University of Reading

Improvement of AlphaFold2-Based Methods for Modelling Quaternary Structures of Proteins

A thesis submitted for the degree of
Doctor of Philosophy
School of Biological Sciences
University of Reading

Ahmet Gurkan Genc
September 2024

Declaration

I confirm that this is my own work and to the best of my knowledge, does not breach copyright law, and has not been taken from other sources except where such work has been cited and acknowledged within the text.

Ahmet Gurkan Genc:

Date: 16 /12 /2024

Abstract

The functions of proteins are determined by their 3D structures, hence different methods have been developed in order to predict protein structures as a stepping stone to better understanding their functions and interactions. Protein structure modelling was a process that often involved two different stages: modelling and refinement. However, the release of the deep neural network-based AlphaFold2 (AF2) as a protein modelling tool in 2020 has enabled significant advances in protein bioinformatics. These advances have made it possible to predict models of monomeric structures that are close to structures derived experimentally. Thus, the effective application of machine learning approaches has reduced the need for the traditional refinement process. Instead, modellers use end-to-end processes for improvements covering both modelling and refinement. One of the most important developments in for such processes was the open-access release of the AF2 code. As a result, almost all recent tools have integrated the AF2 code into their own pipelines using different methods and parameters, aiming to obtain better models than those produced by the default AF2 method. However, the successes achieved for monomeric globular structures have not yet been realised for multimeric globular structures, and this has increased the need for the development of new modelling tools. Although many AF2 versions have been introduced in the process, the full effectiveness of AF2 - and indirectly, which structures it can accurately predict - is not yet fully understood. Therefore, the basis of this research is to investigate the features of this black box and to explore how to use it most effectively for the improvement of quaternary structure models. In this direction, we aimed to design an improved AF2-based multimeric protein modelling pipeline.

The effect of recycling, a key part of the AF2 algorithm, on the refinement of models is investigated in Chapter 2. The results show that in both AF2 versions (AF2_Advanced and AF2_Multimer (AF2M)) the quality of the predicted protein model improves as the number of recycles increases. It is also shown that while 3 cycles is the default value for the AF2 versions, 12 cycles may be more effective for both main versions. With the integration of the custom template option into the AF2M code, the effect of custom templates and recycling methods on protein modelling are examined in Chapter 3. It is shown that providing initial structural information to AF2M as an input and further recycling can lead to better quality structure models. It is also emphasised that using multiple sequence alignment (MSA) inputs is more effective in AF2M compared to providing a single sequence (SS). Another new parameter introduced for AF2M for improving modelling was the custom MSA option. Although the effectiveness of custom templates and custom MSA options have been supported by many

studies, the effect of altering these input features on AF2M rather than using the defaults had not been fully revealed. In Chapter 4, we discovered that when multimeric custom template structures are given to AF2M as a “single-chain” protein structure, a cumulative improvement in TM-scores and IDDT scores are observed, although there is no improvement in interface scores (QS-scores and DockQ_wave scores). Furthermore, in order to obtain custom MSAs, disordered residues in homologous sequences were deleted within MSAs, so that AF2M made its predictions only from residues corresponding to ordered regions. As a result, it was also observed that AF2M obtained higher quality protein structures in more than half of the targets. These two major results emphasise that input changes to AF2M can be more effective for target-specific protein modelling than for general protein modelling. Finally, based on the results from the previous chapters, we designed two successive versions of a protein modelling tool called MultiFOLD, which aims to create a pool of models with conformational sampling using custom template recycling followed by ranking and selection. In Chapter 5, through extensive analysis of benchmarking data we demonstrate that MultiFOLD is particularly effective in modelling multimeric globular structures, and the latest version, MultiFOLD2, outperformed all other servers including AlphaFold3 (AF3) that are participating in the CAMEO-BETA project. With the acquisition of better-quality protein structures, it is now possible to better infer function and to model protein-ligand interactions in downstream analyses.

Contents

List of Figures

List of Tables

List of Abbreviations

AF2	AlphaFold2
AF2M	AlphaFold2-Multimer
AF3	AlphaFold3
BFD	Big Fantastic Database
CAMEO	Continuous Automated Model EvaluatiOn
CASP	Critical Assessment of Structure Prediction
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CRYO-EM	Cryogenic Electron Microscope
EMA	Estimation of Model Accuracy
DNN	Deep Neural Network
FAPE	Frame-Aligned Point Error
FFT	Fast Fourier Transform
FM	Free Modelling
GABAA	Gamma-aminobutyric Acid Type A
GNN	Graph Neural Network
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
IDDT	Local Distance Difference Test
MD	Molecular Dynamics
ML	Machine Learning
MoRFs	Molecular Recognition Features
MSA	Multiple Sequence Alignment
MQA	Model Quality Assessment
NLP	Natural Language Processing
NMR	Nuclear Magnetic Resonance
NN	Neural Network
PAE	Predicted Aligned Score
PDB	Protein Data Bank

pIDDT	Predicted lDDT score
PSSM	Position-Specific Scoring Matrix
pTM-score	Predicted TM-score
QS-score	Quaternary Structure Score
Slims	Short Linear Motifs
SS	Single Sequence
TBM	Template-Based Modelling
TM-score	Template Modelling Score

Acknowledgement

I am profoundly thankful to everyone who supported and guided me during my PhD journey.

First, I am profoundly grateful to my supervisor, Prof. Liam J. McGuffin, for his invaluable and profound guidance, endless patience, and constant encouragement. His unwavering support helped me stand up and keep moving forward every time I felt discouraged during my PhD journey.

I am also deeply thankful to my family, whose love, understanding, and sacrifices made this achievement possible. To my father and mother, I deeply appreciate your confidence in me and your ongoing encouragement. To my brother and his wife, your support has been my foundation, and I am forever grateful.

Special thanks to my dear friend, Dr. Recep Adiyaman, whose encouragement, advice, and companionship made this journey more meaningful.

I would also like to express my heartfelt thanks to all my friends who have been by my side throughout this process. Your support and laughter made the challenges bearable, and your presence was a source of joy and motivation.

Finally, I would like to express my sincere appreciation to the Republic of Turkiye Ministry of National Education for their support and the valuable opportunities extended to me.

The following outputs have arisen from the work presented in this thesis:

McGuffin, L. J., Edmunds N. S., **Genc, A. G.**, Alharbi, S. M. A., Salehe, B. R. and Adiyaman, R. (2023) Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. Nucleic Acids Research, 51, W274–W280. https://doi.org/10.1093/nar/gkad297

My Main Contribution: The design of MultiFOLD

Adiyaman, R., Edmunds, N. S., **Genc, A. G.**, Alharbi, S. M. A. and McGuffin, L. J. (2023) Improvement of protein tertiary and quaternary structure predictions using the ReFOLD refinement method and the AlphaFold2 recycling process. Bioinformatics Advances, 3, vbad078. https://doi.org/10.1093/bioadv/vbad078

My Main Contribution: The analysis of CASP14 Multimeric Targets

Edmunds, N. S., Alharbi, S. M. A., **Genc, A. G.**, Adiyaman, R. and McGuffin, L. J. (2023) Estimation of Model Accuracy in CASP15 Using the ModFOLDdock Server. Proteins, 91, 1871-1878. https://doi.org/10.1002/prot.26532

My Main Contribution: Conceptualization, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing—original draft

Edmunds, N. S., **Genc, A. G.** and McGuffin, L. J. (2024) Benchmarking of AlphaFold2 accuracy self-estimates as indicators of empirical model quality and ranking - a comparison with independent model quality assessment programs. Bioinformatics, 40, btae491. https://doi.org/10.1093/bioinformatics/btae491

My Contribution: The support of methods (obtaining the CASP14 recycling models)

Genc AG, McGuffin LJ. Beyond AlphaFold2: The Impact of AI for the Further Improvement of Protein Structure Prediction. Methods Mol Biol. 2025;2867:121-139. doi: 10.1007/978-1-0716-4196-5_7. PMID: 39576578.

Chapter 1: General Introduction

The book chapter will be published in Springer Nature. The publication is:

Genc AG, McGuffin LJ. Beyond AlphaFold2: The Impact of AI for the Further Improvement of Protein Structure Prediction. Methods Mol Biol. 2025;2867:121-139. doi: 10.1007/978-1-0716-4196-5_7. PMID: 39576578.

Author contributions:

Genc A. G: Conducting extensive literature review, Compiling synthesized findings from various publications, Drafting the manuscript

McGuffin L. J.: Overview and editing, and guidance for designing the last manuscript

1.1 Protein structure

Proteins are the material base of all living organisms and are involved in all cellular processes, such as guidance for biochemical reaction catalysis, ensuring the correct genetic information expression, and the transmission and transduction of signals (Jiang et al., 2017). Proteins are amino acid polymers that are bound by peptide bonds in specific linear sequences. With distinct physiochemical properties, the 20 essential amino acids have different side chains. These features determine the folding of the polymer into a three-dimensional structure: the native conformation (Anfinsen, 1973), which in turn make possible the astonishing variety of molecular functions (Kuhlman & Bradley, 2019). The amino acids are categorised as nonpolar, polar acidic, polar basic and polar neutral amino acids. Nonpolar side chains are the most frequent, whereas other residues contain side chains that range in size, shape, acidity, chemical reactivity, and charges that can be positive or negative (Branden & Tooze, 1991; Williamson, 2012). Understanding the structure of proteins is critical for understanding how they function. Protein structures are generally defined at 4 levels- primary, secondary, tertiary, and quaternary.

1.1.1 Primary structure

In the polypeptide chain, the primary structure of a protein relates to the amino acid sequence. Peptide bonds that are formed during the process of protein biosynthesis keep the primary structure together. The two ends of the polypeptide chain, depending on the existence of the free group on each terminus, are referred to as the amino terminus (N-terminus) and the carboxyl terminus (C-terminus). Amino acids are composed of four chemical groups: a carboxyl (COOH) group, an amino (NH2) group, a hydrogen (H) atom, one changeable group, known as a side chain or R group (Figure 1.1 A). R groups differentiate on the basis of shape, hydrophobicity, size, reactivity, and charge (Sanvictores & Farci, 2023). Residues that make each amino acid unique are bound to a carbon atom, called C-alpha. Two C-alpha atoms are joined by a peptide bond, involved in the atoms of C, O, N, and H in a plane: however, the bonds joining it to the C-alpha can rotate (Figure 1.1 B). Phi (Cα-N bond) and Psi (Cα-C bond) angles are the rotation angles of these bonds, which are the only flexible parts in the peptide chain (Song et al., 2012b).

1.1.2 Secondary structure

Segments of folded polypeptide chains, in general, follow conformations where the main chain's torsion angles (Phi and Psi angle) replicate in a normal fashion and create secondary structure components such as α-helices and/or β-sheets (Pauling et al., 1951). α-helices are helical structures reinforced by the hydrogen bonds between each fourth amino acid. At the same time, β-sheets form from extended strands of parallel or antiparallel segments formed which are stabilized by longer hydrogen bonds between interacting amino acids (Jiang et al., 2017). Turns, which are a flexible part of the protein structure, play a significant role in folds, loops, and interactions, i.e. the elements that bond regular secondary structure units in protein folding (Song et al., 2012a). Protein secondary structure can be thought of as an intermediate between primary sequence and tertiary structure. Hence, many computational methods for secondary structure prediction have been improved, with using by combining both local (adjacent residues) and long-range contact information (Zhang et al., 2018).

1.1.3 Tertiary structure

The tertiary structure level refers to the three-dimensional arrangement of atoms inside a protein. Secondary structures are packed in order to form varied number of folding units called “domains”. Generally, one domain includes approximately 100-150 residues. The 3D structure is folded in the way they reach the lowest energy, known as Anfinsen theory (Anfinsen, 1973), which confers the functional activity. Christian Anfinsen's work in the 1950s demonstrated that the information included in the amino acid sequence determines the three-dimensional structure of a protein. He used urea and beta-mercaptoethanol to break down the secondary and tertiary structures of the protein while working on the ribonuclease enzyme. In his first experiment, when he removed the two substances together, the protein refolded and became active. In his second experiment, when he removed beta-mercaptoethanol first and then urea, the protein misfolded and became biologically inactive. This was due to the failure to form non-covalent bonds. In his third experiment, when he exposed the misfolded protein to trace amounts of beta-mercaptoethanol, the protein formed the correct disulfide bonds and returned to its native and active state. This was because the native structure was the most thermodynamically stable form (Anfinsen, 1973). The folding is stabilised by hydrophobic interactions, hydrogen bonds, disulfide bonds and, electrostatic attractions (Cozzone, 2010).

Disulfide bridges are covalent bonds with strong interactions generated between the sulfhydryl groups of two cysteine residues (Sevier & Kaiser, 2002). Hydrogen bonds and ionic interactions between the polar and charged amino acids allow the tertiary structure to keep a unique shape. However, they can be weaker than other types of interactions (Rehman I et al., 2024).

Protein folding is the physical process by which a protein chain is converted to its native three-dimensional structure, mostly a "folded" configuration that allows the protein to operate biologically functions. From a random coil, a polypeptide folds into its distinctive three-dimensional structure. Many proteins begin folding even during polypeptide chain translation. The amino acid sequence or primary structure determines the resultant three-dimensional structure – this fact is termed “Anfinsen's dogma”. Folding is a process that is mainly guided by hydrophobic interactions, formation of intramolecular hydrogen bonds, van der Waals forces (Pratt & Cornely, 2012). During folding process of a protein, Chaperone proteins give an important role in the cell in order not to aggregate misfolded protein structures from sequences by binding to folded intermediate structures (Mashaghi et al., 2014). This means that chaperone proteins are essential molecules present in all organisms, helping other proteins in proper folding, refolding, and move to their correct cellular locations (Wergin, 2006). Physical contacts of protein-protein interactions are high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect (Titeca et al., 2019). Hydrophobic amino acids are buried within proteins, isolating them from water and this hydrophobic effect cause to makes a protein fold stable (van Dijk et al., 2015).

1.1.4 Quaternary structure

The majority of proteins in a cell interact to form complexes to perform their functions. Such complexes are also known as quaternary structures or oligomeric assemblies. Complexes of proteins, or oligomers, comprise a mixture of several separate monomeric folded tertiary structure chains or subunits (Marsh & Teichmann, 2015). By having various interaction partners, proteins regulate their functions in a cell according to changes in the surrounding environment (Morris et al., 2022). Protein complexes are prevalent in nature and can be separated into two types (Yu et al., 2006). homo-oligomers, which comprise identical subunits such as the homo-tetramer structure of the potassium channel (Doyle et al., 1998), and hetero-oligomers, which comprise varied subunits such as the hetero-pentamer structure of the

gamma-aminobutyric acid type A (GABAA) receptor (Tretter et al., 1997). The four levels leading up to the formation of complex structures are illustrated in Figure 1.1 C.

According to protein-protein interfaces extracted from the PDB, each protein can be divided into three subunits: interface, non-interface, and protein core. Also, these subunits include residue compositions with different solvent accessibility, sequence entropy, size, and preference of contact. Specifically, hydrophobic and aromatic residues are more common in the interface area than hydrophilic residues, which indicates that hydrophobic residues are a critical component for complex stabilisation (Yan et al., 2008).

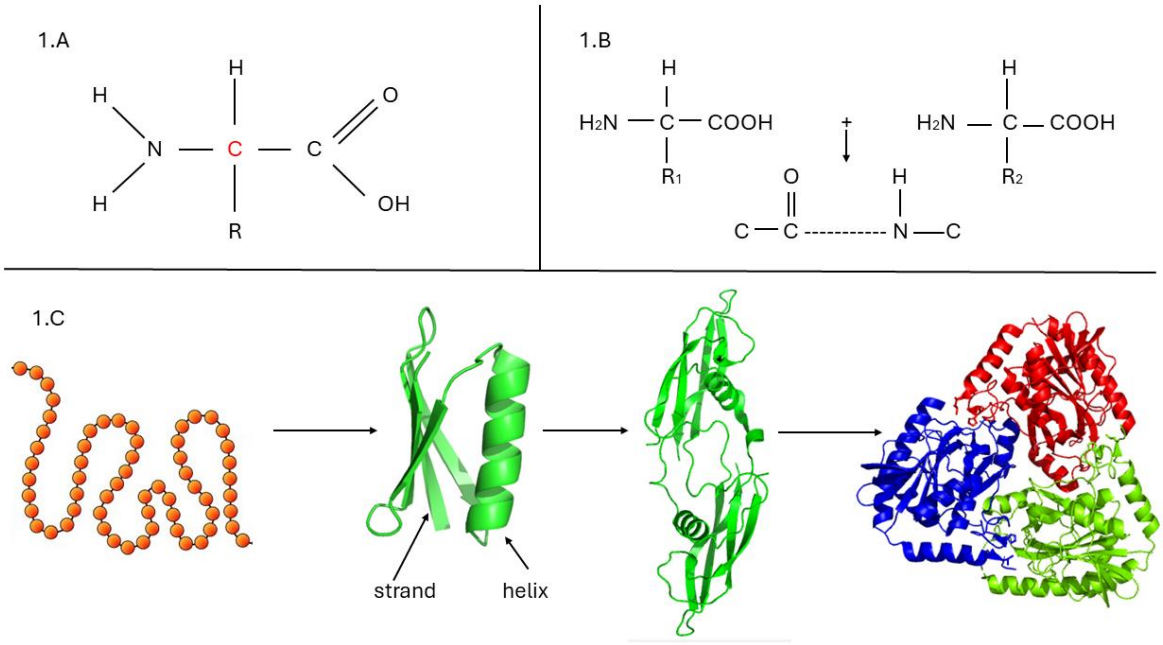


Figure 1.1 The fundamentals of protein structure.
The figure representing the levels of protein structures from sequence to final complex structure. Part A shows the general structure of amino acids. Part B depicts the linkage between two amino acids, where one amino acid is connected to the carboxyl group of the other through a peptide bond. Part C represents various protein structural elements from left to right: The primary structure (where a specific colour represents each amino acid in a linear pattern), secondary structures (with a helical structure on the upper part and a sheet structure on the lower part), tertiary structure (where various combinations of secondary structures are folded in three dimensions), complex structures (where each colour represents a distinct interacting chain, resulting in the assembly of multiple tertiary structures). The molecular graphics in 1.A and 1.B and the protein sequences in 1.C were drawn by using Google Draw, while the parts of protein structures were drawn by using PyMOL.

According to tissue-based protein research, Lehner and Fraser (Lehner & Fraser, 2004) discovered mammalian protein domains as tissue-specific and studied their cellular functions. Bossi and Lehner (Bossi & Lehner, 2009) further showed that protein interactions are tissue-specific and that universally expressed proteins often interact with tissue-specific ones. This

suggests that proteins may undergo structural changes in response to tissue-specific requirements. Protein folding depends on factors such as the length of a protein, domain structures and intrinsically disordered regions (Anfinsen, 1973; Olzscha, 2019). Anfinsen's experiments (Anfinsen, 1973) showed that these factors determine protein folding. However, there are additional factors that influence protein folding within the cell, including molecular crowding and tissue-specific environmental differences (Arndt et al., 2010; van den Berg et al., 1999). Protein structure can also be altered by post-translational modifications, resulting in spatially different conformations (Nussinov et al., 2012).

1.1.5 Disordered regions of protein

The theory that proteins carry out their functions through their unique three-dimensional structures (Anfinsen, 1973) has been challenged by molecular dynamic (MD) simulations (Mao et al., 2010) and sequence analyses (Das et al., 2015), which have shown that certain structures contradict this theory. These structures, known as intrinsically disordered structures, do not fold into a specific 3D conformation under certain physical conditions. Instead, they create a diverse conformational ensemble, allowing them to perform various functions (Uversky & Dunker, 2010). These polypeptide structures, often rich in polar and charged amino acids, do not have a sufficient number of hydrophobic amino acids to form stable folds (Uversky et al., 2000). Nevertheless, the amino acid composition of these disordered structures plays a constraining role in their ability to explore an unlimited conformational space (Mao et al., 2010).

Approximately 40% of eukaryotic proteomes (Wright & Dyson, 2009) consist of disordered regions, and their active involvement in many human diseases such as cancer (Iakoucheva et al., 2002) and neurodegenerative diseases (Uversky, 2014) has been observed. Disordered protein regions can also participate in the formation of functional complexes. Compared to interactions between structured protein regions, disordered regions often exhibit weaker binding affinities, resulting in more transient interactions (Latysheva et al., 2015). Interactions between these disordered regions are facilitated by short linear motifs (Slims) (Davey et al., 2012) and molecular recognition features (MoRFs) (Disfani et al., 2012), which act as interaction sites (Tompa et al., 2014). These features, located within the disordered regions, often induce a specific conformation when interacting with their binding partners (Singh et al., 2007).

1.2 Protein pKa

Protein pKa calculations are used to determine the acid strength of amino acids in free form or protein complexes. pKa values of amino acids play an essential role in defining the pH-dependent characteristics of the protein. The pKa value of a protein can fluctuate as it folds, depending on its three-dimensional structure and the surrounding environment. The determination of pKa is useful in structural bioinformatics, computational biology and molecular modelling. pKa accurately calculates the physical model of protein electrostatics, which can aid structure-based energy estimating approaches (Alexov et al., 2011). The binding changes the pKas of ionizable groups, which causes proton uptake/release and, as a result, the pH dependency of the binding energy (Jensen, 2008). Likewise, the shift in pKas upon protein folding causes pH dependence of the folding energy (Yang & Honig, 1993).

Biological macromolecules are designed to perform certain roles in specific cellular environments (subcellular compartments or tissues); as a result, they must be compatible with the biophysical characteristics of the associated environment, one of which is the characteristic pH. Many macromolecular characteristics, including as stability and activity, are thus pH dependent (Talley & Alexov, 2010) which is as a consequence of the change in protonation of ionise residues. Along with its effect on protein structure, stability and solubility, the types of interactions of polar side chains will have with their surroundings are determined by their protonation state (Grimsley et al., 2009). In the general meaning, the state of ionization depends on the pH of the solution. It is important to note that the resting pH of blood is about 7.4, which is between the pk values of the side chains, histidine and Cysteine. Even so, variation by more than about 0.2 pH unit is hazardous (Lesk, 2016).

1.3 Protein sequence and structure databases

UniRef: The UniRef database is a system that clusters protein sequences based on similarity thresholds of 100%, 90%, and 50%, also aiming to remove redundant proteins. In this way, it serves as a core component of the Universal Protein Resource. The UniRef database leverages sequence clusters, which are produced from UniProtKB and UniParc data (Leinonen et al., 2004) provided by the UniProt Consortium. These clusters are designed to maximize the coverage of the sequence space by grouping related protein sequences, thus facilitating a

more comprehensive and effective analysis of protein variety and function (Suzek et al., 2007; Suzek et al., 2014).

Uniclust: The Uniclust database, much like UniRef, clusters sequences obtained from UniprotKB based on three different similarity thresholds (%90, %50, %30). However, the key difference between these two databases lies in their sequence clustering methods. While UniRef utilizes the CD-HIT software system, Uniclust employs MMseq2 (Steinegger & Söding, 2017). The primary advantage of using MMseq2 is its increased sensitivity to distant homologous sequences, allowing for the lowering of the sequence similarity threshold to as low as 30%.Furthermore, MMseq2 has developed a cascaded clustering model to cluster functionally homologous sequences more effectively than UniRef50 and UniRef90.Additionally, by utilizing HH-suite (Steinegger et al., 2019a) for Uniclust sequences, it has facilitated the recognition of 17% more Pfam domains (Mirdita et al., 2016). The workflow of the core repositories for protein sequences is detailed below.

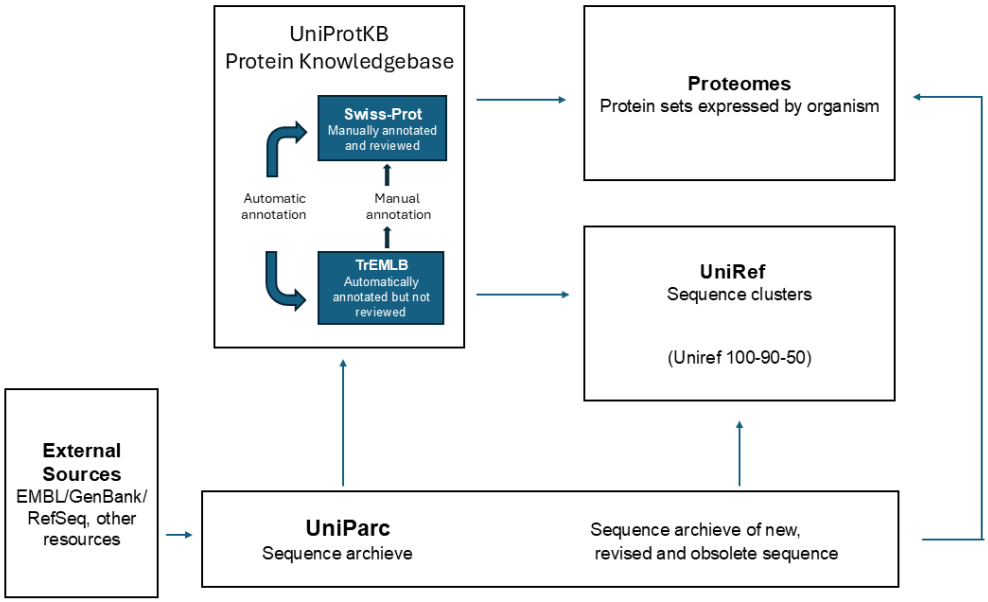


Figure 1.2 The workflow of database organisation.
It illustrates how databases encompassing UniProt interact and operate. This workflow was designed using Microsoft PowerPoint, inspired by the content on www.ebi.ac.uk. Scientific literature, automatic annotation, sequence analysis tools, and other databases enable the integration of protein sequences and related functional information into the UniProt database.

Mgnify: MGnify brings together data from various projects and facilitates their harmonized analysis, aiding in better comprehension and comparison of metagenomic data in an academic context. With the 2022 update, the database currently contains more than 2.4 million non-redundant sequences representing 297 distinct biomes, and more than half of MGnify's analyses originate from just nine of them: human faecal, oral, digestive system, skin, and unspecified human; marine; soil; mammalian digestive systems; and mixed biome samples. As long-read sequencing data from PacBio and Oxford Nanopore Technologies becomes more common alongside short-read technologies like Illumina, the database was expanded to support both long-read-only and hybrid datasets (combining long and short reads from the same sample). The MGnify protein database (the last version of MGnify in 2023) consists of 2,477,479,951 protein sequences, containing all protein sequences from analyses of assembled data – a combination of long and short metagenomic sequences. It plays a vital role as a source of additional sequences for multiple sequence alignments (MSAs) used by AlphaFold2, enhancing protein families underrepresented in conventional databases with sequences from metagenomic sources (Richardson et al., 2022).

In addition to these databases, sequence sets generated using different algorithms have also emerged. MetaClust (Steinegger & Söding, 2018) was created by merging approximately 1.5 billion protein sequence fragments from about 2,200 metagenomic/transcriptomic datasets using Prodigal (Hyatt et al., 2010). and clustering them. Additionally, the Big Fantastic Database (BFD) (Jumper et al., 2021a; Steinegger et al., 2019b) was obtained by clustering 2.5 million protein sequences from the Soil Reference Catalog Marine Eukaryotic Reference Catalog, Metaclust, and Uniprot/TrEMBL+Swissprot. The Linclust (Steinegger & Söding, 2018) algorithm was particularly beneficial in clustering the BFD dataset, as it is independent of the number of resulting clusters (K), thereby revealing the rich source of information in metagenomic and genomic sequences.

Protein Data Bank (PDB): The PDB, which has been organized by the Worldwide Protein Data Bank (wwPDB) since 2003, is a global repository that houses all of three-dimensional experimental structures of biological macromolecules such as proteins and DNA. This essential resource facilitates the collection, validation, and curation of data and provides open access to 3D structures. With the PDBx/mmCIF primary data format, metadata can be handled and stored in the PDB Core Archive (wwPDBconsortium, 2019).

In bioinformatics, in addition to the advantage of a huge number of protein sequences, the presence of similar or homologous sequences in a dataset leads to redundancy, which can

introduce unwanted bias in certain analyses. Hence, their redundancy must eliminate in much research (Carugo, 2008). However, the majority of eukaryotic genes produce transcript isoforms through alternative splicing, which results in functional diversity in interactions between proteins, proteins and DNA, and ligands. These genes also encode numerous protein types through alternative transcription, splicing, 3′ end formation, translation, and post-translational modifications (Pan et al., 2008; Wang et al., 2008). Redundant sequences, which result from the repetition of highly identical sequences must be filtered out in the various research (Hobohm et al., 1992). However, isoforms are sequences derived from the same gene but differentiated by biological processes. Although isoforms share mostly similar sequences, they may have different regions or alterations (Torrens-Fontanals et al., 2021). Such sequences are considered variations that derive from the same gene and have biological significance, so they must not be removed in the sequence analysis. To remove redundant, various computer programs are available and utilize different alignment methods, including local or global alignment and clustering algorithms (Sikic & Carugo, 2010) while the homology detection tools identify isoforms with additional criteria such as gene annotations (Miller et al., 2022) and functional differences (Ferrer-Bonsoms et al., 2020).

1.4 Experimental methods of tertiary and quaternary structures of proteins

Experimental methods that play a crucial role in structural biology for solving protein structures (both tertiary and quaternary structures) at resolutions allowing the elucidation of heavy atom positions (<3 Å), include X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). These methods are widely employed, and the resulting structures are deposited in the PDB for further research by the wider scientific community. X-ray crystallography is the most widely used method for determining the structures of regular protein molecules, constituting approximately 89% of the PDB entries. NMR accounts for approximately 7% of the PDB entries, while cryo-EM represents about 3% of the structural data available in the PDB. However, cyro-EM has gained popularity primarily since it does not require crystallization (Seffernick & Lindert, 2020). Despite their widespread use, each of these methods has its own set of advantages and disadvantages, which are outlined in Table 1.1.

Table 1.1 Advantages and disadvantages of experimental methods for solving tertiary structures of proteins.
The table representing positive and negative side of three experimental methods for solving tertiary protein structures including X-ray, NMR, and cryo-EM (Dokholyan, 2020; Seffernick & Lindert, 2020). X-ray provides high-resolution structures; however, it has trouble with large complexes. Although peak overlapping and line broadening problems restrict NMR to smaller structures, it provides conformational flexibility. Cryo-EM does not require crystallization and can observe many conformational states, but it is often lower resolution and limited to large protein structures.

Experimental method	Advantage	Disadvantage
X-ray crystallography	It provides high resolution structures	Large complex structures, membrane proteins and disordered regions are hard to obtain and it requires samples to be successfully crystallised.
NMR	It provides information on conformational flexibility and can be done in solution	Peak overlapping and line broadening issues restrict its application to smaller structures
cyro-EM	It does not require crystallization and can be used to observe many conformational states in large complexes	It is limited to large protein structures and has lower resolution

According to the multimeric complexes, after the tertiary structures were obtained, it is crucial to know oligomeric information for tertiary structures. While X-ray crystallography cannot directly acquire the functionally relevant multimeric form of a 3D structure, solution NMR or cyro-EM can obtain them. Rather, the PDB Depositor may supply this information as metadata from further experiments, infer it by using software like PISA (Proteins, Interfaces, Structures and Assemblies) (Krissinel & Henrick, 2007) or EPPIC (Evolutionary Protein Protein Interface Classifier) (Duarte et al., 2012) to make predictions before the quaternary structures are stored in the PDB (Korkmaz et al., 2018).

1.5 Computational prediction methods on tertiary protein structures

Computational methods have been growing interest in predicting tertiary structures as they have obtained 3D coordinates of atoms without needing the more resources and time which is indispensable for the experimental methods. However, the main challenge for bioinformatics has been to predict the 3D structures at high accuracy. The general workflow for predicting tertiary structures starts from the prediction of the protein fold and ends with the refinement process of predicted tertiary models. The main computational methods include template-based

and template free modelling. With the help of these methods, many different conformations for given proteins can be obtained, which in turn should select the appropriate structures close to native-like structures using model quality assessment tools. After the selection of them, the refinement process has been used to bring them closer to the native structure (Adiyaman & McGuffin, 2019).

Template based models which use the closest homolog as a template in the database (Roche & McGuffin, 2016) while template free modelling use physical principles to predict tertiary structures when the appropriate template is not available (Dorn et al., 2014). Nevertheless, the template free modelling is capable of predicting proteins with up to 150 residues in order to obtain accurate prediction when there are no available template structures (Lee et al., 2009). In addition, template free method also requires huge resources for large conformational structures search (McGuffin, 2008(b)). Hence, the template based approach has been a more used and accurate method to predict protein structures (Fiser, 2010). After AF2 was released, machine learning approaches have been growing interest in predicting tertiary structures, especially, by obtaining residue-residue distances at high accuracy. AF2 showed great performance in CASP13 for the category for the tertiary structure prediction, this tool managed to solve the folding problem and to predict the tertiary structure at very high accuracy (Senior et al., 2020). Although the main approaches are the same for predicting tertiary and quaternary structures, predicting the quaternary structure of proteins remains the main challenge. Quaternary structures of proteins include interface areas, which request the higher computational calculation and processing time. In the next chapter, the main computational methods for predicting protein complexes were introduced.

Table 1.2 Methods for predicting tertiary structures of proteins.
The table representing the methods and standalone refinement tools for modelling protein tertiary structures. Since the refinement protocols has been required to use after modelling, most of these tool are only used for refinement tools, after the release of AF2. The below tools have been common tool for tertiary structure prediction.

Tool	Method	Reference
AlphaFold2 (AF2)	Machine-learning	(Jumper et al., 2021a)
RoseTTAFold2	Machine-learning	(Baek et al., 2023)
OmegaFold	Machine-learning	(Wu et al., 2022b)
ESMFold	Machine-learning	(Lin et al., 2023)
RaptorX-Contact	Machine-learning	(Xu, 2019)
I-Tasser-MTD	Machine-learning +Template-based	(Zhou et al., 2022)
Phyre2	Template-based	(Kelley et al., 2015)
MODELLER	Template-based	(Webb & Sali, 2016)
SWISS-MODEL	Template-based	(Biasini et al., 2014)
QUARK	Template-free	(Xu & Zhang, 2012)
IntFOLD7	Template-free + Template based	(McGuffin et al., 2023)
ReFOLD3	Refinement	(Adiyaman & McGuffin, 2021)
GalaxyRefine	Refinement	(Heo et al., 2013)
FG-MD	Refinement	(Zhang et al., 2011)
ModRefiner	Refinement	(Xu & Zhang, 2011)
3Drefine	Refinement	(Bhattacharya et al., 2016)

1.6 Computational structure prediction of protein complexes

The prediction of protein quaternary structures or complexes is beneficial for drug design, protein engineering, and function analysis (Quadir et al., 2021). High-throughput experimental methods like yeast two-hybridization can determine whether two proteins establish a permanent or temporary interaction. However, this cannot accurately indicate the complex details of the interacting protein structures. Biophysical experimental methods such as X-ray crystallography, NMR, and cryo-EM can reveal the location and the way in which proteins interact. These methods, however, are costly and time-consuming (Biasini et al., 2014). Hence, various computational methods for rapidly modelling quaternary structures of proteins have been developed over the years. These predictive methods can be broadly classified as template-based modelling (TBM), template-free (TF) docking, and machine learning approaches.

1.6.1 Template-Based Modelling of Complexes

Template-based modelling (TBM) is built on the paradigm that proteins with similar sequences will constitute similar complexes structures (Chakravarty et al., 2020). TBM attempts to model complexes for target proteins with unknown structures by firstly identifying existing protein-protein complexes in the PDB with either similar sequences or tertiary structures to the target. These identified complexes are then used as templates for building the target complex. The main stages of TBM include obtaining one or more available template(s) and aligning them with the given sequence using either template-based or profile-based alignment; constructing a starter model for the given target by cloning the structural fragments from the matching parts of the template(s); changing the side-chains to match the target sequence, building termini regions; and then refining the model by considering its all atomic structure (Szilagyi & Zhang, 2014). Historically, homology modelling has been the most dependable computational technique and it can be applied to both tertiary and quaternary structures if templates with similar sequences can be found. It can be successfully performed for complex models when the structure of a pair of proteins has at least 30% sequence identity (Aloy et al., 2003) previously stored in PDB (Negroni et al., 2014). Fold recognition or threading can be used to successfully build tertiary structures in cases where there is less than 30% identity, but so far it has not been as successful for modelling quaternary structures.

1.6.2 Template Free (FM) Docking

Protein-protein docking usually attempts to solve the problem of modelling complexes by fitting together either known or predicted tertiary structures of the individual protein subunits in the absence of any known quaternary structure templates. The concept of steric complementarity at the protein-protein interface area has been central to the docking process and a driving force in the development of docking techniques. Frequently, physicochemical complementarity including electrostatics (Mandell et al., 2001), hydrophobicity (Berchanski et al., 2004), and tendencies based on statistics (Mintseris et al., 2007) have been included in the process. The docking process comprises two main stages, sampling and scoring. In the sampling process, considering two individual structures, the docking process attempts to sample all potential binding patterns of a complex structure. This can be divided into a rigid body search (global search) and conformational search (local search) of the binding modes, considering protein structure flexibility along with rigid body sampling. In the scoring process, a scoring function is used to rank the sampled binding patterns (Huang, 2014). The rigid docking approach neglects differences between a bound and an unbound structure and takes into account only six degrees of freedom, while flexible docking comprises a larger number of internal coordinates (Vakser, 2014). Additionally, symmetry information is often used to predict homo-oligomer structures (André et al., 2008) and most tools employ information generated by experimental surveys, such as distance restraints (Bonvin et al., 2018). Even though there have been many docking conformational approaches, such as fast Fourier transform (FFT) and particle swarm optimization, it is still difficult to select the nearest to native (i.e., most accurate) models from a huge number of alternative models, frequently referred to as decoys (Moal et al., 2013; Wang et al., 2021).

1.6.3 Machine Learning Approaches

Machine learning (ML) is concerned with improving approaches that help to automatically extract information from training data in order to uncover certain regularities and use them to develop general and accurate models capable of making predictions for hidden data. ML approaches power almost every aspect of modern society from computer science to life science. Deep learning is a subset of ML approaches that rely on deep neural networks (DNNs) based on representational learning, which may be supervised, semi-supervised, or unsupervised (Cios et al., 2007; Schmidhuber, 2015). Specifically, a DNN is a type of machine

learning technique that mimics biological neural networks. Each DNN includes nodes (similar to neurons) communicating with other nodes through connections (similar to dendrites and axons). The weighting of connections between nodes in an DNN is dependent on their capacity to achieve the desired output, mimicking synapses between neurons that are reinforced when their linked neurons produce correlated outputs in a biological neural network (Choi et al., 2020; Hastie et al., 2001).

Deep learning approaches have recently started to be used frequently in protein bioinformatics (Suh et al., 2021). The ability of DNNs to utilise large amounts of suitable datasets and employ appropriate functions for specific tasks has made them suitable for use in protein structure prediction. This has provided a clearer expression of the concepts of both homology and evolutionary information. In addition, the use of advanced computer equipment, such as graphical processing units (GPUs), has further accelerated this process (Kuhlman & Bradley, 2019; Owens et al., 2008; Rost & Sander, 1994). ML approaches generally work by inferring a large number of protein features from MSAs obtained from protein homologous sequences, using intensive computer power (Kandathil et al., 2022).

The most notable method that has emerged in recent years was AlphaFold2 (AF2) from the DeepMind Group, which attained exceptional accuracy in protein structure prediction. This new tool provides insights into the functions of proteins with previously unknown structures and allows for the speedy resolution of modelling difficult targets with accuracy reaching that of X-ray crystallography and Cryo-EM structures. The AF2 method is one of the best examples of a DNN-based approach, consisting of a two-stage neural network architecture capable of end-to-end tertiary structure prediction (Baek et al., 2021; Jumper et al., 2021a). The AF2 approach was further applied to carry out end-to-end prediction of protein complexes with AlphaFold2-Multimer (AF2M) (Evans et al., 2022). The most important success of such approaches is the use of end-to-end optimized models instead of relying on complex and multi-stage processes obtained through human intervention. The neural network model establishes a connection between local and global protein structure using geometric units, which are designed to optimize the overall geometry of the protein without compromising the integrity of its local covalent chemistry (AlQuraishi, 2019).

1.6.3.1 AlphaFold2 (AF2)

Published by DeepMind at CASP13, in 2018, AlphaFold (version 1.0) (AF) modelled target proteins with the highest accuracy, including the hardest FM targets (Senior et al., 2020). Additionally, the method was competitive with the best template-based modelling methods in the TBM category, despite using no templates explicitly as part of its modelling process. AF used a DNN trained using MSA-based features for - 30,000 non-redundant protein structures in the PDB- and it was capable of accurately revealing pairwise distances between all residues in any given target sequence. A large number of estimated distances provided differentiated distance information regarding adjacent residues, covariation, and the local region of the structure. This neural network also allowed for the extraction of the predicted distribution of the backbone torsion angles (Pinheiro et al., 2021). Trained on 180,000 PDB constructs (Thornton et al., 2021), the next version of the method (AF2) also predicted structures more accurately than existing tools, with a median RMSD value of 0.96 Å backbone accuracy. Its strength was confirmed, in particular, by its success in predicting CASP14 targets. After the method was published, it was revealed that several features were key to the success of modelling including the reuse of process losses in iterative cycles to refine predictions (this process was termed “recycling”). DeepMind subsequently open sourced the AF2 code (https://github.com/google-deepmind/alphafold) and this promising tool has since been shown to predict structures with significant domain accuracy and domain packing (Jumper et al., 2021a).

In the AF2 algorithm, a general network pattern aims to predict the Cartesian coordinates of all heavy atoms in a given protein using only the target amino acid sequence and its aligned homologous sequences as an input MSA. The network includes two major parts. The first component is a new type of neural network called an Evoformer. It has the ability to manipulate information in the MSA and Nseq × Nres array (Nseq, number of sequences; Nres, number of residues) that allows direct display about the evolutionary and the spatial relationships. The second component is the structure network comprising each residue of a given protein structure rotationally and translationally in the form of a global rigid body. While the most important invention of this tool is to allow the simultaneous refinement of the entire structure by breaking the chain and to introduce loss terms that provide significant weight on the correctness of the residual orientation, it also introduces a new equivalent transformer to allow the reasoning of the unshown side-chain atoms. The iterative refinement process, namely recycling, presents an especially important part of the AF2 tool, and as the number of cycles increases, recycling feeds the system with the final output, allowing the structure to be

predicted with better quality scores. In this network, there is a relaxation section of the final structure that uses an Amber force field, with gradient descent minimization. This minimization in AF2 does not affect the backbone quality scores (GDT and IDDT-Cα) of the output 3D structure, so this selection is optional. However, it eliminates stereochemical violations without facing the loss of the quality of a structure (Jumper et al., 2021a). Figure 1.2 summarizes the workflow of AF2.

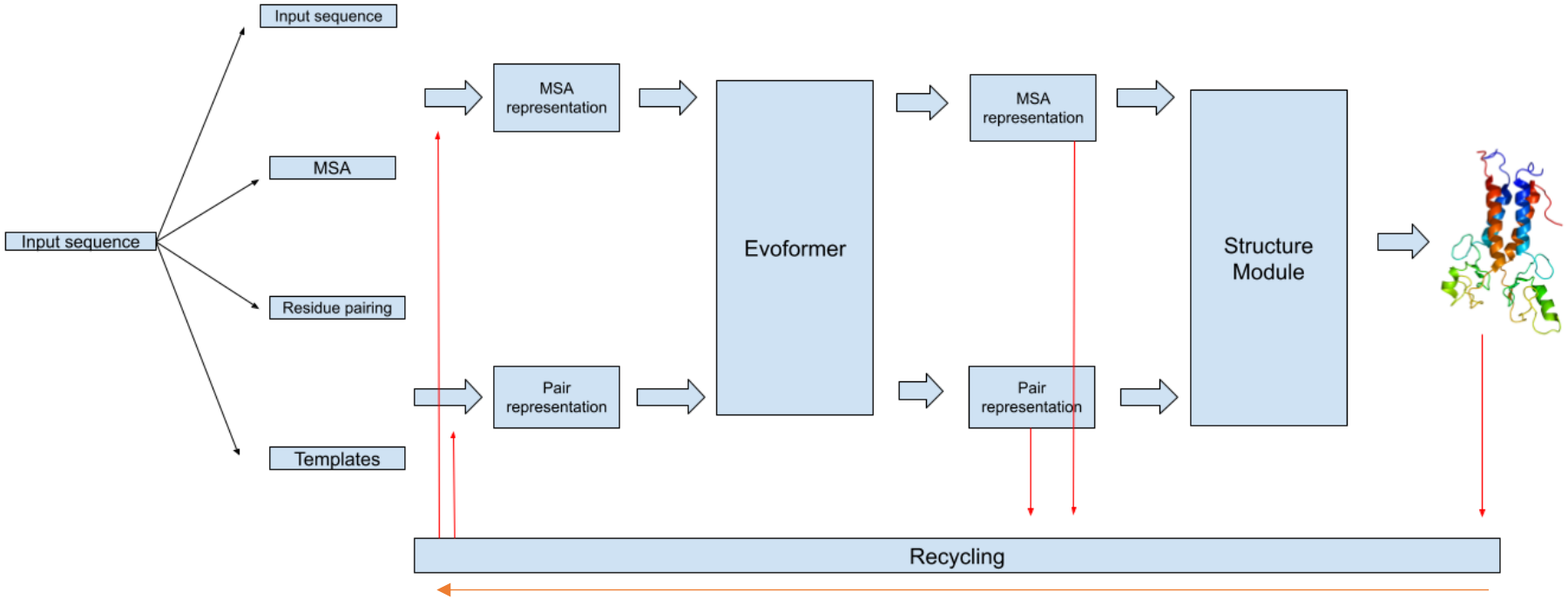


Figure 1.3 Workflow of AF2.
The method starts from input sequence(s) to model quaternary structuress of proteins. AF2 includes two main DNN blocks: Evoformer and Structural Module. AF2 employs the recycling method as improvement methods. The orange arrow demonstrates the recycling process, while the red arrows show the inputs and outputs within the recycling process. The workflow was adapted from Jumper et al. (2021a) and prepared using Microsoft Powerpoint.

1.6.3.2 AlphaFold2-Multimer (AF2M)

Given AF2's effectiveness in predicting protein structures, the obvious issue is whether the approach can also be used to accurately model protein complexes. The CASP14 results demonstrated that most of the residues situated at the domain-domain interfaces of multidomain proteins were predicted exceptionally well by AF2. Despite the lack of training of the methods on interacting protein chains, several of the estimated residue contacts nevertheless appeared to be well suited for predicting interactions between chains. As a result, it was proposed that similar DNN-based approaches could also be adapted directly to predict complex protein structures (Baek et al., 2021; Egbert et al., 2021; Ozden et al., 2021).

AF2M is a new version of AF2, which was specifically trained on protein complexes. Several AF2 versions were released in recent years, including the popular AF2M method, as shown in Table 1.2. AF2M was released with some modification of the basic method of AF2 algorithms as version 2.1. Firstly, permutation symmetry was included in the multimer pipeline, which can pick homomer chains with the best matches relative to ground truth coordinates. Secondly, the aligned sequence was provided explicitly in the receiving pairwise correlation in MSA for heteromeric interaction. Thirdly, AF2's clipping process was modified where the system was trained on a clipped region of the full-length amino acid sequence, taking into account both the clipped region for complex regions and well-balanced portions for interface and non-interface segments. Additional significant modifications addressed the Frame Aligned Point Error of 10 Å where a better gradient signal was provided for the wrong interface followed by consideration of different types of chain structures based on a given amino acid pair. The second change involved the training and inference regimen which included a new weighted combination of predicted TM and interface TM confidence scores (Evans et al., 2022).

Table 1.3 The versions of AlphaFold so far.
The table of the AF2 versions over time, their release dates, and the underlying important differences for each version.

AlphaFold version	Release date	Key differences
1.0.0	16/07/2021	Initial release
2.0.1	30/09/2021	Incorporation of pLDDT scores into the B-factor column of output files (PDBs).
2.1.0	02/11/2021	Adding AF2M weights
2.1.1	05/11/2021	Minor bug fix
2.1.2	28/01/2022	Working the relaxation method on GPU
2.2.0	10/03/2022	Updating of AF2M model parameters
2.2.1	13/06/2022	Integrating new CUDA version (11.1.1)
2.2.2	13/06/2022	Minor bug fix
2.2.3	25/08/2022	Transforming PAE json results in Colab to new output in order to use for AF database (AFDB)
2.2.4	21/09/2022	Minor bug fix
2.3.0	13/12/2022	Updating of AF2M model parameters Smaller GPU memory footprint
2.3.1	12/01/2023	Increasing MSA sequence length as input to 4,000 sequences
2.3.2	05/04/2023	Applying the relaxation method only to the best unrelaxed model

(Note: the AlphaFold3 (AF3) server was released in Summer of 2024, but no source code has been released at the time of writing. See Chapter 5 for benchmarking of AF3 performance against our in-house methods.)

1.6.3.3 ColabFold

Inspired by open source AF2 code, the prediction community has experimented with forks of the original AF2 code. The ColabFold (Mirdita et al., 2022) project is a fork of AF2 which, for efficiency, outsources the generation of MSAs to a remote alignment server called MMseq2. It also provides users with options to adapt the input MSAs for modelling both homo- and hetero-oligomeric complexes using the original AF2 training data. This first code was made freely available and can be run using a Google Colab notebook called AlphaFold2_Advanced (https://github.com/sokrypton/ColabFold).

To harness the power of AF2, powerful computer resources such as a high spec GPU with plenty of RAM are required for running the DNN, and significant Central Processing Unit (CPU) and storage capacity is needed to store the sequence databases required for creating the input MSA. To democratise access to the method, the Google Colaboratory notebook (the merger of Jupyter Notebook with Google) version of AlphaFold (ColabFold), was released, so that general users can now make use of AF2 without the need for investing in their own high computational resources. The main difference in the ColabFold fork of AF2 is the inclusion of fast MMseq2 (Mirdita et al., 2019) searches, so the generation of MSAs is greatly accelerated for both protein tertiary and quaternary structure prediction. The link of last updated Colab notebook of AF2 is : (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) This is different from the original version of the main AF2 method, where searching databases happens via HMMER (Eddy, 2011) and HHblits (Remmert et al., 2012), which still requires a long runtime (Mirdita et al., 2022).

MMseq2 is an open-source software package for searching and clustering large protein and nucleotide sequence libraries. MMseq2 outperforms BLAST, the most widely used sequence similarity web server, by a ratio of 10,000 to 1 in terms of speed and it has the same sensitivity as BLAST. The MMseq2 search process is divided into three parts: 1) a short word (‘’k-mer’’) match process, the most crucial step for improving a search, 2) vectorized ungapped alignment, and 3) gapped (Smith-Waterman) alignment (Steinegger & Söding, 2017). Additionally, the BFD and the Mgnify databases used in the AlphaFold algorithm were reduced in size, and this version of the database was referred to as BFD/MGnify with an extended environmental search database to better include eukaryotic protein diversity. As previously mentioned, BFD is a database that was designed by clustering 2.5 million protein sequences

from different sources including Soil Reference Catalog, Uniprot/TrEMBL + SwissProt, while Mgnify (Mitchell et al., 2020) is the database for the assembly, processing, and storage of microbiome data produced from microbial population sequencing (Mirdita et al., 2022).

1.7 Improvement of protein quaternary structure prediction

New protein structure prediction methods aim to improve upon initial structural models, bringing them closer towards their natural state, thereby achieving accuracy closer to that provided by experimental data. Prior to the AF2M era, protein structure improvement techniques employed conformational sampling to fine-tune a physical force field and move the starting structure closer to its native conformation (Bhattacharya, 2019). After the introduction of AF2M, DNNs that integrate existing experimental datasets such as Cyro-EM (Terwilliger et al., 2022) have also been employed to improve the resolution of structures. In addition, models generated by AF2M can be improved through the integration of sequence-based approaches, such as metagenomic or sequence embedding approaches or protein language models, and structure based approaches, including the use of experimental data, graph neural networks (GNNs) and exploiting the recycling process of AF2M, as shown in Figure 1.3.

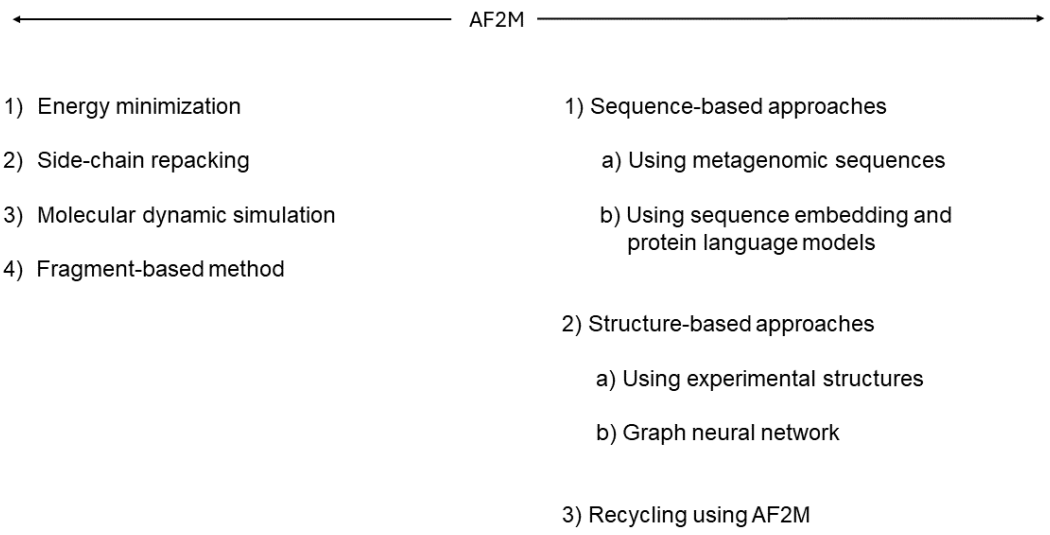


Figure 1.4 Methods for improving model quality before and after AF2M.
Traditional refinement methods (energy minimization, side-chain repacking, MD simulations and fragment-based approach) were utilized prior to AF2M. In contrast, sequence- and structure-based improvements have come into play following AF2M's introduction.

1.7.1 Refinement of 3D structure models before AF2M

Before AF2M was developed, protein structure refinement methods were typically categorized into four groups: MD simulations (Adiyaman & McGuffin, 2021), energy minimization (Xu & Zhang, 2011), side-chain repacking (Heo et al., 2013), and fragment assembly (Bhattacharya & Cheng, 2013) (Figure 1.3). MD simulations generate numerous MD trajectories by adhering to physical laws that govern how atoms interact with each other. Energy minimization techniques aim to discover the structure with the lowest energy by optimizing the arrangement of both core and side-chain atoms using physical and knowledge-driven force fields. Fragment-based approaches leverage template fragment data obtained from the PDB in combination with statistical potentials (Wu et al., 2023). The process of altering the positions of the most probable rotamers within a structure (according to the rotamer library) is referred to as side-chain repacking (Heo et al., 2013).

Traditional refinement methods often necessitate substantial conformational sampling, which can be both time-consuming and computationally resource intensive. Therefore, incorporating restraints has become a common strategy in refinement protocols (Wu et al., 2023) . Various protocols employ this method differently: ReFOLD3's latest version (Adiyaman & McGuffin, 2021) utilizes restraints guided by quality estimation scores, whereas Feig's approaches (Feig & Mirjalili, 2016) apply restraints to backbone atoms. While utilizing one of the aforementioned refinement methods can be beneficial for monomeric structures, more intricate protocols that combine multiple refinement methods are often necessary for multimeric complexes.

The initial quaternary structure models produced by many protein-protein docking processes have been typically formed using rigid-body approaches. However, rigid-body docking does not account for minor conformational changes that may occur when bonding occurs, notably at the interface. Therefore, refining the complex structure with rigid docked approaches is essential if the complex target is to be used for downstream applications, such as determining ‘hotspot’ residues for drug design or conducting more precise protein-protein interaction or structure-function research (Verburgt & Kihara, 2022). Refinement methods should aim to arrange more natural residue-residue contacts at the interface, improve interface complementarity, and provide better shape conformity with respect to the native complex. Additionally, scoring functions should be capable of identifying native or native-like docking patterns (Schindler et al., 2015). A major concern with refinement methods for both tertiary and quaternary models of proteins is the risk of moving the target protein structure away from,

rather than towards, the native protein structure, especially when the subunit already has a high-quality score (Adiyaman & McGuffin, 2019; Verburgt & Kihara, 2022). Schindler *et al.* (2015) recognized the relationship between the small-scale movement of interface residues and the large-scale displacement of protein structures. This residue movement at the interface can create a smooth energy landscape through a structured-based force field, potentially helping the complex protein escape from local minima and transition towards a near-native structure.

1.7.2 Improvement of 3D structure models after AF2M

In the present day, since most protein structure tools relies on DNN-based approaches, investigating which traditional methods still maintain their validity for improving such structures is an ongoing subject of research. It has been noted that MD simulations not only fail to enhance but can even deteriorate 3D structures generated by ML models like AF2M (Heo et al., 2021). Furthermore, within the AF2 algorithm, using an AMBER force field-based energy minimization method is unlikely to yield notable improvements in the structure (Jumper et al., 2021a). Hence, emphasizing the concept of 'improvement' rather than 'refinement' could lead to the establishment of a novel methods that have the potential to enhance AF2M structures. This concept represents a combination of modelling and refinement. Table 1.3 shows the modelling methods with improvement methods or the standalone refinement methods.

Table 1.4 Methods for predicting quaternary structures of proteins.
The table representing the methods and standalone refinement tools for modelling protein quaternary structures. Since the protein model improvement approach is developed as the last stage of structure prediction tools rather than as a standalone tool, both the common prediction tools and their refinement parts (standalone refinement tools, if available) are shown together. Following the introduction of AF2M, most tools provide improvements through the integration of multiple approaches.

Tool	Method	References
AF2M	AI-based	(Evans et al., 2022)
RosettaFold	AI-based	(Baek et al., 2021)
DeepComplex	AI-based	(Quadir et al., 2021)
MultiFOLD	AI-based	(McGuffin et al., 2023)
OpenFold	AI-based	(Ahdritz et al., 2022)
ESMFold	AI-based	(Lin et al., 2023)
OmegaFold	AI-based	(Wu et al., 2022a)
ClusPro	Docking	(Kozakov et al., 2017)
Swarm-Dock	Docking	(Torchala et al., 2013)
LZerD	Docking	(Christoffer et al., 2021)
ZDOCK	Docking	(Pierce et al., 2014)
MEGADOCK 4.0	Docking	(Ohue et al., 2014)
InterEvDock2	Template-based or Docking (depends on the case)	(Quignot et al., 2018)
EGR	standalone-refinement tool	(Morehead et al., 2022)
iATTRACT	standalone refinement tool	(Schindler et al., 2015)
GalaxyRefineComplex	standalone refinement tool	(Heo et al., 2016)
HADDOCK	standalone refinement tool	(Dominguez et al., 2003)
FiberDock	refinement tool based docking	(Mashiach et al., 2010)

1.7.2.1 Sequence based approaches

The biological and functional properties of a protein are determined by specific sequences, which have been shaped by mutations over evolution and recorded in a certain pattern within these sequences (Lin et al., 2023; Thomas et al., 2008; Yanofsky et al., 1964). By aligning related sequences, known as MSA, it is possible to extract the structural and functional features of a protein from the patterns within these sequences (Thomas et al., 2008). The complete sequencing of numerous genomes has opened up new opportunities for the generation of deeper MSAs, leading to the expansion of a reservoir of information encompassing residues that, while not adjacent in the amino acid chain, functionally co-evolve. This multitude of information has been harnessed to predict distance restraints and matrices, which guide the construction of three-dimensional protein structures (AlQuraishi, 2021; Marks et al., 2011). Thus, sequence similarity searches are the main method for protein description and analysis (Steinegger et al., 2019a; Steinegger & Söding, 2017; van Kempen et al., 2023) ,aiming to seek homologous sequences in order to infer characteristics such as functions and structure. Improved inter-residue contact maps from earlier research and the use of distance information between residues (Ji et al., 2019; Jumper et al., 2021b) have enhanced prediction of quaternary structure of proteins. Better contact maps have been beneficial in diminishing the transitive effect, which occurs when two residues contact a third residue (Ji et al., 2019). Following the releasing of AF2M versions, this type of deep MSA data has begun to be used as a training dataset.

Since contact and distance maps are inferred from MSAs derived using various tools for detecting homologous sequences (Peng et al., 2022) such as HHblits (Remmert et al., 2012) and MMseqs2 (Mirdita et al., 2019), searching for variant sequences in sequence space, particularly for remote homologs of proteins, can be challenging for improving protein structure prediction (Ben-Hur & Brutlag, 2003). Typically, standard sequence-structure research is based on a 25% similarity threshold to homologous sequences (Zhang & Skolnick, 2005). However, this approach is not powerful to detect very remote homologs. Hence, instead of relying solely on sequence homology, using a hybrid approach that includes structural alignments (Hamamsy et al., 2023)- preserved over long evolutionary scale (Zhang & Skolnick, 2005) - can be beneficial, especially for orphan proteins that lack homologous sequences in databases (Suzek et al., 2014). In this case, the structural annotation rates can increase by up to 70% as shown by metagenomic studies (Vanni et al., 2022). Therefore, using metagenomic

sequences to improve MSAs in combination with structure data has the potential to improve quaternary structures of proteins.

1.7.2.1.1 The use of metagenomic sequences

Beyond protein sequences, metagenomic data constitutes an extensive resource for the identification of novel proteins characterized by functional structures. Consequently, there is a growing interest in harnessing metagenomic data to extract evolutionary insights into particular proteins. This is accomplished by integrating metagenomic sequences into MSAs, which are then employed as inputs for DNNs, thus improving the accuracy of protein structure predictions. Due to the limitations of data available in UniProtKB (Boutet et al., 2007), combining genomic and metagenomic databases is increasingly important. The importance of incorporating metagenomic data lies in the remarkable increase in the Neff (number of effective sequences) value, which represents the number of protein sequences in the MSA that provides the most homologous information while including the fewest sequences. An Neff value of 128 has been used in DeepMSA (Zhang et al., 2020). Nonetheless, it should be noted that employing additional metagenomic data does not invariably lead to a more precise MSA (Hou et al., 2022). Yang *et al*. (2021) found that utilising one or a few specific microbes associated with the target protein family is more beneficial for constructing MSAs than using all similar metagenomic sequences. Sequences in the metagenomic databases are predominantly of prokaryotic origin. Despite the growing number of sequencing projects focused on fungi and other eukaryotic genome due to modern technology, applying approaches designed for prokaryotic genomes to eukaryotic-specific protein families remains a limitation (Ovchinnikov et al., 2017). The ESM Atlas (https://esmatlas.com/) created by ESMFold and curated by Meta AI, encompasses a comprehensive repository of structures for novel metagenomic sequences (772 million predicted metagenomic structures) that were predicted using a protein language model based approach (Lin et al., 2023).

1.7.2.1.2 The use of sequence embedding and protein language models

The utilization of next-generation sequencing technologies has resulted in a rapid and exponential growth in protein sequence databases, which effectively double in size approximately every two years. Nonetheless, annotating these proteins with precise and meaningful data necessitates effort, expertise, empirical research, and financial investment (Consortium, 2018). Hence, the so-called "sequence-structure gap" (Rost & Sander, 1996) is

progressively expanding. Most protein structure prediction tools rely heavily on sequence information (Kotowski et al., 2021). In traditional structure prediction approaches, sequence data for a target protein is commonly supplied in various formats, including a position-specific scoring matrix (PSSM) (Gribskov et al., 1987), a Hidden Markov Model (HMM) (De Fonzo et al., 2007) , a single protein sequence, or a k-gram (Qiu et al., 2020). In addition to relying on multiple homologous sequences, the emergence of AF2M has firmly established the MSA a crucial input for protein structure prediction (Yuan et al., 2023). However, when the average sequence similarity between sequences is low, MSA methods struggle to produce high-quality alignments, which leads to incorrect inferences in downstream applications. Most alignment methods attempt to find similarities using a substitution matrix-based scoring method; however, this approach often fails to yield an effective MSAs for proteins with low similarity rates (McWhite et al., 2023; Nute et al., 2019).

Recently, there have been important advancements in Natural Language Processing (NLP), notably with the utilization of pre-trained language models (Otter et al., 2021; Qiu et al., 2020). Language models learn from vast amounts of unlabelled linear data through supervised or semi-supervised learning, capturing the structure patterns within sequences. Embedding information has proven to be valuable in downstream applications, notably in protein sequence analysis. Researchers have increasingly begun to employ these techniques, called Protein Language Models, in the analysis of protein sequences (Qiu et al., 2020). Several language models have since been developed specifically for protein analysis.

These deep learning language models are trained on large datasets of protein sequences to predict the identity of hidden amino acids (masked amino acids) based on their surrounding context in the sequence. After training, each amino acid is represented as a high-dimensional vector in the embedded pattern. These vectors capture the sequence context of each amino acid by encoding information about neighbouring amino acids and their relationships within the sequence (McWhite et al., 2023). ML models can then use these embeddings to predict the conservation or similarity of a specific amino acid position across different sequences or to identify homologous sequences with a common evolutionary origin (Marquet et al., 2022).

These pre-trained models can predict structure in an unsupervised manner, whether provided a SS (Rives et al., 2021) or an MSA (Rao et al., 2021a) as input. This capability stems from their extensive training dataset (Bhattacharya et al., 2022) and the use of attention mechanism to transfer learned knowledge (Bahdanau et al., 2014). Attention mechanisms have recently

been used in the field of protein bioinformatics. These mechanisms assign different scores to individual positions within the input, enabling the model to focus on the most relevant information. In protein contact predictions, attention mechanisms assign weights to individual positions on a 2D image, enabling the visualisation of critical residual contacts that play an important role in the structure prediction during inference (Chen et al., 2021(b)). The AF2 algorithm uses transformer structures based on attention mechanisms to predict protein structures with high accuracy. AF2 consists of two main transformer blocks: an encoder and a decoder. These blocks contain multi-head attention layers to learn the relationships between protein sequences. The evoformer block focuses on MSA (Multiple Sequence Alignment) and pair representation. The MSA attention layer performs weight calculations on a large protein symbol matrix. To reduce the computational cost, this attention is divided into row-based closed attention and column-based closed attention components. Row-based attention determines the relationship of amino acid pairs, while column-based attention emphasises pairs that it considers to contain more meaningful information. In the pair representation pathway, an attentional mechanism based on triangle relationships is used, where the sides of each triangle are updated by influencing each other. The structural module of AF2 consists of 8 blocks and each block updates the single representation and spine frames. Invariant Point Attention (IPA), an important component in this module, focuses on updating the singular representation and generates 3D equivalent attention values. IPA is invariant to global rigid motions (rotation and translation) and is not affected by them. IPA predicts the relative rotational and translational motions of each backbone frame, thus enabling more precise modelling of the structure. These attention mechanisms and transformer structures of AF2 are important in protein structure prediction (Yang et al., 2023(b)). Transformer models address numerous challenges encountered by traditional deep learning methods, which rely on homologous sequences. The attention module in the transformer model enables each token (amino acids in sequences) to influence the weights of all other tokens in the sequence. This capability allows the transformer model to focus on distant relationships within input tokens, accounting for the entire context of an input sequence and leading to improved results and sequence embedding. Transformer based protein language models have outperformed AF2M and MSA-based methods (Chandra et al., 2023). While several transformer models incorporate evolutionary information from MSAs during pre-training, this pre-training is usually a one-time process. New protein representations are typically retrieved using the transformer models' pre-trained hidden states. MSA-based tools involve a time-consuming (Hong et al., 2021) alignment process with homologous sequences retrieved from the UniProt database. In contrast, embeddings obtained from language models provide a more extensive and detailed

set of information, even when the target protein has fewer homologous sequences (Wang et al., 2022).

Despite the success of sequence-based homology inference, many proteins remain unannotated as identifying distant evolutionary relationships purely from sequences cannot always be achieved (Mahlich et al., 2018). To address this, OmegaFold (Wu et al., 2022a) was designed as one of the first methods using attention based transformer approaches to predict protein structures from a single sequence (SS). OmegaFold outperformed RoseTTAFold and showed competitive results with AF2M, making it effective for modelling protein structures without relying on MSA, particularly for orphan proteins and fast-evolving proteins, such as those involved in antigen-antibody interactions (Wu et al., 2022a). Fine-tuning is an effective approach for leveraging protein language models. Models such as ESM-1b (Rives et al., 2021) and ProtTrans (Elnaggar et al., 2022) provide a fundamental framework for a wide range of protein-related tasks and are primarily used for feature extraction. However, customizing these pre-trained models for specific downstream applications by adapting them to specific tasks can be a more effective approach (Yang et al., 2023(a)). When labelled data is limited, fine-tuning is usually considerably faster and more effective than training a new model from scratch (Ofer et al., 2021). OpenFold (Ahdritz et al., 2022) is an open-source tool that reconfigures AF2M, enhancing its effectiveness (more rapid, more memory-effective) through the fine-tuning method. It also facilitates training on new datasets.

1.7.2.2 Structure-based improvement approaches

Coevolutionary and homology information has been employed to generate initial models (Hiranuma et al., 2021) as most tools are integrated with AF2M and/or protein language models, especially in the CASP15 competition. (https://predictioncenter.org/casp15/doc/CASP15_Abstracts.pdf). Hence, the main challenge in improving predicted protein structures is to more effectively sample the conformational space of the target proteins. The structure space to search, even in close proximity of the initial protein structure model, is extremely huge (Feig, 2017; Heo & Feig, 2018). However, using data that includes initial protein structural information can be beneficial to DNNs in guiding conformational sampling (Tian et al., 2021). Better structural predictions can be achieved by using computational methods that incorporate structural knowledge of proteins. For instance, AF2 and RoseTTAFold2 obtain residue distance information from sequence covariation obtained through MSA (Baek et al., 2021; Jumper et al., 2021a). Additionally, AlphaFold-

Phenix employs low resolution experimental atomic coordinates from the density maps obtained through Cyro-EM (Terwilliger et al., 2022).

Conformational sampling described by generative DNNs such as autoencoders (Tian et al., 2021), and generated from MD simulations can also be utilized to extract suitable conformational information (Degiacomi, 2019). DNNs have further improved the geometric properties of predicted protein structures. It has been shown that graph representations of protein structures can be more effective than contact or distance representations for capturing global information and residue correlations.

GNNs can thus be utilised to improve models. A graph consists of nodes (such as atoms or residues) connected by edges (such as bonds or contacts). It is often represented with a feature matrix that contains information about each node and an adjacency matrix that describes how nodes are connected. A GNN is an algorithm that repeatedly updates each node's information by considering its connections with neighbouring nodes. This process is called "message passing". Thus, GNNs can be trained to optimize the updating of node information, making them valuable for predicting the characteristics of individual proteins or collections of proteins (Igashov et al., 2021). GNNs are advantageous compared to Convolutional NNs (CNNs) because CNNs, which are typically used for generating contact/distance maps, require larger datasets to learn orientations in 3D input data and face challenges with inconsistencies on large interface surfaces within the predefined 3D grid size, especially when predicting protein-protein interfaces (Réau et al., 2022). SE(3)-Transformer, a recent method to improve the performance of GNNs, is a version of the self-attention method that works on 3D point clouds and graphs. This model is sensitive to the position and rotations of the data in 3D space and can maintain its performance unchanged according to these transformations. The concept of equivariance in SE(3) allows a model to change its output in a similar way despite transformations in the input (e.g. rotation and translation). This makes the model sensitive to unnecessary transformations in the data and makes the results more stable. For example, when an object is rotated or translated, the model reacts in the same way, resulting in consistent performance. This allows the equivariance approach to be used effectively in the modelling of 3D protein structures. This efficiency of the equivariant allows the model to reuse the same weights in the transformer for different transformations as they are sensitive to input transformations, helping the model to learn more efficiently and generalise better with fewer parameters (Fuchs et al., 2020).

For effective prediction of protein structure complexes, the protein-protein interface must be accurately modelled (Shuvo et al., 2023). In most ML approaches, the features obtained from the protein interface dataset and used for interface prediction include: amino acid types, physicochemical properties of amino acids such as hydrophobicity, interface propensity (meaning varied interaction propensities according to physicochemical properties), inter-chain evolutionary information, relative solvent accessibility, and surface shape. Due to the limited discriminatory power of individual features in distinguishing interfacial residues from the remaining residues within a protein, several prediction methods employ a combination of multiple features (Xue et al., 2015). Among these features, the most significant one identified thus far is the evolutionary information encapsulated in PSSMs (Yan & Wang, 2014). However, utilizing structural information, rather than relying solely on sequence data, has been shown to have significant effect on improving performance, as illustrated by studies conducted by BIPSPI (Sanchez-Garcia et al., 2018) and PAIRPred (Minhas et al., 2014). In these studies, it was observed that utilizing structural information instead of sequence data alone led to an average improvement of 10% in both precision and recall (Dai & Bailey-Kellogg, 2021).

1.7.2.3 Recycling using AF2M

The term "recycling" refers to the final process of AF2M's protein structure modelling, evaluated as the important part in ablation studies. Recycling involves reinserting the final and intermediate outputs into the system in an embedding format. During the training phase, recycling makes the network deeper without increasing the number of parameters or extending training time. In the inference phase, it creates a new network initialisation by utilizing the structural output and input features from the previous network.

In the training process of AF2M, a different approach was adopted rather than simply running through every cycle (Jumper et al., 2021a). Instead of using N cycles, Equation 1 below utilizes intermediate outputs as an auxiliary loss to enhance predictions. It is hypothesized that this approach allows the network to iteratively refine its own predictions multiple times. The efficiency of this procedure is notably enhanced by halting gradients for the intermediate outputs, which in turn improves both memory usage and computational efficiency.

$$\frac{1}{Ncycle}\sum_{c=1}^{Ncycle} c = \frac{Ncycle+1}{2} \qquad (1)$$

Initially, an objective was established to determine the average loss across all iterations. Then, the number of iterations between 1 and N cycles was sampled to construct an unbiased Monte Carlo estimate of this objective. Next, the flow of gradients from a specific iteration to the previous iterations was halted, effectively skipping the backward pass - a step in training where gradients are computed and used to update the model's parameters (weights and biases) through backpropagation.

