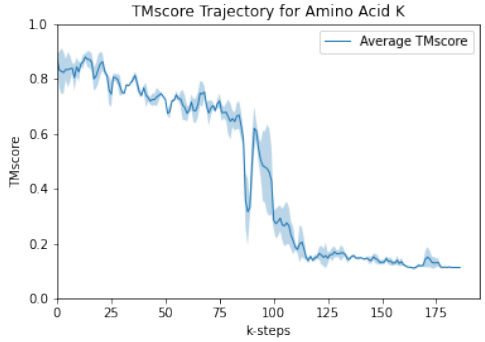


(h) ColabFold SASA

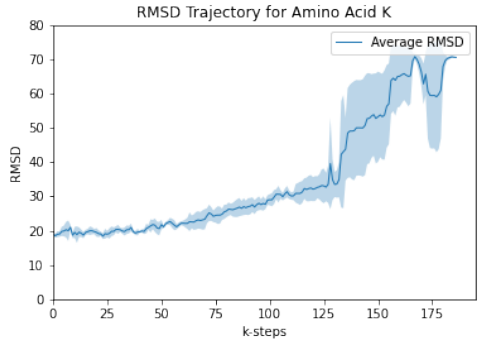
Figure A.8: Trajectory I. Comparison of the structural metrics along trajectory I from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



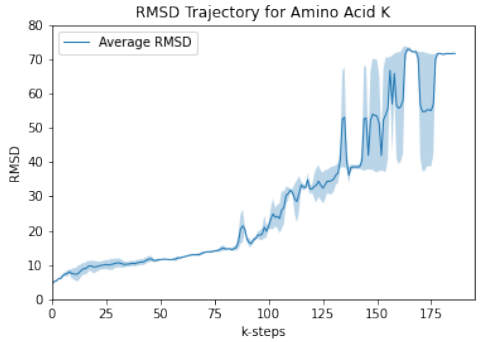
(a) ESMFold TM Score



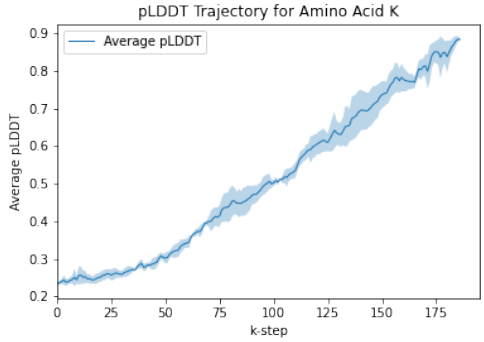
(b) ColabFold TM Score



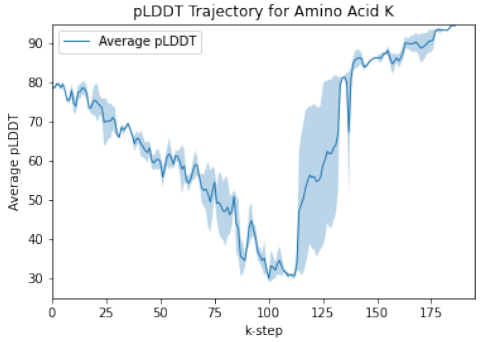
(c) ESMFold RMSD



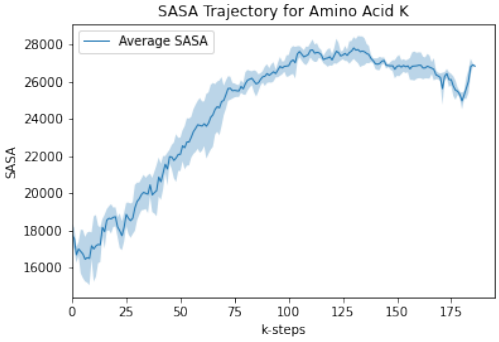
(d) ColabFold RMSD



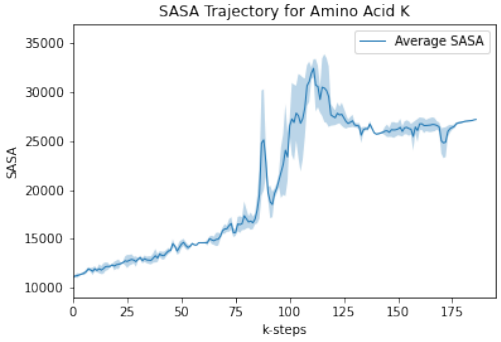
(e) ESMFold pLDDT



(f) ColabFold pLDDT

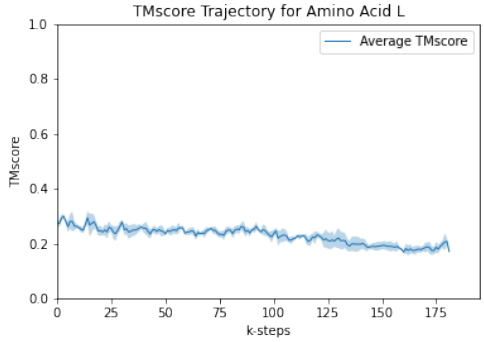


(g) ESMFold SASA

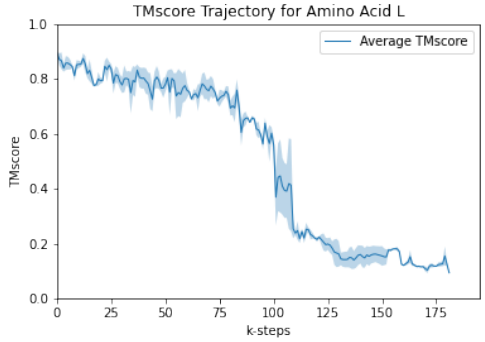


(h) ColabFold SASA

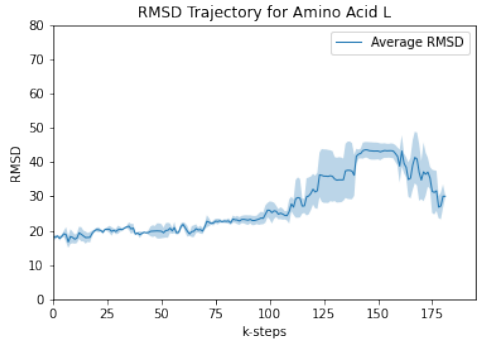
Figure A.9: Trajectory K. Comparison of the structural metrics along trajectory K from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



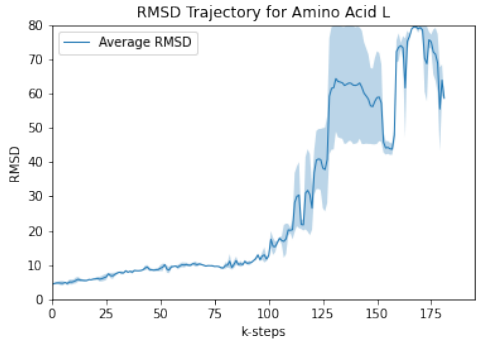
(a) ESMFold TM Score



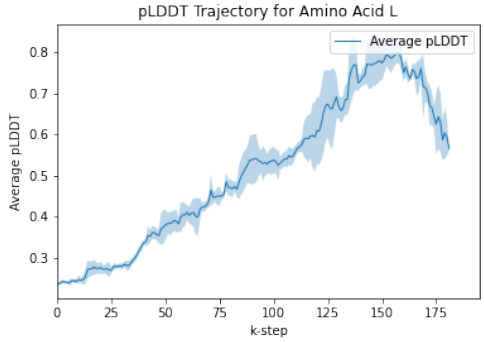
(b) ColabFold TM Score



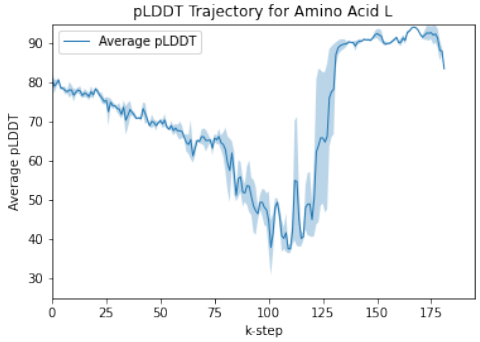
(c) ESMFold RMSD



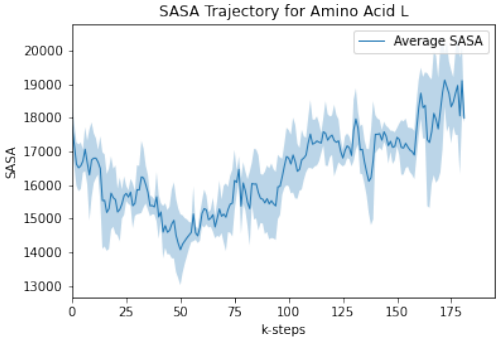
(d) ColabFold RMSD



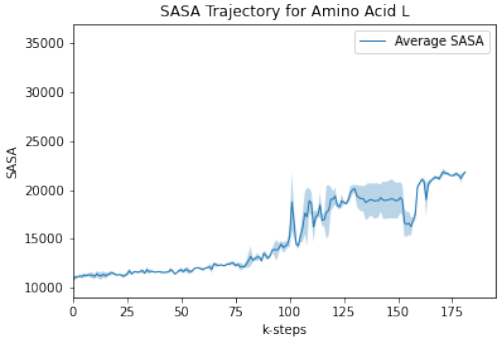
(e) ESMFold pLDDT



(f) ColabFold pLDDT

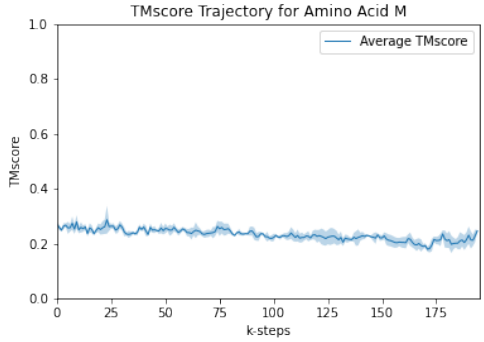


(g) ESMFold SASA

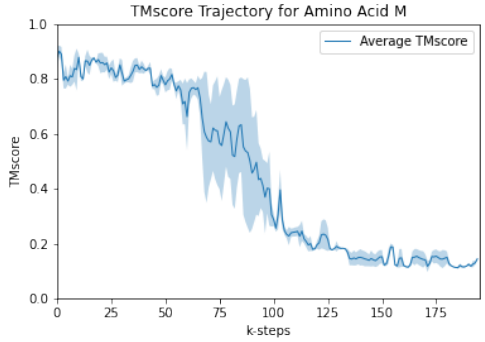


(h) ColabFold SASA

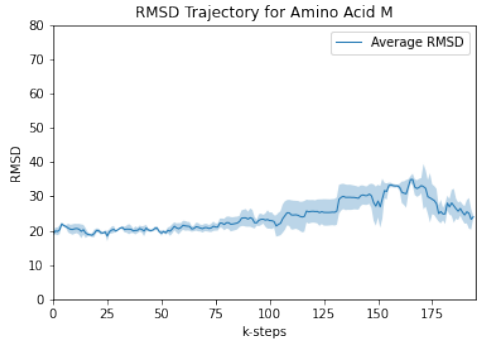
Figure A.10: Trajectory L. Comparison of the structural metrics along trajectory L from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



