Improvements to methods for the quality estimation and refinement of protein quaternary structure models

A thesis submitted for the degree of
Doctor of Philosophy

School of Biological Sciences,
Faculty of Life Sciences
University of Reading

Nicholas S Edmunds

Supervisor: Prof. Liam J McGuffin

April 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 in the text.

Nicholas S Edmunds

Date: 19/04/24

Abstract

Computational protein modelling has increased in public profile following the success of AlphaFold2 at CASP14 in 2020. This led many to proclaim the protein folding problem essentially solved, meaning *in silico* methods could now fill the sequence-structure gap which had grown since the advent of next generation sequencing techniques.

However, proteins which prove problematic to experimental methods like X-ray crystallography and NMR are often multimeric in nature, like trans-membrane proteins or receptor binding interactions and, as the 2020 success was limited to tertiary structures, significant obstacles in quaternary structure elucidation remained. Contemporaneous analysis of assembly modelling showed that atomic contact prediction was a particular weakness and, as model refinement focusses on correcting small errors in atomic positioning, we proposed that a novel refinement method could be realised if full model coordinate files could be successfully submitted and recycled through the AF2 neural network. We present data in this thesis demonstrating that this is possible and that it significantly improved the quality of models including the official AF2 competition models from CASP14.

Model quality assessment programs for quaternary structures had been largely absent with modellers relying on various proprietary accuracy estimates and docking scores. ModFOLDdock was conceived to independently evaluate multimeric model quality from any modelling software. Here we show how ModFOLDdock was improved by neural network training using three conceptual target scores and regression analysis leading to a significant increase in predictive performance. Further optimisation of our three unique combinations of distance-based quality measures resulted in the definition of three ModFOLDdock variants, all of which were subsequently highly placed in the CASP15 EMA competition, ranking 2nd for global score, 1st for interface score and 2nd for interface residue score. Evidence is also presented showing that ModFOLDdock outperforms the AlphaFold2 quality measures plDDT and pTM at quality-ranking quaternary structure models.

Contents

List of Figures

List of Tables

List of Abbreviations

1-D	1 Dimensional
2-D	2 Dimensional
2-DE	2-D gel Electrophoresis
3-D	3 Dimensional
AI	Artificial Intelligence
AF2	AlphaFold2
AFM	AlphaFold2-Multimer
AMBER	Assisted Model Building with Energy Refinement
ANN	Artificial Neural Network
ASE	Accuracy self-estimation
AUC	Area Under the Curve
BLOSUM	Blocks Substitution Matrix
BFD	Big Friendly Database
CAD	Contact Area Difference
CAMEO	Continuous Automated Modelling Evaluation
CAPRI	Critical Assessment of Prediction of Interactions
CASP	Critical Assessment of Protein Structure Prediction
CDA	Contact Distance Agreement
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CM	Comparative Modelling
Cryo-EM	Cryogenic Electron Microscopy
DipDiff	The average between dipeptide unit (Dip) scores
DNN	Deep Neural Network
DNA	Deoxyribonucleic Acid
DSSP	Dictionary of Secondary Structure in Proteins
EMA	Estimation of Model Accuracy
EMBL	European Molecular Biology Laboratory
ET	Electron Tomography
FFT	Fast Fourier Transform
FM	Free Modelling
Fnat	Fraction of native interface contacts
FPR	False Positive Rate
GDP	Guanosine Diphosphate
GDT	Global Distance Test
GDT_TS	Global Distance Test – Total Score

GNN	Graph Neural Network
GPU	Graphics Processing Unit
GRC	GalaxyRefineComplex
GTP	Guanosine Triphosphate
HAART	Highly Active Anti-Retroviral Therapy
HADDOCK	High Ambiguity Driven protein-protein Docking
HER2	Human Epidermal Growth Factor Receptor 2
HIV	Human Immunodeficiency Virus
HM	Homology Modelling
ICS	Interface Contact Score
IPS	Interface Patch Score
iRMS	interface Root Mean Square (deviation)
lDDT	local Distance Difference Test (all atom)
lDDT-Cα	local Distance Difference Test limited to α-carbons.
LRMS	Ligand Root Mean Square (deviation)
LRRK2	Leucine Rich Repeat Kinase 2
mAbs	monoclonal Antibody
mmCif	macromolecular Crystallographic Information File
MD	Molecular Dynamics
MLP	Multi-Layer Perceptron
MS	Mass Spectroscopy
MSA	Multiple Sequence Alignment
MQA	Model Quality Assessment
MQAP	Model Quality Assessment Program
NAMD	Nanoscale Molecular Dynamics
NCBI	National Centre for Biotechnology Information
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
PAE	Predicted Alignment Error.
PD	Parkinson’s Disease
PDB	Protein Data Bank
PDB(e)	Protein Data Bank (in Europe)
PISA	Proteins Interfaces Structures and Assemblies
plDDT	predicted local Distance Difference Test (AlphaFold)
ProtCID	Protein Common Interface Database
PSSM	Position Specific Scoring Matrix

PPI	Protein-Protein Interaction
pTM	Predicted Template Modelling score (AlphaFold)
QS-score	Quaternary Structure Score
RDC	Residual Dipolar Coupling
RF	RoseTTAFold
RMSD	Root Mean Squared Deviation
ROC	Receiver Operating Characteristics
ROESY	Rotating frame NOE
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SAXs	Small Angle X-ray scattering
SNNS	Stuttgart Neural Network Simulator
SVM	Support Vector Machine
TBM	Template-Based Modelling
TM-score	Template Modelling score
TPR	True Positive Rate
TR-EM	Time-Resolved cryo Electron Microscopy
UCL	University College London
UKPDC	UK Parkinson's Disease Consortium

Acknowledgement

Thanks and appreciation to Prof. Liam McGuffin for his continued guidance and support over the last five and a half years. Also, thanks to Recep Adiyaman for technical guidance and Danielle Brackenridge for moral support at difficult times.

Thanks also to my wife, Stephanie, for her patience in allowing me to endlessly explain the finer points of protein model scoring routines, a subject she isn't the least interested in. I must also acknowledge my eldest daughter Dani for providing the inspiration to embark on this adventure as well as advice on write-up and technical points. Thanks also to my two younger daughters, Nicole and Josephine for their patience.

Lastly, I would like to dedicate this work to my mum, Angela, who was an inspiration to our whole family and who supported me and my daughters in everything we decided to do. Sadly, she died before this work could be completed.

Dedicated to Angela Catherine Edmunds, 2nd Oct 1933 to 16th Jan 2022.

CHAPTER 1
Introduction

1.0 An overview of the problem and the broad aim of the thesis

The discipline of computational protein modelling has evolved to address the problems associated with protein structure determination by experimental means. Many of these problems have been overcome with sophisticated practical approaches including X-ray crystallography, cryogenic electron microscopy (Cryo-EM) and nuclear magnetic resonance (NMR) techniques. However, the complexity and technical demands of these processes has made experimental structure determination an expensive and time-consuming process (Nealon *et al.*, 2017). As such, experimental techniques have not kept pace with the rate of identification of new protein sequences, which followed the completion of the Human Genome project in 2003, nor with the subsequent rise of techniques like two-dimensional gel electrophoresis (2-DE) and mass spectroscopy (MS) which have underpinned an expansion of protein expression mapping (Al-Amrani *et al.*, 2021). These kinds of proteomics advances have led to a significant sequence-structure gap resulting in an approximate 0.06% structure representation (Varadi *et al.*, 2022) within the Protein Data Bank (PDB) of the roughly 200 million amino acid sequences deposited in the UniProtKB database (The-UniProt-Consortium, 2021). Further to this, the rate of protein-protein interaction (PPI) identification has been increased by techniques like the yeast two-hybrid process. On the other hand, experimental structural determination methods have been described as showing less success with quaternary structure determination (Lensink *et al.*, 2017). This can be due to harsh preparation procedures like purification and dehydration which may distort or destroy associations between individual protein chains. Multimeric proteins, exhibiting some form of transient or obligate quaternary structure, therefore, represent a particular challenge in terms of closing the sequence-structure gap.

This study was conceived in 2018 with the aim of developing two unpublished, emergent computational pipelines; MultiFOLD for multimeric protein modelling and ModFOLDdock for multimeric protein model quality assessment (MQA). It was the intention that these two pieces of software would combine into a symbiotic pair with ModFOLDdock quality assessment driving continued improvements in MultiFOLD modelling. The ultimate intention was to create a publicly available webserver providing a one-stop multimer modelling and quality assessment tool, underpinned by accepted benchmarking results, which could be used to advance the quality of protein quaternary structure modelling and biomolecular research in general.

1.1 The essentials of protein structure and folding

1.1.1 Amino acid structure

α-amino acids are relatively simple organic molecules, all of which share a backbone or main chain consisting of a nitrogen and two carbon atoms (N-C-C). At one end the nitrogen forms

an amine (NH_2) group with a carboxyl (COOH) group formed by the carbon at the other end. The central α-carbon is attached to a single hydrogen and one other group, often referred to as the R (residue) group or sidechain, which is different for each of the 20 naturally occurring amino acids. This structure is shown in Figure 1.1 for two example amino acids and is important for two reasons; firstly, the amine and carboxyl groups from different amino acids are able to form a (peptide) bond between them, meaning that amino acids can be polymerised into long polypeptide chains. Secondly, the different R-groups confer different chemical and physical properties to each amino acid resulting in amino acid categorisation as aromatic, hydrophilic, hydrophobic, bulky, charged, polar or neutral. One other feature of polypeptides is that there is rotation around the σ-bonds within and between the amino acids, these are known as torsion angles and are called phi (Φ) (N to Cα), psi (Ψ) (Cα to Carboxyl) and omega (Ω) (peptide bond).

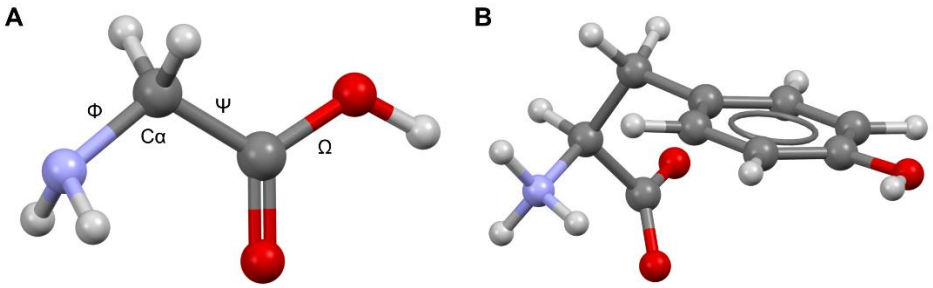


Figure 1.1 The structure of two α-amino acids showing main and sidechains. A. Glycine with a hydrogen sidechain and showing Φ, Ψ and Ω angles (adapted from https://commons.wikimedia.org/wiki/File:Glycine-neutral-lpttt-conformer-3D-bs-17.png), **B.** Tyrosine with a bulky aromatic sidechain. (adapted from https://commons.wikimedia.org/wiki/File:Tyrosine-from-xtal-3D-bs-17.png).

1.1.2 Protein structure (primary to quaternary) and torsion angles

As organised polypeptides, proteins are essentially chains of amino acids joined together by peptide bonds and the order in which the amino acids occur is referred to as a protein's primary structure - also simply called its sequence. The metaphor of beads on a string is sometimes used to visualise this arrangement and primary structure is classified as covalent bonding between main chain atoms. The sequence or primary structure, exemplified for three amino acids in Figure 1.2, is important because the properties of the relative amino acid sidechains will influence the final 3-D structure of the protein.

A simple example of a chemical property influence would be that amino acids with hydrophobic sidechains tend to favour the water-free core of a protein. A simple illustration of a physical property influence would be that sidechain size will dictate the ranges of phi (φ) and psi (ψ) torsion angles possible for any amino acid, meaning that Glycine, for example, is usually found at sharp turns in polypeptide chains. In this way primary structure is thought to govern

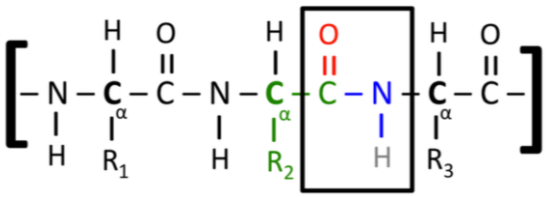


Figure 1.2 A section of primary structure showing the peptide bond. R = sidechains, the carboxyl carbon is now a carbonyl group and the amine group is now an amide (image adapted from EMBL-EBI online training: https://www.ebi.ac.uk/training/online/courses/biomacromolecular-structures/proteins/levels-of-protein-structure-primary/).

spontaneous higher-level protein folding through sidechain interaction, a concept often termed Levinthal's paradox (Zwanzig *et al.*, 1992). Levinthal argued that the short time it takes for a protein to fold evidences a folding pathway or mechanism governing the formation of the correct fold combination. A task that a random approach could theoretically take eternity to achieve.

Higher levels of protein structure are termed secondary, tertiary and quaternary structure. Secondary structures are stabilised by hydrogen bonding between main chain carbonyl groups (coloured red in Figure 1.2) and amide groups (coloured blue). This mostly results in one of two structures, the alpha helix or the beta sheet (although other structures are possible). These structures are determined by the torsion angles adopted by the constituent amino acids. In α-helices each amino acid hydrogen bonds with another four places further along the chain with typical torsion angles of -60° (phi) and -50° (psi) (Sailbil, 2010). In contrast, β-sheets form when torsion angles of -140° (phi) and +130° (psi) allow polypeptide chains to run alongside one another. Both structures are shown in Figure 1.3A.

Tertiary structures, shown in Figure 1.3B, are also stabilised by hydrogen bonding, but this time between amino acid sidechains which have been spatially rearranged following hydrophobic collapse of the structure, rather than main chain atoms. Side chains exert a certain influence over secondary structure via permitted torsion angles but will govern tertiary structure to a much greater extent through their level of hydrophobicity. Tertiary structure is characterised as secondary structure elements folding over each other via bends and twists using unstructured "loop" regions and is heavily influenced by the percentage and positioning of amino acids with sidechains of different properties. The result of the folding is to align linearly distant amino acids to form recognisable motifs some of which will be part of functional domains or active sites, in the case of enzymes.