The output pair and first row MSA representations from the Evoformer, along with the estimated backbone atom coordinates from the Structure module, were recycled in AlphaFold's embedding form. Two types of representations were used: (zij) and (mij). These representations underwent a process known as "LayerNorm processing" a technique used in deep learning methods to normalize the inputs to a neural network layer. To calculate the pairwise distances between elements, predicted coordinates of beta carbon atoms (or alpha carbon for glycine) were used. These pairwise distances were then discretized into distinct intervals. In this process, the distances were divided into 15 bins of equal width, each spanning 1.25 angstroms (Å), covering a total range of approximately 20 Å. A one-hot encoded distogram was processed and modified through linear projection before being added to a pair representation update. Additionally, recycling updates, which contain information from previous iterations, were integrated into the network's operation (Jumper et al., 2021a).

1.8 Critical assessment of structure prediction (CASP)

Established in 1994, CASP is held every two years to drive cutting-edge research and development of new technologies for protein structure prediction and evaluate the progress in accuracy. It is seen as the international gold standard for evaluating structure prediction approaches, and it represents a global community engaged in a competitive effort. Participants must predict the target protein structures from their amino acid sequences in a double-blind procedure, before the experimentally derived coordinates become available. Subsequently, once the observed structures are known, the predicted structures are compared to the experimentally determined ones and the best prediction methods are identified (Moult, 2005). The CASP14 competition (in 2020) included eight categories: high accuracy modelling, topology**,** contact and distance prediction, refinement, assembly, accuracy estimation, data-assisted prediction, and biological relevance while the last CASP competition (CASP15 in 2022) included six categories: single protein and domain modelling, assembly, accuracy estimation, RNA structures and complexes, and protein-ligand complexes. Refinement,

contact and distance prediction, and domain-level estimates of model accuracy were not included. Despite DeepMind's absence from CASP15, AF2 had continued to make a significant impact. The groups that integrated AF2 into their pipelines were among the most successful participants. These groups employed two main strategies to improve upon AF2: 1) utilizing more effective templates and/or MSA techniques, and 2) enhancing AF2 by implementing dropout methods (Elofsson, 2022). In order to evaluate the performance of their server or methods, participants in the CASP competition are provided with protein sequences by the CASP assessors, with their structures disclosed at the end of the competition. In the model prediction part of CASP competition, there are targets ID for each target sequence, which represent prefix “H” represents heteromeric proteins and prefix “T” represents homomeric proteins.

1.9 Project Objectives

Since 2020, AF has been continuously developed, and its evolving models have been adapted to tools designed by the bioinformatics community. As a result, AF2 has become a fundamental approach in structural bioinformatics. In light of this, the objective of this project is first to investigate how to effectively use the AF2 versions in order to improve the accuracy of modelling protein quaternary structures. Especially since both AF2 performs well for globular proteins and CASP targets include only globular context proteins, membrane (lipid layer) proteins were ignored in this research. Hence all the data were about globular proteins. Although the refinement category was removed in CASP15, this stage remains relevant, especially for protein complexes. Traditional refinement methods, when applied to structures generated through DNNs, have often been shown to be detrimental to model accuracy. Thus, ultimately, improvement methods integrated with modelling tools are used more frequently than standalone tools in this process. Furthermore, AF2 itself may be utilized as a tool for improving upon input models. Therefore, this project aims to explore the numerous options for integrating AF2 and utilizing parts of its pipeline to improve model quality. Among these, the recycling process, using custom templates, and/or custom MSAs are particularly prominent. Effectively using these options is crucial for better structure modelling and so the optimal parameters are explored. Finally, these optimal approaches are integrated with our MultiFOLD server versions for improved modelling with performance exceeding that of the default AF2 versions.

The primary focus of protein modelling tools is to obtain the best static structure for a given protein. To achieve this, specific scores are used, and the best one is selected. However, the actual environment in which proteins function is often not mentioned. Proteins are constantly interacting with environmental factors such as water or lipids in their cellular membranes. An important limitation is that current modelling tools do not adequately take these environmental factors into account. For example, some proteins may form a hydration ring on part of their surface, while other regions may interact with lipids. A DNN-based method that takes residue-lipid interactions into account for the interactions of transmembrane proteins with the lipid layer or for the detailed evaluation of fibrous structures has not yet been developed.

A DNN-based method that takes residue-lipid interactions into account for the interactions of transmembrane proteins with the lipid layer or for the detailed evaluation of fibrous structures has not yet been developed. Therefore, it is important to understand the limitations of existing prediction tools. In particular, AF2 bases its predictions on co-evolutionary information rather than the protein's environment, making it more effective for globular proteins. Since AF2 is currently the most powerful algorithm, it would be more efficient to work with globular proteins for AF2-based studies. However, for modelling transmembrane proteins and their complex interactions with the environment, methods that include environmental factors are needed in the future. Modelling the interactions of protein structures with their environment, such as the interaction of surface residues with water molecules, would be more effectively accomplished through physics-based tools like MD simulations, which incorporate environmental factors rather than relying solely on evolutionary information. The specific objectives of each chapter follow.

1.9.1 The impact of recycling on the modelling of quaternary structures of proteins: An evaluation of two AlphaFold2 versions (AF2_Advanced and AF2-Multimer)

In the **second chapter**, the impact of recycling using AF2 based ColabFold versions, AF2_Advanced and AF2M, on structural improvements was investigated. AF2_Advanced, primarily trained on single structures and updated for predicting multimer structures, was compared with AF2M, which was trained exclusively on multimer structures, with respect to their recycling performance. Three standard quality scores were utilized, the global quality score (TM), the local quality score (IDDT), and the interface score (QS-score), to evaluate improvements in performance for two AF2 versions with six different numbers recycles. The

results provided guidance for determining an effective recycling procedure for the multimer modelling tool (MultiFOLD) that we developed prior to CASP15.

1.9.2 The impact of the custom template recycling for the improvement of quaternary structures of proteins

The AF2M versions allows for custom templates and MSA inputs, offering a pathway for further enhancing structural quality. Therefore, in the **third chapter**, the impact of AF2M's further recycling combined with the custom template option on structural improvement was investigated. To assess the performance improvement, the CASP14 and CASP15 datasets were utilized separately, and the quality of protein models generated by AF2M with both MSA and SS methods was evaluated using five scores: the TM-score, the IDDT score, the protein interface scores (DockQ-wave and QS-score), and the MolProbity score. Recycling assessments were carried out using 1, 3, 6, and 12 cycles. The custom templates used were the predicted structures of the top five groups in the CASP 14 competition and four groups in the CASP15 competition with the highest Z-scores, along with those from the NBIS-AF2-Multimer and MultiFOLD groups in CASP15. Using custom templates with recycling is now a crucial part of the pipelines for future versions of MultiFOLD as well as our manual modelling protocols.

1.9.3 The impact of varying custom input options on models generated by AF2M

In our previous studies and published articles, the impact of externally provided structural information on AF2M performance was demonstrated. However, many aspects of AF2 remain unknown. Among the debated topics is what kind of information AF2M internally learns during its training. One of AF2M's successes in protein modelling is attributed to its transformer-like network. The attention mechanism in this network prioritizes relevant information, facilitating its flow through the Evoformer module. Considering that externally provided information passes through this attention mechanism, there are limited studies on how various input data affect protein models generated by AF2M.

Hence, in the **fourth chapter**, investigations were conducted to assess the impact of two types of input modifications on AF2M, particularly examining whether these changes improve the modelled protein structures. Initially, alterations were made to the custom templates of AF2M,

and the effect of providing a custom multimer structure to AF2M as if it were a single chain structure was examined, to investigate whether preserving the relative orientation of structures led to further improvements. This approach was chosen since AF2M processes multimeric template structures one chain at a time, rather than as a complex structure. Therefore, providing complex template structures as single chains could improve the quality of the models generated by AF2M.

Subsequently, changes were made to the AF2M's custom MSA input method to obtain higher-quality protein structures. This was achieved by removing residues corresponding to disordered structures within the MSA, and the impact on the resulting structure models was investigated. This approach was chosen since AF2M's algorithm relies on MSAs with different weights assigned by the attention mechanisms (Jumper et al., 2021a; Skolnick et al., 2021). In addition, AF2M was trained with protein structures in the PDB. Considering that AF2M infers structure from residue co-evolution and that missing residues in the PDB are often treated as disordered regions, focusing solely on MSAs derived from homologous sequences with regular residues can potentially enhance AF2M's performance. The effect of modifications in both custom methods on structural improvement was evaluated by comparing the resulting quality scores.

1.9.4 Performance comparison of MultiFOLD1 and MultiFOLD2 using data from the CAMEO-BETA Project

In the recent CASP15 competition, our newly designed protein structure complex modelling tool, MultiFOLD1, achieved a remarkable ranking. The performance of MultiFOLD1 was also tested weekly using the CAMEO-BETA project. Determining the optimal parameters for the AF2M components of MultiFOLD was essential to enhance the quality of its model outputs. Following the CASP15 competition, MultiFOLD1 was upgraded by combining the dropout method of AF2M with RoseTTAFold2 and RoseTTAFold All-Atom, resulting in the development of MultiFOLD2. Like MultiFOLD1, the new MultiFOLD2 server was also tested using the CAMEO-BETA project. The underlying method involved first creating a structure pool and then selecting the best structure via a quality estimation tool. The **fifth chapter** delves into the performance comparison between MultiFOLD1 and MultiFOLD2, as well as benchmarking the performance of two versions of MultiFOLD compared to other state-of-the-art servers participating in the CAMEO-BETA project.

Chapter 2: The Impact of Recycling on the Modelling of Quaternary Structure of Proteins: An Evaluation of Two AlphaFold2 Versions (AF2_Advanced and AF2-Multimer)

2.1 Background

The majority of proteins in a cell interact to form complexes to perform their functions. By having various interaction partners, proteins regulate their functions in a cell according to changes in the surrounding environment (Swamy et al., 2021). Hence, the prediction of quaternary structures of proteins or complexes is beneficial for downstream analysis including drug design, protein engineering, and function analysis (Quadir et al., 2021).

Historically, modelling protein quaternary structure was solely based on template based and docking approaches. Template based approaches proved to be more effective when a suitable quaternary structure template was available in the PDB, while docking approaches were more accurate in the absence of template and when the corresponding monomer structures have high quality. However, it is a challenge to find suitable templates for multimeric structures rather than monomer structures as there are fewer of them in the PDB (Kozakov et al., 2017; Liu et al., 2023a). Additionally, the initial quaternary structure models produced by many protein-protein docking processes have been formed using rigid-body approaches. However, rigid-body docking does not account for minor conformational changes which may occur when bonding occurs, notably at the interface (Baek et al., 2017). Thus, refining the complex structure with rigid docked approaches is an essential step if the complex target is to be used for downstream application, such as the determination of 'hotspot' residues for drug design in the context of further and more precise protein-protein interaction or structure-function research (Verburgt & Kihara, 2022). With the release of the first version AlphaFold (AF), in 2020, and the subsequent release of AF2 in 2022, AI-based algorithms have become predominant in the field of protein modelling and refinement. AF2 is a DNN-based protein modelling tool where evolutionary and physical information were employed in novel training procedures, resulting in the generation of high-quality protein structures (Jumper et al., 2021a). However, at the present time, following the release of AF2 (detailed in Chapter 1), refining complex structures via deep learning-based methods has become more popular than physical MD based approaches, which are more time consuming.

After Google DeepMind, the developers of the AF2 method, released AF2 to model tertiary structures, Yoshitaka Moriwaki (Moriwaki, 2021) demonstrated, through his own Twitter page (the old version of X platform) on 19.07.2021, that AF2 can be modelled for complex protein structures by adding a linker between two chains. Furthermore, Baek used AF2 by balancing residue index in order to model complex protein structures (Baek, 2021). In the pursuit of these successful steps, AF2_Advanced was designed by the ColabFold team for modelling protein complexes (both homo- and hetero-oligomers) by adding such functionality as well as running

MMseqs2 (Steinegger & Söding, 2018) for generating MSAs (Mirdita et al., 2022), which increased the accessibility of the complex modelling method. The rationale behind this was to insert a 21 residue GGS linker, also called the AlphaFold-Linker between the protein chains. Following the last released version of AF2_Advanced, DeepMind developed AF2M (Evans et al., 2022) by retraining the method on complexes and releasing a new set of weights better adapted for modelling complex structures.

In the new era, it remains uncertain about which type of traditional refinement process is most beneficial to models generated by AI-based methods. Since the AI-based models continue to exhibit shortcomings and there were significant energy barriers in refinement pathways after complex modelling, consideration of inter-domain and oligomeric interactions has become increasingly crucial (Heo et al., 2021). In particular, Heo *et al*. (2021) found that using MD simulation for refinement of AF2 models was not able to improve the initial AF2 models and even decreased their quality. However, another ML based models were improved by physical simulations. Hence, it can be crucial to evolve different approaches regarding the refinement process for initial structures obtained via AI-based methods, especially AF2M structures. The DeepMind group adopted a different approach to “refinement” by evolving to feed the loss function of the last generated model back into the algorithm, which is called ‘recycling’ as mentioned in Chapter 1. In addition, AF2M was used as a refinement method for structures generated by ClusPro based on physical approaches and was observed the improvement in ClusPro structures (Ghani et al., 2021). When DeepMind first released their codes to the community of structural bioinformatics, its algorithm became a research focus for almost all groups. However, there was little information on several stages of the prediction algorithm, notably a key process called “recycling”. In the code, there is an optional default value of 3 as the maximum number of recycles for both monomer and oligomer targets.

2.1.1 The aim of study

Using AF2, it has been verified that ML methods outperforms traditional docking methods based on physical laws and knowledge-based potentials. (Evans et al., 2022). The AF2 method predicted the tertiary protein structures very well in the CASP14 experiment, and 3 recycles were used by default, but it was not used in that experiment to model quaternary structures. Subsequently, this value was also used as the default for modelling complexes with the AF2 versions, however, there was no detailed research on the effect of changing the recycling process for models of quaternary structures. Hence, we hypothesise that the further recycling stage can be used as a standalone refinement process. Also, in the interest of optimizing both the use of computing resources and maximizing model quality, it is worthwhile to investigate the recycling process and determine the optimum number of cycles for different targets. Thus, this study aims to investigate the recycling progress of the AF2 algorithm during the refinement stage and to determine whether altering the number of recycles affects the final quality of modelled protein complexes. Subsequently, it aims to identify the optimal number of effective cycles, comparing them with the default value of 3, which is used by both AF2M and AF2_Advanced.

2.2 Method

2.2.1 Data collection

The sequences of targets for the last CASP competition at the time of the study (CASP14) were downloaded from the CASP website (http://prediction.org/download_area/_) and were used in this analysis to benchmark and compare the recycling procedure for both AF2_Advanced and AF2M versions. At the time of conducting the analyses in this chapter, CASP14 was the last CASP competition, but the findings were used to inform our strategy for CASP15, the results of which are presented in subsequent chapters. Models were generated for 10 homo-oligomers and three hetero-oligomers using two AF2 versions with varying numbers of recycles. These target proteins names is follow: 1-) PEX4-PEX22 (Organism: Arabidopsis thaliana), 2-) N4-Cytosine Methyltransferase (Organism: Serratia marcescens), 3-) Testis-expressed protein 12 (Organism: Homo sapiens), 4-) Structural maintenance of chromosomes flexible hinge domain containing 1 (Organism: Homo sapiens), 5-) Inhibitor of the Yeast Formin Bnr1 (Organism: Saccharomyces cerevisiae), 6-) Tomato Spotted Wilt Virus (TSWV) glycoprotein (Organism: Semliki Forest virus), 7-) BonA (Organism: Acinetobacter baumannii), 8-) Tailspike protein (Organism: Escherichia virus CBA120) 9-) Hypothetical protein predicted by Glimmer/Critica (Organism: Bdellovibrio bacteriovorus), 10-) a small

secreted cysteine-rich protein (Tsp1) (Organism: Trichoderma virens) 11-) Nitro-histidine zipper coiled coils (Organism: Nitrosococcus oceani), 12-) Meio-histidine zipper coiled coils (Organism: Meiothermus silvanus), 13-) Tuna-histidine zipper coiled coils (Organism: Methylobacter tundripaludum). However, in the CASP competition, the organisator use the specific CASP code for each protein target sequence which the participants handle with. In the code, the prefix of "H" represents heteromeric protein targets, while the prefix of "T" represents homomeric protein targets. The CASP14 codes of above protein targets to use are follow: H1045, H1065, H1072, T1032, T1034, T1038, T1054, T1070, T1073, T1078, T1083, T1084, and T1087, respectively. However, a common subset of targets was used in the analysis of quality scores for both AF2 versions.

The target models were generated using the AF2_Advanced (listed as AlphaFold2_Advanced_v2 on ColabFold) and AF2M (listed as AlphaFold2_mmseqs2 on ColabFold) forks of AF2 that are available via the ColabFold Google Colab Notebooks (https://github.com/sokrypton/ColabFold). For each AF2 version, and for every cycle number, five models were generated for every given CASP target. In the process of getting five models for both AF2 versions, the optional stage "refine structures with Amber-relax" was eliminated to control for refinements occurring purely due to the recycling process. (Incidentally, Amber was originally used by the DeepMind team to achieve a marginally better 3D structure, and it seemed to have no significant effect on quality scores (Jumper et al., 2021a)). Therefore, the outputs resulted in five "unrelaxed" models of quaternary structures.

2.2.2 Observed model quality scores

2.2.2.1 TM-score

The TM-score is an evaluation tool developed by Zhang and Skolnick (2004) to measure the observed topological similarity between predicted and experimental structures using structural alignment to superpose coordinate data. The TM-score is designed with an extension of the approaches used in two other global score measures- the Global Distance Score (Zemla et al., 1999) and MaxSub (Siew et al., 2000), however the TM-score is arguably more robust measure for assessing the global score (Zhang & Skolnick, 2004). AF2 tools generated a predicted TM-score (pTM-score) as an output score for ranking modelled complexes. Therefore, in this study, the observed TM-score of all the targets was analysed as the principle evaluated quality score. The observed TM-scores were generated using the MM-Align (Mukherjee & Zhang, 2009) Tool (for protein complex structural alignment) from the Zhang Group by superimposing the predicted models with the experimentally determined quaternary structures.

2.2.2.2 lDDT score

The global superposition-based quality measurement score is limited if the given target is flexible with multiple domains that can change their relative position with respect to each other. In this regard, the superposition-free measurements such as lDDT score are suitable for considering small domain regions rather than large domain regions and for calculating the amount of corresponding reference protein structure within the generated protein structure (Mariani et al., 2013). AF2 versions also generate a predicted lDDT score (plDDT score) as the output of the quality estimate score, so, in this research, the observed lDDT score was the second score to be analysed. The lDDT score was generated using the OpenStructure (OST) package (version 2.3) downloaded from https://www.openstructure.org/download/ and installed locally. It was run in the Ubuntu terminal with the guidance on the OST "actions" from the online documentation (https://www.openstructure.org/docs/2.3/actions). The measured lDDT value from the OST actions used in this chapter was the oligo-lDDT score.

2.2.2.3 QS-score

The QS-score is capable of considering the junctional interface as a whole and it is useful for comparing both homo- and hetero-complexes with different kinds of stoichiometry, various relative chain orientations, and various amino acid sequences. The QS-score is superior to alternative interface quality measurement scores like interface-RMSD (I-RMSD), which provide scores for assessing dimeric interfaces (Bertoni et al., 2017). The QS-score allows us to consider cases where the protein assembly is not only binary, thus it is more useful to evaluate the complete complex for protein targets without dissociating them into binary interactions. Therefore, the QS-score was also used as the observed interface quality score in our study. The method for generating the observed QS-scores is using the same OST tool as for the lDDT score described above.

2.2.2.4 DockQ_wave score

DockQ is a metric used to evaluate the quality of protein-protein docking models. It integrates various docking quality measures, including Fnat (fraction of native contacts), iRMSD (interface root mean square deviation), and LRMSD (ligand root mean square deviation). The DockQ score ranges from 0 to 1, with higher scores indicating better docking models. Recent updates to DockQ have extended its functionality beyond protein-protein interactions to include nucleic acids and small molecules. It also features advanced options like automatic chain mapping and support for multimeric complexes with complex stoichiometries, making it a

versatile tool for analyzing molecular assemblies. However, one limitation of the DockQ score is that it evaluates protein-protein interfaces separately within a structure. For large protein complexes, this may reduce reliability as it does not simultaneously consider all interface regions. To address this, the DockQ_wave score was introduced, which calculates DockQ scores for each protein interface and computes a weighted average. The interface weights are determined by the number of native contacts. In contrast, QS-score places importance on symmetry in the interface contacts, which can pose challenges in unresolved structures. Like DockQ, DockQ_wave scores also range from 0 to 1, with higher values reflecting better models (Studer et al., 2023).

2.2.2.5 Molprobity score

MolProbity is a model validation score for protein structures that does not requires native structures, retrieved at http://molprobity.biochem.duke.edu. The MolProbity score merges the combination of the rotamer, clashscore, and Ramachandran scores into a single score. By integrating a number of stereochemical and geometric factors, it offers a comprehensive evaluation of structural validation. Models with lower Molprobity scores are more stereochemically correct than those with higher values, ranging from 0 to 1 (Chen et al., 2010(a)).

Both DockQ_wave and Molprobity score did not use for this chapter. DockQ_wave was released after CASP14 competition. Since we did not use the Amber relaxation method in AF2, the structures were not evaluated in terms of MolProbity score.

2.2.3 Experimental design

The first analysis aimed to determine the baseline recycling process for both AF2 versions. The second analysis aimed to determine the optimal number of cycles required to generate the best predicted models with the highest observed model quality scores for each AF2 version. On the Google Notebook platform, five models for every CASP target were generated by entering the following values:

Within both AF2 tools, the sequences are entered by putting ':' between every chain sequence.

The AF2M (v.1.2) code was executed by selecting the following parameters:

msa_mode: MMseqs2 UniRef+Environmental, model-type: auto, pair_mode: unparied + paired, num_recycles: 1, 3, 6, 12, 24 and 48 in series (The early versions of AF2 employed the 'auto' model selection, which helped avoid biased predictions for the CASP14 targets. Selecting 'auto' model type provided models generated by the AF2 with v1 model).

The AF2_Advanced was run by selecting the following parameters:

homooligomer: the number of chain, msa_method: MMseq2, msa_format: fas, pair_mode: unpaired, pair_cov: 50, pair_qid: 20, rank_by: pTM, use_turbo: true, max_msa: 512:1024 , num_models: 5, use_ptm: true, num_ensemble: 1, max_recycles: 1, 3, 6, 12, 24 and 48 in series, tol: 0, num_sample: 1(MMseqs2 was preferred over JackHMMER for homology search and MSA design).

Both versions of AF2 was run separately for each recycling value.

The MM-Align and OpenStructure programs were used to obtain observed scores for the TM-score and the QS-score and IDDT score respectively by comparing the generated complex structures with the native complex structures. These quality scores were then used to evaluate the performance of the recycling process. Based on C-alpha atoms superposition with native structure, the TM-score was employed for analysing the global quality of the predicted complexes, which is the best score for measuring the overall quality of the modelled structures, including residues outside of the interface, as well as measuring the accuracy of the relative orientations of the interacting subunits. The IDDT score was used to analyse the overall quality of the given protein structures at a more local level, as it was based on residue accuracy in each chain, and so less dependent on the relative orientation of the interacting subunits.

For measuring the accuracy of the modelled interfaces, the QS-score was used to simultaneously score interfaces between all interacting subunits. (Note: the DockQ-score (Basu & Wallner, 2016) is another popular score for analysing interface area quality. Unlike the QS-score, at the time of the analysis this score evaluated only binary interface areas). At the same time, the predicted model quality score value (pTM-score for both versions) was used to determine the ranking of models from 1 to 5, according to each cycle number. For ranking of the modelled complexes, it was advised that the global pTM-score should be used instead of pIDDT score, which is better for reranking tertiary structures (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb). Specifically, in this study, the analysis focus was on the 1st ranked models for each cycle number, in order to determine the optimum number of cycles to produce the best observed scores.

2.2.4 Statistical Analysis

After evaluating the improvements in cumulative scores, our next aim was to determine whether further AF2 recycling is an effective method for achieving better protein quality scores. To compare the performance of further recycling, a non-parametric statistical test (the paired Wilcoxon signed-rank test) was conducted on paired datasets, since the data did not follow a normal distribution. The methodologies described in AlphaFold-related publications were followed during the statistical analysis. In this analysis, TM-scores, IDDT, and QS-scores obtained using two different options (representing two recycling values) for the same protein targets were compared to assess whether there was a statistically significant improvement in quality scores with increased recycling for each target. For statistical evaluation, we adopted the single type of statistical analysis method mentioned above, based on the approach used in last AlphaFold studies (Abramson et al., 2024). Figure 2.1 summarizes the workflow of methods used in this chapter`s analysis.

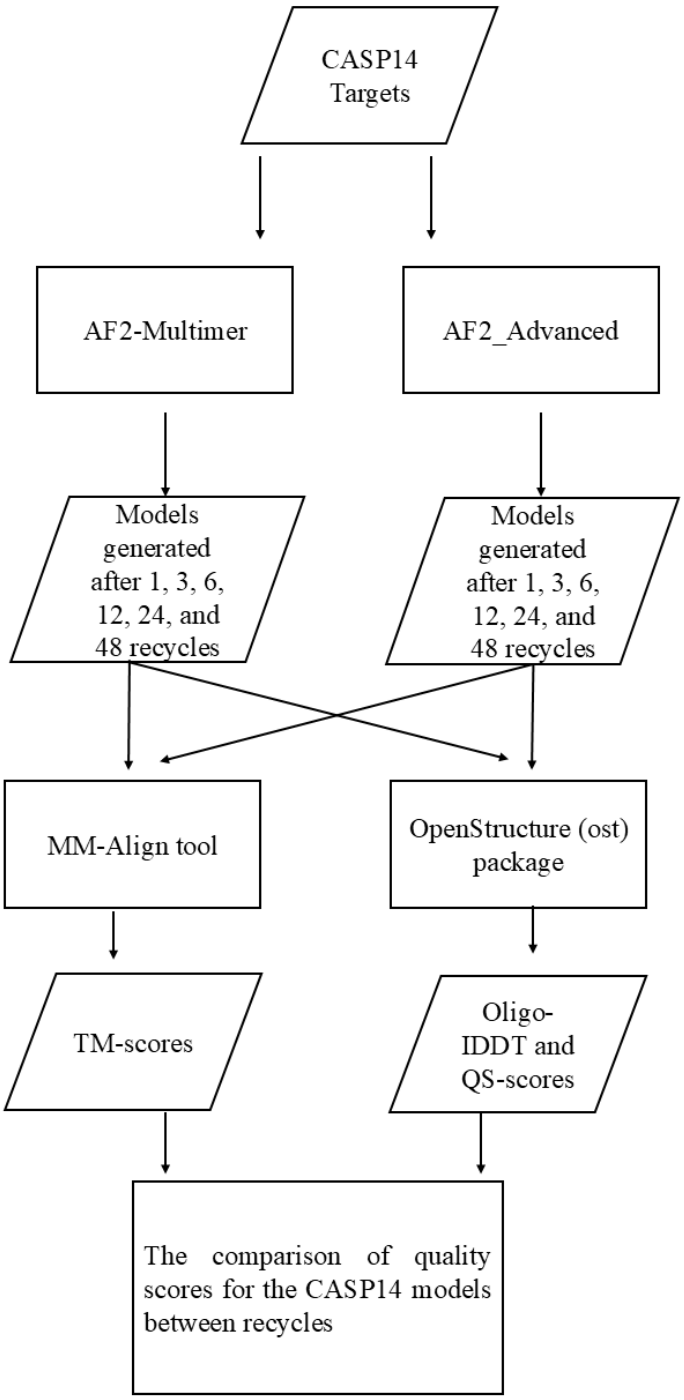


Figure 2.1 The flowchart of the method for determining the optimal recycling parameters from modelling quaternary structures.

Flowchart showing the process for determining optimal number of cycles in AF2_Advanced and AF2M using MM-Align tool and OpenStructure package as well as the comparison of both versions. The observed quality scores, with TM-score from MM-Align and IDDT/QS-score from OpenStructure, were produced by aligning the models with the native structures for each target. Subsequently, the observed quality scores for the rank-1 models in each cycle round were statistically evaluated using the paired Wilcoxon-signed-rank test. Since the high computational load is required for quaternary structures, each model is run once for each recycle model (n=1).

2.3 Results and Discussion

In this study, a total of 13 CASP14 multimeric targets for calculating the observed TM-score and IDDT score, and QS-score, were analysed by separately selecting models obtained from six different cycle numbers from both AF2_Advanced and AF2M. Several CASP targets were not used for this research as there were either too many residues and/or too many chains involved in the complex (Google Colab limits the GPU RAM (16 GB) available on each node for each user). In addition, not all CASP targets had experimentally derived structures of the complexes so some models could not be evaluated against the native target structures.

Firstly, correlation analysis was carried out to assess the relation between the predicted quality scores generated by the AF2 versions (pIDDT score and pTM-score) and the observed scores (IDDT score and TM-score), aiming to determine the reliability of these prediction scores. For the correlation test of the both AF2 versions, only CASP14 targets were selected since at the time of this research, CASP 14 was the last CASP to be held. Since the results of the correlation tests indicate a highly correlation between both prediction and observed scores, highlighting the reliability of both pIDDT and pTM scores. This suggest that the highest predicted scores can correspond to the highest observed scores.

From the observed quality scores of the models which were subject to the recycling algorithm, it can be observed that the recycling procedure using AF2M often resulted in an improvement of the modelled complexes for the protein target (e.g., for T1083 see Figure 2.2). This figure illustrates the increase in three quality scores (TM-score, IDDT, and QS-score) for T1083 up to further recycling rounds. In the same figure, structural changes can be seen by superimposing the initial, refined, and reference models. Furthermore, improvements in model quality obtained from the recycling procedure can be detected using the pIDDT and pTM-score via their correlation with the observed quality scores (IDDT and TM-score) (Figures 2.3-2.6).

The main principle of AF2 lies in integrating co-evolutionary information with novel machine learning techniques. However, it is well-known that proteins perform their functions in nature through different conformational states. AF2, on the other hand, predicts only a single static structure of a protein. The key detail here is the use of different random seeds. Protein prediction tools like AF2, which rely on DNN, may obtain different conformational states when it is run with different seeds. However, the early versions of AF2 did not incorporate random seed selection, and the same seed was used in every run. One significant limitation of these early versions was the need to rerun AF2 for each recycling value (n=1), which allowed for the possibility of using different seeds. To counteract this limitation, AF2 implemented a

mechanism where multiple models were generated during each run. These models were then ranked based on their quality scores (pIDDT and pTM-score), with the best structure being selected as rank-1, even if different seeds are run.

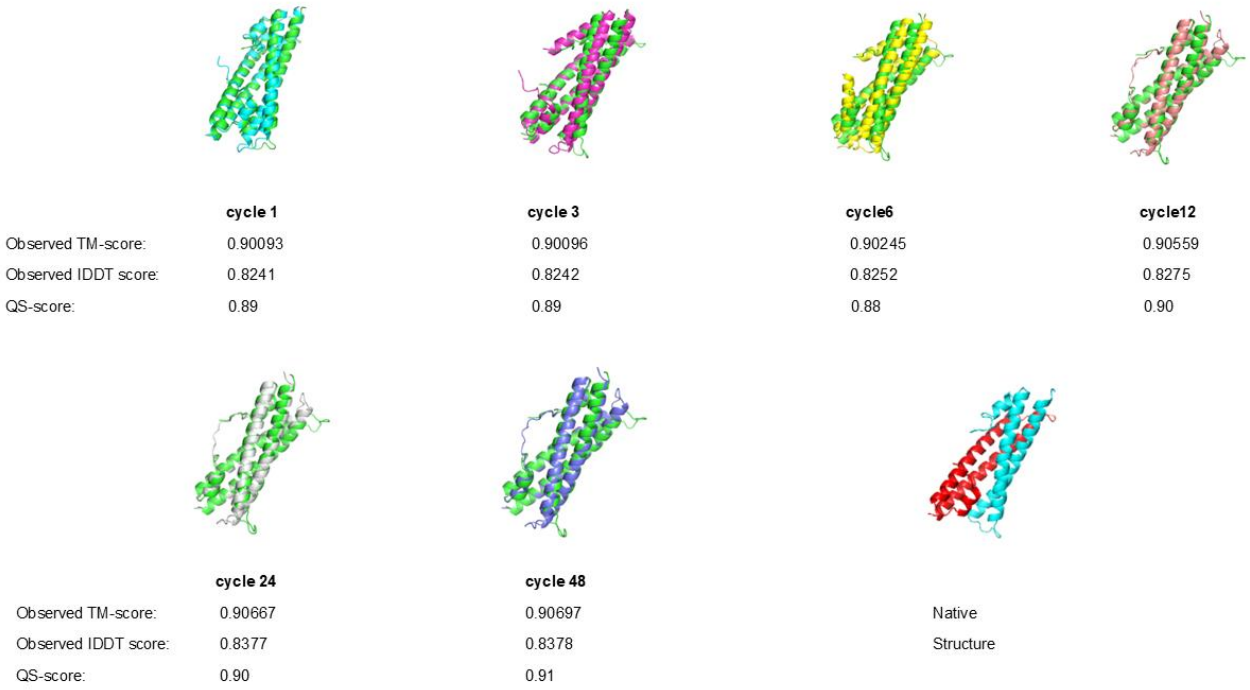


Figure 2.2 The impact of further recycling on quality scores for T1083 (CASP14 target).

A comparison of the native structure with the modelled complexes for target T1083 (Dimer) with cycle 1, 3, 6, 12, 24, and 48 cycles obtained from AF2M. It was rerun for each recycling value (n=1). Each image was generated using PyMOL (https://www.pymol.org). In this example, the score indicates that AF2M has shown an increase in model quality up to cycle 48, except QS-score for cycle 6 was lower than previous cycle (cycle 3). The different chains in the models and native structure are colour coded. In the structural superposition image, the native structure is in cyan.

Figures 2.3 and 2.4 show a positive correlation between the observed and the predicted quality scores. The Pearson's R, Kendall's tau B, and Spearman's rho correlation were used to examine the degree of the relationship between the observed and predicted quality scores of 78 models obtained from 13 targets (CASP14) generated by AF2M. Correlation analysis between the predicted quality scores (pTM-score) and the observed TM-scores shows a significant ($p < 0.05$) high positive linear correlation for the modelled complexes generated by AF2M with Pearson's R = 0.80, Spearman's Rho = 0.76, and Kendall's tau B = 0.60. The high positive linear correlation indicates that the increases in the global pTM-score generated by AF2M will correlate with an increase in the observed quality scores (TM-score). Furthermore, correlation analysis between the pIDDT score and the observed IDDT scores also shows a significant positive correlation ($p < 0.05$). However, this is a stronger correlation than the correlation between pTM-score and observed TM-score, with Pearson's R = 0.89, Spearman's Rho = 0.78, and Kendall's tau B = 0.62. Hence this signifies that the increase in pIDDT scores generated by AF2M is more strongly correlated with an increase in the observed IDDT quality scores.

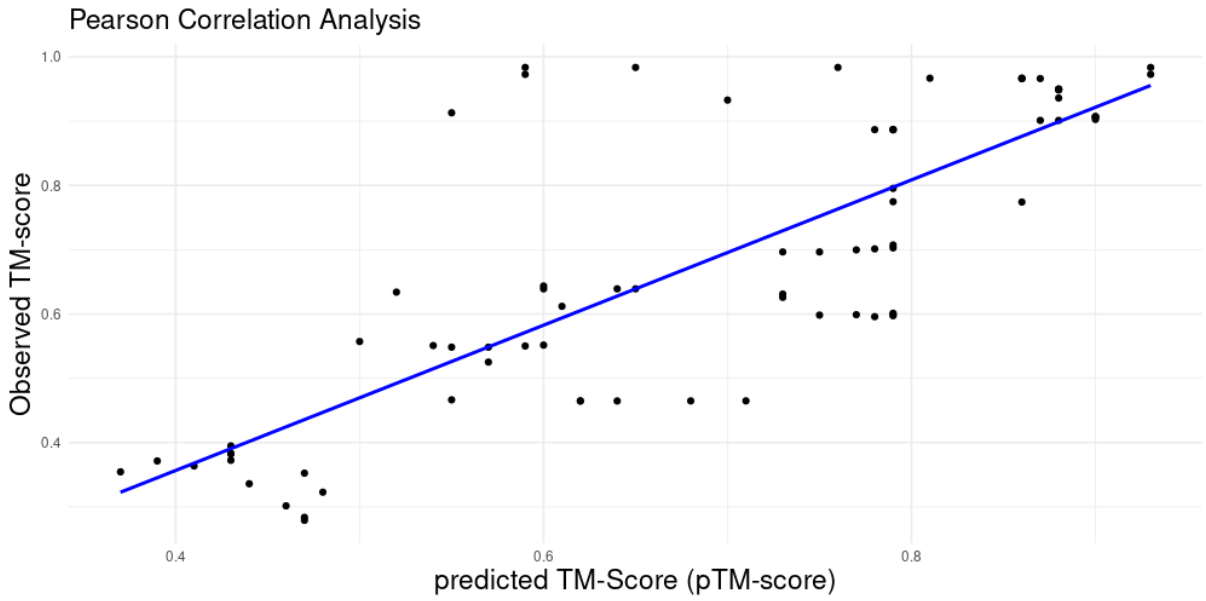


Figure 2.3 The correlation plot between the observed TM-scores and the predicted TM-scores for AF2M.
Scatter plot is showing linear, positive relationship between the predicted global scores (assessed by the pTM-score) (x-axis) versus the observed TM-scores (y-axis) of the models of CASP14 targets generated using AF2M with n = 13 targets (13 models from each of the 6 different numbers of recycles -1, 3, 6, 12, 24, and 48 cycles- for each target = 78 models). The Pearson's R correlation is 0.80, Spearman's Rho correlation is 0.76, and Kendall's tau B correlation is 0.60. P-values for three correlation tests are less than 0.05.

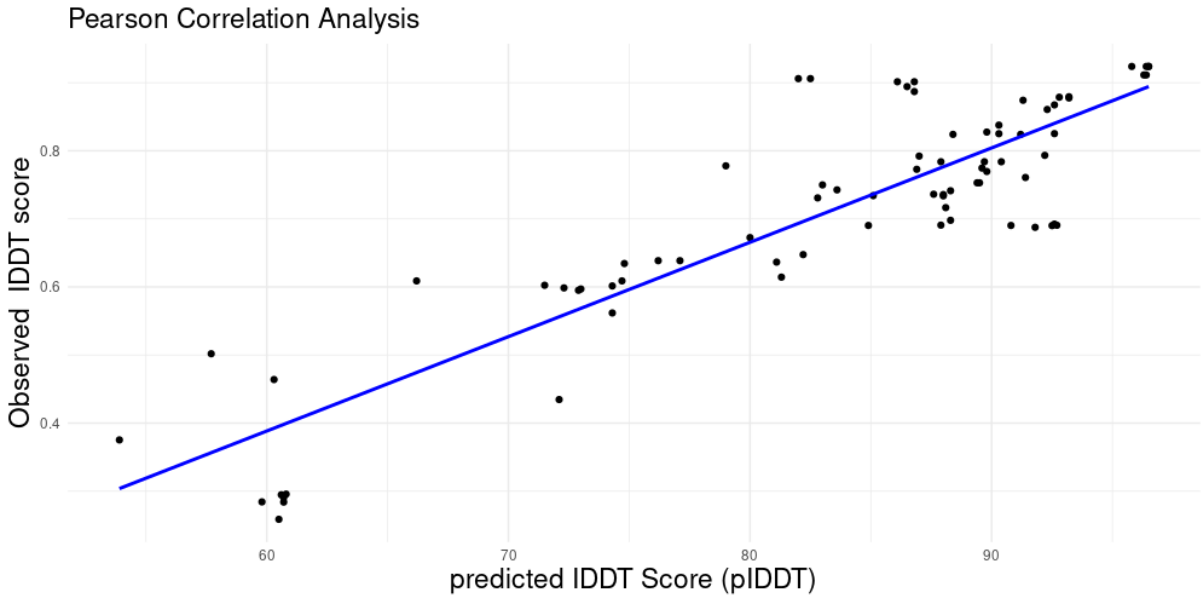


Figure 2.4 The correlation plot between the observed lDDT scores and the predicted lDDT scores for AF2M.
Scatter plot is showing linear, positive relationship between the predicted global scores (assessed by pIDDT score) (x-axis) versus the observed lDDT scores (y-axis) of the models of CASP14 targets generated using AF2M with n = 13 targets (13 models from each of the 6 different numbers of recycles -1, 3, 6, 12, 24, and 48 cycles- for each target = 78 models). The Pearson's R correlation is 0.89, Spearman's Rho correlation is 0.78, and Kendall's tau B correlation is 0.62. P-values for three correlation tests are less than 0.05.

Figures 2.5 and 2.6 show a positive correlation between the observed and the predicted quality scores. This time the Pearson's R, Kendall's tau B and Spearman's Rho correlation were used to examine the degree of relationship between the observed and the predicted quality scores of 78 models which obtained from 13 targets (CASP14) generated by AF2_Advanced. Correlation analysis between the pTM-score and the observed TM-scores shows a significant ($p < 0.05$) linear positive correlation for the modelled complexes generated by AF2_Advanced with Pearson's R = 0.83, Spearman's Rho = 0.84, and Kendall's tau B = 0.65. Again, here the positive linear correlation indicates that the increase in the global pTM-scores from AF2_Advanced correlates with an increase in the observed quality scores (TM-score). However, a weak positive ($p < 0.05$) correlation for complex protein models can be observed between the observed lDDT scores and the pIDDT scores with Pearson's R = 0.39, Spearman's Rho = 0.39, and Kendall's tau B = 0.25. So, these data signify that the increase in the pIDDT scores from AF2_Advanced is more weakly correlated with an increase in the observed lDDT quality scores.

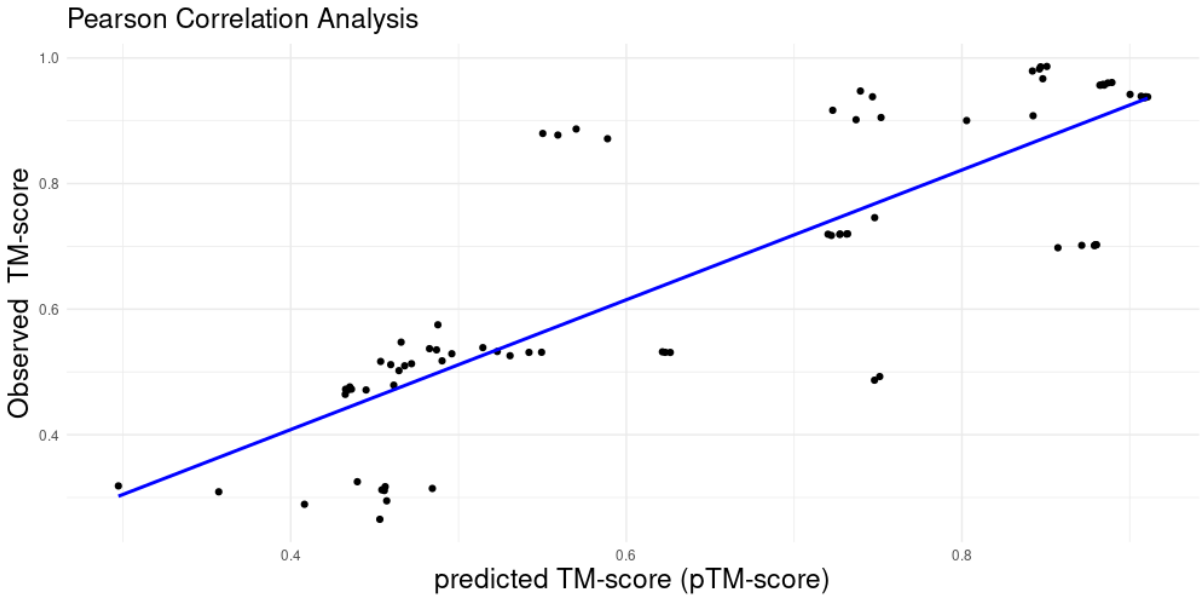


Figure 2.5 The correlation plot between the observed TM-scores and the predicted TM-scores for AF2_Advanced.

Scatter plot is showing linear, positive relationship between the predicted global scores (assessed by pTM-score) (x-axis) versus the observed TM-scores (y-axis) of the models of CASP14 targets generated using AF2_Advanced with n = 13 targets (13 models from each of the 6 different numbers of recycles - 1, 3, 6, 12, 24, and 48 cycles- for each target = 78 models). The Pearson's R correlation is 0.83, Spearman's Rho correlation is 0.84, Kendall's tau B correlation is 0.65. P-values for three correlation tests are less than 0.05.

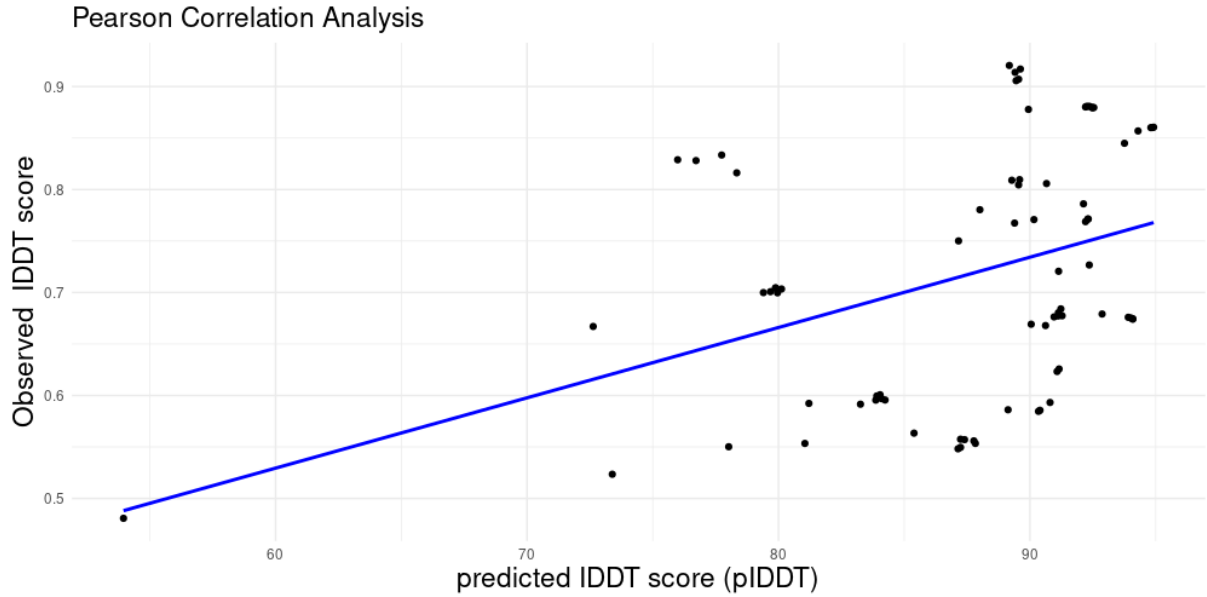


Figure 2.6 The correlation plot between the observed IDDT scores and the predicted IDDT score for AF2_Advanced.

Scatter plot is showing linear, positive relationship between the predicted global scores (assessed by pIDDT) (x-axis) versus the observed IDDT scores (y-axis) of the models of CASP14 targets generated using AF2_Advanced with n = 13 targets (13 models from each of the 6 different numbers of recycles - 1, 3, 6, 12, 24, and 48 cycles- for each target = 78 models). The Pearson's R correlation is 0.39, Spearman's Rho correlation is 0.39, and Kendall's tau B correlation is 0.25. P-values for three correlation tests are less than 0.05.

Considering the correlation test results and the appropriateness of prediction scores, it would be appropriate to consider the top ranked models generated by the different versions of AF2 in order to observe whether there is an improvement in the predicted structures with increased cycle. Evaluating the scores cumulatively shows the overall improvement of existing high-quality scores in structures across all targets evaluated. Observing the improvements obtained for each target cumulatively provides more comprehensive detectable changes, considering that the individual improvement rates for individual targets are expected to be relatively small. Therefore, the cumulative observed quality scores of the top ranked models obtained after different numbers of cycles were compared with AF2's default cycle values.

With the advent of AF2, the prediction of protein structures, including complex structures, has improved, reducing reliance on refinement methods that were essential prior to AF2. However, for downstream analyses such as drug design, predicting binding site regions often requires additional post-modelling procedures. According to the refinement theorem, well-modelled structures are more likely to deviate from the native structure when subjected to these methods (Adiyaman & McGuffin, 2019). Therefore, even minor improvements in cumulative quality scores, which are used to compare modelling tools, hold importance. Since even small increases in cumulative scores are crucial in protein structure modelling, statistical testing can introduce limitations. However, to assess the significance of the differences between scores, we followed the approach used in AlphaFold-related studies (Abramson et al., 2024) , applying only the paired Wilcoxon signed-rank test. The “paired” means before and after scores for model generated by AF2 for each target. Our secondary objective was to compare the individual paired scores across different recycle statistically after the cumulative scores was evaluated. AF2 unexpectedly generates a new model instead of refining an existing one for a given target (Adiyaman et al., 2023), leading to highly variable recycle values for certain targets. Consequently, a recycle value associated with the highest cumulative score may not necessarily yield the most statistically significant result, as in Figure 2.7 in Page 54. The null (H_0) and alternative (H_1) hypotheses are detailed in the legend of Table 2.1 (See Page 63).

Figure 2.7 shows the cumulative observed TM-scores as the cycle number is increased for AF2M. Based on the cumulative TM-scores in rank-1, the different cycles can be ranked as follows: cycle 3 (9.22) > cycle 48 (9.18) > cycle 24 (8.95) > cycle 12 (8.92) > cycle 6 (8.84) > cycle 1 (8.41). The cumulative TM-scores indicate that the general structural quality of the modelled complexes increases as the cycle increases and cumulatively model improvement up to cycle 48 generates a model with quality better than the previous cycles, except for cycle 3. In cycle 3, a model of better quality was generated cumulatively. However, the paired statistical analysis revealed that the cumulative TM-scores between cycles 12 and 24 were significantly greater ($p < 0.05$) than cycle 1 (Table 2.1).

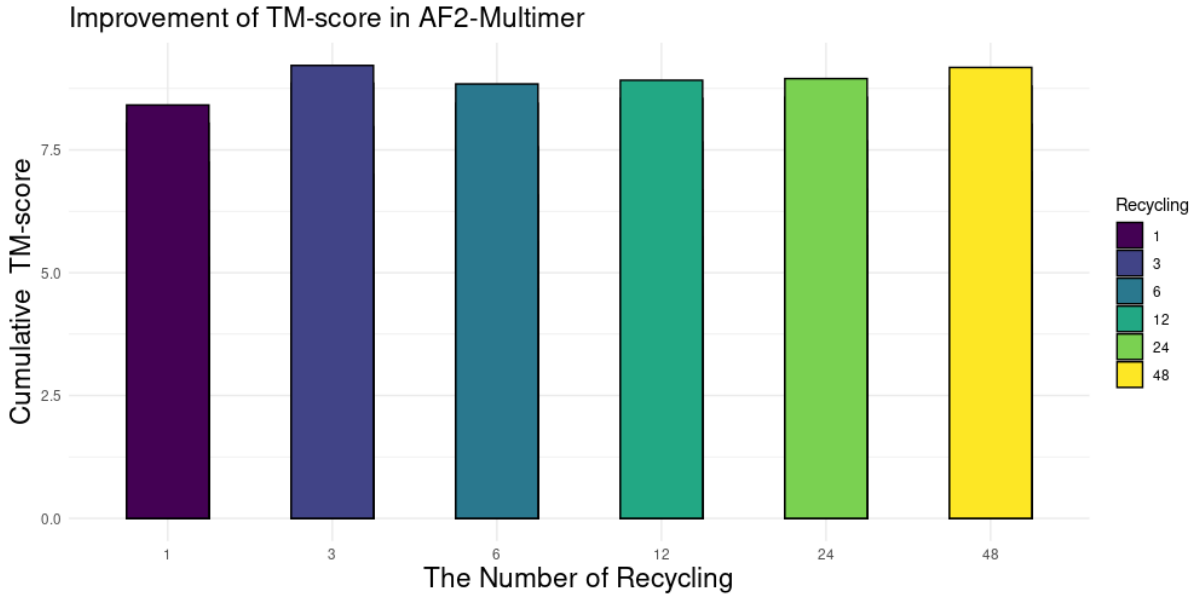


Figure 2.7 The improvement of TM-scores in the CASP14 models for AF2-Multimer (AF2M).
The bar chart displaying the cumulative observed global quality score (assessed by TM-score) of the modelled complex of all targets (CASP14) generated by AF2M, based on rank-1 models. The rank-1 models are from six cycle rounds and each cycle is a colour coded. The rank-1 model corresponds to the model with the best pTM-score. This graphic was drawn using R.

Figure 2.8 shows the cumulative observed IDDT scores as the cycle number is increased for AF2M. Based on the cumulative IDDT scores in rank-1, the different cycles can be ranked as follows: cycle 48 (9.43) > cycle 3 (9.33) > cycle 12 (9.32) > cycle 24 (9.26) > cycle 6 (9.18) > cycle 1 (8.77). The cumulative IDDT scores indicate that the local quality of the modelled complexes increases as the cycle increases and cumulatively model improvement up to cycle 48 generates a model with quality better than cycle 1, similar to TM-score. However, in this time, models generated using cycle 48 had better quality than the 5 other recycling values. The paired statistical analysis revealed that the cumulative IDDT scores up to cycle 12, 24, and 48 were significantly greater ($p < 0.05$) than cycle 1 (Table 2.1).

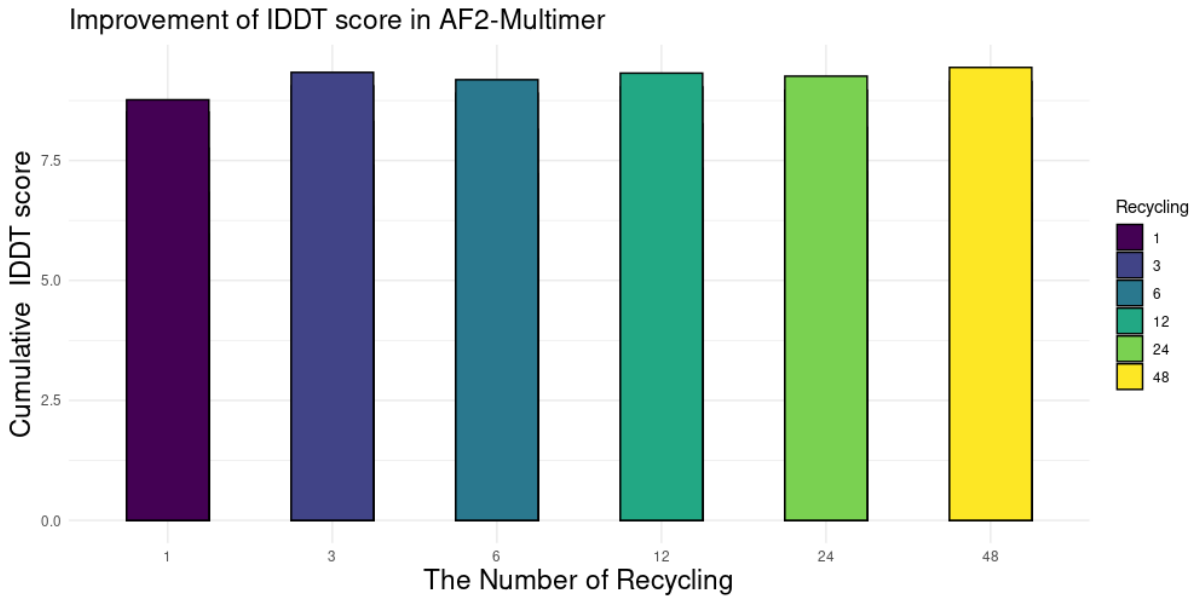


Figure 2.8 The improvement of IDDT scores in the CASP14 models for AF2-Multimer (AF2M).
The bar chart displaying the cumulative observed local quality score (assessed by IDDT score) of the modelled complex of all targets (CASP14) generated by AF2M, based on rank-1 models. The rank-1 models are from six cycle rounds and each cycle is colour coded. The rank-1 model corresponds to the model with the best pTM-score. This graphic was drawn using R.

Figure 2.9 shows the cumulative observed QS-scores as the cycle number is increased for AF2M. Based on the cumulative QS-scores in rank-1, the different cycles can be ranked as follows: cycle 3 (7.05) > cycle 48 (6.98) > cycle 24 (6.53) > cycle 6 (6.48) > cycle 1 (6.44) > cycle 12 (6.16). According to cycle 1, the cumulative QS-scores indicate that the interface quality of the modelled complexes increases as the cycle progresses and cumulatively model improvement up to cycle 6 generates a model with better quality. However, after that, the cumulative score decreased up to cycle 24. Interestingly, with 24 cycles, better quality models were generated again. Cycle 3 yielded the highest cumulative score while cycle 48 was second recycle value that exhibited a high cumulative QS-score. The statistical analysis revealed that the cumulative QS-score for five different cycles (3-6-12-24-48) was not significantly greater ($p > 0.05$) than cycle 1 (Table 2.1).

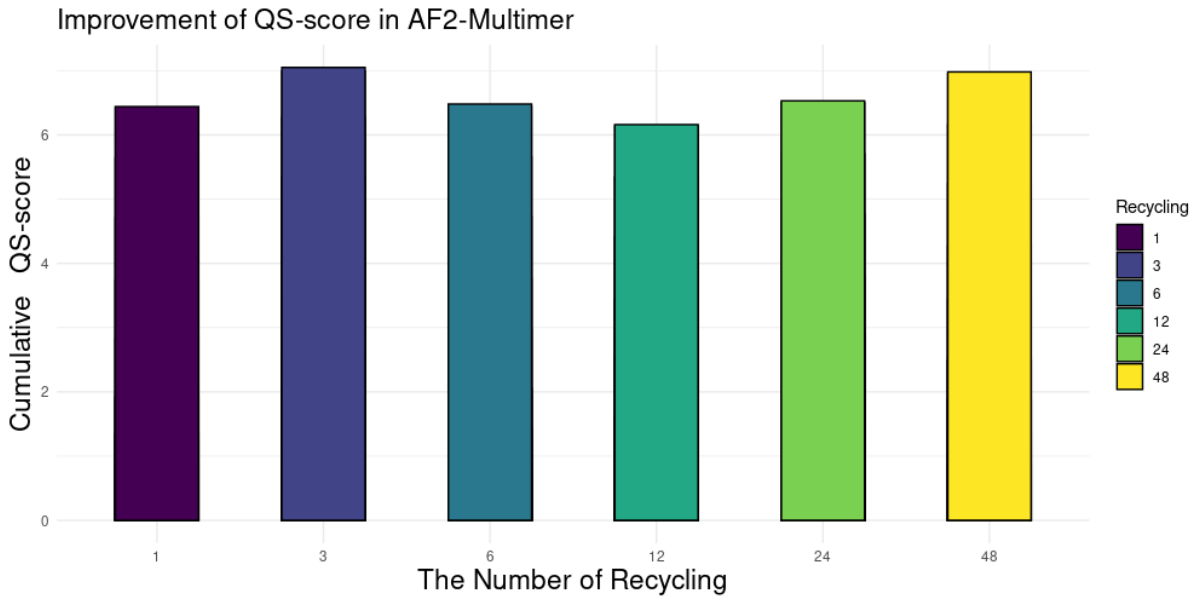


Figure 2.9 The improvement of QS-scores in the CASP14 models for AF2-Multimer (AF2M).
The bar chart displaying the cumulative observed interface quality score (assessed by QS-score) of the modelled complex of all targets (CASP14) generated by AF2M, based on rank-1 models. The rank-1 models are from six cycle rounds and each cycle is colour coded. The rank-1 model corresponds to the model with the best pTM-score. This graphic was drawn using R.

The cumulative observed TM-scores for AF2_Advanced are displayed in Figure 2.10 as the cycle number increases. The following is a ranking of the different cycles based on the cumulative TM-scores in rank-1: cycle 12 (9.12) > cycle 24 (8.97) > cycle 6 (8.91) > cycle 48 (8.77) > cycle 3 (8.35) > cycle 1 (7.80). The cumulative TM-scores demonstrate that the overall structural quality of the modelled complex improves as the cycle increases when compared to cycle 1. Model improvement up to cycle 12 results in a model with superior quality compared to the preceding cycles. However, according to the paired statistical analysis, cycles 24 had the cumulative TM-scores that were significantly higher ($p < 0.05$) than cycle 1 (see Table 2.1).

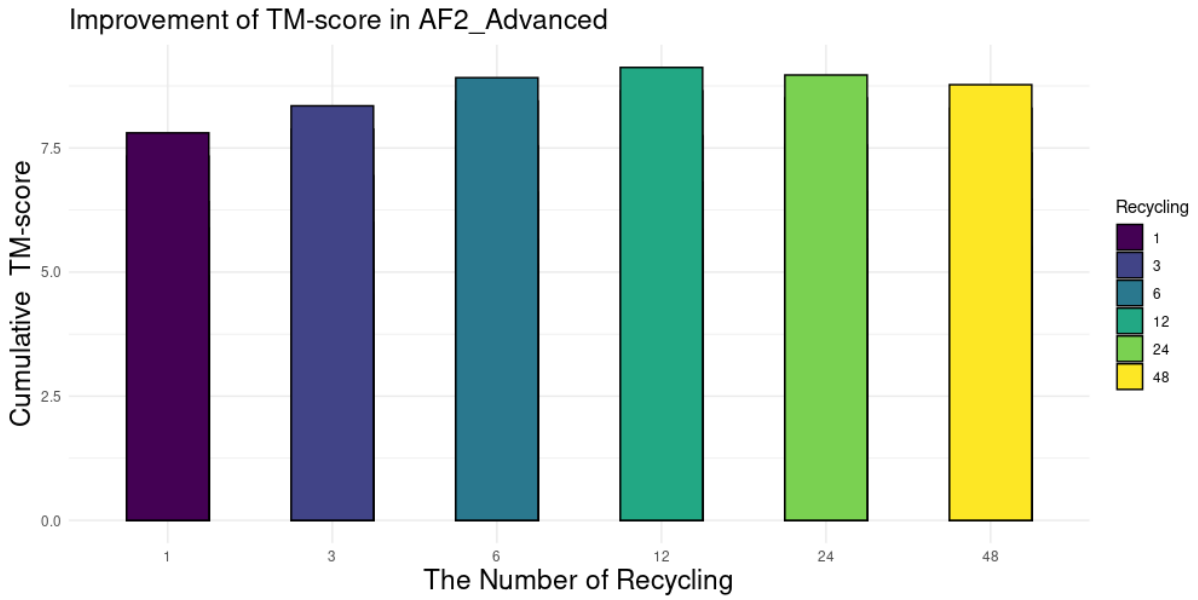


Figure 2.10 The improvement of TM-scores in the CASP14 models for AF2_Advanced. The bar chart presents the cumulative observed global quality score (assessed by TM-score) of the modelled complex structures of CASP14 targets generated by AF2_Advanced, based on rank-1 models. The six cycle rounds from which the rank-1 models are drawn are colour coded. The model with the highest pTM-score is the rank-1 model. This graphic was drawn using R.

The cumulative observed IDDT scores for AF2_Advanced are displayed in Figure 2.11 as the cycle number increases. The following is a ranking of the different cycles based on the cumulative IDDT scores in rank-1: cycle 12 (9.60) > cycle 24 (9.47) > cycle 6 (9.46) > cycle 48 (9.38) > cycle 3 (9.19) > cycle 1 (8.79). The cumulative IDDT scores demonstrate the same incline with the cumulative TM-score in Figure 2.10. According to the paired statistical analysis, all the cumulative IDDT scores up to cycle 48 were significantly greater ($p < 0.05$) than cycle 1. However, between other successive cycles including cycle 3- cycle 6 and cycle 6- cycle 12, there was no significant difference ($p > 0.05$) (Table 2.1).

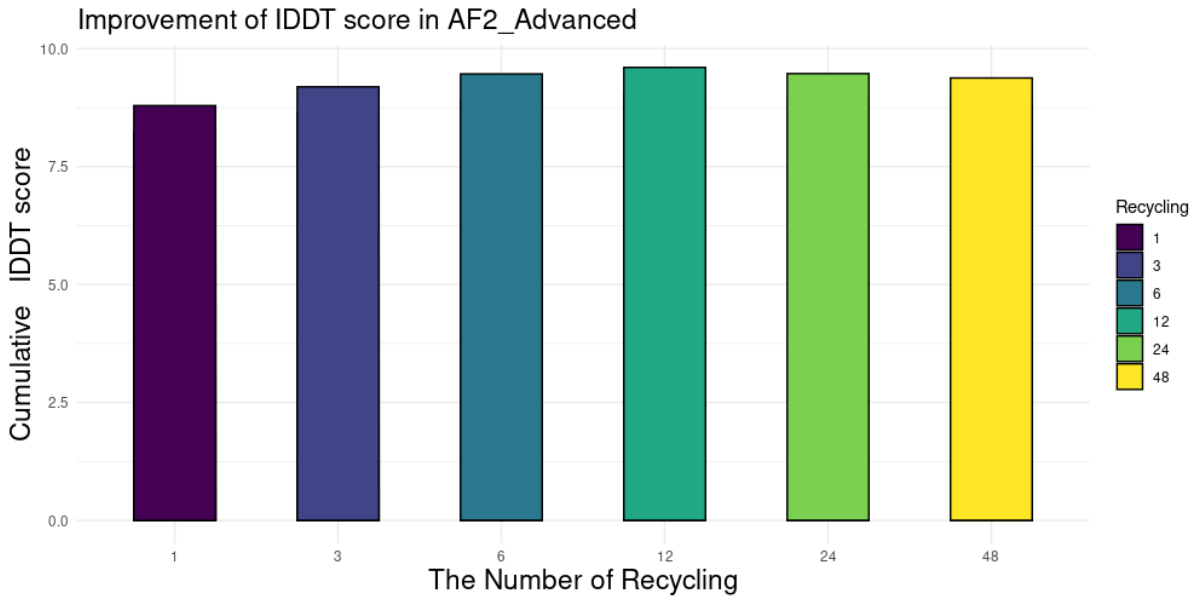


Figure 2.11 The improvement of IDDT scores in the CASP14 models for AF2_Advanced.
The bar chart presents the cumulative observed global quality score (assessed by IDDT score) of the modelled complex structures of CASP14 targets generated by AF2_Advanced, based on rank-1 models. The six cycle rounds from which the rank-1 models are drawn are colour coded. The model with the highest pTM-score is the rank-1 model. This graphic was drawn using R.

The cumulative observed QS-scores for AF2_Advanced are displayed in Figure 2.12 as the cycle number increases. The following is a ranking of the different cycles based on the cumulative QS-scores in rank-1: cycle 12 (6.79) > cycle 24 (6.44) > cycle 6 (6.34) > cycle 48 (6.09) > cycle 3 (5.3) > cycle 1 (4.25). Similar to both cumulative TM-score and IDDT score when compared to cycle 1, the cumulative QS-scores indicate an increase in the quality of the modelled complexes increases as the cycle increases and cumulatively model improvement up to cycle 12 generates a model with quality better than the previous cycles. According to the statistical analysis, the cumulative QS-scores up to cycle 24 were significantly greater ($p < 0.05$) than cycle 1. However, between other successive cycles, there was no significant difference ($p > 0.05$) (Table 2.1).

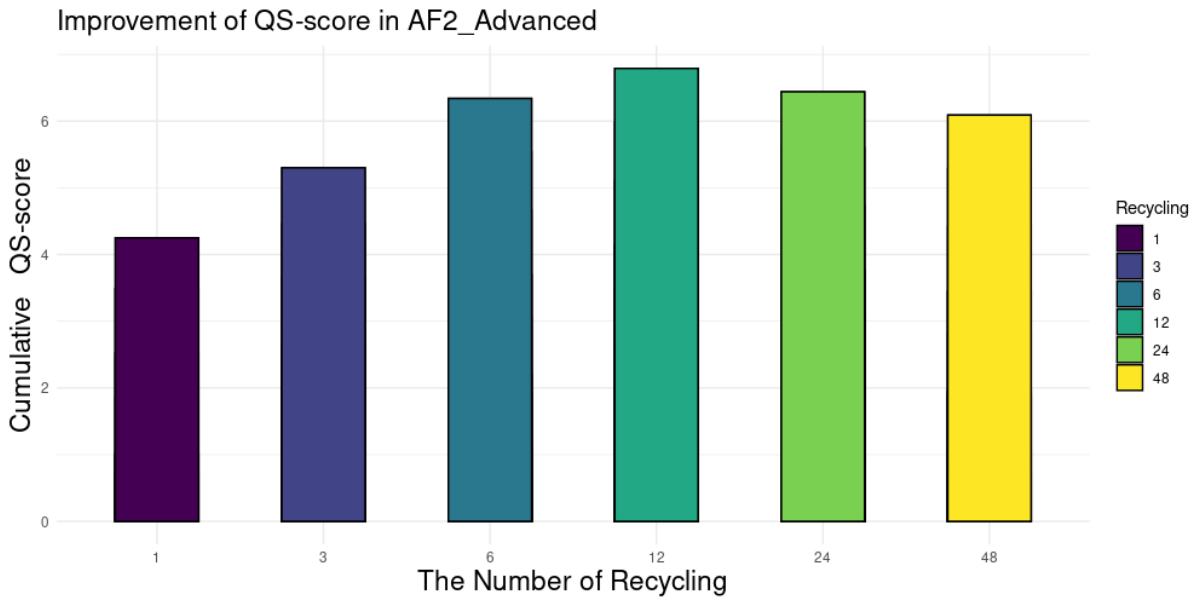


Figure 2.12 The improvement of QS-scores in the CASP14 models for AF2_Advanced. The bar chart presents the cumulative observed global quality score (assessed by QS-score) of the modelled complex structures of CASP14 targets generated by AF2_Advanced, based on rank-1 models. The six cycle rounds from which the rank-1 models are drawn are colour coded. The model with the highest pTM-score is the rank-1 model. This graphic was drawn using R.

Table 2.1 shows whether the differences in the observed quality of models produced using varying numbers of recycles (e.g., models produced using 1 cycle versus 3 cycles, and successive cycles) are statistically significant (note that the model generated in cycle1 was considered as initial/baseline model). For AF2_Advanced, firstly, the IDDT scores exhibited improvement when utilizing up to 48 cycles compared to only cycle 1 whereas the QS-scores of models were significantly improved using up to 24 cycles versus only 1 cycle (p-value < 0.05). Secondly, the TM-scores of models were only significantly improved using 1-12 cycles and 1-24 cycles (p-value < 0.05). However, no significant improvements were observed in the three quality scores (observed TM-score, IDDT, and QS scores) between successive cycles, except for 3-6 cycles and 6-12 cycles for the IDDT score. Therefore, this indicates that the increase in quality scores may continue with increasing cycles following cycle 24. Nevertheless, generating model data beyond 48 cycles was not evaluated, which is a challenging process because of additional computational load. Hence, considering that cycle 12 presents highest cumulative scores for all three-quality scores evaluated and after cycle12 the improvement is already significant versus cycle 1, selecting cycle 12 versus cycle 1 may lead to significant improvements on average for AF2_Advanced (Jumper et al., 2021a).

For AF2M, the statistical analysis revealed that the observed TM-scores of models for cycle 12 and cycle 24 were significantly greater ($p < 0.05$) than those from cycle 1. Furthermore, the observed IDDT scores for models from cycles > 12 to cycle < 48 were significantly greater ($p < 0.05$) than models produced using cycle 1, whereas no significant improvements in QS-scores were gained from increasing cycle number ($p > 0.05$). All p-values between consecutive cycles for AF2M were higher than 0.05, except for the comparison between cycles 6 and 12 for the IDDT score. Cumulatively, complex models exhibited improvement in either cycle 3 or cycle 48 for all scores. Considering the computational load in cycle 48 and the lack of statistical significance in cycle 3, the data suggest that using 12 recycles in the official Multimer version of AF2 rather than the default value of 3, would be a prudent option to increase the quality of multimers according to IDDT and TM-scores.

Although the cumulative score increases in recycle 3 is higher than in recycle 12, it is not statistically significant because the Wilcoxon test results are based entirely on the paired differences between individual pairs of data. The test evaluates the ranks of the differences between each paired pair of data, not the total or average score difference between the two data sets. Even if the cumulative score of recycle 3 is higher than recycle 12, if some of the individual differences are positive and some are negative or close to zero, the test may not find these differences to be ordinal significant. In contrast, if the distribution of paired differences between recycle is more consistently positive (i.e. most of the differences are significantly in

the same direction), the Wilcoxon test may consider this to be statistically significant. The results are therefore determined by the direction and consistency of the individual matched differences rather than the total scores. In the case of AF2, this shows a substantial difference between the scores in the individual matches as a result of refining the structure at some recycle values and generating a new model at the next recycle value.

Table 2.1 A comparison of performance for recycling processes (cycleX-cycleY) according to the cumulative scores of rank-1 models of CASP14 targets.

One-tailed Wilcoxon signed-rank tests were used to compare the effect of recycling with different cycle numbers for rank-1 models. H_0: according to the given pairwise cycles, the observed quality scores of models generated using *y* cycles are equal to or lower those of models generated using *x* cycles by the different AF2 variations, where *x* and *y* are integers between 1 and 48. H_1: according to the given pairwise cycles, the observed quality scores of models produced after *y* cycles are greater than those generated using *x* cycles by the different AF2 variations. P-values ≤ 0.05 indicate significant statistical differences. P-values where H_0 was rejected are in boldface. (n = 78 models for the observed TM-score, IDDT, and QS-score). One-tailed Wilcoxon signed-rank test was performed in the R program.