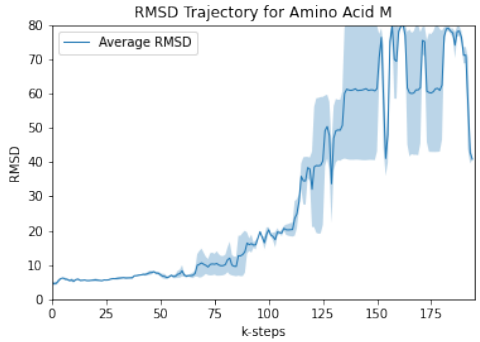
(a) ESMFold TM Score



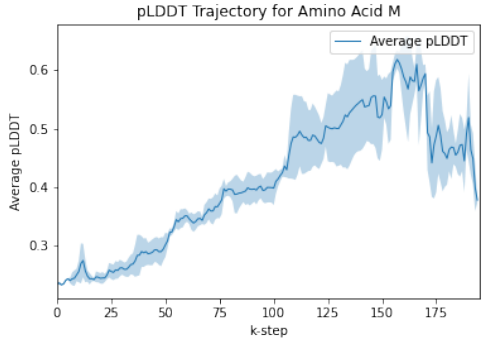
(b) ColabFold TM Score



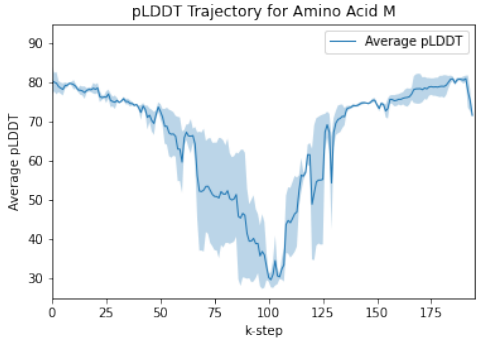
(c) ESMFold RMSD



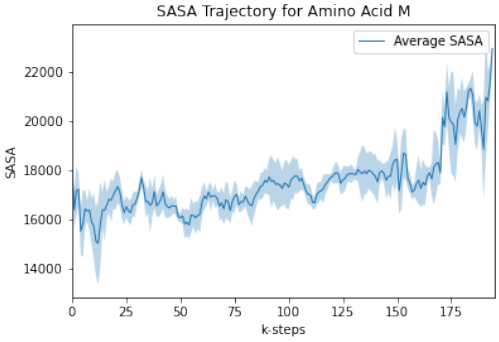
(d) ColabFold RMSD



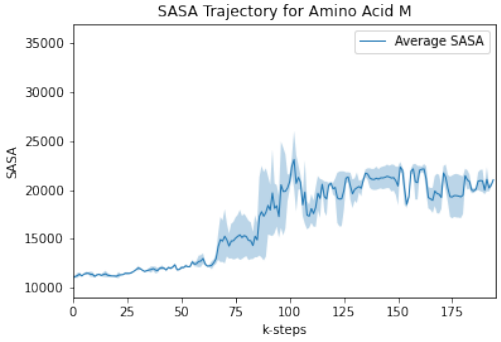
(e) ESMFold pLDDT



(f) ColabFold pLDDT

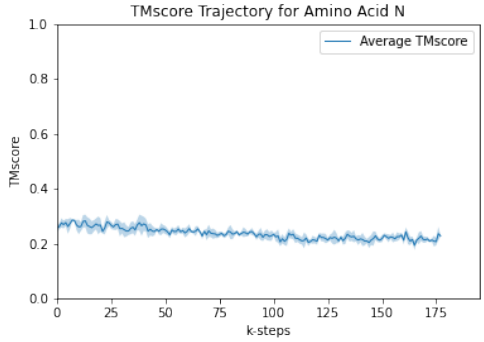


(g) ESMFold SASA

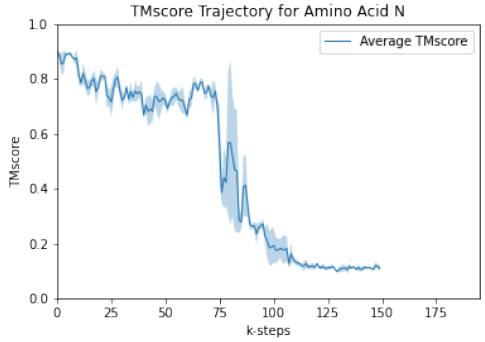


(h) ColabFold SASA

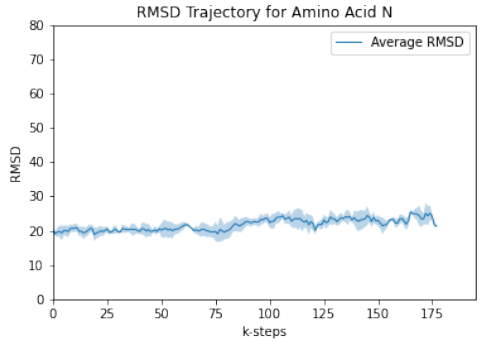
Figure A.11: Trajectory M. Comparison of the structural metrics along trajectory M from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



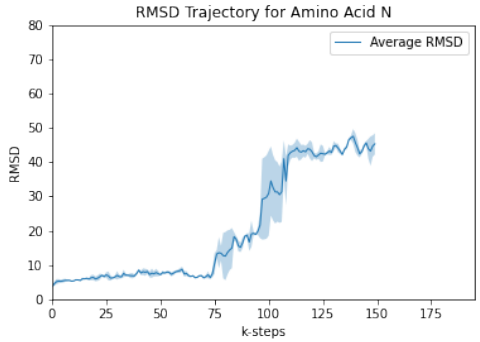
(a) ESMFold TM Score



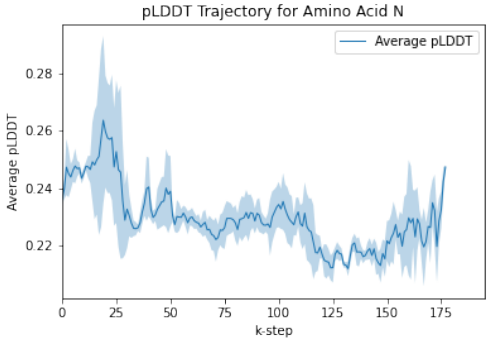
(b) ColabFold TM Score



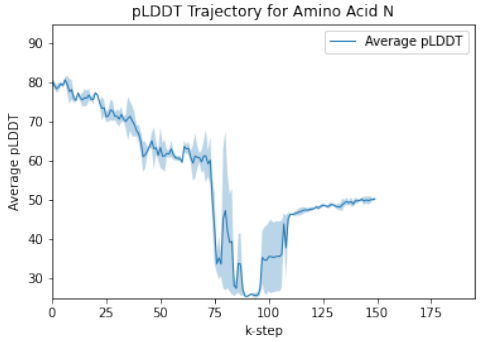
(c) ESMFold RMSD



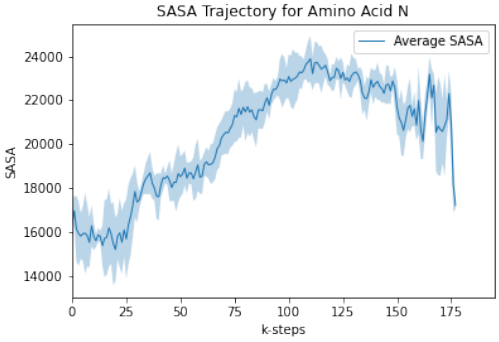
(d) ColabFold RMSD



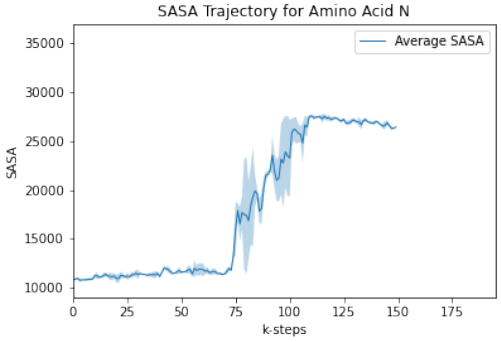
(e) ESMFold pLDDT



(f) ColabFold pLDDT

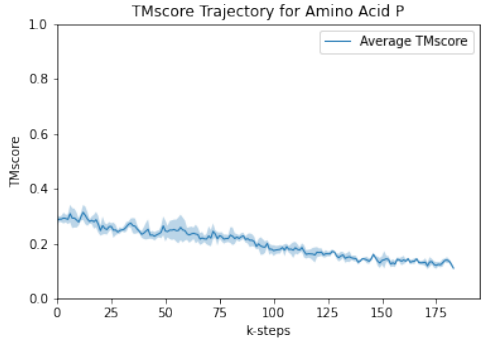


(g) ESMFold SASA

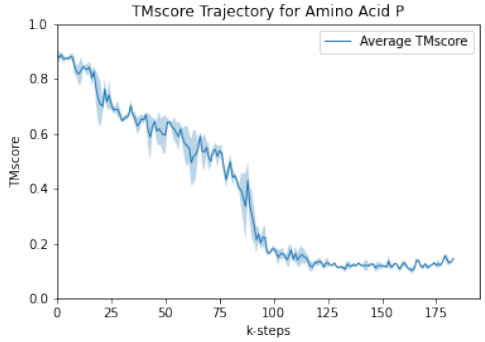


(h) ColabFold SASA

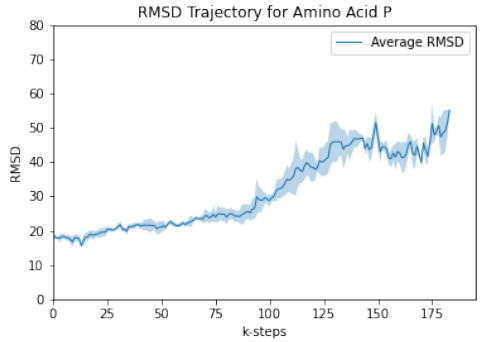
Figure A.12: Trajectory N. Comparison of the structural metrics along trajectory N from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