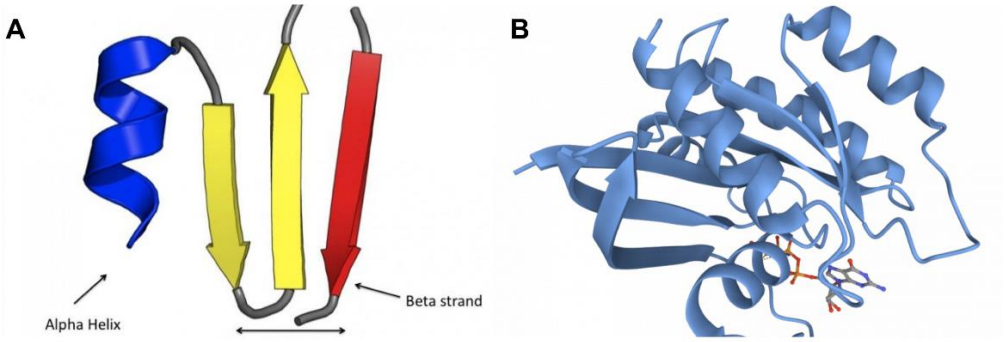


Figure 1.3 A. Secondary structures showing an α-helix and β-sheet made up of β-strands, and B. Tertiary structure showing folding of the secondary structure elements. (image adapted from EMBL-EBI online training: https://www.ebi.ac.uk/training/online/courses/biomacromolecular-structures/proteins/levels-of-protein-structure-primary/levels-of-protein-structure-secondary/)

The final level of protein structure is quaternary structure, a concept first proposed by Bernal *et al.* in 1958 and is the result of two or more individual protein chains binding together either permanently (obligate proteins) or in a transient association (non-obligate). This level introduces some additional structural complexity by having one (or more) interchain interfaces. Also to consider is the stoichiometry of the structure, i.e., the number of sub-units involved (dimer, trimer or higher association), and the symmetry (the orientation that each sub-unit takes relative to the others). Additionally, it is possible that some conformational changes may take place within the individual protein chains upon binding. This particular phenomenon is shown in Figure 1.4 where the unstructured regions in the monomers shown in image B spontaneously form α-helices upon association to form the interface, shown in image A.

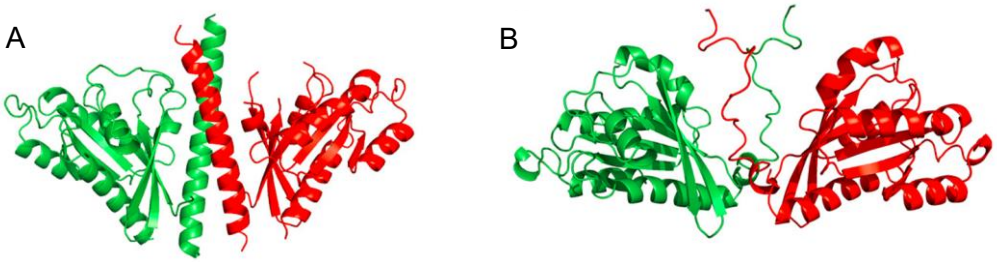


Figure 1.4. The quaternary structure of a simple homodimer. A. α-helices correctly form the interface. **B.** An early MultiFOLD model showing that the loop regions of the TBM tertiary model have not been altered to form the correct α-helix interface. Image taken from (Nealon *et al.*, 2017).

1.1.3 Protein folding

After the cellular processes of transcription and translation, a polypeptide chain rapidly folds into a predetermined structure that minimises the molecule's free energy (Anfinsen and Scheraga, 1975). Only at this stage can the polypeptide chain truly be referred to as a functional protein and its final three-dimensional conformation will depend on a number of factors that are a direct consequence of its primary structure. While this discussion ignores chaperone proteins and post-translational modifications (PTMs) such as phosphorylation, glycosylation

and methylation it is important to note that these can influence the stability and function of the protein (Zhong *et al.*, 2023). However, many proteins can achieve their native conformation without chaperones or PTMs, relying on main chain hydrogen bonding to stabilise their secondary structure (influenced by the preferred Φ and Ψ angles of individual amino acids) and the intra molecular forces resulting from hydrophobic collapse which are a direct consequence of side chain properties. These forces can be broadly characterised as Van der Waals interactions for larger, non-polar side chains; electrostatic interactions between charged side chains; permanent diploe interactions between polar side chains and possibly disulfide bridges (S-S covalent bonds) which are a consequence of Cysteine thiol (S-H) group bonding. All of these interactions are important in stabilising the protein but it is thought that the main thermodynamic driver for protein folding is, in fact, the interaction between the polypeptide chain and the water surrounding it, often referred to simply as the hydrophobic effect (Li *et al.*, 2021). Folding has the effect of increasing the entropy of the whole system due to the release of water molecules, which tend to become ordered around hydrophobic regions and as $\Delta G = \Delta H - T\Delta S$, the greater the entropy (S), the lower the free energy (G) if enthalpy (H) is constant or very small. Li *et al.* estimated that the change in enthalpy for main chain atoms forming H-bonds is approximately +2.7Kcal/mol while the entropy released due to the hydrophobic effect may lower the Gibbs fee energy to -23Kcal/mol (measured for Leu-Leu interactions), thus exemplifying the entropic compensation that hydrophobic collapse is assumed to provide to drive spontaneous folding. In this example only main chain interactions were considered and this is because the formation of secondary structure regions (β-turns, in particular) is thought to be the initial step in protein folding with one turn (sometimes called a foldon) influencing the formation of others in a chain-reaction style process (Englander and Mayne, 2014). Thereafter, further hydrophobic entropy gains are achieved as the protein folds into its final three-dimensional tertiary structure and buries hydrophobic side chains at its core. This hypothesis satisfies both Levinthal's paradox, that the degrees of rotational freedom are too great to allow folding without a pre-defined pathway and Anfinsen's assertion that folding must result in the lowest free energy conformation and be somehow encoded in the primary structure of the protein.

Unfortunately, while the fundamental principles that determine the final folded structure of proteins are better understood, the exact pathway and intermediate steps involved in protein folding remain elusive. Despite recent advances in artificial intelligence (AI) such as AlphaFold2 (AF2), which have improved the ability to predict final structures, understanding the dynamic folding process and its determining factors remains unsolved. Solving the folding problem is important as the function of a protein is generally considered a direct result of its three-dimensional shape which brings distant parts of the polypeptide chain together. These are then able to form structural motifs such as the β-turn which underpins the formation of the β-sheet or the

helix-turn-helix which has a role in DNA interaction and which themselves often form part of a functional domain of the protein. Being able to predict a protein's domain and the exact atomic coordinates within that domain from its primary structure would allow an understanding of the substrate or ligand with which the protein interacts as well as interactions with other proteins. Exact functions and modes of action could then be determined via accurate three-dimensional models which would allow insights into diseases associated with improper protein interactions and aggregation as well as therapeutic drug design where detailed knowledge of binding or active sites allows targeting of molecular pathways involved in disease.

1.2 Protein structure in healthcare and disease mechanisms

It has been estimated that free proteins may interact with up to 10 others to form low-affinity complexes (Chen and Skolnick, 2008), many of which are involved in cell catalysis, signalling or regulatory pathways (Sowmya *et al.*, 2015). An understanding of protein quaternary structure and protein-protein interactions (PPI) would therefore represent an important asset in structure-based drug design, and accurate protein models could be particularly useful in developing new therapeutics targeting cell signalling pathways, for example, which often involve a cascade of protein binding interactions. Some notable current treatments that rely on PPIs are those for cancer treatment, HIV, Alzheimer's, and Parkinson's disease.

For cancer treatment it may be possible to further exploit therapeutic approaches like the design of ligands to disrupt abnormal PPIs which would otherwise result in malignancy, similar to the Bcl-2 inhibitors for apoptosis regulation (D'Aguanno and Del Bufalo, 2020) and monoclonal antibody (mAbs) treatment to bind to specific target proteins in the same way as Trastuzumab targets the HER2 receptor limiting breast cancer cell proliferation (Gajria and Chandarlapaty, 2011). In HIV (HAART) treatment, enzyme inhibitors that target the active site of the HIV-1 protease have been effective in blocking the enzyme binding interactions with its substrates thus preventing viral replication (Lv *et al.*, 2015). Research into new treatments for Alzheimer's has recently employed PPI networks to identify potential repurposing of known drugs Raloxifene and gentian violet (Soleimani Zakeri *et al.*, 2021).

1.2.1 Parkinson's and the LRRK2 protein

Parkinson's disease (PD), a neurodegenerative disorder resulting from the loss of dopaminergic neurons in the substantia nigra, is a particularly interesting example where mutations in the leucine-rich repeat kinase 2 (LRRK2) gene have been identified as a genetic risk factor. The product of this gene, the RoCo (Roc and CoR domain) protein LRRK2, has been suggested as the vector for Parkinson's development. Human LRRK2 shares a homologue with the anaerobic phototrophic bacterium *Chlorobium tepidum,* called Ct.RoCo, which has been structurally solved and is shown in Figure 1.5. It is known that the Ct.RoCo

protein is involved in a GDP-mediated dimerisation cycle in which the monomer is GTP-bound while the dimer is GDP-bound (Deyaert *et al.*, 2019).

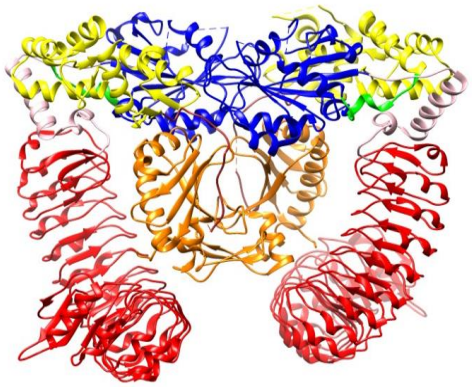


Figure 1.5 The PDB structure *6hlu* showing the *C. tepidum* LRRK2 protein Ct.RoCo, coloured by domain. The Leucine rich repeat (LRR) is shown in red, the linker section in pink, the RoC section in blue, a second linker in green, the N-COR section in yellow and the C-COR section in orange.

The LRRK2 protein is an example of a broader class known as the RoCo proteins which contain three key domains, the leucine rich repeat (LRR), the Ras-like GTPase (Roc) and the C-terminal of Roc (CoR). Conformational changes in the LRR and CoR domains are thought to regulate the dimerisation cycle. As a large, multi-domain protein involved in PPI and potential conformational changes, the LRRK2 protein represents a classic problem for experimental and computational protein modellers alike. As a result, the McGuffin group was contacted in 2020 by a research partnership from the Parkinson's Disease Consortium (UKPDC), the Department of Molecular Neuroscience, UCL and The Royal Veterinary College to model the human LRRK2 protein to allow assessment of its structural similarity to Ct.RoCo and thus whether the same GTP/GDP binding was likely. Although we were able to produce a reasonable quality model of the structure, the technology at the time did not allow the atomic level accuracy that this work required. More recently, advances have been made in producing experimental models of this protein, particularly using Cryo-EM and ET (electron tomography, described further in section 1.3.3) technology, however these have led to either high resolution images of single domains or low resolution images of the full length structure bound to microtubules (Zhang and Kortholt, 2023). The N-terminal domain remains unresolved as does the identification of the domains involved in membrane binding and the dynamic conformational changes involved in dimerisation and phosphorylation and, importantly how these are affected by the PD mutations. The same study describes how AlphaFold2 structures have contributed to the structural knowledgebase but also highlights some disagreements between the predicted and observed structures, which are yet to be resolved. The continued struggle to produce a full-length, atomic resolution model of LRRK2 for disease understanding and potential drug development exemplifies the need to improve the quality of protein quaternary structure modelling.

1.3 Experimental methods of protein structure determination

1.3.1 X-ray crystallography

X-ray crystallography has traditionally been considered the gold standard for biomolecular structural determination due to its ability to produce images at atomic resolution. Briefly, the method involves the crystallisation of the target protein followed by X-ray diffraction and finally mathematical calculations to produce the electron density map and the final molecular structure. Despite its standing as a cornerstone of structural bioinformatics, the technique is not without its challenges and limitations.

In order to produce the required well-ordered crystals, sufficient quantities of the protein must first be expressed by a suitable cellular host and then purified. Thereafter, the sample is subjected to dehydrating and crystallising conditions, which can in themselves prove challenging for some proteins due to size, solubility and flexibility issues. Obtaining suitably high-quality crystals can, therefore, require extensive optimisation of conditions which can be time-consuming and resource intensive, although advances such as microcrystal electron diffraction (MicroED) (Mu *et al.*, 2021) and automated crystallization screening (Shaw Stewart and Mueller-Dieckmann, 2014) have significantly improved efficiency and success rates by allowing the use of smaller crystals and faster identification of optimal conditions.

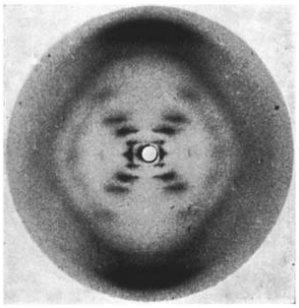


Figure 1.6 The famous *Photo 51* showing the X-ray diffraction pattern of DNA (Image taken from https://en.wikipedia.org/wiki/File:Photo_51_x-ray_diffraction_image.jpg).

The X-ray process involves collecting diffraction patterns resulting from the interaction between the electrons in the sample and the X-ray wave as explained by Bragg's law (Thomas, 2012). However, despite the complexity of this process, the result is merely a pattern of light and dark spots, as shown by Frankin's famous crystallographic Photo 51 of DNA (Figure 1.6). These patterns require mathematical interpretation. Diffraction patterns similar to those shown in Figure 1.6 correspond to the arrangement of atoms within the crystal lattice and the spot intensities contain information about the spatial distribution of electrons. Mathematical techniques such as a Fourier transform or molecular replacement (where a known model of a homologous structure provides starting point values) can be used to construct an electron density map from the diffraction pattern, which is then interpreted into atomic coordinates. In

addition to the potential difficulties mentioned above with the crystallisation process, X-ray crystallography only produces a single snapshot of one conformation of a protein, which may not represent its biological form. Additionally, proteins are not static entities; they exhibit flexibility and undergo conformational changes related to their function and these will not be captured by crystallography. Lastly, as alluded to in Section 1.0, crystallography conditions are extreme and larger proteins, particularly multimeric structures can become damaged during the process. Membrane proteins, due to their size and hydrophobic transmembrane section represent a particular crystallisation challenge.

1.3.2 Nuclear Magnetic Resonance

Nuclear Magnetic Resonance (NMR) is a spectroscopic technique originally used for small organic molecule structure determination, but which has been adapted for larger molecules such as nucleic acids and proteins. The core concept is that all nucleons exhibit a phenomenon known as spin, meaning that a nucleus comprising odd numbers of nucleons will itself exhibit an overall spin moment. Nuclear spin is measurable in Hydrogen atoms as they have only one proton and so proton NMR (also known as ^{1}H NMR) is a useful technique to investigate organic molecules due to their high Hydrogen atom content. Other NMR techniques are also possible using isotopes of carbon (^{13}C NMR) and sometimes nitrogen (^{15}N NMR).