AF2-Multimer (AF2M)

One-tailed Wilcoxon signed-rank test (p-value)

Cycle-cycle	TM-score	IDDT score	QS-score
Cycle 1- cycle 3	3.78E-01	2.04E-01	6.63E-01
Cycle 1- cycle 6	7.34E-02	2.52E-01	4.00E-01
Cycle 1- cycle 12	**1.88E-02**	**4.49E-02**	6.64E-01
Cycle 1- cycle 24	**4.58E-02**	**2.73E-02**	5.56E-01
Cycle 1- cycle 48	2.88E-01	**1.51E-02**	2.21E-01
Cycle 3- cycle 6	4.19E-01	8.47E-01	6.64E-01
Cycle 6- cycle 12	3.80E-01	**1.83E-02**	2.65E-01
Cycle 12- cycle 24	4.12E-01	3.78E-01	6.05E-01
Cycle 24- cycle 48	6.37E-01	6.10E-01	3.36E-01

AF2_Advanced

One-tailed Wilcoxon signed-rank test (p-value)

Cycle-cycle	TM-score	IDDT score	QS-score
Cycle 1- cycle 3	1.47E-01	**1.27E-02**	**3.80E-02**
Cycle 1- cycle 6	1.04E-01	**1.27E-02**	**1.04E-02**
Cycle 1- cycle 12	**4.04E-02**	**3.96E-03**	**1.04E-02**
Cycle 1- cycle 24	**4.03E-02**	**4.86E-03**	**1.50E-02**
Cycle 1- cycle 48	2.21E-01	**7.22E-03**	6.15E-02
Cycle 3- cycle 6	3.38E-01	**5.94E-03**	6.38E-02
Cycle 6- cycle 12	1.47E-01	**4.94E-03**	1.98E-01
Cycle 12- cycle 24	5.83E-01	9.95E-01	8.60E-01
Cycle 24- cycle 48	8.18E-01	7.35E-01	9.61E-01

On the one hand, the recycling process can be impacted by the abundance of disorder regions corresponding to -N and -C termini region of protein structures. These regions lack the contact or distance predictions based on pair residue-residue, hence, predicting -N and -C regions present a challenge for improving models through subsequent recycles. On the other hand, the flexible regions of protein structure can be better generated by more detailed physical approaches rather than AF2's network. Since AF2 is predominantly based on an evolutionary constraint in the MSA where the last 3D protein structure in the previous cycle is used to determine the pair residue constraint in the next cycle, the intermediate structure models with lower quality can generate wrong pair residues for the next cycle. Hence, the lower cycle numbers may be better for protein structures with more flexible regions. In addition, these proteins often have heterogeneous structures (e.g.: transient alpha-helices in the disorder regions) that may only be predicted with the integration of physical laws (MD simulation) with machine learning, therefore, such disordered or highly flexible structures still represent a challenge for the AF2 algorithm. However, it can be highlighted that intertwined regions of protein, such as coiled coils, can be generated well by AF2M (Evans et al., 2022). It is interesting to note that by using AF2 it is often possible to identify the regions with low pIDDT scores as being disordered (natively unstructured) regions rather than being poorly modelled regions of structure. This concept of associating local regions of low model quality with native disorder was pioneered in the DISOclust (McGuffin, 2008) method which is part of the IntFOLD (McGuffin et al., 2019) server.

It is intriguing to assess AF2M in terms of QS-scores, as demonstrated by Figure 2.9 and Table 2.1, which indicate the ongoing need for recycling beyond cycle 48 to better predict interface areas. Firstly, up to cycle 48, the cumulative QS-score for cycle 12 was initially lower, but then increased for cycle 48 compared to cycle 1. However, such an incline was not observed for TM-score and lDDT score. Secondly, the initial AF2M method faced challenges when the complex structure comprised more than two chains (Bryant, Pozzati, Zhu, et al., 2022), potentially causing an increase or decrease in the QS-scores for models generated by AF2M when further recycling was applied. For instance, T1073, including four chains, obtained a QS-score of "0". Moreover, T1073 was evaluated as easy target by CASP assessor, thus the poor orientation of the subunits may result in a lower global observed score for the quaternary structure model after superposition, even if the individual tertiary structures of the subunits are well predicted. Despite the fact that AF2M was validated in terms of pairwise interface quality score (DockQ score), the reason for using QS-score is to evaluate all interface quality scores at once giving a single score. (note: since this initial study was performed, there is now an additional variant of DockQ, DockQ-wave (Kryshtafovych et al., 2023b; Studer et al., 2023), which gives a single score for all interfaces. DockQ-wave was one of the scores used in

CASP15 - see subsequent chapters). The cumulative QS-scores for AF2M were higher than AF2_Advanced. This difference in performance might be explained by the cropping procedure that made training AF2 systems easier. In the training phase, the cropping of up to 384 amino acids in protein sequences for AF2_Advanced was made while for AF2M, it was specifically trained on multimers and designed to maximize chain coverage. In this way, in AF2_Advanced, interface information for multimer targets was somewhat lost, while in AF2M, a good balance was obtained between interface and non-interface regions. However, for AF2M, this cropping is optimized so that binding interfaces in the multiple chains are involved (Evans et al., 2022).

Despite the main difference underlying AF2M, compared with the original AF2, is in its training parameters, which were specifically designed for modelling complexes, there are still notable drawbacks to relying solely on this method for modelling quaternary structures. Despite T1083 and T1084 presenting highly similar structural targets and difference sequences in the CASP14, and both having their monomer structures evaluated as hard predictions, there was a substantial difference between three scores of both. When analysing the reason, T1084 exhibited low sequence identity (< 0.4) across all positions in the MSA, whereas T1083 showed varied sequence coverage for separate positions in the MSA. This indicates that AF2M relies predominantly on the strength of MSA mining tools. The figures in Appendix Figure S.1a depict the difference in sequence coverage by MMseq2 for T1083 and T1084 during cycle 12, along with sequence numbers in Appendix Figure S.1b and reference structures in Appendix Figure S.1c. Furthermore, Dapkūnas et al. (2021) found that the docking approach was better than the template approaches, as AF2M relies on template structures for subsampling (Jumper et al., 2021a) Importantly, it is noteworthy that MMseq2 can force a search leading to false positives due to its masking and filtering function (Mirdita et al., 2022).

Recycling analysis on CASP12 and CASP13 targets as well as CASP14 was analysed as well. However, since the AF2M training phase occurred after CASP13, models from CASP12, 13 were not included in the general evaluation to avoid biased decisions, under the assumption that they would already predict well. However, when analysing the models, although AF2M is trained with these structures, deficiencies continue. The QS-score for the T0991 (A2) targets in CASP13 (PDB code: 6YFJ), for example, could not be calculated, which was unexcepted. The OpenStructure tool detected multiple clashes in the model and all scores displayed 'failed' when chemically matching the reference structure and the target structure. Therefore, this model had QS-score of null, or zero. Upon detailed examination of the modelled complex, many irregularities such as clashes with disorder regions were observed. Even though AF2M was trained with these structures, models with more disorder regions were not successful, suggesting a need to update training processes.

When the differences in the cumulative observed scores for models within the same cycle numbers were compared between the two AF2 versions, the cumulative observed quality scores of models generated by AF2_Advanced were found to be lower than AF2M during the recycling procedure (Figures 2.7-2.12). Subsequently, the significance of the differences for each recycle value was analysed using Wilcoxon signed-rank tests. The observed TM-scores, IDDT and the QS-score for all cycle pairs were p-value > 0.05, except for one between cycle 1 for QS-Score of both of the AF2 versions (See Appendix Table S.1). With the application of additional recycling, both versions seem equally effective in terms of their utilization, as there is no statistical difference between the two AF2 variants. Therefore, the question arising here is which version will undergo how many recycling rounds? Considering the potential updates in forthcoming versions of AF2M with further recycling, it may be suitable to specifically tailor them for modelling quaternary protein structures. It is significant to note that using a metagenomic database increases the predictive power of the AF2 versions. However, for efficiency the metagenomic database (MGnify) had been reduced (Jumper et al., 2021a). This stage of reduction of the database may have led to loss of significant information, Hence, the information lost with MMseqs2 could potentially be regained by providing the structural information of a target proteins as a template to AF2M to enable a more accurate structure.

The last but not least, it should not be forgotten that while high-quality scores are important, the initially obtained structure is crucial for recycling since it is a predicted model. Protein structures can have highly diverse conformational states, particularly in apo and holo forms. Hence, comparison of protein structures with crystal structures may not fully reflect their inherent dynamics, because crystal structures are static, whereas proteins fluctuate and can adopt different conformations in aqueous environments. A crystal structure represents only a snapshot of a protein's state, whereas proteins are flexible and can have multiple stable conformations. During model evaluation, small structural differences can reflect the natural mobility of proteins and these variations can be biologically meaningful. Therefore, although it is useful to compare models to crystal structures, it is important to consider the flexible and dynamic nature of proteins in such evaluations. However, the structure selected and stored in the database is considered to be the one that best represents the conformational structures. However, demonstrating that the individual structure obtained by AF2 is truly the desired one for our purposes remains open to debate.

2.4 Conclusions

Historically, the modelling methods for predicting the quaternary structures of proteins have involved either template-based approaches or docking approaches, starting from the individual tertiary structures, either observed or predicted (Mandell et al., 2001). Conventional protein-protein docking methods have so far used knowledge-based interaction “energies” (Tovchigrechko & Vakser, 2006) and/or protein-protein interaction physics (Mashiach et al., 2010) to predict quaternary structures, using this information as either constraints or as scoring functions for ranking various model poses. Recently, end-to-end deep learning-based approaches have become more of a focus as they provide results without the need for feature selection or handcrafting intermediate process. AF2, which uses, has been a major milestone for the protein structure prediction field, as it uses end-to-end deep learning methods instead of the traditional methods which have thus far taken the centre stage in this field.

Using end-to-end deep learning, AF2 and related algorithms are capable of predicting a tertiary structure very close to a local structure. However, this predictive power is highly dependent on the internal structure of the given target. The predictive power of AF2 was validated through the CASP14 and CASP15 experiments. AF2 is dependent on the availability of comprehensive MSAs which reveal the coevolutionary information of the given protein structure, allowing the method to model the pairwise residue interactions through template structures, and ultimately recycle them repeatedly leading to performance gains (Jumper et al., 2021b). With DeepMind’s release of the code for the AF2 algorithm, the bioinformatics community has been able to explore how to integrate its core algorithm with existing tools.

The recycling process of the AF2 method is akin to a model iterative refinement stage (Bhattacharya, 2019) in a traditional modelling pipeline; it is the final part of the algorithm, and it repeatedly provides intermediate losses back to the system (Jumper et al., 2021a). Indeed, the refinement capability of AF2’s has been demonstrated through its use to improve docking models from ClusPro. By using ClusPro models as “template” inputs, the AF2 recycling process was used which resulted in docking models with higher observed quality scores (Ghani et al., 2021). In both AF2 versions tested here (AF2_Advanced and AF2M), the maximum number of recycles is set to 3 by default. However, at the time of our study, this value had not been validated in detail or optimised for modelling protein complexes. In this regard, our research is based on statistical analysis in order to determine the optimal maximum cycle number for both AF2 versions, and furthermore, we compare the modelling performance of each version of AF2.

In the initial version of AF2M and AF2_Advanced, utilizing increasing cycle values may lead to improved results. As a result of the Wilcoxon signed-rank test, cycle 12 and 24 for AF2M appear to be the optimal value set for the “max_cycle” parameter, despite cycle 3 or cycle 48 showing highest cumulative scores. It seems model improvement can be influenced by target context rather than obtaining the same increase rate for all protein models during recycling. When it comes to consecutive cycles for IDDT score, the improvement between cycle 6 and 12 is statistically significant, hence selecting a maximum cycle number of 12 is prudent without prohibitive additional computational burden. For AF2_Advanced, significant differences were observed in the observed quality scores (TM-score, IDDT and QS-score) between models produced using cycle 1 and cycles > 1, and this improvement persisted up to cycle 12 and 24 for all observed quality scores. However, again it is reasonable to use a maximum cycle number of 12, as it produces the highest cumulative improvements for all three observed quality scores. Based on the cumulative scores of the observed quality scores in Figures 2.7-2.12 and the statistical analysis in Appendix Table S.1, it is not expected that the performance of the first version of AF2M will be superior for modelling quaternary structures compared to the AF2_Advanced version.

Further improvements could be made to AF2 variants for more accurately modelling of complexes. This may involve firstly further optimisation of NN model parameters, retraining on alternative datasets, and/or integrating significant specific features (such as the post-translational modification of residues) of protein complexes into the algorithms. In doing so, higher quality proteins may be obtained with fewer recycles, improving efficiency. Secondly, given the QS-score results, there is room for AF2 to improvements in this regard, perhaps by encoding the interface dynamics as input features for better estimation of interface space. Thirdly, to model the best protein, it might be possible to determine the optimal number of cycles from the quality (sequence depth) of MSA, or the templates used. Additionally, better quality MSA and templates may provide AF2 variants (input parameters referred to as ‘custom_MSA’ and ‘custom_Template’ in ColabFold last version (v.1.5.5) of AF2M will make this a possibility to explore). This concept of "recycling" explored in this chapter invokes images of an extracting usefulness by the cleaning up initial messy/garbage information. To extend this metaphor, the question we have addressed here is: how much recycling is required to clean up the model enough to obtain a useful output? If MSAs are of sufficient quality or there are enough available templates with high-quality scores for a particular target, then models may require much less of a cleanup process. Conversely, if only shallow MSAs can be generated or few suitable templates exist for a target, then resulting models may require a more intense cleanup during recycling.

Chapter 3: The Impact of the Custom Template Recycling for the Improvement of Quaternary Structures of Proteins

The results from the CASP14 data provided in this chapter have been published in *Bioinformatics Advances:*

Recep Adiyaman, Nicholas S Edmunds, **Ahmet G Genc**, Shuaa M A Alharbi, Liam J McGuffin, Improvement of protein tertiary and quaternary structure predictions using the ReFOLD refinement method and the AlphaFold2 recycling process, *Bioinformatics Advances*, Volume 3, Issue 1, 2023, vbad078, https://doi.org/10.1093/bioadv/vbad078

Author contributions:

Recep Adiyaman: Idea development, Data management, Structured analysis, Securing funding, Conducting research, Applying research methods, Project management, Utilizing resources and computational tools, Providing oversight, Verification, Data representation, Drafting, Revising, and Editing the manuscript.

Nicholas S. Edmunds: Idea development, Data management, Structured analysis, Securing funding, Conducting research, Applying research methods, Project management, Utilizing resources and computational tools, Providing oversight, Verification, Data representation, Drafting, Revising, and Editing the manuscript.

Ahmet G. Genc: Idea development, Data management, Structured analysis, Securing funding, Conducting research, Applying research methods, Project management, Utilizing resources and computational tools, Providing oversight, Verification, Data representation, Drafting, Revising, and Editing the manuscript.

Shuaa M. A. Alharbi: Securing funding, Providing oversight, and Data representation.

Liam J. McGuffin: Conceptualization, Formal analysis, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing (original draft), and Review and editing.

3.1 Background

In the CASP14 experiment, AF2`s performance represented unprecedented success in modelling monomeric protein structures, irrespective of the degree of structural complexity (Ozden et al., 2023). The success of AF2 comes from its ability to interpret the co-evolutionary structural information embedded in the MSAs used as inputs for the modelling process. In addition, AF2 can succeed in monomer protein modelling without structural template when sufficient homolog sequences in an MSA are provided(Jumper et al., 2021a). However, the performance drops off for more difficult targets, when evolutionary information is weaker (Bryant et al., 2022). In this instance, it can be possible to generate complex structures of higher quality through either more conformational sampling (Wallner, 2023b) or the incorporation of structural templates to constrain the sampling process (Adiyaman et al., 2023).

Inspired by the success of AF2 in modelling monomeric structures, DeepMind enhanced AF2 by implementing several modifications (as indicated in Chapter 1) and released AF2M. Nevertheless, the success of AF2M diminishes when there is insufficient evolutionary signal and/or limited structural template availability. However, it has been noted that AF2M are successfully used to obtain refined models using similar structural templates and the augmented MSA (Liu et al., 2023a). Therefore, it can be suggested that modelling using suitable template structures from external sources may also be beneficial for certain targets.

One of the most significant studies after AF2M was released was the study of (Terwilliger et al., 2022). Terwilliger *et al.* argued that the information from Cyro-EM density maps, along with the AF2 recycling process, could be used to obtain higher quality protein structure models. To demolish this, the AF2M models were trimmed and superimposed with the density maps before each subsequent cycle. Following each cycle, the resulting rebuilt model was then used as the template to guide the prediction of the final protein structure. Following this success of using low resolution experimental data for iterative structure prediction via AF2M recycling, Sergey et al. added “custom template” option to the ColabFold (Mirdita et al., 2022) platform. The algorithm ran the same as Terwilliger et al`s AF-Phenix (Terwilliger et al., 2022), the only difference being that the input structural template guided the algorithm rather than a density map. Thus, the approach exploits AF2M`s attention mechanism, which focuses on the most pertinent structural template information from the input data to construct a protein structure model.

3.1.1 The aim of study

The success of AF2 was demonstrated following CASP14, and the wider community began exploiting the AF2 algorithm in different ways including by our own group, where we used it as a refinement tool to generate improved models for CASP15. After a certain period, due to AF2M's algorithmic approach of treating each chain of the final output as individual monomers, there exists a likelihood of degradation in models during successive recycles, as each iteration necessitates re-modelling. Consequently, recycling the generated model can either further improve or worsen it. Moreover, since a specific numerical value for the number of recycles has not yet been determined, stopping at the default number of cycles poses a high risk of missing the best model. Therefore, finding the optimal number of cycles is crucial both in terms of time and efficiency.

In Chapter 2, it was demonstrated that AF2 versions, utilizing input structural templates solely obtained from the structural databank via HHpred (Söding et al., 2005), exhibit a greater improvement effect on structural modelling through further recycling. This chapter aims to investigate the impact of providing externally structural models as “custom templates” on the improvement of models through recycling. To demonstrate this, two different template populations were gathered: Firstly, the pre-AF2M models on the CASP14 data which generated by the methods developed prior to the release of AF2 code and, secondly, the post-AF2M models on the CASP15 generated by the methods that were developed after AF2 code was released. The effect of recycling on structure modelling has been examined from various perspectives. Wallner et al. (2023) investigated the recycling of AF2M using different versions and optional parameters alongside MSA, focusing primarily on the Global-DockQ score by using the average of DockQ score for each protein interface weighted based on an interface size (Basu & Wallner, 2016; Wallner, 2023a). However, our study aims to evaluate models through different quality scores (TM-score, IDDT score, QS-score, DockQ_wave score, Molprobity) and also to examine whether recycling can still be used to refine input template models even in the absence of an input MSA using the SS option. Thus, this investigation is unique in its aims to explore the effects of exploiting various custom template inputs for AF2M, with the goal of enhancing recycling effectiveness while considering performance and efficiency aspects.

3.2 Methods

3.2.1 Data collection

3.2.1.1 Data collection of CASP14 models

The CASP14 multimeric models were downloaded from CASP website (https://predictioncenter.org/download_area/CASP14/predictions/), according to the CASP target codes. 10 CASP14 targets, including H1045 (PEX4-PEX22 (Organism: Arabidopsis thaliana)), H1065 (N4-Cytosine Methyltransferase (Organism: Serratia marcescens)), H1072 (Testis-expressed protein 12 (Organism: Homo sapiens)), T1032 (Structural maintenance of chromosomes flexible hinge domain containing 1 (Organism: Homo sapiens)), T1054 BonA (Organism: Acinetobacter baumannii)), T1070 (Tailspike protein (Organism: Escherichia virus CBA120)), T1073 (Hypothetical protein predicted by Glimmer/Critica (Organism: Bdellovibrio bacteriovorus)), T1078 (a small secreted cysteine-rich protein (Tsp1) (Organism: Trichoderma virens)), T1083 (Nitro-histidine zipper coiled coils (Organism: Nitrosococcus oceani)), and T1084 (Meio-histidine zipper coiled coils (Organism: Meiothermus silvanus)) were selected. The criteria for target inclusion in the study were the submission of same target by six different groups during CASP14, the presence of experimentally validated structures, and the limit on the number of residues which could be modelled using AF2M. The top models (model-1) for each target were taken from the top-performing groups (according to the assessor`s Z score ranking), including Baker-experimental (Baek et al., 2021), Venclovas (Dapkūnas et al., 2021), Takeda-Shitaka, Seok (Park et al., 2021), and DATE. For each target, each model was used as an input template, or initial model, for AF2M (v.1.2) prior to recycling. Also, as the DeepMind group did not submit models for multimeric targets in CASP14, AF2M models were generated for the same targets and were also used as input templates for AF2M.

3.2.1.2 Data collection of CASP15 models

The CASP15 multimeric models were also downloaded from the CASP website (https://predictioncenter.org/download_area/CASP15/predictions/), according to the CASP target codes. 24 CASP15 targets, including H1106 (YscY-YscX protein (Organism: Yersinia enterocolitica)), T1109 (D180A isocyanide hydratase (Organism: Ralstonia solanacearum)), T1110 (wild-type isocyanide hydratase (Organism: Ralstonia solanacearum)), T1113 (Glycoprotein 2 (GP2) (Organism: Bacteriophage PA1C)), T1121 (the Wadjet nuclease subunit JetD (Organism: Pseudomonas aeruginosa PA14)), T1123 (Capsid protein (Organism: Human Astrovirus MLB1)), H1129 (Receptor-binding protein pb5 (Organism: enterobacteriophage T5)), T1132 (Antibiotic biosynthesis monooxygenase (Organism: Pseudomonas aeruginosa)), H1134 (Chymotrypsin digested toxin/immunity complex for a T6SS lipase effector (Organism: enterobacter cloacae)), H1140 (CNPase-Nb (Organism: mouse/alpaca)), H1141 (CNPase-Nb7e (Organism: mouse/alpaca)), H1142 (CNPase-Nb8c (Organism: mouse/alpaca)), H1143 (CNPase-Nb10e (Organism: mouse/alpaca)), H1144 (CNPase-Nb8d (Organism: mouse/alpaca)), H1151 (Probable

transcriptional regulator WhiB6 (Organism: Mycobacterium tuberculosis)), T1153 (Endonuclease/exonuclease/phosphatase family domain-containing protein 1(Organism: Human)), T1160 (The mk2h_deltaMILPYS peptide(Organism: Ancient protein reconstruction)), T1161 (The dimeric DZBB fold protein Ph1(Organism: Ancient protein reconstruction)), H1166 (Human Fab S24-188 in the complex with the N-teminal Domain of Nucleocapsid protein from SARS CoV-2 (Organism: Human)), H1167 (Human Fab S24-1379 in the Complex with the N-teminal Domain of Nucleocapsid Protein from SARS CoV-2 (Organism: Human)), H1168 (Human Fab S24-1063 in the Complex with the N-teminal Domain of Nucleocapsid Protein from SARS CoV-2(Organism: Human)), T1173 (Cell wall surface anchor family protein (Organism: Bdellovibrio bacteriovorus)), T1174 (the C-terminal domains of the Bdellovibrio bacteriovorus Bd2133 fibre (Organism: Bdellovibrio bacteriovorus)), and T1179 (the murine astrovirus capsid spike (Organism: Murine astrovirus)) were selected. Again, the criteria for target inclusion in the study were the submission of same target by six groups during CASP15, the presence of experimentally validated structures, and the limit on the number of residues accepted by AF2M. The top models were again taken from the top-performing groups (according to the assessor`s Z score ranking), which this time included Zheng, Venclovas, Wallner, and Yang-Multimer. Additionally, the NBIS-AF2-Multimer group was selected to evaluate the performance of AF2M models while MultiFOLD was selected (our group's automated server multimer prediction tool), which was ranked 8th in CASP15. For each target, each model was used as an input template for AF2M prior to recycling.

3.2.2 Experimental design

Each model PDB file was transformed to mmCIF format using the RSCB PDB MAXIT conversion tool (https://mmcif.pdbj.org/converter for CASP14 target). The transformed model files were then uploaded to the Google Colaboratory hosted by ColabFold [v1.3.0 (4-March-2022)] for CASP14 target while the model structure was then uploaded to the Google Colaboratory hosted by ColabFold [v1.5.3 (After 4-March-2023)] for the CASP15 targets as "custom templates" along with their corresponding amino acid sequences. Two separate sub-populations of recycled models were then created for each individual model: "MSA models" for which ColabFold was permitted to construct an MSA prior to recycling and "Single-Sequence (SS)" models for which an MSA was not used. Every initial model was subjected to 1, 3, 6, and 12 recycles separately in both the MSA and SS modes. In all cases, Amber relaxation was disabled to ensure that we were solely testing for the recycling effect. Rank-1 models created for each ColabFold run were collected along with their pTM-scores and pIDDT scores.

The ColabFold settings used were:

For the CASP14 models, Template_mode: custom; msa_mode: MMseqs2 (UniRef+Environmental) OR single sequence; pair_mode: unpaired+paired; model-type: auto; num_recycles: 1, 3, 6, 12. (N.B. auto was specified as the model type, which at the time defaulted to the original pre-CASP14 model `alphafold2_Multimer_v1`, avoiding biased structure prediction for the test set).

For the CASP15 models, Template_mode: custom; msa_mode: MMseqs2 (UniRef+Environmental) OR single sequence; pair_mode: unpaired+paired; model-type: alphafold2_Multimer_v2; num_recycles: 1, 3, 6, 12. (N.B. 'alphafold2_Multimer_v2' was specified as the model type, which was the original pre-CASP15 model, in order to avoid biased structure prediction to the test set).

3.2.3 Structure analysis

The baseline initial models (the input template models prior to recycling) and the models generated by each number of recycles were then directly compared with the native structure obtained from the PDB, to generate observed quality scores (expect for the Molprobity score (Chen et al., 2010(a)), which evaluated the stereochemistry of each model). The scores were then compared between the baseline models and the recycled models. For evaluating each modelled multimeric structure, the MM-Align (Mukherjee & Zhang, 2009) and OpenStructure (https://openstructure.org/). Programs were employed to obtain observed scores for the TM-score and IDDT score, respectively. The IDDT score referred to in this chapter is the Oligo-IDDT. Additionally, the QS-score from the OpenStructure program was used to measure the interfaces of multimers. The QS-score proves valuable in comparing homo- and hetero-complexes with different stoichiometries, diverse orientations of relative chains, and varied amino acid sequences (Bertoni et al., 2017). At the time of the analysis of the CASP14 targets the latest version, OpenStructure 1.1, was utilized. However, for the CASP15 targets, OpenStructure 2.1 was available and used to produce equivalent scores. In addition, DockQ_wave scores (Studer et al., 2023) were obtained for the CASP15 targets. The DockQ_wave scores for CASP15 targets were acquired using OpenStructure version 2.1 (DockQ_wave was not available in the 1.1 version, while the Molprobity scores were obtained using the server (http://molprobity.biochem.duke.edu/). Statistical analysis was conducted by comparing the performance of methods based on the increase in observed scores for models when using to custom template recycling options. The statistical method used is explained in detail in the methods section of Chapter 2. Figure 3.1 summarises the workflow of methods used in the analysis for this chapter.

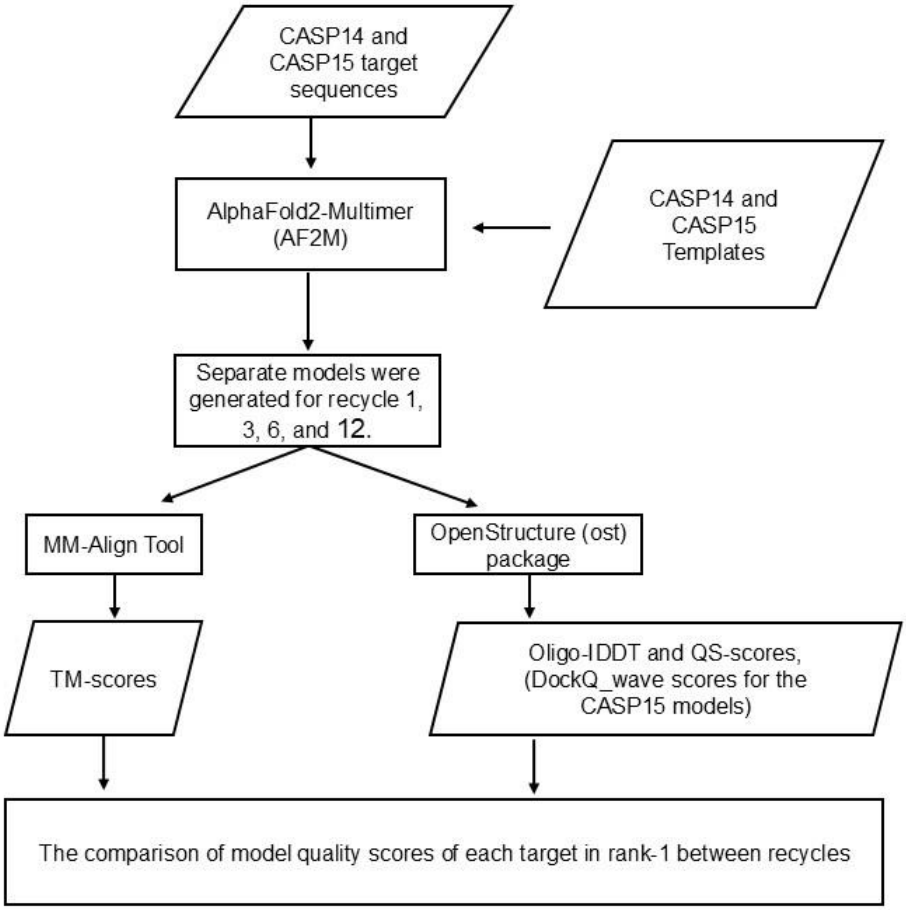


Figure 3.1 The flowchart of the method for using the custom template recycling options from modelling quaternary structures.
Flowchart showing the process for evaluating custom template cycles in AF2M using MM-Align tool and OpenStructure package. The observed quality scores for CASP14 and CASP15 models, with TM-score from MM-Align and IDDT/QS-score from OpenStructure, were produced by aligning the models with the native structures for each target. Also, DockQ_wave score was generated via OpenStructure only for CASP15 models. Subsequently, the observed quality scores for rank-1 models for each group in each cycle round were statistically evaluated using the paired Wilcoxon-signed-rank test (compared with the initial scores).

3.3 Results and Discussion

The results comprise the top modelling results for the 10 CASP14 and 24 CASP15 targets. Our initial aim is to seek the effectiveness of using the custom template option with further recycles in AF2M with a view to integrate these improvements as part of our complex protein structure prediction tool, called MultiFOLD. Additionally, in order to control for the effect of the inclusion of an MSA, models were processed both with MSA and in SS methods specified prior to recycling. The custom template option of AF2M is the most important input to guide conformational space searching of the target proteins. Hence the models from the top groups were used as the custom templates, and these initial models were generated by different methods other than AF2M, thereby giving the AF2M algorithm a different starting point for modelling. The most effective way of evaluating multimeric structure improvement with further recycle is to align the models with the known structures and then compare

the resulting observed quality scores. For this reason, the TM-score and IDDT scores were used as scores to evaluate the observed model quality at the overall fold level while QS-score and DockQ_wave scores are used for comparing the quality of the modelled interfaces compared with the experimentally determined structure. The TM-score was used as the principal evaluation score to observe the improvement from one cycle to the next, as AF2M relies on pTM-score and then ipTM (predicted interface TM-score) to select the best multimeric prediction target. Additionally, the Molprobity score was employed to analyse prediction models in each cycle, which considers the stereochemistry of all atoms in a model and does not require comparison of experimental structures.

Firstly, the effect of custom template recycling on the 60 models (CASP14) and 144 models (CASP15) generated by AF2M using MSA and SS methods was investigated. Table 3.1 demonstrates that the initial models were refined when used the custom template along with further recycling. This refinement is evident in terms of four cumulative quality scores: TM-score, IDDT, Qs_score, and Molprobity score. The improvement was linear during further recycling when compared to the scores for the refined models and baseline models. In addition, without MSA information, the improvement remained consistent, although all quality scores for the MSAs method were higher than that of the SS method during each cycle. These results suggest that structural information can be valuable when AF2M is run using MSAs. Without using MSA, custom templates with further recycling were used to generate models. The results show that improvements were observed as the number of recycles was increased; however, no improvement was observed when the initial models were compared. To analyse the effect of the quality of initial protein structure on models generated AF2M using custom template recycling, the analysis of refinement during recycling process starts by comparing the quality of the initial baseline AF2M models (models generated by ColabFold for the CASP14 targets and those generated by NBIS-AF2M for the CASP15 targets) and non-AF2M models (models generated by the 5 top-ranked groups in the CASP14 competition and those generated by the four top-ranked groups plus, MultiFOLD in the CASP15 competition). Testing of both the CASP14 and CASP15 models has provided an opportunity to historically examine the recycling algorithm of AF2M for use on pre-AF2 and post-AF2 models. Table 3.2 supports the refinement effect of custom template recycling on the CASP models. Compared to the baseline cumulative scores, the cumulative TM-score, IDDT, QS-score for the models generated by AF2M using the MSA method are statistically significant ($p < 0.05$) for recycles 3, 6, and 12. However, the differences in the cumulative IDDT and QS-score between baseline and recycle 1 are not statistically significant. Moreover, these appears to be a significant increase in the quality of models generated during three separate recycles after recycle 1, compared to its previous recycle.

Table 3.1 A comparison of the cumulative quality scores for the CASP models versus the baseline models after recycling.
The table showing the cumulative scores of the starting models, known as, baseline and of the CASP14 and CASP15 models obtained after four various recycling (1-3-6-12) using the MSA and SS methods. The increase in cumulative scores for TM-scores, IDDT, QS-scores indicates that the refined models were obtained following further recycling (between cycle 6 and cycle 12). It is significant to note that the lower Molprobity score presents higher quality model.

Method	Baseline	Cycle-1	Cycle-3	Cycle-6	Cycle-12
			∑TM-score		
MSA	163.595	167.067	170.364	170.978	171.181
SS	160.482	145.749	152.019	156.602	150.913
			∑IDDT		
MSA	159.625	160.351	164.33	166.974	167.037
SS	156.216	140.092	143.284	145.259	155.004
			∑QS-score		
MSA	115.47	119.5	126.26	128.21	128.78
SS	112.38	87.92	101.78	108.02	111.47
			∑Molprobity score		
MSA	341.07	528.14	520	507.81	505.6
SS	348.95	583.17	575.47	572.76	571.12

Table 3.2 A statistical comparison of the cumulative quality scores for the CASP models versus baseline models after recycling.
Computed p-values resulting from the Wilcoxon signed-rank test for TM-scores (a), IDDT (b), and QS-scores (c) in relation to both baseline and recycled models, considering both CASP14 and CASP15 models generated by AF2M.

Recycle type	Baseline to 1 recycle	1 recycle to 3 recycles	Baseline to 3 recycles	3 recycles to 6 recycles	Baseline to 6 recycles	6 recycles to 12 recycles	Baseline to 12 recycles
a)			TM-score				
MSA	**9.147e-3**	8.547e-2	**8.924e-4**	4.816e-1	**3.789e-4**	9.671e-1	**1.102e-3**
b)			IDDT				
MSA	8.869e-1	**1.307e-08**	**1.464e-2**	**1.531e-07**	**7.678e-3**	7.357e-1	**1.937e-3**
c)			QS-score				
MSA	2.335e-1	**1.56e-4**	**8.676e-3**	**1.676e-2**	**1.04e-3**	**4.959e-2**	**2.815e-4**

*MSA: Multiple Sequence Alignment, Ho: Recycling, as stated by column, produces models of equivalent or lower quality than the template structures used as baseline templates, or the models generated in previous cycles. H1: Recycling, as stated by column, produces models of higher quality than the template structures used as baseline templates, or the models generated in previous cycles. P-values ≤0.05 indicate there is significant differences, as in highlighted in bold. The one-tailed Wilcoxon signed-rank tests were used to measure the (A) TM-scores, (B) IDDT scores, (C) QS-scores for 60 models (CASP14) and 144 models (CASP15).

3.3.1 The impact of custom template recycling with MSA on multimeric CASP14 models

In Chapter 2, it was observed that further recycling has a refinement effect on multimeric targets. Here, the effect of recycling on the refinement of custom template models where initial structure guides to the AF2M modelling process was measured. Analysing the TM-scores, 70% of AF2M models improved, whereas 98% non-AF2M models demonstrated a more substantial improvement. In terms of IDDT scores, the recycling of multimeric models resulted in an improvement in 80% of AF2M models, compared to 94% of non-AF2M models. Additionally, considering QS-scores, 50% of AF2M models exhibited improvement, while 86% of non-AF2M models showed an improvement. When it comes to the SS approach, analysing the TM-scores showed that 80% of AF2M models refined, whereas 82% of non-AF2M models demonstrated a more substantial improvement. In terms of IDDT scores, the recycling of multimeric models resulted in an improvement of 30% of AF2M models, compared to 64% of non-AF2M models. Lastly, considering QS-scores, %30 of AF2M models had better quality after recycling, while 60 % of non-AF2M models also showed refinement (See Appendix Figure S.2).

In Adiyaman et al. (2023), these above results were published and the effect of custom template recycling on the multimeric CASP14 targets was showed. The results for multimers were consistent, indicating a higher quality of protein models was achieved through custom templates along with further recycle. The reason why the number of high-quality AF2M models is lower than the number of models that could not be predicted with AF2M is that the DeepMind group did not participate in CASP14 for multimeric targets. Therefore, AF2M models were obtained by us using the updated version of AF2 (AF2_Advanced) available at that time. AF2M ranks the models based on pTM-score, however, TM-score was used for the calibration of AF2M (Evans et al., 2022), while IDDT score was employed for the calibration of the first version AF2 (AF2_Advanced) (Jumper et al., 2021a; Tunyasuvunakool et al., 2021). As a result, both AF2M and non-AF2M models may demonstrate a different trend in terms of rank-1 model quality.

Examples of the refinement effect of the recycling process on models based on the quality scores are shown in Figure 3.2. AF2M for T1078 in Figure 3.2A exhibited a notably lower QS score in the initial model and a considerably higher QS score in the refined model. In addition, when comparing other two baseline and refined models shown in Figures 3.2B and 3.2C, all tree quality scores was substantially improved. Interestingly, it was observed remodelling rather than more effective improvement in model quality as seen in T1078 in the Figure 3.2. It appears that the quality of initial model can affect the rate of improvement after an increase in cycles, potentially leading to divergence from the original structure.

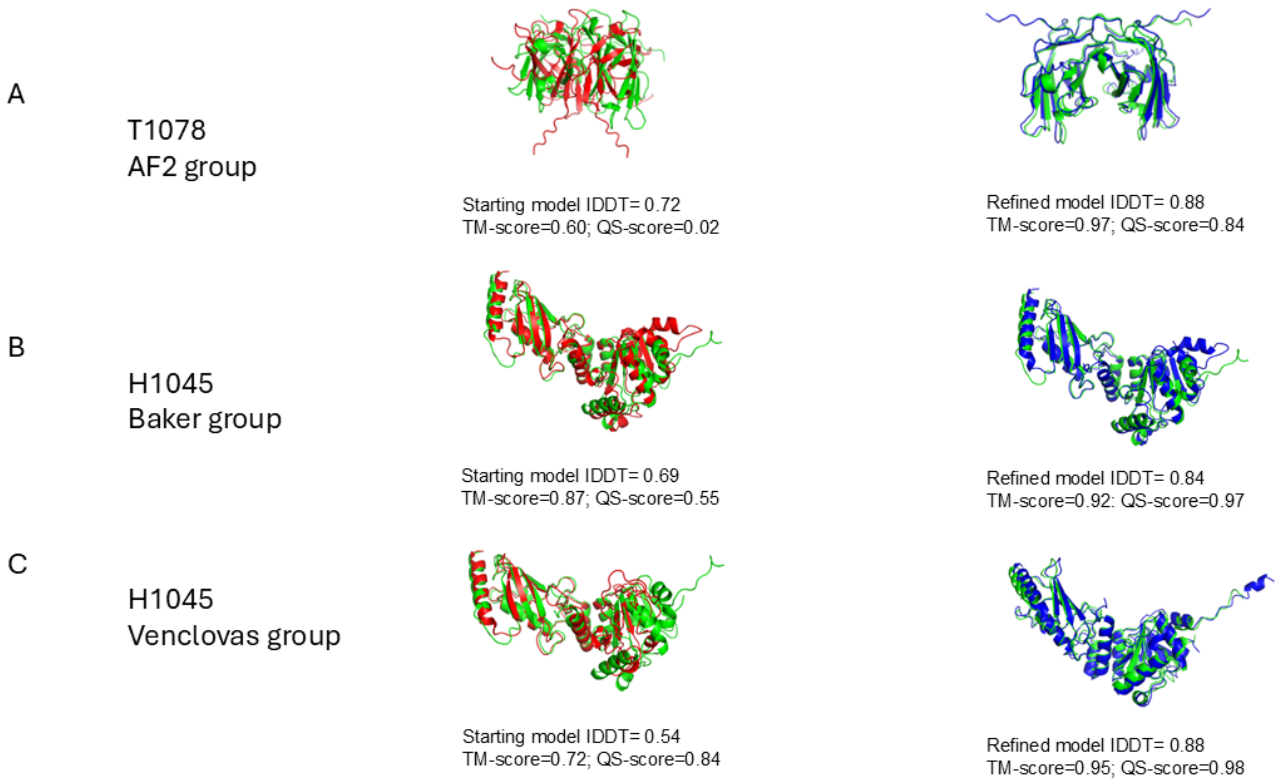


Figure 3.2 Example of the refinement effect of the recycling on three CASP14 targets.
The comparison of the refined multimeric models and the starting models with the observed structures in the superposition way. The starting superposition models are located in the first column, while the refined superposition models are situated in the last column. In the columns, the figures of the starting models (red), the refined figures generated by AF2M (blue) and the reference structures (cyan) were represented. These figures were generated by PyMOL.

The scatter plot in Figure 3.3A shows the improvement conditions based on the initial models (baseline). Figure 3.3B indicates that the majority of models (86%) exhibit improvement according to TM-scores. When compared based on the number of recycles, 87%, 85%, 90%, 84% of targets were improved after recycle-1, 3, 6, 12, respectively (See Appendix Figure S.3). These results suggest that further recycling can improve starting models for the majority of target proteins in terms of TM-scores. While non-AF2 models generally showed more pronounced improvements in the overall structures, the AF2M models exhibited the poorest performance improvement for all number of recycles. Among these structures are H1045, H1065, and T1054 CASP14 models. Since AF2 did not participate in multimer category of CASP14, there is no independent validation of its performance. The recycled AF2M models for H1065 and T1054 fell below baseline. These targets were classified as difficult by CASP14 assessors. Additionally, earlier versions of AF2M show decreased predictive power beyond two chains and T1054 has a tetrameric structure.

Figure 3.3B illustrates that the cumulative improvement per group in model quality exhibited a non-linear trend with increasing numbers of recycles; higher recycle numbers (>3) did not consistently result in greater improvement across all model types. Nevertheless, models from almost all groups except for Takeda-Shitaka and Baker demonstrated noticeable improvement with further recycling. The data show a more pronounced improvement in non-AF models, and this can be attributed to two factors. Firstly, the initial lower quality of the baseline template models offers more room for improvement. Secondly, there's the likelihood of a certain degree of remodelling occurring during the recycling process. It is imperative to ascertain the extent of this remodelling and strive to illustrate that substantial improvement occurred solely through recycling. To determine whether the improvement comes from the template structures themselves or from evolutionary information, a direct comparison between SS recycling and MSA recycling was conducted.

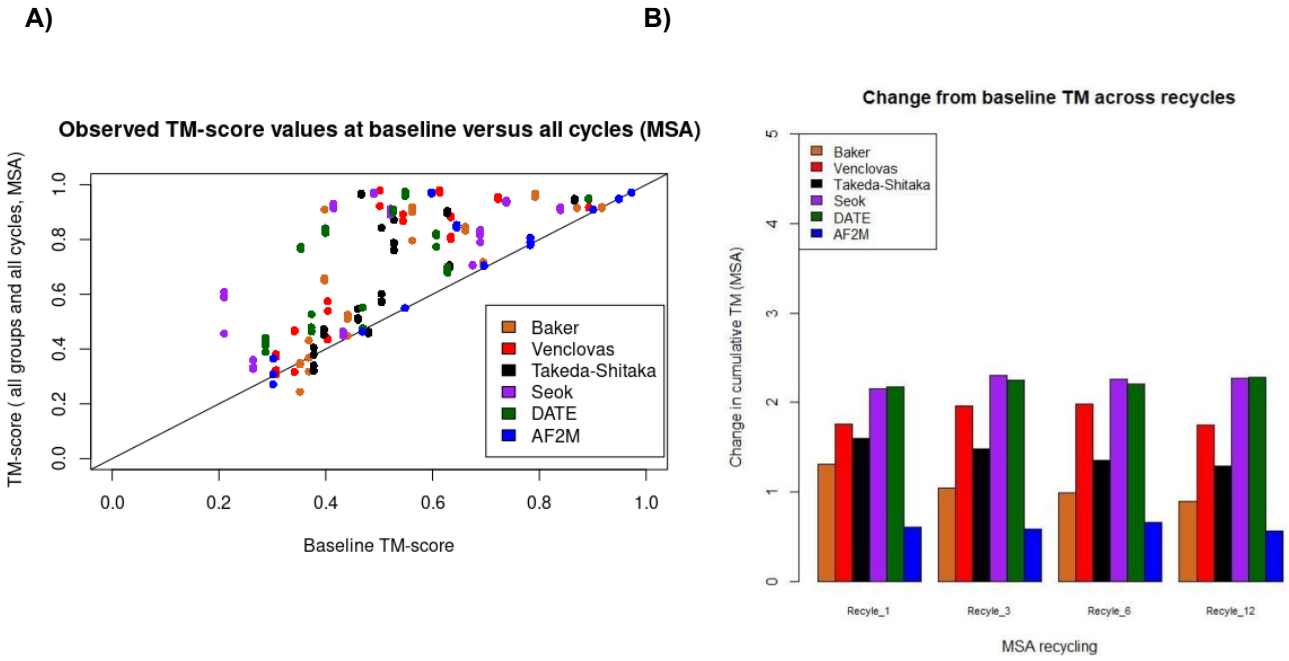


Figure 3.3 A comparison of the observed and baseline TM-scores for the CASP14 models during all recycles and based on group.

A) Scatter plot showing the improvement of models following recycling. The plot compares the observed TM-scores for the improved models (y-axis) versus the baseline TM-scores (x-axis) for CASP14 models generated during all recycles (1-3-6-12) for six group models using MSAs in the AF2M recycling process. The minimum value for TM-scores is 0, while the maximum value is 1. **B)** Bar charts representing the change of cumulative in observed TM-scores generated from the baseline models and the models generated through varying numbers of recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda_Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. This scatter plot was drawn using R.

The scatter plot in Figure 3.4A shows the improvement conditions based on the initial models (baseline). Figure 3.4B indicates that the majority of models (79%) show enhancement according to the IDDT scores. When compared based on the number of recycles, 75%, 78%, 78%, and 85% targets were improved after recycle-1, 3, 6, and 12, respectively (See Appendix Figure S.4). These results suggest that further recycling can improve initial models for the majority of target proteins in terms of the IDDT scores. When the Takeda-Shitaka group models were subjected to further recycling, the modelled structures exhibited the most pronounced deterioration. The Takeda-Shitaka models which demonstrated that the most deterioration were for three targets (T1070, T1073, T1083).

Figure 3.4B illustrates that the cumulative enhancement per group in model quality exhibited a linear trend (up to 6 recycle) with increasing numbers of recycles. Even though higher recycle numbers (>1) consistently resulted in greater improvement across all model types, models from all groups except for Baker demonstrated noticeable improvement with 6 and 12 recycles. Nevertheless, deterioration in AF2M models was observed with further recycling (Cumulative IDDT score for 6 recycle > cumulative TM-score for 12 recycle). There was the deterioration for AF2M models in cycle 1, yet these models were improved with further recycling (> 3).

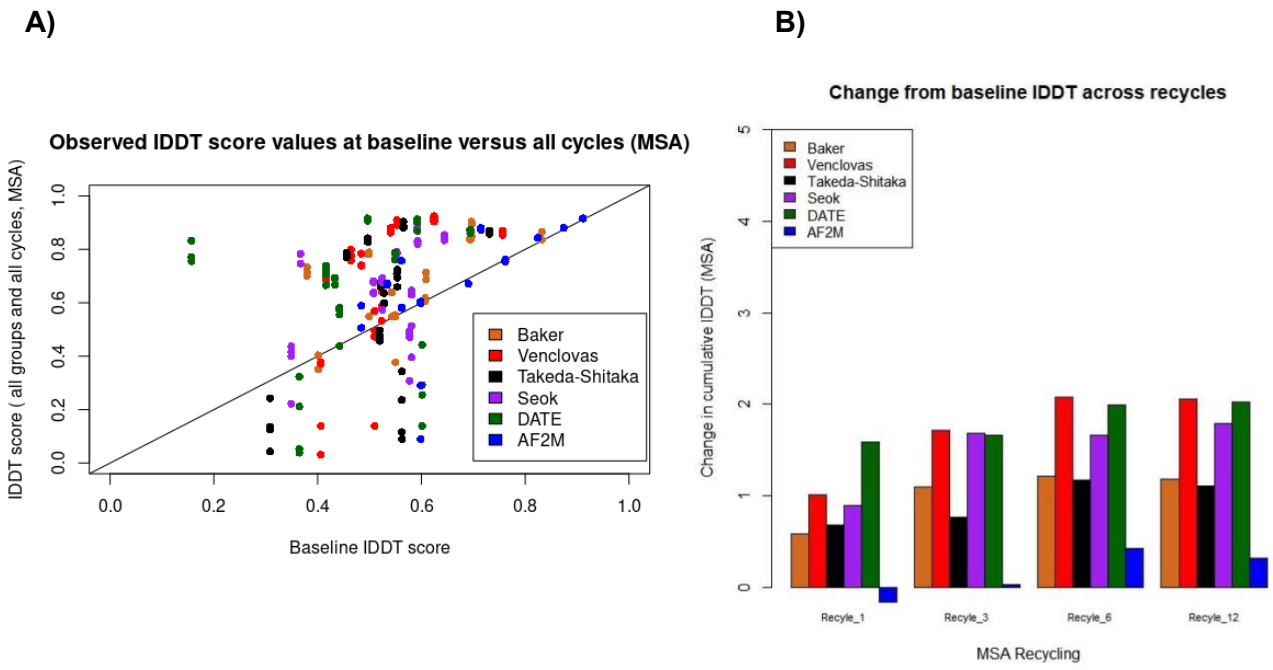


Figure 3.4 A comparison of the observed and baseline IDDT scores for the CASP14 models during all recycles and based on group.

A) Scatter plot showing the improvement of models following recycling. The plot compares the observed IDDT scores for the improved models (y-axis) versus the baseline IDDT scores (x-axis) for CASP14 models generated during all recycles (1-3-6-12) for six group models using MSAs in the AF2M recycling process. The minimum value for IDDT is 0, while the maximum value is 1. **B)** Bar charts representing the change of cumulative in observed IDDT scores generated from the baseline models and the models generated through varying numbers of recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda-Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. This scatter plot was using R.

The scatter plot in Figure 3.5A shows the improvement conditions based on the initial models (baseline). Figure 3.5B indicates that the majority of models (85%) exhibit enhancement according to QS-scores. When compared based on the number of recycles, 85%, 83%, 85%, and 87% of targets were improved after recycle-1, 3, 6, 12 (See Appendix Figure S.5). This result suggests that further recycling can improve starting models for greater part of target proteins in terms of QS-scores. When the Takeda-Shitaka group models were subjected to all four recycle, the modelled structures exhibited the most pronounced deterioration. The Takeda-Shitaka models which demonstrated the most deterioration was for three targets (T1054, T1070, T1073). Figure 3.5B illustrates that the cumulative refinement per group in model quality exhibited a non-linear trend with increasing numbers of recycles; higher recycle numbers (>1) did not consistently resulted in greater improvement across all model types. The improvement in models for Baker, Venclovas, and AF2M groups were approximately same, while there was an improvement in the models for Seok and DATE groups. However, only one group (Takede-Shitaka) belonged to the deterioration models after recycling (<3). The main reason for the increased deterioration of the QS-score for Takeda-Shitaka may lie in the utilization of proper monomers first and then template-based docking, relying on TM-align and residue-residue score (CAB-align) (Terashi & Takeda-Shitaka, 2015). The other methods incorporate both template docking and MSA searching (https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf). Interestingly, cumulative change in QS-score for the two lower quality models except for AF2M models was observed greater, which is likely to search better conformational structure for lower model whose structure is not close to natural structure in the sampling space.

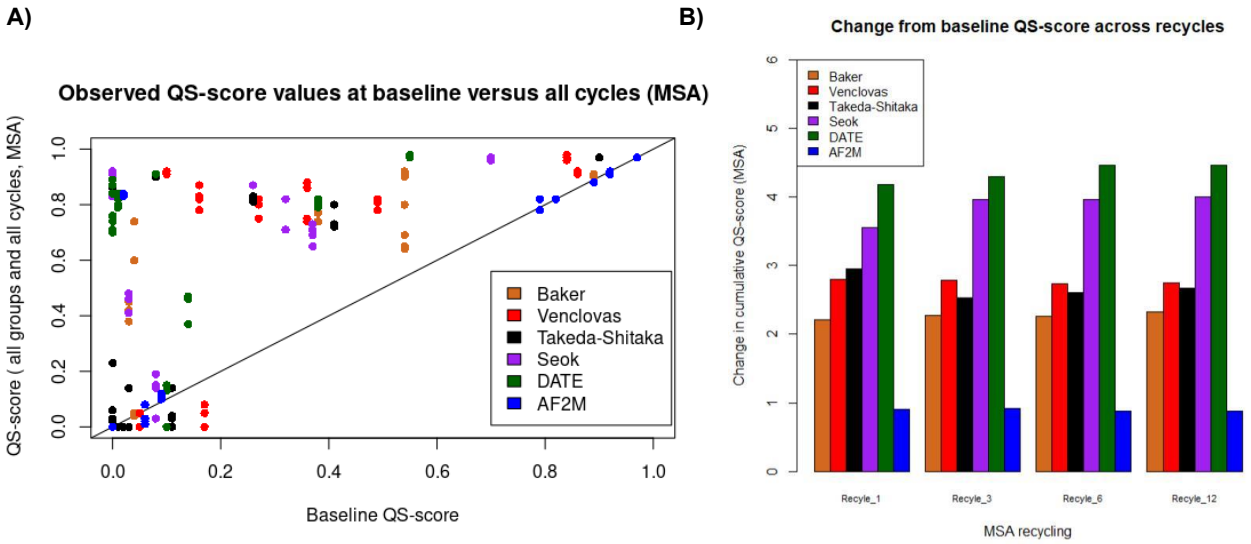


Figure 3.5 A comparison of the observed and baseline QS-scores for the CASP14 models during all recycles and based on group.

A) Scatter plot showing the improvement of models following recycling. The plot compares the observed QS -scores for the improved models (y-axis) versus the baseline QS-scores (x-axis) for CASP14 models generated during all recycles (1-3-6-12) for six group models using MSAs in the AF2M recycling process. The minimum value for QS-score is 0, while the maximum value is 1. **B)** Bar charts representing the change of cumulative in observed QS-scores generated from the baseline models and the models generated through varying numbers of recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda_Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. This scatter plot was drawn using R.

All in all, the scatter plots and bar charts in Figure 3.3 to 3.5 indicate that the majority of models exhibit enhancement in terms of three quality scores, when the further recycling was applied. However, it is evident once again that the improvement in IDDT scores (Figure 3.4A) is less uniform rather than TM-scores (Figure 3.3A). Figure 3.3 highlights a great improvement in TM-scores for both AF2M and non-AF2M models, with fewer models experiencing decrease in their own quality following increases in the recycling. The output models generated by AF2M tend to have higher TM scores because it selects a best model as rank-1 based on the ipTM scores, making it more likely to choose models with lower IDDT scores. After a comparison of structural differences between models and experimental observed protein structures with the help of TM-score, IDDT, and QS-score, Molprobity score was employed to evaluate geometric form for models.

MolProbity score (Chen et al., 2010) is calculated by analysing bond length, torsion angles, atom contacts in the structure, atom clashes, and sidechain rotamers, which is frequently used to analyse the geometric validity of experimentally observed structures before they get stored in the PDB. A lower MolProbity for model implies it has greater all-atom quality and belongs to a more natural protein structure. Figure 3.6 demonstrates a majority of CASP14 models exhibit that Molprobity scores were around 2 and lower than baseline after four recycling (1-3-6-12), indicating clashes, particularly between sidechains, which are reasonably acceptable at this level. Both IDDT and Molprobity scores, being independent of the superposition state, were lower compared to the TM-score (See Figures 3.3A and 3.4A), further supporting the abundance of clashes between sidechains. It can be highlighted that the CASP14 models obtained are unrelaxed form, hence, it is likely that there can be a decrease in Molprobity scores upon the application of relaxation methods. Nevertheless, rather than relying solely on energy minimization methods or recycling, physical approaches aimed at broadening conformational sampling in modelled structures could significantly reduce clashes between sidechains in AF2M.

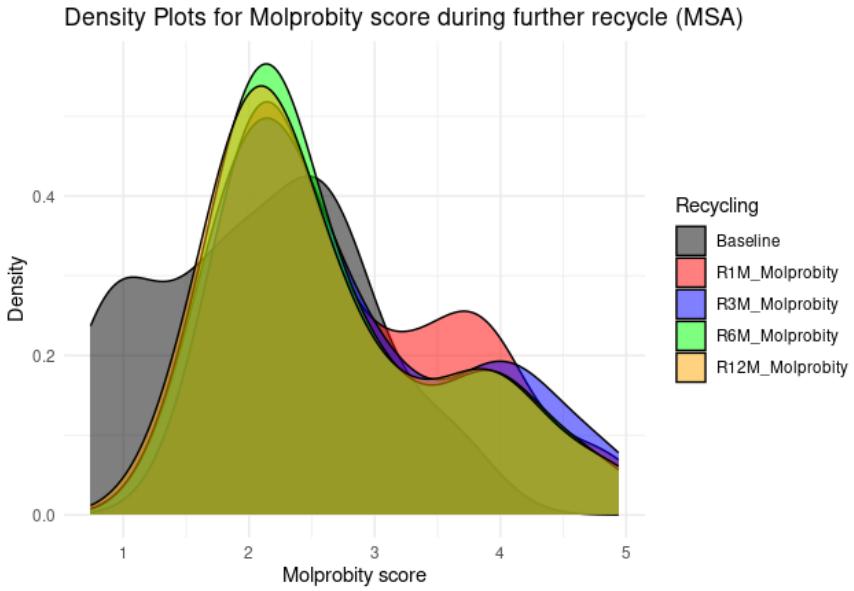


Figure 3.6 A comparison of the observed and baseline Molprobity scores for the CASP14 models after recycling.
The density plot showing the Molprobity scores (lower Molprobity scores are better) for the CASP14 models generated by AF2M using MSA, with red for cycle 1 (R1M), blue colour for cycle 3 (R3M), green colour for cycle 6 (R6M), magenta colour for cycle 12 (R12M) and black colour for baseline as starting model. This plot compares the geometric correctness rate for models after recycles, without using experimentally observed protein structure. Molprobity scores were generated by http://molprobity.biochem.duke.edu/. The density plot was drawn using R.

Unlike IDDT score, TM-scores and QS-scores for SS methods in Appendix Figure S.6 show similarity compared to that of the MSA method. This can vary depending on the content of the structure, as a significant portion of the secondary structures is predicted by AF2M necessarily relying on MSA-derived information. The increase in both scores with the SS method also indicates effective research even without MSA, as the provided models serves as a template. Structural representations have especially proven to be more effective than sequence-based methods, as information obtained by MSA of protein sequence can lead to false positives, potentially resulting in the loss of valuable data. Interestingly, the template structure of higher quality did not generate better quality structures.

It was observed that TM-scores and QS-scores for models generated by MSA methods have shown greater improvement following further recycle (See Figure 3.3B, 3.5B). Interestingly, models generated by SS methods exhibited a non-linear development when to apply advanced recycling, except for IDDT score (See Appendix Figure S.6 (middle)). Furthermore, interface scores of models produced with SS methods, except for the top two groups (Baker and Venclovas groups), have also shown greater development after further recycling. However, when compared based on the group level, it is evident that the lower the initial structure quality, the more susceptible the structure is to improvement. Models of lower quality (both non-AF2M and AF2M models) are more prone to improvement with SS methods. This may be due to the constraining effect of coevolutionary information derived from MSA. The same findings were applicable for interface region of models. The only difference was that the higher the initial quality of the structure, the greater the difference

between MSA and SS methods. This phenomenon may stem from MSA subsampling, which is more effective in generating structural information for interface regions.

It was also observed that models generated by the SS method had lower IDDT values following further recycling. In fact, it shows that there was no improvement in IDDT scores for models, but rather a deterioration. The reason for the decrease in the IDDT scores, despite the improvement in TM-scores and QS-scores without utilizing coevolution information from MSA, could be attributed to be the custom template. The Molprobity scores of models generated by AF2M using the SS method were higher compared to the scores for models generated by AF2M using the MSA methods. A greater number of models had scores between 4 and 5, which may not be acceptable score for protein structure quality score (See Appendix Figure S.7).

Table 3.3 compares the models generated using further recycling to the baseline models which were used as custom templates. For the non-AF2M models, statistically significant enhancements were observed across three quality scores (TM-score, IDDT, and QS-score) after four different recycles when utilized the MSA method ($p > 0.05$). In contrast, the AF2M models consistently exhibited significant improvements in the IDDT (recycle 6 and 12) and TM-score (recycle 1 and 6) ($p < 0.05$). Additionally, when employed the SS method, significant improvements were evident in both TM-score and QS-scores for the non-AF2M models generated by four various recycles ($p < 0.05$). There was only a statistically significant difference between the TM-scores for the AF2M models generated by three type of recycling and the baseline models. Analysing the p-values for all models and scores appears that both recycles 6 and 12 may be optimal parameters for achieving higher quality multimeric protein structures.

Table 3.3 A statistical comparison of the cumulative scores for the CASP14 models versus the baseline models after recycling.

Computed p-values resulting from the Wilcoxon signed-rank test for IDDT scores (a), TM-scores (b), and QS-scores (c) in relation to both baseline and recycled models, considering both non-AF2M and AF2M models, specifically for CASP14 multimeric targets.

a		IDDT-scores						
Models	Recycle type	Baseline to 1 recycle	1 recycle to 3 recycles	Baseline to 3 recycles	3 recycles to 6 recycles	Baseline to 6 recycles	6 recycles to 12 recycles	Baseline to 12 recycles
AF2M	MSA	1.106e-1	5.203e-1	1.795e-1	7.679e-2	**4.157e-2**	1.106e-1	**5.146e-2**
	SS	9.966e-1	**4.157e-2**	9.369e-1	6.177e-2	9.369e-1	9.736e-1	9.369e-1
non-AF2M	MSA	**3.748e-3**	**4.267e-05**	**1.398e-05**	**6.915e-3**	**1.02e-06**	9.565e-1	**4.935e-07**
	SS	8.492e-1	**1.475e-2**	5.116e-1	1.611e-1	4.197e-1	**1.013e-2**	3.285e-1
b		TM-scores						
AF2M	MSA	**3.327e-2**	2.704e-1	6.314e-2	6.314e-2	**2.075e-2**	8.205e-1	1.54e-1
	SS	**5.146e-2**	1.795e-1	**2.075e-2**	6.201e-1	**1.247e-2**	6.202e-1	**2.075e-2**
non-AFM2	MSA	**2.066e-0**9	2.715e-1	**1.453e-09**	9.428e-1	**2.926e-09**	9.858e-1	**6.889e-09**
	SS	**3.338e-3**	**3.97e-3**	**5.516e-4**	7.458e-1	**1.367e-4**	3.75e-1	**2.946e-4**
c		QS-scores						
AF2M	MSA	4.161e-1	5.724e-1	1.976e-1	4.27e-1	5e-1	5e-1	5e-1
	SS	7.992e-1	**5.017e-2**	5e-1	1.855e-1	3.422e-1	8.618e-1	3.375e-1
non-AFM2	MSA	**1.577e-07**	2.268e-1	**2.578e-07**	1.575e-1	**1.09e-07**	2.326e-1	**6.799e-08**
	SS	**3.491e-2**	**1.118e-2**	**4.175e-3**	3.083e-1	**2.548e-3**	2.406e-1	**4.089e-3**

*MSA: Multiple Sequence Alignment, SS=Single sequence. Ho: Recycling, as stated by column, produces models of equivalent or lower quality than the template structures used as baseline templates, or the models generated in previous cycles. H1: Recycling, as stated by column, produces models of higher quality than the template structures used as baseline templates, or the models generated in previous cycles. P-values ≤0.05 indicate there is significant differences, as in highlighted in bold. The one-tailed Wilcoxon signed-rank tests were used to measure the (A) IDDT scores, (B) TM-scores, (C) QS-scores for 10 AF2 models and 50 non-AF2 models from various CASP14 targets.

In Table 3.4, concerning all quality scores, the baseline models were refined following further recycling. However, for the TM-scores and IDDT scores, deterioration was observed after reaching recycle 6. Conversely, for the QS-score, it appears that the models with better interface quality scores could be obtained when run AF2M using further recycling. Given all the scores and statistical analysis, a specific recycle number, such as 6, can be suggested to achieve an improved multimeric structure. It is noteworthy that during the training of neural networks, increasing the number of hidden layers and nodes can lead to issues such as trapping in local minima (Wang & Cao, 2018), which can persist during inference. Since AF2M uses a DNN, this might explain why AF2M does not always

generate a more refined protein structure after each cycle. When model improvement was analysed based on the type of multimeric structure, the heteromeric models were observed to exhibit better improvement compared to the homomeric models (See Appendix Figure S.8).

Table 3.4 A comparison of the cumulative scores for the CASP14 models versus the baseline models after recycling.

The table showing the cumulative scores of the starting models, known as, baseline and of models obtained after four various recycling (1-3-6-12) using MSA and SS methods. The cumulative scores for TM-scores, IDDT, QS-scores highlighted in red represented the reduction in cumulative score for TM-scores and IDDT scores when applied to further recycling (between cycle 6 and cycle 12), unlike other quality scores. It is significant to note that the lower Molprobity score presents higher quality model.

Method	Baseline	Cycle-1	Cycle-3	Cycle-6	Cycle-12
			∑TM-score		
MSA	34.205	43.797	43.814	43.958	43.431
SS	31.092	34.559	35.659	36.292	28.663
			∑IDDT		
MSA	33.315	37.911	40.510	41.944	41.897
SS	29.906	26.192	27.554	28.219	36.194
			∑QS-score		
MSA	15.71	32.23	32.39	32.52	32.7
SS	12.62	16.18	18.94	19.61	19.74
			∑Molprobity score		
MSA	124.08	169	166.44	160.03	159.88
SS	131.96	202.55	199.72	198.72	197.12

3.3.2 The impact of custom template recycling with MSAs on multimeric CASP15 models

The impact of custom template recycling was measured using four quality scores (TM-score, lDDT, QS-score, and DockQ_wave). Different from the quality scores of the CASP14 models, DockQ_wave score was included as a quality score of the CASP15 models, since DockQ_wave score was not introduced at the time of the analysis of the CASP14 targets. The DockQ_wave score was released after the CASP15 competition, which is more sensitive than the QS-score (Kryshtafovych et al., 2023b; Studer et al., 2023). Notably, the CASP15 competition involved more multimeric targets and thus focused on the modelling of the multimeric structures. The CASP14 competition included 84 monomeric and 29 multimeric targets (https://predictioncenter.org/casp14/numbers.cgi), while the CASP15 competition included 81 monomeric and 47 multimeric targets (https://predictioncenter.org/casp15/numbers.cgi). Similar to the CASP14 targets, 576 models from six groups which participated in the CASP15 competition were used to observe any improvements following further recycling. Out of 576 models, 262 models (45%) in terms of TM-score and 115 models (20%) in terms of lDDT were improved, while 153 models (27%) in terms of QS-score and 197 models (35%) in terms of DockQ_wave score showed improvement after further recycling. When it comes to the SS method, 231 models (40 %) and 98 models (17%) out of 576 models showed improvement in terms of TM-score and lDDT, respectively. Meanwhile, 142 models (25%) and 135 models (25%) were refined in terms of QS-score and DockQ_wave score, respectively, following further recycling (See Appendix Figure S.9 - S.10). These results indicate that when compared to the MSAs with the SS approaches, using MSAs can be more effective, along with further recycling.

Three CASP15 models were selected as examples to demonstrate the improvements in the quality scores, as shown in Figure 3.7. The H1151 model belongs to the best group (Zheng group), whilst the H1143 models belong to NBIS-AF2-Multimer and our group (MultiFOLD). All three models demonstrated an increase in all four quality scores after recycling. However, the rate of increase in all scores for the NBIS-AF2-Multimer and Zheng models was lower than that of the MultiFOLD model. Interestingly, a similar increase was observed in scores for other models which may be explained as the consequence of limited conformational sampling. Specifically, when a starting model is very close to natural structure of a protein and within the conformational space, the rate of increase in scores may be affected. This increase could either be very minimal, or in some cases, model deterioration may even be observed. The Zheng model was the best models, whereas the NBIS-AF2-Multimer model may have evaluated as the high-quality model when used as custom template, given that these models were already generated by AF2M before. Consequently, it is highly possible to deviate from the correct refinement path while searching through the conformational space.

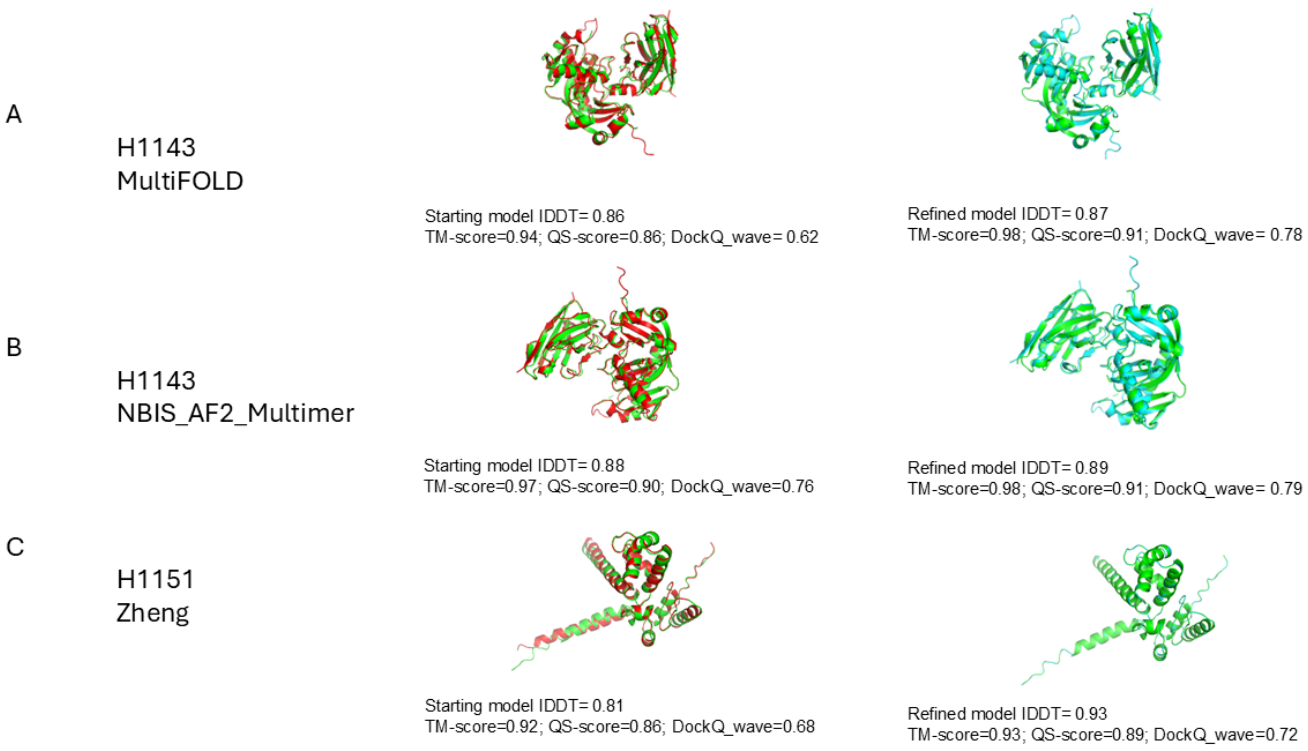


Figure 3.7 Examples of the refinement effect of the recycling on three CASP15 targets.
The comparison of the refined multimeric models with the starting models. The starting models are in the first column, while the refined models are situated in the second column. The figures in both columns were coloured based on the reference and modelled structures. The columns represent the alignment figures of the starting models (red), the refined figures generated by AF2M (magenta) and the reference structures (cyan). The figures were generated by PyMOL.

The scatter plot in Figure 3.8A indicates that less than half of models (45%) exhibit enhancement in terms of TM-score during at least one of the four types of recycling (1, 3, 6 ,12). When analysed based on the number of recycles in Figure 3.9, the percentage of improved models after recycles 1, 3, 6, and 12 were 46%, 45%, 56%, and 55%, respectively. These results suggest that further recycling can improve general structure of target proteins in terms of TM-scores. Interestingly, most of the NBIS-AF2-Multimer models seemed to deteriorate during recycle 1, and the Zheng models showed the least improvement when further recycling was applied. Figure 3.8B illustrates that the cumulative change in model quality based on groups exhibited a linear trend after further recycling; higher recycles consistently resulted in greater improvement in the TM-scores for five group models. The MultiFOLD models only showed improvement during further recycling (≥ 3 recycles). Notably, the H1143 and T1132 models were refined after recycle 1, while the H1166 model was not improved after recycling. It is noteworthy that the cumulative score changes exhibited a negative trend when compared to the quality scores for the baseline models; the difference reduced following further recycling. The improvement of the MultiFOLD models reflects what we expect with traditional refinement. However, the majority of the models used were generated by the most successful groups in the CASP15 competition, which are expected to have the very highest-quality initial scores. It is interesting to note that the homomeric models (48%) exhibited a greater rate of improvement compared to the heteromeric models (41%) (See Appendix Figure S.11(top)).