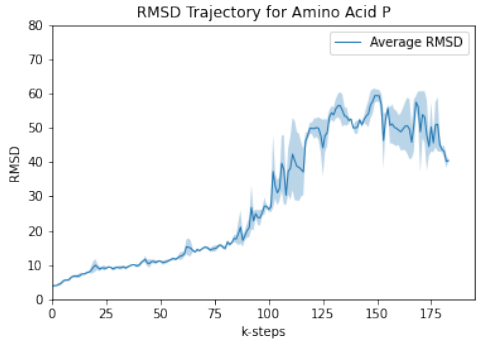
(a) ESMFold TM Score



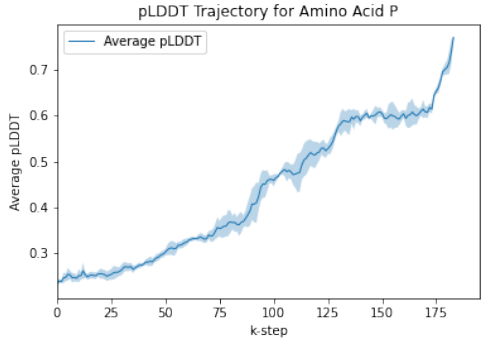
(b) ColabFold TM Score



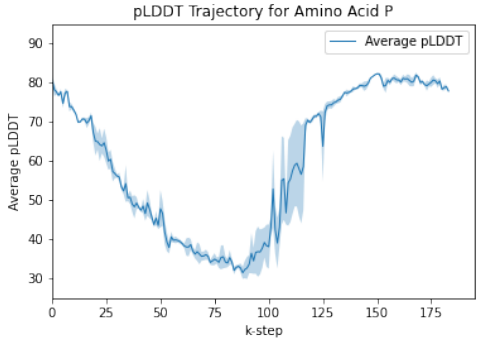
(c) ESMFold RMSD



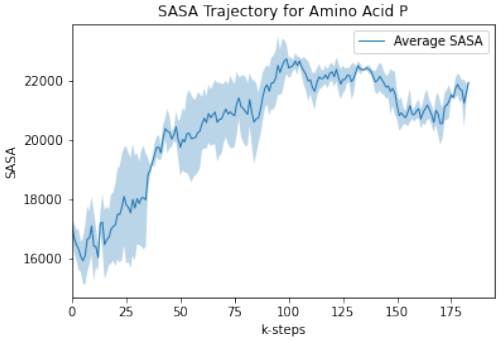
(d) ColabFold RMSD



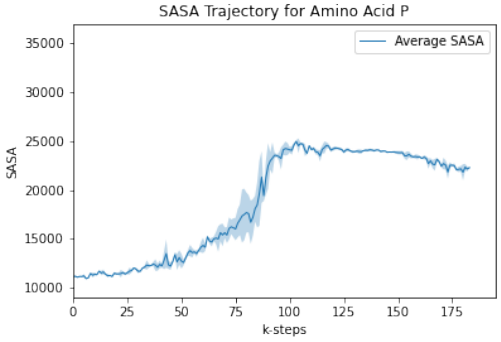
(e) ESMFold pLDDT



(f) ColabFold pLDDT

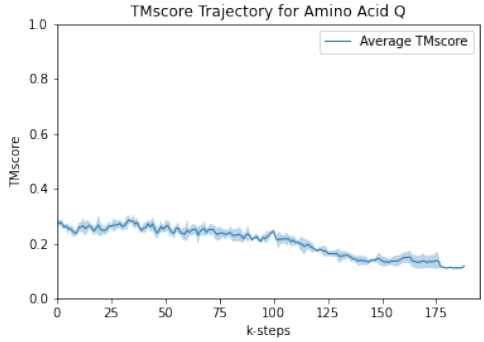


(g) ESMFold SASA

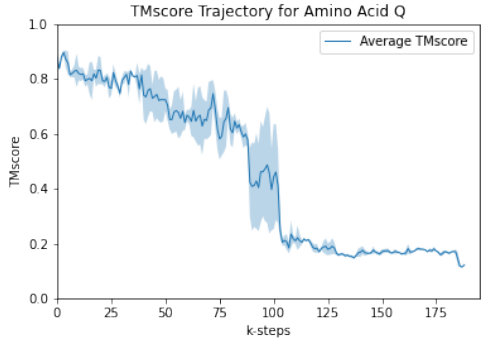


(h) ColabFold SASA

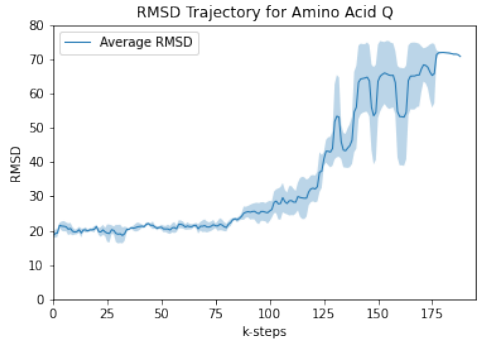
Figure A.13: Trajectory P. Comparison of the structural metrics along trajectory P from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



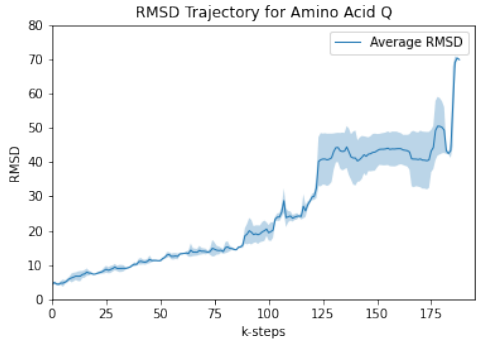
(a) ESMFold TM Score



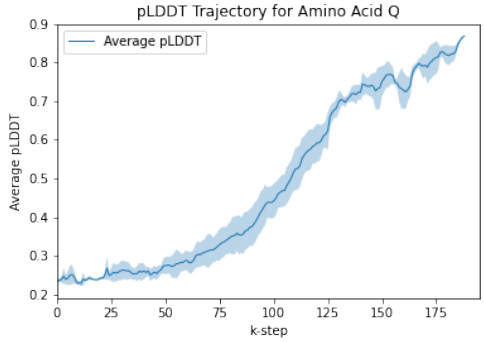
(b) ColabFold TM Score



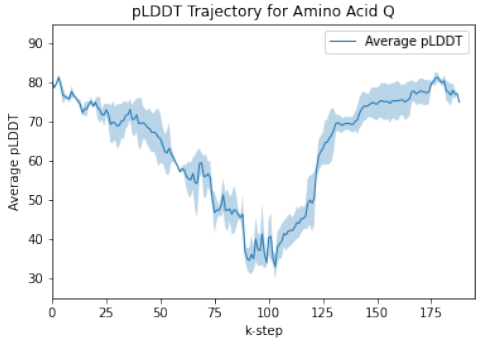
(c) ESMFold RMSD



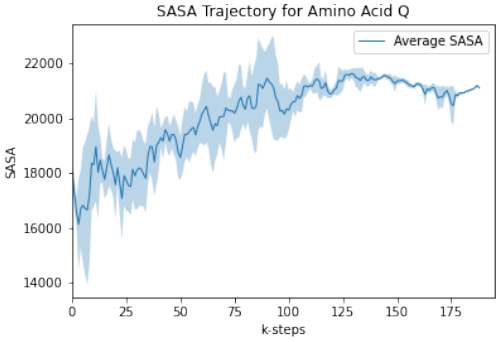
(d) ColabFold RMSD



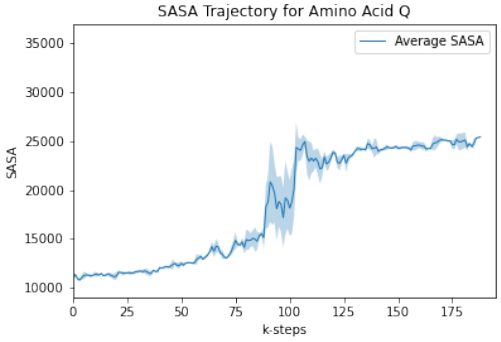
(e) ESMFold pLDDT



(f) ColabFold pLDDT

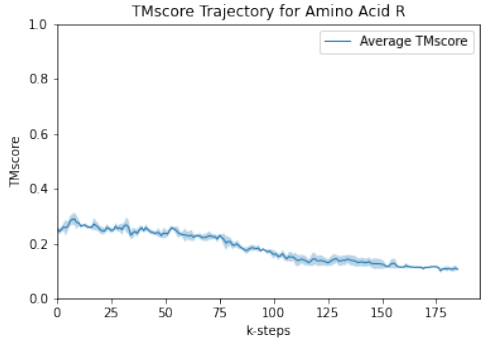


(g) ESMFold SASA

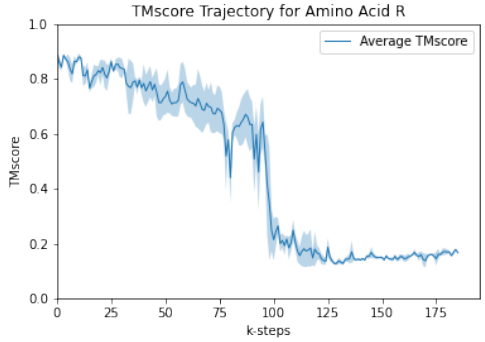


(h) ColabFold SASA

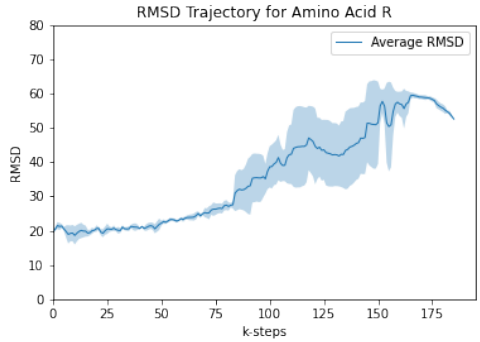
Figure A.14: Trajectory Q. Comparison of the structural metrics along trajectory Q from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



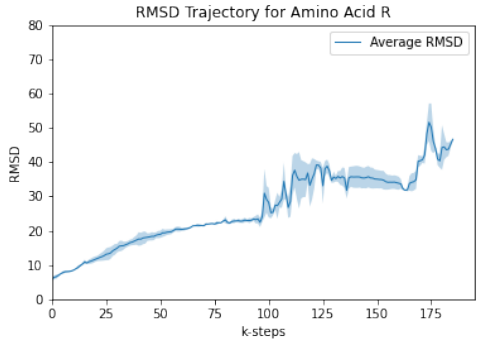
(a) ESMFold TM Score



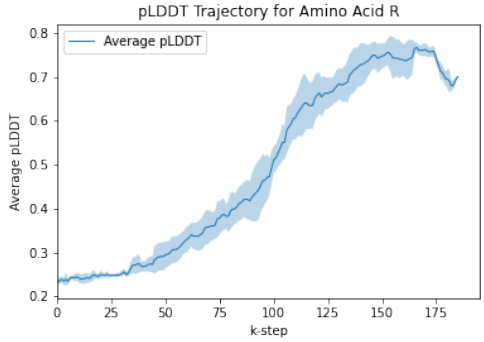
(b) ColabFold TM Score



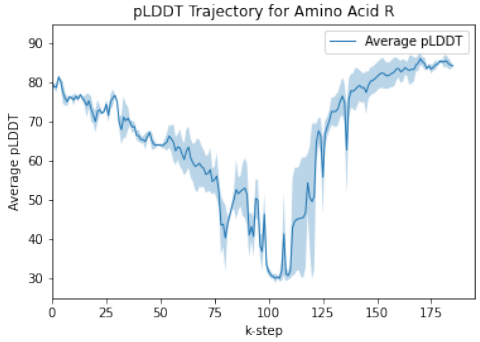
(c) ESMFold RMSD



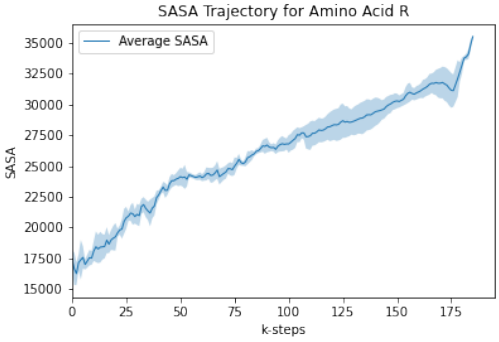
(d) ColabFold RMSD



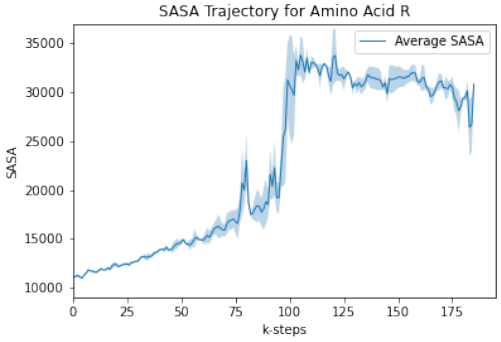
(e) ESMFold pLDDT



(f) ColabFold pLDDT

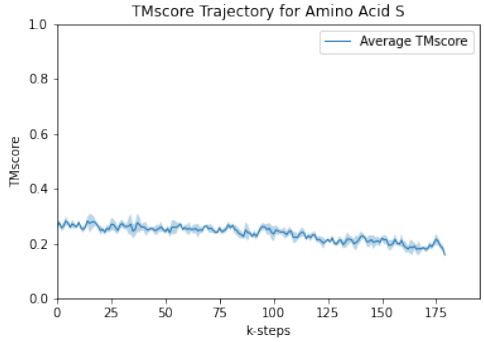


(g) ESMFold SASA

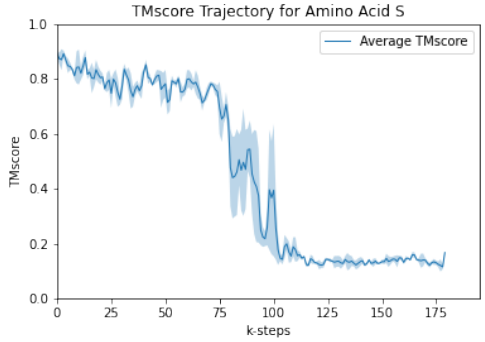


(h) ColabFold SASA

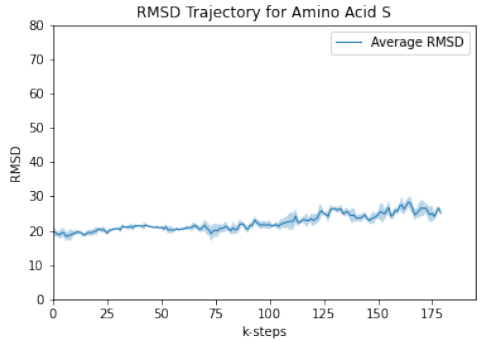
Figure A.15: Trajectory R. Comparison of the structural metrics along trajectory R from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