Structure determination for small organic molecules, where the identity of the molecule is unknown, centres on two key concepts, that of carbon environments in ^{13}C NMR and spin (or J) coupling for ^{1}H NMR. The former allows the user to identify the number of carbon atoms in a molecule, whereas the latter allows the assessment of the number of hydrogens on each carbon by interpretation of peak splitting within the trace. Along with the chemical shift which helps to identify different organic functional groups, the identity of molecules can be determined. For proteins, the identity of the molecule is not in question as it is described entirely by the primary structure. The important aspect is the spatial arrangement of the amino acids within the protein. For this a slightly different spin characteristic is used, one called the Nuclear Overhauser Effect (NOE). The NOE is essentially the transfer of spin, called cross-relaxation, between atoms that are in close proximity, usually defined as ≤ 6Å (Hu *et al.*, 2021). In this way local spatial relationships can by determined via either a 2-D tracing technique known as NOESY or a slightly more complicated rotational version called ROESY. Spin-coupling can, however, play a role in refinement of the structure suggested by the NOE. This technique, called residual dipolar coupling (RDC), essentially involves comparisons of peak splitting patterns for identical molecules measured under different anisotropic (orientation) conditions. NMR requires less harsh conditions than crystallography and is therefore more suitable for structure determination in environments resembling physiological conditions. It is

also more sensitive to alternative protein conformations and protein dynamics meaning that investigations into PPIs are possible. NMR has also been used successfully to characterise membrane proteins (Opella and Marassi, 2017) and the number of NMR structures deposited in the PDB has risen year-on-year since 1991, currently totalling 14,189 (result of a search on 02/03/2024). However, NMR can encounter resolution limitations with larger proteins and complexes in excess of 80kDa due to spectral overlap (Hu *et al.*, 2021) and requires relatively high sample concentrations and stability of its target protein (Benjin and Ling, 2020). Another limiting factor has traditionally been the length of time required for specialist data interpretation, taking months in some cases to convert measurements into structures (Klukowski *et al.*, 2022), particularly for proteins with multiple conformations or significant dynamics.

1.3.3 Cryogenic Electron Microscopy

Cryogenic electron microscopy (Cryo-EM) has recently emerged as a potentially revolutionary technique allowing structural determination of large proteins, complexes and membrane proteins at near-atomic resolution without the need for crystallisation. This immediately resolves many issues surrounding X-ray crystallography, making Cryo-EM suitable for studying challenging targets and those with multiple dynamic conformations under near native conditions (Murata and Wolf, 2018). A recent and notable example of this was its use to model the multimeric SARS-CoV spike protein trimer (Alsaadi and Jones, 2019). The basic technique centres around flash-freezing a solution of the target protein in vitreous ice prior to examination by electron bombardment, but there have been low resolution issues surrounding the structures produced for many years (Callaway, 2020). Consequently, the number of structures resolved by EM techniques in the PDB has been slow to develop, standing at just a single structure in 1991 and climbing to 320 by 2010 following Richard Henderson's resolution review in 1995 (Henderson, 1995). However, following the work of Dubochet, Frank, and Henderson in 2017, 2020 saw a breakthrough in Cryo-EM techniques, allowing true atomic-level resolution (below 3Å (Ashmore *et al.*, 2021)) to be obtained for the first time (Yip *et al.*, 2020;Nakane T, 2020) with structures reaching a maximum resolution of 1.2Å. Accordingly, the number of structures in the PDB has risen to 19,106 (03/03/24) with 4582 deposited in 2023 alone. However, Cryo-EM is not without its challenges; problems may yet be encountered with unstable, aggregated or low homogeneity samples, buffer contamination or freezing issues, all of which can reduce contrast and resolution. Sample preparation therefore continues to be time intensive, requiring a high level of expertise coupled with high-quality instrumentation, especially for small (<500kDa) or flexible proteins (Benjin and Ling, 2020).

Despite this, recent advances have led to the development of new techniques like time-resolved cryo-electron microscopy (TR-EM) and Cryo-electron tomography (Cryo-ET). These

two innovative methods enable visualisation of biological molecules in dynamic states and in cellular conditions. With TR-EM, conformational changes can be captured by freezing and immobilising molecules at different time points during their dynamic transition. This could reveal further details of PPIs or even the mechanism of protein folding itself. Cryo-ET allows investigation into whole cells and, through tilted imaging, can show the 3-D location of large biomolecules within the cell. Additionally, recent work following CASP15 suggested that mechanisms for both validating Cryo-EM structures using AF2-style distance predictions (Sanchez Rodriguez *et al.*, 2022) as well as resolving poorly modelled loop regions by refinement can be realised using predicted computational structures (Mulvaney *et al.*, 2023).

1.4 Computational solutions to protein structure prediction

Although the discipline of protein modelling began experimentally with Kendrew's 1957 model of myoglobin interpreted from X-ray analysis, it wasn't long before computational methods were developed, initially in the form of probabilistic secondary structure prediction by the Chou–Fasman method in the early 1970s (Chou and Fasman, 1974). The Chou-Fasman method was based on amino acid frequencies determined by X-ray crystallography demonstrating that, from the earliest days, computational methods have relied on experimental data to make predictions. Thus, as the availability of experimental data increased, the potential for complementary computational methods also rose, with the single most significant source of data being the Protein Data Bank (PDB), established in 1971 at the Brookhaven National Laboratory, which dovetailed with the development of the first sequence alignment algorithm by Needleman and Wunsch (Needleman and Wunsch, 1970). The first successful homology model, meaning the construction of a model of a protein with an unresolved structure entirely by comparison with evolutionarily related homologues, is generally considered to be Greer's 1980 structure of the haptoglobin heavy chain (Greer, 1980). From that point, using the increasing number of experimental structures in the PDB, which hit 25,000 in around 2003 and currently stands at 217,157 (PDB search on 13/3/24), as well as the growing availability of sequence databases (UniProtKB held 190 million in 2021, (The-UniProt-Consortium, 2021)) and increasing computational power, homology or comparative modelling has become a useful method of protein structure determination. Additionally, it has been estimated that at least 70% of known protein sequences have at least one domain related to another protein (Fiser, 2010) meaning that, as more structures are determined experimentally, many more structures become available for homology modelling.

1.4.1 A summary of tertiary structure comparative modelling

The terms comparative modelling (CM), homology modelling (HM) and template-based modelling (TBM) have become almost interchangeable, although strictly speaking homology modelling describes the process of using structural templates with an established evolutionary

relationship to the target sequence. Regardless of terminology differences, the method has been a popular technique driving computational tertiary structure modelling. Rangwala and Kapris (Rangwala and Karypis, 2011) defined the process in terms of five distinct stages; selection of templates, alignment of sequences, model building, quality evaluation and refinement.

Identification of suitable templates is often the most important part of the TBM process and can be achieved by sequence alignment tools such as PSI-BLAST (Altschul *et al.*, 1997) using the NCBI database (Sayers *et al.*, 2022) to produce paired alignments between two sequences. Often attempting a global alignment of the whole target sequence using the Needleman and Wunsch algorithm or similar, results in few or poor matches owing to the potential for protein domains to swap places over time. Therefore alignment routines often use local sequence alignment techniques, first devised by Smith and Waterman (Smith and Waterman, 1981b), where sequences are considered in segments and then cross-aligned to allow a search of the whole sequence for matches. Again, due to the nature of protein evolution, even successful alignments encounter missing sequence sections (deletions), additional sections (insertions) or substitutions where amino acids have been replaced with others. It can then become difficult to directly compare sequence alignments and a BLOSUM matrix (Henikoff and Henikoff, 1992) is often used to contextualise each alignment by scoring conserved amino acids well and penalising missing sections or those where replacements have occurred, particularly in ordered secondary structure regions. Some programs also use a secondary structure consensus predictor like PSIPRED (Jones, 1999) at this point to increase confidence in the final template selection. Despite some structural diversion with increasing evolutionary distance, protein structure has remained surprisingly stable (Chothia and Lesk, 1986) and, in general, sequence identities above 30% have been successful in establishing similar structures via evolutionary relationships (Buenavista *et al.*, 2012), although this threshold is somewhat length-dependent and may depend on the absolute number of shared residues. For sequences with very low identities, a technique known as fold recognition or threading can be employed. In this approach, the query sequence is used to generate a position-specific scoring matrix (PSSM), which captures evolutionary information by scoring each position based on a multiple sequence alignment. The PSSM is then used to search the PDB for compatible structures by aligning the sequence to known structural templates, thus predicting the fold of the query sequence (Bowie *et al.*, 1991).

Once templates are identified it is usual to perform a second alignment, often a multiple sequence alignment (MSA) is used to align the target protein sequence with one or more template structures. The goal is to identify structurally conserved regions between the target

and templates to guide the construction of the model. From this it is possible for modelling software to construct an initial model, most commonly by spatial restraints as used by the popular modelling software MODELLER (Eswar *et al.*, 2006). During this process inter residue distances, and a host of stereochemical constraints including bond lengths and dihedral angles are used to construct complementary structures guided by the template and then select the best structure on the basis of minimum violation of the constraints. Unfortunately, this is rarely sufficient to build the complete model unless very close and high-quality templates are available. The parts missing tend to consist of the unstructured loop regions which occur between areas of organised secondary structure which, in the main, make up the fold and domains of the protein. Loop modelling can be achieved either by Ab initio modelling entirely guided by physics-based rules (often represented by the CHARMM (Brooks *et al.*, 1983) force field) to predict the shape from first principles (e.g., ModLoop (Fiser and Sali, 2003) or Rosetta (Simons *et al.*, 1997)) or by using a loop-fragment database (e.g., ArchPRED) (Barozet *et al.*, 2021). It is worth mentioning here that unstructured loops also present a problem for experimental methods, with estimations that up to 69% of structures in the PDB have missing fragments, rising to 80% for very high resolution structures (Djinovic-Carugo and Carugo, 2015). This can be due to inherent flexibility, making loops difficult to resolve accurately by crystallography or NMR. Flexibility can lead to incoherent X-ray scattering and a subsequently weak contribution to the electron density map or a weak NOE signal in NMR, resulting in an ensemble of differing conformations (Kwan *et al.*, 2011).

Despite the sophistication of modelling software, it is not uncommon for models to contain both local and global errors like unrealistic contacts or hydrogen bonds, steric clashes, incorrect bond lengths or unfavourable dihedral angles (Bhattacharya and Cheng, 2013). Despite this it can be challenging to improve models and attempts can result in the deterioration of model quality, particularly for TBM models (Adiyaman, 2021). Refinement is the process of improving a model by making small changes to the 3D structure with the aim that the new model will be closer to the native protein than the original. Refinement programs can be broadly split into two types; stand-alone stereochemical force fields like AMBER (Cornell *et al.*, 1996) which can be used to directly optimise for bond-length and geometry and full molecular dynamics (MD) simulations (of which the full AMBER package is also an example). MD programs can be further sub-divided into manual programs which tend to perform computationally intensive simulations and are available to download and run locally, requiring some technical familiarity with the software. Alternatively, automated server-style programs are available via public webpages which tend to be quicker and less computationally intensive, using methods like side-chain optimisation and less stringent energy minimisation functions (Feig, 2017).

To briefly explain molecular dynamics simulations. These simulate the motion of atoms over time when the model is programmatically solvated in water with an ionic component designed to mimic physiological conditions. The stereochemical force fields (AMBER or CHARMM) are again used to govern the potential energy of the system and so dictate the atomic positions. The simulation traditionally consists of two stages; equilibration - where the protein is allowed to adjust to the environment via energy minimization to fix clashes and achieve a stable starting point, and the main simulation - which may include a perturbation step, thereafter allowing the model to settle into a thermodynamically favourable state over a short period of time. The goal is to allow the model to explore different conformations and interactions arriving at the thermodynamically most favourable. After the simulation, the programs perform two further functions; the first is sampling, meaning to create a range of refined models, the second is scoring, using an energy function such as DFIRE (Zhang *et al.*, 2004) or a stereochemical checker like MolProbity (Williams *et al.*, 2018), to identify any improvements.

1.4.2 CASP competitions and the success of different modelling strategies

The **C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction (CASP) experiment is a biennial blind structure prediction competition created by John Moult and colleagues in 1994 (Moult, 2005). Its aim is to objectively assess the prediction capability of modelling groups worldwide and to create a forum for shared practice. Organisers source unpublished experimental structures and invite predictor groups to model the structures, the native structures are revealed some months later along with scores for each submitted model (Moult *et al.*, 2011). The experiments have attracted increasing participation over the years; CASP1 consisted of 35 invited predictor groups (Moult *et al.*, 1995) whereas CASP8 (2008), for example, received predictions from 253 groups across 24 countries.

In the earlier years, many predictor groups favoured *ab initio* modelling using physics-based methods including free energy calculations, electrostatic interactions, hydrogen bonding and solvation energy scoring functions to empirically solve the folding problem (Moult, 2005). By CASP10 (2012) the number of solved structures in the PDB had reached 87,000, representing 1393 unique folds and researchers had changed their focus to comparative modelling to exploit the available templates. The rise of TBM methods allowed the creation of ever greater numbers of models for each target, potentially at the expense of the understanding of folding mechanics, but also creating an increasing requirement for model quality assessment (MQA) and ranking programs. Consequently, two future challenges highlighted in the CASP9 report (Moult *et al.*, 2011) were the improvement in accuracy of regions not easily derived from a template and improvements in methods for selecting the best model from those generated by TBM programs.

By 2016, these challenges had started to be met and CASP12 and 13 models showed an increase in accuracy (Kryshtafovych *et al.*, 2018) which was attributed partly to the increased number and quality of templates available in the PDB, but also to improved model selection by MQA programs (Croll *et al.*, 2019), highlighting the importance of quality assessment in driving modelling advances. The next significant increase in model quality was seen at CASP14 (2020) with the participation of Google DeepMind's AlphaFold2 (AF2) (Jumper *et al.*, 2021b) deep learning software. The increased levels of accuracy and methods by which they were attained are covered later in Section 1.5.

Quaternary structure modelling, known as assembly modelling, was included as an assessed category from 2016 (CASP12) and, in 2022 (CASP15), the estimation of model accuracy (EMA) category was modified to focus on scoring quaternary structure models. This, again, demonstrated the value that CASP organisers placed on quality estimates in advancing protein modelling quality.

1.4.3 Docking and the docking problem

As described in Section 1.2, correctly predicting protein assembly binding orientations using docking methods may provide a knowledge base for medical development. One route could be via drug development, particularly those designed to disrupt protein-protein binding interactions but a second, equally important route, could be via the generation of antigen-antibody complexes for the treatment of autoimmune conditions or vaccine development, the latter exemplified by work supporting the recent Covid vaccines (Bansal *et al.*, 2021).

Docking and screening routines, in which an initial phase of protein docking is followed by scoring each docking pose, were popular methods in early CASP experiments (Vasker, 2014). This was due to a number of factors, firstly that docking programs had been developed for protein-ligand docking studies (Sousa *et al.*, 2013) and these were easily repurposed for protein quaternary structure modelling and, secondly, that TBM approaches had experienced limited success due to a lack of multimeric structures of sufficient quality to use as templates (Lensink *et al.*, 2016).

The docking problem is one where, using only the 3D atomic coordinates the native positional and rotational orientations between two protein molecules must be identified (Vasker, 2014). This must be achieved without significant overlap of atomic space (clashes) nor by leaving gaps between the chains. Shape complementarity could rely on flexibility so it would be useful if docking algorithms allowed flexible chain binding. However, this has for the most part, remained too computationally expensive (Marze *et al.*, 2018) and grid-based rigid-body docking methods became the core technology, representing a less complicated but affordable

compromise (Garzon *et al.*, 2009), although soft docking approaches permit a certain steric overlap to represent flexibility (Bonvin, 2006).