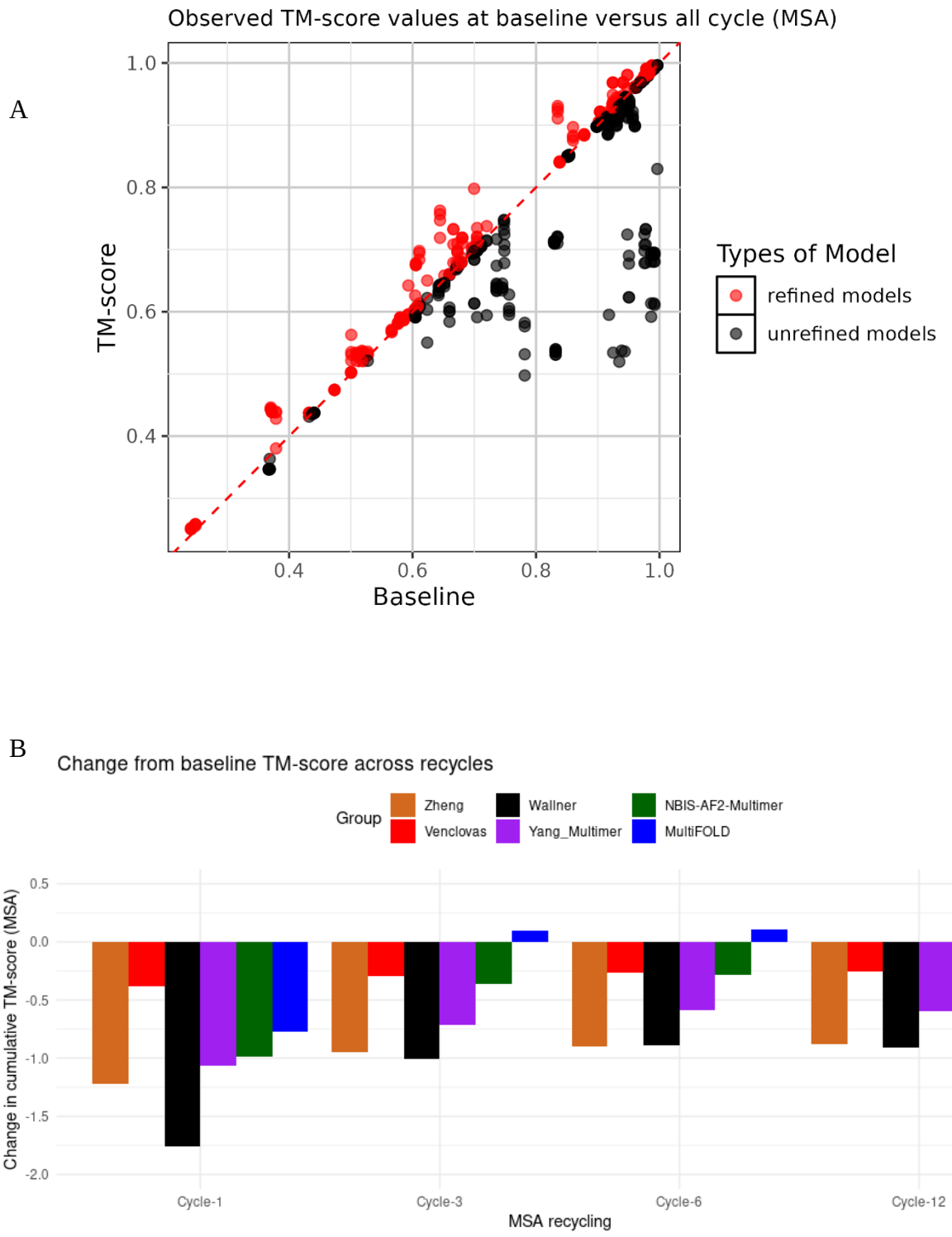


Figure 3.8 A comparison of the observed and baseline TM-scores for the CASP15 models after recycling.

A) Scatter plot representing the comparison of the observed TM-scores for the improved models (y-axis) versus baseline TM-scores (x-axis) for the CASP15 models generated during all recycles (1-3-6-12) for six group models using the MSA method. The minimum value for TM-score is 0, while the maximum value is 1. The red circles represent the refined models, while the black ones represent the unrefined models. **B)** Bar chart representing the cumulative change in the observed TM-scores generated from the baseline models and the models generated by recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. Both the scatter plot and bar chart were drawn using R.

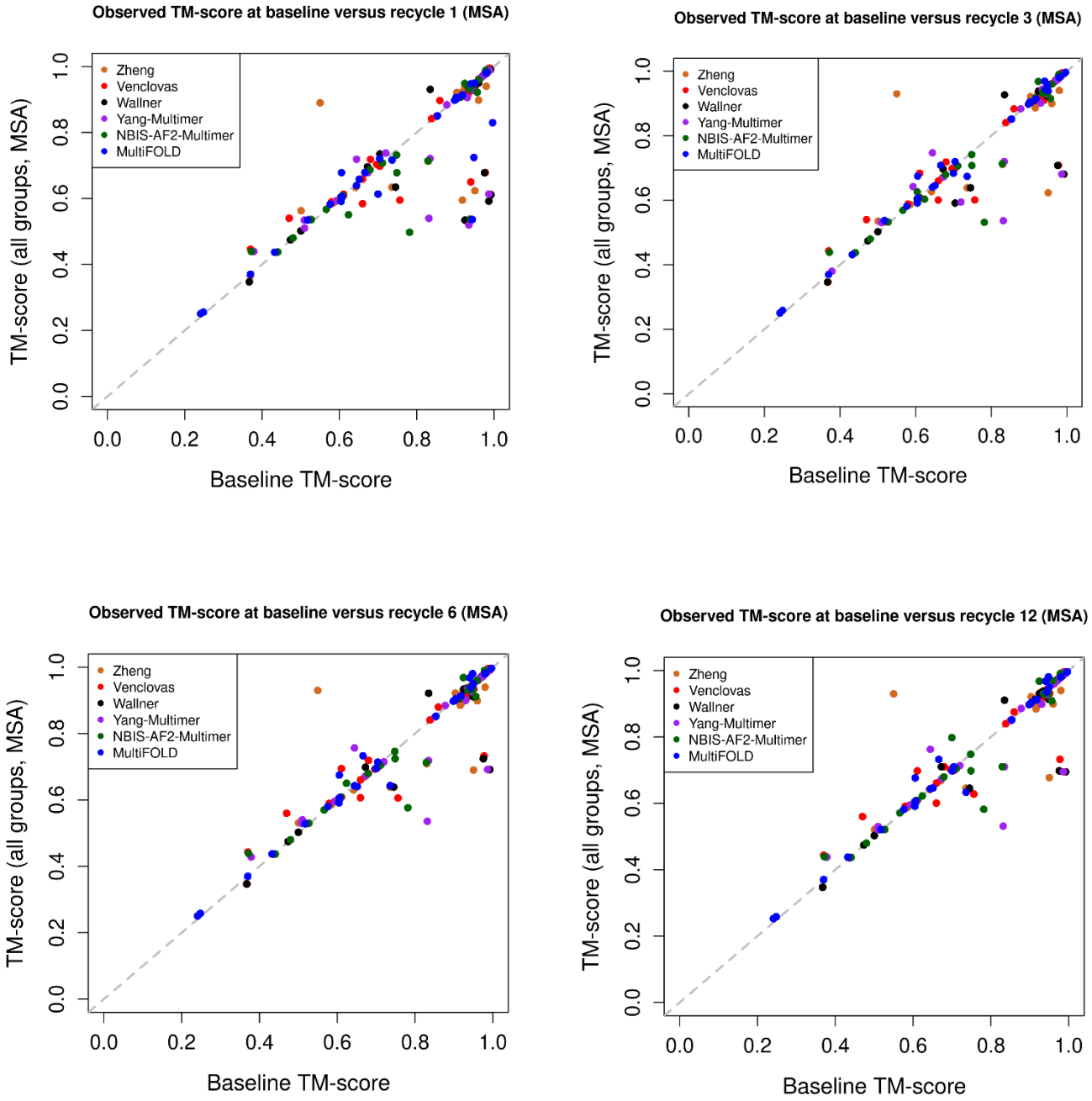


Figure 3.9 A comparison of the observed and baseline TM-scores for the CASP15 models during each recycles (1-3-6-12).

Four scatter plots representing the comparisons of the observed TM-scores for the improved models of six groups (y-axis) versus the baseline TM-scores (x-axis) for the CASP15 models generated during recycles 1 (top-left), 3 (top-right), 6 (bottom-left), 12 (bottom-right), separately, using the MSA method. Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. The scatter plots were drawn using R.

The scatter plot in Figure 3.10A indicates that the majority of models (20%) exhibit enhancement in terms of IDDT score during at least one of the four types of recycling (1, 3, 6 ,12). When analysed based on the number of recycles in Figure 3.11, the percentage of improved models after recycles 1, 3, 6, and 12 were 45%, 55%, 62%, and 63%, respectively. These results suggest that further recycling can improve general structure of target proteins in terms of IDDT score. When the NBIS-AF2-Multimer models were subjected to further recycling, the final models generated by AF2M exhibited the most pronounced deterioration. Figure 3.10B illustrates that the cumulative change in model quality based on the CASP15 groups exhibited a linear trend after further recycling (<3); higher recycle numbers consistently resulted in greater improvement in the IDDT scores for five group models. Even though the cumulative score for the models was less than that of the baseline models, as seen in TM-score, the Yang-Multimer models inclined toward a positive difference in score during recycle 6 and 12. The positive trend observed in the Yang-Multimer models generated by AF2M during further recycling may be attributed to the exclusion of MSA pairing in the methods utilized by Yang-Multimer group (*CASP15 Abstracts*, 2022). This improvement can enhance the heteromeric models, as supported by Bryant *et al.* (Bryant, Pozzati, & Elofsson, 2022). Furthermore, the improved number of the heteromeric models (53%) were higher than the number of the homomeric models (42%) (See Appendix Figure S.11(bottom)). The MultiFOLD models exhibited the lowest negative cumulative change among the other groups, indicating that slightly lower-quality models (compared to the top groups) are more likely to be improved during recycling. This suggest that there may be more room to for improvement of local errors in such models.

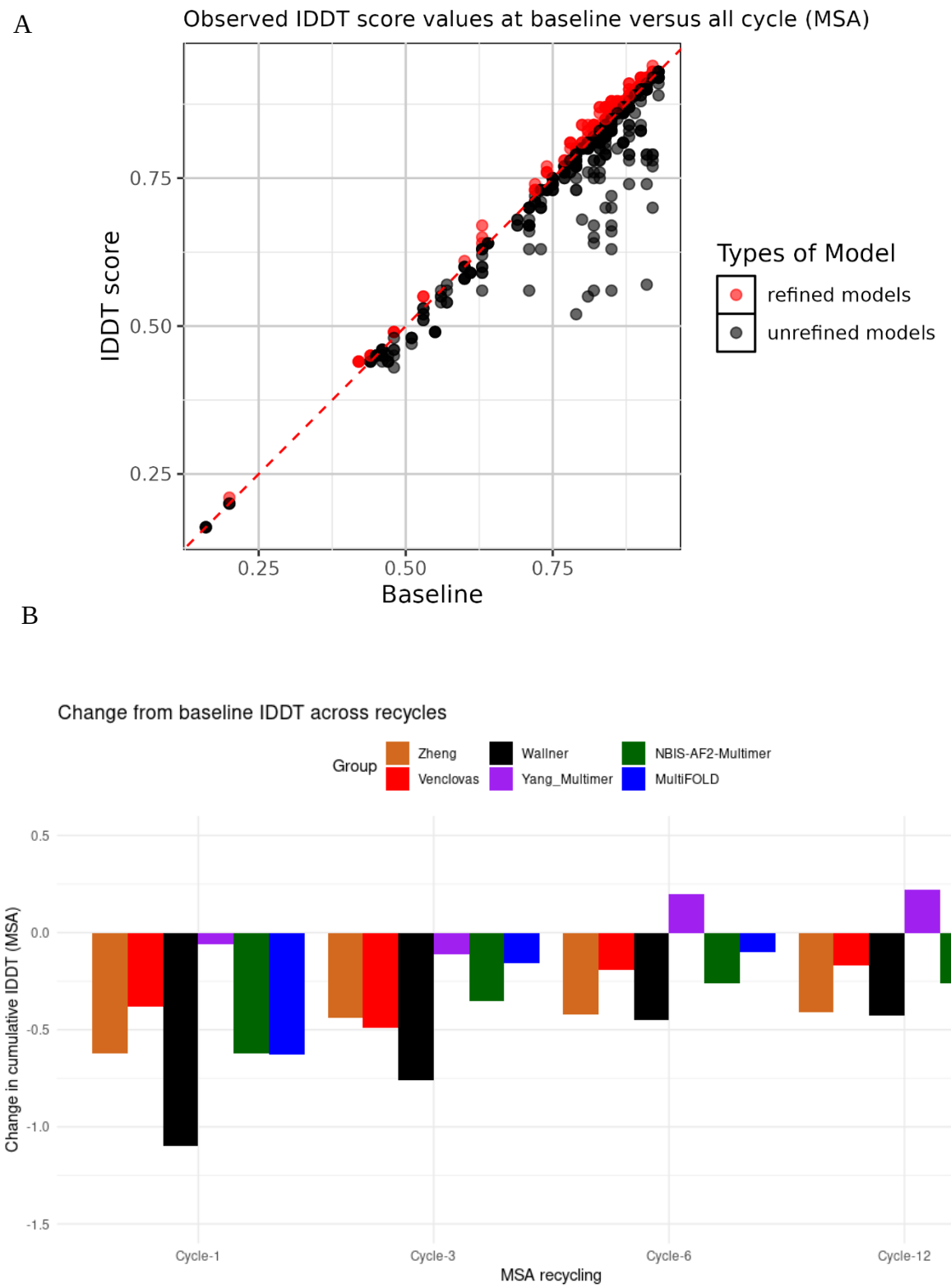


Figure 3.10 A comparison of the observed and baseline IDDT scores for the CASP15 models after recycling.

A) Scatter plot representing the comparison of the observed IDDT scores for the improved models (y-axis) versus the baseline IDDT scores (x-axis) for the CASP15 models generated during all recycles (1-3-6-12) for six group models using the MSA method. The minimum value for IDDT score is 0, while the maximum value is 1. The red circles represent the refined models, while the black ones represent the unrefined models. **B)** Bar chart representing the cumulative change in the observed IDDT scores generated from the baseline models and the models generated by recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. Both the scatter plot and bar chart were drawn using R.

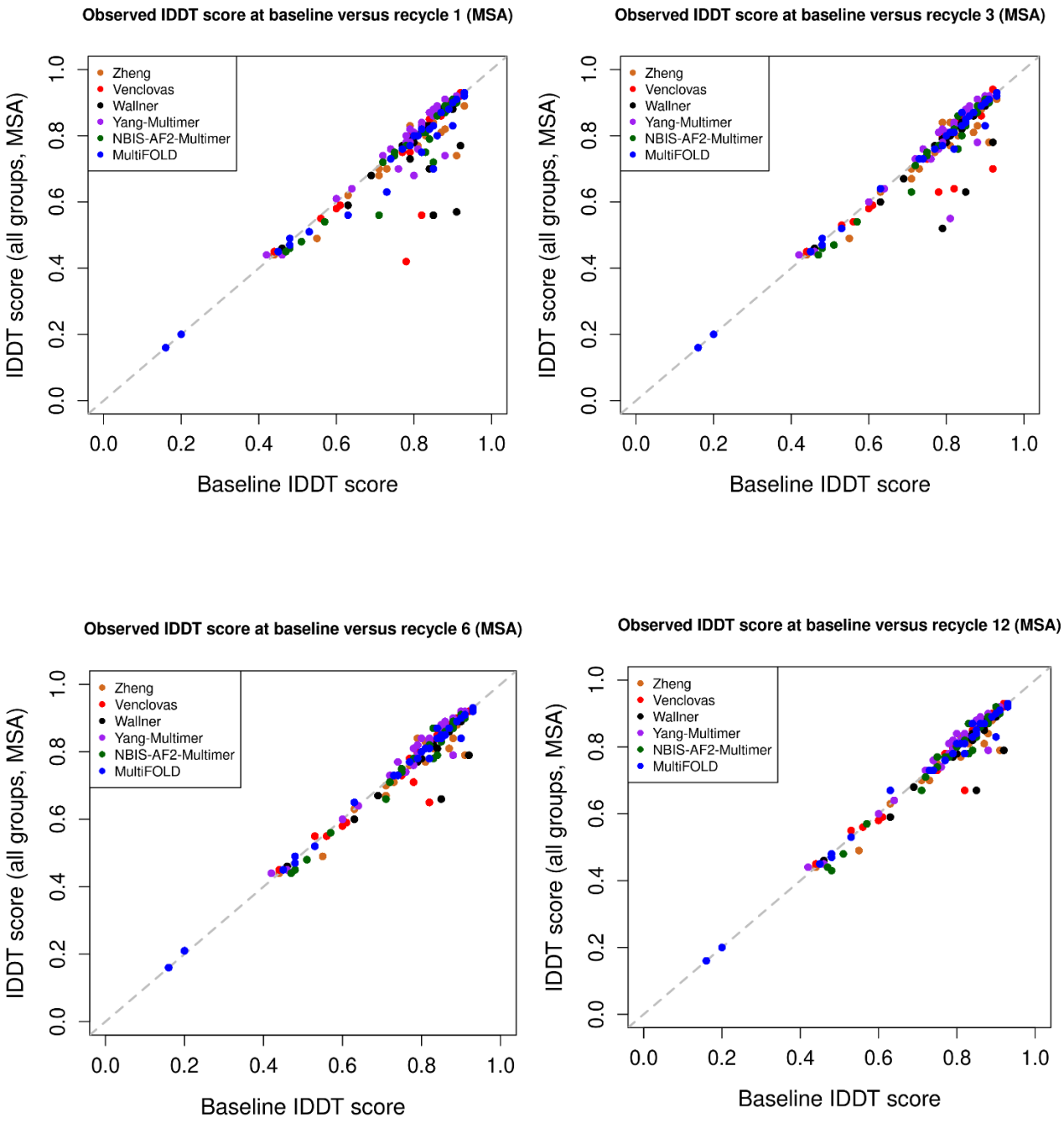


Figure 3.11 A comparison of the observed and baseline IDDT scores for the CASP15 models during each recycles (1-3-6-12).

Four scatter plots representing the comparisons of the observed IDDT scores for the improved models of six groups (y-axis) versus the baseline IDDT scores (x-axis) for the CASP15 models generated during recycles 1 (top-left), 3 (top-right), 6 (bottom-left), 12 (bottom-right), separately, using the MSA method. Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. The scatter plots were drawn using R.

The scatter plot in Figure 3.12A indicates that the majority of models (27%) exhibit enhancement in terms of QS-score during at least one of the four types of recycling (1, 3, 6 ,12). When analysed based on the number of recycles in Figure 3.13, the percentage of improved models after recycles 1, 3, 6, and 12 were 58%, 65%, 70%, and 68%, respectively. The Yang-Multimer models generated by AF2M using recycles 1 and 3 showed more deterioration, whereas the number of Zheng models exhibited consistent improvement during more recycles (recycles 6 and 12). It was observed that after further recycling, the models that initially had very low interface scores were found to have poor interface quality. Furthermore, If AF2M managed to generate higher quality scores for the models than QS-scores for the baseline models, this improvement persisted following further recycling. Nevertheless, if this did not occur during recycle 1, then the interface quality for the models did not improve. Figure 3.12B illustrates that the cumulative change in model quality based on groups exhibited a linear trend after further recycling; higher recycle numbers consistently resulted in greater improvement in QS-scores for all group models. Even though the cumulative QS-score for the CASP15 models exhibited a negative difference, similar to the global quality scores (TM-score and IDDT) for the models, the NBIS-AF2-Multimer and MultiFOLD models showed an inclination towards a positive difference when recycles 3 and more were used. The highest positive difference for both tools was during recycles 6. When the impact of further recycling on homomeric and heteromeric models was analysed, it was observed that a greater number of improvements were evident in the homomeric models (%70), rather than the heteromeric models (61%) (See Appendix Figure S.12(top)).

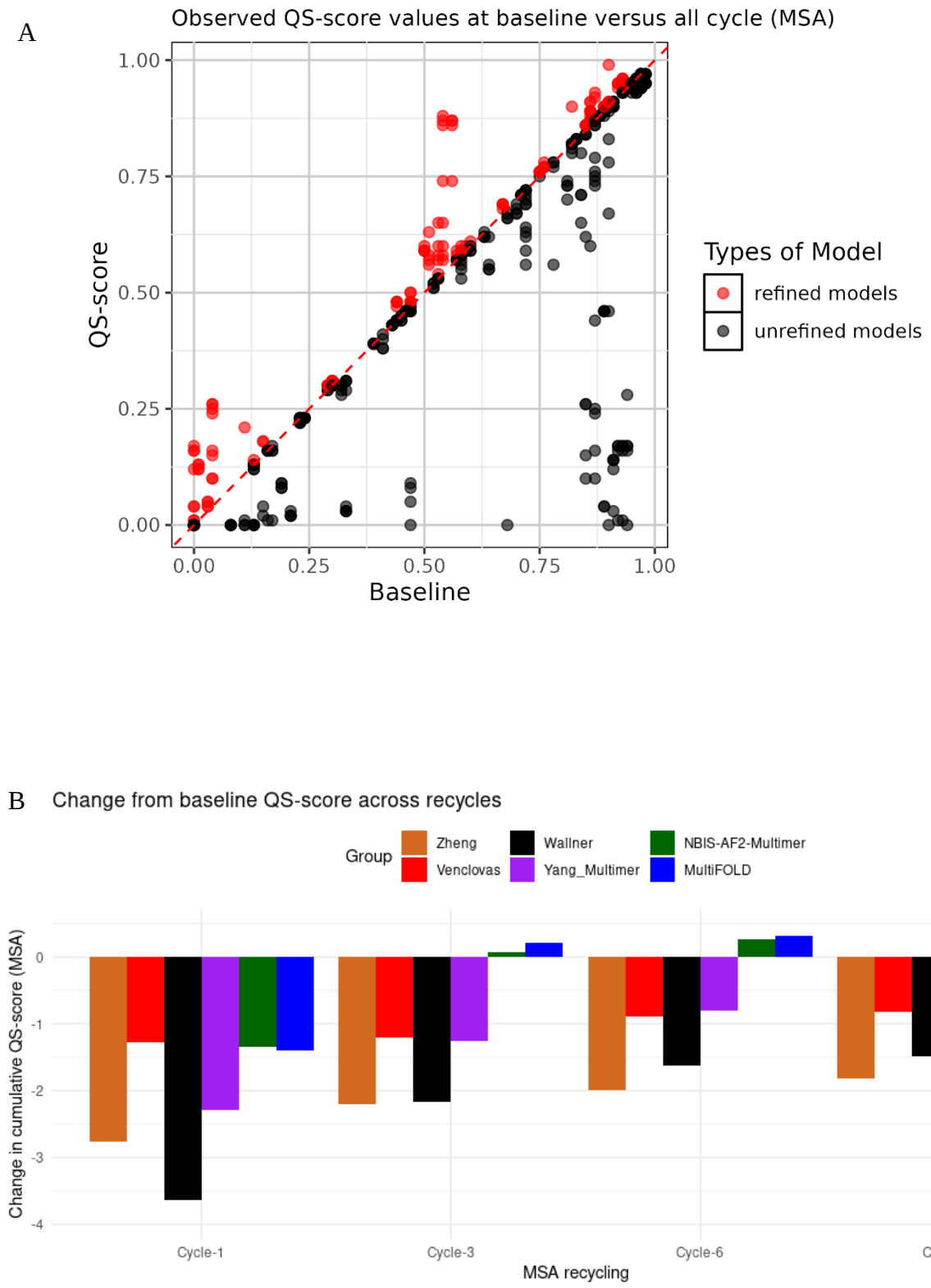


Figure 3.12 A comparison of the observed and baseline QS-scores for the CASP15 models after recycling.

A) Scatter plot representing the comparison of the observed QS-scores for the improved models (y-axis) versus the baseline QS-scores (x-axis) for the CASP15 models generated during all recycles (1-3-6-12) for six group models using the MSA method. The minimum value for QS-score is 0, while the maximum value is 1. The red circles represent the refined models, while the black ones represent the unrefined models. **B)** Bar chart representing the cumulative change in the observed QS-scores generated from the baseline models and the models generated by recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. Both the scatter plot and bar chart were drawn using R.

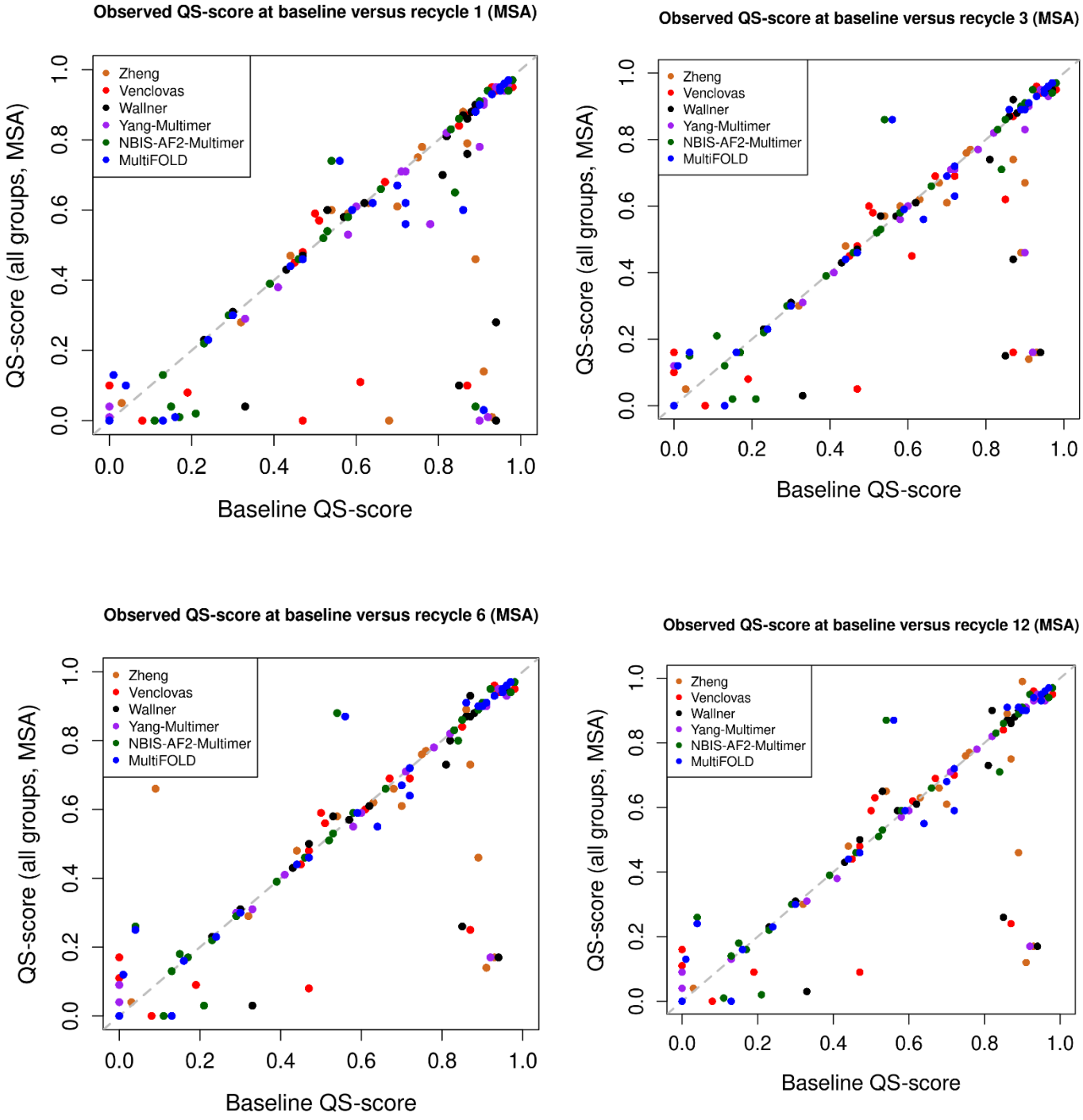


Figure 3.13 A comparison of the observed and baseline QS-scores for the CASP15 models during each recycles (1-3-6-12).

Four scatter plots representing the comparisons of the observed QS-scores for the improved models of six groups (y-axis) versus the baseline QS-scores (x-axis) for the CASP15 models generated during recycles 1 (top-left), 3 (top-right), 6 (bottom-left), 12 (bottom-right), separately, using the MSA method. Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. The scatter plots were drawn using R.

DockQ_wave score (Studer et al., 2023) is an interface quality metric that emerged after the CASP15 competition. Therefore, when analysed the effect of AF2M's recycling on protein model refinement historically, DockQ_wave score alongside QS-score for the CASP15 models was also investigated. The scatter plot in Figure 3.14A indicates that the majority of models (35%) demonstrated refinement in terms of DockQ_wave during at least one of the four types of recycling (1, 3, 6 ,12). When analysed based on the number of recycles in Figure 3.15, the percentage of improved models after recycles 1, 3, 6, and 12 were 57%, 64%, 63%, and 61%, respectively, suggesting that further recycling can refine the general structure of target proteins in terms of the interface quality scores. However, when the Venclovas models were subjected to more than recycles 3, the models exhibited the most pronounced deterioration. In Figure 3.14B even though the cumulative score for the models showed a negative cumulative difference similar to the global quality scores (TM-score and IDDT) and another interface score (QS-score), the MultiFOLD models showed an inclination towards positive cumulative score differences during recycles 3, 6, and 12. The highest positive difference for the MultiFOLD models were observed during recycles 6 and 12. Only the cumulative change in the Venclovas models exhibited an inverse trend among all groups` models. Again, it was observed that a greater number of improvements were evident in the homomeric models (71%), rather than the heteromeric models (51%) (See Appendix Figure S.12(bottom)).

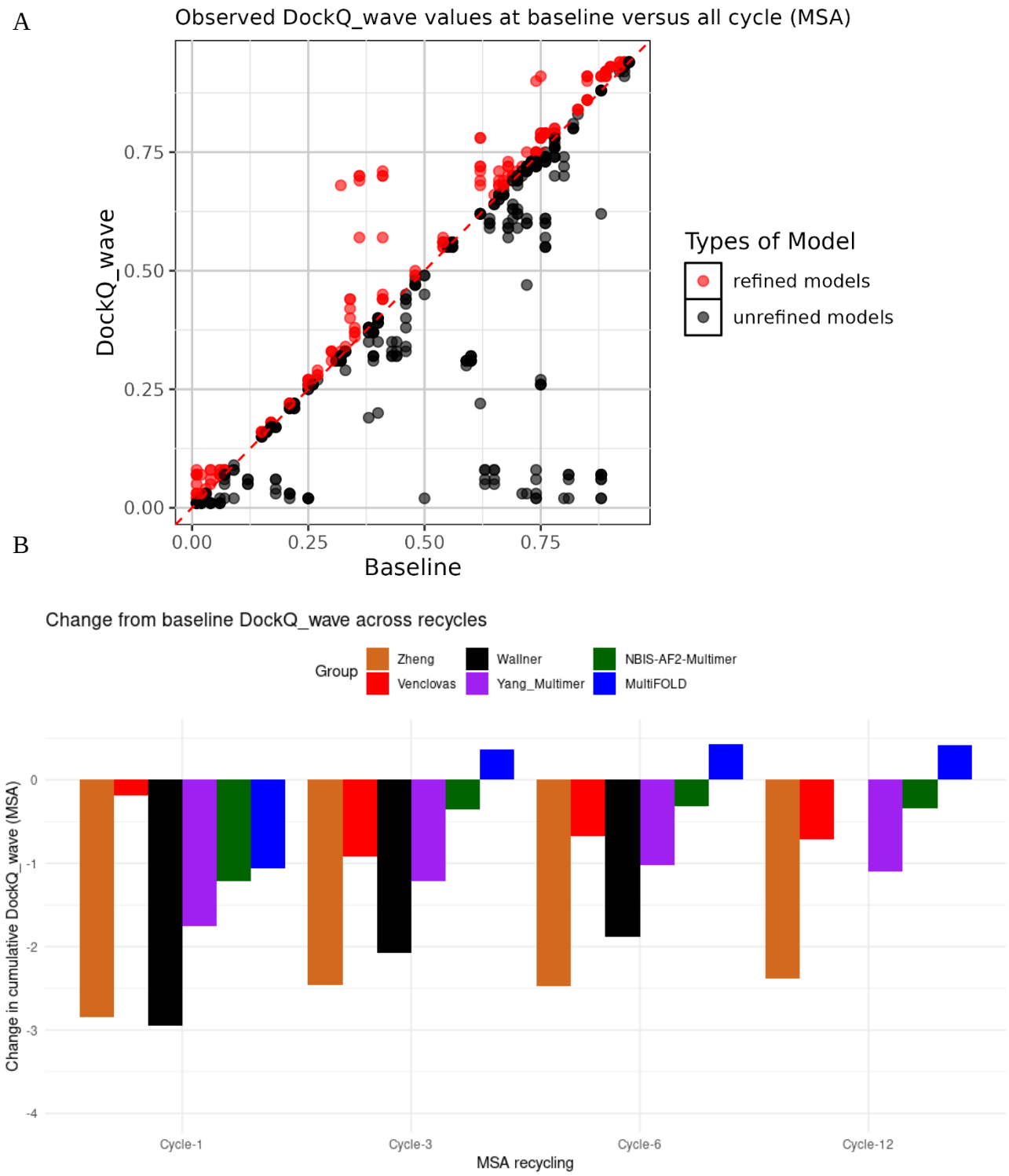


Figure 3.14 A comparison of the observed and baseline DockQ_wave scores for the CASP15 models after recycling.

A) Scatter plot representing the comparison of the observed DockQ_wave scores for the improved models (y-axis) versus the baseline DockQ_wave scores (x-axis) for the CASP15 models generated during all recycles (1-3-6-12) for six group models using the MSA method. The minimum value for DockQ_wave score is 0, while the maximum value is 1. The red circles represent the refined models, while the black ones represent the unrefined models. **B)** Bar chart representing the cumulative change in the observed DockQ_wave scores generated from the baseline models and the models generated by recycling (1-3-6-12). Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. The scatter plot and bar chart were drawn using R.

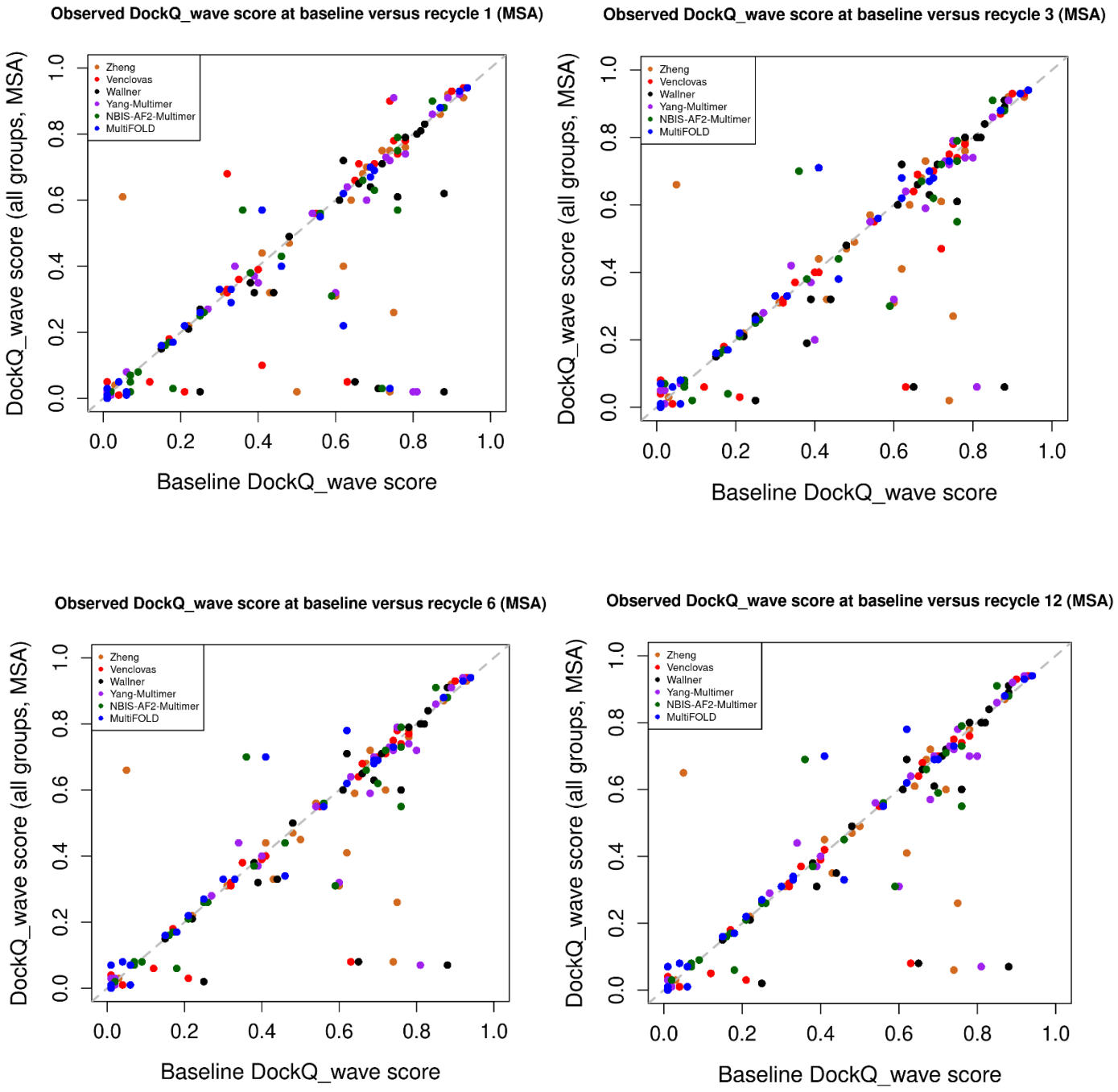


Figure 3.15 A comparison of the observed and baseline DockQ_wave scores for the CASP15 models during each recycles (1-3-6-12).

Four scatter plots representing the comparisons of the observed DockQ_wave scores for the improved models of six groups (y-axis) versus the baseline DockQ_wave scores (x-axis) for the CASP15 models generated during recycles 1 (top-left), 3 (top-right), 6 (bottom-left), 12 (bottom-right), separately, using the MSA method. Each colour corresponds to different group models, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. The scatter plots were drawn using R.

The residue clashes for the CASP15 models were investigated using Molprobity score after modelled via AF2M, similar to the CASP14 models. In Figure 3.16, Molprobity scores of the models were worse than the baseline models since the models were purposefully generated in an unrelaxed form to control for the effect of Amber relaxation (see methods). However, after subsequent recycling, an improvement in the MolProbity scores was observed. The results suggest that there is still the need for MD simulations to resolve clashes and relaxation may be useful rather than relying solely on effective MSAs and templates for refining protein structures. When compared MSAs versus SS methods, a decrease in residue clashes is generally observed during recycles 6 and 12, indicating the acquisition of more geometrically accurate structures (See Appendix Figure S.13).

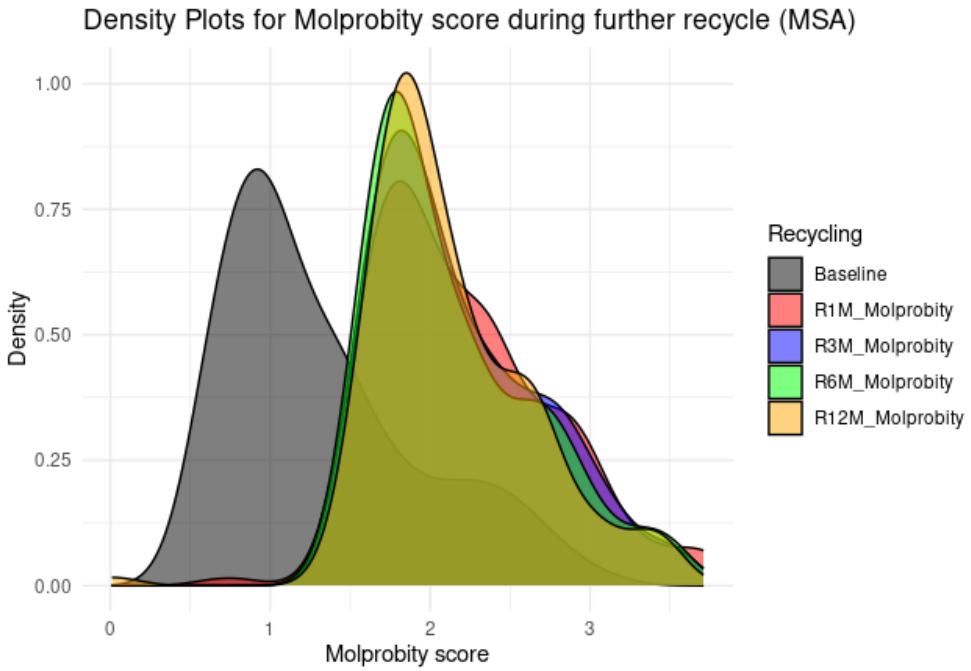


Figure 3.16 A comparison of the observed and baseline Molprobity scores for the CASP15 models after recycling.

Density plot showing the Molprobity scores (lower Molprobity scores are better) for the CASP15 models generated by AF2M using MSA method, with red for cycle 1 (R1M), blue colour for cycles 3 (R3M), green colour for cycles 6 (R6M), magenta colour for cycles 12 (R12M) and black colour for baseline as the initial model. This plot compares the geometric correctness rate for models after recycling, without using experimentally observed protein structure. The Molprobity scores were generated by http://molprobity.biochem.duke.edu/. The density plot was drawn using R.

Following further recycling, MSA was disregarded to understand whether the improvement in the final output was solely attributable to the feedback network of AF2M, similar to the CASP14 models. In table 3.5, the MSA method can be better than SS method in terms of five quality scores. AF2M using SS method, did not show improvement in the initial models during recycles 1, 3, 6, and 12 (See Appendix Figure S.14). It can be note that the cumulative scores for models generated by AF2M using custom template recycling with SS method were expected to be lower. According to Molprobity score, even in the absence of MSA, with the increasing recycling, the template structure demonstrates the potential for obtaining more accurate geometric structures.

3.3.3 What was the wrong for CASP15 models when AF2M custom template recycling was used?

AF2M succeeded in improving the initial models after further specific recycling, in terms of IDDT, QS-score, and DockQ_wave score. In addition, around half of the models were improved in terms of TM-score. In Table 3.5, the cumulative score improvement for the CASP15 models was observed for the five different scores (TM-score, IDDT, QS-score, DockQ-wave, and Molprobity score) following further recycling. However, when compared the initial models with the CASP15 models generated by AF2M using four different recycling (1-3-6-12), a cumulative improvement was not observed in terms of all scores. It seems that the CASP14 and CASP15 models differ in their susceptibility to be successfully refined when used as custom templates for AF2M with further recycling. This difference may be explained by two main reasons:

- **The methods used for generating the CASP models:** with the release of the AF2M code, almost every group in CASP15 integrated AF2M into their own servers or prediction methods. Hence, the CASP14 models were generated by the methods designed by group that attended in the CASP14 competition prior to the availability of AF2M, while the CASP15 models were generated by the methods derived from or integrated with AF2M. As a result, AF2M may already (often correctly) perceive these models as being of high quality. This implies that the initial models of high quality may lead to the generation of alternative conformations, rather than making small refinements to the existing models. It seems it is compatible with traditional refinement phenomenon that the higher quality score initial model has, the worse quality score can obtain following refinement. The likelihood of increasing structural deterioration, similar to the trend observed in global quality scores for refinement, rises with further recycling, especially when starting with a good-quality initial structure (Adiyaman & McGuffin, 2019). However, positive changes were observed in cumulative scores despite using

the best-quality models from the CASP14 competition as initial structures. In addition, AF2M may undertake re-modelling after each cycle. That is to say, remodelling the structure implies exploring the conformational space of the target proteins again, which increases the likelihood of getting trapped in a local minimum, if the initial structure is closer to the native structure.

- **Version difference of AF2M:** During each cycle, AF2M recycling can conduct sampling, which can explore various conformational structures and select the best one based on its own scoring term, such as ipTM-score and Frame Aligned Point Error (FAPE) (Evans et al., 2022). However, the difference in versions of AF2M may limit sampling in order to search the conformational space. Specifically, AF2M trained_v2 mode may restrict the search to a smaller conformational space compared to AF2M trained_v1 mode, since v2 mode is more sensitive to residue clashes. Notably, the difference between two modes of AF2M is that the loss function of v1 mode includes residue clash penalty term. Hence, the effect of custom recycling template on the CASP15 models generated AF2M using v2 mode may be limited in terms of refining the CASP15 models. AF2M using v1 mode may have more room to search for the higher quality CASP14 models when compared the baseline CASP14 models. Wallner (2023a) found similar results in their research.

To investigate why the initial models could not be improved despite obtaining improved models following further recycles, the cumulative scores for only TBM hard, FM, and FM/TBM hard targets were calculated, as these targets are typically easier to refine (as the starting models are often lower in quality). However, the results of Table 3.5 and 3.6 showed the same trend. Again, this could be due to the two reasons mentioned above as post CASP14, there are more methods available, which have higher success at accurate modelling of targets with no known templates.

Table 3.5 A comparison of the cumulative quality scores for the CASP15 models versus the baseline models after recycling.

The table showing the cumulative score of the starting models, known as baseline, and the CASP15 models generated during all recycles (1-3-6-12) for both the MSAs and the SS methods. The cumulative scores include TM-score, IDDT, QS-score, DockQ_wave, and Molprobity score. Note that the lower Molprobity score means higher quality.

Method	Baseline	Cycle-1	Cycle-3	Cycle-6	Cycle-12
			∑TM-score		
MSA	129.39	123.27	126.55	127.02	127.07
SS	129.39	111.19	116.36	120.31	122.25
			∑IDDT		
MSA	126.31	122.44	123.82	125.03	125.14
SS	126.31	113.90	115.73	117.04	118.81
			∑QS-score		
MSA	99.76	87.27	93.87	95.69	96.08
SS	99.76	71.74	82.84	88.41	91.73
			∑DockQ_wave		
MSA	78.17	68.20	71.94	72.65	73.02
SS	78.17	55.61	64.21	67.98	71.15
			∑Molprobity score		
MSA	216.99	359.14	353.56	347.78	345.72
SS	216.99	381.17	375.75	374.04	374.00

Table 3.6 A comparison of the cumulative quality scores for the only TBM hard, FM, and FM/TBM hard CASP15 models versus the baseline models after recycling.
The table showing the cumulative score of the starting models, known as baseline, and the CASP15 models (only TBM hard, FM, and FM/TBM hard targets) generated during all recycles (1-3-6-12) for both the MSAs and the SS approaches. The cumulative scores include TM-score, IDDT, QS-score, DockQ_wave, and Molprobity. Note that the lower Molprobity score means higher quality.

Method	Baseline	Cycle-1	Cycle-3	Cycle-6	Cycle-12
			∑TM-score		
MSA	85.14	79.52	82.33	82.77	82.79
SS	85.14	74.40	77.40	79.537	80.71
			∑IDDT		
MSA	85.21	82.1	83.14	84.08	84.19
SS	85.21	76.34	77.18	78.39	79.32
			∑QS-score		
MSA	60.96	49.71	55.56	57.10	58.00
SS	60.96	45.49	51.47	55.08	57.78
			∑DockQ_wave		
MSA	47.38	37.63	41.75	42.24	43.00
SS	47.38	34.43	39.82	42.29	44.45
			∑Molprobity score		
MSA	151.36	241.99	237.41	235.55	233.01
SS	151.36	255.56	253.60	252.33	252.30

Last but not least, another reason for the discrepancy between the high number of developed models and the lower cumulative scores compared to the initial models could be that AF2M does not detect the available recycle number for each protein target. Thus, due to differences in internal protein structures, the specific number of cycles required for improvement may vary. This implies that a default or optional cycle number may not yield the positive outcome for each protein target, and a better quality models may be found at different recycles. This rationale can be attributed to the impact of the content of protein targets on AF2M confidence scores. In other words, disordered or flexible regions of protein models can influence pTM-score, which in turn can negatively affect ipTM-score, even if the multimeric models are predicted accurately.

Since AF2M primarily ranks models based on ipTM-score and, in such case, it may prioritize the lower quality protein models.

The refinement effect of recycling is based on an underlying deep neural network, which predominantly requires structural information during inference. The utilization of structural information as a template was researched by Liu *et al.* (2023). They demonstrated that AF2M can generate refined models when provided with different similar structures as templates and utilized this method in their own tool, MULTICOM. However, MultiCOM (Liu et al., 2023a) did not focus on recycling and instead ran AF2M five times. Wallner *et al.* suggest the positive effect of using templates, as demonstrated by the differences between AF2M trained _v1 and _v2 modes, along with recycling. We also tested custom template recycling in the CASP15 competition by integration it into our multimeric protein modelling tool, MultiFOLD, which ranked 8th among server groups and outperformed NBIS-AF2-Multimer, which employs the baseline method. More detailed analysis for MultiFOLD can be found in Chapter 5. The recycling method of AF2M resembles iterative refinement methods. Before the release of AF2M, this type of refinement approach was used to improve protein quality (Bhattacharya et al., 2016).

Comparing the recycling process with MD simulation, a powerful tool (Zhang et al., 2009) that reflects protein behaviour in nature (Hollingsworth & Dror, 2018), can be prominent for evaluating the refinement effects of AI-based techniques. However, evaluating the refinement effect of MD simulation on multimeric targets can be challenging due to the need to update force fields, such as better representations of solvents (Yu & Dalby, 2020) for the interfaces of multimeric targets. At the time of writing this chapter, there is no standalone MD simulation tool like ReFOLD3 (Adiyaman & McGuffin, 2021) for multimeric targets. Our research on the monomeric targets indicated that MD simulations (the latest version of ReFOLD) can be used as an alternative to the recycling approach of AF2M and is particularly effective with use of restraints based on local quality scores. ReFOLD3 employs the per-residue scores generated from ModFOLD9 (McGuffin & Alharbi, 2024) as restraints. Making these types of comparisons is not currently possible for multimeric targets. Possibly, a comparison in terms of energy minimization, called relaxation, as used by GalaxyRefineComplex (Heo et al., 2016), could be made. However, the AF2 paper (Jumper et al., 2021a) mentioned that relaxation does not make sense for AF2M targets. Additionally, making comparisons with unrelaxed AF2M models (relaxation was omitted – as a control), can generate misleading results compared with relaxed models. The high Molprobity scores in Figures 3.6 and 3.16 support this idea.

3.4 Conclusions

This chapter investigates the impact of AF2M`s custom template recycling on multimeric protein structures using four separate quality scores (TM-score, IDDT, QS, Molprobity) for models generated from all CASP targets. The results show that the improved models can be obtained when used AF2M using both custom template recycling and MSA method. For CASP14 models, it was observed that using the AF2M_v1 mode resulted in model improvements following further recycling and particularly, with recycle 6 or more, yielding more effective results. However, when the CASP15 models were refined using the AF2M_v2 mode, this led to improvements in models during further recycling, but it did not improve upon the initial models. This can be due to difference in training between v1 and v2 modes of AF2M. Furthermore, all initial CASP15 models were generated by the methods that were already integrated with or based on AF2M. Similar to the CASP14 models, recycle 12 or more could generate higher quality models than initial models, however determining an effective recycle number beyond 12 for the CASP15 targets remains challenging. Generally, the IDDT scores were prone to change around baseline, while the TM-scores exhibited varied change as the number of cycles increased. The change in interface scores occurred for models generated during recycle 1. It was observed that if lower quality models than the initial model were generated during recycle 1, then they did not show improvement following further recycling.

Apart from the AF2M models (from the CASP14 dataset) and the NBIS-AF2-Multimer and MultiFOLD models (from the CASP15 dataset), the models used for research included those submitted by the very top groups in each CASP competition. It was demonstrated that even the highest quality models, expect for AF2M models, were improved when the AF2M_v1 mode was used, while the models generated by AF2M_v2 mode showed the least improvement due to limited conformational sampling in the recycling. Interestingly, our group models, the MultiFOLD models, were improved during further recycling except for IDDT scores. Hence, custom template recycling was used to refine an initial model, and different AF2M versions were employed to generate more conformational structures were in MultiFOLD. Considering that AF2M`s MSA subsampling was used for modelling different conformational structures, it is conceivable that within the algorithm, optimization may become stuck in a local minimum during conformational sampling. Additionally, if the intermediate structure closely resembles the native structure, further optimization may worsen the structure, which aligns with phenomena observed with previous refinement methods (Adiyaman & McGuffin, 2019).

Along with custom template recycling, the models generated by the MSA method exhibit better quality structures after recycling compared to the SS method. In addition to the MSA methods, the given structural information leads to improved models, especially using the AF2M_v1 mode, even with the SS approach. Regarding the type of models, the heterometric CASP14 and homomeric CASP15 models showed better improvement following further recycling. This could be attributed to AF2M using template structures as single chains even when they are multimeric, leading to a loss of information between two chains and potential deterioration in heterometric models. Furthermore, minimizing residue clashes within the models is crucial for improving protein models. More intensive relaxation methods, such as MD simulations, may be more effective for this purpose than simple relaxation methods like energy minimization. However, for AF2M, the simple relaxation method is deemed unnecessary due to a small increase in quality scores. Additionally, the scope of the study does not encompass extensive conformational sampling. Therefore, Molprobity score is not the most appropriate quality score for this chapter, as models generated by AF2M without relaxation option support a decline in residue clashes following further recycling.

In conclusion, AF2M demonstrates effectiveness in refining models, particularly coupled with MSAs and custom template recycling. However, our detailed results highlight the variability in the required number of recycles for each protein targets. Moreover, in the post-AF2 era, downstream analyses like protein-ligand prediction have gained importance over modelling. To enhance model refinement with AF2M, it is crucial to determine specific recycles tailored to protein families or individual structures rather than applying a generic approach. With the introduction of AF2M_v3 version, this challenge has been addressed somewhat by automating the determination of the recycling value (--num_recycles auto--recycle_early_stop_tolerance_auto) for each protein structure, eliminating the need for manual intervention and streamlining the refinement process.

Chapter 4: The Impact of Varying Custom Input Options on Models Generated by AF2M

4.1 Background

Deep neural networks (DNNs) have played a crucial role in the field of protein structure prediction, perhaps most notably after the release of AF2, which was as a major milestone of structural bioinformatics (Osadchy & Kolodny, 2021). The remarkable progress achieved can be attributed to the availability of large amounts of labelled data and the development of increasingly powerful computational hardware (GPUs). DNNs learn the complex features of datasets in models consisting of millions of parameters, drawing certain inferences such as classification or regression from the features they have learned. When training the parameters within the network, DNNs internally learn these features either by minimizing the loss function through backpropagation or through optimization techniques (Kingma & Ba, 2014). This is what sets DNNs apart from classical ML applications. Namely, in traditional ML applications, hand-crafted features are used, and when these features are not well-designed, there is a likelihood of the model making incorrect inferences (Osadchy & Kolodny, 2021). Hence, end-to-end DNNs have been the key factor behind the power of AF2M due to their ability to extract necessary structural information for a given target from MSA.

The unexpectedly high accuracy of AF2M, end-to-end DNN, is mainly based on information from MSA and structural templates. One of the other major factors of AF2M's success is the iterative process from improving models through multiple passes through the network, which is called recycling. However, a broader search for conformational changes in important proteins, such as G-Protein Coupled Receptors (GPCRs), could broaden the scope of possible model solutions, resulting in lower convergence and greater structural variance. The initial trials to utilize the default implementation of AF2M to obtain a collection of structures spanning holo and apo states or to capture the flexibility of disordered regions in protein structures failed because the default options did not sample the anticipated structural heterogeneity. Researchers have begun examining modifications to the AF2M platform, including custom options to obtain multiple conformational structures (Sala et al., 2023; Saldaño et al., 2022). Detailed analysis using the AF2M “custom template” option was provided in Chapter 3. The AF2M “custom MSA” option will be explored in this chapter.

MSAs are fundamentally based on an amino acid substitution matrices and when homologous sequences obtained from databases exhibit 20% or higher similarity (Rost, 1999), protein sequence alignments can be used to reveal similarities and differences between protein structures. Amino acid substitutions are vital for revealing information about protein evolution and function (Dayhoff, 1978), and they are a crucial consideration in designing MSAs (Fox et al., 2015). Additionally, within protein MSAs, there are residues corresponding to both ordered

and disordered structures. Ordered structures provide significant information in terms of evolutionary context, whereas disordered structures may offer less evolutionary information due to a higher frequency of mutations (Brown et al., 2002).

An intrinsically disordered protein may contain short or long regions of disorder. Short disordered regions are frequently noted in the shape of hinges, which facilitate controlled movement of a domain, or loops that exhibit both open and closed shapes (DeForte & Uversky, 2016). Molecular Recognition Features (MoRFs) are also considered short disorder region, which experience a contextual shift from disorder to order upon binding (Vacic et al., 2007). Theoretical analysis of disordered regions led to categorizations of short disorder as ≤30 residues and long disorder as >30 residues. Additionally research has uncovered distinct tendencies towards specific amino acid residues (Zhang et al., 2012). The amino acid content of disordered regions typically shows a distinctive character, with a higher rate of residues that promote disorder (such as A, R, G, Q, S, P, E, K) and a lower rate of residues that promote order (such as W, C, F, I, Y, V, L, N) (Szilágyi et al., 2008). In addition, intrinsically disordered regions (IDRs), spanning a length of 20 to 30 residues or more, may also exhibit associations with globular protein partners, akin to structured domains. These segments of disordered proteins or regions are termed disordered interacting domains (Tompa et al., 2009). These regions possess conserved functions, sequences, and disorder (Chen et al., 2006). Types of disordered residue is demonstrated in Table 4.1.

Table 4.1 Types of disordered residues within protein structure.
This table presents three types of disordered regions within protein structures according to the number of disordered residues within the protein sequence. Regions within the protein structure that include fewer than 30 subsequent disordered residues are known as short disordered regions, while more than 30 disordered residues within structure create long disordered structures. There is one other group known as intrinsically disordered region with a length of 20 to 30 residues.

Types of disordered residues	The number of disordered residues
Short disordered region	≤30
Long disordered region	>30
Intrinsically disordered region	a length of 20 to 30

Eukaryotic protein structures are known to exhibit a higher prevalence of disorder regions compared to prokaryotic structures (Ward et al., 2004). However, the presence of high disorder rates in single-cell structures indicates that there is still much to understand about the functions and evolution of disordered structures (Kastano et al., 2020). Although disordered regions contain fewer co-evolved residues compared to globular structures, EVfold (Toth-Petroczy et al., 2016) was updated to uncover evolutionary coupling information within disordered protein

chains (Pancsa et al., 2018). Recent advancements in the field of deep learning that have significantly impacted protein structure modelling have also included co-evolution or contact-based methods. At present, the most successful protein structure modelling has been entirely dependent on ML methods in order to understand how evolutionarily coupled residues govern protein structure (Suh et al., 2021).

Various approaches have been used to improve the quality of protein structures by integrating MSAs into DNNs. The first method (A DNN-based method that employs an end-to-end differentiable method to extract information directly from an MSA) involves inputting embedded protein sequences (Kandathil et al., 2022). The rawMSA (Mirabello & Wallner, 2019) takes the MSA directly as input and transforms the entire MSA into a numerical vector form (embedding method). This approach does not require the handcrafted features typically used to extract evolutionary information from sequence profiles. Traditional methods rely on expert-designed features to extract such information, which involves specific computations to understand the similarities and evolutionary relationships among the sequences. In contrast, the rawMSA approach eliminates the need for these additional processing steps by incorporating raw sequence data directly into the model's input. This allows the model to automatically learn evolutionary information, offering a more efficient learning process by removing the need for manual feature engineering. Another tool used is MSA Transformers (Rao et al., 2021b), distinguished by its incorporation of both row and column attention mechanisms, providing an advantage over standard transformers. Further to these approaches, AF2 (Jumper et al., 2021a) and RoseTTAFold (Baek et al., 2021) have emerged, which produce protein structures of excellent quality, especially for single chains, via the end-to-end method using the embedding method directly to input MSA (Kandathil et al., 2022).

The use of an end-to-end DNN exemplifies AF2M's strength. One of the most significant advancements is the use of AF2M’s transformer neural networks. Transformer models play an important role in addressing the drawback of traditional DNN-based methods. Specifically, they can determine which information in a dataset is most important for a specific task, thanks to the attention mechanism. This is accomplished by assigning weights to different elements within the data. In the field of protein modelling, transformers leverage valuable information within protein sequences. AF2M`s transformer combines the information generated from residue co-evolution in homologous sequences within MSAs with the information obtained from pair presentation, determining which residue pairs are more effective after 48 iterations in the Evoformer block. Namely, the underlying algorithm combines the residual pairs information obtained from the template database with the knowledge in the MSA transformer in the Evoformer block, strengthening the attention mechanism by adding bias, and revealing key

residue pairs. In AlphaFold, the attention mechanism called the ‘triangle multiplicative update’ is used to find missing links between residues (amino acids) in a protein. In a triangle of three residues, this mechanism uses the information from two sides to predict the missing third side. Thus, the relationships between residues are completed, and a protein structure is modelled more accurately. This method helps the model understand the complex interactions within the protein and better predict the structure. In addition, during the training of AF2M, the 'masked token in the MSA' method was employed, involving masking 15% of the sequence, to improve performance (Jumper et al., 2021a).

Another significant advancement is to combine homologous sequences in the MSA obtained from a combination of databases and to use exact information via Transformers. ColabFold (Mirdita et al., 2022) can be used as an more efficient alternative to AF2M which processes many homologous sequences extracted from UniRef30, BFD/Mgnify including a combination of BFD and Mgnify, ColabFold DB including BFD/Mgnify, MetaClust2 (Steinegger & Söding, 2018), Metagenomic Gut Virus catalogue (Nayfach et al., 2021), MetaEuk (Levy Karin et al., 2020), Human gut bacteriophage catalogue (Camarillo-Guerrero et al., 2021), Marine planktonic eukaryotes-SMAG (Delmont et al., 2022), TOPAZ (Alexander et al., 2023) databases. ColabFold employs MMseqs2 (Mirdita et al., 2019) in order to create an MSA by aligning homologous sequences extracted from these databases, which is more practical compared with the default AF2M MSA generation.

A large number of disordered proteins have been revealed experimentally. Most of these proteins are collected in two sources rich in disorder proteins, one of which is DisProt (Quaglia et al., 2021) and the other is the PDB (Rose et al., 2017). However, since the PDB database generally hosts ordered structures, missing regions in X-ray experimental structures and high mobility regions in NMR are considered as disordered segments (Monastyrskyy et al., 2014). Intrinsic disordered residue (IDR) predictors can utilise amino acid sequence only without the need for sequence dependent emergent properties such as backbone dynamic of protein structures (Orlando et al., 2022). However, most research indicates that AF2M predict only single state model (Jumper et al., 2021b), while Guo et al. (2022) suggest that AF2M translates protein sequences to residue flexibility though predicted aligned score (PAE) and pIDDT score (Guo et al., 2022). While deeper MSAs lead to increased protein structure quality, a high Neff value (Guo et al., 2021), the number of effective homologous sequence in the MSA, may not always produce protein structures with better quality scores (Yang et al., 2021). Therefore, the success of a prediction using AF2M can be affected by depth and the number of homologous sequences in the MSA. AF2M models can be improved by MD-based tools, such as ReFOLD4 (Adiyaman et al., 2023). This type of physical approach can be more suitable for the refinement

of disordered regions in models and the alternative models produced can reflect structural dynamics. Heo et al. (2021) demonstrated that MD-based tools in CASP14 could be used to improve models from many AI-based protein structure modelling methods, with the exclusion of AF2M models. However, our later results (Adiyaman et al., 2023) demonstrated that most of CASP14 monomer models from AF2M can be refined using ReFOLD4.

Several gaps and challenges still exist in improving AF2M multimeric models through physical approaches, such as tailored force fields for multimeric structures. In addition to the challenges of physical approaches, aligning tens of thousands of homologs using computer-based tools, can lead to alignment problems (Iantorno et al., 2014). Nevertheless, numerous studies have been conducted to enhance the quality of MSAs. In the latest CASP (CASP15), these studies involved the combination of homologous sequences obtained from various databases, a method that can be termed horizontal MSA filtering. Recently, the optimal effective number of sequences (Neff) using the DeepMSA tool (Zhang et al., 2020) was found to be 128. With this approach, sequences that are not effective are excluded from the MSA. Another method, SpliVert (Zhan et al., 2020), focuses solely on vertical alignment and filtering to achieve a more efficient MSA. The technique of filtering entire columns within the MSA is also frequently employed, although such methods may lead to information loss. However, there is a lack of detailed studies on the impact of filtering sequences on AF2M model quality, as most research on filtered MSAs focuses on evolutionary analysis (Ashkenazy et al., 2019; Steenwyk et al., 2020; Tan et al., 2015; Zhang et al., 2021).

4.1.1 The aim of study

In Chapters 3, it was demonstrated that AF2M can model multimeric structures more effectively by using the custom template recycling method. In the current version, the number of recycles in the AF2M network is set to 'auto' by default, meaning that AF2M can determine the effective number of recycles automatically. This determination is provided by the “early stop tolerance” which activates to stop recycling if the angstrom differences in distance matrices are below the specified tolerance value. In addition, by default all chains in a template structure are considered individually in the modelling process and so relative chain orientation information may be lost. Hence, the first aim is to evaluate the impact of using “single-chain” custom template on AF2M models, whereby all of the template chains are considered as a single entity, rather than as separate templates for each chain in the modelling process. In summary, single-chain templates were a type of template created by converting multi-chain structures into single-chain forms using the methods described in the "Methods" section with the PyMOL

program, while standard custom templates, on the other hand, were external protein structures that are supplied straight to AF2 as input without any processing or changes.

Another input option for AF2M is the “custom MSA”. By inputting custom MSAs into AF2M, there is the potential to generate better models. Thus, the second aim here will be to the “custom MSA” and investigate whether AF2M may be able to produce higher quality models as a result. Petti et al. (2022) investigated the impact of “low-quality” self-inconsistent sequences in the MSAs on AF2M models. Seemingly paradoxically, the self-inconsistent MSAs, with higher complexity, were found to have a positive adversarial effect on models predicted by AFM2. Inspired by study, the effectiveness of excluding disorder information in MSAs used by AF2M will also be evaluated by applying a filtering method, whereby disordered residues are ignored in each protein sequences. The rationale for filtering out the low complexity disordered regions is to introduce self-inconsistency into the MSA inputs for AF2M. Overall, this research aims to contribute to the understanding of the potential effect of various custom inputs on protein models during AF2M`s inference time.

4.2 Methods

4.2.1 Data collection

For testing the effect of using “single-chain” custom templates, the best four groups of models, the NBIS-AF2-Multimer models, and our prediction models (MultiFOLD) were selected from the CASP15 website, according to the assessor evaluation in the CASP competition (z-score). The dataset included a total of 120 models for 20 targets (Table 4.2), which were transformed into “single-chain” custom templates using PyMOL. In the models’ PDB files, all chain letters were changed to ‘A’ and all residues in the PDB file were numbered consecutively from the beginning to the end of the files. To test for the effect of varying the MSA complexity, the available method for assessing the improvement effect on protein modelling is to observe the score difference between the models and the reference structures. Again, the most appropriate targets for this assessment are those associated with CASP, particularly the last CASP (CASP15) targets, as they provide observed structures for the given targets. 13 CASP15 targets and six relevant targets from the CASP14 competition were included to further investigate homomeric models (Table 4.3). For the target selection process, consideration was given to the presence of the observed structures, the presence of residues corresponding to disorder structures in the MSA, and whether the total number of residues was within the bounds that AF2M could handle in a single run. In MSAs for heteromeric models generated by AF2M, there are both separate homologous sequences and paired homologous sequences for given protein sequences. Since the impact of paired homologous sequences on model quality is more complex, homomers rather than heteromers were preferred for this initial analysis.

Table 4.2 "single-chain" custom template targets from the CASP15 competitions.
The table listing the CASP15 multimeric targets used as the "single-chain" custom template. There are several information (description of proteins including PDBID in blanket if there are any, stoichiometry, PDBID about the CASP15 targets.

Targets	Descriptions	Stoichiometry
T1109	D180A isocyanide hydratase (Organism: Ralstonia solanacearum)	A2
T1110	wild-type isocyanide hydratase (Organism: Ralstonia solanacearum)	A2
T1113	Glycoprotein 2 (GP2) (Organism: Bacteriophage PA1C)	A2
T1121	The Wadjet nuclease subunit JetD (Organism: Pseudomonas aeruginosa PA14)	A2
T1123	Capsid protein (Organism: Human Astrovirus MLB1)	A2
T1127	L-ornithine N5-acetyltransferase NATA1 (Organism: Arabidopsis thaliana)	A2
T1132	PA0709 with glyoxal and BME modifications (Organism: Pseudomonas aeruginosa)	A6
T1153	Endonuclease/exonuclease/phosphatase family domain-containing protein 1 (Organism: Human)	A2
T1160	The mk2h_deltaMILPYS peptide homodimer (Organism: HAncient protein reconstruction)	A2
T1161	The dimeric DZBB fold protein Ph1 (Organism: HAncient protein reconstruction)	A2
T1174	(the C-terminal domains of the Bdellovibrio bacteriovorus Bd2133 fibre (Organism: Bdellovibrio bacteriovorus)	A3
T1178	Neuronal HAstV VA1 capsid spike domain (Organism: Human Astrovirus VA1)	A2
T1179	GenBank: QBQ83077.1 (8tn8)	A2
T1187	Tobacco lectin Nictaba in complex with triacetylchitotriose (Organism: Nicotiana tabacum)	A2
H1106	(YscY-YscX protein (Organism: Yersinia enterocolitica)	A1B1
H1134	(Chymotrypsin digested toxin/immunity complex for a T6SS lipase effector (Organism: enterobacter cloacae)	A1B1
H1140	CNPase-Nb (Organism: mouse/alpaca)	A1B1
H1141	CNPase-Nb7e (Organism: mouse/alpaca)	A1B1
H1142	CNPase-Nb8c (Organism: mouse/alpaca)	A1B1
H1151	Probable transcriptional regulator WhiB6 (Organism: Mycobacterium tuberculosis)	A1B1

Table 4.3 The custom MSA targets from the CASP14 and the CASP15 competitions.

The table shows the multimer targets of the CASP competition used to filter the MSA. The first column shows the target name, the second column shows the description of protein with PDBID in the blankets if there are any, the third column shows stoichiometry, and the last column indicates the CASP to which targets belong. There are 19 multimer targets.

Multimeric Targets

Target	Name of Protein	Stoichiometry	CASP
T1032	Structural maintenance of chromosomes flexible hinge domain containing 1 (Organism: Homo sapiens)	A2	14
T1034	Inhibitor of the Yeast Formin Bnr1 (Organism: Saccharomyces cerevisiae)	A4	14
T1038	Tomato Spotted Wilt Virus (TSWV) glycoprotein (Organism: Semliki Forest virus)	A2	14
T1078	a small secreted cysteine-rich protein (Tsp1) (Organism: Trichoderma virens)	A2	14
T1083	Nitro-histidine zipper coiled coils (Organism: Nitrosococcus oceani)	A2	14
T1087	Tuna-histidine zipper coiled coils (Organism: Methylobacter tundripaludum)	A2	14
T1109	D180A isocyanide hydratase (Organism: Ralstonia solanacearum)	A2	15
T1110	wild-type isocyanide hydratase (Organism: Ralstonia solanacearum)	A2	15
T1113	Glycoprotein 2 (GP2) (Organism: Bacteriophage PA1C)	A2	15
T1121	The Wadjet nuclease subunit JetD (Organism: Pseudomonas aeruginosa PA14)	A2	15
T1123	Capsid protein (Organism: Human Astrovirus MLB1)	A2	15
T1124	MfnG (Organism: Streptomyces drozdowiczii)	A2	15
T1127	L-ornithine N5-acetyltransferase NATA1 (Organism: Arabidopsis thaliana)	A2	15
T1132	PA0709 with glyoxal and BME modifications (Organism: Pseudomonas aeruginosa)	A6	15
T1153	Endonuclease/exonuclease/phosphatase family domain-containing protein 1 (Organism: Human)	A2	15
T1160	The mk2h_deltaMILPYS peptide homodimer (Organism: HAncient protein reconstruction)	A2	15
T1161	The dimeric DZBB fold protein Ph1 (Organism: HAncient protein reconstruction)	A2	15
T1178	Neuronal HAstV VA1 capsid spike domain (Organism: Human Astrovirus VA1)	A2	15
T1187	Tobacco lectin Nictaba in complex with triacetylchitotriose (Organism: Nicotiana tabacum)	A2	15

4.2.2 Experimental design

Firstly, AF2M was run on the CASP targets producing for a total of 120 models using “single-chain” templates with the “custom template” option (renaming the multiple chains in the PDB as one chain). Secondly, AF2M was run for 19 targets using low quality custom MSAs (filtering MSA to remove residues corresponding to disordered regions in the 3D structure). Low quality MSAs were obtained by filtering residues corresponding to the various types of disorder separately from the initial MSA of the given protein. Various methods have been developed to identify residues corresponding to disordered structures within protein sequences. Among these methods, IUPred3 (Erdős et al., 2021) has been widely utilized. IUPred3 typically provides probabilistic values, considering a residue as disordered if the probability value is above 0.5. These probability values are generated based on energy-based methods. The baseline MSA was obtained via AF2M using the default values, while the filtering methods were implemented using an in-house python script (See Appendix Table S.2). The code for running IUPred3 is provided in Appendix Table S.3. After determining the potential disorder for each target residue, if the disorder score of a residue is 0.5 or higher, a residue was removed and replaced with "-" in the corresponding position. This ensures uniform length across homologous sequences. Subsequently, the obtained MSA was adjusted for compatibility with the custom MSA input of AF2M by adding residue and chain numbers, and primary protein sequences to the first three lines of the MSA file.

Models were generated using ColabFold (version 1.5.3), which was executed with the following parameters for multimeric structures:

*** template_mode: (none for custom MSA); msa_mode: (MMseqs2 (UniRef+Environmental) for baseline and custom for the filtered MSA); pair_mode: unpaired+paired; model-type: v1 for CASP14 and v2 for CASP15; num_recycles: auto. (N.B. Selecting 'alphafold2_Multimer_v1 or _v2' from the model type was intended to avoid bias in structure prediction).

For “single-chain” custom templates, the template mode ”custom” was selected, while msa_mode was selected as “MMseqs2 (UniRef+Environmental)”.