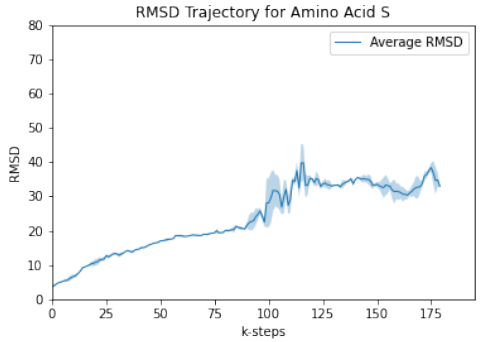
(a) ESMFold TM Score



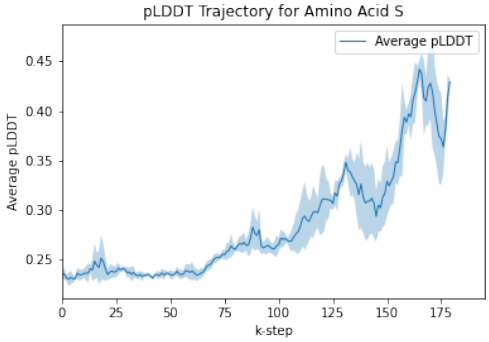
(b) ColabFold TM Score



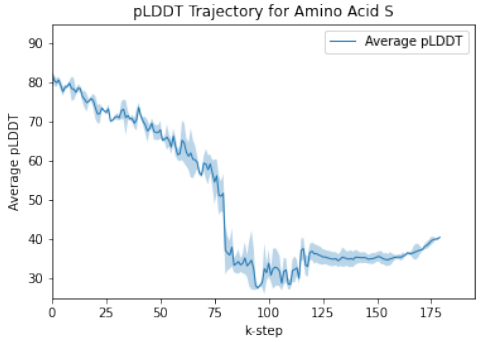
(c) ESMFold RMSD



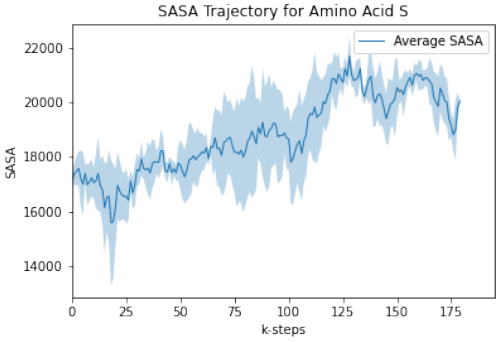
(d) ColabFold RMSD



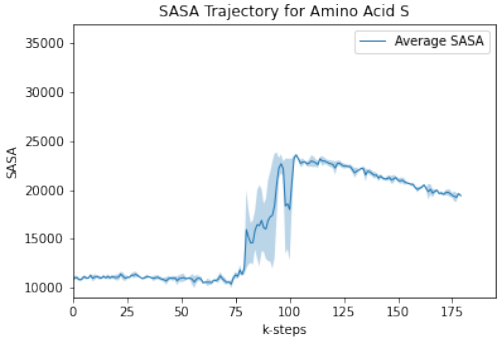
(e) ESMFold pLDDT



(f) ColabFold pLDDT

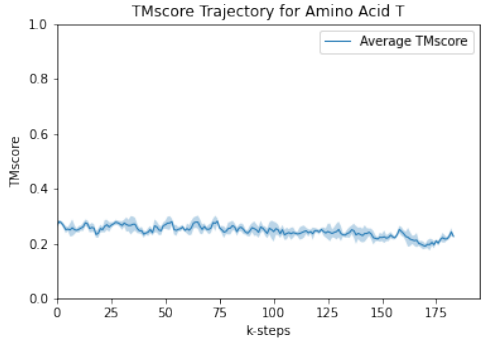


(g) ESMFold SASA

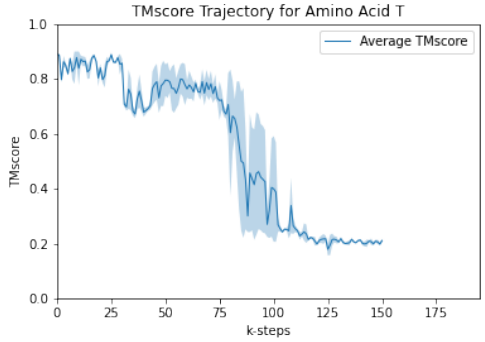


(h) ColabFold SASA

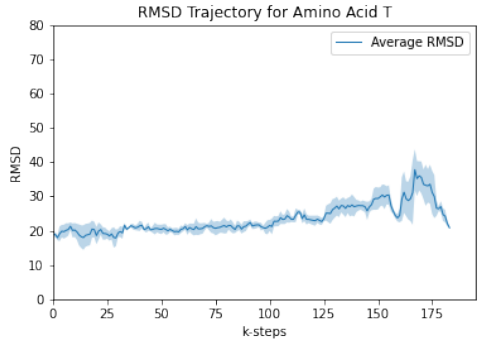
Figure A.16: Trajectory S. Comparison of the structural metrics along trajectory S from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



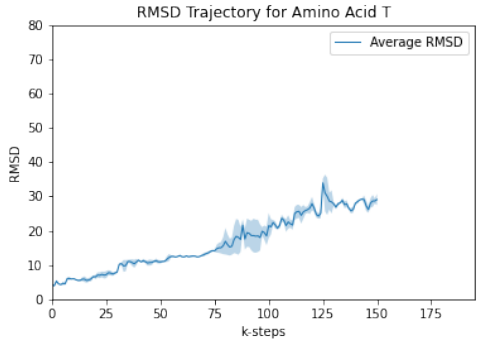
(a) ESMFold TM Score



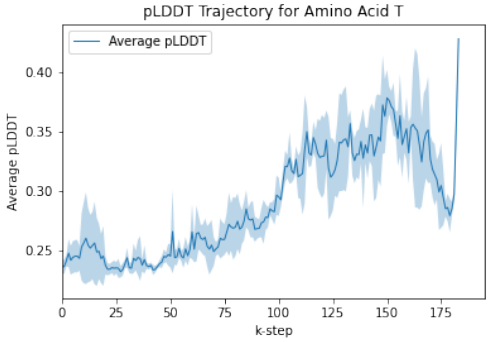
(b) ColabFold TM Score



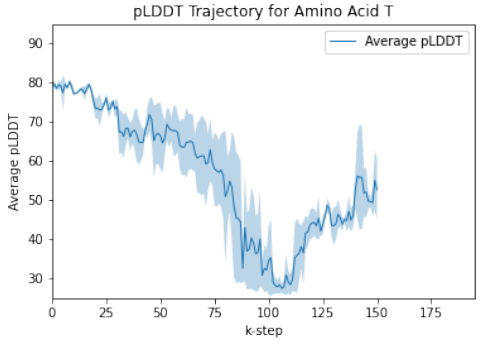
(c) ESMFold RMSD



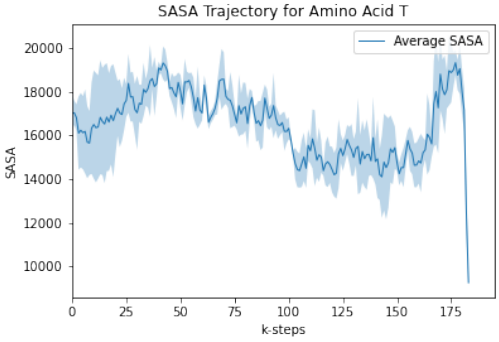
(d) ColabFold RMSD



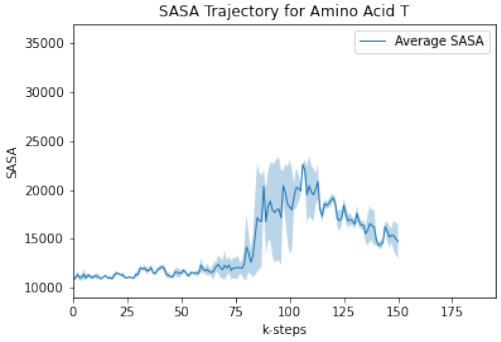
(e) ESMFold pLDDT



(f) ColabFold pLDDT

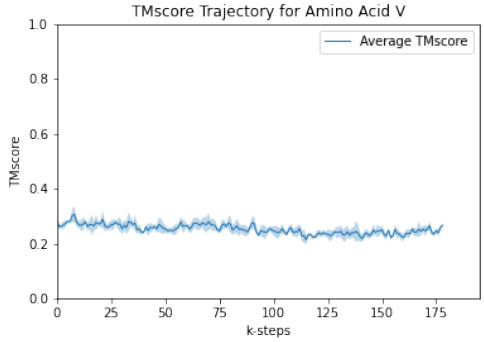


(g) ESMFold SASA

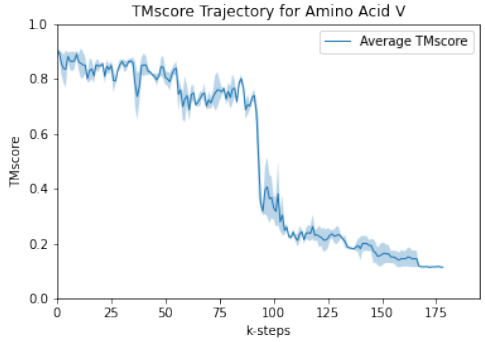


(h) ColabFold SASA

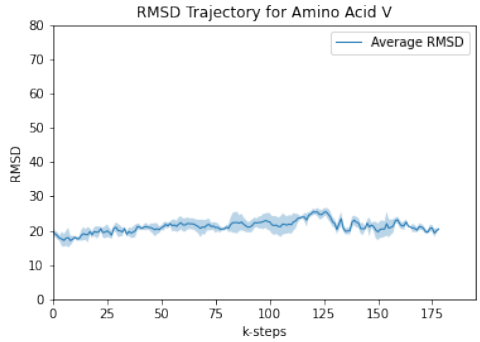
Figure A.17: Trajectory T. Comparison of the structural metrics along trajectory T from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