Solving this problem requires the sampling of many thousands or even millions of potential poses to account for the many translational and rotational orientations possible between the two proteins, often referred to as receptor and ligand regardless of size difference. To facilitate this within the capabilities of most servers, a fast Fourier transform (FFT) algorithm was used (Katchalski-Katzir *et al.*, 1992). This involved representing proteins as 3D projections consisting of nodes and edges where each point is defined by a range of scores. These include definitions of surface (1), internal (-1) or external (0) space as well as a number of amino acid properties like hydrophobicity, side-chain size and electrostatic interactions, for example. These values are then discretised as a matrix where they can be converted by a Fourier transform into a frequency-space representation (Yin and Yau, 2017). It is then much simpler to compare frequencies to find potential matches than it would have been to compare all of the individual scores, on the assumption that areas with complementary properties are likely to represent binding sites. Promising poses which show high surface to surface definition (rather than surface to internal or external atomic space) can then be scored on a shape-complementarity or energy basis. A list of high-scoring poses and scores can then be output by the program.

1.4.4 Quaternary structure prediction at CASP

Assuming it is possible to replicate a high percentage of docking poses via the FFT method, the success of docking methods is then governed by the ability to select the native-like poses from the decoys. Docking success therefore becomes a function of MQA accuracy, a reason for the pressing need to develop reliable quaternary structure MQA methods.

In an early joint CASP/CAPRI experiment run as part of CASP11 (2014), despite the difficulties with reliable scoring and selection techniques, docking was considered a superior method to early multimeric TBM attempts, mainly due to the relatively low numbers and quality of available templates in the PDB and specialist databases like PISA (Krissinel and Henrick, 2007), (Lensink *et al.*, 2016).

At CASP13 (2018), participating groups in the assembly competition employed a mix of docking and TBM strategies (Kryshtafovych *et al.*, 2019), but the success remained somewhat varied, leading the authors to conclude that, although good models were seen when closely related templates existed for the whole assembly structure, the approach of building separate monomers and then docking them via rigid body methods was essentially flawed. This opinion was somewhat reinforced by the CASP13 official results (Duarte and Guzenko, 2018) which

showed that assembly modelling had only a 31% success rate (measured by all aspects of assessor total score >0.5) which could further be broken down into global relatedness (measured by TM-score >0.5) of 80% but interface similarity (measured by interface contact score (ICS) >0.5) of only 34%. Despite the development of hybrid techniques employing both TBM and docking methods, exemplified by software like GALAXY (Lee *et al.*, 2017) and data-driven approaches like HADDOCK (Vangone *et al.*, 2017) as well as the availability of interface fragment libraries like Swiss-Model (Waterhouse *et al.*, 2018) and ProtCID (Xu and Dunbrack, 2020), CASP14 assembly modelling resulted in only marginal improvements in accuracy, with the percentage of TM-scores >0.5 rising to 86% and those for ICS rising to 38% (Karaca, 2020).

It was clear that the formation of correct interfaces was a problem for quaternary structure prediction. The Venclovas group, who had achieved first and second place in CASP13 and 14 assembly modelling respectively, further demonstrated this problem by breaking down their CASP14 modelling results by method and comparing them by QS-score, which is particularly sensitive to interface orientation. They found that for free docking 80% of models scored 0.3 or lower, for hybrid docking this reduced to 55%, further reducing down to only 9% for TBM modelling (Dapkunas *et al.*, 2021). Although this analysis used a very limited number of models and docking methods were only employed where good templates could not be found, it nevertheless exemplified the difficulty in locating good multimeric templates as well as the continued problems with docking model interfaces.

1.4.5 Refinement and a gap in quaternary prediction methods

The status of multimer or quaternary structure modelling at that point in time was the motivation for the title of this project, that is to say that the data pointed to the need for a reliable multimeric MQA method to improve the selection of native structures from long lists of decoy models and also that errors in multimeric models appeared to centre around the interface contacts, an area potentially sensitive to the resolution of clashes by refinement. At the time, there were limited options for multimer refinement and of the two methods explored, SymmRef (Mashiach-Farkash *et al.*, 2011) and GalaxyRefineComplex (Heo *et al.*, 2016), the latter was chosen due to the former's specialisation for symmetrical structures which possibly limited its use. GalaxyRefineComplex is a side-chain repacking algorithm from the Seok lab in which models are relaxed using molecular dynamics (MD) simulations. It was the initial intention to improve multimer models by this method and incorporate a similar approach into our fledgling MultiFOLD pipeline. In the end, due to rapid advances in computational methods, a different method of refinement and model improvement was eventually developed for MultiFOLD.

1.5 Advances in computational methods

1.5.1 The importance of multiple sequence alignments (MSA)

The concept of a multiple sequence alignment (MSA) has been known to the protein modelling community since the early days of computational biology research in the mid to late 20th century, with seminal contributions by pioneers such as Margaret O. Dayhoff (Strasser, 2010) and advancements in alignment algorithms by researchers like Smith and Waterman (Smith and Waterman, 1981a). The ability of an MSA to reveal patterns not seen in simple pairwise alignments has made them useful in fold recognition or threading approaches, where amino acid probability profiles are created to identify similar folds in different templates or secondary structure similarities (Jones, 1999). Consequently a number of algorithms like Divide and Conquer (DCA) (Tonges *et al.*, 1996), MUSCLE (Edgar, 2004) and Kalign (Lassmann and Sonnhammer, 2005) were developed with Clustal Omega (Sievers and Higgins, 2014) becoming a popular choice a little later.

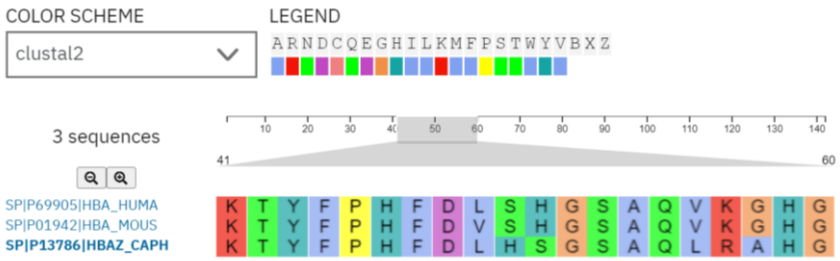


Figure 1.7 A multiple sequence alignment (MSA). An example the output for the test amino acid sequence supplied on the Clustalw webpage (https://www.ebi.ac.uk/jdispatcher/msa/clustalo).

However, it wasn't until later that deep alignments were used specifically to establish evolutionary relationships (de Juan *et al.*, 2013). The theory is essentially that conserved residues can be used to highlight evolutionarily stable regions of the protein and that where sequentially distant amino acid residues are shown to co-mutate, the likelihood is that there is a relationship between them, based on a contact formed upon folding. By charting these co-evolutionary mutations, it is possible to construct a contact map which can be used to guide (template) Free Modelling (FM) predictions (Li *et al.*, 2019), that is, modelling using energy functions and conformational mapping rather than that relying on the availability of similar structures in the PDB. However, creating deep MSAs can be computationally expensive, and interpreting the coevolution data via a Potts model requires sufficient depth meaning that early methods like PSICOV (Jones *et al.*, 2012) and GREMLIN (Kamisetty *et al.*, 2013) could fail for shallow alignments. To solve this, an element of machine learning was added which was able to distinguish between conservative (little structural effect) and non-conservative (significant effect) mutations (Lupo *et al.*, 2022) using fewer sequences by training on prior data. Another of the pioneering methods linking MSA information with supervised machine learning for

accuracy gains was MetaPSICOV (Jones *et al.*, 2015). The use of deep learning methods for contact prediction pushed the accuracy of contacts maps even further, a concept which was later adapted by DeepMind with well documented success in their first version of AlphaFold (Senior *et al.*, 2019).

1.5.2 Machine learning

The key concepts of machine learning, the technology underpinning Artificial Intelligence (AI), are that collections of data points are defined by distinct unitary parameters such as time, volume or temperature, for example. At its most basic, machine learning is simply a case of defining these parameters as either inputs or outputs and setting a computer the task of predicting the latter from the former. Machine learning is routinely categorised into three main types: supervised, unsupervised and reinforcement learning. Briefly, supervised learning uses labelled data and the algorithm is asked to find the best way to associate the input parameters with the true labels which form the output parameter. True labels are often supplied via experimental processes, and in the case of MQA, these would be the observed quality scores of the model that are generated by comparison with the native structure. In unsupervised learning, the algorithm will be supplied with unlabelled data where the emphasis is on learning how to cluster like data together, reduce the range of data to focus on important patterns or to detect anomalous values (Parasa *et al.*, 2021). Reinforcement learning, on the other hand, is more focussed on decision-making in a reward-penalty paradigm with the aim of maximising reward over penalty. Reinforcement learning is commonly associated with gaming-style algorithms.

Of the three types, supervised learning is most often used in protein modelling scenarios as it is suited to either classification or regression tasks (Greener *et al.*, 2022). Classification tasks are used if the true labels are mutually exclusive, like identifying protein sequences that represent the distinct secondary structure conformations helix, sheet or strand, perhaps. Regression tasks are more suited to data that are continuous in nature, like quality assessment scores. Sometimes it is appropriate to convert continuous data into categorical data to add a classification task to the regression task, this is usually done to allow the user to collect a single score enabling direct comparisons between different scenarios. A common example of this is creating binary data used to construct contingency tables from which a true positive rate (TPR) value can be calculated or, additionally, receiver operator curves (ROC) from which an area under the curve (AUC) value can be calculated. When using supervised learning with continuous data, analysed as a regression task, a simple multi-layer perceptron (MLP) is the recommended machine learning architecture (Greener *et al.*, 2022).

1.5.3 Support vector machines (SVMs)

Support Vector Machines (SVMs) are explained here as they are mentioned in Chapter 3 as a machine learning class that had an initial impact on protein MQA. They are supervised learning-based algorithms which are particularly suited to classification tasks. They work by classifying data into two classes based on a specified feature. In order to visualise the concept, it is useful to consider a number of points plotted on an axis with a line drawn to separate the two distinct classes, as depicted in Figure 1.8A. In this case, the further the points are from the dividing line the more confident we can be that they truly belong to their respective class. The margin is a conceptual area between the dividing line and the first true point in each class, the boundaries of which are described by the two support vectors and which is the focus of the SVM. As more data is added to the model it is conceivable that anomalies will encroach into the margin and some will be misclassified on the wrong side of the dividing line. An SVM will then attempt to redraw the dividing line as a plane of separation by replotting the data from one dimension to two dimensions (or higher if required) thus finding a new separating line to minimise misclassification. This is essentially support vector regression using a hyperplane to optimally separate the data. The power of SVMs can be used to decide where a data point fits into a pattern, for example, whether a model agrees with the native structure or not. Powerful computers can be used to apply hyperplanes to higher level distributions to make decisions for thousands or millions of data points.

1.5.4 Neural network (NN) architecture and training

An artificial neural network (ANN, often abbreviated simply to NN) works on the principle of programmed nodes and connectors with the whole arrangement often referred to as a model, (a term avoided whenever possible in this document to minimise confusion with a protein 3D model, instead we refer to the learned NN model as the “weights”, see below). Nodes (representing artificial neurons) are arranged in layers and are interconnected by a series of connectors (representing artificial synapses). A simple representation of this architecture can be found in Figure 1.8B, along with the SVM diagram. The output from one node represents the input to one in the next layer, this output is a number and is usually referred to as a weight. During supervised training, a NN learns by adjusting the weights between nodes in individual layers to minimise the difference between the final predicted value and the true label. This difference is often termed the loss function which is usually measured by mean squared error for regression tasks. NNs always have an input layer for the initial input values and an output layer for the output prediction but vary by the number of hidden layers that separate them. Additionally, NN architecture may vary by the number of nodes within each hidden layer. Deep neural networks (DNN) such as the DeepMind network behind AlphaFold consisting of many hidden layers and requiring considerable computational power.

The simplest form of NN is a feed-forward network which only transmits the signal forward, from input to output layer without feedback to previous layers. This is essentially a multi-layer perceptron (MLP) and uses a concept called backpropagation to calculate the loss function and thus automatically adjust the weights. In order to ensure this is done appropriately, the first step in training is to set the network hyperparameters. If using a package like the Stuttgart Neural Network Simulator (SNNS), many of the more complicated parameters are controlled automatically. However, it is still necessary to set the number of neurons in each layer, the number of layers, the learning rate (which defines the step size for the weight updates that occur with each iteration), the maximum difference between prediction and true label considered an error (Max Diff) and the maximum number of iterations that the optimisation algorithm allows during training (Max It). This can be accomplished by calculating the maximum performance from a number of test runs with varied hyperparameter settings. The point of this is to avoid overfitting and underfitting. Overfitting is where the MLP is essentially too powerful for the data presented to it and will proceed to learn the dataset rather than the relationships within, leading to perfect performance on the training data but poor performance on testing data. Underfitting is the opposite, where the MLP fails to learn the relationships and performs poorly on all data. This is explained more comprehensively in Chapter 3 (3.3.5).

The next step is the training of the MLP itself. To avoid overfitting, one commonly used strategy is called N-fold cross-validation. In this technique, data are split into training and testing datasets so that the data used for predicting (testing dataset) are separate from the training dataset. In this way the true labels for the testing dataset are never seen by the MLP. N-part cross-validation results in the dataset being split into N parts, the MPL will train on N -1 parts and predict on the remaining part of the dataset. Often, N versions of the MLP are created so that every part of the dataset is equally used for both training and testing.

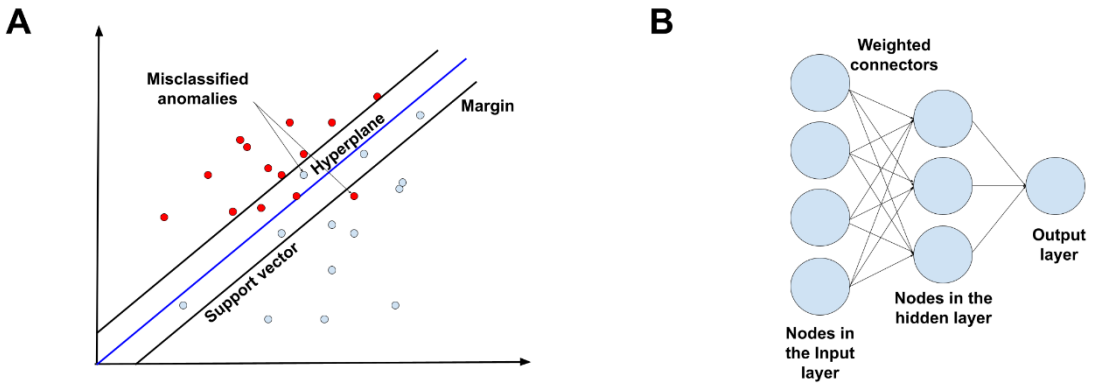


Figure 1.8 Representations of two types of machine learning (ML). A. A representation of an SVM showing the hyperplane and two support vectors. **B.** The architecture of a simple feed forward MLP with one input, one hidden and one output layer.