4.2.3 Evaluation

In order to investigate the impact of the “single-chain” custom template and of the low quality MSA, four different model quality scoring metrics were employed. The commonly used metrics for complex model quality were the TM-score, IDDT. The IDDT scores refers to the Oligo-IDDT scores. However, for the “single-chain” custom template analysis, Molprobity score was not

included due to the need for relaxation of models compared to the initial models. Additionally, the QS-scores and DockQ_wave scores, which specifically demonstrate improvements in the interface structures, were obtained. The TM-scores for the models were obtained through the MM-Align server and the IDDT and interface scores were obtained using the OpenStructure package (2.1). The improvements or deteriorations in these scores were examined both cumulatively for the models generated AF2M using the "single-chain" custom template and on a target-by-target basis for models predicted by AF2M using filtered custom MSAs. To assess the statistically difference in model quality using the "single-chain" custom template approach, the paired Wilcoxon signed-rank test was conducted using R, as in previous chapters. This analysis aimed to ascertain whether there were significant improvements in the scores of models generated by AF2M with the "single-chain" custom template compared to those generated by the standard custom template approach and the initial models. The statistical method used is explained in detail in the method part of Chapter 2. Figures 4.1 and 4.2 summarise the workflow of methods used in the analysis for this chapter, subsequently.

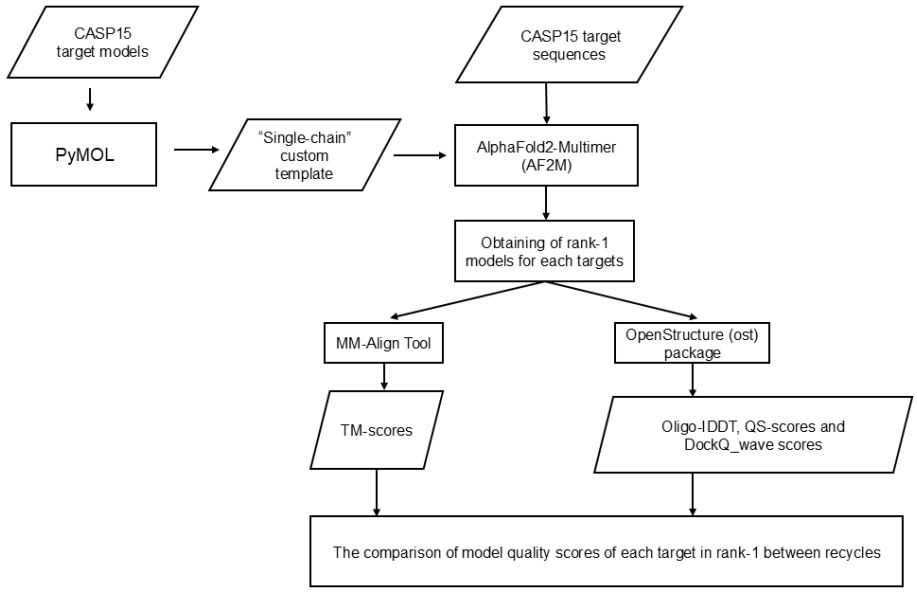


Figure 4.1 The flowchart of the method for evaluating the effect of a "single-chain" custom template on modelling quaternary structures.
Flowchart showing the process in which multi-chain CASP15 models are first converted into single-chain forms using PyMOL. These single-chain structures are then used as template inputs for AF2M to generate rank-1 models. Four different quality scores are applied to evaluate these models. The observed quality scores, with TM-score from MM-Align and IDDT/QS-score/DockQ_wave from OpenStructure, were produced by aligning the models with the native structures for each target. Initially, the cumulative scores are assessed, followed by the evaluation of individual pairwise scores for each model using the Wilcoxon-signed-rank test.

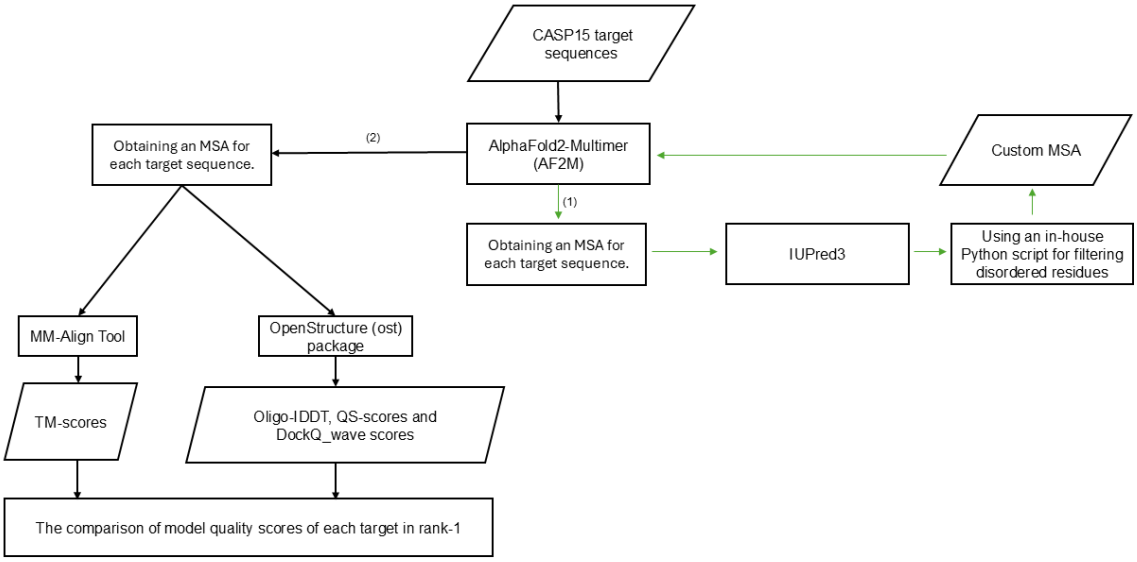


Figure 4.2 The flowchart of the method for evaluating the effect of an MSA without disordered residues on modelling quaternary structures.
Flowchart showing a two-step process demonstrating the impact of removing disordered residues from sequences comprising the MSA on the model quality of AF2 predictions. The green arrows represent the first step, where an MSA devoid of disordered residues is generated, referred to as the "custom MSA" in the subsequent phase. The black arrows indicate how the custom MSA is used in the second step and how models are evaluated using four distinct scoring metrics. The observed quality scores, with TM-score from MM-Align and lDDT/QS-score/DockQ_wave from OpenStructure, were produced by aligning the models with the native structures for each target.

4.3 Results and Discussion

4.3.1 The impact of using “single-chain” custom templates” for quaternary structures modelling

The primary aim is to compare the results of models generated by AF2M using different custom template inputs to evaluate the effectiveness of treating all chains in a template complex as a single chain. Thus, the various quality scores (TM-score, lDDT, QS-score, and DockQ_wave) for the models predicted by AF2M using the “single-chain” custom templates are compared to those of models generated by AF2M using the default custom template and to the initial models (the template structures). Table 4.4 demonstrates the cumulative effects on the model improvement for models generated by AF2M using two different custom template inputs. The cumulative TM-scores and lDDT scores of the models generated by AF2M using the “single-chain” custom templates were higher than those of the models obtained by AF2M using the standard custom templates, as well as those of initial models. In terms of interface scores, the initial models were not improved; however, the models generated by AF2M using “single-chain” custom templates were better predicted than the models generated by AF2M using standard custom template. The results suggest that using “single-chain” custom templates can be more effective than using the standard custom template options. In addition, AF2M using a “single-chain” custom template input can produce protein models with higher global and local quality

compared to the initial models. The paired Wilcoxon signed-rank test (See Method section in Chapter 2) result demonstrated that the increase in TM-scores and IDDT scores for the models generated by AF2M using "single-chain" custom template were significantly different compared to both the standard custom templates and the initial models (Statistically, $p=3.0E{-}02 < 0.05$ for TM-score and $p=4.04E{-}02 < 0.05$ for IDDT-score between "single-chain" custom template models and initial models while $p=1.04E{-}02 < 0.05$ for TM-score and $p=3.96E{-}03 < 0.05$ for IDDT-score between "single-chain" custom template models and initial models. No other statistically significant differences were observed between the given two variables). This result also supports Figures 4.3, 4.4, 4.5 and 4.6.

Table 4.4 The cumulative global and interface scores for the AF2M models and initial models.

The table comprising of the cumulative TM-scores, IDDT scores, and interface scores (QS-scores and DockQ_wave scores) for the homomeric and heteromeric models generated by AF2M using both "single-chain" custom templates and standard custom templates, compared to the initial models.

Cumulative scores											
Single-chain custom template				**Standard custom template**				**Initial model**			
TM-score	IDDT	QS-score	DockQ_wave	TM-score	IDDT	QS-score	DockQ_wave	TM-score	IDDT	QS-score	DockQ_wave
95.08	**92.82**	73.87	57.44	94.72	91.71	73.52	57.27	94.87	92.75	**74.4**	**57.86**

Table 4.5 shows the effect of using a "single-chain" custom template on AF2M models in terms of different types of multimeric models. Nearly half of homomeric models generated using "single-chain" custom template exhibited improved predictions across all four quality metrics compared to the initial models. Furthermore, greater than half of the homomeric models achieved higher quality scores with "single-chain" custom templates compared with those generated by AF2M using the standard custom templates. The cumulative scores were generally higher when the "single-chain" custom templates compared to both the initial models and the models generated AF2M using the standard custom templates. The standard custom template method also showed good performance except for the cumulative IDDT scores, where the models generated by AF2M using standard custom templates surpassed the initial models (Table 4.5a). According to the heteromeric models, more of the heteromeric models generated by AF2M using "single-chain" custom templates exhibited higher quality in terms of TM-score and DockQ_wave score when compared to initial models, rather than the heteromeric models generated by AF2M using the standard custom template. Due to the higher sensitivity of DockQ_wave scores, fewer models were expected to show better quality compared to analysis using the QS-scores. Interestingly, when the DockQ_wave scores are considered, more initial models were of higher quality compared to AF2M using "single-chain" custom templates. The cumulative scores of initial heteromeric models were the highest among all considered models (Table 4.5b).

Table 4.5 The number of improved models and cumulative global, local, and interface scores for the AF2M and initial homomeric and heteromeric models.

The tables demonstrating the number of improved models and comparing the cumulative TM-scores, IDDT scores, and interface scores (QS-scores and DockQ_wave) for the models. Tables a and b compare the number of improved models and cumulative scores separately for the homomeric and heteromeric models generated by AF2M using both "single-chain" custom templates and standard custom templates, compared to the initial models. The blank scores indicate the highest cumulative scores for quality scores.

a)

"Single-chain" custom template and standard custom template (Homomers) The percentage of the improved models			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
55	62	58	51
"Single-chain" custom template and initial model (Homomers) The percentage of the improved models			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
42	43	46	44
Cumulative scores of the homomeric models (using "single-chain" custom template)			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
65.90	**63.49**	**55.79**	**42.63**
Cumulative scores of the homomeric models (using standard custom template)			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
65.60	62.53	55.55	42.52
Cumulative scores of the initial homomeric models			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
65.25	63.35	54.92	42.01

b)

"Single-chain" custom template and standard custom template (Heteromers) The percentage of the improved model			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
37	45	28	37
"Single-chain" custom template and initial model (Heteromers) The percentage of the improved model			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
53	45	28	48
Cumulative scores of the heteromeric models (using "single-chain" custom template)			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
29.18	29.32	18.08	14.8
Cumulative scores of the heteromeric models (using normal custom template)			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
29.12	29.18	17.97	14.75
Cumulative scores of the initial heteromeric models			
TM-score	**IDDT**	**QS-score**	**DockQ-wave**
29.61	**29.40**	**19.48**	**15.85**

The above scores indicate that the input of multi-chain templates as a single chain is more effective in both global and local folding and interface areas in the homomer models. However, in the heteromeric models, according to the TM-scores and DockQ_wave scores, although the models generated using "single chain" custom templates are more than the initial models, the cumulative scores decrease, indicating that more models are distorted than the initial models. The AF2M developers highlighted that the standard AF2M predicted homomeric interface regions better than heteromeric interface regions (Jumper et al., 2021b). In addition, the first version of AF2M faced the issue of stoichiometry. Namely, AF2M did not manage to predict well for models with more than two chains. Therefore, this observation helps to explain why improvements were also more noticeable here in homomeric models, since the symmetry information for homomeric models can be more advantageous (Gaber & Pavšič, 2021).

Figure 4.3A illustrates that out of 120 initial models, only 54 models were not improved when the “single-chain” custom templates were used. The remaining initial models achieved higher quality (63 models) scores. However, when compared to the models generated by AF2M using standard custom template, fewer models showed improvement, and 38 models have higher TM-scores (Figure 4.3B). This suggests that using “single-chain” custom templates may be more effective than using standard templates in terms of improving the TM-score of models. In addition, the plots show the distribution of TM-scores for both refined and non-refined models, which peak around 0.9 to 1.0. They demonstrate a strong agreement between the TM-scores of the models obtained using the standard custom template option and those obtained using the “single-chain” custom templates, particularly at higher scores (Figure 4.3A). This agreement also suggests that both methods are fairly consistent in their TM-scores. High TM-scores with one method mostly correspond to high scores in the other.

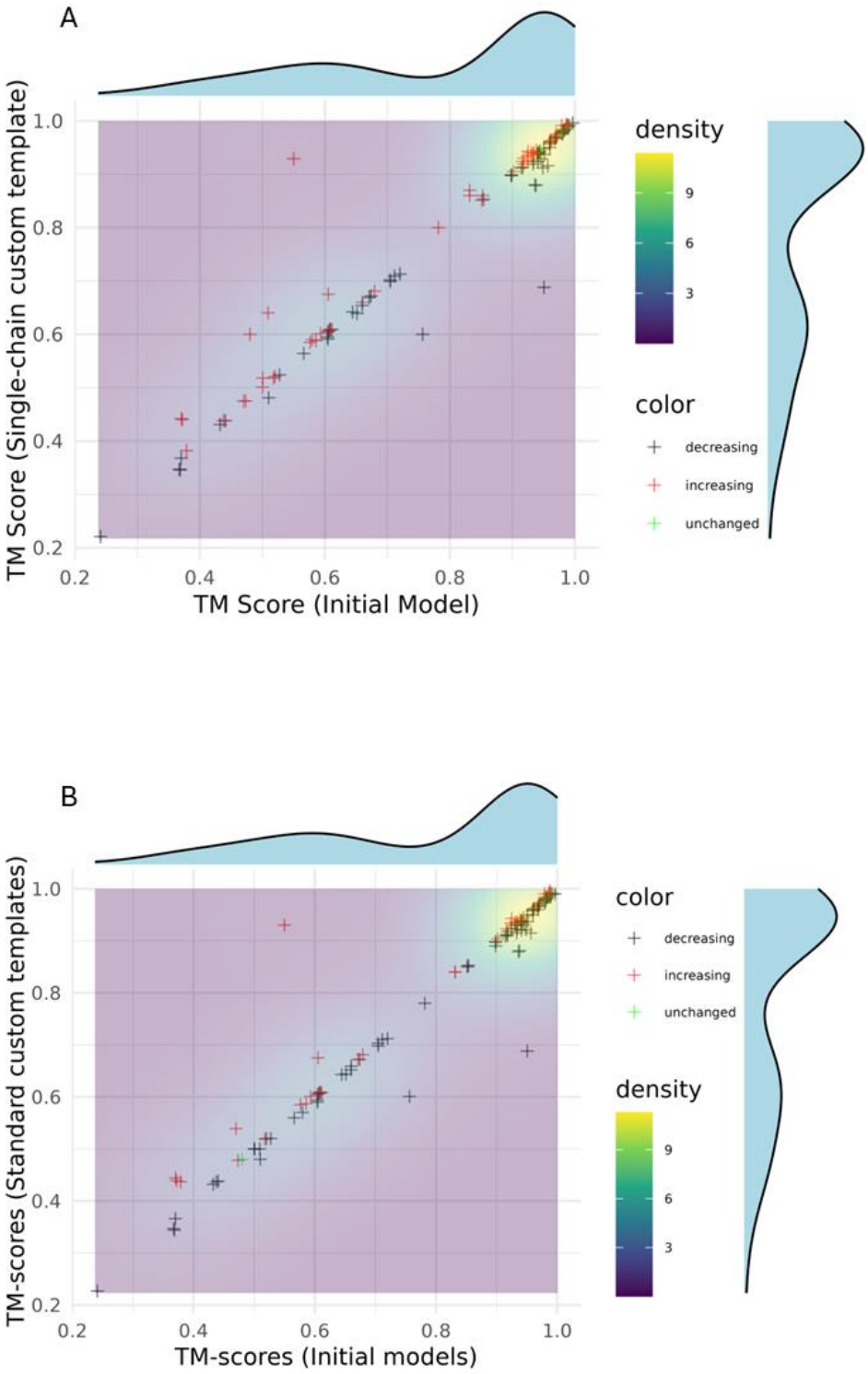


Figure 4.3 Comparison of observed TM-scores between the AF2M models using the custom template option with recycling and the initial models.

Density scatter plot presenting the comparison of TM-scores for (A) the models generated by AF2M using "single-chain" custom template versus the initial models, and also (B) the models generated by AF2M using the standard custom templates versus that of the initial models. The red pluses indicate the refined models. The black pluses indicate the unrefined models while the green pluses represent that the model quality do not change. The density scale ranges from purple to yellow. The scatter plots were drawn through R.

Figure 4.4 shows the improvement in protein models according to the IDDT scores after using the “single-chain” custom template option. The IDDT scores of all models ranged from 0.35, with the majority falling between 0.75 and 1. However several models exhibited the lower TM-scores and high IDDT scores. Typically, initial models with TM-score as low as 0.5 might be considered random; however, AF2M can potentially generate models with TM-scores of 0.5 or higher following the recycling process. For the “single chain” custom templates compared to the initial models, 52 models were deteriorated, and 65 models were improved (Figure 4.4A), whereas for the standard custom templates, 36 out of 120 models showed improvement and 75 models had lower IDDT score (Figure 4.4B). This suggests that using “single-custom” custom templates may influence local improvements of initial models. There are two models which are outliers. One model (T1179 MultiFOLD) showed a very low IDDT score and did not improve after AF2M using both types of custom template inputs. The second model (T1110 Zheng) exhibited very high improvement when the “single-chain” custom template was used, which had lower IDDT score when using the standard custom template.

Modelling performed using single-chain custom templates showed higher global and local quality compared to both standard custom templates and initial models Presenting results on a target-specific basis visualisation even minor improvements in modelling scores. Additionally, while AF2 recycling can improve the structure in one cycle value, it can model a different conformational structure in the next cycle. In given cases, it is possible that the initial models the initial models may remain unchanged (neither improvement nor deterioration). However, in the latest versions of AF2, the recycling process is controlled automatically, allowing the model to perform recycling until it identifies the best structure based on the given template. As a result, the differences between the generated structures can be minimal, leading to stacking in certain regions.

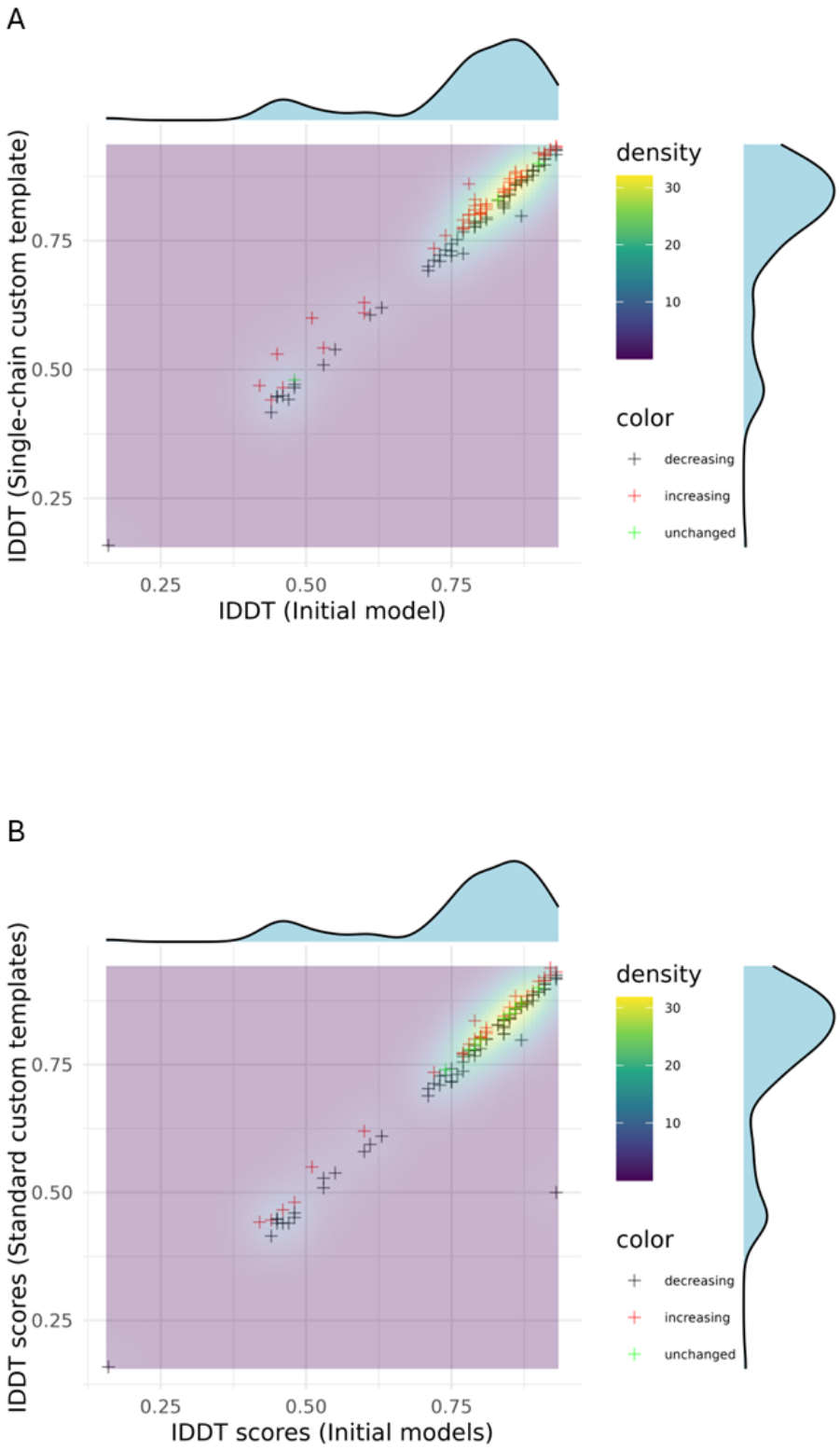


Figure 4.4 Comparison of observed lDDT scores between the AF2M models using the custom template option with recycling and the initial models.

Density scatter plot presenting the comparison of lDDT scores for (A) the models generated by AF2M using “single-chain” custom template versus the initial models, and also (B) the models generated by AF2M using the standard custom templates versus that of the initial models. The red pluses indicate the refined models. The black pluses indicate the unrefined models while the green pluses represent that the model quality do not change. The density scale ranges from purple to yellow. The scatter plots were drawn through R

Interface quality scores were crucial for comparing the improvement effect of using a “single-custom” template against using a standard custom template. With the “single-chain” custom templates, AF2M can refine the entire structure simultaneously, potentially viewing these structures as a single chain rather than distinct chains and thereby maintaining the modelled interface. In such cases AF2M will recognize the separate chains as “domains” within a single structure, which can lead to improvements. Firstly, in terms of the QS-score, 51 models showed deterioration, and 48 models show improvement when the models generated by AF2M using the “single-chain” custom template were compared to the initial models (Figure 4.5A). However, 57 of the initial models were deteriorated while 32 models showed improvement when using the “standard custom template” (Figure 4.5B). In general, most of the QS-scores for the models predicted by both custom template methods tended to be high and around the baseline, similar to TM-score and IDDT scores. This is because a normal custom template processes AF2M structures chain by chain, whereas a “single-chain” custom template treats the structure as a whole. However, since it may not correctly detect the interface regions that are crucial for forming the complex structure, its effectiveness might be limited. For models that include more than two chains the AF2M success rate can decrease. For example, T1179 consists of four chains, and the lowest performing model was T1179 MultiFOLD across all quality scores. Thus, the success of the approach may decrease as the number of chain increases. It should be note that the MultiFOLD versions are AF2M based tools.

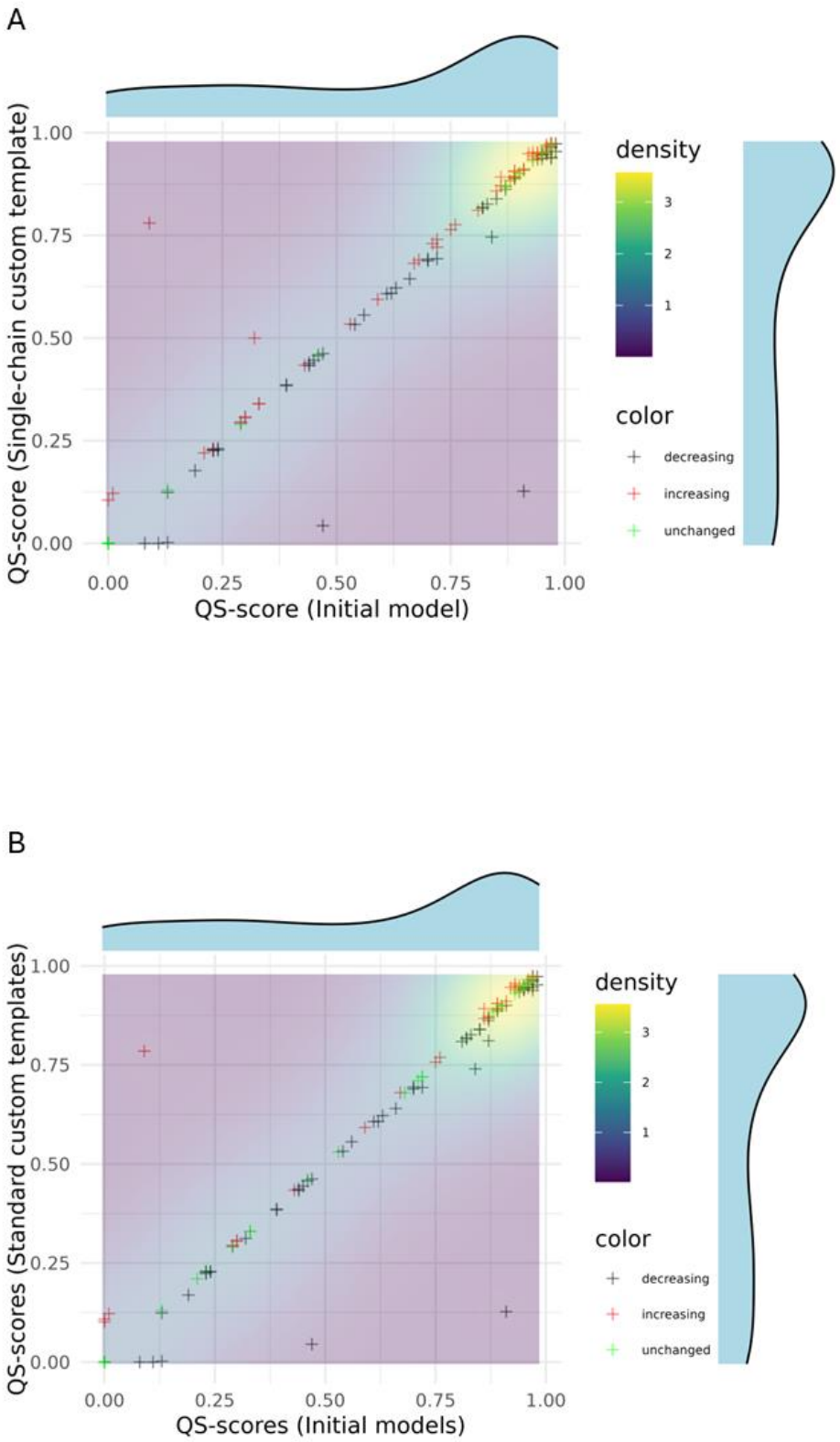


Figure 4.5 Comparison of observed QS-scores between the AF2M models using the custom template option with recycling and the initial models.
Density scatter plot presenting the comparison of QS-scores for (A) the models generated by AF2M using "single-chain" custom template versus the initial models, and also (B) the models generated by AF2M using the standard custom templates versus that of the initial models. The red pluses indicate the refined models. The black pluses indicate the unrefined models while the green pluses represent that the model quality do not change. The density scale ranges from purple to yellow. The scatter plots were drawn through R.

In terms of DockQ_wave score, it was observed that more models had lower scores compared to the QS-scores for the models. Additionally, for the models generated by AF2M using “single-chain” custom templates, 54 models showed improvement, and 57 models showed deterioration (Figure 4.6A) compared to the initial models, while for the models generated by AF2M using standard custom templates, there were improvement in 56 models and deterioration in 17 models (Figure 4.6B). It can be observed that there are more outlier models compared to the initial models, and less of a difference between using a standard or “single-chain” custom templates in terms of interface score, as evidenced by the lower number of models that show improvement. The data both after and before the use of custom template methods are summarised in Appendix Table S.4. In our previous research, it was shown that the models recycled with AF2M do not show any improvement in terms of initial scores according to Molprobity scores, mainly due to the lack of the relaxation protocol, which was removed as a control (see Chapter 3; Amber relaxation was not enabled as a control measure to ensure we were testing for the recycling effect only). Therefore, the Molprobity score evaluations were not conducted for the models in this section.

In general, the interface quality scores for NBIS-AF2-Multimer initial models tended to either improve or remain consistent after running AF2M with the “single-chain” custom templates. Nevertheless, all quality scores for the initial MultiFOLD models showed improvement. It is noteworthy that MultiFOLD, except for NBIS-AF2-Multimer, was selected as the last group. AF2M using the “single-chain” custom templates generated models with varying quality scores for the best initial group models. In Chapter 3, the MultiFOLD initial models were showed improvement with increasing recycling. However, in this chapter, AF2M run ‘auto’ recycle number for all initial models, suggesting that AF2M needs more room to explore new conformations. Yet, the coordinates of structures in good quality models may impose stringent restraints for AF2M`s optimization protocol.

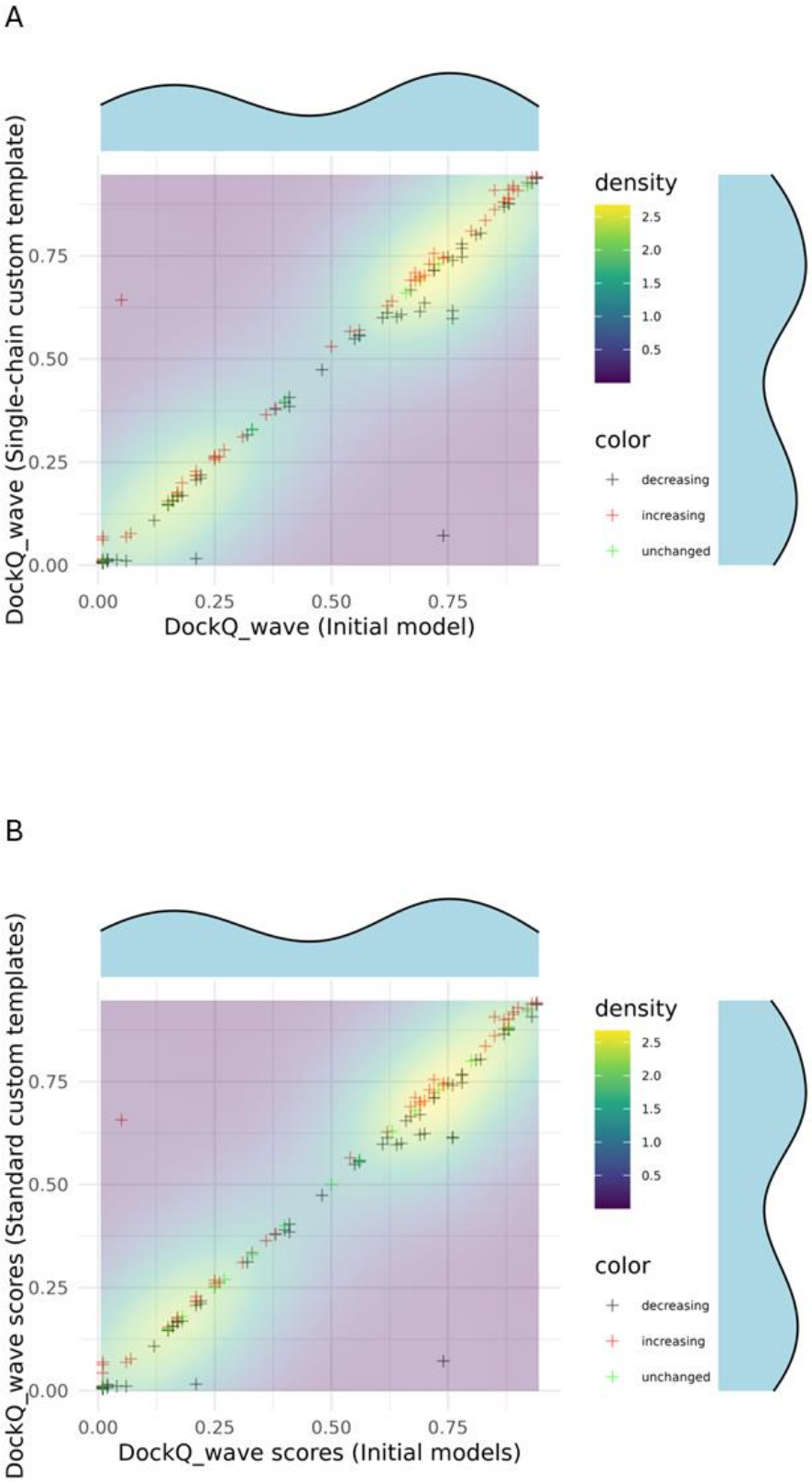


Figure 4.6 Comparison of observed DockQ_wave scores between the AF2M models using the custom template option with recycling and the initial models.

Density scatter plot presenting the comparison of DockQ_wave scores for (A) the models generated by AF2M using “single-chain” custom template versus that of initial models, and also (B) the models generated by AF2M using the standard custom templates versus that of the initial models. The red pluses indicate the refined models. The black pluses indicate the unrefined models while the green pluses represent that the model quality do not change. The density scale ranges from purple to yellow. The scatter plots were drawn through R.

4.3.2 The impact of using "Filtered Custom MSAs" on the quality of predicted quaternary structure of proteins

The evaluation was based on structure quality using the four different scores for the CASP14 and CASP15 multimeric models. An improvement in the structure quality score indicates that using MSAs with filtered disordered residues can generate higher quality models compared with including the entire homologous sequences in the MSA. Hence, the quality of 3D protein structures generated by AF2M using the different filtered MSA methods can be evaluated by comparing them with the observed structures, in the same way as we have previously done. However, to evaluate the effect of using disorder filtered custom MSAs, a target-by-target based evaluation was preferred rather than the cumulative score of models since each target includes different disorder types. As detailed in Table 4.3, 13 CASP15 and 6 CASP14 homomeric targets were used. However, in addition to the research indicating the effectiveness of MSA-pairing (Bryant, Pozzati, & Elofsson, 2022), there is also research suggesting that it may not be particularly effective in large protein structures (Bryant, 2023); this dilemma constitutes a separate topic of discussion. Hence, heterometric CASP targets were not evaluated since AF2M uses MSA-pairing to obtain MSAs for heteromeric models. In the remaining section, if the AF2M is provided with the MSA inputs where short, long, domain disordered residues are filtered, it is referred to as AF2M-SF, AF2M-LF, and AF2M-DF, respectively. In addition, if AF2M utilizes its standard MSA, it is referred to as AF2-MSA (baseline).

4.3.2.1 Evaluation of AF2M score reliability with disorder filtered MSAs

To assess the predicted score reliability of multimer structures generated by AF2M-SF, AF2M-LF, and AF2M-DF, correlations were measured between the predicted model quality scores of AF2M-SF, AF2M-LF, and AF2M-DF and the observed model quality scores. These results demonstrate the score reliability of AF2M`s predicted quality is maintained using a filtered MSA. The assessment of correlation results was based on the interpretation of correlation coefficient ranges from (Akoglu, 2018), corresponding to none, weak, moderate, strong, and perfect. Figures 4.7A, B, and C show a moderate positive correlation between the observed and the predicted quality scores for three different filtered methods. In addition to Pearson's R, Kendall's tau B and Spearman's Rho correlation test were used to examine the degree of the relationship between the observed and the predicted scores of the models. The correlation analysis between the pTM-scores and the observed TM-scores shows a moderate linear positive correlation for the modelled complexes generated by AF2M-SF, AF2M-LF, AF2M-DF with Pearson's R= 0.66, 0.57, and 0.65, respectively. The positive linear correlation indicates

that the increases in the pTM-scores generated by AF2M-SF, AF2M-LF, AF2M-DF correlate with an increase in the observed TM-scores. Furthermore, the correlation analysis between the pIDDT scores and the observed IDDT scores in Figure 4.8A, B, and C show a weak and moderate linear positive correlation. Additionally, the Pearson's R = 0.28 for AF2M-LF was observed, while Pearson's R = 0.43 for AF2M-DF and Pearson's R = 0.65 for AF2M-SF was observed. This signifies that the increase in pIDDT scores generated by AF2M-SF is more highly correlated with an increase in the observed IDDT quality scores. Further test results (Kendall's tau B and Spearman's Rho correlation) for TM-scores and IDDT scores are included in the legend of Figure 4.7 and 4.8. In the general context, MSAs without disorder co-evolution information may be used to model structures, although the AF2M-SF approach may maintain more reliable predicted model quality scores due to the highest correlations between both predicted and observed scores.

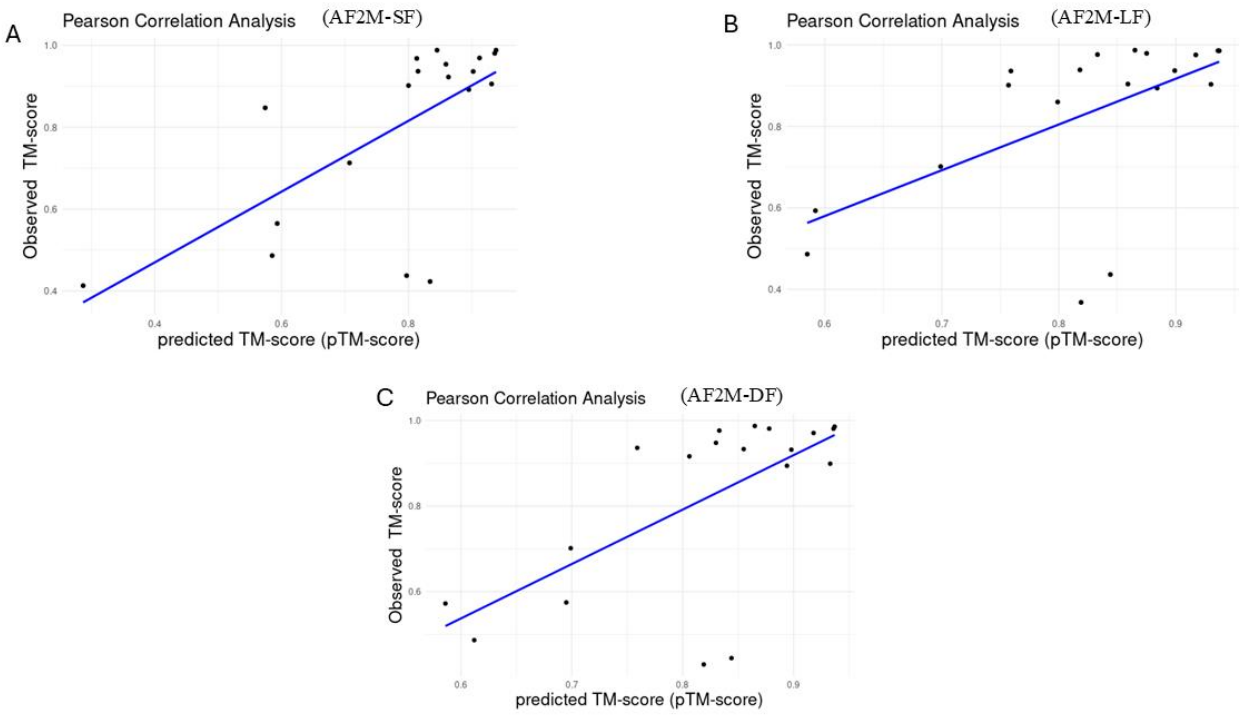


Figure 4.7 The correlation between the observed and predicted TM-score for three filtered MSA methods.

Scatter plots showing linear, positive relationship between the predicted global scores (pTM-scores) versus the observed TM-scores of the models of CASP14-15 targets generated using A) AF2M-SF B) AF2M-LF C) AF2M-DF with n = 19 multimer targets. These above scatter plots belong to the Pearson's R correlation test as an example. The Pearson's R correlation is 0.66, Kendall's tau B correlation is 0.51 and Spearman's Rho correlation is 0.70 for AF2M-SF, the Pearson's R correlation is 0.57, Kendall's tau B correlation is 0.46 and Spearman's Rho correlation is 0.62 for AF2M-LF, the Pearson's R correlation is 0.65, Kendall's tau B correlation is 0.44 and Spearman's Rho correlation is 0.59 for AF2M-DF. This plot was drawn via R.

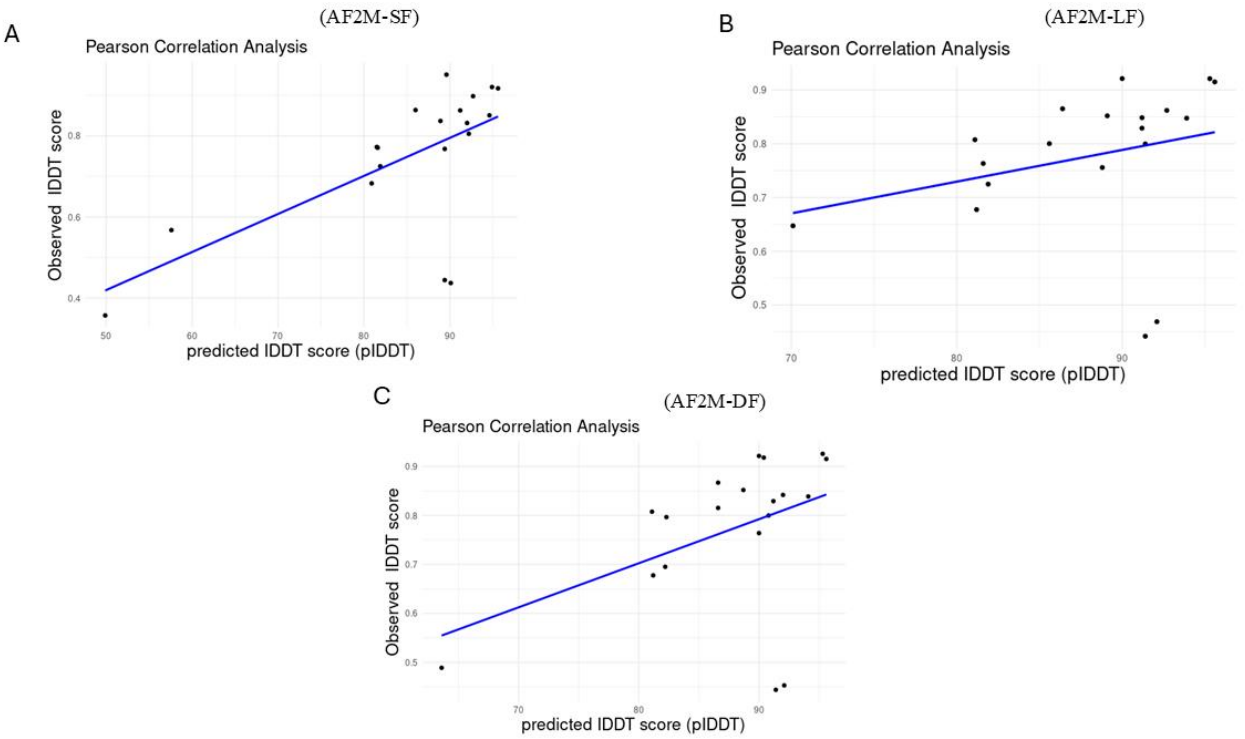


Figure 4.8 The correlation between the observed and predicted lDDT score for three Filtered MSA methods.

Scatter plots showing linear, positive relationship between the predicted global scores (pIDDT) versus the observed oligo-lDDT scores of the models of CASP14-15 targets generated using A) AF2M-SF B) AF2M-LF C) AF2M-DF with n = 19 multimer targets. These above scatter plots belong to the Pearson's R correlation test as an example. The Pearson's R correlation is 0.65, Kendall's tau B correlation is 0.40 and Spearman's Rho correlation is 0.65 for AF2M-SF, the Pearson's R correlation is 0.28, Kendall's tau B correlation is 0.48 and Spearman's Rho correlation is 0.41 for AF2M-LF, the Pearson's R correlation is 0.43, Kendall's tau B correlation is 0.30 and Spearman's Rho correlation is 0.30 for AF2M-DF. This plot was drawn via R.

In order to have confidence in the approach, the predicted scores for the three types of filtered MSA methods need to show correlations with the predicted scores for AF2M with standard MSA. Hence, the predicted scores of targets generated by AF2M using standard MSA were correlated with those of AF2M using the filtered methods. In terms of the pTM scores, the score for all the filtered methods showed a great correlation with the score for AF2M with standard MSA. Pearson's R = 0.94, 0.88, and 0.94 for AF2M-LF, AF2M-SF, and AF2M-DF were observed respectively. In terms of pIDDT score, the trend was the same and also showed the great correlation. Pearson's R = 0.89, 0.92, and 0.89 for AF2M-LF, AF2M-SF, and AF2M-DF were observed respectively. These results suggest that the predicted scores obtained by AF2M with all filtered MSA methods can be consistent with the predicted scores obtained by AF2M using the standard MSA method (See Appendix Figure S.15 and S.16).

4.3.2.2 Improvement of the multimeric structures generated by AF2M using the custom MSA complexity

The first aim was to investigate whether only including ordered residues within the custom MSA for AF2M is sufficient for generating homomeric structures. Thus, the TM-scores of the models produced by AF2M using both default values and the custom MSA were initially calculated. The reason this score was selected first is that AF2M uses the pTM based confidence scores to rank the models for multimeric targets. Additionally, the TM-score is crucial for providing information on how well the overall structure of the relevant protein improves or deteriorates compared to the observed structure after filtering in MSA. In addition to the TM-score, the IDDT is used as a superposition free score, which gives more of an indication if the improvement of the local regions within models.

An improvement in the models was observed for 19 targets according to the four different values when the short, long, or domain options of IUPred3 were used for screening the input MSA. When the filtered disorder sequences were applied to sequences in the MSA, models for 11 out of the 19 targets (58%) were improved, as shown in Figure 4.9. The remaining models did not exhibit a significant decrease in the TM-scores, except for T1038 for AF2M-SF and T1123 for AF2M-DF. Notably, among these models, 9, 11, and 9 models were improved using AF2M with the short, long, and domain option in terms of the TM-scores. In terms of the IDDT scores, the models for over half of targets were improved, (12 out of 19 models; 63%), as shown in Figure 4.10. The models for the rest of the targets did not exhibit a significant decrease in the IDDT scores, except for T1038 which had a low IDDT score for AF2M-SF (<0.4). Among 19 models, 9 models were improved using AF2M-SF. In addition, 13 models were refined using AF2M-LF, while the IDDT scores for 11 models were increased using AF2M-DF. The IDDT and TM-scores for T1038 exhibited the same trend. However, the IDDT of T1187 was decreased for AF2M-DF, although the TM-score for T1187 remained the same across all filtering methods.

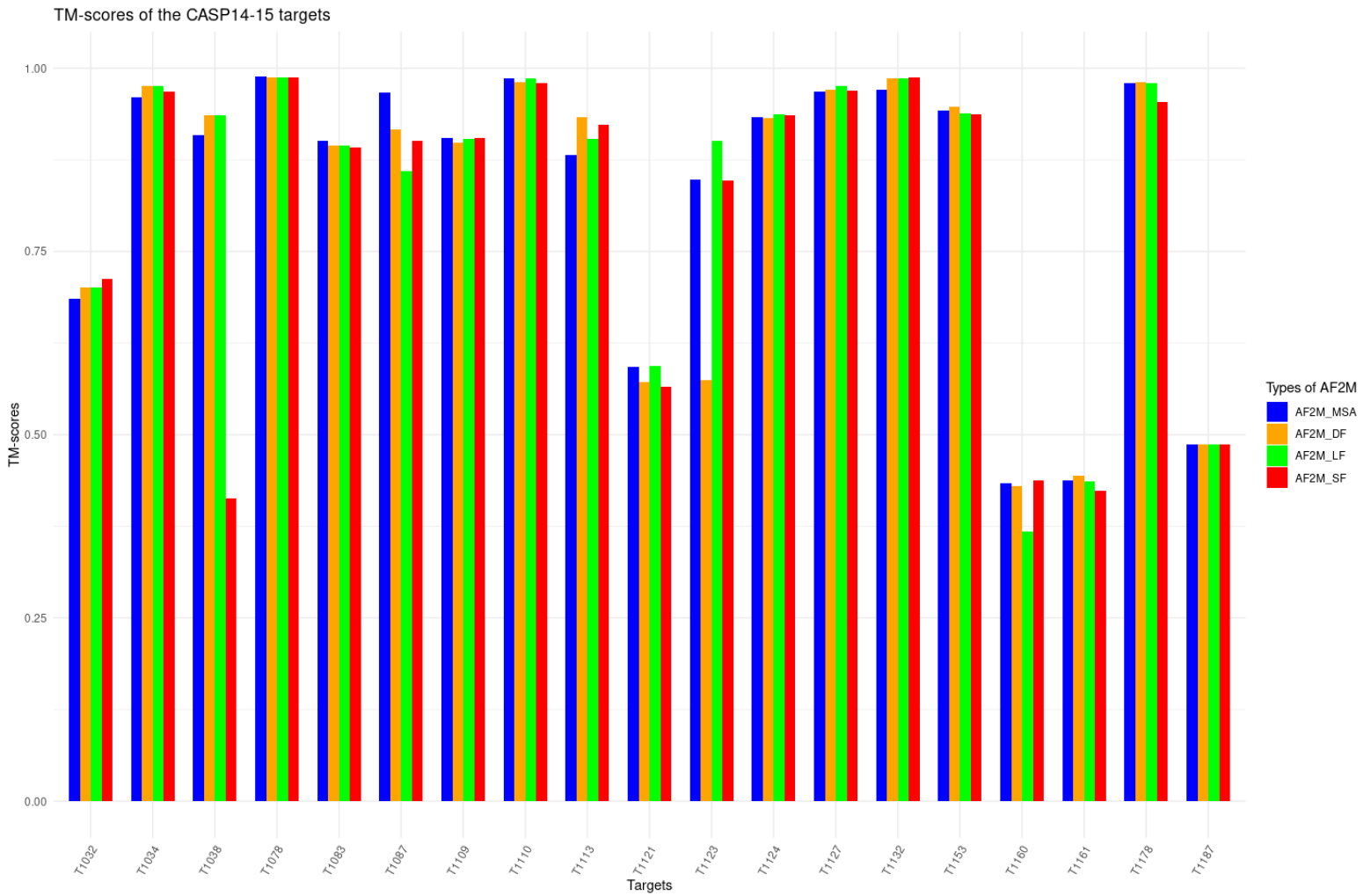


Figure 4.9 The global scores of models generated by AF2M, and AF2M with three filtered MSAs.

The model-based bar charts comparing the TM-scores for the models generated by AF2M using the filtered MSA methods to the models generated by AF2M using the default MSA. Each colour represents a different filter method for disordered residues in the MSA. Blue, red, green, and orange bar charts represent the quality scores for the CASP14-15 models generated by AF2M-MSA (indicating the use of standard MSA method), AF2M-SF (indicating the use of filtered short disordered MSA method), AF2M-LF (indicating the use of filtered long disorder MSA method), and AF2M-DF (indicating the use of filtered domain disordered MSA method), respectively. All quality scores range from a minimum value of 0 to a maximum value of 1. The bar chart was created using R.

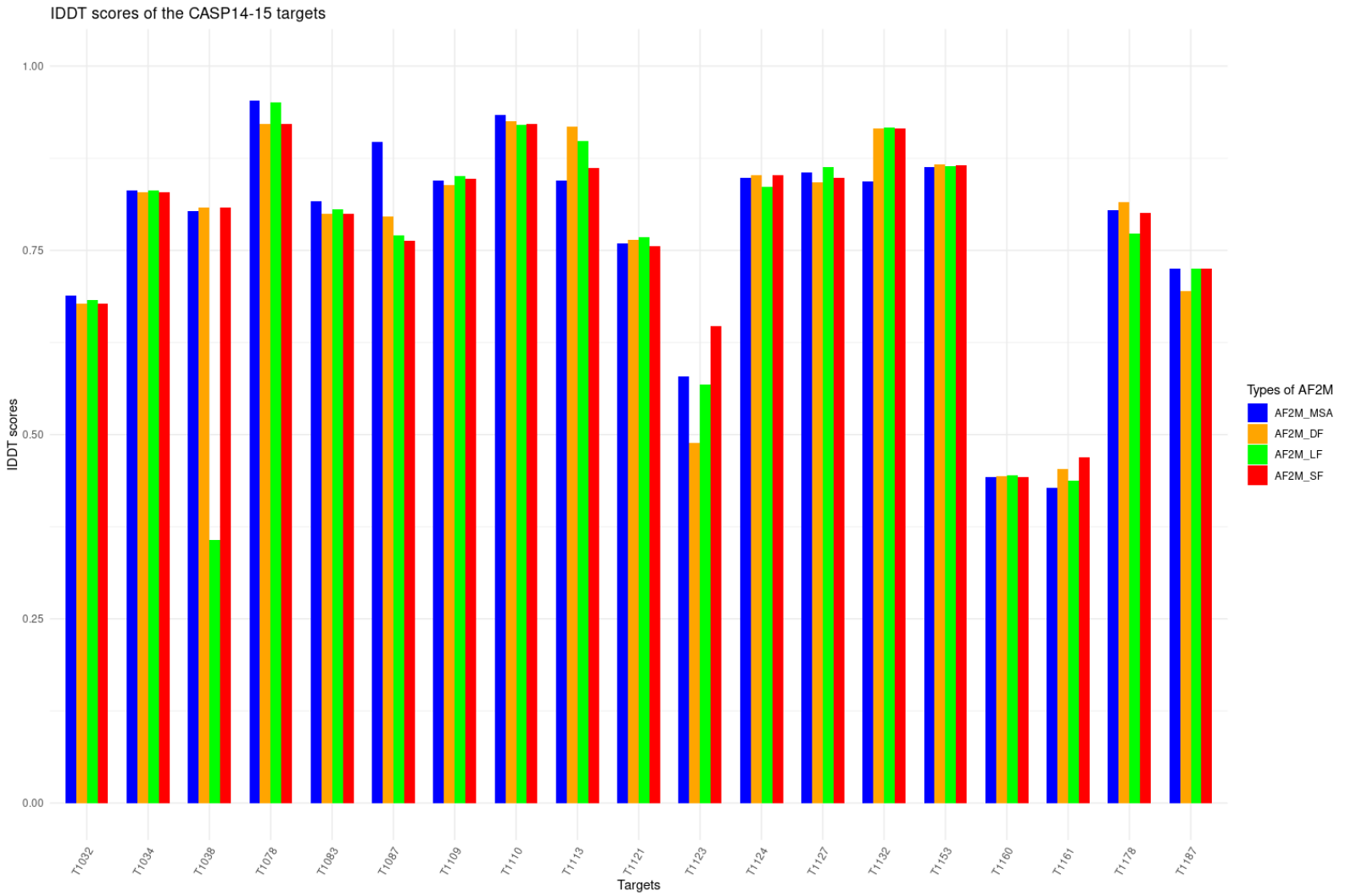


Figure 4.10 The local scores of models generated by AF2M, and AF2M with three filtered MSAs.

The model-based bar charts comparing the IDDT scores for the models generated by AF2M using the filtered MSA methods to the models generated by AF2M using the default MSA. Each colour represents a different filter method for disordered residues in the MSA. Blue, red, green, and orange bar charts represent the quality scores for the CASP14-15 models generated by AF2M-MSA (indicating the use of standard MSA method), AF2M-SF (indicating the use of filtered short disordered MSA method), AF2M-LF (indicating the use of filtered long disorder MSA method), and AF2M-DF (indicating the use of filtered domain disordered MSA method), respectively. All quality scores range from a minimum value of 0 to a maximum value of 1. The bar chart was created using R.

Two distinct interface quality scores were employed to assess the improvement of the interface regions in the models generated by AF2M, particularly when disorder information was not provided. The first score utilized was the QS-score, and the second one was the DockQ_wave score. In Figure 4.11, 12 out of 19 models (63%) were observed to improve for at least one filtering method. It is noteworthy that, unlike general structure scores (TM-scores and IDDT scores), when there was no disordered residue information was included in the MSA, either a strict improvement or strict deterioration in the models was observed, except for T1083, T1109 and T1153. The TM-scores and IDDT scores did not change greatly when the QS-scores were not much lower or much higher than that of baseline. When the DockQ_wave scores are examined in Figure 4.12 it can be seen that models from AF2M-DF demonstrated improvements in their interface regions, with 12 out of 19 models (63%) showing improvements with at least one filtering methods. Out of the 19 targets, 9, 6, and 8 targets showed improved models using AF2M with the short, long, and domain residue options, respectively. When comparing both interface scores, although the total improved models is same, discrepancies in the number of improved models were observed in terms of the filtered MSA methods. This could be because DockQ_wave score is more sensitive than QS-score. Especially, when targets are not fully resolved, QS-score can be problematic for targets due to non-symmetric interface contacts. However, DockQ_wave averages weight of interface DockQ scores representing the number of interface contacts (Studer et al., 2023) Ultimately, it can be observed that AF2M model quality does not necessarily depend on whether disordered residue information is included in the MSA or not, as the scores exhibit variability depending on the target structure.

The models for T1123 showed intriguing trend where the use of filtered domain disordered residues decreased in the TM-score and IDDT score, yet the QS-score for the model increased. One explanation for the improved TM-scores could be the presence of structures transitioning from disorder to order in the interface area when they form a complex. Considering that short disordered residues can be flexible linkers or loops (Monzon et al., 2020; Necci et al., 2018), in the monomeric structure for T1123, these residues were flexible linkers (up to 30 residues), which was also a domain disordered area (See Appendix Figure S.17). With transition from disorder to order, the increase in QS-score was observed. Moreover, While the TM-score increased as a result of applying long disordered residue filtering, no change was observed in the TM-score when short disordered residue filtering was applied. This indicates that the regular regions within the long disordered region are better improved. In addition, the TM-score is above 0.5 for both filtering methods, indicating that the predicted structures are not randomly predicted (Figure 4.9). This supports the result showing the increase in local residue improvement (IDDT) when using the short disordered residues filter

(Figure 4.10). It should be noted that short disordered residues can also be present within the long disordered regions (See Appendix Figure S.17) (Monzon et al., 2020). Target T1123 was Human Astrovirus MLB1 protein. Interestingly, for this target, it has been emphasized that template-based ML methods are more effective than homologous sequenced-based ML methods (Delgado-Cunningham et al., 2022). Thus, the need for a template that narrows down the sampling of disordered structures was indicated to provide better guidance in modelling than protein sequences (Delgado-Cunningham et al., 2022).

Most tools provide numerous decoys for given protein sequences. Among the most intriguing tools for obtaining the decoys are MassiveFold (Brysbaert et al., 2024) and AFSample (Wallner, 2023b). Generating lots of conformational structure for protein sequence requires huge computational calculation, which can be time consuming. MassiveFold addressed this issue by using lots of batches to model each conformational structure of protein in parallel, along with utilizing the combination of AF2 versions (Brysbaert et al., 2024). In addition, using different parameters they managed to obtain a wider variety of protein structures. This supports the importance of for using different versions of AF2M, as discussed in Chapter 2 and Chapter 3. Conformational sampling, especially for multichain protein structures like antigen-antibody interactions, has become more feasible with advances in computational methods such as AF2. Hence, it can be crucial to evaluate the efficiency of AF2, along with using different parameters, for generating structures specifically for challenging protein targets, rather than just general protein structure prediction.

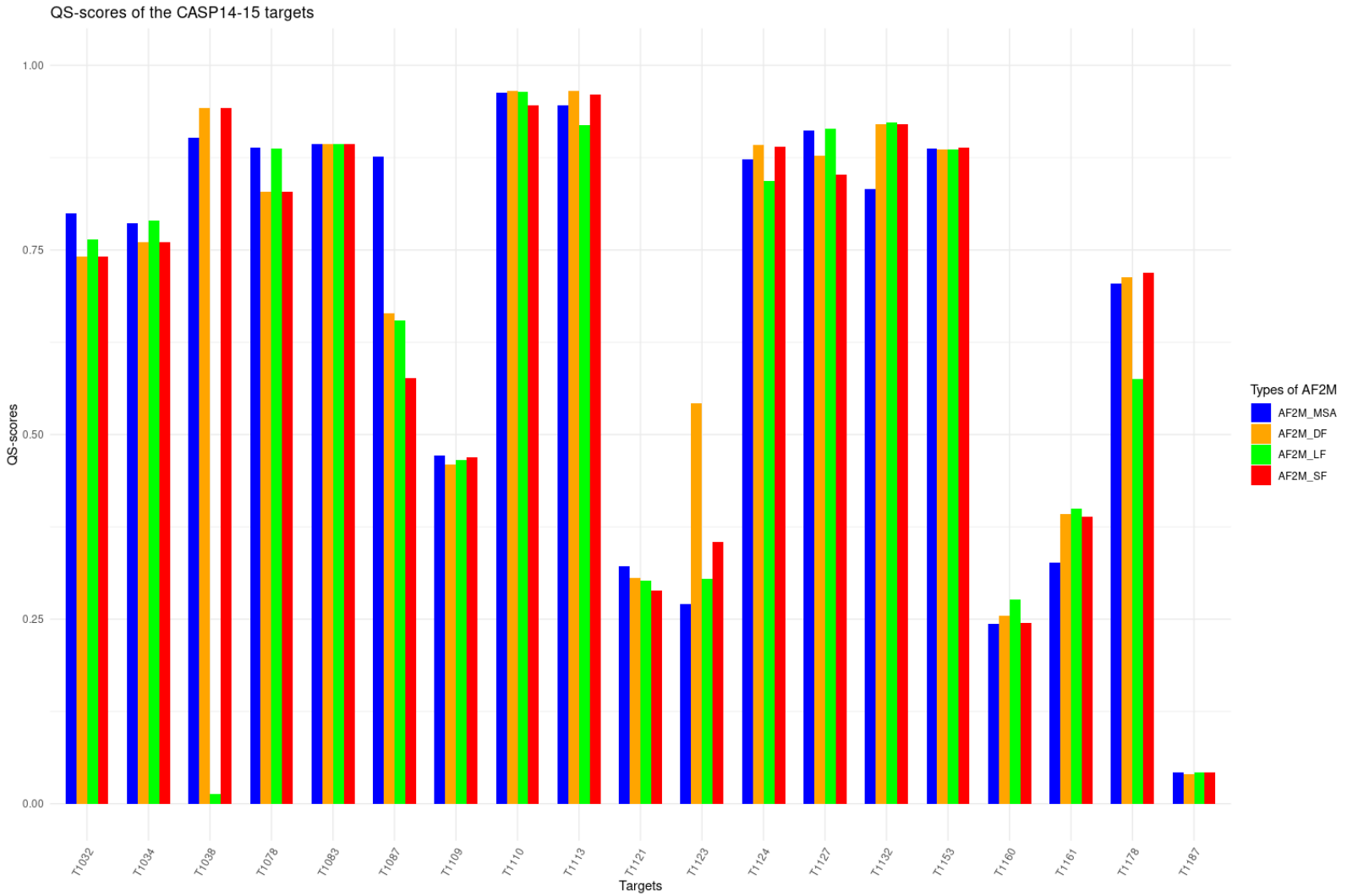


Figure 4.11 The interface QS-scores of models generated by AF2M, and AF2M with three filtered MSA.

The model-based bar charts comparing the QS-scores for the models generated by AF2M using the filtered MSA methods to the models generated by AF2M using the default MSA. Each colour represents a different filter method for disordered residues in the MSA. Blue, red, green, and orange bar charts represent the quality scores for the CASP14-15 models generated by AF2M-MSA (indicating the use of standard MSA method), AF2M-SF (indicating the use of filtered short disordered MSA method), AF2M-LF (indicating the use of filtered long disorder MSA method), and AF2M-DF (indicating the use of filtered domain disordered MSA method), respectively. All quality scores range from a minimum value of 0 to a maximum value of 1. The bar chart was created using R.

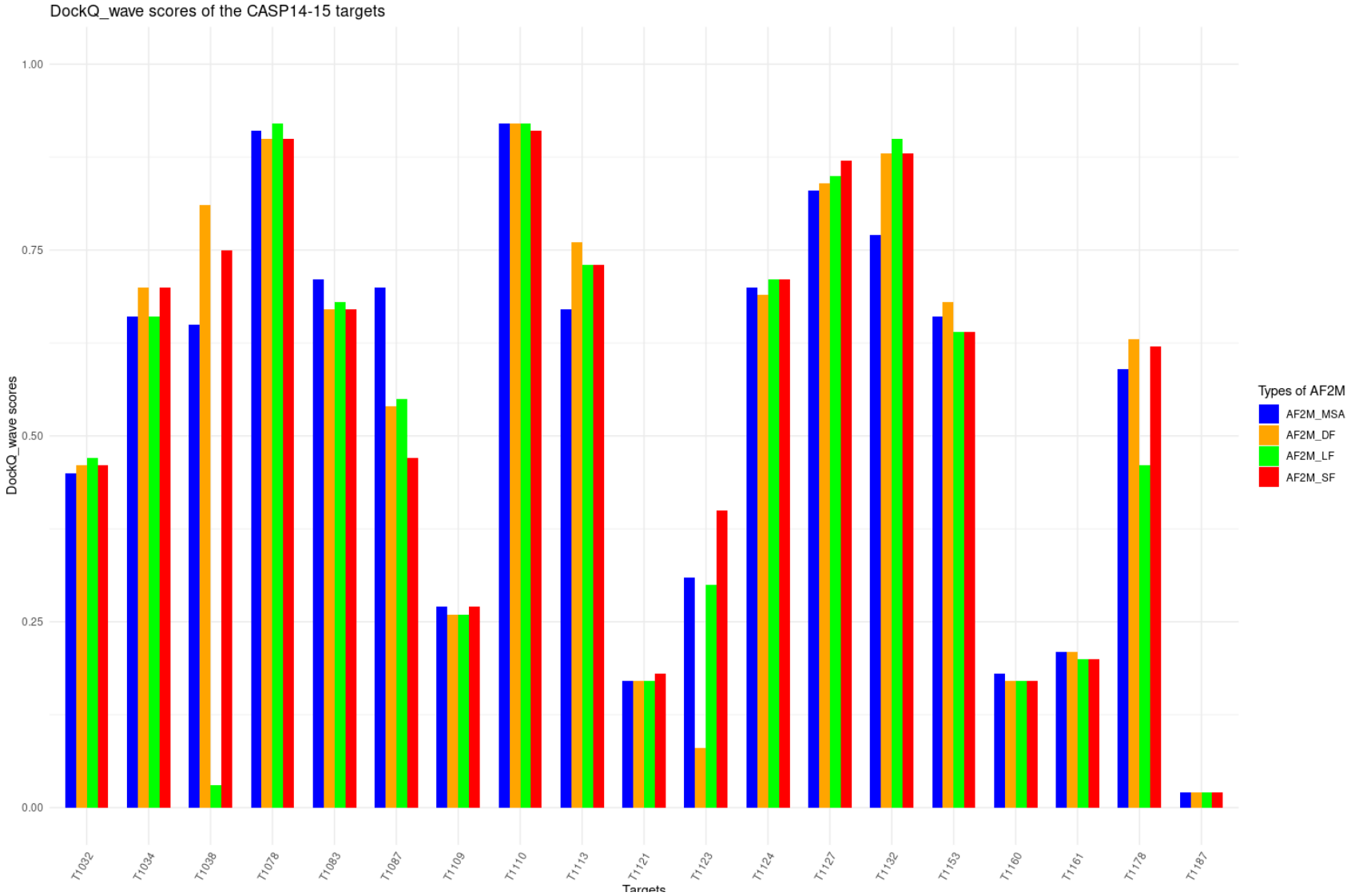


Figure 4.12 The interface DockQ_wave scores of models generated by AF2M, and AF2M with three filtered MSA.

The model-based bar charts comparing the DockQ_wave scores for the models generated by AF2M using the filtered MSA methods to the models generated by AF2M using the default MSA. Each colour represents a different filter method for disordered residues in the MSA. Blue, red, green, and orange bar charts represent the quality scores for the CASP14-15 models generated by AF2M-MSA (indicating the use of standard MSA method), AF2M-SF (indicating the use of filtered short disordered MSA method), AF2M-LF (indicating the use of filtered long disorder MSA method), and AF2M-DF (indicating the use of filtered domain disordered MSA method), respectively. All quality scores range from a minimum value of 0 to a maximum value of 1. The bar chart was created using R.

Obviously, during the generation of AF2M models, the relaxation method to alleviate clashes is left as an optional choice for the user. This leads to an increase in clash scores within the structures. However, Wallner's research (Wallner, 2023a) has highlighted the need for more effective conformational sampling based on dynamic simulations, rather than simple relaxation methods like minimization, to improve modelled structures. Currently, the force fields are not too sufficient to introduce interchain information. This result supports ignoring the evaluation of the Molprobity scores for models (See Figure 4.13 on Page 143). By applying filters to sequences in the MSA, the decrease in the MSA depth could potentially affect the quality of protein structures. However, it was observed that most models showed improved quality. Yin et al. (2022) indicate that for large proteins with small interface sizes relative to their overall size, these factors may be more significant than the limited MSA depth in achieving better protein structures. While general disorder filtering or increasing MSA complexity did not lead to optimization for AF2M on multimeric models, a notable improvement was observed when specific disorder patterns unique to each target were identified beforehand, and a custom MSA was prepared for AF2M. Chains that form homomeric structures often undergo structural rearrangements during complex formation. One frequently observed phenomenon is the transition of disordered regions to ordered structures, aiming to achieve the lowest energy confirmation (Mendoza-Espinosa et al., 2009).

Our observations from Chapters 2 and 3 indicate that homomeric structures require a longer modelling period with AF2M. Therefore, the impact of MSA filtering on the modelling of homomers was investigated. The models generated by AF2M-MSA (cumulative time: 384 m) exhibited shorter modelling time for homomers compared to the other three filtering methods (cumulative time of AF2M-SF, AF2M-LF, AF2M-DF: 520 m, 470 m, 408 m, respectively). This difference can be attributed to AF2M potentially needing more extensive exploration of conformational space for multimer models, resulting in a higher tendency for recycling and consequently a longer modelling time. In addition, considering proteins have various conformers to fulfil their specific function, disregarding residues corresponding to small disorder regions in the MSA can result in the lost disorder-to-order information. Especially, conformational changes within side chain and backbone including disorder-to-order structures can be challenging for obtaining high quality models for docking methods. In response to this challenge, DNNs based methods like AF2 shows an advantage due to its end-to-end method (Yin et al., 2022).

Quaternary structures within a protein family exhibit less conservation compared to monomeric structures. When two chains interact, and if these complexes play functional roles, preserving the interface structures becomes crucial. Hence, there will be an increase in evolutionary

constraints in the interface region. In such cases, residues involved in contacts are expected to be structured rather than unstructured like disordered residues. However, even with high sequence similarity, protein chains can form a diverse array of possible quaternary structures. Therefore, using homologous structures as templates in homology modelling, may be preferable over employing various MSAs, under the assumption that the interface evolution is more predictable (Bertoni et al., 2017). Moreover, by ignoring disordered residues in the MSAs, it is possible to redirect AF2M`s attention towards structured residues in the attention mechanism. R.Yin *et al.* (2022) emphasized that the performance of AF2M and ColabFold (Mirdita et al., 2022) correlates with the specific case studies. AF2M achieves accurate structure predictions using its own scoring metrics, such as pTM-score, pIDDT, and PAE, which can be influenced by the characteristics of the protein under study. Moreover, differences among various MSA inputs may be masked when the AF2M training set includes the relevant complexes to be analysed (Yin et al., 2022). Therefore, utilizing the CASP data can be particularly effective in order to being most updated models in the PDB.

T1038 was one of the most striking targets with models that had drastically worse TM-scores after filtering the MSA for the short disordered regions. When the MSA of T1038 was filtered in terms of the long and domain disorder residues, improvements in models were observed, however the model generated using the short disorder filtered MSA resulted in a sharp decline in the TM-score. Although all three quality scores were identical for the model, the DockQ_wave scores were different when comparing AF2M-LF to AF2M-DF in Figure 4.13. T1038 in the CASP14 competitions was a homodimer structure of TSWV glycoprotein. In its monomeric form, a short disorder structure is experimentally observed at the N- and C-terminal. However, despite the presence of long disorder residues in the N-terminal of its homomeric form (indicated by missing electron density between residues 36-107 in the experimental structure) (Bahat et al., 2020), the removal of residues from the MSA corresponding to this region may suggest that AF2M can generate the region of structures by utilizing the learned information from protein tunnel indirectly (Chakravarty et al., 2023). Another possibility could be related to the complexity of the MSA. By partially designing complexity in the MSA through filtering disordered co-evolution information, MSA complexity can be used to force AF2M to predict better protein structures.