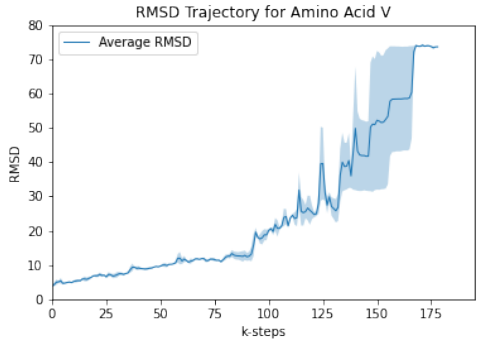
(a) ESMFold TM Score



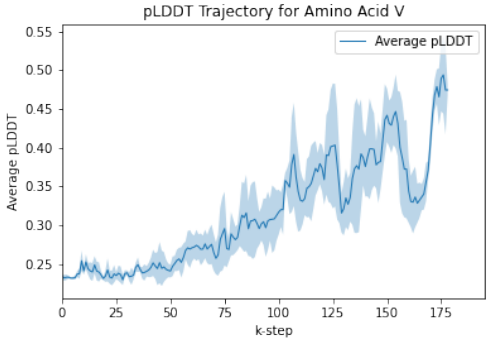
(b) ColabFold TM Score



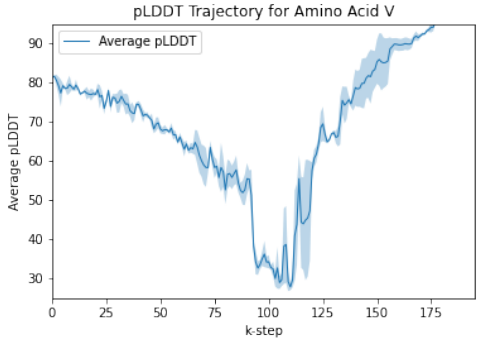
(c) ESMFold RMSD



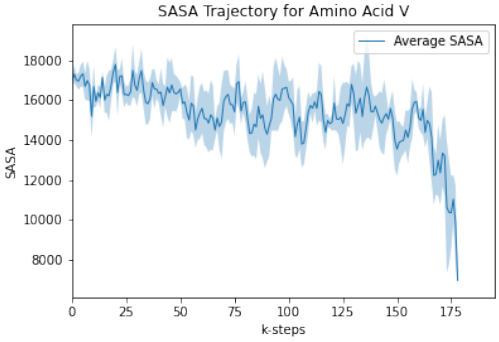
(d) ColabFold RMSD



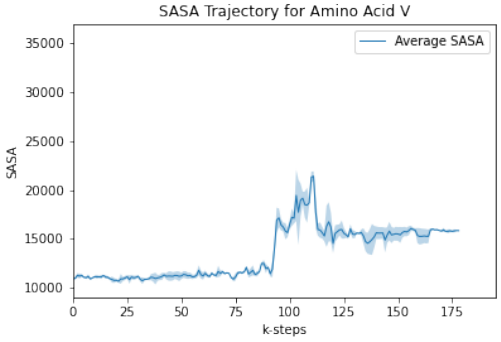
(e) ESMFold pLDDT



(f) ColabFold pLDDT

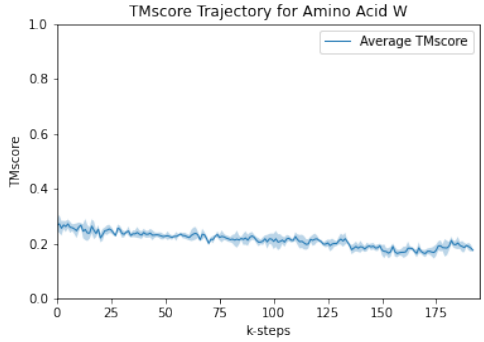


(g) ESMFold SASA

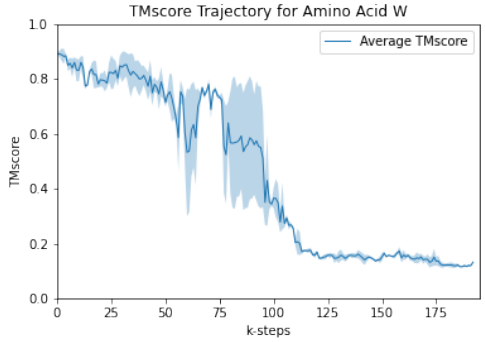


(h) ColabFold SASA

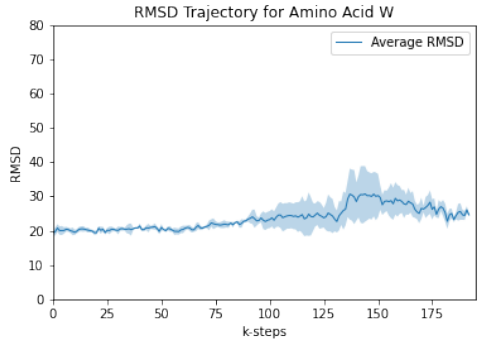
Figure A.18: Trajectory V. Comparison of the structural metrics along trajectory V from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



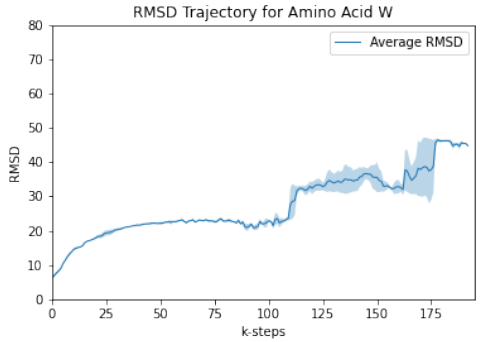
(a) ESMFold TM Score



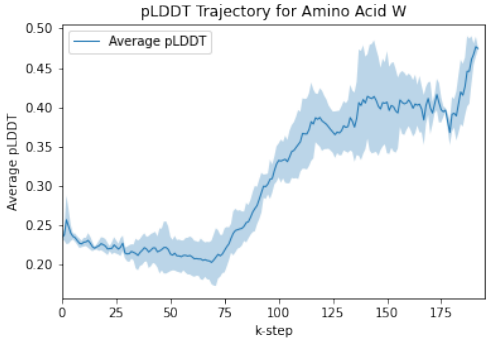
(b) ColabFold TM Score



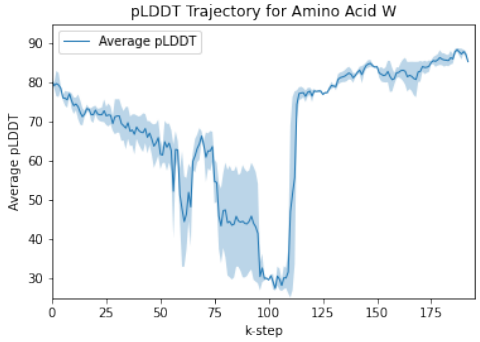
(c) ESMFold RMSD



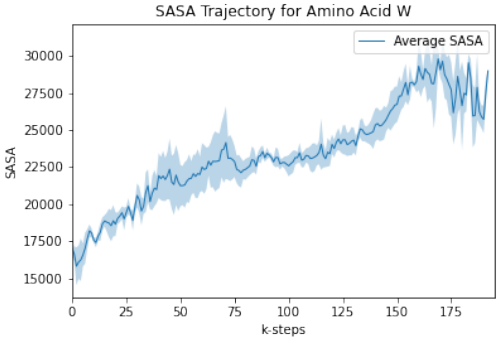
(d) ColabFold RMSD



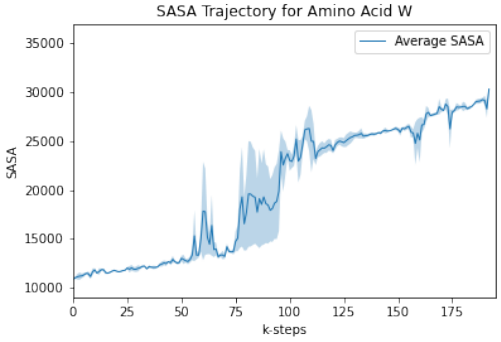
(e) ESMFold pLDDT



(f) ColabFold pLDDT

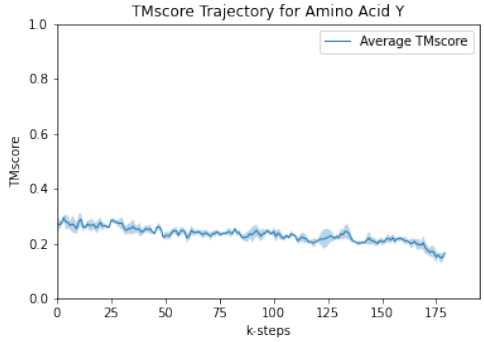


(g) ESMFold SASA

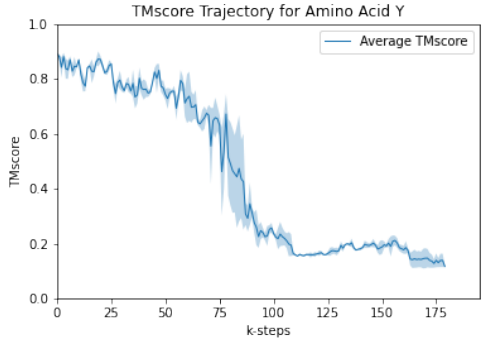


(h) ColabFold SASA

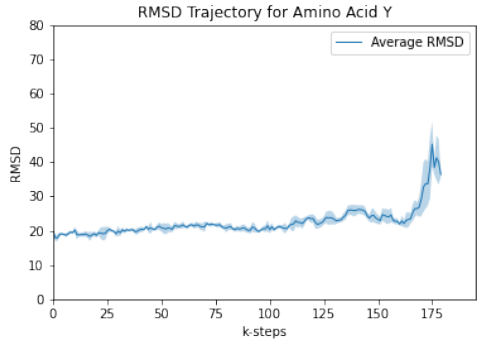
Figure A.19: Trajectory W. Comparison of the structural metrics along trajectory W from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.