1.5.5 AlphaFold2 (AF2) and new levels of accuracy in CASP14

At CASP14 (2020) Google DeepMind submitted tertiary structure models using their new method AlphaFold2 (AF2), which represented a significant improvement in tertiary structure model quality. In fact, the high accuracy they achieved in the FM (no templates available) and FM-TBM (limited templates available) classes (Kryshtafovych *et al.*, 2019) has been described as “atomic level” (Yang *et al.*, 2023) with median GDT_TS (see Appendix 1 for the definition) scores of 87.0 and 92.4 respectively (Jumper *et al.*, 2021a) (scores >75 are considered to have mostly correct atomic coordinates (Kryshtafovych *et al.*, 2019)). These were impressive figures when contextualised against the previous experiment (CASP13 in 2018) where the FM average GDT_TS score for the highest scoring group was 61.4 (Senior *et al.*, 2019).

AlphaFold2 achieved this impressive jump in performance with the unique union of two key ideas. The first was a deep multiple sequence alignment (MSA) which was made accessible by clustering, where similar sequences are clustered together and a single representative of each cluster is submitted for consideration. This technique reduced the computational resources required to detect evolutionary relationships between amino acids and also added. AF2 also constructed detailed pair representations in the form of residue pair relationships like type, position and distance measures. This information was combined with that from the MSA and used to create a “distogram” (Li, 2022) from which the basis for a residue contact map of the target protein could be formed. The second was a deep neural network (DNN), or more correctly, a pair of DNNs (Jumper *et al.*, 2021b) running on Google DeepMind’s powerful servers. The attention-based transformer (Evoformer) was used to interpret the MSA and distogram information (Lupo *et al.*, 2022) into contacts and then a graph representation of a starting model, with information then passed to the Structure Module to construct a final real-world structure from the starting model by applying protein modelling constraints such as torsion angles and side-chain preferences as which were obtained using a set of residue triangulation calculations . DeepMind also programmed a feedback or recycle pathway into the algorithm, allowing AlphaFold2 to repeatedly pass information about the newly forming model created by the Structure Module back to Evoformer for further evaluation. This clever idea allowed a shuttling of information backwards and forwards between the modules allowing the DNNs to reinterpret results and adjust structures accordingly. This resulted in high-accuracy modelling being achievable for FM structures for the first time, a term previously only associated with TBM modelling when closely related homologs were available as templates. Moreover, these models required CASP assessors to develop a new high-accuracy score (DipDiff) to assess whether differences in GDT scores were due to model or native structure deficiencies, the models were also shown to be accurate enough for use in molecular

replacement techniques (Pereira *et al.*, 2021). However, one crucial question is whether AlphaFold2 would ever be able to predict novel structures, considering its reliance on MSA's.

1.5.6 AF2-Multimer

When DeepMind released AF2 as a Jupyter notebook on Google's Co-laboratory platform (Colab) in July 2021, following their GitHub code release a few days earlier, there were a number of attempts by developers to adapt the technology to model multimeric proteins. Developers were encouraged by realistic-looking interfaces in CASP14 AF2 models of monomeric structures, which were known to form quaternary interactions (Egbert *et al.*, 2021). Two popular techniques were to either add an amino acid linker (usually Glycine due to its potential flexibility) between dimer chains to simulate a dual domain tertiary structure (Ghani *et al.*, 2022) or to add a 32 amino acid long gap between the individual chains. The latter technique exploited some programming within the AF2 code which allowed a maximum 32 residue gap between relative amino acid positions meaning that an offset greater than this forced AF2 to treat the amino acid indexes as separate chains (Mirdita *et al.*, 2022). Some success was seen with both of these techniques (Gao *et al.*, 2022) before a new version of AF2 called AlphaFold-Multimer (AFM) was released in late 2021. In the paper describing the release of this updated method (Evans *et al.*, 2022), it was confirmed that this version had been retrained on multimeric data and that superior performance had been achieved over the AF2 linker method. The results showed that 67% of heteromeric models in a 4433-model test dataset were scored as acceptable, with a DockQ score (Basu and Wallner, 2016a) of 0.23 or greater, 23% of which achieved higher accuracy defined as DockQ scores reaching the 0.8 threshold (see Section 3.1.3 for a full description of DockQ). The results were similar for homomeric targets with 69% of models ≥0.23 of which 31% were ≥0.8. Although there were no comparisons for the AF2 linker method using this dataset, comparative performance using a template-restricted dataset of 17 heterodimers was included. AFM achieved good models (DockQ ≥0.49) for 14 models, 6 of which met the ≥0.8 high-quality threshold compared to 9 (≥0.49) and 4 ≥0.8 for the AF2 linker method. While AFM achieved better results than the linker method it was not clear if this was consistent throughout both heteromer and homomer populations. What was clear from these results was that AlphaFold Multimer was not able to replicate the outstanding quality that AF2 had achieved for tertiary structures at CASP14 which may, to some extent, reflect the lower total number and variety of complexes available to make up datasets for training quaternary structure methods. Consequently, some structural models in existing datasets were originally generated by protein docking methods whose quality is lower than state of the art tertiary structure predictors (Chen *et al.*, 2023).

1.5.7 Other MSA-NN methods, RoseTTAFold and ColabFold

RoseTTAFold (RF) (Baek *et al.*, 2021) is a tertiary structure prediction method from the Baker laboratory which was inspired by the AF2 success at CASP14 and represented an evolution of their trRosetta method which used a neural network to predict inter-residue geometries and use them as modelling restraints in their popular Rosetta algorithm (hence “tr” for transform restrained) (Yang *et al.*, 2020). RF went a step further than trRosetta, using a three-track neural network to assess sequence data (1-D), a 2-D distance matrix and also the 3-D atomic coordinates. The method achieved similar but slightly lower performance to AF2 (Baek *et al.*, 2021) but was able to run on a modest server, although this could incur a time penalty. This method treats the MSA differently by using distinct aspects to represent different parts of the protein structure, rather like the 3D-shotgun method which focussed on different structural aspects at the scoring stage (Fischer, 2003). The consequence of this was that RF was able to model both mono and multimeric proteins, as whole MSAs did not necessarily need to equate to each single chain entity, thus, not only was multimeric modelling possible, an element of flexible backbone modelling was introduced in which chains are built in a complementary fashion rather than via single chain construction and docking, which is essentially the AFM way. For the initial iteration of RF there was a small quality gap between its models and AF2 models, however RF2 was redesigned to include a number of AF2 features and closed the quality gap to almost zero, with RF2 actually out-performing AFM on CASP14 target structures (Baek *et al.*, 2023). This study also suggested that RF2 was now outperforming AF2 on computing time, particularly noticeable for longer structures.

ColabFold (Mirdita *et al.*, 2022) is a reimplementation of the AF2 and RF algorithms which also runs on the Google Colab platform. It was developed by a consortium from Harvard university with the intention of making MSA-NN based modelling technology readily available to the wider community. The major difference between ColabFold and AF2 is that the former reduces large-scale database searches and therefore saves computing memory and runtimes by replacing JackHMMER (Johnson *et al.*, 2010) and HHblits (Remmert *et al.*, 2012) used by AF2 with the fast homology search algorithm MMseqs2 (Steinegger and Soding, 2017). This resulted in an estimated 40-60-fold faster search speed thus optimising MSA construction time. Rather than using the extensive databases used by AF2 (Uniref90 (Suzek *et al.*, 2015), Uniclust30 (Mirdita *et al.*, 2017), MGnify (Mitchell *et al.*, 2020) and BFD (Jumper *et al.*, 2021b)), MMSeqs2 searches the sequence identity-clustered UniRef30 (30% identity) database, the results of which are used to search a merged and clustered version of the BFD/MGnify databases which is also filtered to keep the 10 most diverse sequences in each cluster. The result is a user-friendly community resource described as achieving very similar results to the full AF2 installation in most cases (Mirdita *et al.*, 2022).

1.5.8 The potential downsides of MSA-NN modelling

The depth of an MSA is important, it's widely held that less than 30 hits leads to reduced accuracy (Jumper *et al.*, 2021b), although the quality of the alignments will also be a factor. This becomes important when proteins with no previously solved homologues or those underrepresented in the various databases are modelled. Training for all AI systems centres on the structures in the PDB, it is known that experimental structures represent only snapshots of many potential protein conformations and also are a product of their experimental preparation methods (like crystallisation), which may not represent cellular conditions. These issues introduce a margin of error into PDB structures, which is likely to be repeated by predictive AI methods. Larger assemblies tend to cause particular problems with either amino acid number exceeding system limits or multiple chains extending GPU memory use beyond capacity. This was seen during CASP15, where many of the larger complexes (>2000-3000 total residues) required additional human input (Ozden *et al.*, 2023). Lastly, the NN modelling processes that the methods use do not appear to be shedding any light on folding pathways or the underlying mechanism linking primary structure to tertiary or quaternary structure (Outeiral *et al.*, 2022).

1.6 Model quality assessment (MQA) – the philosophy and intention

The increase in accuracy attained by AlphaFold at CASP14 was only quantifiable due to the existence of model quality assessment programs. In this case model quality was assessed absolutely, that is by reference to experimentally determined structures which were deemed of sufficient resolution (<2Å (Kryshtafovych and Fidelis, 2009)) to act as a proxy for the native conformation. Scores obtained by this method are referred to as observed quality scores and are deemed to be accurate in describing the model in terms of its similarity to the native structure. However, there are still two potential sources of error, even with observed scores. Firstly, is the question of how accurate the experimental structure is, in terms of resolution but also in terms of how representative the crystalline (or other) image is of the native biological conformation. As the saying attributed to George Box goes, "*All models are wrong, some are promising*" and there may be known flexibility in the native protein, which would render any snapshot as unrepresentative or conformational anomalies in the experimental structure that are attributable to crystal packing artefacts, for example. Secondly, is the question of what the MQA program actually measures and whether it is sufficient for the model to be proclaimed accurate. The former point is one of philosophy concerning the acceptance of experimental structures as the ground truth and will be difficult to resolve until developing experimental technologies like TR-EM and Cryo-ET allow true native conformations to be sampled or indeed developing computational methods like RoseTTAFold diffusion (Watson *et al.*, 2023) or hybrid AF2-NMR methods (Ma *et al.*, 2023) allow modelling which is independent of crystal structures. Advances

on this scale would represent a further step change in protein modelling accuracy and appear, as yet some way distant. However, the latter point relating to which aspect of the model to measure has been one that the protein modelling community has been striving to resolve since MQA methods were conceived.

The intention of observed MQA is many fold; to allow the objective assessment of models for purposes of ranking to find the most representative model from a decoy group; to allow fair comparisons between models from different sources; to benchmark the accuracy of modelling techniques in general and to allow an objective assessment of the usefulness of models or local parts of models for applied research as briefly described in Section 1.2. From the earliest CASP competitions, observed MQA has been effected using a variety of quality measures with the aim of providing a balanced overview of quality. These measures were mostly based on the superposition scores root mean square deviation (RMSD) and the Global Distance Test (GDT) with the template modelling score (TM-score) (Zhang and Skolnick, 2005) as well as the superposition free local distance difference test (lDDT) (Mariani *et al.*, 2013) being introduced later. The exact definition of different scores is important and detailed descriptions of these and other scores are included in Chapter 3, Section 3.1.2 and also in Appendix 1. In an attempt to give a balanced assessment of model quality, CASP assessors have routinely used an assessor's formula to combine individual scores, for example, the CASP11 assessors formula combined two GDT-based scores with lDDT, SG (sphere grinder score, (Kryshtafovych *et al.*, 2014)) and a weighted contribution of the stereochemical-based score MolProbity (Chen *et al.*, 2010) to give a final ranking score for submitted models. In order to visualise how different scores differently represent the models, one superposition score (TM-score) and one superposition-free score (lDDT) are described below.

The TM-score is based on the TM-align (Zhang and Skolnick, 2005) algorithm, which is a pairwise alignment of Cα atoms in the protein chain backbone. It works via a number of iterative alignments, starting with a secondary structure alignment based on the dictionary of secondary structure of proteins DSSP (Kabsch and Sander, 1983) distance definitions, followed by a gapless and then a gapped threading algorithm to complete the initial alignment. The structure is then subject to a number of rotational and scoring rounds until no further improvement in alignment score is achieved. The TM-score is calculated as the distance between residue pairs normalised by a factored chain length value which means that the score is not length dependent. The local Distance Difference Test (lDDT) is designed to be super-position-independent and is calculated as the fraction of contacts between atoms of different residues present in the model that are also present in the reference structure. For example, if a contact exists between atoms of residue A and B in the reference structure and is also evident in the

model (regardless of any difference in the actual orientation) the contact is said to be conserved as shown in Figure 1.9B where the greyed structure represents an alternative formation in which the contact is still present. The global lDDT score is a mean of all residue-level scores.

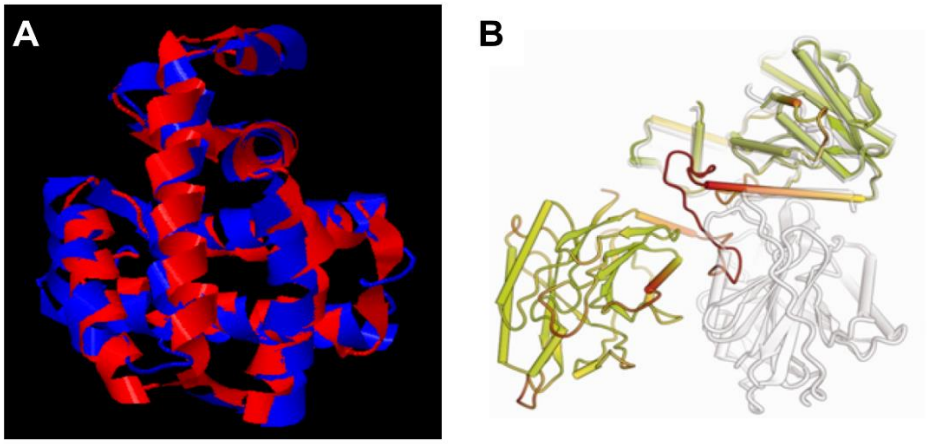


Figure 1.9 Two contrasting methods of scoring. A. The superposition alignment by TM-align on which TM-score is based, where scoring relies on the closeness of the alignment. **B.** The superposition-free distance score lDDT showing that the lower domain will score equally whether it occupies the coloured position (lower left) or the greyed position (lower right) with respect to the upper domain. (TM-align image adapted from the example page at https://seq2fun.dcmb.med.umich.edu//TM-align/example/, lDDT image adapted from (Mariani *et al.*, 2013)).

For a protein with well modelled domains but incorrect inter-domain orientation, for example, scoring by TM-score may heavily penalise the model on the basis of misalignment, whereas the lDDT score could remain consistent regardless of differences in the orientations of the domains. The best score would depend on whether the overall shape or the local domain was considered more important. Again see, Chapter 3, Section 3.1.2 for details of the score calculations.