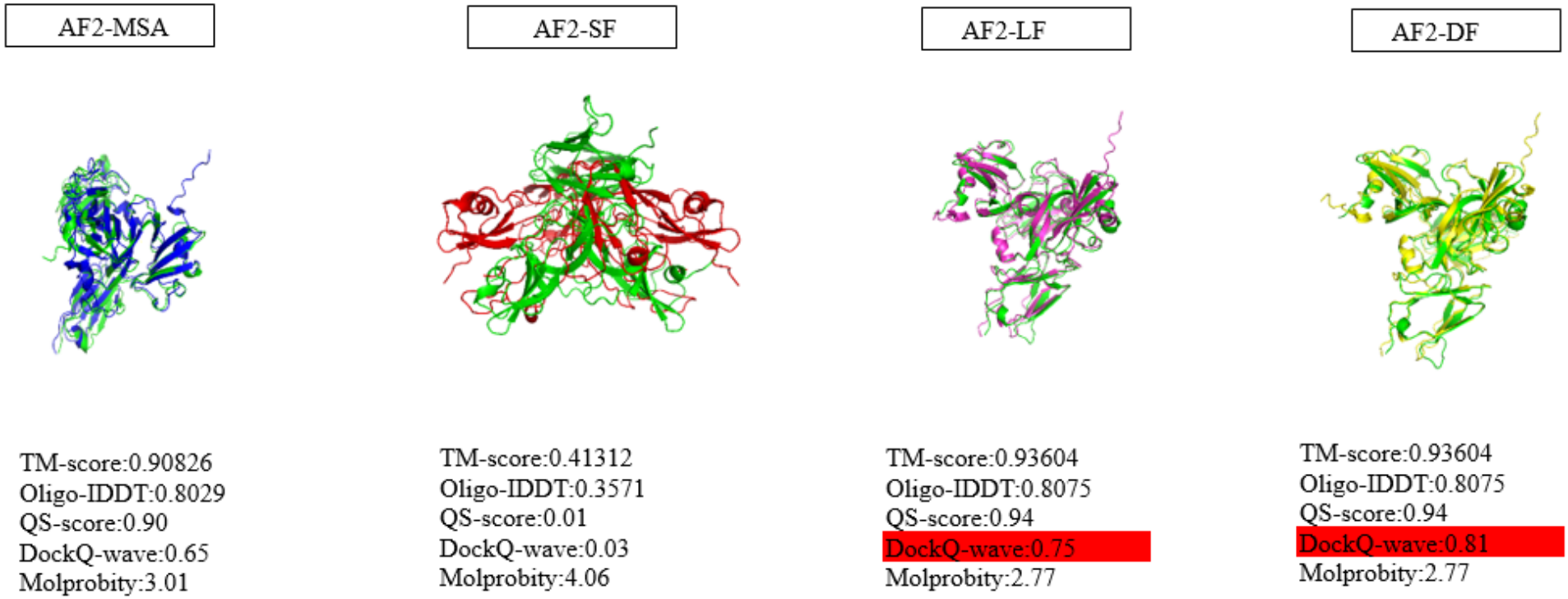


Figure 4.13 The comparison of five quality scores for the models generated by AF2M using default MSA and three filtered MSA.

The models of the T1038 multimeric target (CASP14-A2) are depicted above. The models which generated by AF2M-MSA (indicating the use of standard MSA method), AF2M-SF (indicating the use of filtered short disordered MSA method), AF2M-LF (indicating the use of filtered long disorder MSA method), and AF2M-DF (indicating the use of filtered domain disordered MSA method) were aligned with the reference structure. While the reference structures are shown in green, the AF2M-MSA, AF2M-SF, AF2M-LF, AF2M-DF models are represented in blue, red, magenta, and yellow colours, respectively. Below each model, the observed scores corresponding to the models and the baseline structure. Notably, the AF2M-SF model exhibits a markedly different alignment compared to the other models, resulting in considerably lower observed scores. For the models generated by AF2M-LF and AF2M-SF, all scores except for the DockQ_wave scores, are identical. This underscores DockQ_wave`s greater sensitivity in evaluating interface regions. The detailed methods for scoring the T1038 multimeric model are mentioned in the "method" section. The PDB structures were visualized and aligned using PyMOL.

4.4 Conclusions

In this chapter, the impact of using different custom inputs on AF2M model quality was investigated. Specifically, two types of custom inputs were examined: “single-chain” custom templates and disordered residue filtered custom MSAs. The improvement effect was measured for the models generated by AF2M using “monomeric” template models derived from multi chain models, as well as custom MSAs with the residues corresponding to disorder regions filtered out. The rationale for choosing these two custom input methods is as follows: for custom templates, by default AF2M consider each chain separately, even if multimer templates are provided, leading to a potential loss of interface information which may not occur if they are considered instead as a single-chain templates, for custom MSAs, ignoring the disordered residues in the MSA will reduce the redundancy of evolutionary information, increasing complexity and potentially leading to lower quality MSA, which may lead to prediction of better models.

Firstly, when “single-chain” custom templates were used, the cumulative TM-scores and IDDT scores were higher than those of the initial models, however the initial models had higher interface quality scores. Hence, using AF2M with a “single-chain” custom templates may be more beneficial for models with fewer chains. However, the “single-chain” custom template inputs led to greater improvements of the AF2M models compared to the standard custom template inputs. In addition, different custom template applications may vary the model structures. Rather than relying solely on better custom templates, the AF2M architecture could possibly be adjusted to improve interface regions via an interface-based attention mechanism.

Secondly, when the disorder filtered custom MSAs were employed, higher quality models were generated compared to standard AF2M-MSA when at least one of the filtered MSAs was used. Our study suggests that, rather than employing the standard MSA for all protein structures, an effective strategy for AF2M may involve initially filtering homologous sequences of the relevant protein based on disorder information. Target-specific information, particularly for custom MSAs, has proven to be beneficial for AF2M, indicating its usefulness in a targeted manner rather than applied generally for all protein models. This suggests that while target-specific disorder-filtered MSA may be advantageous for AF2M models, it should not be applied universally for all models. Following the launch of AF2M, the process of protein structure modelling has transitioned from a challenging phase to one that facilitates specific analyses, such as protein binding site identification and protein design. The general process now involves the need for a more improved model to effectively utilize structures generated by AF2M, emphasizing the preference for target-specific methods to improve models. In our approach,

we introduce a level of disorder in the MSA input specifically tailored for AF2M. Similar work conducted by (Petti et al., 2022) and colleagues also supports our findings, noting the more effective modelling capabilities of AF2M with custom MSAs.

The markers of disorder in the PDB structures may not be ideal for the specified definitions, as defining disorder solely based on regions lacking spatial coordinates or exhibiting high mobility may no longer be suitable. Considering that AF2M modelling derives its power from co-evolutionary information, truly disordered residues may persist in the MSA even after filtering for disorder-order separation, potentially causing a bias effect in the results. In the past, it was observed that the spatial coordinates of flexible hinged structural domains in crystal lattices were unclear, leading to annotation errors (Huber, 1979). With the current version of ColabFold, the v3 weights were released. By utilizing a combination of these weight, more conformation structures can be researched. Alternatively, a strategy aiming to identify the best structure could involve generating decoy structures using either AF2M with different custom template or with MSA options (Brysbaert et al., 2024; Wallner, 2023a). Model quality prediction tools could then be employed to determine the optimal structure among these decoy structures. Although there has been substantial research on the effect of templates on AF2M models, a more detailed analysis is required to evaluate the custom MSAs without disorder information on AF2M models. Additionally, the lack of data for benchmarking AF2M on proteins with disordered regions is an issue, given that it was trained on PDB structures. However, an NMR dataset could be designed for benchmarking AF2M since it was trained on X-Ray and Cryo-EM datasets. This approach would provide a more comprehensive evaluation of the performance of AF2M models in handling disorder information.

Finally, the results in this chapter suggest that it may not be possible to create the ideal MSAs for every target using generic setting. Filtering out specific types of disordered regions within the MSA can lead to better quaternary structure predictions for specific targets. Additionally, preserving the quaternary structural information for interacting chains with “single-chain” custom template inputs can lead to greater improvements in AF2M models with fewer chains compared to using the standard custom template inputs.

Chapter 5: Performance Comparison of MultiFOLD1 and MultiFOLD2 Servers in the CAMEO-BETA project

Performance data for our MultiFOLD server has been published in *Nucleic Acids Research:*

Liam J McGuffin, Nicholas S Edmunds, **Ahmet G Genc**, Shuaa M A Alharbi, Bajuna R Salehe, Recep Adiyaman, Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers, *Nucleic Acids Research*, Volume 51, Issue W1, 5 July 2023, Pages W274–W280, https://doi.org/10.1093/nar/gkad297

Author contributions:

Liam J. McGuffin: Idea development, Data management, Structured analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualisation drafting, Revising, and Editing the manuscript.

Nicholas S. Edmunds: Idea development, Conducting research, Project management, Utilizing resources and computational tools, Providing oversight, Drafting, Revising, and Editing the manuscript.

Ahmet G. Genc: Idea development, Data management, Structured analysis, Securing funding, Conducting research, Applying research methods, Project management, Utilizing resources and computational tools, Providing oversight, Verification, Data representation, Drafting, Revising, and Editing the manuscript.

Shuaa M. A. Alharbi: Idea development, Securing funding, Conducting research, Project management, Providing oversight, Drafting, Revising, and Editing the manuscript.

Bajuna R. Salehe: Software and Revising.

Recep Adiyaman: Idea development, Conducting research, Project management, Utilizing resources and computational tools, Providing oversight, Drafting, Revising, and Editing the manuscript.

.

5.1 Background

In the post-AF2 era, various adaptations of AF2M have become integrated into different tools to address issues adjacent to the tertiary structure prediction problem, including complex protein structure prediction (McGuffin et al., 2023), protein-ligand binding sites (Gazizov et al., 2023), protein design (Goverde et al., 2023), the conformational sampling of proteins (Wallner, 2023b), drug design (Borkakoti & Thornton, 2023), protein-DNA (Yuan et al., 2022), and protein-RNA (Darai et al., 2023) interaction prediction. With each updated version of the AF2M method, better results have been obtained for the remaining problems. At the time of writing this thesis, the latest version of AF2M has been used with the version 3 weights, and ColabFold (5.5.1) has been released incorporating this weight set. The other versions of the methods continue to be tested for reliability in blind competitions like the CASP competition. In the latest CASP (CASP15), almost all prediction tools integrated a version of the AF2M algorithm into their own approaches (https://predictioncenter.org/casp15/doc/CASP15_Abstracts.pdf). As a result, the average prediction performance was higher than that of the previous CASPs. Specifically, in the assembly modelling part of the CASP15 competition, the interface contact score (ICS, also known F1-score) was twice as high as in the CASP14 competition, while the Oligo-IDDT score was approximately 33% higher than in the CASP14 competition (https://predictioncenter.org/index.cgi).

The accurate prediction of protein structures paves the way for obtaining the improved results in downstream analysis. Hence, after AF2 successfully predicted monomeric structures, it became necessary to design new tools that predict higher quality multi-chain protein structures. The efficacy of these emerging tools must be continuously tested due to the dynamic nature of protein bioinformatics field. Therefore, the prediction community organized various blind competitions to continuously evaluate new and existing prediction tools, beyond just CASP, which just provides a snapshot every 2 years (Kryshtafovych et al., 2023a). In this regard, the Continuous Automated Model EvaluatiOn project (CAMEO) serves as a more up-to-date standard for the blind evaluation of current prediction methods.

5.1.1 CAMEO-BETA: evaluation of methods for modelling complexes

CAMEO is a project that aims to provide easy access to performance information for the current modelling methods, so that life scientists may predict 3D structures more effectively.

With this project, it will be possible to obtain evaluation and validation criteria for all protein models through a comprehensive comparison of theoretical and experimental structures. The Protein Model Portal (PMP) is an adjacent project, which serves as a point of entry for the best new techniques established by CAMEO project and the community, such as novel modelling servers for producing homology models and new quality estimation servers for model validation (Haas et al., 2013).

CAMEO is similar to CASP, in that it provides a platform for blind predictions to be made and evaluated, however CAMEO makes fully automated assessments of 3D models for *weekly* releases of protein sequences, prior to the experimental structures being published in the PDB. CAMEO publishes the weekly results based on models collected over a 4-day period, evaluating approximately 100 targets over about 5 weeks. The assessment data is consistently generated simultaneously for all participating techniques, enabling developers to evaluate and cross-validate the efficiency of their techniques. Furthermore, their benchmark data are open and can be directly referenced in publications (Haas et al., 2018).

The rapid advancements of methods alongside AF2M have not marked the end of structural bioinformatics but rather have sparked new beginnings. These new beginnings, primarily downstream analyses of complexes, should encourage greater community involvement and more frequent testing of protein modelling methods. Consequently, despite the CASP competition being considered as a valuable biennial milestone and the “world championships” that galvanises the community, it has become a bottleneck in terms of large scale continuous evaluation. However, while the CASP competition evaluates methods biennially, CAMEO performs this task weekly based on the data it publishes, serving as a complementary role to the CASP competition (Haas et al., 2019). Furthermore, with the performance demonstrated by AF2M in de novo protein predictions during CASP14, many challenges related to protein folding have been largely addressed. However, issues persist particularly concerning interactions in complex structures. To shift the focus towards multi-chain complexes rather than single chains, the CAMEO-BETA project, a branch of CAMEO, has been launched (Robin et al., 2021).

With the expansion of CAMEO, CAMEO-BETA tests the capacity of participant servers to properly model the oligomeric form of a target sequence and predict its proper assembly based on amino acid sequences. Since targets are provided as only a single amino acid sequences, participants need to predict the protein's correct stoichiometry before modelling the correct folds of the subunits and interfaces between them. Thus, the modelling challenges in the complex structure effort require: (1) predicting the complex's stoichiometry; (2) predicting the 3D structures of all entities: proteins-peptides, protein-DNA, protein-RNA, and protein-ligands,

indicating their orientation and interfaces; and (3) providing per-residue confidence estimations for the model. The CAMEO-BETA category operates on an opt-in model, allowing tools that generate single-chain protein structures to participate and receive targets. The beta version of CAMEO `Structures & Complexes` is accessible at https://beta.cameo3d.org/ and registration is available to all. Since October 2020, it has been sending multiple targets comprising protein-protein, protein-RNA, protein-DNA, and protein-ligand complexes to enrolled servers once a week. Predicted structures may be submitted by servers in either PDB or mmCIF format and they are subsequently evaluated utilising an entirely automated workflow that includes measure of local fold accuracy (Oligo-lDDT) and interface acuracy scores (QS-scores). Following the expiry of targets and the release of the experimental structures in the PDB, the weekly predicted models and the observed experimental structures are provided, along with the assessment of the results through the CAMEO-BETA website (Robin et al., 2021).

5.1.2 Model quality assessment (MQA)

Protein modelling tools, designed with various algorithms and scoring techniques, can predict a wide and diverse variety of alternative structural models for a given protein sequence. Therefore, methods for assessing a protein model's quality are needed in order to select the very best predicted models. The provision of additional conformational structures through various applications of AF2M, such as dropout (Srivastava et al., 2014), has further highlighted these needs. Initially, structure prediction techniques incorporated methods for selecting the best model as a component, but in recent years, an increasing number of stand-alone techniques have been developed (Chen & Siu, 2020). Methods for model ranking include both consensus and single-model methods. Consensus methods in model evaluation calculate an average similarity score among models, with the presumption that better models exhibit greater similarity with others in the pool. Single-model approaches consider models on an individual basis, extracting output features from other tools (predicted secondary structures, solvent accessibility scores), utilizing evolutionary information from homologous sequences and physics-based information, such as energy scores (Ouyang et al., 2020).

ML-based techniques for estimating model accuracy (EMA) combine various forms of information, whereas traditional EMA methods are mainly based on energy, physicochemical or statistical factors. In recent years, significant advances in protein structure prediction have been achieved through the integration of features via DNNs, especially in accurately predicting inter-residue structural constraints. For instance, in the CASP13 competition, the use of inter-residue contact information and deep learning techniques minimised the loss of GDT-TS score

in ranking protein structural models. This approach played an important role in the success of DeepRank (Renaud et al., 2021). Similarly, in the subsequent CASP14 competition, DeepAccNet (Hiranuma et al., 2021), a deep residual network for the prediction of local quality scores, was more successful than other networks in minimising IDDT score loss (Chen et al., 2021).

When it comes to AF2M, it employs an encoder and decoder network based on evolutionary blocks to establish the connection between sequence, structure, and model quality, thereby enhancing the accuracy of model quality assessment by integrating structure, sequence, physicochemical information, and DNN architecture, as demonstrated by previous research (Liu et al., 2023). AF2M generates scoring metrics such as the pIDDT score and pTM-scores and provides information on the quality of protein structure without the need for external EMA methods (Jumper et al., 2021a). However, independent MQA still plays a critical role in protein structure prediction, because methods such as ModFOLD9 (McGuffin & Alharbi, 2024) produce more consistent scores that can be used to directly compare models ranging in quality from a variety of different methods more accurately. Therefore, they offer an unbiased comparison of models from multiple different methods regardless of the modelling approach used. This independent assessment ensures the reliability and efficiency of prediction models, which in turn affects the efficiency of target discovery and drug design (Hiranuma et al., 2021).

5.1.3 Dropout algorithm in AF2M

Further to improving upon the quality assessment of models, an additional focus of research is to improve the sampling of the conformational space for model generation. The dropout technique can be applied to standard NNs. What distinguishes the dropout method from using a default NNs is the ignoring of certain neurons within a layer, thereby increasing the attention on the remaining neurons (Figure 5.1). Using the dropout technique during inference creates a perturbation in the model by randomly excluding some information from each prediction. In this way, the robustness of the network is increased by enabling a variety of outputs to be obtained. In NN based methods, such as AF2M, randomisation of features during training is used to encourage the network to adopt different learning strategies. Activating dropout layers during the inference phase encourages the network to use alternative solutions that might otherwise be ignored and may sacrifice some predictive power to achieve a wider range of solutions. The use of dropout during inference has been proposed as a method of incorporating uncertainty and building an ensemble of models without the need for additional training time. (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). The "is_training" parameter in

ColabFold enables users to select dropout during inference, triggering the stochastic part of the model and generating a wider range of predictions. Iterating with different seeds captures the uncertainty resulting from co-evolutionary constraints obtained from MSA, allowing

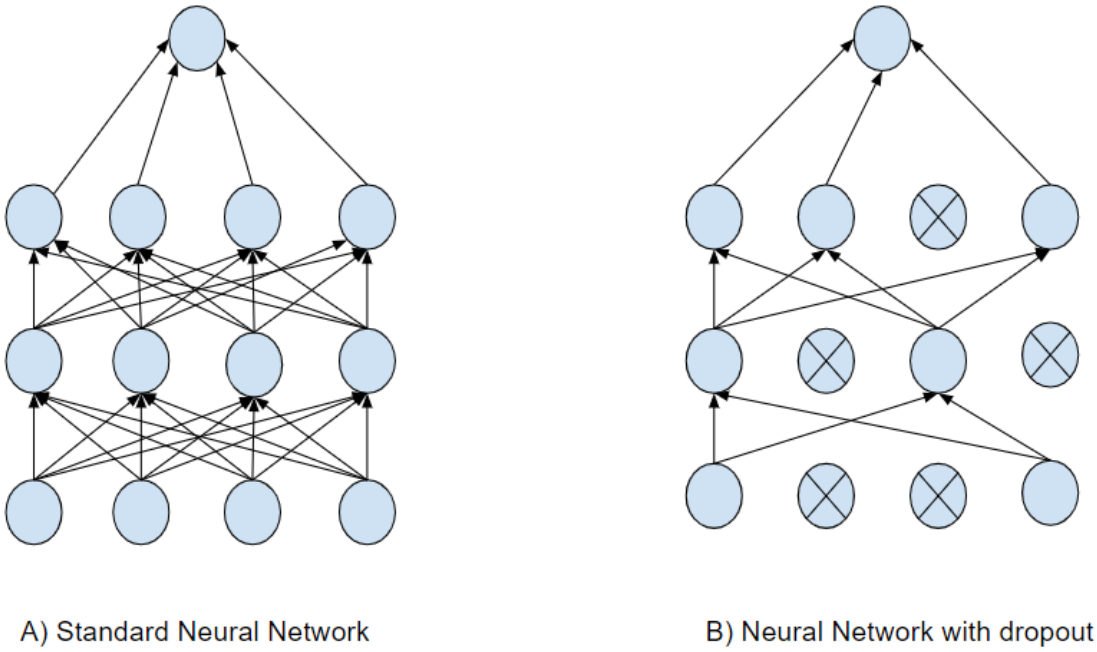


Figure 5.1 The difference between a standard NN and a NN with dropout.
improved sampling of various structure predictions (Mirdita et al., 2022) and uncertainty from the model (Gal & Ghahramani, 2016).

These images represent a) a standard fully connected NN b) a NN after applying dropout. In the standard NN, each neuron (circles) connects each neuron within the next layer, while only activated neurons connect to each other between layers in the neural network with dropout. The deactivated neurons are presented as 'Xs' in the circles.

In the AF2M algorithms, dropout is implemented with customized modifications specific to various self-attention methods and residual revisions, incorporating a dropout rate ranging from 10% to 25% according to different network modules (Jumper et al., 2021a). The dropout parameters in AF2M are as follows as shown in Figure 5.2 and 5.3:

****DropoutRowwisex:** This variation employs a dropout method known as DropoutRowwise, in which dropout masks are shared between rows [1, *N*res, *N*channel]. This technique is denoted by the operator "DropoutRowwisex," with "x" representing the dropout rate. It is used in triangular self-attention around the initial node and in residual revisions after row-wise self-attention in MSA. Similar channels are set to zero for every row in every residue (column) throughout these updates.

****DropoutColumnwisex:** This variation employs a dropout method known as DropoutColumnwise, in which dropout masks are shared between columns [*N*res, 1, *N*channel]. This technique is denoted by the operator "DropoutColumnwisex. It is employed in triangular self-attention around the last node, after residual revision. For each row, identical channels are deactivated across all columns.

These parameters are used in the two main modules of AF2M:

****Evoformer Stack:** Row-wise dropout is utilised in the main Evoformer stack during residual revisions for triangular multiplicative calculations on the pair of stacks. This configuration is preserved for both the template pair stack and the unclustered MSA stack.

****Structure Module:** Within the Structure module, dropout is implemented on the outcomes of the Transition layer and the Invariant Point Attention.

These dropout adjustments improve the capacity of the model to adapt and adjust to different structural attributes across various components. This provides higher efficiency and robustness in the face of the model sampling challenges encountered in protein structure prediction (Jumper et al., 2021a).

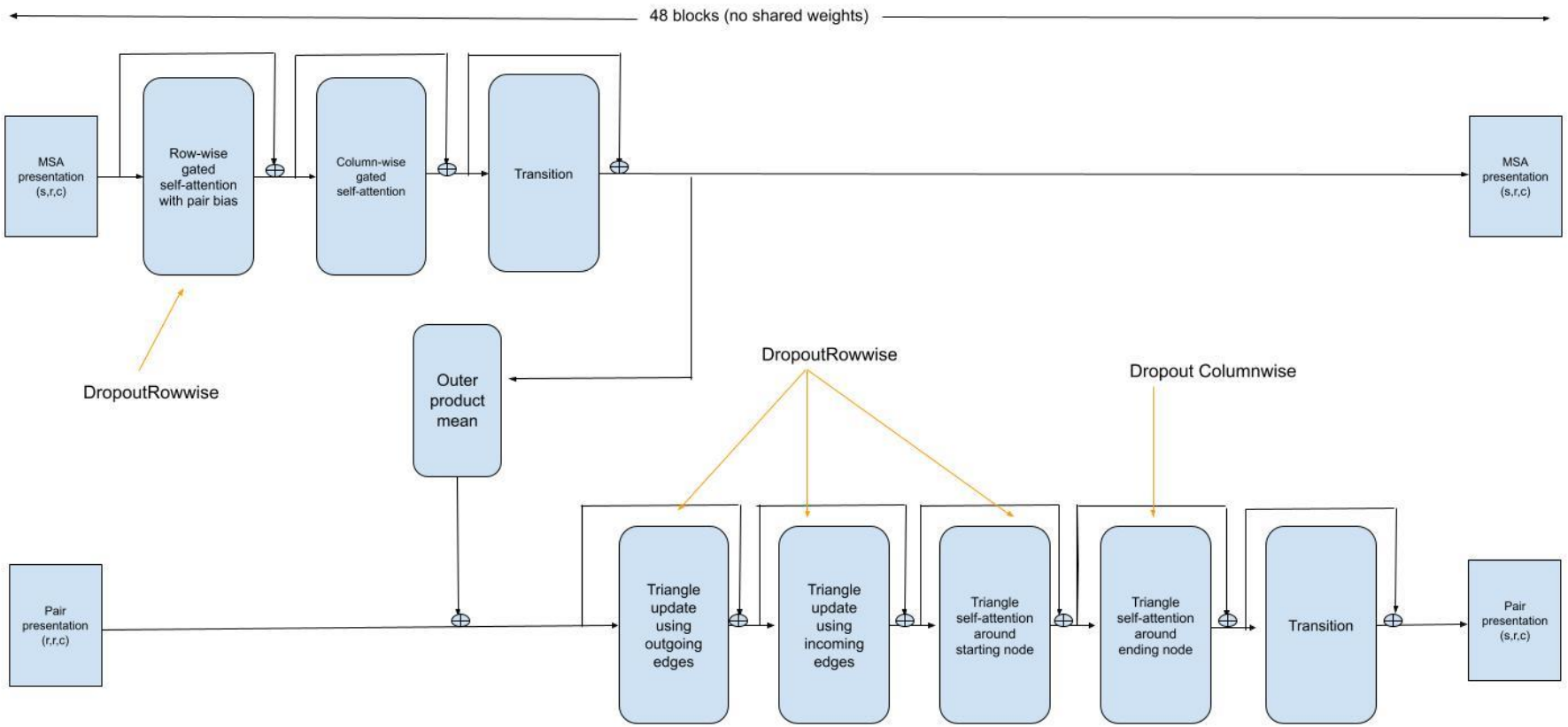


Figure 5.2 The Evoformer module of AF2M.

The figure shows one of the main modules of AF2M known as the Evoformer. Evoformer takes as input MSA and pair presentations and gives as output updated MSA and pair presentations via self attention mechanism and triangle multiplicative updates. In the Evoformer module, dropout methods specially designed for AF2M are used. DropoutRowwise is used within the row-wise self attention, the triangle self attention around initial node and the triangle updates, while DropoutColumnwise is employed the triangle self attention around last node in order to residue updates. This graphic was taken from (Jumper et al., 2021a) and prepared using Microsoft Powerpoint.

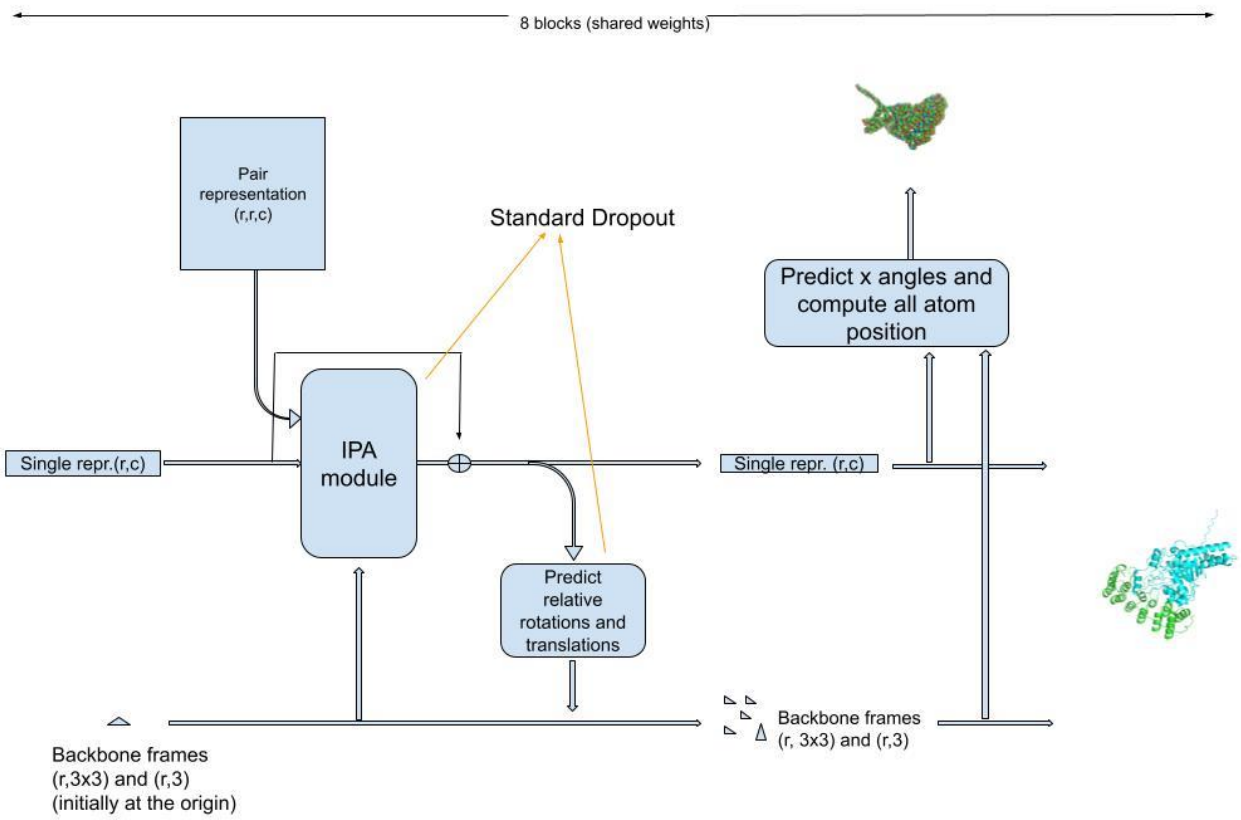


Figure 5.3 The Structure module of AF2M.

The figure shows one of the main modules of AF2M known as the Structure module. This module takes as input the first sequence within the MSA and backbone frames and gives as output 3D protein structure via IPA module and updating the backbone frames by predictig their rotations and translations. In the Structure module, the standard dropout method are used for AF2M. This graphic was taken from (Jumper et al., 2021a) and prepared using Microsoft Powerpoint.

5.1.4 RoseTTAFold2 and RoseTTAFold All-Atom

RoseTTAFold All-atom is an extension of RoseTTAFold2, designed to model proteins and other non-protein complex, as well as protein-protein interactions. The key aspect of this extension is the integration of diffusion models to simultaneously predict interactions among proteins and other components (Krishna et al., 2024). Diffusion models have gained popularity, leading to the release of a new version of AlphaFold, AF3, which integrates diffusion models following AF2M (AF3 code is not available at the time of writing). Since diffusion models were designed to predict protein-nonprotein interactions, RoseTTAFold2 remains the main specialised method for predicting protein complexes. RoseTTAFold2 was developed following the incorporation of several AF2M methods, including recycling, using FAPE loss as the predicted structure loss function, and leveraging distillation data, where highly reliable network results were used as a new training dataset to feed the network (Baek et al., 2023). Furthermore, RoseTTAFold2 incorporates various methods not found in AF2M, and both tools have been validated for their strong performance in predicting complex protein structures (Baek et al., 2023; Evans et al., 2022). These different methods include: a) the integration of a third track in the RoseTTAFold1 main block based on 3D structure, b) using biaxial attention in the 2D pair track, replacing triangle multiplication and attention in AF2M, c) the adoption of the SE3-equivariant transformer for structural revisions instead of Invariant Point Attention (IPA) used in AF2M (Baek et al., 2023).

5.1.5 The aim of study

Since the CASP14 competition, with the diminished emphasis on refinement methods such as MD simulations and the increased use of ML applications, the refinement process has been integrated into the modelling process. End-to-end DNNs, such as AF2M, produce higher quality protein structures without requiring additional refinement methods. However, given that AF2M relies on MSA mining methods and is unable to predict more than two chains effectively, there remains a need for new methods to predict multi-chain structures. Hence, a standalone multi-chain structure prediction method known as MultiFOLD1 has been developed, building on previous research presented in this thesis and elsewhere. MultiFOLD1 was blind tested and achieved good results in the CASP15 competition (McGuffin et al., 2023). Currently, MultiFOLD1 is utilized for downstream analysis such as protein-ligand interaction modelling; however, it still requires ongoing testing against other state-of-the-art and emerging methods. Therefore, MultiFOLD1 is currently being tested in the ongoing CAMEO-BETA competition.

According to the CAMEO-BETA results, MultiFOLD1 needed updates using current applications, particularly to improve the prediction of multimeric models. In addition to following the CASP15 competition, the latest version of MultiFOLD, MultiFOLD2, was released. This new version includes improved sampling using the dropout technique from AF2M, RoseTTAFold2 and RosettaFold All-Atom, the new version of ModFOLDdock (Edmunds et al., 2023) known as ModFOLDdock2 for MQA, as well as refinement of the final selected models using recycling. Overall, this chapter critically evaluates the performance of MultiFOLD1 and MultiFOLD2 in comparison with other servers and with each other, using the CAMEO-BETA data.

5.2 Methods

5.2.1 MultiFOLD1 and MultiFOLD2

MultiFOLD is a specialised tool developed for modelling multimeric structures of proteins. In the CASP15 competition, MultiFOLD performed well, ranking among the top ten servers (https://predictioncenter.org/casp15/index.cgi) (McGuffin et al., 2023). MultiFOLD consists of three main components: structure prediction and assembly, evaluation of assembled structures using quality estimation, and refinement of the quaternary structure models. In the first stage, 20 different multimeric structures were generated based on the protein sequence using localColabFold (v1.0.0) and (v1.3.0).

**The parameters to be used for LocalColabFold v.1.0.0 are listed below:

“homooligomer”, “use_ptm”, “use_turbo” “max_recycle 3", and " num_relax Top5

**The parameters to be used for LocalColabFold v.1.3.0 are listed below:

“templates”, “amber”, “num-recycle 3”, and “model-type auto”

The primary goal of using both versions was to explore the conformational space more broadly using different weights to improve sampling. It has been previously indicated that different versions have an impact on the structure. (For v1.3.0, the 'templates' and 'amber' parameters were not applied to targets longer than 1000 residues but shorter than 2500 residues.)

In the second stage, the 20 different multimeric structures modelled (5 unrelaxed and 5 relaxed models for each LocalColabFold version) were scored and ranked using ModFOLDdockR, our model quality evaluation tool, and the top 5 models were selected. Subsequently, these top 5 models were fed back into LocalColabFold as templates. All models were prepared as mmCIF

files via MAXIT, and then following appropriate recycling, the final refined models were obtained. For the refinement protocol, LocalColabFold v1.3.0 was used with the template option. (In CASP15, any targets with sequences longer than 2500 residues were split into segments and modelled separately, due to limited GPU capacity. After submission to MultiFOLD, these divided structures were manually combined using PyMOL to generate the final models of the protein.)

The two sets of refinement parameters to be used for LocalColabFold v.1.3.0 are:

**“custom_templates”, “amber” if < targets with 1000 residues, “num-recycle 12”, and “model-type auto”

**“custom_templates”, “no_amber if > targets with 1000 residues”, “num-recycle 3”, and “model-type auto”

For MultiFOLD2, the method was similar to MultiFOLD1. The main difference was the integration of RoseTTAFold2, RoseTTAFold All-atom, AF2M dropout into along with the two previous LocalColabFold methods to improve sampling. 10 models were generated with each method resulting in 50 models, which were then ranked with ModFOLDdock2. The top 5 ranked models were used as templates for custom template recycling, resulting in the final 5 refined models.

5.2.2 ModFOLDdock

ModFOLDdock is our server for the quality assessment of quaternary structures, utilizing three different variants that are optimised for the different facet of the quality estimation problem: ModFOLDdock, ModFOLDdockS, and ModFOLDdockR. The ModFOLDdock server employs a hybrid consensus method that integrates seven different quality scores to produce estimates of both local (interface) and global model accuracy. These quality scores include the DockQJury, QSscoreJury QSscoreOfficialJury, lDDTOfficialJury, voronota-js-voromqa, CDA, and ModFOLDIA. Various optimal combinations of these scores are combined to derive the variant of ModFOLDdock. ModFOLDdockR is used in MultiFOLD to select the top models. More detailed information about ModFOLDdock can be found in the server articles (Edmunds et al., 2023; McGuffin et al., 2023).

The standard ModFOLDdock scores were optimised for correlations of the predicted scores with the observed quality scores. However, ModFOLDdockR was optimized for ranking, i.e., to

just score the very best models as the highest, regardless of how well scores correlate with the actual mode quality. Using ModFOLDdockR for ranking, the total fold correctness was determined by the mean of the lDDTOfficialJury, QSscoreJury, and voronota-js-voromqa scores. In addition, the total interface correctness was determined by the mean of the QSscoreOfficialJury, DockQJury, and voronota-js-voromqa scores. Confidence scores for all interface residues in each model were calculated as the mean of the per-residue scores from ModFOLDIA and voronota-js-voromqa. The calculation of final interface accuracy is detailed in the server article (Edmunds et al., 2023). However, due to CPU and GPU limitations, it was not feasible to perform structural comparisons for large complexes (>1500 residues). Therefore, scores for all models were calculated using only a single model. Finally, to compare the generated structures, 30 reference structures were selected.

Additionally, the ModFOLDdockS variant used a quasi-single method, which uses the 30 reference models generated by MultiFOLD1 to score each model. For ModFOLDdockS, the mean of DockQJury and lDDTOfficialJury scores were used to determine the overall fold accuracy, while the mean of DockQJury and QSscoreOfficialJury scores were used to calculate the overall interface accuracy score. For individual residue confidence scores, the mean of the CDA scores, ModFOLDIA, and voronota-js-voromqa were used.

For MultiFOLD2, ModFOLDdock2 is used. Several changes were made to the ModFOLDdockS variant, including the addition of a NN for improved interface residue scoring, which integrates the input from VoroMQA, VoroIF, lDDTJury, CADJury, PatchQSJury, PatchDockQJury, ModFOLDIA, and CDA score. The NN includes 48 inputs comprising these eight input scores for the six closest residues to each interface. For ModFOLDdock2S, the mean of DockQ_waveJury and QS-bestJury scores were used to determine the global interface score, while the mean of TM-scoreJury and oligo-GDTJury scores were used to calculate the global fold score. A flow chart comparing the processes used in versions of MultiFOLD is shown in Figure 5.4.

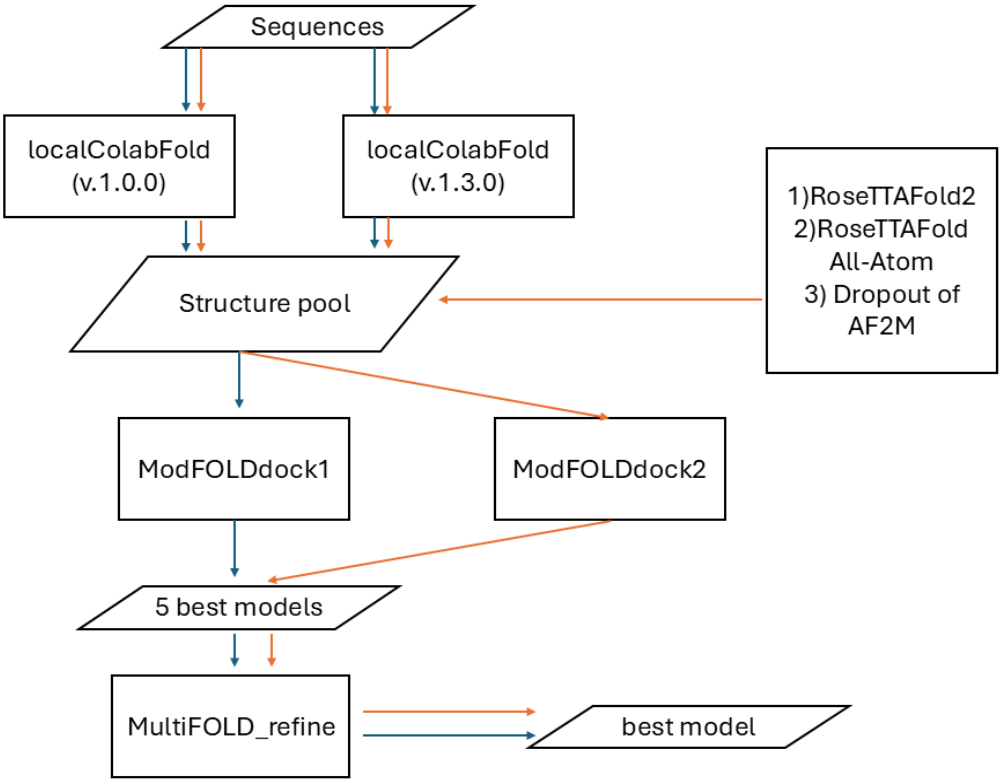


Figure 5.4 Flowchart of MultiFOLD1 and MultiFOLD2.
The flowchart illustrates the process from inputting at least two sequences to selecting the best model. The path for MultiFOLD1 is indicated by a blue line, while MultiFOLD2 is represented by a red line. Initially, a pool of modelled structures is generated using two versions of LocalColabFold. The five best models are then selected through ModFOLDdock1. Each model is subsequently refined with MultiFOLD_refine (using the custom template recycling methods), and the final 5 refined models are generated. Key differences for MultiFOLD2 include the incorporation of RoseTTAFold2, RoseTTAFold All-Atom, and the dropout approach from AF2M, which improves the sampling for the structure pool. Additionally, ModFOLDdock2 is used instead of ModFOLDdock1 for improved scoring.

5.2.3 Experimental analysis

Every Thursday, result files for weekly models generated by two version of MultiFOLD and the four other anonymous servers were downloaded after logging in to the CAMEO-BETA website. While the 3D structures in these files were in PDB format, the result files were JSON files in OpenStructure format. The Oligo-IDDT and QS-scores for each model from Server 1, Server 2, Server 4, Server 5, Server 76, and AF3 were extracted from the JSON files using in-house python code and then organized into common subsets to fairly compare the performance of each server against one another. The statistical analysis of the increase in quality scores for both MultiFOLD1 (Server1) and MultiFOLD2 (Server5) was performed using a one-tailed Wilcoxon signed-rank test, comparing to the quality scores for each of the other servers. The statistical method was explained in detail in the methods section of Chapter2.

5.3 Results and Discussion

5.3.1 Performance comparison of MultiFOLD1 against other servers using the CAMEO-BETA data

We initially compared the performance of four protein structure prediction methods, including Server 1 (MultiFOLD1), Server 2, Server 4, and Server 76, using the Oligo-IDDT and the QS-scores obtained from the JSON files for the multimer targets. Only these two scores were used to compare the modelling tools, as CAMEO-BETA only provides these two scores as part of their official evaluation. The dataset was obtained using results from CAMEO-BETA over the last 1.5 years for all targets (Table 5.1). Comparing all servers at once was not possible due to the limited number of common targets among them. Therefore, pairwise server comparisons were performed on common subsets of the data. Since CAMEO-BETA does not disclose information about the methodology used by Server 2 and Server 4, it is impossible to explain the differences in performance these anonymous group from an algorithmic standpoint. However, they serve as useful benchmarks as to the current state-of-the-art performance in complex modelling servers that are being developed and tested.

Table 5.1 Number of CAMEO-BETA common targets submitted by servers.
Table showing the total number of multimeric targets provided by CAMEO-BETA for the performance comparison of Server 1 (MultiFOLD1) with Servers 2, Server 4, and Server 76. The number of targets reflects those submitted to CAMEO-BETA by both servers. Collection of common targets between Server 4 and Server 1 commenced in December. The "After" refers to the number of targets collected following the resolution of the stoichiometry problem in MultiFOLD1 (see main text). The targets were collected from January 2023 to April 2024.

	The number of common targets	
	Before bug fix	After bug fix
MultiFOLD1-Server 2	77	50
MultiFOLD1-Server 4	--	104
MultiFOLD1-Server 76	226	146

The prediction of quaternary structure of complex protein generated by MultiFOLD1 is continually evaluated by the CAMEO-Beta project. No models were delivered for any target on server 4 until December 2023. Initially, the comparison was conducted among MultiFOLD1, Server 2, and Server 76. According to Table 5.2a, when MultiFOLD1 was compared to Server 2, the Oligo-IDDT and QS-score for Server 2 were higher. Comparing MultiFOLD1 to Server 76, the Oligo-IDDT for MultiFOLD1 was higher than that for Server 76, while the QS-score for MultiFOLD1 was lower. In Table 5.2b, when comparing Server 1 to other servers in terms of structure types (homomeric or heteromeric models), MultiFOLD1 outperformed Server 76 in the Oligo-IDDT for both heteromeric and homomer structures, as well as in the QS-score for heteromeric structures. However, Server 2 was found to predict models better than MultiFOLD1 according to both quality scores. Based on these results, we evaluated what went wrong and identified a problem with the stoichiometry prediction, where all homomers were mistakenly predicted as monomers by MultiFOLD1. This situation explains why the QS-score for MultiFOLD1 was lower compared to other servers, resulting in a cumulative QS-score of zero for homomeric models in MultiFOLD1. However, evaluating the comparison between two servers in terms of heteromeric structures was not possible, as there are no common heteromeric targets between MultiFOLD1 and Server 2.

Table 5.2 Performance comparison of multimeric structure predictions servers using CAMEO-BETA data before bug fixing.

(a) Table showing the performance results of three servers over a period of approximately one year, from 1 January 2023 to 5 December 2023. The CAMEO-BETA project provided direct results for the oligo-IDDT and QS-scores, and the table shows the cumulative scores for the paired servers. The highest cumulative scores are highlighted in bold. **(b)** Table comparing the cumulative Oligo-IDDT and QS-scores for multimeric structure prediction tools separated into the homomeric and heteromeric target types. No heteromeric targets for Server2 were submitted to CAMEO-BETA.

a)

	Cumulative Oligo-IDDT		Cumulative QS-score	
Paired servers	MultiFOLD1	Server 2	MultiFOLD1	Server 2
MultiFOLD1-Server2	26.54	**36.18**	0	**7.91**
Paired servers	MultiFOLD1	Server 76	MultiFOLD1	Server 76
MultiFOLD1-Server76	**92.50**	45.84	7.28	**17.50**

**Server4 Data were collected from December 2023

b)

Cumulative Oligo-IDDT Heteromer		Cumulative QS-score Heteromer		Cumulative Oligo-IDDT Homomer		Cumulative QS-score Homomer	
MultiFOLD1	Server 2	MultiFOLD1	Server 2	MultiFOLD1	Server 2	MultiFOLD1	Server 2
--	--	--	--	26.54	**36.18**	0	**7.91**
MultiFOLD1	Server 76	MultiFOLD1	Server76	MultiFOLD1	Server76	MultiFOLD1	Server76
41.78	13.60	**24.29**	6.65	**50.72**	32.23	0	**11.61**

So, until 5 December 2023, MultiFOLD1 incorrectly predicted multimer targets as monomers, resulting in lower cumulative oligo-IDDT score and QS-score. The accuracy of monomer prediction for MultiFOLD1 was evaluated due to the misprediction of homomers, since homomeric models consist of two or more copies of the same chain. The Oligo-IDDT score is calculated based on superposition-free alignments. Hence, even when the stoichiometry is incorrectly predicted as monomer, Oligo-IDDT is always measured. Therefore, comparing the QS-scores for multimeric models and the monomer-IDDTs for monomeric models produced by MultiFOLD1 against other servers will provide more information about its performance.

MultiFOLD1`s monomeric models outperformed those of Server 2 and Server 76 in Table 5.3 and 5.4. These results indicate that if MultiFOLD1 had correctly predicted these homomer targets, the current oligo-IDDT would have been higher than that of Server 2. Moreover, MultiFOLD1 could achieve better result, possibly obtaining higher cumulative QS-score.

Table 5.3 Performance comparison of monomer structure predictions for MultiFOLD1 using the CAMEO-BETA data.

The performance results of three servers are displayed over a period of about one year, from 1 January 2023 to 5 December 2023. CAMEO-BETA provides monomer-IDDT as an output for the models. The table displays the cumulative scores for all servers, along with the number of joint targets. The cumulative scores for the MultiFOLD1 are highlighted in bold. Results of all Wilcoxon signed-rank tests that were statistically significant ($p < 0.05$) are also highlighted in bold.

Servers	Number of targets	Cumulative Monomer-IDDT	Cumulative Monomer-IDDT
MultiFOLD1-Server 2	127	**94.53**	80.26
MultiFOLD1-Server 76	180	**151.17**	136.23
**Server 4	Data were collected from December 2023		

Table 5.4 Performance comparison of monomer structure predictions for MultiFOLD1 using the CAMEO-BETA data, after the stoichiometry issue was fixed.

The performance results of three servers are displayed over a period of about seven months, from 5 December 2023 to 8 June 2024. CAMEO-BETA provides monomer-IDDT as an output for the models. The table displays the cumulative scores for all servers, along with the number of joint targets. The cumulative scores for the MultiFOLD1 server are highlighted in bold. Additionally, the mean scores for each server are shown in the parentheses and the highest mean value highlighted in black. Unlike the comparison of MultiFOLD1 and Server 4, the results of all Wilcoxon sign-ranked test that were statistically significant ($p < 0.05$) are also highlighted in bold.

Servers	Number of targets	Cumulative Monomer-IDDT	Cumulative Monomer-IDDT	Wilcoxon Test
MultiFOLD1-Server 2	45	**32.71 (0.73)**	27.25 (0.61)	**p = 1.95e-07< 0.05**
MultiFOLD1-Server 4	36	**24.42 (0.68)**	24.06 (0.67)	p = 5e-1 > 0.05
MultiFOLD1-Server 76	100	**46.11 (0.46)**	41.94 (0.41)	**p = 3.34e-06< 0.05**

Data were then collected weekly from the CAMEO-BETA website starting from 5 December 2024, until the release of MultiFOLD2. MultiFOLD1 was determined to be the best performing server overall compared with the other three competing servers based on the cumulative oligo-IDDT and QS-scores in pairwise common subset server comparisons (Table 5.5a). According to the Wilcoxon signed-rank test in Table 5.5c, there were no statistical differences in the oligo-IDDT scores between MultiFOLD1 and Server 4 ($p=0.33>0.05$), which supports all the aforementioned results ($p<0.05$).

These results highlighted the stoichiometry issue with MultiFOLD1 had been resolved. According to Table 5.5b, both the oligo-IDDT and QS-score for homomeric structures generated by MultiFOLD1 outperformed the other three servers. Specifically, it is these data that indicate the stoichiometry bug was fixed (Table 5.5a). Regarding the heteromeric models generated by MultiFOLD1, MultiFOLD1 was not compared with Server 2 again due to the absence of any common heteromeric targets in the CAMEO_BETA dataset. The oligo-IDDT for heteromeric models generated by MultiFOLD1 was higher than that of Server 76, while it was lower than that of Server 4 (Table 5.5b). However, in all heteromeric models, the QS-score for MultiFOLD1 indicated poor performance. MultiFOLD1 utilises various versions of AF2M in order to obtain high-quality protein structures, with the best models selected by ModFOLDdock1. Therefore, it is crucial for MultiFOLD1 to achieve higher quality models before selecting the best ones. The performance of Server 4 was better than that of MultiFOLD1. In addition, given that Server 4 was released in December 2023, while the algorithms underlying MultiFOLD1 were based on progress in DNN before the CASP15 competition, Server 4 is likely to use later, more up-to-date methods. Server 76 is one of the main and older servers in the CAMEO-BETA project. The lower cumulative QS-score for MultiFOLD1 heteromeric models compared to that of Server 76 indicates a need for improvement in predicting higher quality interface heteromeric models. Thus, MultiFOLD1 may need to explore more extensive conformational sampling to generate higher quality models.

Table 5.5 Performance comparison of multimeric structure predictions servers using CAMEO-BETA data after bug fixing.

(a) Table showing the performance results of three servers over a period of approximately six months, from 5 December 2023 to 23 March 2024. The CAMEO-BETA project provided direct results for the oligo-IDDT and QS-scores, and the table shows the cumulative scores for the paired servers. The highest cumulative scores are highlighted in bold. Additionally, the mean scores for each server are shown in the parentheses and the highest mean value highlighted in black. **(b)** Table comparing the cumulative Oligo-IDDT and QS-scores for multimeric structure prediction tools separated into the homomeric and heteromeric target types. No heteromeric targets for Server 2 were submitted to CAMEO-BETA. **(c)** Table showing the statistical significance (p-value) of the cumulative Oligo-IDDT and the cumulative QS-scores for MultiFOLD1 compared to the three servers. The p-values with below 0.05 are highlighted in black.

a)

	Cumulative Oligo-IDDT		Cumulative QS-score	
Paired servers	MultiFOLD1	Server 2	MultiFOLD1	Server 2
MultiFOLD1-Server 2	**23.70 (0.33)**	16.95 (0.24)	**11.44 (0.16)**	8.02 (0.11)
Paired servers	MultiFOLD1	Server 4	MultiFOLD1	Server 4
MultiFOLD1-Server 4	**49.27 (0.47)**	43.31 (0.41)	**27.35 (0.26)**	23.36 (0.22)
Paired servers	MultiFOLD1	Server 76	MultiFOLD1	Server 76
MultiFOLD1-Server 76	**81.61 (0.56)**	61.63 (0.42)	**44.76 (0.30)**	33.79 (0.23)

b)

Cumulative Oligo-IDDT Heteromer		Cumulative QS-score Heteromer		Cumulative Oligo-IDDT Homomer		Cumulative QS-score Homomer	
MultiFOLD1	Server 2	MultiFOLD1	Server 2	MultiFOLD1	Server 2	MultiFOLD1	Server 2
--	--	--	--	**26.29**	19.28	**12.42**	8.99
MultiFOLD1	Server 4	MultiFOLD1	Server 4	MultiFOLD1	Server 4	MultiFOLD1	Server 4
26.12	**26.53**	13.26	**16.38**	**34.90**	28.86	**20.85**	13.93
MultiFOLD1	Server 76	MultiFOLD1	Server76	MultiFOLD1	Server76	MultiFOLD1	Server76
38.52	37.72	19.66	**27.28**	**54.73**	33.87	**30.84**	12.89

c)

	Wilcoxon signed-rank test (p-values)	
Paired servers	Cumulative Oligo-IDDT	Cumulative QS-score
MultiFOLD1-Server 2	**1.05e-05**	**8e-3**
MultiFOLD1-Server 4	3.3e-1	**3.6e-2**
MultiFOLD1-Server 76	**2.25e-10**	**1e-2**

According to the CAMEO-BETA results, overall MultiFOLD1 was the best performing server, even for monomeric models. However, despite the fix, the heteromeric models generated by MultiFOLD1 did not exhibit good interface quality, indicating a need to integrate more up-to-date methods. As of 23 March 2024, MultiFOLD2 was released as Server 5 and we began testing it in the CAMEO-BETA project, alongside MultiFOLD1.

5.3.2 Performance comparison of MultiFOLD2 against other servers using the CAMEO-BETA data

After building on the lessons learned from MultiFOLD1, the new version, MultiFOLD2, was developed (see Methods section), and tested in the CAMEO-BETA project (as Server 5). The main goal with the new method was to improve the protein structure modelling performance compared to the previous version of MultiFOLD. Consequently, the differences in the cumulative scores (oligo-IDDT and QS-score) between MultiFOLD1 and MultiFOLD2 were evaluated. Table 5.6 presents the number of CAMEO-BETA targets for the paired servers. Servers 2 and 4 were excluded from the evaluation of MultiFOLD2's performance due to no target submission for both servers. During the data collection period, AF3 was released, and the CAMEO-BETA community began publishing results based on the models obtained from the AF3 server. Therefore, the results of AF3 were included to the comparison of MultiFOLD2.

Table 5.6 Number of CAMEO-BETA common multimeric targets submitted by servers.
This table shows the total number of models for common multimeric targets provided by the CAMEO-BETA project for comparisons between Server 5 (MultiFOLD2), Server 1 (MultiFOLD1), Server 76, and AF3.

Servers	The number of common targets
MultiFOLD2 - MultiFOLD1	139
MultiFOLD2 - Server 76	116
MultiFOLD2 - AF3	55

Following the success of MultiFOLD1, MultiFOLD2 demonstrated superior performance in terms of Oligo-IDDT and QS-score (Table 5.7a). The primary method in MultiFOLD1 involved designing an extensive structure model pool and then selecting the best models from this pool using ModFOLDdock1. Given that ModFOLDdock1 ranked as one of the best performing quality estimation servers in the CASP15 competition (Edmunds et al., 2023), the more structures generated by the MultiFOLD, the better the quality of the best model obtained. AF2M's limitation in predicting a single state of the proteins is noted in the literature (Sala et al., 2023). The higher performance of MultiFOLD2 can be attributed to the integration of three components: RoseTTAFold2, RoseTTAFold All-Atom, and a dropout approach, which aimed to achieve better sampling of higher quality protein models. These approaches enabled MultiFOLD2 to obtain higher quality pool of conformations to select from. RoseTTAFold2 and RoseTTAFold All-Atom likely provided orthogonal model sampling, which could yield more variety in conformations. The dropout method also introduced minor pertubations to the existing intermediate conformational structures in the network. This extended sampling potentially achieved more optimal performance in predicting higher quality disorder regions or transient interactions (Johansson-Åkhe et al., 2019). However, searching the conformational landscape could still lead to traps in local minima while aiming for a broader range of conformational structures (Roney & Ovchinnikov, 2022). This issue was somewhat addressed in the last stage of the pipelines for part of both versions of MultiFOLD, known as MultiFOLD_refine, by employing recycling approaches to refine the final models.

AFsample (Wallner, 2023b) utilized a combination of the dropout method and extensive sampling, resulting in an increase in the average DockQ score in the CASP15 competition. It is noteworthy that MultiFOLD2 relies on AF2M`s inference dropout rate, which was 15-20%. Hence, the diversity generated by MultiFOLD2 was comparable to AF2M's success. However, increasing the dropout rate could further enhance diversity (Wallner, 2023b). In light of this, two versions of RoseTTAFold2 were used to generate more diversity in MultiFOLD2, along with the dropout method. Furthermore, it was demonstrated that RoseTTAFold All-Atom can predict larger protein complexes more effectively (Krishna et al., 2024) due to its use of structure based attention rather than triangle attention for updating pair features. For better conformational sampling, FAPE loss can be highly beneficial, as used in RoseTTAFold All-atom similar to AF2M. The FAPE loss compares the predicted atom positions with the actual positions. The predicted frames are aligned to the actual frame and the distances between the atom positions of each frame are calculated. These distances are penalised by an L1 loss. This method ensures that atoms are correct with respect to the local frame and that side chain interactions work properly (Jumper et al., 2021a). Specifically, FAPE loss allows models to follow the right path to the native structure in conformational space by providing smooth gradients (Baek et al., 2023). MD simulations are the most convenient way to sample proteins. Before the advent of AF2M, MD simulations were often used to refine modelled protein structures, and the resulting trajectories could be used to obtain different structures in conformational space. However,

the time-dependent nature of MD trajectories makes them CPU/GPU intensive and thus time-consuming to obtain for complex proteins with current computational techniques (Kožić & Bertoša, 2024).

In addition to the success of MultiFOLD2 compared to the first version of MultiFOLD, MultiFOLD2 outperformed Server 76 and even AF3 (Table 5.7b). The success of MultiFOLD2 against other servers may also be attributed to our built-in quality estimation tool (Edmunds et al., 2024), ModFOLDdock (Edmunds et al., 2023). Since most tools rely on AF2M`s main method, each tool implicitly benefits from the strengths of DNN methods. In addition, groups tend to obtain better conformational ensembles by using MSA subsampling (Wayment-Steele et al., 2024). Hence, almost all methods manage to predict targets with a certain quality for downstream analysis. The combination of tools used can provide a variety of protein structures, with the main challenge being to detect the best one. ModFOLDdock1 address this issue by using a combination of different scores generated by various tools, whereas AF2M rely solely on pIDDT and pTM-score for ranking after the generation of structure models, which is inferior to ModFOLDdock for model ranking (Edmunds et al., 2024). Along with multimeric model predictions, MultiFOLD2 also performed better than Server1 and Server76 for monomeric models; however, it did not show better performance compared to AF3 on monomers (Table 5.8). The Wilcoxon signed-rank test showed that there was a significant difference in the cumulative scores between MultiFOLD2 and MultiFOLD1, Server 76, and AF3 ($p < 0.05$) for both multimeric and monomeric models, except for the monomeric models generated by MultiFOLD2 and AF3 (Table 5.7c).

Table 5.7 Comparison of the cumulative scores for MultiFOLD2, MultiFOLD1, and other servers.

(a) The table representing the comparison of cumulative oligo-IDDT and QS-score for the models generated by MultiFOLD2 and MultiFOLD1. **(b)** The table showing the comparison of the same cumulative scores for the models generated by MultiFOLD2 against those of Server76 and AF3. The cumulative scores for the best performing servers are highlighted in bold. Additionally, the mean scores for each server are shown in the parentheses and the highest mean value highlighted in black in (a) and (b) parts. **(c)** Table showing the statistical significance (p-value) of the cumulative Oligo-IDDT and the cumulative QS-scores for MultiFOLD2 compared to the three servers. The p-values with below 0.05 are highlighted in black.

a)

	Cumulative oligo-IDDT		Cumulative QS-score	
Paired servers	MultiFOLD2	MultiFOLD1	MultiFOLD2	MultiFOLD1
MultiFOLD2 – MultiFOLD1	**81.83 (0.59)**	79.24 (0.57)	**52.56 (0.37)**	50.22 (0.36)

b)

	Cumulative Oligo-IDDT		Cumulative QS-score	
Paired servers	MultiFOLD2	Server 76	MultiFOLD2	Server 76
MultiFOLD2 - Server76	**69.52 (0.60)**	53.50 (0.46)	**45.66 (0.40)**	34.72 (0.30)
Paired servers	MultiFOLD2	AF3	MultiFOLD2	AF3
MultiFOLD2 - AF3	**34.33 (0.62)**	27.03 (0.50)	**24.28 (0.44)**	13.98 (.025)

c)

	Wilcoxon signed-rank test (p-values)	
Paired servers	Cumulative Oligo-IDDT	Cumulative QS-score
MultiFOLD2-MultiFOLD1	**1.41e-06**	**3.566e-2**
MultiFOLD2-Server 76	**1.39e-09**	**4.729e-3**
MultiFOLD2-Server AF3	**4.984e-2**	**3.242e-3**

Table 5.8 Performance comparison of monomer structure predictions for MultiFOLD2 using the CAMEO-BETA data.

The performance results of four servers are displayed over the period from 30 March 2024 to 08 June 2024. CAMEO-BETA provides monomer-IDDT as an output for the models. The table displays the cumulative scores for all servers, along with the number of joint targets. The highest cumulative scores for each server are indicated in bold. Additionally, the mean scores for each server are shown in the parentheses and the highest mean value highlighted in black. Unlike the comparison of MultiFOLD2 and AF3, the results of all Wilcoxon signed-rank tests that were statistically significant ($p < 0.05$) are also highlighted in bold.

Servers	Number of targets	Cumulative monomer-IDDT	Cumulative monomer-IDDT	Wilcoxon Test
MultiFOLD2-MultiFOLD1	41	**35.45 (0.86)**	35.24 (0.85)	**p = 9e-3<0.05**
MultiFOLD2-Server 76	46	**36.29 (0.79)**	35.64 (0.77)	**p = 3e-4<0.05**
MultiFOLD2-AF3	22	18.59 (0.85)	**18.68** (0.85)	p = 9e-1>0.05

MultiFOLD2 outperformed MultiFOLD1 and other servers in terms of both homo and hetero types of complexes. Only the cumulative QS-score for the heteromeric models generated by Server 76 was higher than that of the MultiFOLD2 heteromeric models (Table 5.9). Server 76 (disclosed by the CAMEO authors as being Swiss-Model) (Biasini et al., 2014) is based on homology modelling, which is successful for modelling the conserved interface of complex proteins; hence, homomeric models may be better predicted than heteromeric models. However, heteromeric predictions for Server 76 outperformed MultiFOLD2 in terms of QS-score, which could be due to incorrect stoichiometric information, despite the fact that the Oligo-IDDT score for MultiFOLD2 was higher. When analysing the heteromeric models generated by both servers (MultiFOLD2 and Server 76), MultiFOLD2 tended not to predict well for the heteromeric models with more than four chains. The QS-score for the only model (target:8JLC) generated by MultiFOLD2 was zero despite it being a di-heteromeric model, while the QS-score for the same target generated by Server 76 was 0.46. Interestingly, the Oligo-IDDT score for 8JLC generated by MultiFOLD2 was twice as high as that of Server 76. This suggests that MultiFOLD2 can predict the monomeric structures of dimers well; however, it does not predict the interface accuracy for dimeric interface as effectively. An observation for AF3 was that it predicted all homomeric models as monomers, similar to MultiFOLD1, resulting in an QS-score of zero.

Table 5.9 Comparison of MultiFOLD2, MultiFOLD1, and other servers in terms of the type of multimeric structures.

The table representing the comparison of cumulative oligo-IDDT and QS-scores for the homomeric and heteromeric models generated by MultiFOLD2 and those generated by MultiFOLD1, Server76, and AF3. The cumulative scores for the best servers are indicated in bold.

Cumulative Oligo-lDDT Heteromer		Cumulative QS-score Heteromer		Cumulative Oligo-lDDT Homomer		Cumulative QS-score Homomer	
MultiFOLD2	MultiFOLD1	MultiFOLD2	MultiFOLD1	MultiFOLD2	MultiFOLD1	MultiFOLD2	MultiFOLD1
46.43	44.51	**28.10**	27.28	**35.40**	34.73	**24.45**	22.93
MultiFOLD2	Server 76	MultiFOLD2	Server 76	MultiFOLD2	Server 76	MultiFOLD2	Server 76
34.12	33.29	21.21	**26.67**	**35.40**	20.21	**24.45**	8.053
MultiFOLD2	AF3	MultiFOLD2	AF3	MultiFOLD2	AF3	MultiFOLD2	AF3
21.99	21.15	**14.28**	13.98	**12.33**	5.87	**10.00**	0

MultiFOLD1 tended to predict the models that either underestimated or overestimated when compared to reference structures in terms of the number of chains. This was an issue of stoichiometry, which was addressed for MultiFOLD2. However, the stoichiometry of 57%, 50%, and 59% of the models in the common subsets were predicted correctly when MultiFOLD2 was compared to MultiFOLD1, Server 76, and AF3, respectively. Specifically, as the number of chains increased, the likelihood of an issue with correct stoichiometry prediction occurring also increased. Compared to MultiFOLD1, MultiFOLD2 predicted the wrong stoichiometry for 69 models out of 139 targets (Table 5.10). MultiFOLD2 was not designed in order to solve the stoichiometry issue. Similar to MultiFOLD1, MultiFOLD2 relies on template-based stoichiometry prediction. Instead of predicting a greater number of chains compared to the reference structures, MultiFOLD2 tended to predict fewer chains compared to the reference structures, especially when compared to the other three methods. In addition, when comparing common targets between MultiFOLD1 and MultiFOLD2, MultiFOLD2 predicted 11 multimer targets as monomers, while MultiFOLD1 predicted 12 such targets, with seven of these being the same for both. It is clear that stoichiometry prediction is an area for further improvement for all methods.

Table 5.10 Evaluation of stoichiometry for joint models generated by servers.
The table representing incorrect estimations of chain numbers for models generated by the servers. It also includes the target numbers for joint targets submitted by the servers. Comparisons are made between Server5 (MultiFOLD2) to Server 1 (MultiFOLD1), Server 76 and AF3. The servers` predictions were assessed for either underestimation or overestimation. In addition, multimers predicted as monomeric structures are included in the underestimation column.

Servers	Target Number	Underestimation	Overestimation
MultiFOLD2 - MultiFOLD1	139	38 (11 monomer)	11
MultiFOLD2 - Server 76	116	40 (11 monomer)	8
MultiFOLD2 - AF3	55	25 (5 monomer)	4

To create a diverse structure pool, various approaches have been employed in MultiFOLD, including different versions of AF2M and varying recycling parameters in the MultiFOLD_refine section. In addition, to access a more effective conformational space, the efficacy of generative models as a replacement for AF2M has been investigated. Research has demonstrated that these models can produce functional structures that are not present in the Protein Data Bank (PDB) (Tian et al., 2021). Since MultiFOLD2's structure search relies on AF2M and other similar approaches, there is a possibility of getting trapped in local minima, potentially limiting the quality of the obtained models. However, generative models can overcome such issues as they are not constrained by kinetic barriers (Janson et al., 2023). This issue might be addressed by the diffusion model used in the latest stage of AF3, however the code for AF3 is not available at the time of writing. Despite these short comings, MultiFOLD2 has outperformed AF3 in predicting complex structures without employing generative models. This can be attributed to MultiFOLD2’s use of built-in stoichiometry prediction, independent scoring, along with improved sampling, which allows for a diverse conformational search strategy.

5.4 Conclusions

Despite the breakthroughs in protein structure modelling achieved using DNNs, such as AF2M and RoseTTAFold2, significant challenges remain in accurately predicting protein complexes. These challenges include modelling stoichiometrically correct multi-chain structures and generating comprehensive conformational ensembles. Addressing these fundamental issues has driven the development of new methods. Consequently, MultiFOLD1 was developed in order to predict complex protein structures with accuracy beyond that of AF2M (McGuffin et al., 2023). While the algorithm of MultiFOLD1 relies on AF2M for generating multiple structures from the given protein sequence, it improves upon them through better selection using an independent MQA known as ModFOLDdock, and refinement through custom template recycling. The further success of the latest version of MultiFOLD, MultiFOLD2, can be attributed to several additional factors: accessing a broader conformational space by using two version of RoseTTAFold and a dropout approach, utilizing different combinations of quality assessment scores, and employing recycling with the improved version of ColabFold. The positive impact of using varied recycling values on complex structures was demonstrated in the previous chapters.

Our primary objective was to demonstrate the superiority of MultiFOLD2 over MultiFOLD1 and to benchmark the performance of both versions against other servers using the CAMEO-BETA data. Results indicate that MultiFOLD2 has emerged as the top-performing server according to the CAMEO-BETA project, surpassing even the latest version of the AF methods, known as AF3. Analysis the cumulative performance of the servers shows improved modelling of homo and hetero multimeric structures. The only exception was that AF3 outperformed MultiFOLD2 in monomeric structure predictions. This discrepancy may be attributed to our ModFOLDdock method, which was optimised with scores tailored for multimer protein models, which potentially affecting the accurate selection of the best homomer model.

The most critical problem for MultiFOLD2 is the stoichiometry issue, which is a common challenge for all similar methods. MultiFOLD2 often predicted multi-chain reference structures with fewer chains, sometimes even as monomers, compared to the reference structures. When MultiFOLD1 was first introduced, it predicted all homomer structures as monomers, resulting in poor QS-scores. However, after a bug in the template-based stoichiometry assignment was fixed, this issue was resolved in December 2023, leading to improved scores. While MultiFOLD1 exhibits lower performance for heteromeric structures, MultiFOLD2 has shown improved success in this area compared to other servers, with the exception of the Oligo-IDDT scores from Server 76. Efforts to improve complex protein structure predictions are inherently constrained by computational limitations. Currently, MultiFOLD can only process protein sequences up to the residue capacity of AF2M. Therefore, manual prediction is required for protein sequences exceeding ~6000 residues, which is the current size limit based on the latest 48GB GPU cards.

The development of MultiFOLD can be advanced through three distinct avenues. Firstly, if Google DeepMind release their code, then AF3 method will be integrated with the pipeline. Secondly, the integration of the next version of ModFOLDdock, trained using different DNNs and new quality scores, may be expected to enhance the performance beyond that of MultiFOLD2. With the high-quality protein structures obtained from AF2M and RoseTTAFold2, research has increasingly shifted focus towards studying interactions between proteins and non-protein structures. Consequently, accurately identifying binding sites on proteins has become increasingly important. Therefore, integrating scores that identify and score binding site regions into ModFOLDdock could further enhance its capabilities. Additionally, incorporating generative models to explore previously uncharted regions of conformational space would enrich the structural model pool. At the time of writing, MultiFOLD2 is currently being tested in the CASP16 competition, and its performance will be announced at the CASP16 conference in December.

Chapter 6: Synthesis, Conclusions and Next Directions

6.1 Synopsis of study

Structural bioinformatics has been revolutionised following the release of AF2, which predicts both monomeric and multimeric structures with up to two chains at close to experimental accuracy. Hence, following the CASP14 competition and the release of the AF2 code (https://github.com/google-deepmind/alphafold), most existing tools were upgraded to integrate various AF2 versions into their own methods, and new specific functional tools were released using the different parts of the AF2 code. Over the last few years, several different AF2 versions have been released, and each new version has outperformed the previous versions incrementally. Although it is known that the most recent AF2 versions predict monomeric structures very well, it remains uncertain how the code can be used effectively to obtain high quality models for larger multimeric structures.

With the growing use of DNNs in protein structure modelling, many new methods have managed to obtain high quality models for both monomeric and multimeric structures, and in turn, the use of traditional refinement methods has decreased. Alternatively, predictors have sought to include other techniques for the improvement of intermediate models rather than refining the final modelled structures to fix the errors using traditional methods such as MD simulation (Heo et al., 2019). This change in practice occurred along with the removal of the CASP15 refinement category. However, research has shown that the models generated by even the very latest AF2 versions are still imperfect and that procedures need to be developed to improve them (Heo & Feig, 2020). Rather than using the traditional refinement tools for the improvement of multimeric models, the various AF2 input parameters can be effectively exploited to obtain higher quality multimeric structures. This study first focused on the optimisation of parameters for the AF2M version, particularly the use of further recycling, custom MSAs and custom templates, to improve models of the quaternary structures of proteins. Subsequently, these optimisations, described in Chapters 2, 3 and 4, were integrated with our MultiFOLD (McGuffin et al., 2023) server versions, which have shown leading performance in gold-standard blind prediction experiments, such as the CASP competition and the CAMEO project.

6.1.1 The impact of recycling on the modelling of quaternary structures of proteins: An evaluation of two AlphaFold2 versions (AF2_Advanced and AF2-Multimer)

Inspired by the success of AF2 in predicting single chains (Jumper et al., 2021a), research was conducted to apply AF2 methods to protein complexes. AF2_Advanced and AF2M were developed as forks of the original AF2 version to model the quaternary structures of proteins, but the issue of improving the accuracy of the models remained. One of the most intriguing parts of the AF2 code

was the concept of the recycling process, which was used to improve models by passing them through the DNN multiple times. Three recycling steps were found to be sufficient for accurately predicting single chains (Jumper et al., 2021a); however, the optimal number of recycling steps for modelling protein complexes is not yet fully understood.