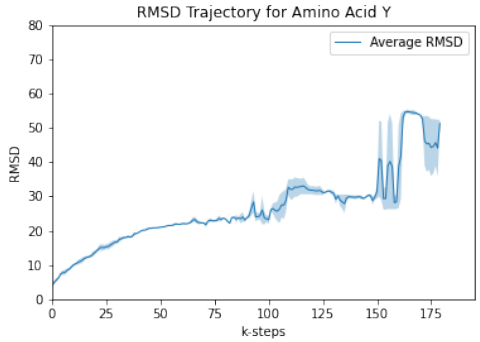
(a) ESMFold TM Score



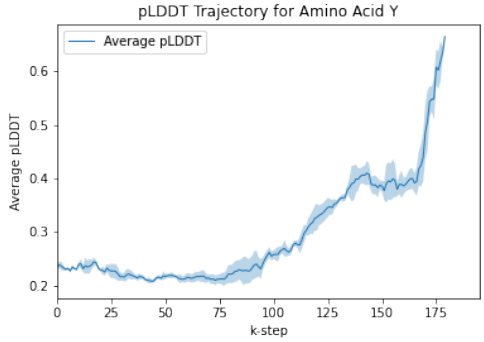
(b) ColabFold TM Score



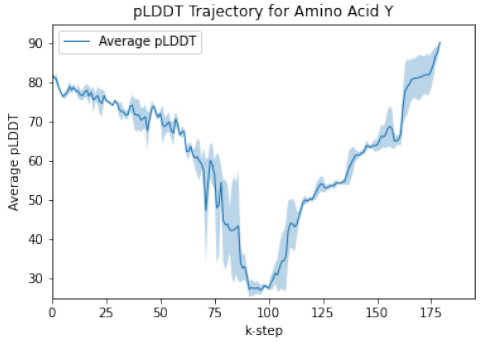
(c) ESMFold RMSD



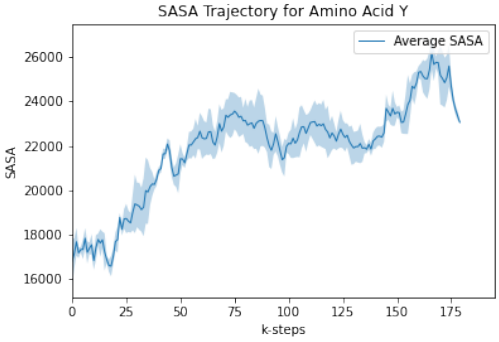
(d) ColabFold RMSD



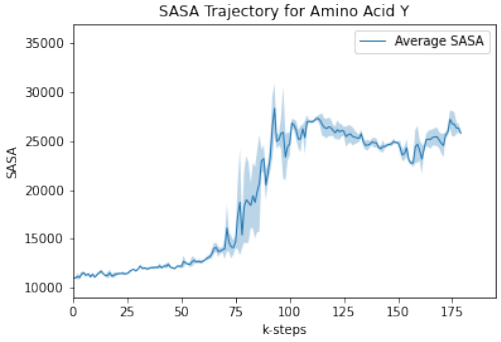
(e) ESMFold pLDDT



(f) ColabFold pLDDT



(g) ESMFold SASA

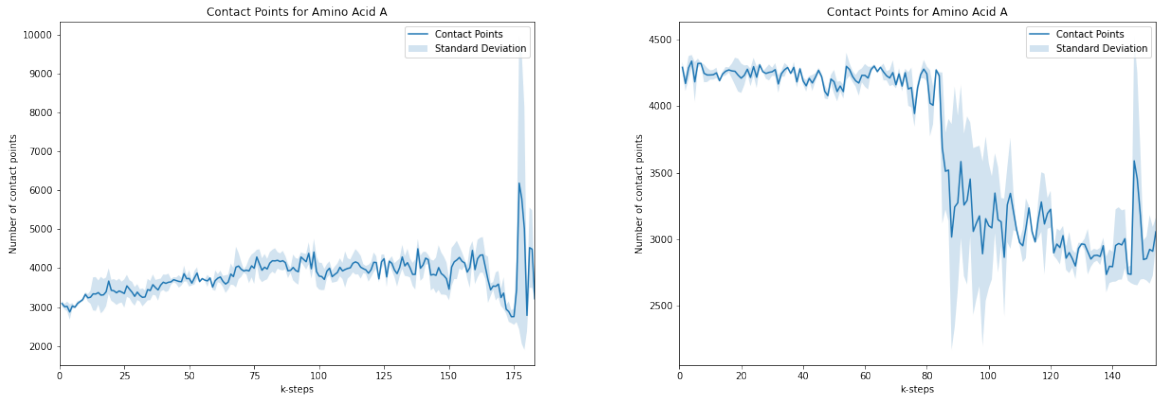


(h) ColabFold SASA

Figure A.20: Trajectory Y. Comparison of the structural metrics along trajectory Y from ESMFold and ColabFold. The left column shows the results from ESMFold, and the right column shows the results from ColabFold.

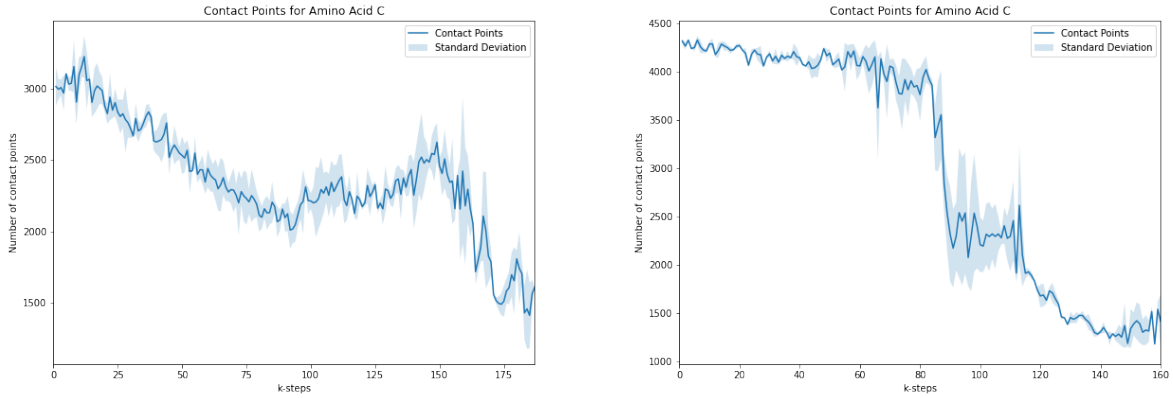
Appendix B

Contact Matrix Plots



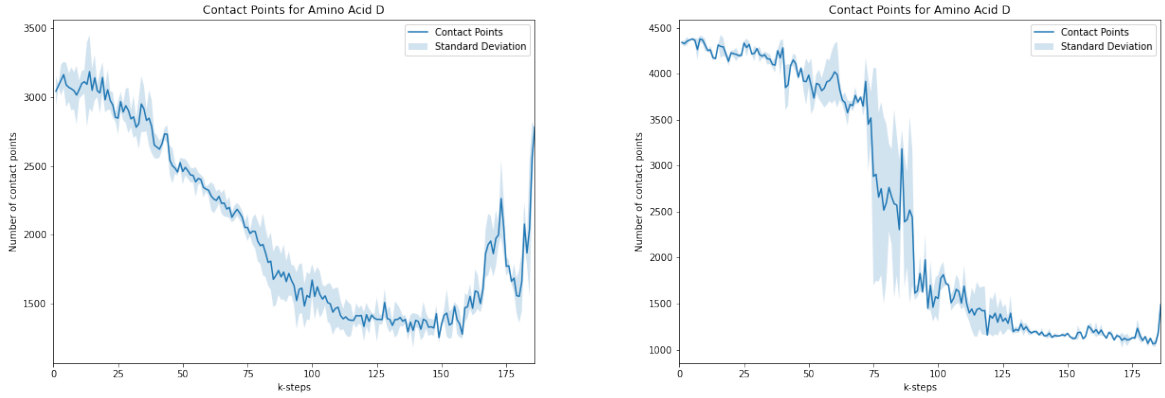
(a) ESMFold **(b)** ColabFold

Figure B.1: Trajectory A. Contact Matrix along the trajectory towards the homopolymer of amino acid A.



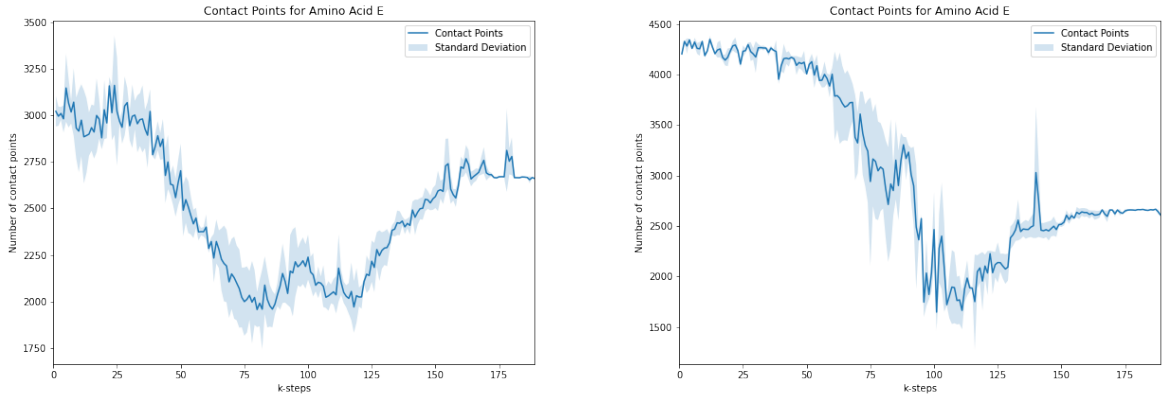
(a) ESMFold **(b)** ColabFold

Figure B.2: Trajectory C. Contact Matrix along the trajectory towards the homopolymer of amino acid C.



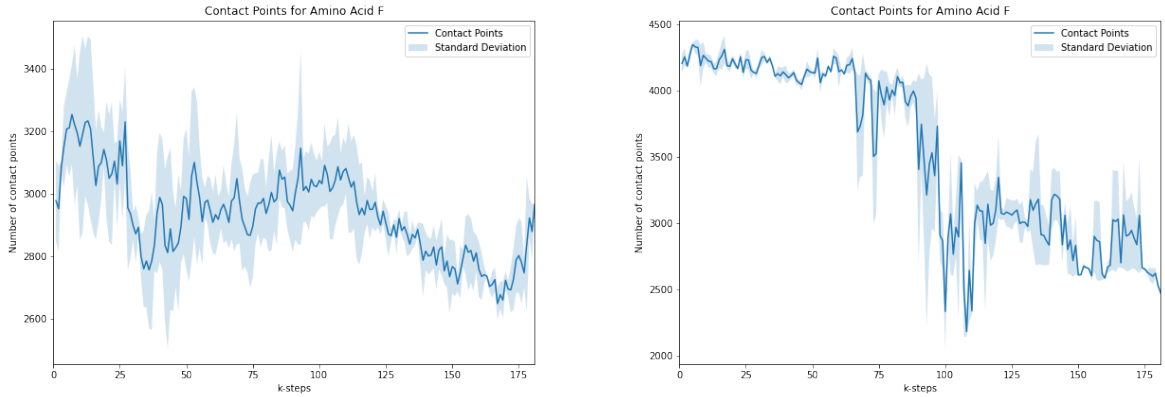
(a) ESMFold **(b)** ColabFold

Figure B.3: Trajectory D. Contact Matrix along the trajectory towards the homopolymer of amino acid D.



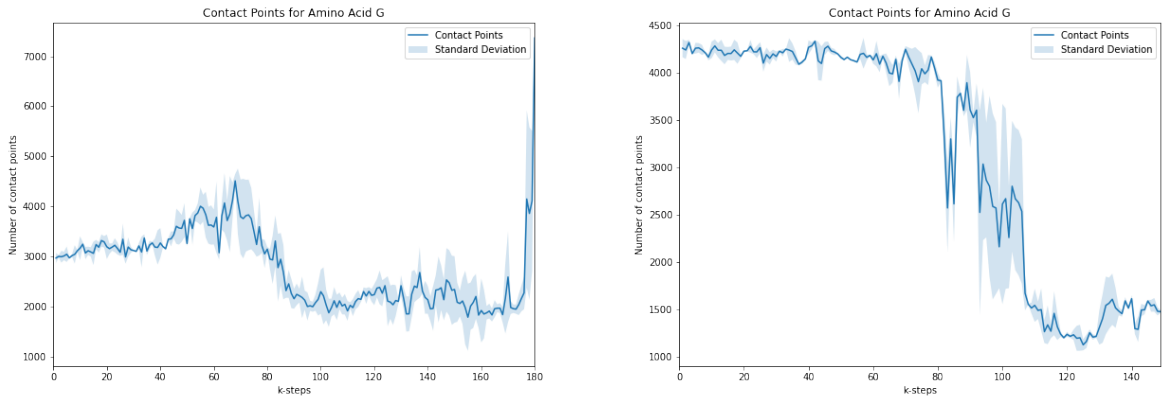
(a) ESMFold **(b)** ColabFold

Figure B.4: Trajectory E. Contact Matrix along the trajectory towards the homopolymer of amino acid E.