The Critical Assessment of PRedicted Interactions (CAPRI) group (Janin *et al.*, 2003), a similar competition to CASP but focussed on PPI and protein quaternary structure, uses a similar approach of multiple quality indicators but limits them to three scores called Fnat, LRMS and iRMS. Fnat is defined as the fraction of native interface contacts observed in the model, LRMS is the root mean square deviation (RMSD) of the chain denoted the ligand (smaller chain of a complex) after superposition of the larger chain and iRMS is the RMSD between interface residues seen in the native structure compared to the model (again this definition is included in Chapter 3, Section 3.1.2).

In this way the protein modelling community has used the observed scores from successive CASP and CAPRI competitions, as well as the on-going server competition CAMEO (Contin-

uous Automated Model EvaluatiOn) (Haas *et al.*, 2018) to benchmark the quality of their models, assess new modelling technology and drive the improvement in modelling that has taken place since 1994.

1.6.1 Predicting model quality and MQAPs

Rating and benchmarking the quality of models against experimental structures is a valuable exercise, but if computational modelling is to truly fill the sequence-structure gap and create reliable models of proteins with no experimentally solved homologues, MQAPs must be able to predict a model's accuracy equally reliably *without* a reference structure. This is the problem that the CASP blind estimation of model accuracy (EMA) competition has been attempting to address since 2006 (CASP7) (Kwon *et al.*, 2021).

There is an important difference between MQAPs and quality assessment scores; MQAPs are programs developed to predict model quality using one or many individual quality assessment scores. MQAPs can be categorised in a number of ways; one popular method is by the number of models they require in order to formulate an accurate score. Thus, MQAPs can be separated into single-model and consensus or clustering methods. Single-model methods use molecular scoring functions which they apply to each model individually, making them suitable for scoring one or only a few models. Some use physiochemical features, often referred to simply as physics-based methods such as Ramachandran torsion angle constraints, bond lengths, environment compatibility (hydrophobicity or solvent accessibility) or structural features (such as secondary structure compatibility), to determine a model's conformity to expected values. Others, such as VoroMQA (Olechnovic and Venclovas, 2014) rely on a single structural feature, in this case the distance between Voronoi cells defined using van der Waals radii. Consensus or clustering methods tend to focus on pairwise distance comparisons and often employ a mix of proprietary and established quality scores from which a consensus is calculated. The algorithms usually measure distances between residue pairs and compare them on an all against all basis. The results are then clustered by distance similarities and the best models are scored on the basis that recurrent patterns are likely to be more like native proteins than random occurrences (Kryshtafovych and Fidelis, 2009). Although these methods have been described as performing better than single-model methods (Pages *et al.*, 2019), their efficacy relies on the model population size and quality, with accuracy decreasing with fewer or less diverse models. Notable proponents of this method have been Pcons (Lundström *et al.*, 2001) and ModFOLD (McGuffin *et al.*, 2021) and, despite their success, one long-standing problem has been finding the optimal weighting and combination of quality measures to create a representative consensus score (Kryshtafovych and Fidelis, 2009). Two adaptations of the clustering category are the quasi single-model and hybrid methods. Quasi single-model methods are

designed to retain the accuracy of clustering methods while allowing both multiple and single-model inputs. They do this by creating their own set of reference models to act as a set of comparators (McGuffin *et al.*, 2013) and work well as long as the decoy set are diverse enough to allow differentiation. Hybrid methods are consensus methods combining a range of individual approaches such as clustering, single-model, traditional stereochemical measurements or ML, with the aim of creating a consensus score which is more accurate than any of the individual contributing scores (Chen and Siu, 2020).

One alternative method of MQAP classification is whether local (residue level) scores are output in addition to the global score relating to the whole model (Chen and Siu, 2020). This distinction became increasingly relevant with the rise of modelling by contact prediction methods from around 2010. These methods were shown to have plateaued at only 20% precision (FM modelling) up to CASP11 (2014), but increased to 40% in CASP12 and again to 70% by CASP13 (Kryshtafovych *et al.*, 2019). The authors attributed the initial increase in success to better interpretation of transitivity (linking two proteins not previously considered homologous via a shared intermediate (Bolten *et al.*, 2001)) shown by the MSA, while the second increase was attributed to the rise of ML techniques, particularly deep neural networks (DNN). This, shift in focus was largely responsible for the increase in popularity of observed quality assessment by lDDT, which is sensitive to distances in the local environment. As single-model predictive MQA methods were considered to estimate lDDT more reliably than clustering methods (Kwon *et al.*, 2021), an increase in single-model methods was also seen at this time, demonstrating how different scores and methods vary with prevailing modelling technology. Latterly, with the advent of AF2, the lDDT score and TM-score have both seen renewed popularity to complement AF2's predicted quality measures, plDDT and pTM.

1.6.2 MQA for multimeric proteins

An important and continuing problem for accurate multimer modelling remains reliable MQA to rank and select the highest quality predicted models (Kinch *et al.*, 2021). This statement refers to the lack of reliable independent predictive quaternary structure MQAPs prior to CASP15, with the possible exception of the ProQDock program (Basu and Wallner, 2016b) and VoroMQA, the latter designed for tertiary structures but able to assess multimers (a more comprehensive history of early multimeric MQA is given in Chapter 3, Section 3.1.1). In 2020, however, there was a gradual change in focus from tertiary structure to quaternary structure MQA, following the success of AF2 at CASP14 prompting some groups to declare that the tertiary structure prediction problem was essentially solved (Kwon *et al.*, 2021). At this time, and continuing the trend favouring single-model methods, Han *et al* offered a classification of MQAPs as either physical energy, statistical potential or machine learning (ML) based (Han *et al.*, 2021). In

Han's definition, traditional distance-based and physical energy methods were essentially consigned to the past along with Boltzmann statistical-potential methods, which had suffered issues with defining a hypothetical reference state to compare observed frequencies (Rykunov and Fiser, 2010)). In place of these and other multi-model methods Han *et al.* argued in favour of graph-based neural network (GNN) technology. Indeed single-model GNN based methods featured highly at CASP15 where both VoroIF-GNN (Olechnovic and Venclovas, 2023), an updated version of the Voronoi tessellation program VoroMQA and GuijunLab-RocketX (Liu *et al.*, 2023), using the latest version of DeepUMQA (Guo et al., 2022) an Ultrafast Shape Recognition-based system, both used the technology to predict local residue contacts well (Studer *et al.*, 2023). Two other notable deep learning methods used deep neural networks to understand PPI interfaces rather than GNNs. These were DeepRank (Renaud *et al.*, 2021) and MULTICOM_qa (Cheng *et al.*, 2023) and, whereas DeepRank did not feature at CASP15, MULTICOM_qa used a hybrid pairwise similarity method linked to interface deep learning to rank first in the CASP15 global score category.

As mentioned in Section 1.5.6, the structural models in many existing datasets used for training quaternary structure methods were generated by protein docking methods whose quality is lower than state of the art tertiary structure predictors, for example (Chen *et al.*, 2023) and training deep learning MQA methods on these datasets could lead to lower accuracy on with higher quality structures. The McGuffin group were consequently somewhat circumspect about the wisdom of relying on deep learning exclusively, favouring the view that it is not possible to describe the quality of a protein or protein complex model by a single measure (Kwon *et al.*, 2021). Therefore, the ModFOLDdock methods were designed to increase prediction accuracy by using a combination of individual established and bespoke algorithms. This approach focussed on the all-important weighting of a calculated consensus score from a range of single-model and clustering methods as well as an element of deep-learning input (Edmunds *et al.*, 2023). The relative success of this approach at CASP15, compared to the other methods described in this section is covered in detail in Chapter 4.

1.7 Original hypothesis and project objectives

This project has evolved over the five or so years since its beginning in late 2018, however the fundamental philosophy, aims and principles underpinning it remain largely unaltered. The philosophy has been that the whole is greater than the sum of its parts, meaning that optimal combinations of methodologies of proven quality are likely to be significantly better than any single method. The overall aim has always been to create easy-to-use pipelines for modelling and quality assessment, bringing together state-of-the-art technologies in publicly available servers, which provide better performance than any single constituent method. The principles

governing this aim have been to survey and critically analyse the available technology and to use blind competition benchmarking to objectively assess performance progress.

In 2018 a gap in the protein multimer modelling landscape was identified - there were few publicly available multimer or quaternary structure modelling methods, which didn't require the installation of specialist docking software. There were even fewer independent multimer model quality assessment programs (see Chapter 3 for fuller account of the multimer modelling landscape). Therefore, in accordance with the above, the specific aims of the project became:
1. To investigate methods for the improvement of MultiFOLD, an unpublished multimer modelling pipeline to include, but not limited to, the concept of refinement to reduce atomic overlap and clashes and thus improve interface quality.

2. To analyse the performance of and optimise ModFOLDdock MQA scoring routines in order to close the gap between predicted and observed score accuracy.

It was reasoned that observed scores could be used to continually assess improvements in both MultiFOLD model quality and ModFOLDdock predicted score accuracy between blind benchmarking experiments.

The radical improvement in modelling accuracy achieved by AlphaFold2 at CASP14 in 2020 represented a new benchmark for state-of-art tertiary structure modelling. It was not clear, however, whether this accuracy level could be reproduced for multimeric proteins. Although the fundamental aims of the project did not change, new tools such as ColabFold were now available with which to achieve them, although the baseline for modelling accuracy and predicted quality assessment had now increased substantially. During the long process of experimental modelling that ensued it was noticed that multimer modelling using AFM was less accurate than that achieved for tertiary modelling with AF2. In addition, it was noticed that the AFM accuracy self-estimates (ASEs) were similarly inaccurate in some cases (see Chapter 5 for both). To address these continued accuracy gaps the aims of the project were extended and now became:
1. To investigate methods to improve MultiFOLD to include the concept of refinement to produce a measurable improvement over baseline modelling using AFM alone.
2. To optimise ModFOLDdock MQA scoring routines in order to close the gap between predicted and observed score accuracy and also beyond the accuracy of AFM plDDT and pTM scores.

CHAPTER 2

MultiFOLD: Improvement of protein tertiary and quaternary structure modelling using the AlphaFold2 recycling process

Work presented in this chapter has been published in the following paper:

Improvement of protein tertiary and quaternary structure predictions using the ReFOLD refinement method and the AlphaFold2 recycling process. *Adiyaman R., Edmunds N S., Genc A G., Alharbi S M A., & McGuffin L J.* Bioinformatics Advances, Volume 3, Issue 1, 2023.

Individual author contributions are as follows.
Adiyaman R: ReFOLD4 refinement.
Edmunds N S: AlphaFold2 recycling proof of concept work using tertiary structures.
Genc A G: Extension of AlphaFold2 recycling to quaternary structures.
Alharbi S M A: Rendering of images in PyMOL.
McGuffin L J: Overview and guidance from conception to publication.

Cited as (Adiyaman *et al.*, 2023) in the text.

2.1 Background and historical context

This chapter describes the development of MultiFOLD from a hybrid-docking pipeline to an AI-based tool incorporating scoring of multiple alternative models followed by a recycling-refinement routine designed to improve model quality beyond levels attainable by AlphaFold2 alone.

As recently as early 2020, template-based modelling (TBM) and rigid grid-based docking methods remained the mainstay of multimer modelling pipelines. These had been in existence since at least 2004 (Pierce *et al*., 2014) and much of the intervening research had been concerned with the use of so-called data-driven approaches. TBM methods had shifted focus from early methods which considered lower resolution techniques like SAXs or cryo-EM to provide clues to the overall shape and structure of target multimeric proteins (van Dijk *et al.*, 2005) towards interface prediction methods (Xue *et al.*, 2015). These included use of fragment libraries such as Swiss-Model (Waterhouse *et al*., 2018) and interface libraries like ProtCID (Xu and Dunbrack, 2020) intended to improve TBM accuracy and guide interface identification for docking routines. At this time the MultiFOLD pipeline was described as a hybrid-docking modelling tool incorporating both TBM and docking technology (McGuffin *et al.*, 2020), although it was necessary to run each process individually and manually collate results to form a single model population.

2.1.1 The early MultiFOLD pipeline used for CASP13 (2018)

The McGuffin group's method for the creation, scoring and ranking of quaternary structure models can be simplified into four phases; template identification, tertiary structure modelling, oligomeric modelling and quality assessment (McGuffin *et al.*, 2018). In order to maximise the number of tertiary models feeding into the oligomeric pipeline, a dual input approach was used by pooling CASP server models with our own IntFOLD (McGuffin *et al.*, 2019) tertiary structure models. The CASP13 MultiFOLD modelling pipeline is summarised in Figure 2.1.

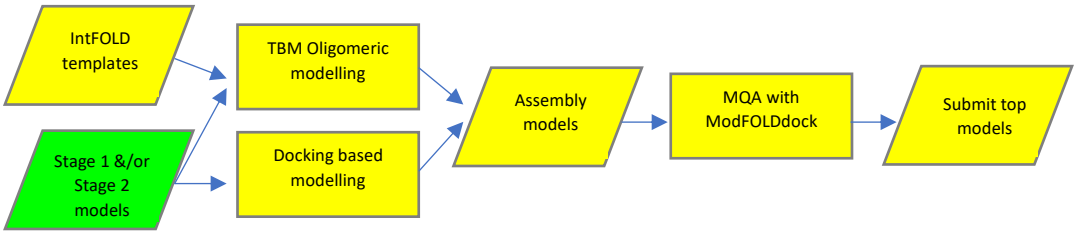


Figure 2.1. An overview of the MultiFOLD CASP13 oligomeric modelling process. This shows how the templates identified by IntFOLD, and both the IntFOLD and CASP server tertiary models (highlighted in green), fed into the TBM and docking pipelines.

Stage 1 models were created from sequence via a two-step process using the IntFOLD server and this was followed by model ranking and selection rounds. In the initial step, tertiary

templates were identified using six individual fold-recognition programs and the 8 threading programs in the LOMETS package (Wu and Zhang, 2007) before being quality assessed with ModFOLDclust2 (McGuffin and Roche, 2010). In the second step, an initial model was built from the two top-ranked templates which was iteratively compared to models built using all other templates. The best model was then selected on amino acid coverage and the process was performed twice more with a second ModFOLDclust2 scoring round. Stage 2 models underwent an additional refinement and re-ranking step in which I-TASSER (Yang and Zhang, 2015) and HHpred (Soding *et al.*, 2005) were used to build three separate models each. These were added to the group of models and fed into a loop of molecular dynamics based refinement by ReFOLD (Shuid *et al.*, 2017) and ranking by ModFOLD7_rank. The final top ranked tertiary model (or models for heteromers), along with the list of IntFOLD templates was then input into the oligomeric modelling process.