One aim of Chapter 2 was to identify the effect of further recycling on multimeric structures and to find, if available, the most suitable number of recycles for both main AF2 versions in order to obtain multimeric models with the highest model quality. In terms of quality score (TM-scores (Zhang & Skolnick, 2004), IDDT scores (Mariani et al., 2013) and QS-scores (Bertoni et al., 2017)), an increase in the number of recycles has been observed to have a positive impact on the quality of quaternary structure models of proteins. However, increases in the number of cycles do not always result in higher-quality models of complexes. Additionally, there is no significant difference between AF2_Advanced and the first versions of AF2M regarding model improvement with increased recycling. The analysis indicated that 12 recycles can be more effective for both AF2M and AF2_Advanced than the default 3 recycles. This result shows that higher quality complex structures can be obtained through the use of further recycling.

(Following this study, the “auto” recycling parameter was introduced by the ColabFold developers, which now automatically determines the optimal number of according to the score produced for a given protein structure; hence, a user does not necessarily need to manually select the number of recycling cycles when running AF2M).

6.1.2 The impact of the custom template recycling for the improvement of quaternary structures of proteins

The AF2M method has demonstrated the benefits of the use of end-to-end DNNs for protein structure prediction (Evans et al., 2022). Over time, changes have been made to AF2's methods to improve its ability to predict structures. New versions of AF2M, v1 and v2, have been developed that offer two options, mainly in terms of different trained weights. Another important change was the addition of the custom template option, which can greatly enhance the initial quality of the structure predicted by AF2M. This is because template structures can be created using other modelling methods, particularly those based on physical laws, leading to better results. In Chapter 3, the best-modelled structures from the CASP14 and CASP15 targets were used as templates, as well as the AF2-NBIS-Multimer and MultiFOLD models. Separate weights were employed to ensure that these targets were not part of the AF2M training set. The result shows that custom templates with further recycling can be employed to obtain higher quality multimeric structures. However, for the CASP14 models, the improved model quality correlated with an increase in further recycling with the custom templates,

while the CASP15 models show that the way AF2M is trained affects the quality of multimeric models (Wallner, 2023a). This discrepancy may be attributed to the use of v2 for the CASP15 models; AF2M_v2 is more recent than AF2M_v1 and was adjusted to improve MolProbity scores (Chen et al., 2010(a)), which were designed to reduce clashes. This aligns with traditional refinement theory, suggesting that higher-quality initial structures tend to degrade rather than improve during conformational sampling. Therefore, there may be connection between the quality of the template used and which version (v1 or v2) is used. Additionally, the CASP15 templates were higher quality starting models (https://predictioncenter.org/casp15/results.cgi?tr_type=multimer) than the CASP14 templates (https://predictioncenter.org/casp14/results.cgi?view=targets&tr_type=multimer) since tools in the CASP15 competition employed AF2M in their methods. Along with custom templates, the use of MSAs is advantageous for models generated by AF2M using further recycling compared to using a SS approach (Adiyaman et al., 2023).

6.1.3 The impact of varying custom input options on models generated by AF2M

The most important factors in the model development of AF2M are the input features designed within AF2, namely, custom templates and custom MSAs. These two inputs have been shown to be used in different ways, particularly in the CASP15 competition, demonstrating the advanced performance of AF2M (Kryshtafovych et al., 2023a). Therefore, in Chapter 4, the impact of changing the way inputs are given on AF2M models was examined. When the custom templates were provided as “single-chain” custom templates instead of the default setting being used, it was demonstrated that AF2M can be influenced by external custom structural input. When AF2M uses template structures as guides for conformational sampling, it individually processes each chain for multichain protein structure templates (Evans et al., 2022), which affects multimeric model quality. Therefore, providing complex templates to AF2M as “single-chain” custom templates was examined, as such provision may allow the simultaneous evaluation of the entire structure, preserving the interfaces between chains. While the results showed improvements in TM-scores and IDDT scores, no improvement was observed in the interface scores (QS-scores and DockQ-waves).

Co-evolution information derived from MSAs has significantly contributed to AI-based tools for structure prediction. AF2 is among the most impactful methods developed for extracting valuable insights from MSAs by leveraging coevolutionary data. Therefore, a high quality MSA must be provided as input for the power of AF2 to be of benefit. The most important factor here is that residues corresponding to ordered regions give strong coevolution signals. However, residues corresponding to disordered regions, or natively unstructured regions, lack some coevolution information (Iserte et al., 2020). Parts that cause such a weak signal are also likely to cause incorrect alignment within the MSA, which reduces the accuracy of predictions during the inference phase of AI-based tools. MSAs

that lack disordered structures also mean that they have increased complexity. However, Petti et. al (2022) have shown that high complexity MSAs are more effective inputs for AF2M, and thus AF2M can be forced to make better use of co-evolution information. In the case of the MSAs filtered to screen out disordered residues, it was observed that more than half of the multimeric models were improved across the four quality scores. The results show that when suitable disorder filtering is applied to input an MSA, AF2M can produce higher quality protein models by strengthening the co-evolutionary signal within the ordered regions.

6.1.4 Performance comparison of MultiFOLD1 and MultiFOLD2 using data from the CAMEO-BETA project

In the CASP15 competition, AF2M-based tools modelled the quaternary structures of proteins very well. Although AF2M provides higher quality protein models, its drawback is that it predicts only a single conformation (Sala et al., 2023), highlighting the need for a new multichain protein modelling tools that improve conformational sampling. This issue can be solved by designing an extended structure pool with models obtained from other tools or methods. To this end, we developed an AF2M-based server called MultiFOLD (McGuffin et al., 2023), employing custom templates with further recycling to model complexes. The main MultiFOLD server algorithm performs more effective conformational sampling to generate diverse structures (Brysbaert et al., 2024; Wallner, 2023b), selecting the best ones from a structure pool using our quality estimation tool known as MultiFOLDdock (Edmunds et al., 2023). ModFOLDdock, designed by our group, ranked second in the CASP15 competition. Consequently, selecting the best models from the generated multichain protein structure pool using ModFOLDdock enabled MultiFOLD to achieve better results than other servers. Both MultiFOLD1 and the subsequently developed MultiFOLD2 outperformed other methods in modelling monomeric and multimeric models. While the latest version of AlphaFold (AF), AlphaFold3 (AF3), achieves more effective results for monomeric models, the two versions of MultiFOLD have shown superior performance for multimeric models.

Initially, in MultiFOLD1, the structure pool was created by utilizing two different local versions of ColabFold (Mirdita et al., 2022), which is a local and website variant of AF2M, with varying parameters. This approach is based on the hypothesis that different versions of AF2M with different parameters can produce more orthogonal sampling of models through recycling, as explored in Chapters 2 and 3. In MultiFOLD2, the conformational sampling space was further enhanced by incorporating RoseTTAFold2 (Baek et al., 2023) and RoseTTAFold2 All-Atom (Krishna et al., 2024) and dropout methods of AF2M (Jumper et al., 2021a), resulting in a more extensive structure pool.

6.2 Conclusions

AF2M was examined in terms of the combination of modelling and refinement, and it's appropriate and effective use were determined. AF2 has evolved, with new versions also appearing since the first day it was released. Here, the most effective way to use AF2M in terms of improvement is by increasing conformational sampling, which produces a large number of conformational structures in the same way that traditional refinement methods do. It has been observed that this generally involves using either recycling or dropout methods. Additionally, more recycling gave better results in both main versions of AF2. Assuming that the external templates provided to AF2M are beneficial input, the quality of the template structure - that is, the starting structures - obtained by externally templating AF2M_v1 and AF2M_v2 varies, depending on which AF2M version is used and whether the MSA information is used or not. If the template structure is of good quality, AF2M_v1 tends to give better structures, while the inclusion of MSA information along with a custom template and further recycling has been observed to increase the quality of both the global assembly and the modelled interface residues in the structures.

Since it is important to provide an effective MSA to AF2M which benefits from clear co-evolutionary information, a custom MSA input has been provided. In this way, residues that lack co-evolutionary information on the MSA produced by AF2M are removed and MSAs are replaced with those that have strong coevolutionary information. When an MSA input is given, AF2M ordered structure signal will be improved, and higher quality structures will be likely to emerge. However, there is not a single set of parameters for MSA filtering that can be applied generally to each protein structure. Therefore, performing generic MSA filtering for each protein structure does not always provide improvements in models, and it is not possible to determine what kind of disordered structures filtering a given the protein target will benefit from. We wanted a consistently effective improvement process in a protein modelling tool, so we designed the AF2M-based MultiFOLD server by focusing on both the sampling and refinement phases. To ensure effective use of the structural information of the AF2 versions, different versions of AF2M, as well as RoseTTAFold2 and dropout methods of AF2M, were used to produce different conformations in modelling, and in the refinement process with initial models as custom template inputs, 12 recycles of AF2M was used to ensure that higher quality structures were prepared. The effectiveness of two MultiFOLD servers was benchmarked with other servers in the CAMEO project based on both Oligo-IDDT and QS-scores. MultiFOLD2 was found to outperform MultiFOLD1 and furthermore, it was determined the best among the other servers, even outperforming AF3, according to the modelling of the quaternary structures of proteins.

6.3 Future directions

As the use of AI-based methods on multichain protein structure modelling has increased and higher quality models of complex proteins have been predicted, the focus on modelling has now given way to specific areas, the most important of which are protein design and protein ligand prediction. Sampling low-energy conformational structures of proteins for protein design and modelling binding sites in protein structures for ligands are likely to be the main focuses for “improvements” from now on. Hence, future research will focus on extending the existing structure pool used for MultiFOLD2 by examining the following major topics.

A) I will aim to develop a standalone mechanism that will integrate AF methods with the models in trajectories obtained by using open source MD simulations, such as OpenMM (Eastman et al., 2024). Since force fields are insufficient for multimeric models, I will follow developments in this field.

B) MultiFOLD2 is currently being tested in the CASP16 competition. After the CASP16 conference, new methods introduced by the community will be evaluated and integrated into MultiFOLD version 3 to produce higher quality structures. Comparisons with AF3 (Abramson et al., 2024) indicate that, while MultiFOLD2 performs well, there is still a need for better tools to create even higher-quality models of complex structures.

   At the time of this writing, the AF3 code has not been released, so it is impossible to make use of the strength of the diffusion model in AF3 (Abramson et al., 2024). Therefore, integrating deep generative models into future versions of MultiFOLD is a potential further step. By doing so, I may achieve more effective conformational sampling for multimeric structures.

C) It was shown in Chapter 4 that modifications to the inputs given to AF2M can affect the quality of the model predicted by AF2M. The main factor is the transformer method, which decides which information in the input structures is more beneficial. Transformers determine this by using neurons weighted according to certain information and allowing AF2M to benefit effectively from coevolution information. Furthermore, even if disorder residue information is not available in MSAs, the improvement in model quality changes may be attributed to changes in the weights created in the neurons of the DNNs within the transformer structure. Consequently, investigating this effect can provide faster modelling and increased efficiency in terms of time by examining whether other information can be reduced.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E.,…Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493-500. https://doi.org/10.1038/s41586-024-07487-w

Adiyaman, R., Edmunds, N. S., Genc, A. G., Alharbi, S. M. A., & McGuffin, L. J. (2023). Improvement of protein tertiary and quaternary structure predictions using the ReFOLD refinement method and the AlphaFold2 recycling process. *Bioinformatics Advances*, *3*(1). https://doi.org/10.1093/bioadv/vbad078

Adiyaman, R., & McGuffin, L. J. (2019). Methods for the Refinement of Protein Structure 3D Models. *Int J Mol Sci*, *20*(9). https://doi.org/10.3390/ijms20092301

Adiyaman, R., & McGuffin, L. J. (2021). ReFOLD3: refinement of 3D protein models with gradual restraints based on predicted local quality and residue contacts. *Nucleic Acids Res*, *49*(W1), W589-w596. https://doi.org/10.1093/nar/gkab300

Ahdritz, G., Bouatta, N., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M. M., Zhang, S., Ojewole, A., Guney, M. E., Biderman, S., Watkins, A. M., Ra, S.,…AlQuraishi, M. (2022). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*, 2022.2011.2020.517210. https://doi.org/10.1101/2022.11.20.517210

Akoglu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med*, *18*(3), 91-93. https://doi.org/10.1016/j.tjem.2018.08.001

Alexander, H., Hu, S., Krinos, A., Pachiadaki, M., Tully, B., Neely, C., & Reiter, T. (2023). Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio*, *14*, e0167623. https://doi.org/10.1128/mbio.01676-23

Alexov, E., Mehler, E. L., Baker, N., Baptista, A. M., Huang, Y., Milletti, F., Nielsen, J. E., Farrell, D., Carstensen, T., Olsson, M. H., Shen, J. K., Warwicker, J., Williams, S., & Word, J. M. (2011). Progress in the prediction of pKa values in proteins. *Proteins*, *79*(12), 3260-3275. https://doi.org/10.1002/prot.23189

Aloy, P., Ceulemans, H., Stark, A., & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, *332*(5), 989-998. https://doi.org/10.1016/j.jmb.2003.07.006

AlQuraishi, M. (2019). End-to-End Differentiable Learning of Protein Structure. *Cell Syst*, *8*(4), 292-301.e293. https://doi.org/10.1016/j.cels.2019.03.006

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1-8. https://doi.org/https://doi.org/10.1016/j.cbpa.2021.04.005

André, I., Strauss, C. E. M., Kaplan, D. B., Bradley, P., & Baker, D. (2008). Emergence of symmetry in homooligomeric biological assemblies. *Proceedings of the National Academy of Sciences*, *105*(42), 16148-16152. https://doi.org/doi:10.1073/pnas.0807576105

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223-230. https://doi.org/10.1126/science.181.4096.223

Arndt, V., Dick, N., Tawo, R., Dreiseidler, M., Wenzel, D., Hesse, M., Fürst, D. O., Saftig, P., Saint, R., Fleischmann, B. K., Hoch, M., & Höhfeld, J. (2010). Chaperone-assisted selective autophagy is essential for muscle maintenance. *Curr Biol*, *20*(2), 143-148. https://doi.org/10.1016/j.cub.2009.11.022

Ashkenazy, H., Sela, I., Levy Karin, E., Landan, G., & Pupko, T. (2019). Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction. *Syst Biol*, *68*(1), 117-130. https://doi.org/10.1093/sysbio/syy036

Baek, M. (2021). *Twitter post: adding a big enough number for residue_index feature is enough to model hetero-complex using AlphaFold (green&cyan: crystal structure/magenta: predicted model w/residue_index modification)*Twitter. https://twitter.com/minkbaek/status/1417538291709071362

Baek, M., Anishchenko, I., Humphreys, I. R., Cong, Q., Baker, D., & DiMaio, F. (2023). Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv*, 2023.2005.2024.542179. https://doi.org/10.1101/2023.05.24.542179

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C.,…Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871-876. https://doi.org/doi:10.1126/science.abj8754

Baek, M., Park, T., Heo, L., Park, C., & Seok, C. (2017). GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Res*, *45*(W1), W320-w324. https://doi.org/10.1093/nar/gkx246

Bahat, Y., Alter, J., & Dessau, M. (2020). Crystal structure of tomato spotted wilt virus G_N reveals a dimer complex formation and evolutionary link to animal-infecting viruses. *Proceedings of the National Academy of Sciences*, *117*(42), 26237-26244. https://doi.org/doi:10.1073/pnas.2004657117

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Basu, S., & Wallner, B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE*, *11*(8), e0161879. https://doi.org/10.1371/journal.pone.0161879

Ben-Hur, A., & Brutlag, D. (2003). Remote homology detection: a motif based approach. *Bioinformatics*, *19 Suppl 1*, i26-33. https://doi.org/10.1093/bioinformatics/btg1002

Berchanski, A., Shapira, B., & Eisenstein, M. (2004). Hydrophobic complementarity in protein-protein docking. *Proteins*, *56*(1), 130-142. https://doi.org/10.1002/prot.20145

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports*, *7*(1), 10480. https://doi.org/10.1038/s41598-017-09654-8

Bhattacharya, D. (2019). refineD: improved protein structure refinement using machine learning based restrained relaxation. *Bioinformatics*, *35*(18), 3320-3328. https://doi.org/10.1093/bioinformatics/btz101

Bhattacharya, D., & Cheng, J. (2013). *Protein Structure Refinement by Iterative Fragment Exchange* Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, Wshington DC, USA. https://doi.org/10.1145/2506583.2506601

Bhattacharya, D., Nowotny, J., Cao, R., & Cheng, J. (2016). 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Res*, *44*(W1), W406-409. https://doi.org/10.1093/nar/gkw336

Bhattacharya, N., Thomas, N., Rao, R., Dauparas, J., Koo, P. K., Baker, D., Song, Y. S., & Ovchinnikov, S. (2022). Interpreting Potts and Transformer Protein Models Through the Lens of Simplified Attention. *Pac Symp Biocomput*, *27*, 34-45.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., & Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*, *42*(Web Server issue), W252-258. https://doi.org/10.1093/nar/gku340

Bonvin, A. M. J. J., Karaca, E., Kastritis, P. L., & Rodrigues, J. P. G. L. M. (2018). Defining distance restraints in HADDOCK. *Nature Protocols*, *13*(7), 1503-1503. https://doi.org/10.1038/s41596-018-0017-6

Borkakoti, N., & Thornton, J. M. (2023). AlphaFold2 protein structure prediction: Implications for drug discovery. *Current Opinion in Structural Biology*, *78*, 102526. https://doi.org/https://doi.org/10.1016/j.sbi.2022.102526

Bossi, A., & Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Mol Syst Biol*, *5*, 260. https://doi.org/10.1038/msb.2009.17

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods Mol Biol*, *406*, 89-112. https://doi.org/10.1007/978-1-59745-535-0_4

Branden, C., & Tooze, J. (1991). *Introduction to Protein Structure*. Garland Pub. https://books.google.co.uk/books?id=MERrAAAAMAAJ

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., & Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*, *55*(1), 104-110. https://doi.org/10.1007/s00239-001-2309-6

Bryant, P. (2023). Deep learning for protein complex structure prediction. *Current Opinion in Structural Biology*, *79*, 102529. https://doi.org/https://doi.org/10.1016/j.sbi.2023.102529

Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), 1265. https://doi.org/10.1038/s41467-022-28865-w

Bryant, P., Pozzati, G., Zhu, W., Shenoy, A., Kundrotas, P., & Elofsson, A. (2022). Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nature Communications*, *13*(1), 6028. https://doi.org/10.1038/s41467-022-33729-4

Brysbaert, G., Raouraoua, N., Mirabello, C., Thibaut, Blanchet, C., Wallner, B., & Lensink, M. (2024). *MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling*. https://doi.org/10.21203/rs.3.rs-4319486/v1

Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D., & Lawley, T. D. (2021). Massive expansion of human gut bacteriophage diversity. *Cell*, *184*(4), 1098-1109.e1099. https://doi.org/https://doi.org/10.1016/j.cell.2021.01.029

Carugo, O. (2008). Amino acid composition and protein dimension. *Protein Sci*, *17*(12), 2187-2191. https://doi.org/10.1110/ps.037762.108

CASP15 Abstracts. (2022). CASP15, https://predictioncenter.org/casp15/doc/CASP15_Abstracts.pdf

Chakravarty, D., McElfresh, G. W., Kundrotas, P. J., & Vakser, I. A. (2020). How to choose templates for modeling of protein complexes: Insights from benchmarking template-based docking. *Proteins*, *88*(8), 1070-1081. https://doi.org/10.1002/prot.25875

Chakravarty, D., Schafer, J. W., Chen, E. A., Thole, J. R., & Porter, L. L. (2023). AlphaFold2 has more to learn about protein energy landscapes. *bioRxiv*. https://doi.org/10.1101/2023.12.12.571380

Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, *12*, e82819. https://doi.org/10.7554/eLife.82819

Chen, C., Wu, T., Guo, Z., & Cheng, J. (2021(b)). Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. *Proteins: Structure, Function, and Bioinformatics*, *89*(6), 697-707. https://doi.org/https://doi.org/10.1002/prot.26052

Chen, J., & Siu, S. W. I. (2020). Machine Learning Approaches for Quality Assessment of Protein Structures. *Biomolecules*, *10*(4). https://doi.org/10.3390/biom10040626

Chen, J. W., Romero, P., Uversky, V. N., & Dunker, A. K. (2006). Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res*, *5*(4), 879-887. https://doi.org/10.1021/pr060048x

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010(a)). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, *66*(Pt 1), 12-21. https://doi.org/10.1107/s0907444909042073

Chen, X., Liu, J., Guo, Z., Wu, T., Hou, J., & Cheng, J. (2021). Protein model accuracy estimation empowered by deep learning and inter-residue distance prediction in CASP14. *Scientific Reports*, *11*(1), 10943. https://doi.org/10.1038/s41598-021-90303-6

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol*, *9*(2), 14. https://doi.org/10.1167/tvst.9.2.14

Christoffer, C., Bharadwaj, V., Luu, R., & Kihara, D. (2021). LZerD Protein-Protein Docking Webserver Enhanced With de novo Structure Prediction [Methods]. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.2021.724947

Cios, K. J., Kurgan, L. A., & Reformat, M. (2007). Machine learning in the life sciences. *IEEE Eng Med Biol Mag*, *26*(2), 14-16. https://doi.org/10.1109/memb.2007.335579

Consortium, T. U. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506-D515. https://doi.org/10.1093/nar/gky1049

Cozzone, A. J. (2010). Proteins: Fundamental Chemical Properties. In *Encyclopedia of Life Sciences*. https://doi.org/https://doi.org/10.1002/9780470015902.a0001330.pub2

Dai, B., & Bailey-Kellogg, C. (2021). Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, *37*(17), 2580-2588. https://doi.org/10.1093/bioinformatics/btab154

Dapkūnas, J., Olechnovič, K., & Venclovas, Č. (2021). Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1834-1843. https://doi.org/https://doi.org/10.1002/prot.26167

Darai, N., Hengphasatporn, K., Wolschann, P., Wolfinger, M. T., Shigeta, Y., Rungrotmongkol, T., & Harada, R. (2023). A Structural Refinement Technique for Protein-RNA Complexes Based on a Combination of AI-based Modeling and Flexible Docking: A Study of Musashi-1 Protein. *Bulletin of the Chemical Society of Japan*, *96*(7), 677-685. https://doi.org/10.1246/bcsj.20230092

Das, R. K., Ruff, K. M., & Pappu, R. V. (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol*, *32*, 102-112. https://doi.org/10.1016/j.sbi.2015.03.008

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Mol Biosyst*, *8*(1), 268-281. https://doi.org/10.1039/c1mb05231d

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. matrices for detecting distant relationships. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (Vol. 5, pp. 345-358). National biomedical research foundation Washington.

De Fonzo, V., Aluffi-Pentini, F., & Parisi, V. (2007). Hidden Markov Models in Bioinformatics. *Current Bioinformatics*, *2*(1), 49-61. https://doi.org/http://dx.doi.org/10.2174/157489307779314348

DeForte, S., & Uversky, V. N. (2016). Order, Disorder, and Everything in Between. *Molecules*, *21*(8). https://doi.org/10.3390/molecules21081090

Degiacomi, M. T. (2019). Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure*, *27*(6), 1034-1040.e1033. https://doi.org/10.1016/j.str.2019.03.018

Delgado-Cunningham, K., López, T., Khatib, F., Arias, C. F., & DuBois, R. M. (2022). Structure of the divergent human astrovirus MLB capsid spike. *Structure*, *30*(12), 1573-1581.e1573. https://doi.org/10.1016/j.str.2022.10.010

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A. M., Kourlaiev, A., d'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., Sunagawa, S.,…Jaillon, O. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, *2*(5), 100123. https://doi.org/https://doi.org/10.1016/j.xgen.2022.100123

Disfani, F. M., Hsu, W. L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., Uversky, V. N., & Kurgan, L. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, *28*(12), i75-83. https://doi.org/10.1093/bioinformatics/bts209

Dokholyan, N. V. (2020). Experimentally-driven protein structure modeling. *J Proteomics*, *220*, 103777. https://doi.org/10.1016/j.jprot.2020.103777

Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, *125*(7), 1731-1737. https://doi.org/10.1021/ja026939x

Dorn, M., MB, E. S., Buriol, L. S., & Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Comput Biol Chem*, *53pb*, 251-276. https://doi.org/10.1016/j.compbiolchem.2014.10.001

Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T., & MacKinnon, R. (1998). The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity. *Science*, *280*(5360), 69-77. https://doi.org/doi:10.1126/science.280.5360.69

Duarte, J. M., Srebniak, A., Schärer, M. A., & Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC Bioinformatics*, *13*, 334. https://doi.org/10.1186/1471-2105-13-334

Eastman, P., Galvelis, R., Peláez, R. P., Abreu, C. R. A., Farr, S. E., Gallicchio, E., Gorenko, A., Henry, M. M., Hu, F., Huang, J., Krämer, A., Michel, J., Mitchell, J. A., Pande, V. S., Rodrigues, J. P., Rodriguez-Guerra, J., Simmonett, A. C., Singh, S., Swails, J.,…Markland, T. E. (2024). OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials.

The Journal of Physical Chemistry B, *128*(1), 109-116. https://doi.org/10.1021/acs.jpcb.3c06662

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edmunds, N. S., Alharbi, S. M. A., Genc, A. G., Adiyaman, R., & McGuffin, L. J. (2023). Estimation of model accuracy in CASP15 using the ModFOLDdock server. *Proteins*, *91*(12), 1871-1878. https://doi.org/10.1002/prot.26532

Edmunds, N. S., Genc, A. G., & McGuffin, L. J. (2024). Benchmarking of AlphaFold2 accuracy self-estimates as indicators of empirical model quality and ranking—a comparison with independent model quality assessment programs. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btae491

Egbert, M., Ghani, U., Ashizawa, R., Kotelnikov, S., Nguyen, T., Desta, I., Hashemi, N., Padhorny, D., Kozakov, D., & Vajda, S. (2021). Assessing the binding properties of CASP14 targets and models. *Proteins*, *89*(12), 1922-1939. https://doi.org/10.1002/prot.26209

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell*, *44*(10), 7112-7127. https://doi.org/10.1109/tpami.2021.3095381

Elofsson, A. (2022). Protein Structure Prediction until CASP15. *ArXiv*, Article /abs/2212.07702.

Erdős, G., Pajkos, M., & Dosztányi, Z. (2021). IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, *49*(W1), W297-W303. https://doi.org/10.1093/nar/gkab408

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E.,…Hassabis, D. (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034. https://doi.org/10.1101/2021.10.04.463034

Feig, M. (2017). Computational protein structure refinement: Almost there, yet still so far to go. *Wiley Interdiscip Rev Comput Mol Sci*, *7*(3). https://doi.org/10.1002/wcms.1307

Feig, M., & Mirjalili, V. (2016). Protein structure refinement via molecular-dynamics simulations: What works and what does not? *Proteins*, *84 Suppl 1*(Suppl 1), 282-292. https://doi.org/10.1002/prot.24871

Ferrer-Bonsoms, J. A., Cassol, I., Fernández-Acín, P., Castilla, C., Carazo, F., & Rubio, A. (2020). ISOGO: Functional annotation of protein-coding splice variants. *Scientific Reports*, *10*(1), 1069. https://doi.org/10.1038/s41598-020-57974-z

Fiser, A. (2010). Template-based protein structure modeling. *Methods Mol Biol*, *673*, 73-94. https://doi.org/10.1007/978-1-60761-842-3_6

Fox, G., Sievers, F., & Higgins, D. G. (2015). Using de novo protein structure predictions to measure the quality of very large multiple sequence alignments. *Bioinformatics*, *32*(6), 814-820. https://doi.org/10.1093/bioinformatics/btv592

Fuchs, F. B., Worrall, D. E., Fischer, V., & Welling, M. (2020). *SE(3)-transformers: 3D roto-translation equivariant attention networks* Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada.

Gaber, A., & Pavšič, M. (2021). Modeling and Structure Determination of Homo-Oligomeric Proteins: An Overview of Challenges and Current Approaches. *Int J Mol Sci*, *22*(16). https://doi.org/10.3390/ijms22169081

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning* Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v48/gal16.html

Gazizov, A., Lian, A., Goverde, C., Ovchinnikov, S., & Polizzi, N. F. (2023). AF2BIND: Predicting ligand-binding sites using the pair representation of AlphaFold2. *bioRxiv*, 2023.2010.2015.562410. https://doi.org/10.1101/2023.10.15.562410

Ghani, U., Desta, I., Jindal, A., Khan, O., Jones, G., Kotelnikov, S., Padhorny, D., Vajda, S., & Kozakov, D. (2021). Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro. *bioRxiv*, 2021.2009.2007.459290. https://doi.org/10.1101/2021.09.07.459290

Goverde, C. A., Wolf, B., Khakzad, H., Rosset, S., & Correia, B. E. (2023). De novo protein design by inversion of the AlphaFold structure prediction network. *Protein Sci*, *32*(6), e4653. https://doi.org/10.1002/pro.4653

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, *84*(13), 4355-4358. https://doi.org/10.1073/pnas.84.13.4355

Grimsley, G. R., Scholtz, J. M., & Pace, C. N. (2009). A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci*, *18*(1), 247-251. https://doi.org/10.1002/pro.19

Guo, H.-B., Perminov, A., Bekele, S., Kedziora, G., Farajollahi, S., Varaljay, V., Hinkle, K., Molinero, V., Meister, K., Hung, C., Dennis, P., Kelley-Loughnane, N., & Berry, R. (2022). AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Scientific Reports*, *12*(1), 10696. https://doi.org/10.1038/s41598-022-14382-9

Guo, Z., Wu, T., Liu, J., Hou, J., & Cheng, J. (2021). Improving deep learning-based protein distance prediction in CASP14. *Bioinformatics*, *37*(19), 3190-3196. https://doi.org/10.1093/bioinformatics/btab355

Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., & Schwede, T. (2018). Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, *86 Suppl 1*(Suppl 1), 387-398. https://doi.org/10.1002/prot.25431

Haas, J., Gumienny, R., Barbato, A., Ackermann, F., Tauriello, G., Bertoni, M., Studer, G., Smolinski, A., & Schwede, T. (2019). Introducing "best single template" models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins*, *87*(12), 1378-1387. https://doi.org/10.1002/prot.25815

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., & Schwede, T. (2013). The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database (Oxford)*, *2013*, bat031. https://doi.org/10.1093/database/bat031

Hamamsy, T., Morton, J. T., Blackwell, R., Berenberg, D., Carriero, N., Gligorijevic, V., Strauss, C. E. M., Leman, J. K., Cho, K., & Bonneau, R. (2023). Protein remote homology detection and structural alignment using deep learning. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01917-2

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. https://books.google.co.uk/books?id=VRzITwgNV2UC

Heo, L., Arbour, C. F., & Feig, M. (2019). Driven to near-experimental accuracy by refinement via molecular dynamics simulations. *Proteins*, *87*(12), 1263-1275. https://doi.org/10.1002/prot.25759

Heo, L., & Feig, M. (2018). Experimental accuracy in protein structure refinement via molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, *115*(52), 13276-13281. https://doi.org/doi:10.1073/pnas.1811364115

Heo, L., & Feig, M. (2020). High-accuracy protein structures by combining machine-learning with physics-based refinement. *Proteins*, *88*(5), 637-642. https://doi.org/10.1002/prot.25847

Heo, L., Janson, G., & Feig, M. (2021). Physics-based protein structure refinement in the era of artificial intelligence. *Proteins*, *89*(12), 1870-1887. https://doi.org/10.1002/prot.26161

Heo, L., Lee, H., & Seok, C. (2016). GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. *Scientific Reports*, *6*(1), 32153. https://doi.org/10.1038/srep32153

Heo, L., Park, H., & Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res*, *41*(Web Server issue), W384-388. https://doi.org/10.1093/nar/gkt458

Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., & Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun*, *12*(1), 1340. https://doi.org/10.1038/s41467-021-21511-x

Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci*, *1*(3), 409-417. https://doi.org/10.1002/pro.5560010313

Hong, L., Sun, S., Zheng, L., Tan, Q., & Li, Y. (2021). *fastMSA: Accelerating Multiple Sequence Alignment with Dense Retrieval on Protein Language*. https://doi.org/10.1101/2021.12.20.473431

Hou, Q., Pucci, F., Pan, F., Xue, F., Rooman, M., & Feng, Q. (2022). Using metagenomic data to boost protein structure prediction and discovery. *Computational and Structural Biotechnology Journal*, *20*, 434-442. https://doi.org/https://doi.org/10.1016/j.csbj.2021.12.030

Huang, S. Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov Today*, *19*(8), 1081-1096. https://doi.org/10.1016/j.drudis.2014.02.005

Huber, R. (1979). Conformational flexibility in protein molecules. *Nature*, *280*(5723), 538-539. https://doi.org/10.1038/280538a0

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., & Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*, *323*(3), 573-584. https://doi.org/10.1016/s0022-2836(02)00969-5

Iantorno, S., Gori, K., Goldman, N., Gil, M., & Dessimoz, C. (2014). Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol Biol*, *1079*, 59-73. https://doi.org/10.1007/978-1-62703-646-7_4

Igashov, I., Olechnovič, K., Kadukova, M., Venclovas, Č., & Grudinin, S. (2021). VoroCNN: deep convolutional neural network built on 3D Voronoi tessellation of protein structures. *Bioinformatics*, *37*(16), 2332-2339. https://doi.org/10.1093/bioinformatics/btab118

Iserte, J. A., Lazar, T., Tosatto, S. C. E., Tompa, P., & Marino-Buslje, C. (2020). Chasing coevolutionary signals in intrinsically disordered proteins complexes. *Sci Rep*, *10*(1), 17962. https://doi.org/10.1038/s41598-020-74791-6

Janson, G., Valdes-Garcia, G., Heo, L., & Feig, M. (2023). Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, *14*(1), 774. https://doi.org/10.1038/s41467-023-36443-x

Jensen, J. H. (2008). Calculating pH and salt dependence of protein-protein binding. *Curr Pharm Biotechnol*, *9*(2), 96-102. https://doi.org/10.2174/138920108783955146

Ji, S., Oruç, T., Mead, L., Rehman, M. F., Thomas, C. M., Butterworth, S., & Winn, P. J. (2019). DeepCDpred: Inter-residue distance and contact prediction for improved prediction of protein structure. *PLOS ONE*, *14*(1), e0205214. https://doi.org/10.1371/journal.pone.0205214

Jiang, Q., Jin, X., Lee, S. J., & Yao, S. (2017). Protein secondary structure prediction: A survey of the state of the art. *J Mol Graph Model*, *76*, 379-402. https://doi.org/10.1016/j.jmgm.2017.07.015

Johansson-Åkhe, I., Mirabello, C., & Wallner, B. (2019). Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Scientific Reports*, *9*(1), 4267. https://doi.org/10.1038/s41598-019-38498-7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J.,…Hassabis, D. (2021a). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J.,…Hassabis, D. (2021b). Applying and improving AlphaFold at CASP14. *Proteins*, *89*(12), 1711-1721. https://doi.org/10.1002/prot.26257

Kandathil, S. M., Greener, J. G., Lau, A. M., & Jones, D. T. (2022). Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins. *Proc Natl Acad Sci U S A*, *119*(4). https://doi.org/10.1073/pnas.2113348119

Kastano, K., Erdős, G., Mier, P., Alanis-Lobato, G., Promponas, V. J., Dosztányi, Z., & Andrade-Navarro, M. A. (2020). Evolutionary Study of Disorder in Protein Sequences. *Biomolecules*, *10*(10). https://doi.org/10.3390/biom10101413

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845-858. https://doi.org/10.1038/nprot.2015.053

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Korkmaz, S., Duarte, J. M., Prlić, A., Goksuluk, D., Zararsiz, G., Saracbasi, O., Burley, S. K., & Rose, P. W. (2018). Investigation of protein quaternary structure via stoichiometry and symmetry information. *PLoS One*, *13*(6), e0197176. https://doi.org/10.1371/journal.pone.0197176

Kotowski, K., Smolarczyk, T., Roterman-Konieczna, I., & Stapor, K. (2021). ProteinUnet-An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. *J Comput Chem*, *42*(1), 50-59. https://doi.org/10.1002/jcc.26432

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., & Vajda, S. (2017). The ClusPro web server for protein–protein docking. *Nature Protocols*, *12*(2), 255-278. https://doi.org/10.1038/nprot.2016.169

Kožić, M., & Bertoša, B. (2024). Trajectory maps: molecular dynamics visualization and analysis. *NAR Genomics and Bioinformatics*, *6*(1). https://doi.org/10.1093/nargab/lqad114

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Kang, A., Brackenbrough, E., Bera, A. K.,…Baker, D. (2024). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, *384*(6693), eadl2528. https://doi.org/doi:10.1126/science.adl2528

Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, *372*(3), 774-797. https://doi.org/10.1016/j.jmb.2007.05.022

Kryshtafovych, A., Antczak, M., Szachniuk, M., Zok, T., Kretsch, R. C., Rangan, R., Pham, P., Das, R., Robin, X., Studer, G., Durairaj, J., Eberhardt, J., Sweeney, A., Topf, M., Schwede, T., Fidelis, K., & Moult, J. (2023b). New prediction categories in CASP15. *Proteins*, *91*(12), 1550-1557. https://doi.org/10.1002/prot.26515

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2023a). Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins*, *91*(12), 1539-1549. https://doi.org/10.1002/prot.26617

Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, *20*(11), 681-697. https://doi.org/10.1038/s41580-019-0163-x

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). *Simple and scalable predictive uncertainty estimation using deep ensembles* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Latysheva, N. S., Flock, T., Weatheritt, R. J., Chavali, S., & Babu, M. M. (2015). How do disordered regions achieve comparable functions to structured domains? *Protein Sci*, *24*(6), 909-922. https://doi.org/10.1002/pro.2674

Lee, J., Wu, S., & Zhang, Y. (2009). Ab Initio Protein Structure Prediction. In (pp. 3-25). https://doi.org/10.1007/978-1-4020-9058-5_1

Lehner, B., & Fraser, A. G. (2004). Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends in Genetics*, *20*(10), 468-472. https://doi.org/https://doi.org/10.1016/j.tig.2004.08.002

Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., & Apweiler, R. (2004). UniProt archive. *Bioinformatics*, *20*(17), 3236-3237. https://doi.org/10.1093/bioinformatics/bth191

Lesk, A. M. (2016). *Introduction to Protein Science: Architecture, Function, and Genomics*. Oxford University Press. https://books.google.co.uk/books?id=bbo6zQEACAAJ

Levy Karin, E., Mirdita, M., & Söding, J. (2020). MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, *8*(1), 48. https://doi.org/10.1186/s40168-020-00808-x

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123-1130. https://doi.org/doi:10.1126/science.ade2574

Liu, D., Zhang, B., Liu, J., Li, H., Song, L., & Zhang, G. (2023). Assessing protein model quality based on deep graph coupled networks using protein language model. *Briefings in Bioinformatics*, *25*(1). https://doi.org/10.1093/bib/bbad420

Liu, J., Guo, Z., Wu, T., Roy, R. S., Quadir, F., Chen, C., & Cheng, J. (2023a). Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15. *Communications Biology*, *6*(1), 1140. https://doi.org/10.1038/s42003-023-05525-3

Mahlich, Y., Steinegger, M., Rost, B., & Bromberg, Y. (2018). HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics*, *34*(13), i304-i312. https://doi.org/10.1093/bioinformatics/bty262

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., & Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Engineering, Design and Selection*, *14*(2), 105-113. https://doi.org/10.1093/protein/14.2.105

Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., & Pappu, R. V. (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci U S A*, *107*(18), 8183-8188. https://doi.org/10.1073/pnas.0911107107

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722-2728. https://doi.org/10.1093/bioinformatics/btt473

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE*, *6*(12), e28766. https://doi.org/10.1371/journal.pone.0028766

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., & Rost, B. (2022). Embeddings from protein language models predict conservation and variant effects. *Hum Genet*, *141*(10), 1629-1647. https://doi.org/10.1007/s00439-021-02411-y

Marsh, J. A., & Teichmann, S. A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*, *84*, 551-575. https://doi.org/10.1146/annurev-biochem-060614-034142

Mashaghi, A., Kramer, G., Lamb, D. C., Mayer, M. P., & Tans, S. J. (2014). Chaperone Action at the Single-Molecule Level. *Chemical Reviews*, *114*(1), 660-676. https://doi.org/10.1021/cr400326k

Mashiach, E., Nussinov, R., & Wolfson, H. J. (2010). FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Res*, *38*(Web Server issue), W457-461. https://doi.org/10.1093/nar/gkq373

McGuffin, L. J. (2008). Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, *24*(16), 1798-1804. https://doi.org/10.1093/bioinformatics/btn326

McGuffin, L. J. (2008(b)). Aligning sequences to structures. *Methods Mol Biol*, *413*, 61-90. https://doi.org/10.1007/978-1-59745-574-9_3

McGuffin, L. J., Adiyaman, R., Maghrabi, A. H. A., Shuid, A. N., Brackenridge, D. A., Nealon, J. O., & Philomina, L. S. (2019). IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Research*, *47*(W1), W408-W413. https://doi.org/10.1093/nar/gkz322

McGuffin, L. J., & Alharbi, S. M. A. (2024). ModFOLD9: A Web Server for Independent Estimates of 3D Protein Model Quality. *Journal of Molecular Biology*, 168531. https://doi.org/https://doi.org/10.1016/j.jmb.2024.168531

McGuffin, L. J., Edmunds, N. S., Genc, A. G., Alharbi, S. M. A., Salehe, Bajuna R., & Adiyaman, R. (2023). Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. *Nucleic Acids Research*, *51*(W1), W274-W280. https://doi.org/10.1093/nar/gkad297

McWhite, C. D., Armour-Garb, I., & Singh, M. (2023). Leveraging protein language models for accurate multiple sequence alignments. *Genome Res*, *33*(7), 1145-1153. https://doi.org/10.1101/gr.277675.123

Mendoza-Espinosa, P., García-González, V., Moreno, A., Castillo, R., & Mas-Oliva, J. (2009). Disorder-to-order conformational transitions in protein structure and its relationship to disease. *Mol Cell Biochem*, *330*(1-2), 105-120. https://doi.org/10.1007/s11010-009-0105-6

Miller, R. M., Jordan, B. T., Mehlferber, M. M., Jeffery, E. D., Chatzipantsiou, C., Kaur, S., Millikin, R. J., Dai, Y., Tiberi, S., Castaldi, P. J., Shortreed, M. R., Luckey, C. J., Conesa, A., Smith, L. M., Deslattes Mays, A., & Sheynkman, G. M. (2022). Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biology*, *23*(1), 69. https://doi.org/10.1186/s13059-022-02624-y

Minhas, F., Geiss, B. J., & Ben-Hur, A. (2014). PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins*, *82*(7), 1142-1155. https://doi.org/10.1002/prot.24479

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., & Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. *Proteins*, *69*(3), 511-520. https://doi.org/10.1002/prot.21502

Mirabello, C., & Wallner, B. (2019). rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLOS ONE*, *14*(8), e0220182. https://doi.org/10.1371/journal.pone.0220182

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679-682. https://doi.org/10.1038/s41592-022-01488-1

Mirdita, M., Steinegger, M., & Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, *35*(16), 2856-2858. https://doi.org/10.1093/bioinformatics/bty1057

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., & Steinegger, M. (2016). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, *45*(D1), D170-D176. https://doi.org/10.1093/nar/gkw1081

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., & Finn, R. D. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res*, *48*(D1), D570-d578. https://doi.org/10.1093/nar/gkz1035

Moal, I. H., Torchala, M., Bates, P. A., & Fernández-Recio, J. (2013). The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics*, *14*(1), 286. https://doi.org/10.1186/1471-2105-14-286

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., & Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins*, *82 Suppl 2*(0 2), 127-137. https://doi.org/10.1002/prot.24391

Monzon, A. M., Necci, M., Quaglia, F., Walsh, I., Zanotti, G., Piovesan, D., & Tosatto, S. C. E. (2020). Experimentally Determined Long Intrinsically Disordered Protein Regions Are Now Abundant in the Protein Data Bank. *Int J Mol Sci*, *21*(12). https://doi.org/10.3390/ijms21124496

Morehead, A., Chen, X., Wu, T., Liu, J., & Cheng, J. (2022). EGR: Equivariant Graph Refinement and Assessment of 3D Protein Complex Structures. *ArXiv*, *abs/2205.10390*.

Moriwaki, Y. (2021). *AlphaFold2 can also predict heterocomplexes. All you have to do is input the two sequences you want to predict and connect them with a long linker.*Twitter. https://twitter.com/Ag_smith/status/1417063635000598528

Morris, R., Black, K. A., & Stollar, E. J. (2022). Uncovering protein function: from classification to complexes. *Essays Biochem*, *66*(3), 255-285. https://doi.org/10.1042/ebc20200108

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, *15*(3), 285-289. https://doi.org/10.1016/j.sbi.2005.05.011

Mukherjee, S., & Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Research*, *37*(11), e83-e83. https://doi.org/10.1093/nar/gkp318

Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., Proal, A. D., Fischbach, M. A., Bhatt, A. S., Hugenholtz, P., & Kyrpides, N. C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology*, *6*(7), 960-970. https://doi.org/10.1038/s41564-021-00928-6

Necci, M., Piovesan, D., Dosztányi, Z., Tompa, P., & Tosatto, S. C. E. (2018). A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics*, *34*(3), 445-452. https://doi.org/10.1093/bioinformatics/btx590

Negroni, J., Mosca, R., & Aloy, P. (2014). Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone. *Structure*, *22*(9), 1356-1362. https://doi.org/https://doi.org/10.1016/j.str.2014.07.009

Nussinov, R., Tsai, C.-J., Xin, F., & Radivojac, P. (2012). Allosteric post-translational modification codes. *Trends in biochemical sciences*, *37*(10), 447-455.

Nute, M., Saleh, E., & Warnow, T. (2019). Evaluating Statistical Multiple Sequence Alignment in Comparison to Other Alignment Methods on Protein Data Sets. *Syst Biol*, *68*(3), 396-411. https://doi.org/10.1093/sysbio/syy068

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750-1758. https://doi.org/https://doi.org/10.1016/j.csbj.2021.03.022

Ohue, M., Shimoda, T., Suzuki, S., Matsuzaki, Y., Ishida, T., & Akiyama, Y. (2014). MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers. *Bioinformatics*, *30*(22), 3281-3283. https://doi.org/10.1093/bioinformatics/btu532

Olzscha, H. (2019). Posttranslational modifications and proteinopathies: how guardians of the proteome are defeated. *Biological Chemistry*, *400*(7), 895-915. https://doi.org/doi:10.1515/hsz-2018-0458

Orlando, G., Raimondi, D., Codicè, F., Tabaro, F., & Vranken, W. (2022). Prediction of Disordered Regions in Proteins with Recurrent Neural Networks and Protein Dynamics. *J Mol Biol*, *434*(12), 167579. https://doi.org/10.1016/j.jmb.2022.167579

Osadchy, M., & Kolodny, R. (2021). How Deep Learning Tools Can Help Protein Engineers Find Good Sequences. *The Journal of Physical Chemistry B*, *125*(24), 6440-6450. https://doi.org/10.1021/acs.jpcb.1c02449

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE transactions on neural networks and learning systems*, *32*(2), 604-624. https://doi.org/10.1109/tnnls.2020.2979670

Ouyang, J., Huang, N., & Jiang, Y. (2020). A single-model quality assessment method for poor quality protein structure. *BMC Bioinformatics*, *21*(1), 157. https://doi.org/10.1186/s12859-020-3499-5

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, *355*(6322), 294-298. https://doi.org/doi:10.1126/science.aah4043

Owens, J., Houston, M., Luebke, D., Green, S., Stone, J., & Phillips, J. (2008). GPU computing. *Proceedings of the IEEE*, *96*, 879-899. https://doi.org/10.1109/JPROC.2008.917757

Ozden, B., Kryshtafovych, A., & Karaca, E. (2021). Assessment of the CASP14 assembly predictions. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1787-1799. https://doi.org/https://doi.org/10.1002/prot.26199

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, *40*(12), 1413-1415. https://doi.org/10.1038/ng.259

Pancsa, R., Zsolyomi, F., & Tompa, P. (2018). Co-Evolution of Intrinsically Disordered Proteins with Folded Partners Witnessed by Evolutionary Couplings. *Int J Mol Sci*, *19*(11). https://doi.org/10.3390/ijms19113315

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, *37*(4), 205-211. https://doi.org/doi:10.1073/pnas.37.4.205

Peng, Z., Wang, W., Han, R., Zhang, F., & Yang, J. (2022). Protein structure prediction in the deep learning era. *Current Opinion in Structural Biology*, *77*, 102495. https://doi.org/https://doi.org/10.1016/j.sbi.2022.102495

Petti, S., Bhattacharya, N., Rao, R., Dauparas, J., Thomas, N., Zhou, J., Rush, A. M., Koo, P., & Ovchinnikov, S. (2022). End-to-end learning of multiple sequence alignments with differentiable Smith–Waterman. *Bioinformatics*, *39*(1). https://doi.org/10.1093/bioinformatics/btac724

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T., & Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, *30*(12), 1771-1773. https://doi.org/10.1093/bioinformatics/btu097

Pinheiro, F., Santos, J., & Ventura, S. (2021). AlphaFold and the amyloid landscape. *J Mol Biol*, *433*(20), 167059. https://doi.org/10.1016/j.jmb.2021.167059

Pratt, C. W., & Cornely, K. (2012). *Essential Biochemistry*. John Wiley & Sons, Incorporated. https://books.google.co.uk/books?id=FVABCgAAQBAJ

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897. https://doi.org/10.1007/s11431-020-1647-3

Quadir, F., Roy, R. S., Soltanikazemi, E., & Cheng, J. (2021). DeepComplex: A Web Server of Predicting Protein Complex Structures by Deep Learning Inter-chain Contact Prediction and Distance-Based Modelling. *Front Mol Biosci*, *8*, 716973. https://doi.org/10.3389/fmolb.2021.716973

Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., Pajkos, M., Lazar, T., Peña-Díaz, S., Santos, J., Ács, V., Farahi, N., Fichó, E., Aspromonte, Maria C., Bassot, C., Chasapi, A., Davey, Norman E., Davidović, R., Dobson, L.,…Piovesan, D. (2021). DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research*, *50*(D1), D480-D487. https://doi.org/10.1093/nar/gkab1082

Quignot, C., Rey, J., Yu, J., Tufféry, P., Guerois, R., & Andreani, J. (2018). InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res*, *46*(W1), W408-w416. https://doi.org/10.1093/nar/gky377

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., & Rives, A. (2021a). *MSA Transformer* Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/rao21a.html

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., & Rives, A. (2021b). MSA transformer. *International Conference on Machine Learning*, 8844-8856.

Réau, M., Renaud, N., Xue, L. C., & Bonvin, A. M. J. J. (2022). DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, *39*(1). https://doi.org/10.1093/bioinformatics/btac759

Rehman I, Kerndt CC, & Botelho S. (2024). Biochemistry, Tertiary Protein Structure. In. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; (Updated 2024 May). https://www.ncbi.nlm.nih.gov/books/NBK470269/

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173-175. https://doi.org/10.1038/nmeth.1818

Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M. F., Bonvin, A. M. J. J., & Xue, L. C. (2021). DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. *Nature Communications*, *12*(1), 7068. https://doi.org/10.1038/s41467-021-27396-0

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, Maxwell L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, Lucy J., Curtis, T., Escobar-Zepeda, A., Gurbich, Tatiana A., Kale, V., Korobeynikov, A., Raj, S., Rogers, Alexander B., Sakharova, E., Sanchez, S.,…Finn, Robert D. (2022). MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, *51*(D1), D753-D759. https://doi.org/10.1093/nar/gkac1080

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*, *118*(15). https://doi.org/10.1073/pnas.2016239118

Robin, X., Haas, J., Gumienny, R., Smolinski, A., Tauriello, G., & Schwede, T. (2021). Continuous Automated Model EvaluatiOn (CAMEO)-Perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins*, *89*(12), 1977-1986. https://doi.org/10.1002/prot.26213

Roche, D. B., & McGuffin, L. J. (2016). Toolbox for Protein Structure Prediction. *Methods Mol Biol*, *1369*, 363-377. https://doi.org/10.1007/978-1-4939-3145-3_23

Roney, J. P., & Ovchinnikov, S. (2022). State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys Rev Lett*, *129*(23), 238101. https://doi.org/10.1103/PhysRevLett.129.238101

Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Green, R. K., Goodsell, D. S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A. S., Shao, C.,…Burley, S. K. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res*, *45*(D1), D271-d281. https://doi.org/10.1093/nar/gkw1000

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, *12*(2), 85-94. https://doi.org/10.1093/protein/12.2.85

Rost, B., & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, *19*(1), 55-72. https://doi.org/10.1002/prot.340190108

Rost, B., & Sander, C. (1996). Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct*, *25*, 113-136. https://doi.org/10.1146/annurev.bb.25.060196.000553

Sala, D., Engelberger, F., McHaourab, H. S., & Meiler, J. (2023). Modeling conformational states of proteins with AlphaFold. *Current Opinion in Structural Biology*, *81*, 102645. https://doi.org/https://doi.org/10.1016/j.sbi.2023.102645

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., Gonik, E., García Melani, A., Novomisky Nechcoff, J., & Salas, M. N. (2022). Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics*, *38*(10), 2742-2748.

Sanchez-Garcia, R., Sorzano, C. O. S., Carazo, J. M., & Segura, J. (2018). BIPSPI: a method for the prediction of partner-specific protein–protein interfaces. *Bioinformatics*, *35*(3), 470-477. https://doi.org/10.1093/bioinformatics/bty647

Sanvictores, T., & Farci, F. (2023). Biochemistry, Primary Protein Structure. In *StatPearls*. StatPearls Publishing Copyright © 2023, StatPearls Publishing LLC.

Schindler, C. E., de Vries, S. J., & Zacharias, M. (2015). iATTRACT: simultaneous global and local interface optimization for protein-protein docking refinement. *Proteins*, *83*(2), 248-258. https://doi.org/10.1002/prot.24728

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw*, *61*, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Seffernick, J. T., & Lindert, S. (2020). Hybrid methods for combined experimental and computational determination of protein structure. *The Journal of Chemical Physics*, *153*(24). https://doi.org/10.1063/5.0026025

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710. https://doi.org/10.1038/s41586-019-1923-7

Sevier, C. S., & Kaiser, C. A. (2002). Formation and transfer of disulphide bonds in living cells. *Nat Rev Mol Cell Biol*, *3*(11), 836-847. https://doi.org/10.1038/nrm954

Shuvo, M. H., Karim, M., Roche, R., & Bhattacharya, D. (2023). PIQLE: protein-protein interface quality estimation by deep graph learning of multimeric interaction geometries. *bioRxiv*. https://doi.org/10.1101/2023.02.14.528528

Siew, N., Elofsson, A., Rychlewski, L., & Fischer, D. (2000). MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics (Oxford, England)*, *16*, 776-785. https://doi.org/10.1093/bioinformatics/16.9.776

Sikic, K., & Carugo, O. (2010). Protein sequence redundancy reduction: comparison of various method. *Bioinformation*, *5*(6), 234-239. https://doi.org/10.6026/97320630005234

Singh, G. P., Ganapathi, M., & Dash, D. (2007). Role of intrinsic disorder in transient interactions of hub proteins. *Proteins*, *66*(4), 761-765. https://doi.org/10.1002/prot.21281

Skolnick, J., Gao, M., Zhou, H., & Singh, S. (2021). AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *J Chem Inf Model*, *61*(10), 4827-4831. https://doi.org/10.1021/acs.jcim.1c01114

Song, J., Tan, H., Wang, M., Webb, G. I., & Akutsu, T. (2012b). TANGLE: Two-Level Support Vector Regression Approach for Protein Backbone Torsion Angle Prediction from Primary Sequences. *PLOS ONE*, *7*(2), e30361. https://doi.org/10.1371/journal.pone.0030361

Song, Q., Li, T., Cong, P., Sun, J., Li, D., & Tang, S. (2012a). Predicting Turns in Proteins with a Unified Model. *PLOS ONE*, *7*(11), e48389. https://doi.org/10.1371/journal.pone.0048389

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929–1958.

Steenwyk, J. L., Buida, T. J., III, Li, Y., Shen, X.-X., & Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biology*, *18*(12), e3001007. https://doi.org/10.1371/journal.pbio.3001007

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019a). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, *20*(1), 473. https://doi.org/10.1186/s12859-019-3019-7

Steinegger, M., Mirdita, M., & Söding, J. (2019b). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, *16*(7), 603-606. https://doi.org/10.1038/s41592-019-0437-4

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026-1028. https://doi.org/10.1038/nbt.3988

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, *9*(1), 2542. https://doi.org/10.1038/s41467-018-04964-5

Studer, G., Tauriello, G., & Schwede, T. (2023). Assessment of the assessment—All about complexes. *Proteins: Structure, Function, and Bioinformatics*, *91*(12), 1850-1860. https://doi.org/https://doi.org/10.1002/prot.26612

Suh, D., Lee, J. W., Choi, S., & Lee, Y. (2021). Recent Applications of Deep Learning Methods on Evolution- and Contact-Based Protein Structure Prediction. *International Journal of Molecular Sciences*, *22*(11), 6032. https://www.mdpi.com/1422-0067/22/11/6032

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282-1288. https://doi.org/10.1093/bioinformatics/btm098

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & Consortium, t. U. (2014). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926-932. https://doi.org/10.1093/bioinformatics/btu739

Swamy, K. B. S., Schuyler, S. C., & Leu, J.-Y. (2021). Protein Complexes Form a Basis for Complex Hybrid Incompatibility [Review]. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.609766

Szilágyi, A., Györffy, D., & Závodszky, P. (2008). The twilight zone between protein order and disorder. *Biophys J*, *95*(4), 1612-1626. https://doi.org/10.1529/biophysj.108.131151

Szilagyi, A., & Zhang, Y. (2014). Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol*, *24*, 10-23. https://doi.org/10.1016/j.sbi.2013.11.005

Talley, K., & Alexov, E. (2010). On the pH-optimum of activity and stability of proteins. *Proteins*, *78*(12), 2699-2706. https://doi.org/10.1002/prot.22786

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015). Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology*, *64*(5), 778-791. https://doi.org/10.1093/sysbio/syv033

Terashi, G., & Takeda-Shitaka, M. (2015). CAB-Align: A Flexible Protein Structure Alignment Method Based on the Residue-Residue Contact Area. *PLOS ONE*, *10*(10), e0141440. https://doi.org/10.1371/journal.pone.0141440

Terwilliger, T. C., Poon, B. K., Afonine, P. V., Schlicksup, C. J., Croll, T. I., Millán, C., Richardson, J. S., Read, R. J., & Adams, P. D. (2022). Improved AlphaFold modeling with implicit experimental information. *Nature Methods*, *19*(11), 1376-1382. https://doi.org/10.1038/s41592-022-01645-6

Thomas, J., Ramakrishnan, N., & Bailey-Kellogg, C. (2008). Graphical models of residue coupling in protein families. *IEEE/ACM Trans Comput Biol Bioinform*, *5*(2), 183-197. https://doi.org/10.1109/tcbb.2007.70225

Thornton, J. M., Laskowski, R. A., & Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med*, *27*(10), 1666-1669. https://doi.org/10.1038/s41591-021-01533-0

Tian, H., Jiang, X., Trozzi, F., Xiao, S., Larson, E. C., & Tao, P. (2021). Explore Protein Conformational Space With Variational Autoencoder [Original Research]. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.2021.781635

Titeca, K., Lemmens, I., Tavernier, J., & Eyckerman, S. (2019). Discovering cellular protein-protein interactions: Technological strategies and opportunities. *Mass Spectrometry Reviews*, *38*(1), 79-111. https://doi.org/https://doi.org/10.1002/mas.21574

Tompa, P., Davey, N. E., Gibson, T. J., & Babu, M. M. (2014). A million peptide motifs for the molecular biologist. *Mol Cell*, *55*(2), 161-169. https://doi.org/10.1016/j.molcel.2014.05.032

Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I., Dunker, A. K., & Uversky, V. N. (2009). Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, *31*(3), 328-335.

Torchala, M., Moal, I. H., Chaleil, R. A. G., Fernandez-Recio, J., & Bates, P. A. (2013). SwarmDock: a server for flexible protein–protein docking. *Bioinformatics*, *29*(6), 807-809. https://doi.org/10.1093/bioinformatics/btt038

Torrens-Fontanals, M., Stepniewski, T. M., Gloriam, D. E., & Selent, J. (2021). Structural dynamics bridge the gap between the genetic and functional levels of GPCRs. *Current Opinion in Structural Biology*, *69*, 150-159. https://doi.org/https://doi.org/10.1016/j.sbi.2021.04.005

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T. A., Berger, B., Sander, C., & Marks, D. S. (2016). Structured States of Disordered Proteins from Genomic Sequences. *Cell*, *167*(1), 158-170.e112. https://doi.org/10.1016/j.cell.2016.09.010

Tovchigrechko, A., & Vakser, I. A. (2006). GRAMM-X public web server for protein–protein docking. *Nucleic Acids Research*, *34*(suppl_2), W310-W314. https://doi.org/10.1093/nar/gkl206

Tretter, V., Ehya, N., Fuchs, K., & Sieghart, W. (1997). Stoichiometry and assembly of a recombinant GABAA receptor subtype. *J Neurosci*, *17*(8), 2728-2737. https://doi.org/10.1523/jneurosci.17-08-02728.1997

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A.,…Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590-596. https://doi.org/10.1038/s41586-021-03828-1

Uversky, V. N. (2014). The triple power of D³: protein intrinsic disorder in degenerative diseases. *Front Biosci (Landmark Ed)*, *19*(2), 181-258. https://doi.org/10.2741/4204

Uversky, V. N., & Dunker, A. K. (2010). Understanding protein non-folding. *Biochim Biophys Acta*, *1804*(6), 1231-1264. https://doi.org/10.1016/j.bbapap.2010.01.017

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics*, *41*(3), 415-427. https://doi.org/https://doi.org/10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., & Dunker, A. K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res*, *6*(6), 2351-2366. https://doi.org/10.1021/pr0701411

Vakser, I. A. (2014). Protein-protein docking: from interaction to interactome. *Biophys J*, *107*(8), 1785-1793. https://doi.org/10.1016/j.bpj.2014.08.033

van den Berg, B., Ellis, R. J., & Dobson, C. M. (1999). Effects of macromolecular crowding on protein folding and aggregation. *The EMBO Journal*, *18*(24), 6927-6933. https://doi.org/https://doi.org/10.1093/emboj/18.24.6927

van Dijk, E., Hoogeveen, A., & Abeln, S. (2015). The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLOS Computational Biology*, *11*(5), e1004277. https://doi.org/10.1371/journal.pcbi.1004277

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01773-0

Vanni, C., Schechter, M. S., Acinas, S. G., Barberán, A., Buttigieg, P. L., Casamayor, E. O., Delmont, T. O., Duarte, C. M., Eren, A. M., Finn, R. D., Kottmann, R., Mitchell, A., Sánchez, P., Siren, K., Steinegger, M., Gloeckner, F. O., & Fernàndez-Guerra, A. (2022). Unifying the known and unknown microbial coding sequence space. *eLife*, *11*, e67667. https://doi.org/10.7554/eLife.67667

Verburgt, J., & Kihara, D. (2022). Benchmarking of structure refinement methods for protein complex models. *Proteins*, *90*(1), 83-95. https://doi.org/10.1002/prot.26188

Wallner, B. (2023a). Improved multimer prediction using massive sampling with AlphaFold in CASP15. *Proteins*, *91*(12), 1734-1746. https://doi.org/10.1002/prot.26562

Wallner, B. (2023b). AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*, *39*(9). https://doi.org/10.1093/bioinformatics/btad573

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470-476. https://doi.org/10.1038/nature07509

Wang, Q., Wei, J., Zhou, Y., Lin, M., Ren, R., Wang, S., Cui, S., & Li, Z. (2022). Prior knowledge facilitates low homologous protein secondary structure prediction with DSM distillation. *Bioinformatics*, *38*(14), 3574-3581. https://doi.org/10.1093/bioinformatics/btac351

Wang, X., Flannery, S. T., & Kihara, D. (2021). Protein Docking Model Evaluation by Graph Neural Networks [Original Research]. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.2021.647915

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, *337*(3), 635-645. https://doi.org/10.1016/j.jmb.2004.02.002

Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., & Kern, D. (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, *625*(7996), 832-839. https://doi.org/10.1038/s41586-023-06832-9

Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*, *54*, 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3

Wergin, W. (2006). Essential cell biology (2nd edition). By Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Garland Science/Taylor and Francis Group, New York and London (2003). ISBN 0-8153-3480-X; hardbac. *Scanning*, *27*, 213-213. https://doi.org/10.1002/sca.4950270409

Williamson, M. (2012). *How Proteins Work*. CRC Press. https://books.google.co.uk/books?id=TSsWBAAAQBAJ

Wright, P. E., & Dyson, H. J. (2009). Linking folding and binding. *Current Opinion in Structural Biology*, *19*(1), 31-38. https://doi.org/https://doi.org/10.1016/j.sbi.2008.12.003

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022a). High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.2007.2021.500999. https://doi.org/10.1101/2022.07.21.500999

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022b). High-resolution *de novo* structure prediction from primary sequence. *bioRxiv*, 2022.2007.2021.500999. https://doi.org/10.1101/2022.07.21.500999

Wu, T., Guo, Z., & Cheng, J. (2023). Atomic protein structure refinement using all-atom graph representations and SE(3)-equivariant graph transformer. *Bioinformatics*, *39*(5). https://doi.org/10.1093/bioinformatics/btad298

wwPDBconsortium. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, *47*(D1), D520-D528. https://doi.org/10.1093/nar/gky949

Xu, D., & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J*, *101*(10), 2525-2534. https://doi.org/10.1016/j.bpj.2011.10.024

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715-1735. https://doi.org/https://doi.org/10.1002/prot.24065

Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, *116*(34), 16856-16865. https://doi.org/doi:10.1073/pnas.1821309116

Xue, L. C., Dobbs, D., Bonvin, A. M. J. J., & Honavar, V. (2015). Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters*, *589*(23), 3516-3526. https://doi.org/https://doi.org/10.1016/j.febslet.2015.10.003

Yan, C., & Wang, Y. (2014). A graph kernel method for DNA-binding site prediction. *BMC Systems Biology*, *8*(4), S10. https://doi.org/10.1186/1752-0509-8-S4-S10

Yan, C., Wu, F., Jernigan, R. L., Dobbs, D., & Honavar, V. (2008). Characterization of protein-protein interfaces. *Protein J*, *27*(1), 59-70. https://doi.org/10.1007/s10930-007-9108-x

Yang, A. S., & Honig, B. (1993). On the pH dependence of protein stability. *J Mol Biol*, *231*(2), 459-474. https://doi.org/10.1006/jmbi.1993.1294

Yang, P., Zheng, W., Ning, K., & Zhang, Y. (2021). Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. *Proceedings of the National Academy of Sciences*, *118*(49), e2110828118. https://doi.org/doi:10.1073/pnas.2110828118

Yang, W., Liu, C., & Li, Z. (2023(a)). Lightweight Fine-tuning a Pretrained Protein Language Model for Protein Secondary Structure Prediction. *bioRxiv*, 2023.2003.2022.530066. https://doi.org/10.1101/2023.03.22.530066

Yang, Z., Zeng, X., Zhao, Y., & Chen, R. (2023(b)). AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduction and Targeted Therapy*, *8*(1), 115. https://doi.org/10.1038/s41392-023-01381-z

Yanofsky, C., Horn, V., & Thorpe, D. (1964). PROTEIN STRUCTURE RELATIONSHIPS REVEALED BY MUTATIONAL ANALYSIS. *Science*, *146*(3651), 1593-1594. https://doi.org/10.1126/science.146.3651.1593

Yin, R., Feng, B. Y., Varshney, A., & Pierce, B. G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci*, *31*(8), e4379. https://doi.org/10.1002/pro.4379

Yu, X., Wang, C., & Li, Y. (2006). Classification of protein quaternary structure by functional domain composition. *BMC Bioinformatics*, *7*, 187. https://doi.org/10.1186/1471-2105-7-187

Yuan, Q., Chen, S., Rao, J., Zheng, S., Zhao, H., & Yang, Y. (2022). AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Brief Bioinform*, *23*(2). https://doi.org/10.1093/bib/bbab564

Yuan, Z., Shen, T., Xu, S., Yu, L., Ren, R., & Sun, S. (2023). AF2-Mutation: Adversarial Sequence Mutations against AlphaFold2 on Protein Tertiary Structure Prediction. *arXiv preprint arXiv:2305.08929*.

Zemla, A., Venclovas, C., Moult, J., & Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins*, *Suppl 3*, 22-29. https://doi.org/10.1002/(sici)1097-0134(1999)37:3+<22::aid-prot5>3.3.co;2-n

Zhan, Q., Fu, Y., Jiang, Q., Liu, B., Peng, J., & Wang, Y. (2020). SpliVert: A Protein Multiple Sequence Alignment Refinement Method Based on Splitting-Splicing Vertically. *Protein Pept Lett*, *27*(4), 295-302. https://doi.org/10.2174/0929866526666190806143959

Zhang, B., Li, J., & Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics*, *19*(1), 293. https://doi.org/10.1186/s12859-018-2280-5

Zhang, C., Zhao, Y., Braun, E. L., & Mirarab, S. (2021). TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution*, *12*(11), 2145-2158. https://doi.org/https://doi.org/10.1111/2041-210X.13696

Zhang, C., Zheng, W., Mortuza, S. M., Li, Y., & Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, *36*(7), 2105-2112. https://doi.org/10.1093/bioinformatics/btz863

Zhang, J., Liang, Y., & Zhang, Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure*, *19*(12), 1784-1795. https://doi.org/10.1016/j.str.2011.09.022

Zhang, T., Faraggi, E., Xue, B., Dunker, A. K., Uversky, V. N., & Zhou, Y. (2012). SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn*, *29*(4), 799-813. https://doi.org/10.1080/073911012010525022

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, *57*(4), 702-710. https://doi.org/10.1002/prot.20264

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, *33*(7), 2302-2309. https://doi.org/10.1093/nar/gki524

Zhou, X., Zheng, W., Li, Y., Pearce, R., Zhang, C., Bell, E. W., Zhang, G., & Zhang, Y. (2022). I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and

function prediction. *Nature Protocols*, *17*(10), 2326-2353. https://doi.org/10.1038/s41596-022-00728-0

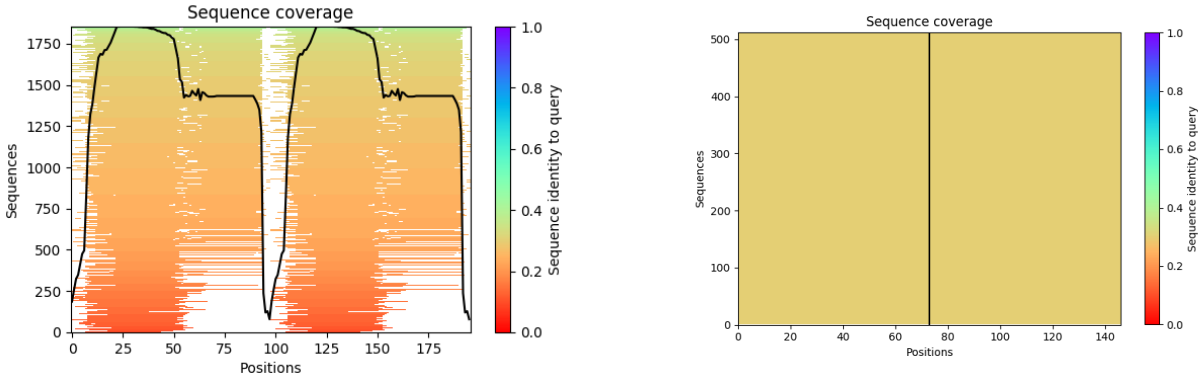
Data availability:

A Data is open access from The Critical Assessment of Protein Structure Prediction (CASP) Community Resource CASP14 DATA: (https://predictioncenter.org/download_area/CASP14/) and CASP15 DATA (https://predictioncenter.org/download_area/CASP15). Since the data for chapters (2, 3, and 4) based on open access, data can be reproducible using Alphafold2. The data in Chapter 5 is maintained by the Continuous Automated Model EvaluatiOn (CAMEO) organization.

Appendices

Appendix 1

a)



b)

T1083(A2):

GAMGSEIEHIEEAIANAKTKADHERLVAHYEEEAKRLEKKSEEYQELAKVYKKITDVYPNIRSYMVLHYQNLTR
RYKEAAEENRALAKLHHELAIVED

T1084(A2):

MAAHKGAEHHHKAAEHHEQAAKHHHAAAEHHEKGEHEQAAHHADTAYAHHKHAEEHAAQAAKHDAEHHA
PKPH

c)

Figure S.1 The drawback of AF2M in terms of homology sequence mining.

These graphics demonstrated that the differences between sequence identity generated by AF2M integrated with MMseq2. The difference means the MSA in T1084 target can be failed. A) The sequence coverage of T1083 (left) and T1084 (right). B) Number of sequences of T1083 (above), of T1084 (below) C) Reference structures of T1083 (left) and T1084 (right).

Appendix 2

Table S.1 Performance comparison between AF2M and AF2_Advanced using the same recycling process, according to the cumulative scores of the modelled complexes of CASP 14 targets.

The Wilcoxon signed-rank tests were used to evaluate whether the quality scores of models generated by AF2M are statistically different from those of models predicted by AF2_Advanced, given a certain recycle value. H_0: The observed quality scores of models generated using *y* cycles by AF2M are equal to or lower those of models generated using *x* cycles by AF2_Advanced, where *x* and *y* are same integers between 1 and 48. H_1: The observed quality scores of models produced after *y* cycles AF2M are greater than those generated using *x* cycles by AF2_Advanced. P-values ≤ 0.05 indicate significant statistical differences. P-value where H_0 was rejected are in boldface. (n = 78 models for the observed TM-score, IDDT, and QS-score) Wilcoxon signed-rank test was performed in the R program.