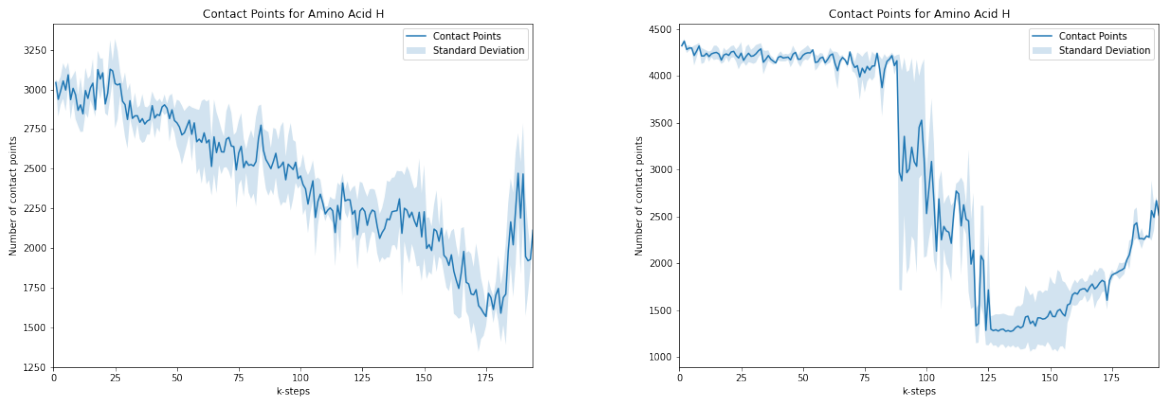
(a) ESMFold **(b)** ColabFold

Figure B.5: Trajectory F. Contact Matrix along the trajectory towards the homopolymer of amino acid F.



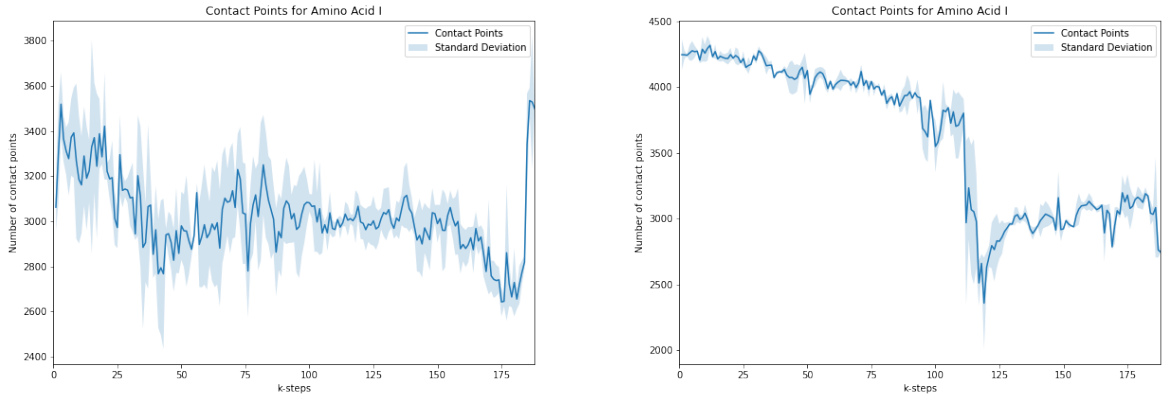
(a) ESMFold **(b)** ColabFold

Figure B.6: Trajectory G. Contact Matrix along the trajectory towards the homopolymer of amino acid G.



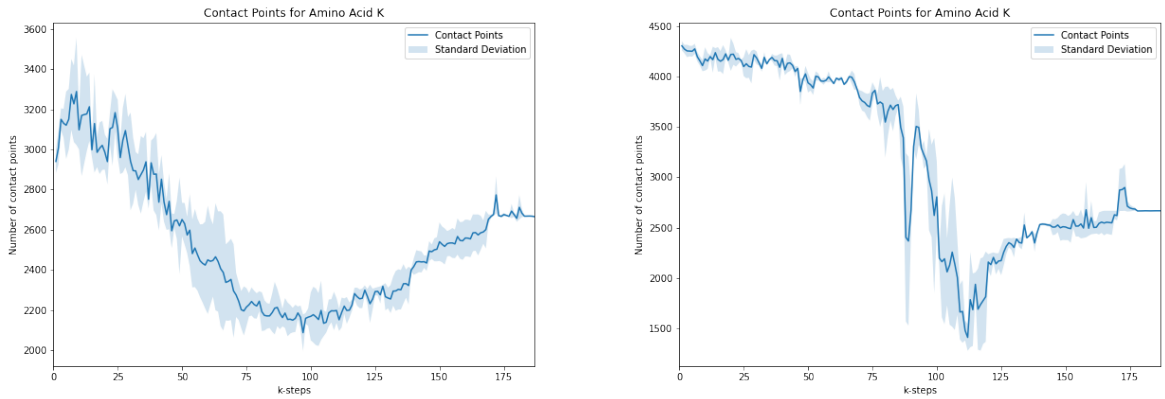
(a) ESMFold **(b)** ColabFold

Figure B.7: Trajectory H. Contact Matrix along the trajectory towards the homopolymer of amino acid H.



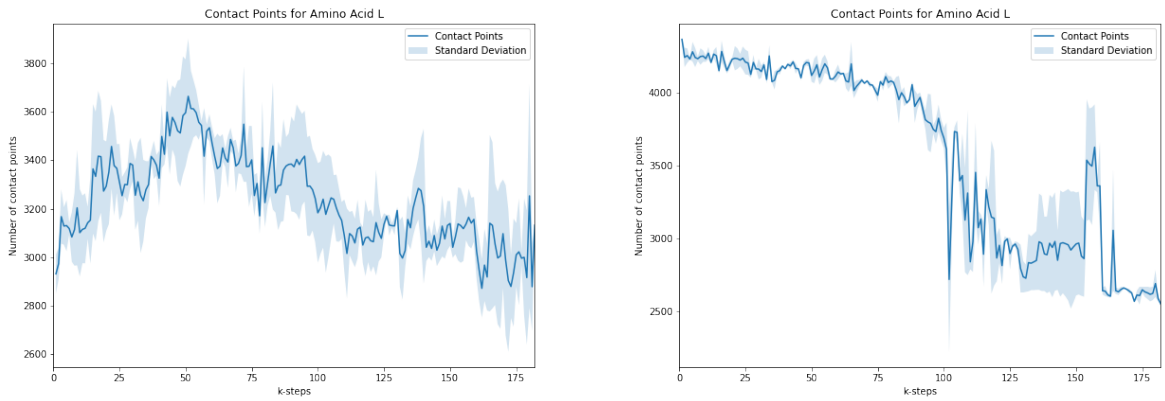
(a) ESMFold **(b)** ColabFold

Figure B.8: Trajectory I. Contact Matrix along the trajectory towards the homopolymer of amino acid I.



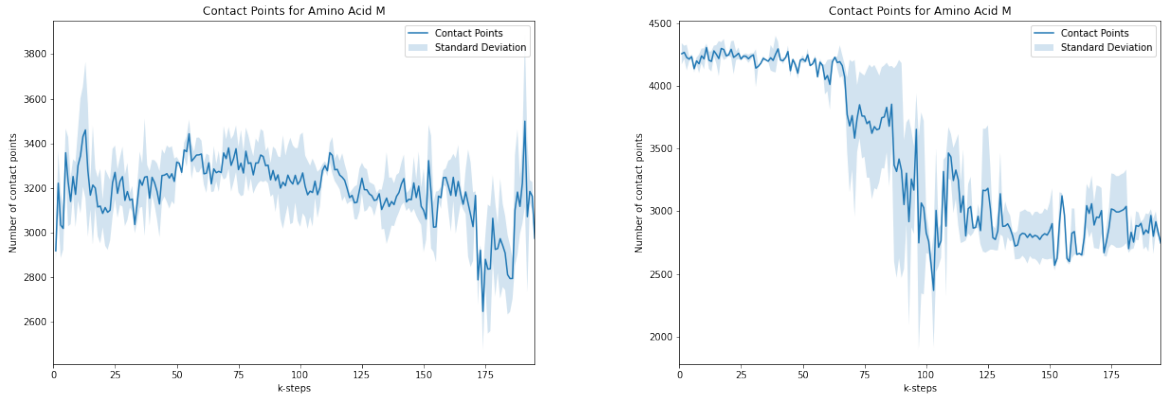
(a) ESMFold **(b)** ColabFold

Figure B.9: Trajectory K. Contact Matrix along the trajectory towards the homopolymer of amino acid K.



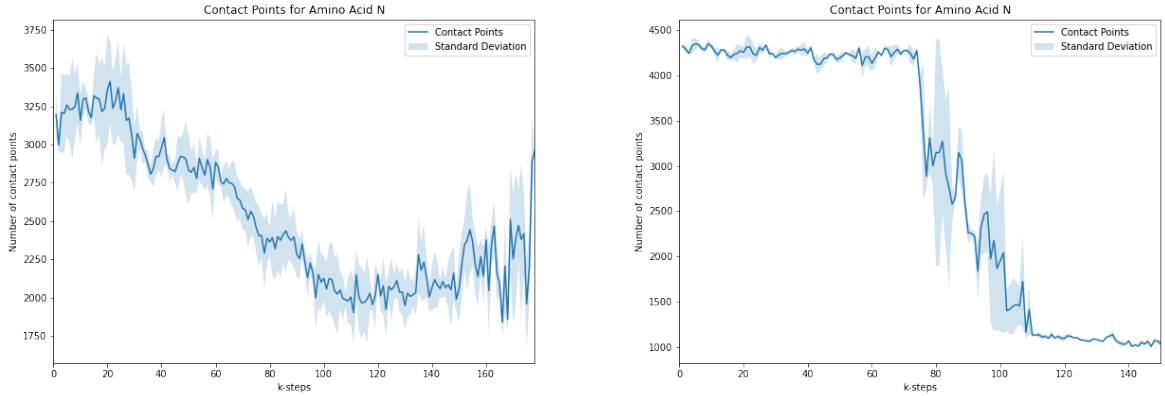
(a) ESMFold **(b)** ColabFold

Figure B.10: Trajectory L. Contact Matrix along the trajectory towards the homopolymer of amino acid L.



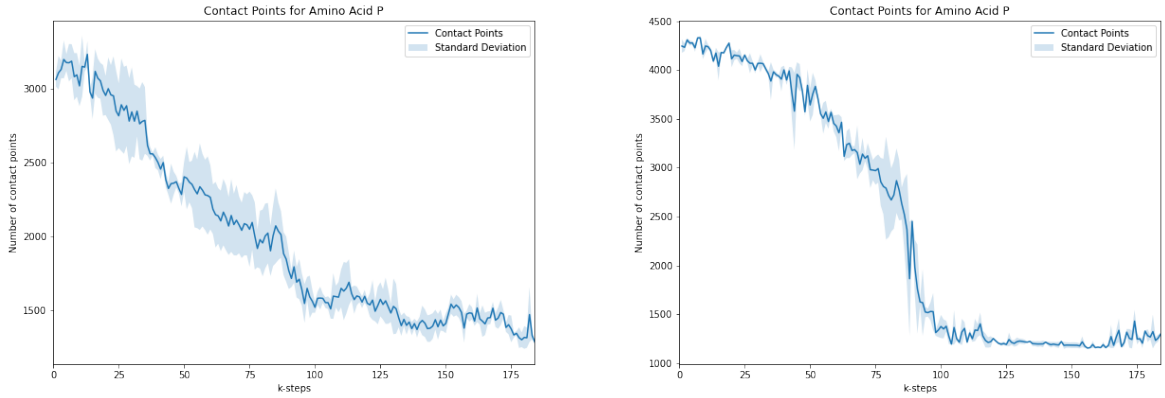
(a) ESMFold **(b)** ColabFold

Figure B.11: Trajectory M. Contact Matrix along the trajectory towards the homopolymer of amino acid M.