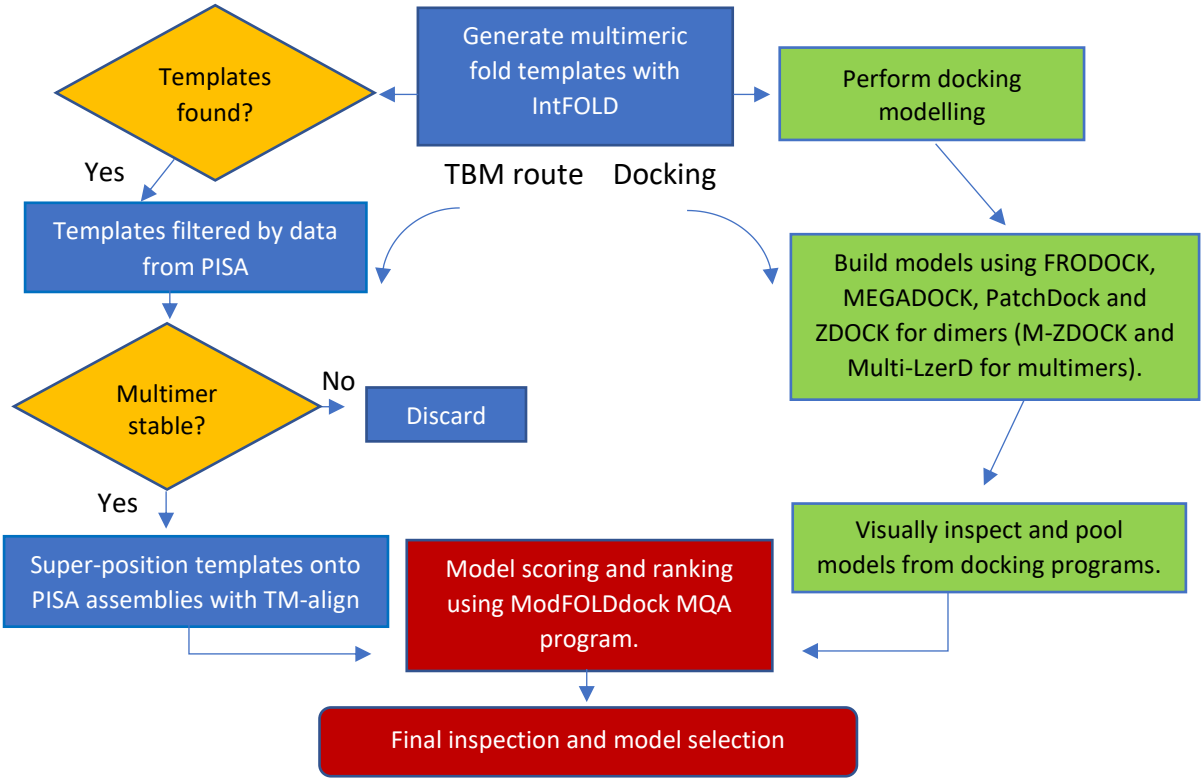


Figure 2.2. A flowchart showing the oligomeric TBM and docking routes within the CASP13 MultiFOLD pipeline. Decision points in the Docking and TBM pipelines are represented by rectangles in Figure 2.1. Docking was always performed but TBM may not have always produced suitable templates.

As shown in Figure 2.2, the PDBe PISA database (Krissinel, 2010) was referenced to validate the templates for the TBM process. For each template verified as stable, quaternary structure models were built by alignment using TM-align (Zhang and Skolnick, 2005) using the top tertiary structures identified earlier in the process. In the complementary docking process, the same tertiary structures were submitted to a range of established docking programs,

increasing the number and variety of oligomeric models available. These were, ZDOCK (Pierce *et al.*, 2014), MEGADOCK (Masahito Ohue *et al.*, 2014), FRODOCK (Garzon *et al.*, 2009), PatchDock (Schneidman-Duhovny *et al.*, 2005) and LZerD (Venkatraman *et al.*, 2009) for dimers and M-ZDOCK and Multi-LZerD for multimers. The top docking models (determined upon visual inspection) and the TBM models were then pooled into one population which was then scored and ranked using an earlier version of ModFOLDdock. Top ranked models were visually inspected in PyMOL (Schrödinger, 2018) for obvious clashes or alignment errors prior to submission.

2.1.2 Overall performance at CASP13

Since 2014 (CASP11) the competition has included a quaternary structure or *assembly* category. CASP13 included 42 assembly targets comprising 30 homomers (18 dimers, 9 trimers, 1 tetramer, 1 hexamer and 1 octamer) along with 12 heteromers. CASP13 ran from April to August 2018 and native structures and scores were revealed during the conference in December 2018 (https://predictioncenter.org/casp13/index.cgi).

CASP group rankings are based on a calculated overall Z-Score which is a combination of Z-scores for four CASP measures; F1 (interface contact score, ICS), Jaccard (interface patch score, IPS), lDDT-oligo and GDT_TS (definitions of scores can be found in Appendix 1). Z-scores are based on the standard deviation (SD) from the mean and in a model population the Z-score is calculated as:

$$Z = \frac{x - \mu}{\sigma}$$

Where Z is the standardised Z-score, x is the observed value (in this case the model score), μ is the mean value (mean score for the sample of models being considered) and σ represents the standard deviation (SD) for the sample. Therefore, the Z-score is a measure of distance from the mean in SD units where 0 represents the mean value while 2 would represent a model in the outer 5% of the distribution (assuming the rule for normal distribution where 1 SD accounts for 68% and 2 SD, 95% of results).

CASP reduce Z-score bias by first, only including Z-scores > 0.0. A higher Z-score therefore always means a better than average model. Secondly, rankings are calculated for both summed and averaged Z-scores as not all groups submit models for all targets. Whereas a summed Z-score potentially favours groups submitting models for more targets, average Z-score may disadvantage groups who attempt a greater number of difficult targets. The final rankings are given in terms of summed Z-score. Figure 2.3, below, displays summed Z-score results calculated for CASP13 assembly modelling.

The McGuffin group submitted models for homomeric complexes only and Table 2.1 shows that the group was ranked between 12th and 16th depending on Z-score calculation, the only exception occurring for Hard targets where the group was ranked 6th by Average Z-score, although it must be noted that models were submitted for only 4 out of the 13 hard targets. Overall, the group was placed 14th by summed Z-score as shown in the final ranking plot in Figure 2.3. See Appendix 2 for definitions of CASP difficulty categories and a list of individual targets and scores for models submitted at CASP13.

Table 2.1. McGuffin group multimeric modelling Z-scores by CASP13 target difficulty. Highlighted scores show the best ranking achieved by the McGuffin group per difficulty rating. "Max. score" is the maximum score attained by any group in the competition.

Target Difficulty	Measure	Score	Rank	Max score
Easy	Sum Z-score (>0.0)	**1.25**	14	10.77
	Average Z-score (>0.0)	0.12	16	0.89
Medium	Sum Z-score (>0.0)	**2.93**	11	12.95
	Average Z-score (>0.0)	0.20	15	1.05
Hard	Sum Z-score (>0.0)	1.85	12	12.23
	Average Z-score (>0.0)	**0.47**	6	0.96
All	Sum Z-score (>0.0)	**6.03**	14	35.97
	Average Z-score (>0.0)	0.20	16	0.86

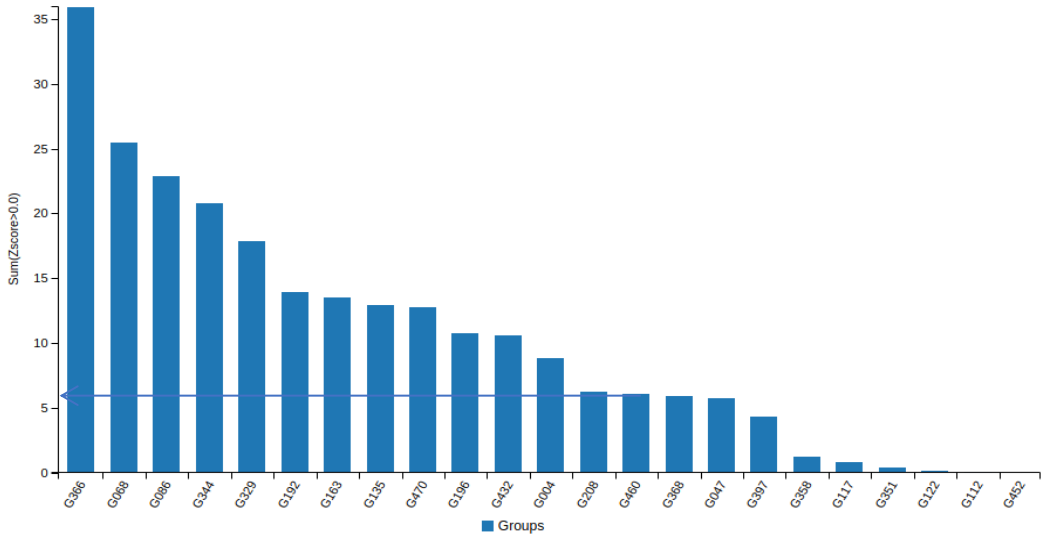


Figure 2.3. CASP13 final group rankings by summed Z-score for assembly modelling. The McGuffin group is G460 and the horizontal arrow shows the Z-score achieved in comparison to other groups. (Image taken from https://predictioncenter.org/casp13/ zscores_multimer.cgi). Group identities above McGuffin are (from 1st): 366:Venclovas, 068:Seok, 086:Baker, 344:Kiharalab, 329:D-Haven, 192:Elofsson, 163:Bates-BMM, 135:SBROD, 470:Seok-assembly(S), 196:Grudinin, 432:Seok-native-assembly(S), 004:YA SARA, 208:KIAS-Gdansk. (S=server group)

2.1.3 Analysis of CASP13 performance

Closer analysis revealed that the performance of both MultiFOLD modelling and ModFOLDdock model selection were variable. Figure 2.4 shows ModFOLDdock predicted and observed scores side by side for each target. Consensus6 scores are an unweighted mean of all six ModFOLDdock predicted scores, observed scores are an unweighted mean of five observed scores (see Section 3.1.3) calculated with reference to native structures.

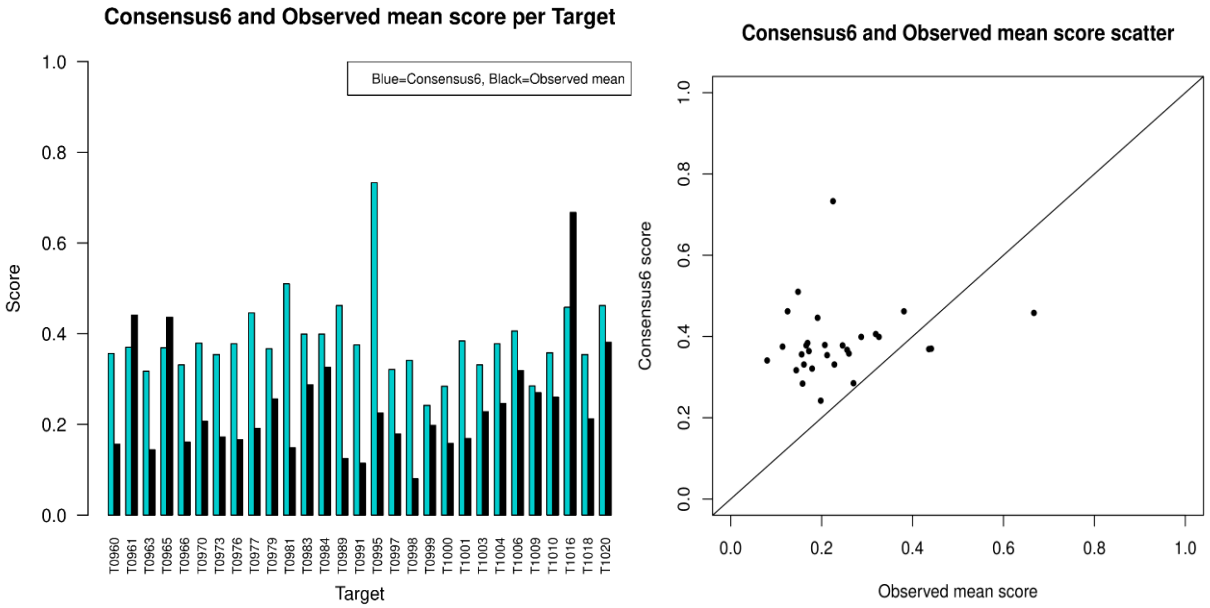


Figure 2.4. MultiFOLD CASP13 multimeric modelling performance as determined by predicted ModFOLDdock "Consensus6" score versus an observed mean score calculated retrospectively with reference to native structures. Left. A bar plot of ModFOLDdock Consensus6 (coloured light blue) versus mean observed scores (coloured black). **Right.** The same data as a scatter plot.

The first observation from both plots in Figure 2.4 is that the predicted and observed scores were generally below 0.5, suggesting a potential for improvement in many models. Any suggestion, however, that the predicted scores were good measures of the observed scores is dispelled by the magnitude differences between the bars representing the two scores in the left-hand bar plot as well as the clustering of most of the scores above the equivalence line in the right-hand scatter plot. Table 2.2 shows a more formal comparison of the results using a Wilcoxon signed rank test. This provides good evidence that the predicted scores were significantly greater than the equivalent observed scores.

Table 2.2. Wilcoxon signed rank test values for ModFOLDdock predicted versus calculated observed scores for MultiFOLD CASP13 multimer models. Significance is calculated at the 95% confidence level meaning P-values <0.05 are considered significant.

Scores compared	Independence and distribution symmetry	p-value
Predicted and observed	Paired; 2-sided test	**4.37×10^{-05}**
Predicted and observed	Paired; 1-sided test, predicted > observed	**2.18×10^{-05}**

There are two notable examples which also serve to highlight the differences between predicted and observed scores. The model for T1016 was underpredicted with a consensus score of 0.458 but achieved a mean observed score of 0.667. CASP official scores of 76.73, 0.689 and 0.693 for GDT_TS, lDDT and QS-score respectively all agreed that the model was underpredicted.

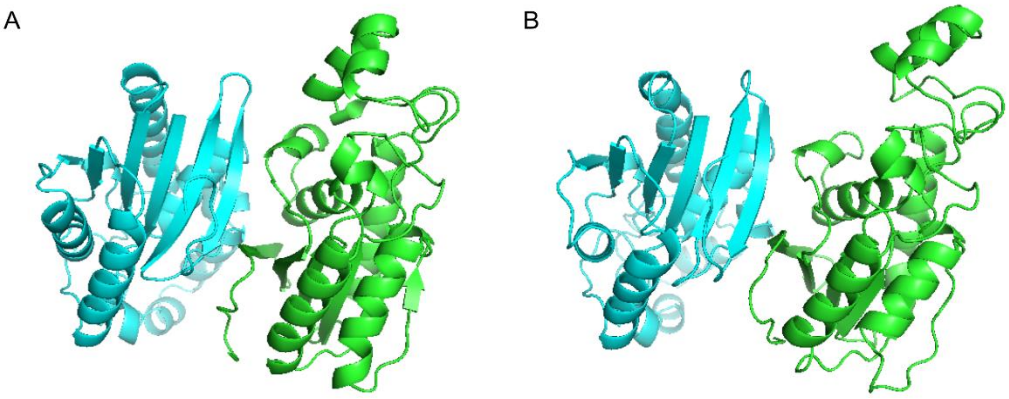


Figure 2.5. A comparative illustration of two models for CASP13 target T1016. A. The under-predicted MultiFOLD model. **B.** The equivalent CASP13 native structure. Models coloured by chain.