The cumulative scores of CASP targets generated by the AF2 versions			
Wilcoxon signed-rank test	p-value		
For pairwise cycles (AF2-Multimer and AF2_Advanced)	TM-score	IDDT score	QS-score
cycle 1 - cycle 1	2.42E-01	5.83E-01	**3.78E-02**
cycle 3 - cycle 3	2.01E-01	2.88E-01	9.26E-02
cycle 6 - cycle 6	5.83E-01	7.12E-01	2.07E-01
cycle 12 - cycle 12	6.63E-01	7.35E-01	4.19E-01
cycle 24 - cycle 24	5.28E-01	7.35E-01	4.19E-01
cycle 48 - cycle 48	3.63E-01	5.28E-01	1.54E-01

Appendix 3

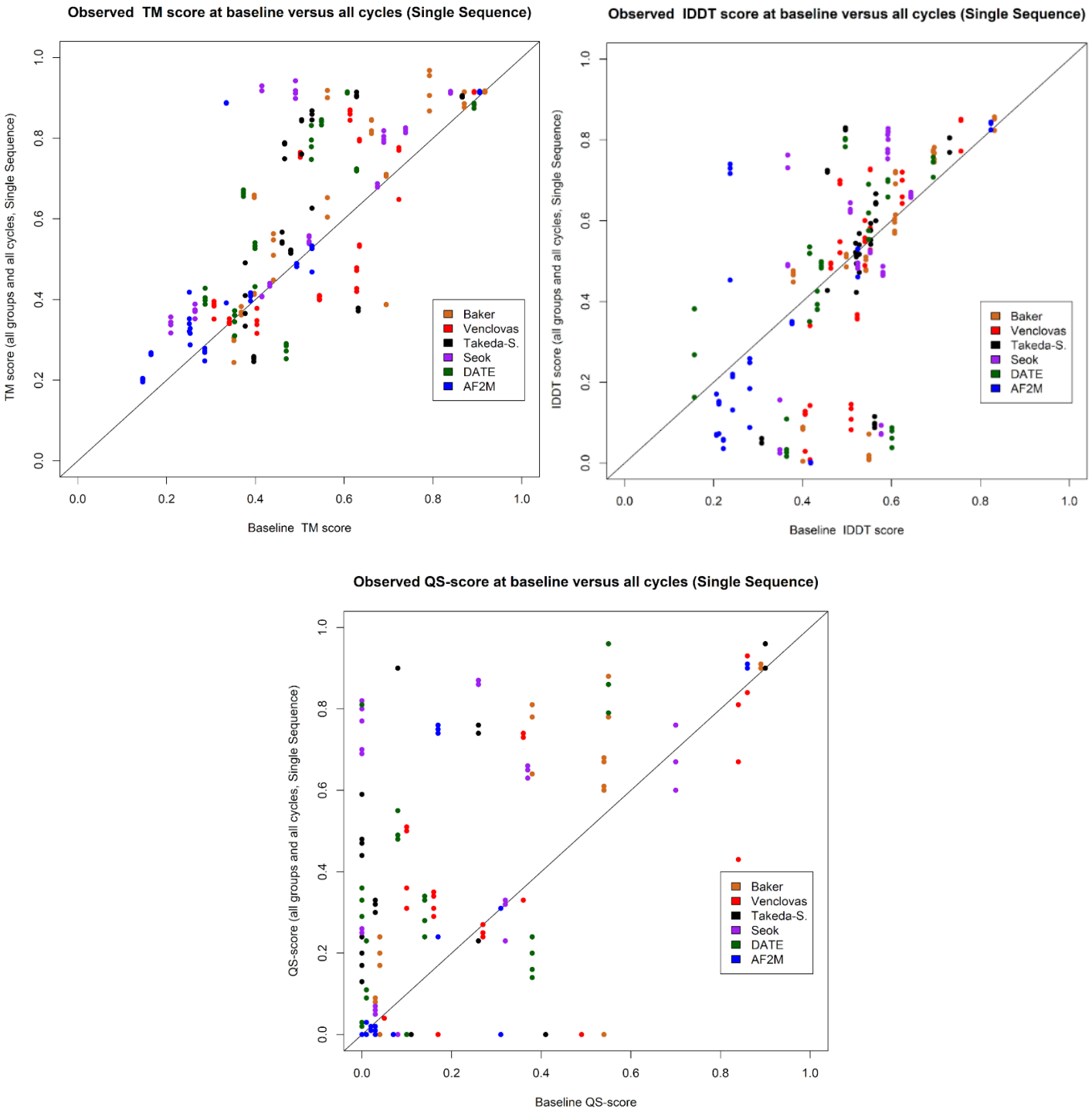


Figure S.2 A comparison of the observed and baseline three quality scores for the CASP14 models after recycling in the SS method.

Three scatter plots showing the improvement of models following recycling. The plots compare the observed TM-score, IDDT, and QS -scores for the improved models (y-axis) versus the baseline TM-score, IDDT, and QS-scores (x-axis) for the CASP14 models generated during all recycles (1-3-6-12) for six group models generated by AF2M using the SS method with recycling, respectively. The minimum values for TM-score, IDDT, and QS-score are 0, while the maximum values are 1. Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda_Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. The scatter plots were drawn using R.

Appendix 4

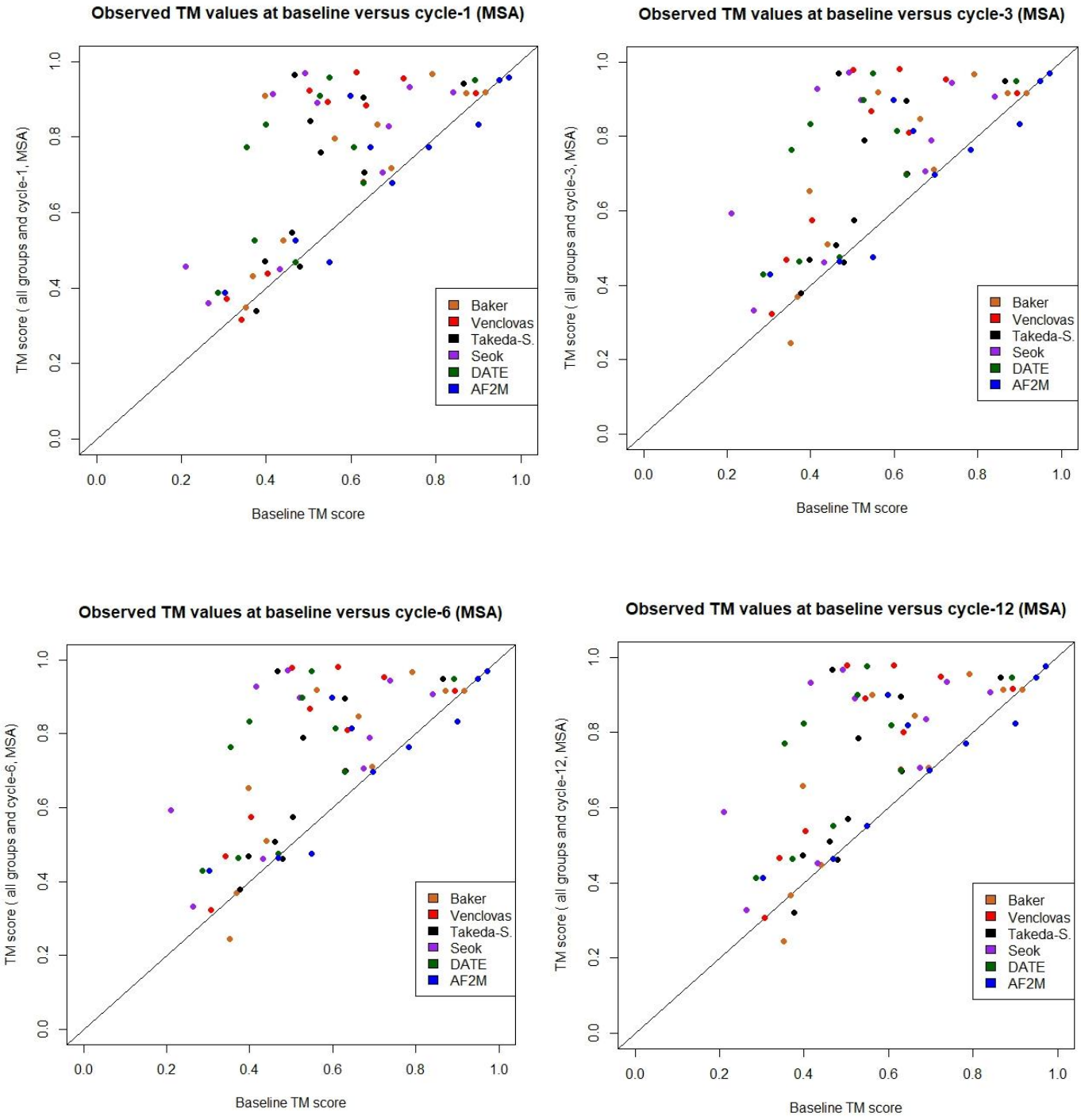


Figure S.3 A comparison of the observed and baseline TM-scores for the CASP14 models during each recycles (1-3-6-12) in the MSA method.

Four scatter plots representing comparisons of the observed TM-scores for the improved models of six groups (y-axis) versus the baseline TM-scores (x-axis) for the CASP14 models generated during recycles 1-3-6-12, separately, using the MSA method. Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda-Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. The scatter plots were drawn using R.

Appendix 5

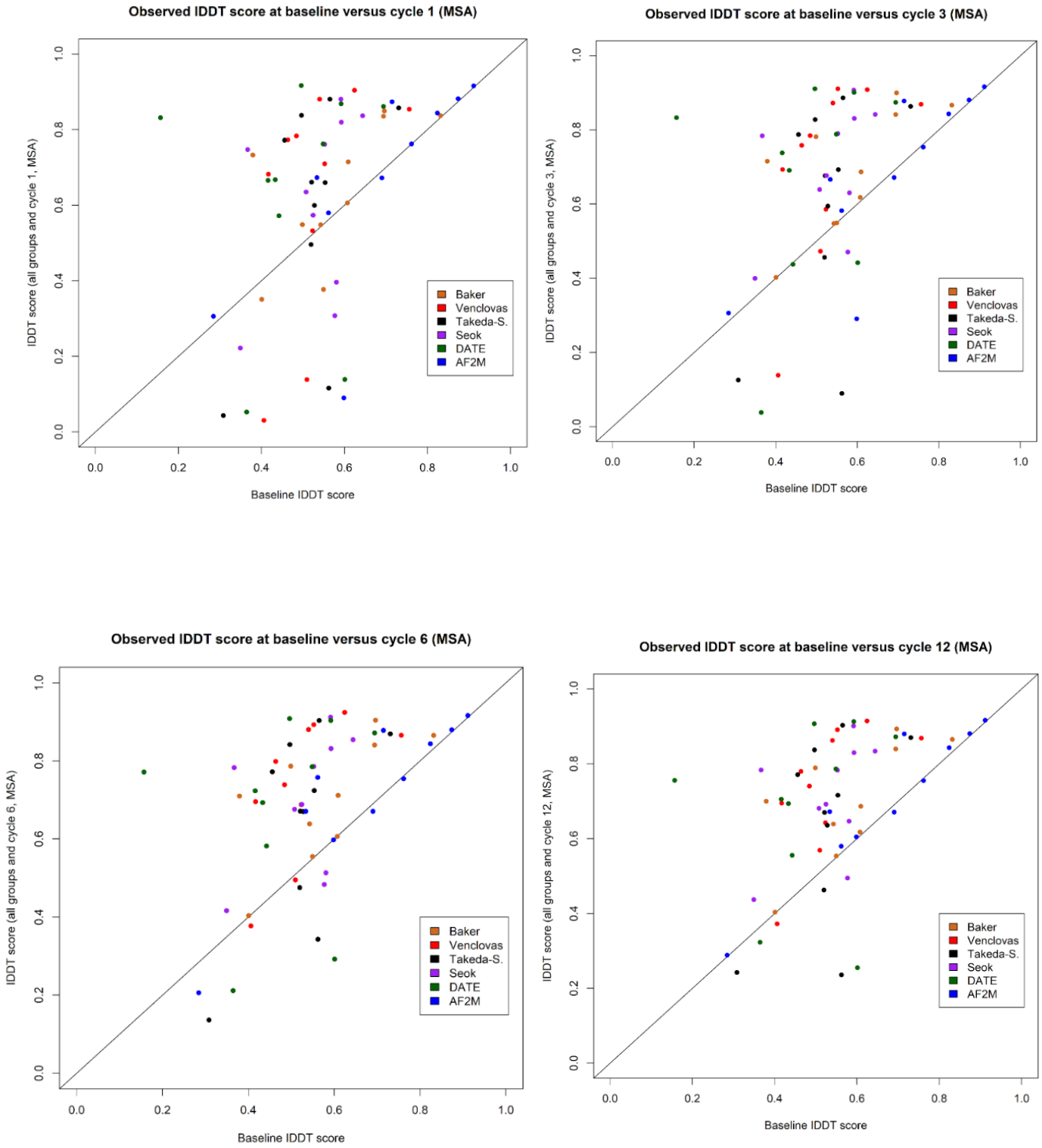


Figure S.4 A comparison of the observed and baseline IDDT scores for the CASP14 models during each recycles (1-3-6-12) in the MSA method.

Four scatter plots representing comparisons of the observed IDDT scores for the improved models of six groups (y-axis) versus the baseline IDDT scores (x-axis) for the CASP14 models generated during recycles 1-3-6-12, separately, using the MSA method. Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda-Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. The scatter plots were drawn using R.

Appendix 6

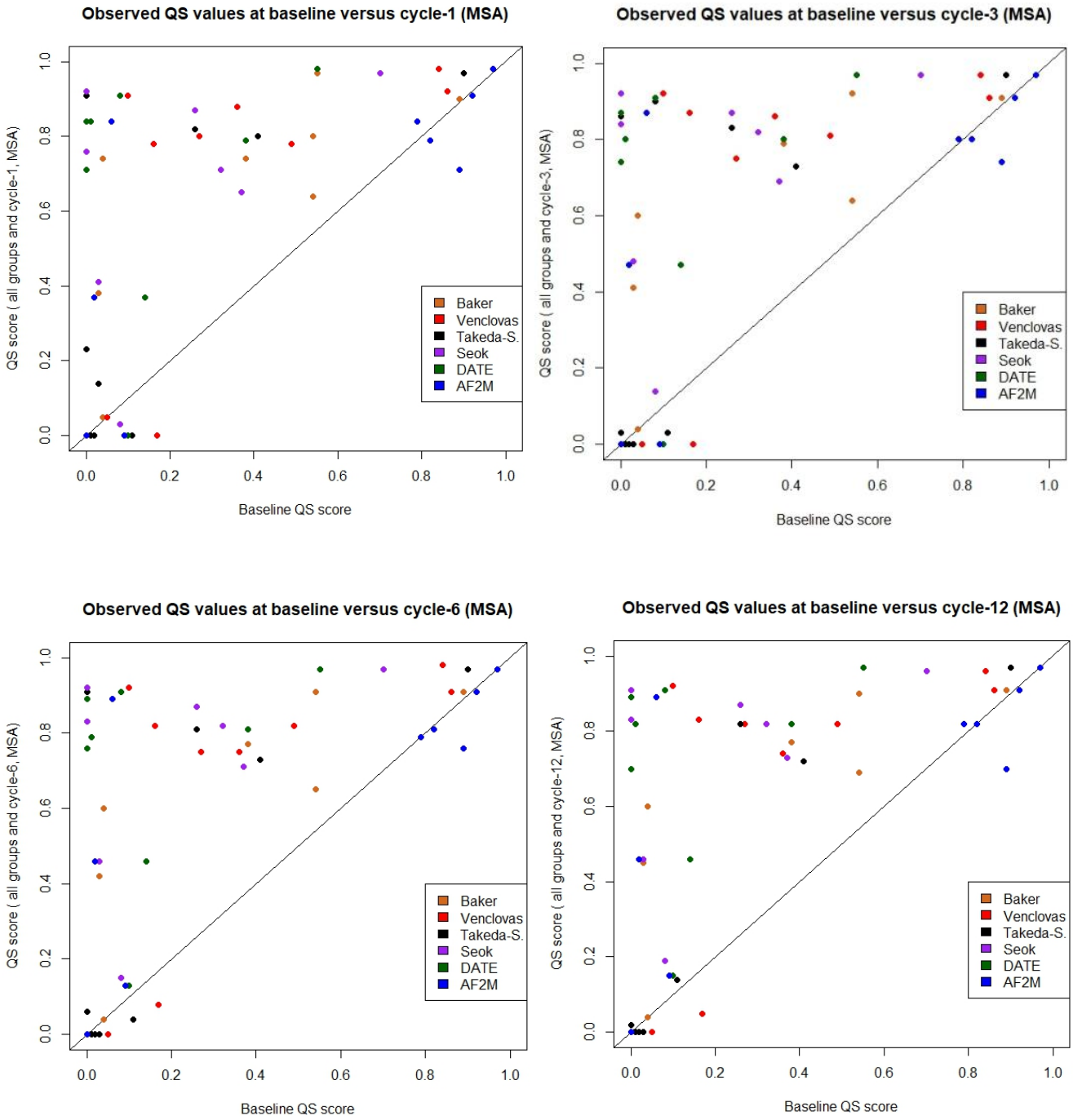


Figure S.5 A comparison of the observed and baseline QS-scores for the CASP14 models during each recycles (1-3-6-12) in the MSA method.

Four scatter plots representing comparisons of the observed QS-scores for the improved models of six groups (y-axis) versus the baseline QS-scores (x-axis) for the CASP14 models generated during recycles 1-3-6-12, separately, using the MSA method. Each colour corresponds to different group models, with orange representing Baker, red representing Venclovas, black representing Takeda-Shitaka, purple representing Seok, green representing DATE, and blue representing AF2M. The scatter plots were drawn using R.

Appendix 7

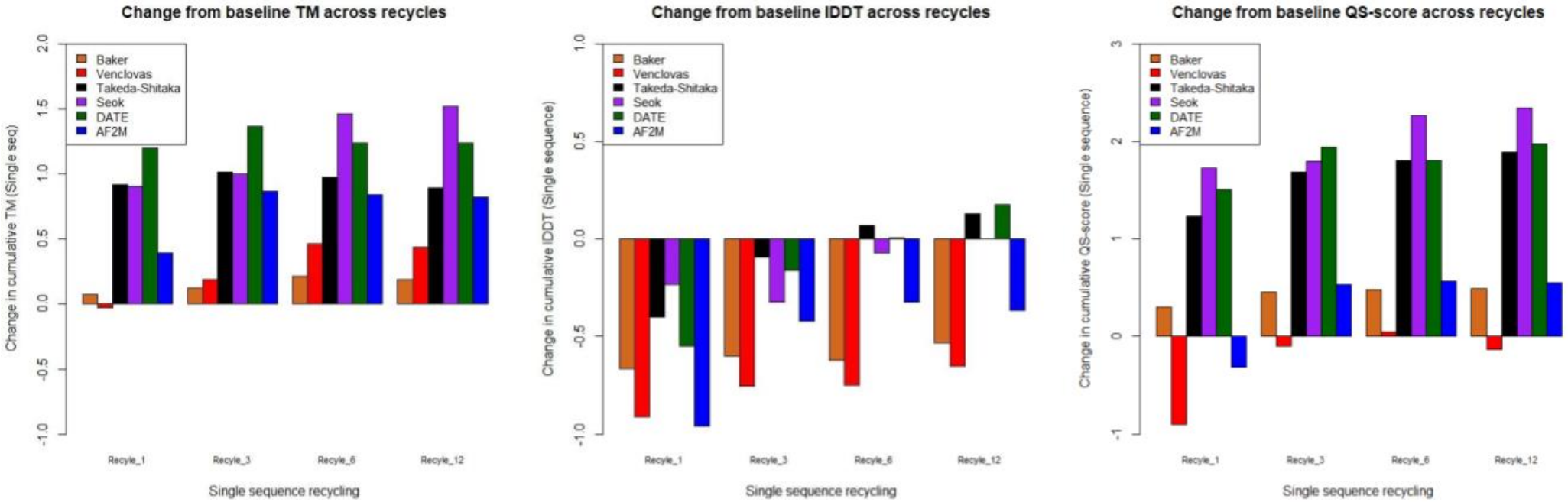


Figure S.6 A comparison of the observed and baseline three quality scores for the CASP14 models after recycling in the SS method.

Bar charts representing the cumulative change in observed A) TM-score (left), B) lDDT (middle), C) QS-score (right) generated from alignment between the baseline models and the CASP14 models generated by AF2M using the SS method after recycling (1-3-6-12). Each colour-coded bar corresponds to a distinct group. The bar charts were drawn using R.

Appendix 8

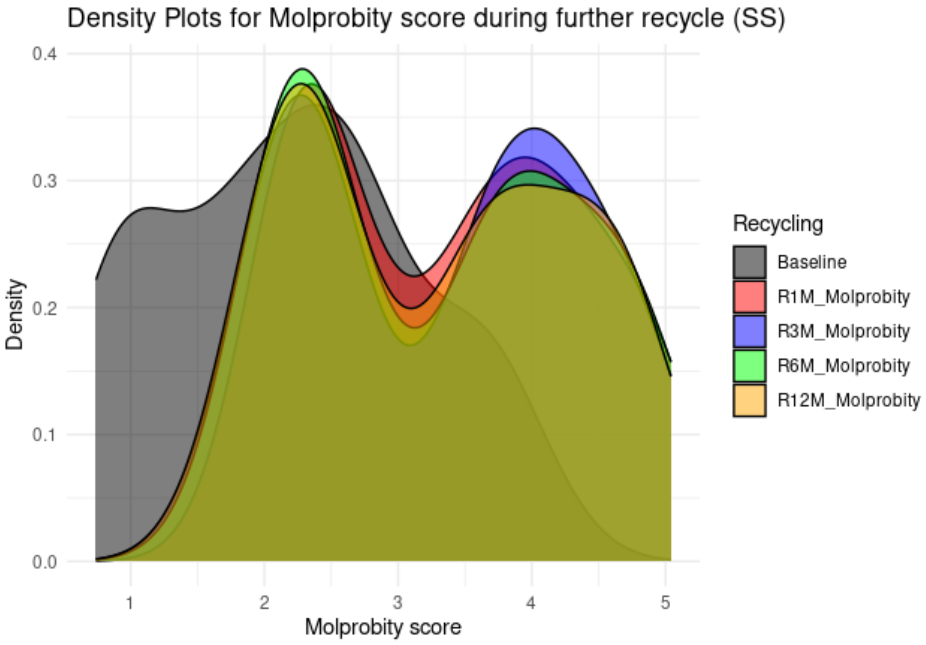


Figure S.7 A comparison of the observed and baseline Molprobity scores for the CASP14 models after recycling in the SS method.
The density plots showing Molprobity scores (lower Molprobity scores are better) for the CASP14 models generated by AF2M using SS, with red for cycle 1 (R1M), blue colour for cycles 3 (R3M), green colour for cycles 6 (R6M), magenta colour for cycles 12 (R12M) and black colour for baseline as starting model. These plot compares the geometric correctness rate for models after cycles, without using experimentally observed protein structure. The Molprobity scores were generated by http://molprobity.biochem.duke.edu/. The density plots were drawn using R.

Appendix 9

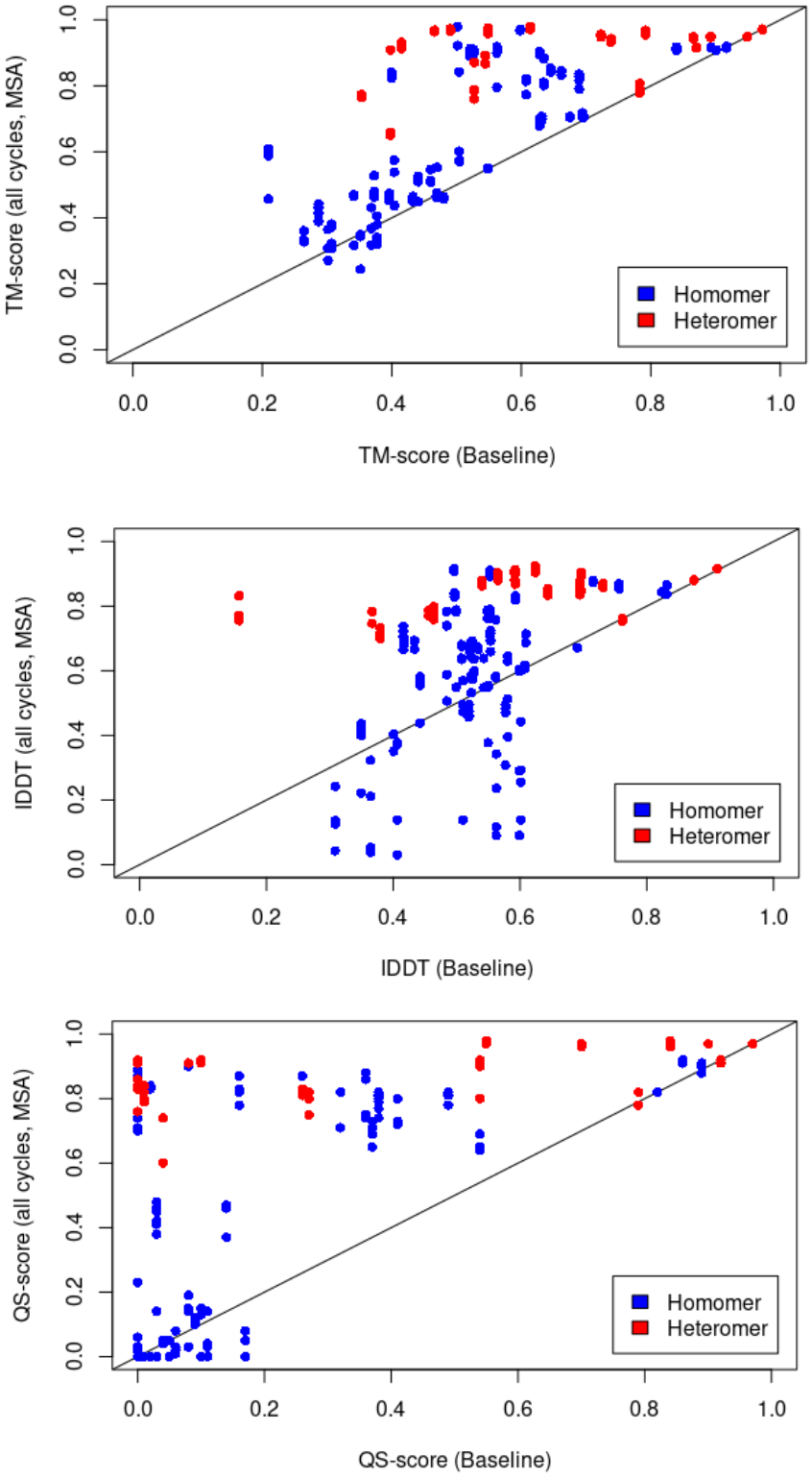


Figure S.8 A comparison of the observed TM-scores, IDDT score, and QS-scores with the baseline in the MSA method, in terms of types of the CASP14 protein targets.

Three scatter plots showing the improvement of homomeric and heteromeric models following further recycling. The plot for TM-scores (top), plot for IDDT (middle), and plot for QS-scores (bottom) compare the observed scores for the improved models (y-axis) versus the baseline scores (x-axis) for the CASP14 models generated by AF2M using the MSAs during all recycles (1-3-6-12) for six group models. The minimum values for TM-score, IDDT, and QS-scores are 0, while the maximum values are 1. The homomeric models highlight in blue, the heterometric models highlight in red. The scatter plots were drawn using R.

Appendix 10

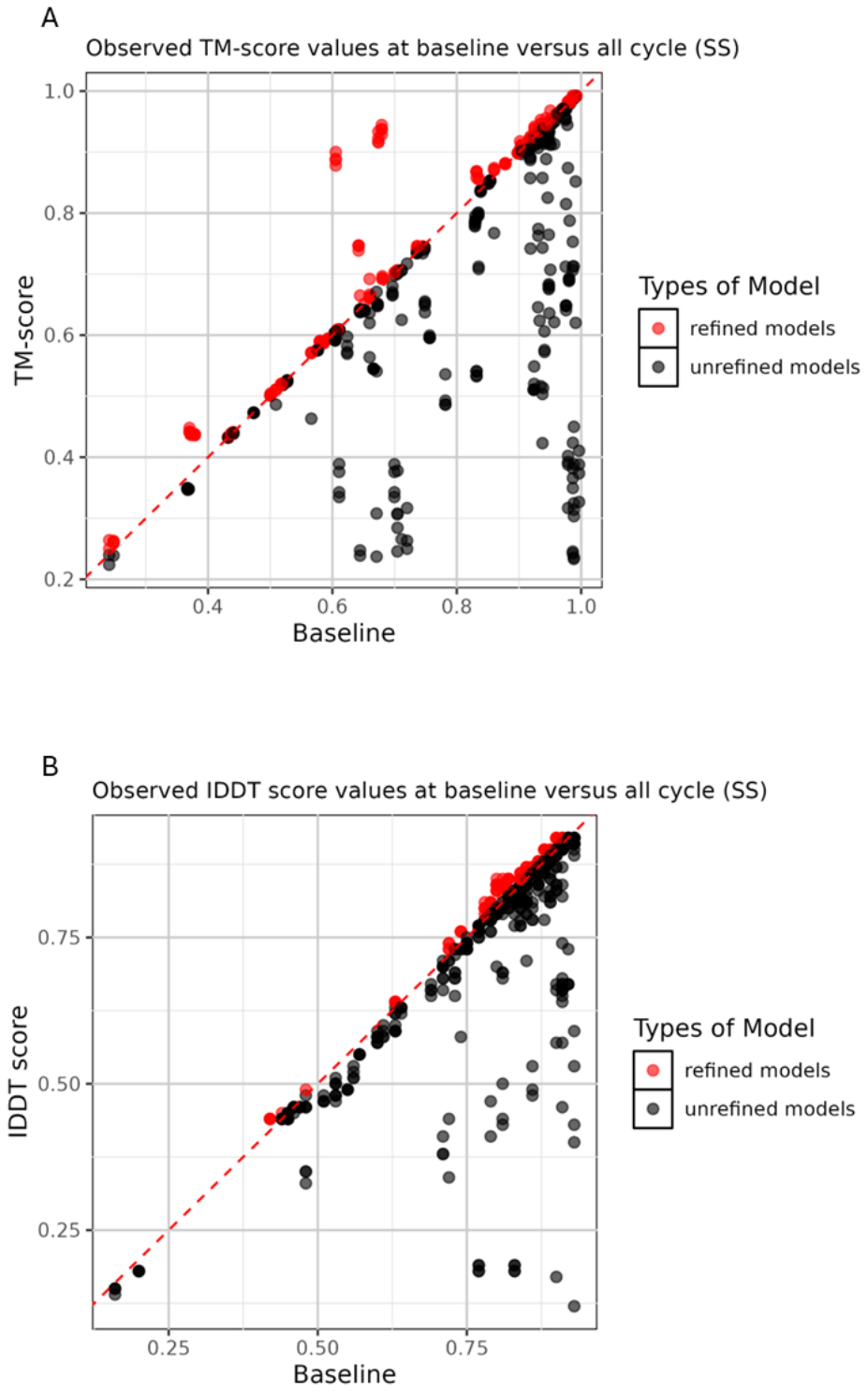


Figure S.9 A comparison of the observed TM-scores and IDDT scores for the CASP15 models with the baseline in the SS method.

Scatter plot representing the comparison of the (A) observed TM-scores (y-axis) and (B) observed IDDT scores (y-axis) with the baseline (x-axis) and models generated during all recycle (1-3-6-12) for six group models in the SS method. The minimum values for TM-scores and IDDT scores are 0, while the maximum values are 1. The red circles represent the refined models, while the black ones represent the unrefined models. This scatter plot was drawn using R

Appendix 11

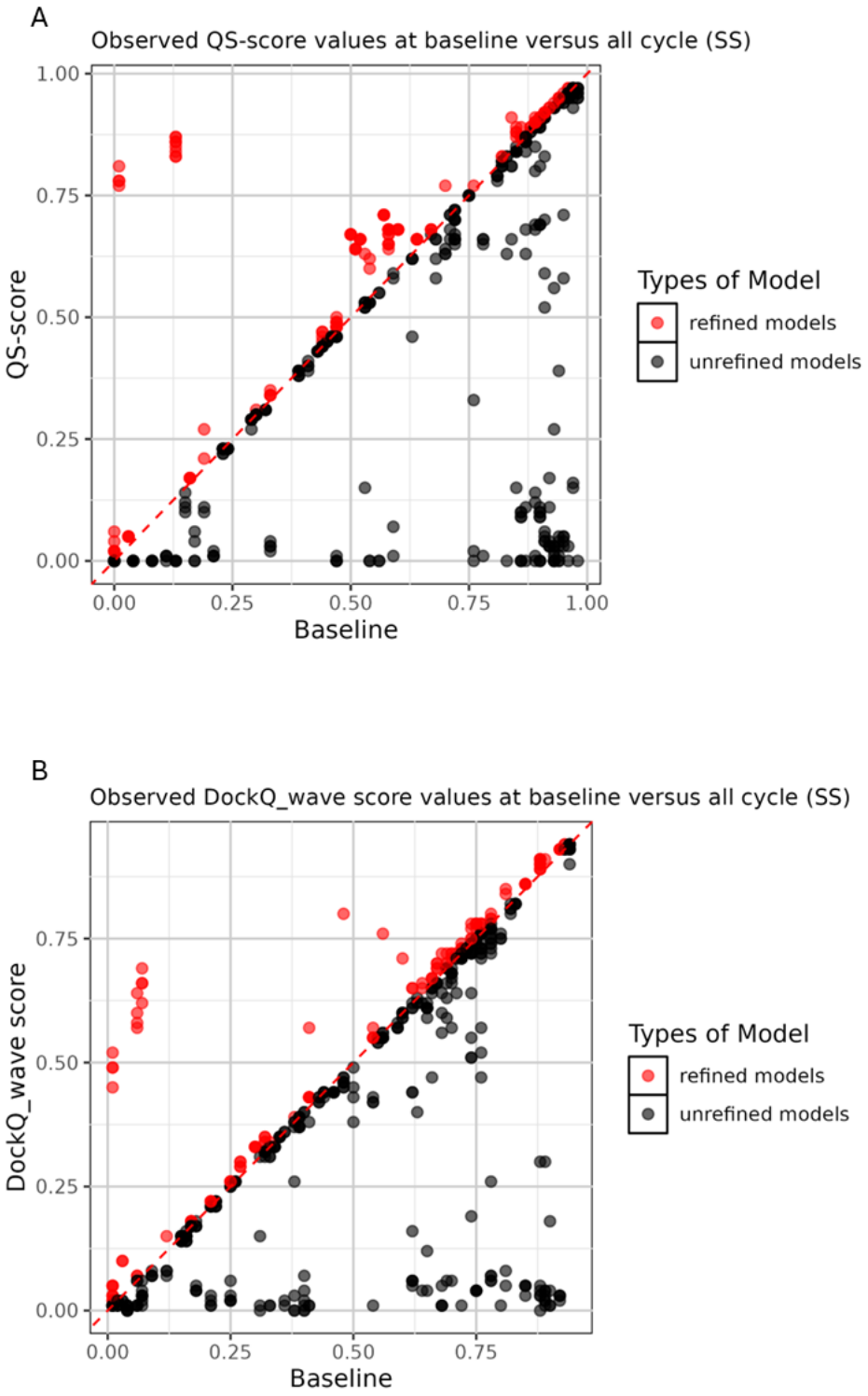


Figure S.10 A comparison of the observed QS-scores and DockQ_wave scores for the CASP15 models with the baseline in the SS method.

Scatter plot representing the comparison of the (A) observed QS-scores (y-axis) and (B) observed DockQ_wave scores (y-axis) with the baseline (x-axis) and models generated during all recycle (1-3-6-12) for six group models in the SS method. The minimum values for QS-scores and DockQ_waves are 0, while the maximum values are 1. The red circles represent the refined models, while the black ones represent the unrefined models. This scatter plot was drawn using R.

Appendix 12

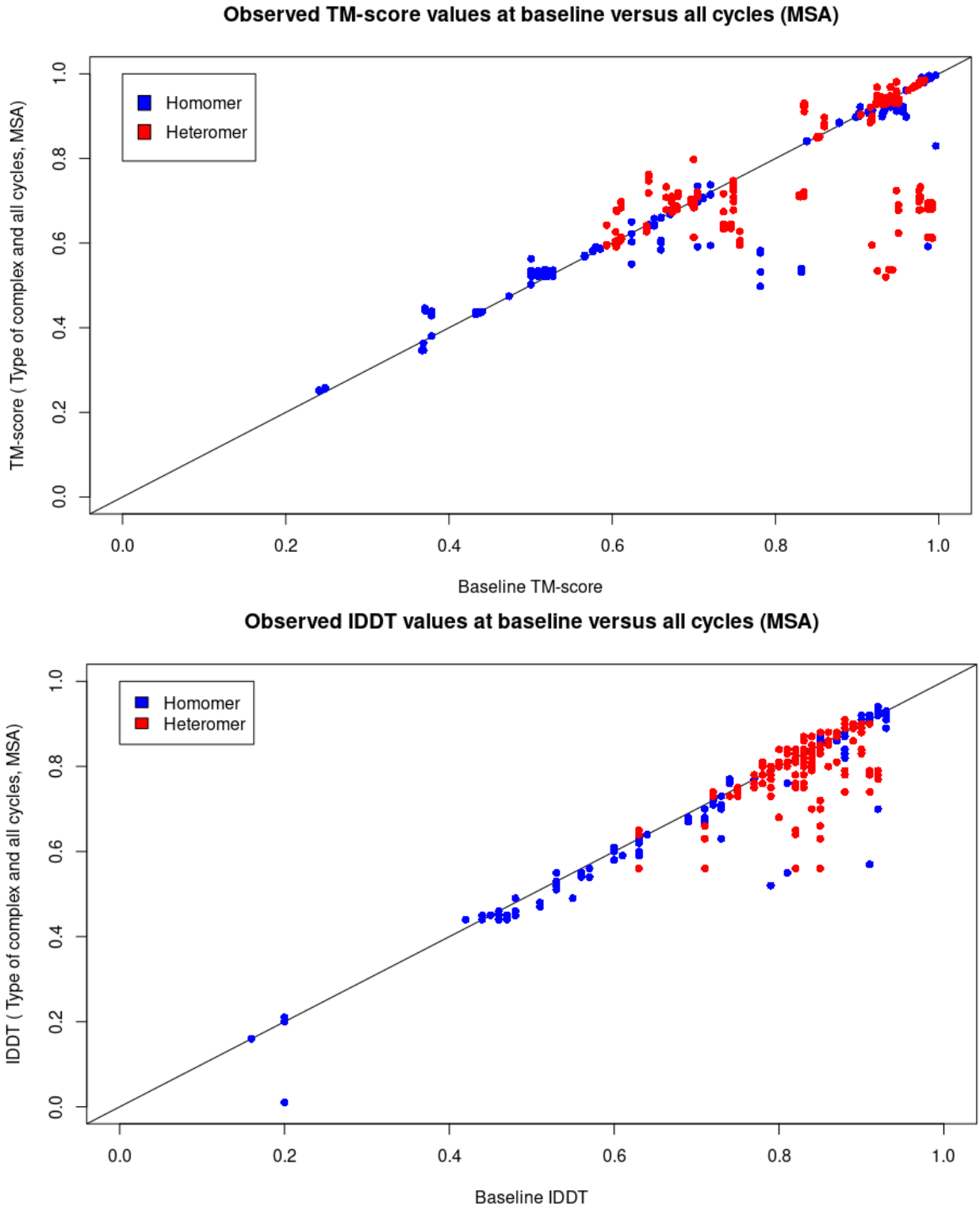


Figure S.11 A comparison of the observed TM-scores and IDDT scores with the baseline in the MSA method, in terms of types of the CASP15 protein targets.

Two scatter plots showing the improvement of homomeric and heteromeric models following recycling. The plot for TM-scores (top) and plot for IDDT (bottom) compare the observed scores for the improved models (y-axis) versus the baseline scores (x-axis) for the CASP15 models generated by AF2M using the MSAs during all recycles (1-3-6-12) for six group models. The minimum values for both TM-scores and IDDT scores are 0, while the maximum values are 1. The homomeric models highlight in blue, the heterometric models highlight in red. The scatter plots were drawn using R.

Appendix 13

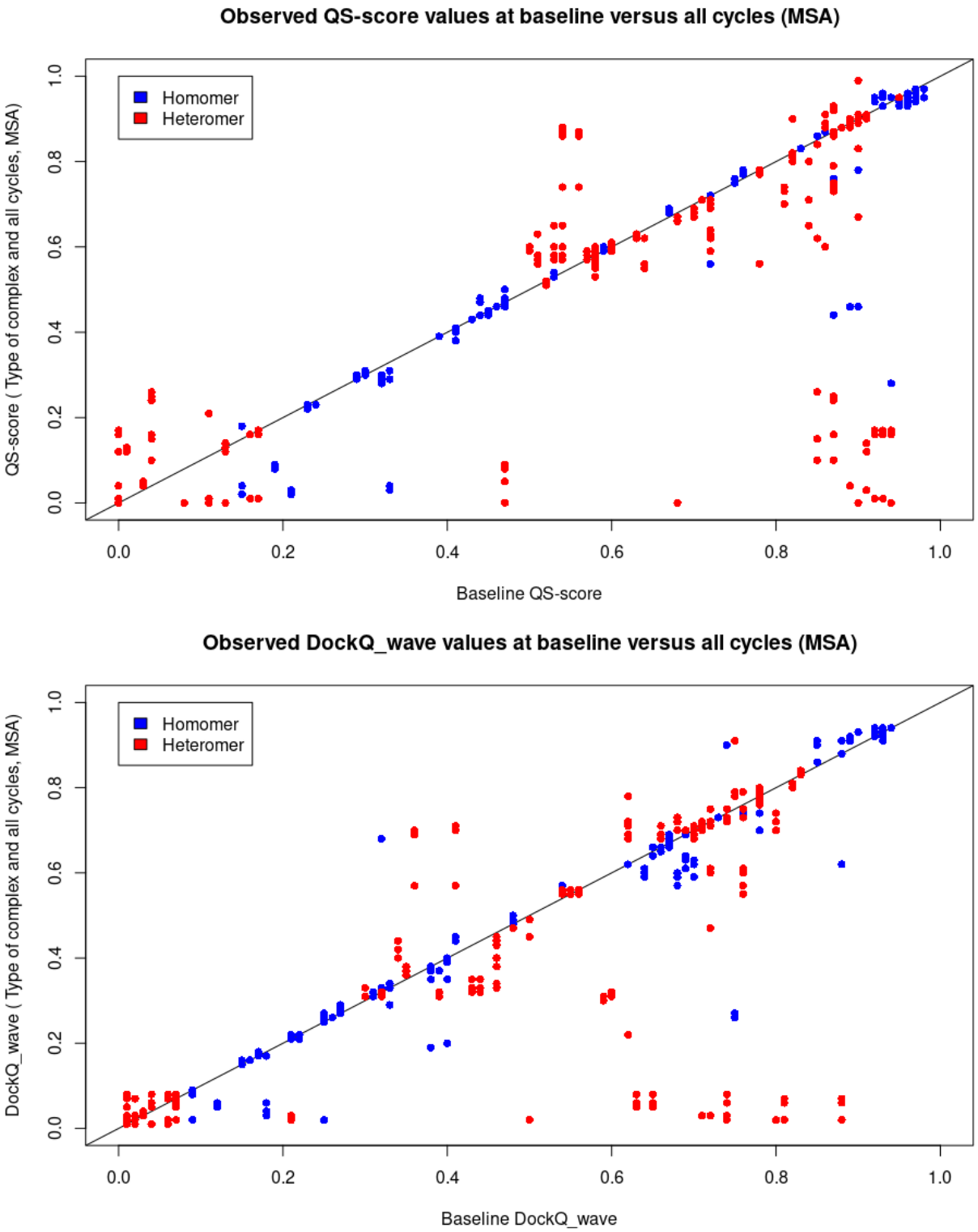


Figure S.12 A comparison of the observed QS-scores and DockQ_wave scores with the baseline in the MSA method, in terms of types of the CASP15 protein targets.

Two scatter plots showing the improvement of homomeric and heteromeric models following recycling. The plot for QS-scores (top) and plot for DockQ_wave scores (bottom) compare the observed scores for the improved models (y-axis) versus the baseline scores (x-axis) for CASP15 models generated by AF2M using the MSAs during all recycles (1-3-6-12) for six group models. The minimum values for both QS-scores and DockQ_wave scores are 0, while the maximum values are 1. The homomeric models highlight in blue, the heterometric models highlight in red. The scatter plots were drawn using R.

Appendix 14

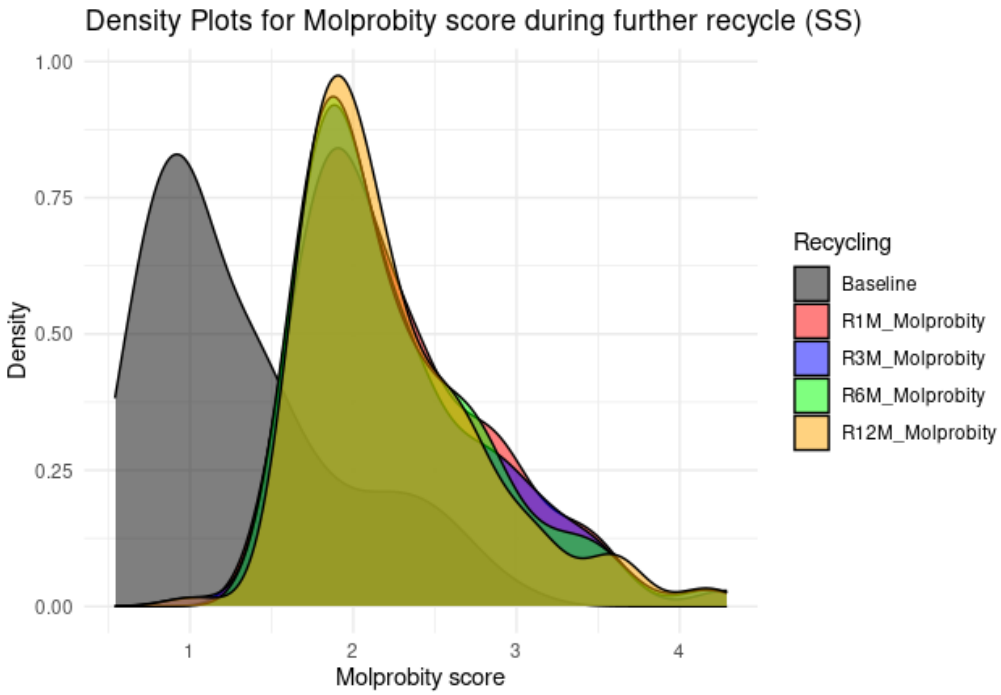


Figure S.13 A comparison of the observed and baseline Molprobity scores for the CASP15 models after recycling in the SS method.

The density plots showing Molprobity scores (lower Molprobity scores are better) for the CASP15 models generated by AF2M using SS, with red for cycle 1 (R1M), blue colour for cycles 3 (R3M), green colour for cycles 6 (R6M), magenta colour for cycles 12 (R12M) and black colour for baseline as starting model. These plot compares the geometric correctness rate for models after cycles, without using experimentally observed protein structure. The Molprobity scores were generated by http://molprobity.biochem.duke.edu/. The density plots were drawn using R.

Appendix 15

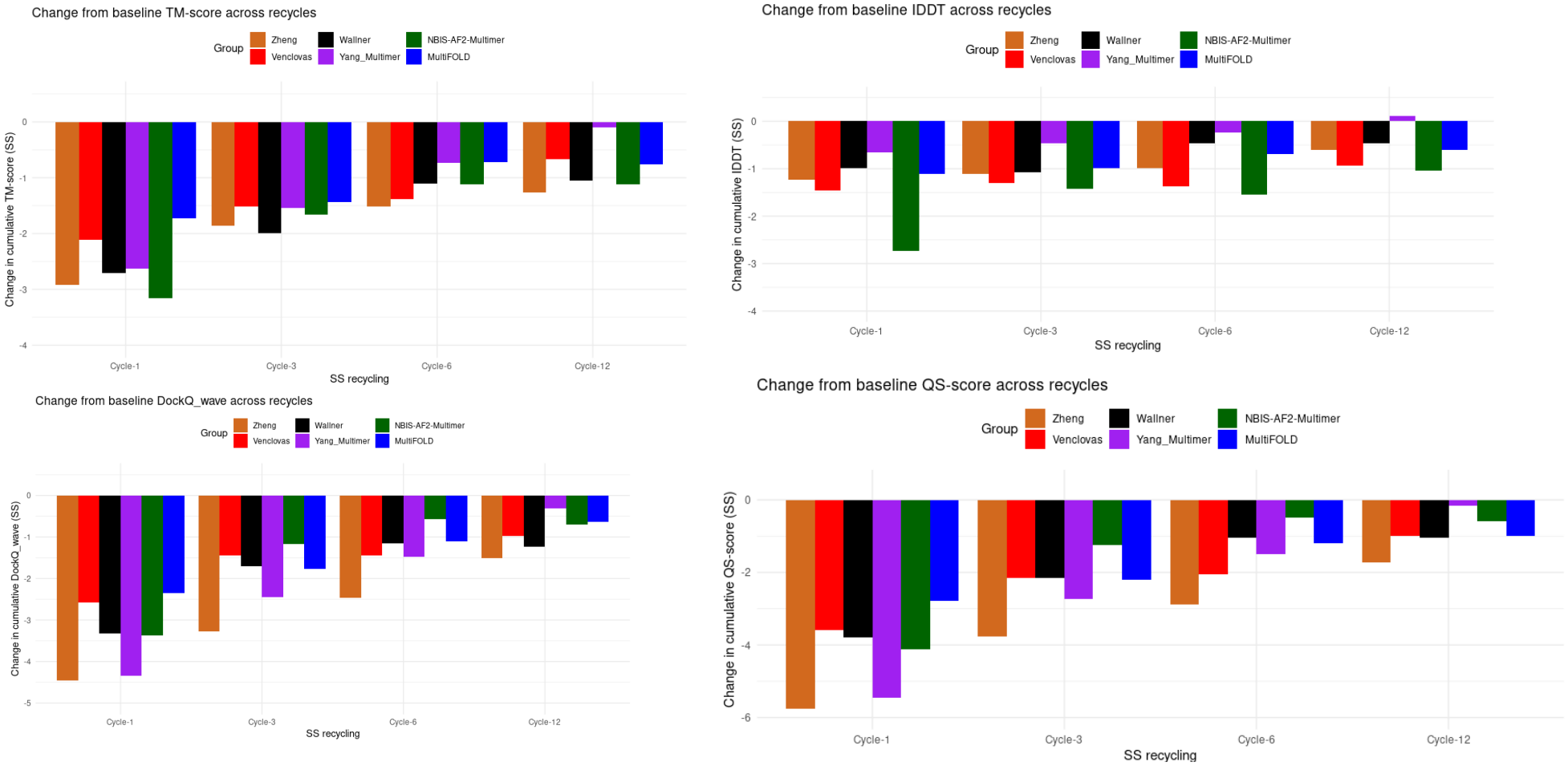


Figure S.14 A comparison of the observed and models baseline four quality scores for the CASP15 after recycling in the SS method.

Bar charts representing the cumulative changes in the observed TM-score (Top-left), IDDT (Top-right), DockQ_wave (Bottom-left), QS-score (Bottom-right) generated from alignment between the baseline models and the CASP15 models generated by AF2M using the SS method after recycling (1-3-6-12). Each colour-coded bar corresponds to a distinct group, with orange representing Zheng, red representing Venclovas, black representing Wallner, purple representing Yang-Multimer, green representing NBIS-AF2-Multimer, and blue representing MultiFOLD. The bar charts were drawn using R.

Appendix 16

Table S.2 The python script for preprocessing for the disorder filtering MSAs.

The script provides to ignore the specific types of disordered residues within homologous sequences and designs MSAs as an input for AF2M.

```
import requests
from Bio import SeqIO

# Specify the name of the FASTA format file
fasta_file = "input.fasta"

# Open the FASTA file and retrieve the sequence names
sequences = SeqIO.to_dict(SeqIO.parse(fasta_file, "fasta"))

# Open a file to write the results
with open("result1.txt", "w") as output_file:
    # For each sequence name, search the UniRef100 database and write the sequence to the file
    for seq_name in sequences:
        output_file.write("Sequence Name: " + seq_name + "\n")

        # Create a URL for searching using the UniProt REST API
        url = f"https://www.uniprot.org/uniref/{seq_name}.fasta"

        # Send a GET request and receive the response
        response = requests.get(url)

        # Check the response
        if response.status_code == 200:
            # Retrieve the text of the response
            fasta_text = response.text

            # Write the FASTA format sequence to the file
            output_file.write("Sequence:\n")
            output_file.write(fasta_text + "\n")
            output_file.write("--------------\n")
        else:
            output_file.write("An error occurred during the search.\n")

# Read the output file and extract the desired sections into a new file
inside_desired_section = False

with open("result1.txt", "r") as input_file, open("result2.txt", "w") as output_95_file:
    for line in input_file:
        if line.startswith(">"):
            inside_desired_section = True

        if inside_desired_section:
            output_95_file.write(line)

        if line.startswith("-------"):
            if inside_desired_section:
                inside_desired_section = False
                output_95_file.write(line)
```

```
with open('result2.txt', 'r') as input_file:
    lines = input_file.readlines()

filtered_lines = [line for line in lines if line.count('-') < 6]

with open('result3.txt', 'w') as output_file:
    output_file.writelines(filtered_lines)

with open('result3.txt', 'r') as input_file:
    lines = input_file.readlines()

filtered_lines = [line.strip() for line in lines if line.strip() != ""]

with open('result4.txt', 'w') as output_file:
    output_file.writelines('\n'.join(filtered_lines))


#### After the result4.txt, a code for IUPRED3 is run as shown in Appendix Table 1.3 result5.txt is
obtained


def merge(file_path):
    combined_line = ''
    with open(file_path, 'r') as file, open('result6.txt', 'w') as result_file:
        for line in file:
            if '>' in line:
                if combined_line:
                    result_file.write(combined_line.rstrip(',') + '\n')  # Removing the last comma
                    combined_line = ''
                result_file.write(line)
            else:
                combined_line += line.strip() + ', '  # Adding a comma at the end of each line

        if combined_line:
            result_file.write(combined_line.rstrip(',') + '\n')  # Removing the last comma

file_path = 'result5.txt'
merge(file_path)


def merge_sequences(file1, file2, output_file):
    seq1 = {}
    seq2 = {}

    # Process the lines of file 1
    with open(file1, 'r') as f1:
        lines1 = f1.readlines()
        i = 0
        while i < len(lines1):
            line = lines1[i].strip()
            if line.startswith('>'):
                seq_id = line[1:].split()[0]  # Skip the '>' character and get the sequence ID
                i += 1
                seq = ''
                while i < len(lines1) and not lines1[i].startswith('>'):
                    seq += lines1[i].strip()
                    i += 1
                seq1[seq_id] = seq

    # Process the lines of file 2
```

```
    with open(file2, 'r') as f2:
        lines2 = f2.readlines()
        i = 0
        while i < len(lines2):
            line = lines2[i].strip()
            if line.startswith('>'):
                seq_id = line[1:].split()[0]  # Skip the '>' character and get the sequence ID
                i += 1
                seq = ''
                while i < len(lines2) and not lines2[i].startswith('>'):
                    seq += lines2[i].strip()
                    i += 1
                seq2[seq_id] = seq

    # Merge common sequences
    merged_seqs = []
    for seq_id in seq1:
        if seq_id in seq2:
            merged_seq = f">{seq_id}\n{seq1[seq_id]}\n{seq2[seq_id]}\n"
            merged_seqs.append(merged_seq)

    # Write the result to the output file
    with open(output_file, 'w') as result_file:
        result_file.writelines(merged_seqs)

    print(f"Result file created as '{output_file}'.")


# Change file paths here or use the file names directly
file1 = 'result6.txt'
file2 = 'input.fasta'
output_file = 'result7.txt'

merge_sequences(file1, file2, output_file)

def main():
    try:
        with open("result7.txt", "r") as file:
            lines = file.readlines()

        new_lines = []
        prev_lines = []

        for i in range(2, len(lines), 3):
            prev_line = lines[i - 2].strip()
            prev_lines.append(prev_line)  # Store the value of prev_line
            current_line = lines[i - 1].strip().split(", ")
            next_line = lines[i].strip()

            extra_line = ""
            for character in next_line:
                score_char = current_line[0]
                char_and_score = score_char.strip("[]").split(",")
                y = float(char_and_score[1].strip("[]"))
                x = char_and_score[0]
                if character == x:
                    if y >= 0.50:                    #### It is the line of code that gives the ratio corresponding
to the disorder residues and shows that values above this ratio will be ignored
                        extra_line += "-"
                    else:
```

```
                    extra_line += character
                else:
                    extra_line += character
            new_lines.append(extra_line)

        with open("Filtered_MSA.txt", "w") as output_file:
            for prev_line, extra_line in zip(prev_lines, new_lines):
                output_file.write(prev_line + "\n")
                output_file.write(extra_line + "\n")

    except FileNotFoundError:
        print("File not found.")

if __name__ == "__main__":
    main()
```

Appendix 17

Table S.3 The script for application of IUPRED3 for residue detection and filtering in the MSAs.

The script was used to run IUPred3 in the Ubuntu terminal, enabling the measurement of the potential disorder rate for each residue within every homologous sequence in the MSA. Residues scoring 0.5 or higher are deemed disordered and consequently deleted from the MSA in order to design a low quality MSAs.

```
#!/bin/bash

input_file="result4.txt"
output_folder="output_files"
total_output="result5.txt"

# Create the output folder
mkdir -p "$output_folder"

# Read the FASTA file line by line and process it
while IFS= read -r line
do
  if [[ $line == ">"* ]]; then
    # Header line
    seq_name="${line#*>}"  # Get all characters after the ">" symbol
    output_file="${output_folder}/$(echo $seq_name | awk '{print $1}').txt"
    seq=""
  else
    # Sequence line
    seq="${seq}${line}"
    echo "$seq" > temp_seq.txt
    python3 iupred3.py temp_seq.txt long > "$output_file"
    rm temp_seq.txt
  fi
done < "$input_file"

# Merge all output files into one
cat "${output_folder}"/*.txt > "$total_output"
```

In the script, (…….) indicates an optional part that includes ‘long’ for long disordered residues, ‘short’ for short disordered residues, ‘glob’ for domain disordered residues.

Appendix 18

Table S.4 Comparison of model quality for AF2M models generated using "single-chain" and standard custom templates.

The table summarises the number of models generated by AF2M using both "single-chain" custom template and standard custom template compared to the initial models' numbers. The comparison is presented in terms of decreasing, increasing and unchanged numbers. In terms of TM-score and IDDT score, the increasing number of models is greater than the decreasing number of models when comparing the "single-chain" custom templates to the initial models. However, when using the standard custom template methods, the decreasing number of models is more than the increasing number of models in terms of all four quality metrics.

Comparing "single-chain" custom template with initial models
The number of the models

Scores	Increasing	Decreasing	Unchanged
TM-score	63	54	3
IDDT	65	52	3
QS-score	48	51	21
DockQ_wave	54	57	9

Comparing standard custom template with initial models
The number of the models

Scores	Increasing	Decreasing	Unchanged
TM-score	38	77	5
IDDT	36	75	9
QS-score	32	57	31
DockQ_wave	56	17	47

Appendix 19

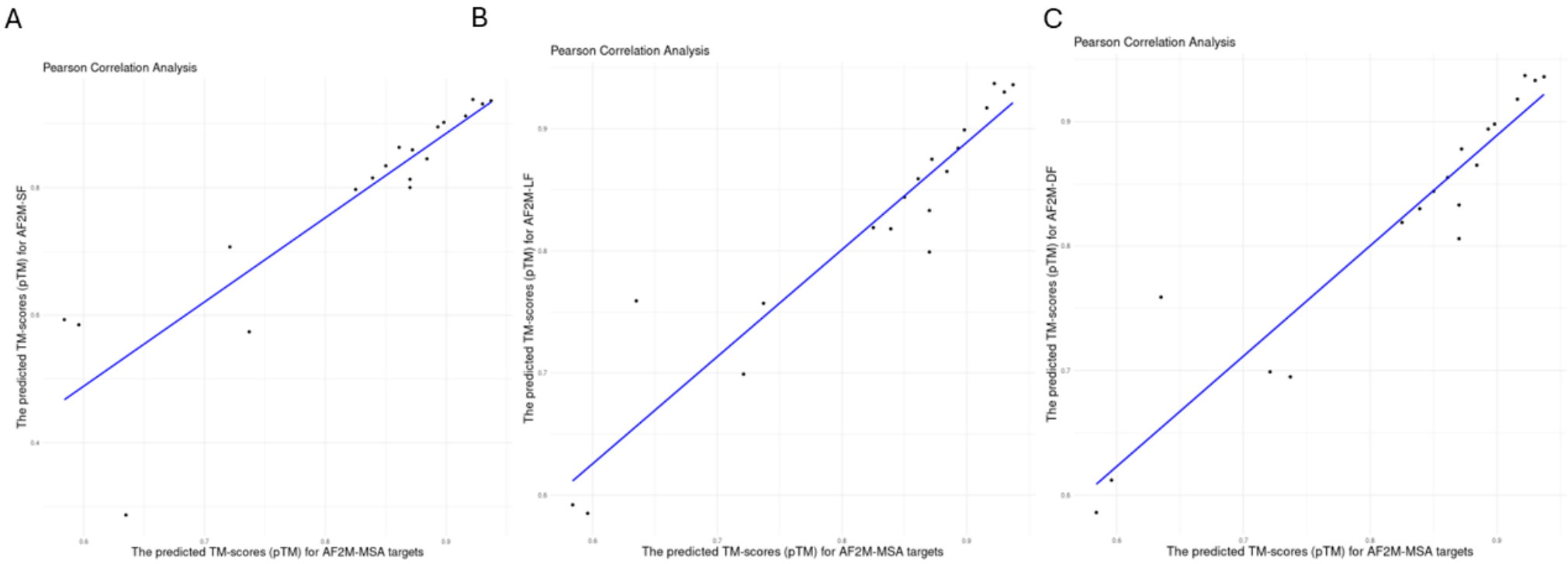


Figure S.15 The correlation between the pTM scores for three filtered MSA methods and the pTM scores for standard MSA methods.
Scatter plots showing linear, positive relationship between the pTM-scores generated by AF2M with standard MSA method versus the pTM scores of the models of CASP14-15 targets generated using A) AF2M-SF B) AF2M-LF C) AF2M-DF with n = 19 multimer targets. These above scatter plots belong to the Pearson's R correlation test as an example. The Pearson's R correlation is 0.88, Kendall's tau B correlation is 0.79 and Spearman's Rho correlation is 0.93 for AF2M-SF, the Pearson's R correlation is 0.94, Kendall's tau B correlation is 0.95 and Spearman's Rho correlation is 0.95 for AF2M-LF, the Pearson's R correlation is 0.94, Kendall's tau B correlation is 0.85 and Spearman's Rho correlation is 0.95 for AF2M-DF. This plot was drawn via R.

Appendix 20

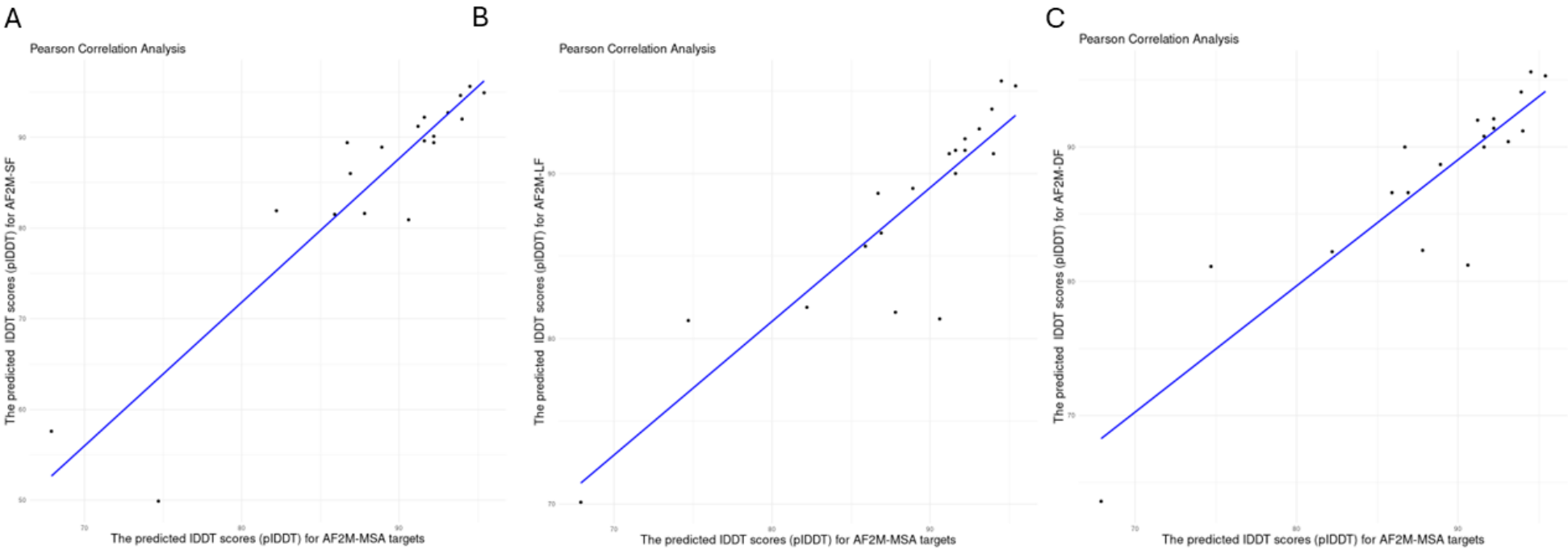


Figure S.16 The correlation between the pIDDT scores for three filtered MSA methods and the pIDDT scores for standard MSA methods. Scatter plots showing linear, positive relationship between the pIDDT scores generated by AF2M with standard MSA method versus the pIDDT scores of the models of CASP14-15 targets generated using A) AF2M-SF B) AF2M-LF C) AF2M-DF with n = 19 multimer targets. These above scatter plots belong to the Pearson's R correlation test as an example. The Pearson's R correlation is 0.92, Kendall's tau B correlation is 0.71 and Spearman's Rho correlation is 0.89 for AF2M-SF, the Pearson's R correlation is 0.89, Kendall's tau B correlation is 0.77 and Spearman's Rho correlation is 0.90 for AF2M-LF, the Pearson's R correlation is 0.89, Kendall's tau B correlation is 0.71 and Spearman's Rho correlation is 0.87 for AF2M-DF. This plot was drawn via R.

Appendix 21

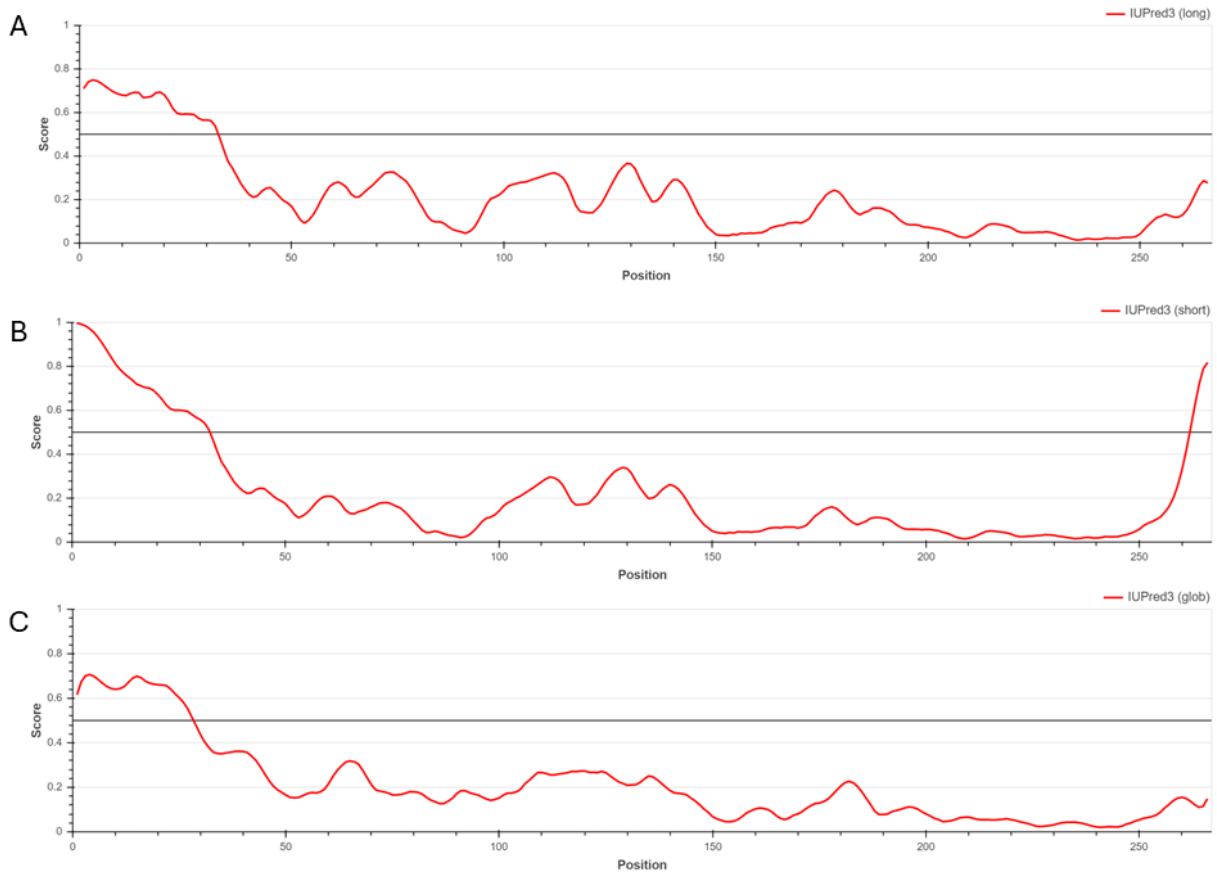


Figure S.17 The disorder/order residues within amino acid positions for the T1123 CASP15 target.
The three figures show the long (A), short (B) and glob (C) disordered residues within monomeric structure for the T1123 target, respectively. A cut off value of 0.5 and above represents the likelihood of a residue being disorder. The figures were obtained through IUPred3.

Glossary

Artificial Intelligence (AI): AI describes the replication of human intelligence in machines that are programmed to understand, learn, and solve issues.

CAMEO: It is a project that aims to provide easy access to weekly performance information for the current modelling methods.

CASP: It is a project held every two years to drive cutting-edge research and the development of new technologies for protein structure modelling and evaluate progress in accuracy, which is seen as the international gold standard.

CD-HIT software system: Cluster Database at High Identity with Tolerance (CD-HIT) is a tool intended for clustering and analyzing biological sequence datasets.

Consensus Method: Consensus methods in model evaluation calculate an average similarity score among models, with the presumption that better models exhibit greater similarity with others in the pool.

CPU: Central Processing Unit is the main component that performs the majority of the operations by executing commands from programs through fundamental arithmetic, logic, control, and input/output operations.

Deep Learning (DL): DL is a distinct subset of ML that leverages multi-layered ANN to process complex data and extract features implicitly.

Diffusion Model: A model is a probabilistic ML model that is trained to produce data by repeatedly denoising a noisy sample, effectively reversing the noise addition method.

Disorder-to-order structure: IDRs or IDPs evolving from a flexible, unstructured state to a more ordered and organized conformation—often due to interaction with a binding partner or specific environmental changes—is referred to as a disorder-to-order transition in protein structures.

Dropout: It is a regularization strategy that enhances generalization using random selection to deactivate a subset of neurons during training, reducing the network's reliance on particular neurons.

Embedded protein sequences: These types of proteins are numerical representations of protein sequences. They are converted into numerical vectors using protein language models (e.g. ProtT5, ESM, and ProtBERT). Such methods are used to better understand and analyse biological meaning and structural features in protein sequences.

Fine-tunning: Rather than training from scratch, fine-tuning in deep learning is the process of fine-tuning a pre-trained model for a particular task by training it for a few more epochs on a smaller, task-specific dataset.

Force Field: A force field for protein refinement is a quantitative model used to compute atomic interactions within a protein in order to stabilize and optimize its structure.

GPU: Graphics Processing Unit is a customized processor designed to accelerate the building and rendering of visuals, animations, and videos.

Homomeric proteins: They are proteins formed by the assembly of subunits of the same type, i.e. all subunits consist of the same polypeptide chain.

Heteromeric (non-homomeric) proteins: They are formed by the combination of different types of subunits and may have more complex structure and function.

Hidden Markow Model (HMM): A HMM is a statistical model employed to describe systems that transition between a limited number of hidden states, in which the state of system at a given time is not directly observable but can be deduced from observed data.

Invariant Point Attention (IPA): IPA is a mechanism in models dealing with data that exhibit symmetry or invariance, such as GNNs. It focuses on "invariant" or "stable" points to guarantee that the model performs consistently, regardless of the transformations applied to the input data.

Language Model (LM): LM is a statistical or NN-based model trained to predict the likelihood of a sequence of words in a language, playing a fundamental role in natural language processing (NLP) tasks.

Machine Learning (ML): ML is a category of AI that provides systems to learn patterns and make decisions from information within data without explicit programming.

Neff value: It refers to the ‘substantial number of sequences’ used in bioinformatics in the context of MSA and protein structure prediction, which estimates the sequence diversity or information content in an MSA.

PDBx/mmCIF: This file format and its corresponding data dictionary serve as the foundation for wwPDB data upload, annotation, and documentation. This template offers PDB data derived from all experimental methods.

Position-Specific Scoring Matrix (PSSM): For bioinformatics applications like protein domain analysis, sequence alignment or motif discovery, a PSSM is a matrix employed to represent sequence motifs or patterns, encoding the probability of different amino acids or nucleotides appearing at each position of a sequence in applications such as sequence alignment, protein domain analysis, or motif discovery.

Quasi-single method: The quasi-single method for protein model quality estimation describes techniques that, without the need for multiple models, assess the quality of a single model based on characteristics like structural geometry, energy-based evaluations, and consistency with the statistical or physical properties of proteins.

Root-mean-square deviation (RMSD): A value that expresses the difference between the the observed measurement and the value predicted by a model.

Single-chain template: It was a type of template created by converting multi-chain structures into single-chain forms using the PyMOL program

Standard custom template: It was external protein structures that are supplied straight to AF2 as input without any processing or changes.

Stoichiometry: The number and composition of the subunits in the assembly are described by stoichiometry. In the CASP competition, “A” is an abbreviation for one type of protein chain. For example, “AB” corresponds to two different chains, while A represents a first chain.