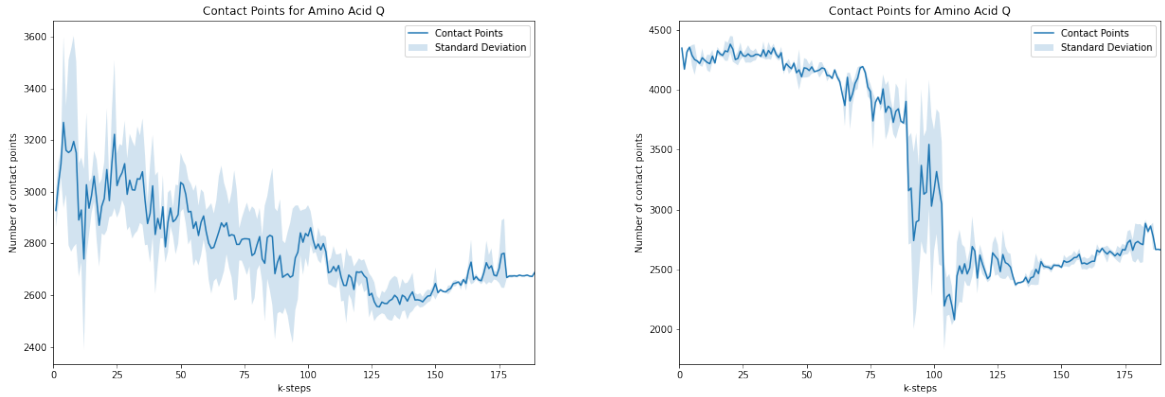
(a) ESMFold **(b)** ColabFold

Figure B.12: Trajectory N. Contact Matrix along the trajectory towards the homopolymer of amino acid N.



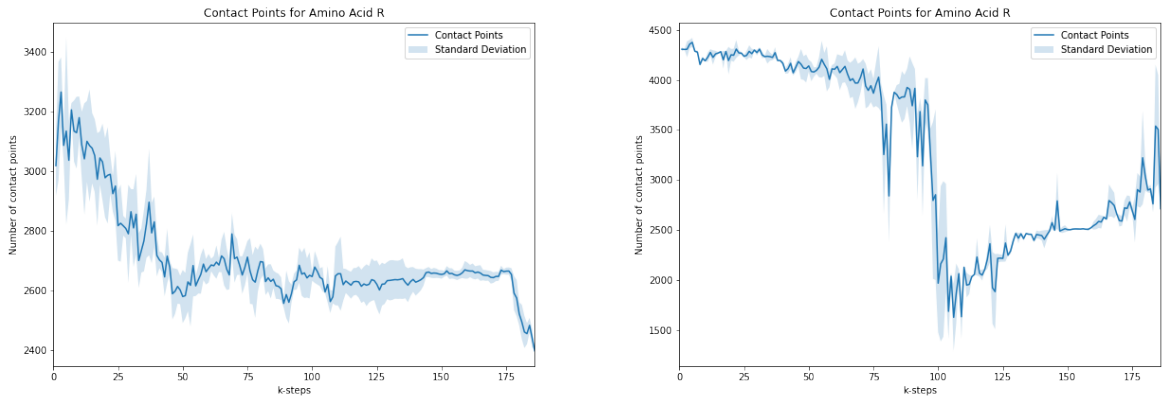
(a) ESMFold **(b)** ColabFold

Figure B.13: Trajectory P. Contact Matrix along the trajectory towards the homopolymer of amino acid P.



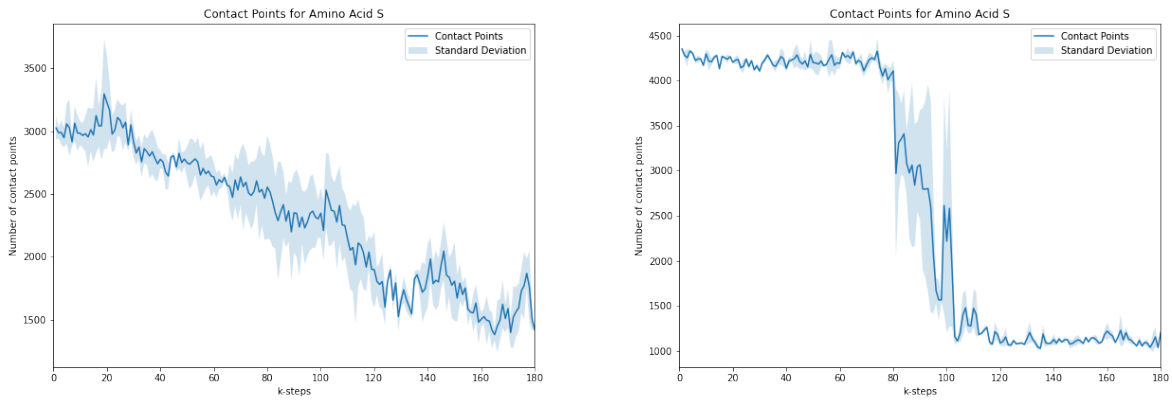
(a) ESMFold **(b)** ColabFold

Figure B.14: Trajectory Q. Contact Matrix along the trajectory towards the homopolymer of amino acid Q.



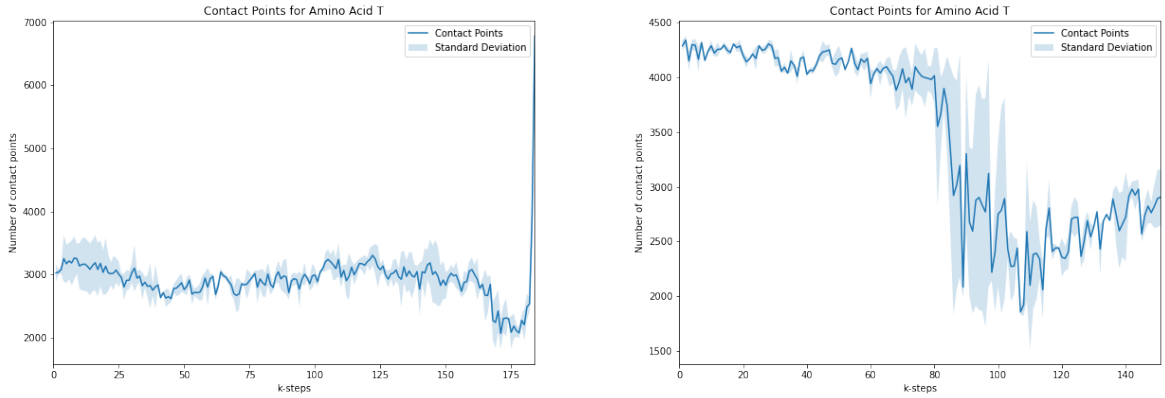
(a) ESMFold **(b)** ColabFold

Figure B.15: Trajectory R. Contact Matrix along the trajectory towards the homopolymer of amino acid R.



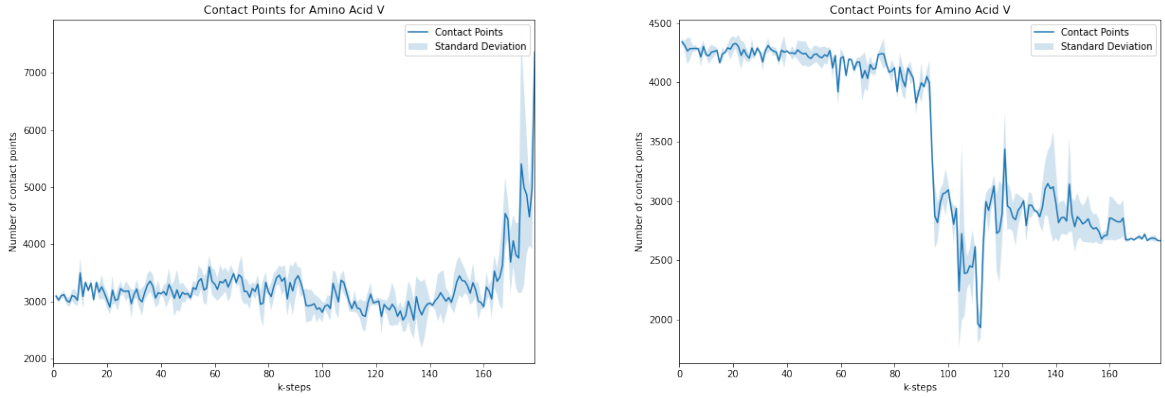
(a) ESMFold **(b)** ColabFold

Figure B.16: Trajectory S. Contact Matrix along the trajectory towards the homopolymer of amino acid S.



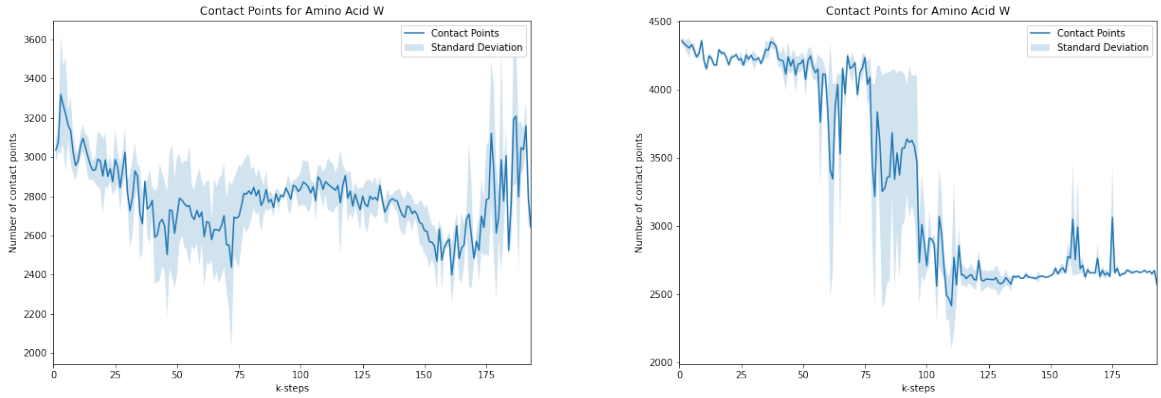
(a) ESMFold **(b)** ColabFold

Figure B.17: Trajectory T. Contact Matrix along the trajectory towards the homopolymer of amino acid T.



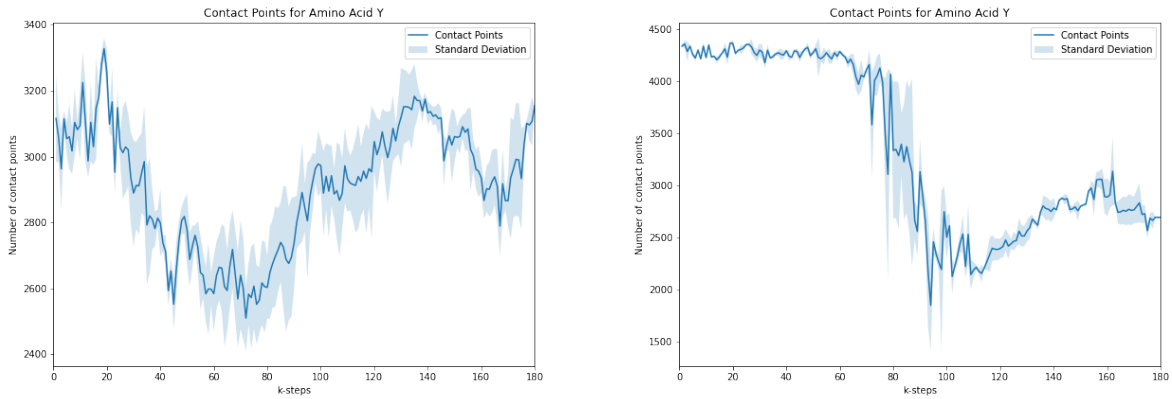
(a) ESMFold **(b)** ColabFold

Figure B.18: Trajectory V. Contact Matrix along the trajectory towards the homopolymer of amino acid V.



(a) ESMFold **(b)** ColabFold

Figure B.19: Trajectory W. Contact Matrix along the trajectory towards the homopolymer of amino acid W.



(a) ESMFold **(b)** ColabFold

Figure B.20: Trajectory Y. Contact Matrix along the trajectory towards the homopolymer of amino acid Y.