The model for T0995, by contrast, was scored highly at the prediction stage (0.733) but turned out to suffer problems with global orientation and interface accuracy. These were confirmed by an observed score of 0.225 and low CASP scores of 10.40 for GDT TS and 0.018 for QS-score.

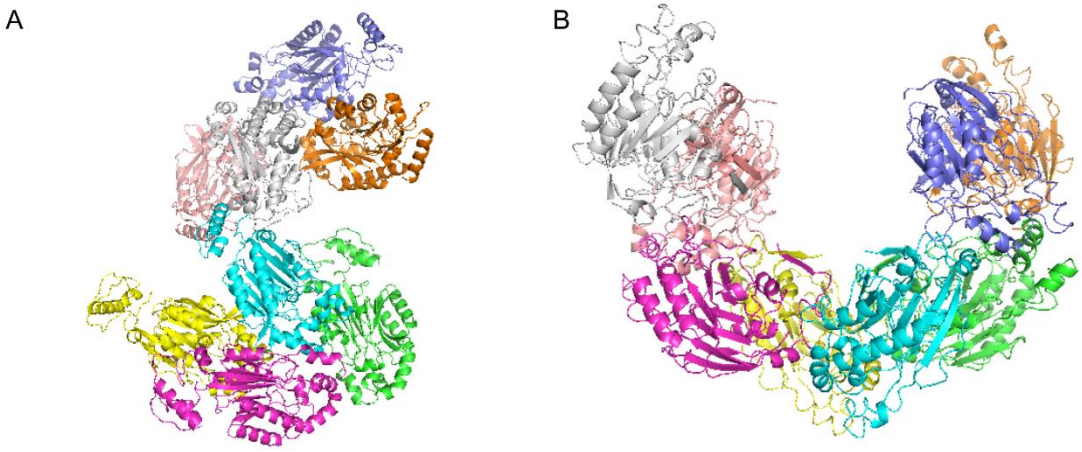


Figure 2.6. A comparative illustration of models for the CASP13 homomeric target T0995 (categorised as A8). A. The MultiFOLD model. **B.** The CASP native structure showing 8 monomers as part of the cyclic homo-18-mer Cyanide dihydratase from Bacillus pumilus C1 variant (PDB 8C5I).

2.1.4 Overview of CASP13 performance

Figure 2.4, along with the above two examples shown in Figures 2.5 and 2.6, highlight inconsistencies in ModFOLDdock predicted scoring leading to inaccurate model ranking. This resulted in variable discernment between good and poorer models making it difficult to select the best model from the range of decoys. In addition to the models highlighted, there were also

a number of cases where a significantly better model (defined as having an observed score >0.1 compared with the submitted model) existed in the decoy population but it was not selected.

In terms of modelling, the results also show that MultiFOLD models tend to be rated more highly with the position independent lDDT score with an average of 0.501 (see Appendix 3 for supporting data) than with the interface implicit QS-score with an average of only 0.053. This suggests that the tertiary structure models constructed by IntFOLD and fed into MultiFOLD were of generally good quality, but that the TBM and docking oligomeric modelling procedures were either failing to orientate these correctly in the multimeric model or failing to produce a sufficiently accurate interface. Both of these problems are typical of rigid-body systems where monomer construction and docking or alignment are performed in separate steps. Further examples of CASP13 models compared to their native structure can be found in Appendix 4.

2.1.5 An exploratory investigation into quaternary structure refinement

One method for eliminating minor errors in protein models is to use refinement techniques (see Introduction 1.4.1). Although refinement can have a variable effect on tertiary structure improvement, sometimes leading to a degradation in quality (Fan and Mark, 2004;Terashi and Kihara, 2018), there have been some positive results, particularly seen with FM models (Adiyaman and McGuffin, 2019), and it was considered that, despite a lack of documented support for quaternary structure model refinement at the time, there were likely to be some advantages to this approach.

GalaxyRefineComplex (GRC) (Heo *et al*., 2016) is a molecular dynamics-based side-chain repacking algorithm that was one of only a few refinement programs designed specifically for protein complexes. In their description of the software the authors explain that many docking programs employ relatively low-resolution scoring functions to perform their orientation analysis in order to conserve computational power. This potentially leaves room for improvement in interface and chain orientation and in their paper, Heo *et al*. found that GalaxyRefineComplex compared favourably with established refinement programs such as RosettaDock and SymmRef in improving a ZDOCK benchmark set. The effect of GalaxyRefineComplex on our CASP13 models was investigated.

Sixteen CASP13 homodimers were selected for this exploratory study (T0965, T0966, T0970, T0973, T0976, T0983, T0984, T0997, T1000, T1001, T1003, T1006, T1010, T1016 and T1018) as it was estimated that refinement of these should prove less CPU intensive than higher order structures. For TBM models, three models per target were selected by observed score: the highest-scoring model, a mid-scoring model and the lowest-scoring model making a total of 48 individual models. In addition, the 100 docking models created for target T0976 were

investigated. These comprised 25 models each from the FRODOCK, MEGADOCK, PatchDock and ZDOCK programs. Again, to conserve processing power a test sample of 36 models was created by calculating the minimum, 25%, 50%, 75% and maximum quartiles using observed scores and then selecting models with scores within 10% of the minimum and maximum value and +/- 5% either side of each of the quartile values. A total of 36 docking models was selected. A working hypothesis was that refined models would show an overall improvement (measured by calculated mean observed score) compared to baseline models. A secondary consideration was whether improvement varied by model construction method (TBM or docking (FM)). To test this, the mean observed scores for the unrefined and refined population were compared using a paired Wilcoxon signed rank test (analysis was carried out in R version 3.6.3).

Table 2.3. Results of a paired Wilcoxon signed rank test on GRC refined versus original models using calculated observed scores. TBM models numbers 48 across 16 CASP13 targets and docking models numbered 100 for target T0976. Again the 95% confidence level was used and P-values <0.05 are considered significant.

Model models compared	Independence and distribution symmetry	p-value
TBM. Refined versus original models	Paired; 2-sided test	**0.328**
Docking. Refined versus original models	Paired; 2-sided test	**2.91×10^{-11}**
Docking. Refined versus original models	Paired; 1-sided test; refined > unrefined	**1.45×10^{-11}**

For TBM models the p-value of 0.328 obtained was above the accepted 95% confidence significance cut-off of 0.05 and therefore the null hypothesis must be accepted: TBM models were not significantly improved by refinement with GalaxyRefineComplex. However, for docking models the p-values obtained of 2.91×10^{-11} for a two sided test followed by that for a one sided test (refined scores are greater than unrefined scores) of 1.45×10^{-11} meant that this time the null hypothesis can be rejected, and it can be concluded that T0976 docking models show a significant improvement upon refinement with GalaxyRefineComplex. Notwithstanding the difference in model populations, a possible explanation for this difference is that TBM models are based on templates of known proteins, and as such their atomic coordinates are less likely to result in clashes or disallowed torsion angles, leaving less room for improvement by physics-based refinement procedures. TBM models may therefore respond variably to refinement depending on the closeness of fit between the template and the native protein. This agrees with later findings (Adiyaman, 2021) showing that TBM models are often more difficult to successfully refine than FM models. In contrast, rigid body docking algorithms arrange individual chains without reference to a template. It is possible that, as Heo, Lee and Seok predicted, the low-resolution scoring functions employed in this process present an opportunity for refinement routines to improve docking models to a greater degree. Indeed, considering the absolute changes in best and median scores across the range of high, mid and low starting

model quality shown in Table 2.4, evidence exists for low-scoring models showing a greater margin of improvement.

Table 2.4. Differential improvement of the 100 T0976 docking models refined with GRC as measured by change in best and median observed score. Models were grouped by the quality of the unrefined starting model measured by mean observed score.

Starting model quality	High (>0.7)	Medium (0.7-0.3)	Low (<0.3)
Mean improvement in best score	0.03	0.29	0.21
Mean improvement in median score	0.06	0.04	0.12

Although this prospective study on its own, did not present strong enough evidence for successful quaternary structure refinement, the concept of the positive effect of refinement on FM and lower scoring models was influential in the design of the subsequent investigation into the AlphaFold2 custom template recycle pipeline explained in section 2.1.7.

2.1.6 Comparative analysis of CASP14 (2020) assembly modelling

CASP14 is arguably the most significant of the CASP experiments to-date due to the introduction of the AlphaFold2 software and its impact in increasing the accuracy of tertiary structure modelling. However, before this is considered, it is worth briefly describing the assembly modelling that took place. The competition was disrupted due to Covid-19 with all meetings taking place online and a truncated population of assembly targets (22, down from 42 in the previous round). Appendix 5 provides a full list and categorisation of all the assembly models submitted by the McGuffin group.

The methodology used to create, score and select McGuffin models for CASP14 was similar to that described for CASP13. There were, however, a number of minor differences; the MultiFOLD program code was reinstalled on the group server which involved a number of updates to underlying programs; FRODOCK (from v1.05 to v3.12), MEGADOCK (from v4.0.2 to v4.1.1) and a replacement version of Multi-LZerD. Secondly an additional scoring step using the Voronoi tessellation program VoroMQA (Olechnovic and Venclovas, 2017) was introduced alongside the older version of ModFOLDdock. The VoroMQA score was combined with the ModFOLDdock Concensus6 score to create a hybrid unweighted mean of both scores which was used as the primary ranking value. Assembly modelling results are summarised in Table 2.5 and CASP13 equivalent values are supplied in grey for comparison.

Table 2.5. McGuffin group CASP14 assembly modelling Z-scores by Target difficulty. CASP 13 scores in grey for comparison.

Target Difficulty	Measure	Score	Rank	Max score
Easy (Z-score >0.0)	CASP14 Sum Z-score	0.25	17	1.67
	(CASP13 for comparison)	1.25	14	10.77
	CASP14 Average Z-score	0.13	17	0.84
	(CASP13 for comparison)	0.12	16	0.89
Medium (Z-score >0.0)	CASP14 Sum Z-score	2.19	17	21.16
	(CASP13 for comparison)	2.93	11	12.95
	CASP14 Average Z-score	0.37	9	1.11
	(CASP13 for comparison)	0.20	15	1.05
Difficult (Z-score >0.0)	CASP14 Sum Z-score	1.39	17	8.44
	(CASP13 for comparison)	1.85	12	12.23
	CASP14 Average Z-score	0.28	15	1.17
	(CASP13 for comparison)	0.47	6	0.96
All (Z-score >0.0)	CASP14 Sum Z-score	3.85	19	31.27
	(CASP13 for comparison)	6.03	14	35.97
	CASP14 Average Z-score	0.30	16	1.17
	(CASP13 for comparison)	0.20	16	0.86

As can be seen the McGuffin group (220) ranked 19th by summed Z-score (see Appendix 5 for a bar plot of full CASP rankings) across all difficulty categories and 16th by average Z-score with a value of 0.3 (max for any group was 1.17). This compared with 14th and 16th respectively achieved in CASP13. Although the CASP14 rankings were lower it must be stated that the McGuffin group submitted models for only 13 homomeric targets (and was therefore naturally penalised by the summed Z-score value) and that the CASP14 targets were rated as more difficult due to their generally higher oligomeric state, including two large icosahedral structures and the classification of four structures in a new extreme category (Karaca, 2020). Competing groups scored an average of 0.86 for TM-score but only 0.38 for ICS (F1) score, showing that CASP14 assembly structures continued to present challenging interfaces for modellers.

In the CASP13 analysis, successful modelling was defined as having any model for a target scored as acceptable quality, i.e., QS-score > 0.1 and this was based on a slightly more stringent definition by the Venclovas group (Dapkunas *et al*., 2019) in their CASP13 analysis.

While in CASP13 this applied to 3/30 (10%) of the McGuffin group's models, in CASP14 it applied to 3/13 models, a higher rate of 23%.

2.1.7 Tertiary structure model quality improvement using AlphaFold2 custom template recycling

At CASP14 DeepMind's AlphaFold group submitted tertiary structure models which were widely accepted as a significant advancement in predicted model quality. They achieved high accuracy in both FM and FM-TBM classes with median GDT_TS scores of 87.0 and 92.4 respectively. Measured on a 0-100 scale, GDT_TS scores above 50 are considered correct in overall topology with scores over 75 considered to have mostly correct atomic coordinates (Kryshtafovych *et al*., 2019). These are clearly impressive values especially when contextualised against CASP13 where the average tertiary GDT_TS score for the best performing FM group (A7D) was 61.4 (Senior *et al*., 2019).

The AlphaFold group achieved these improvements using a machine learning model (AlphaFold2) based on two key factors; a multiple sequence alignment (MSA), used to highlight potential evolutionary relationships between amino acids, and a deep neural network (DNN) used to interpret them. While both of these concepts are familiar to the protein modelling community, AlphaFold2's success appeared to be their unique combination in the construction of an accurate residue distance map. This is then used to construct a detailed contact map which can be interpreted by further neural network (NN) input into a starting model to which the emerging structure can be compared. A schematic of AlphaFold2 is shown in Figure 2.7 below. However, there was a third interesting process within the AlphaFold2 model; the existence of a recycle route intended to allow repeated iterations of the partially completed

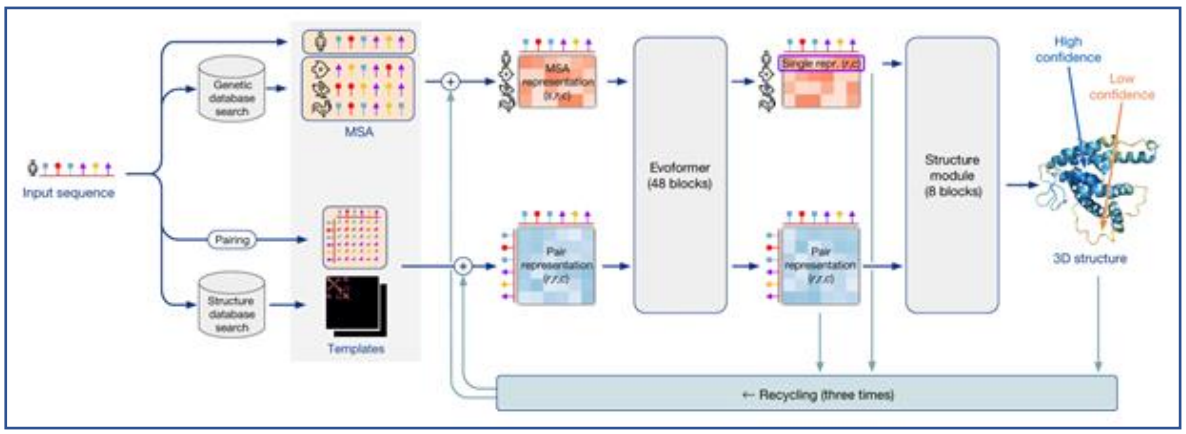


Figure 2.7. A schematic of AlphaFold2 architecture. Taken from (Jumper *et al.*, 2021). This shows how MSA, and pair representation data is processed and iterated via a recycling feedback loop.

proto model through the DNNs until no further improvement was detectable. One early-identified adaptation of this was to input electron density maps to enhance experimental

modelling accuracy (Terwilliger, 2022). Inspired by this, an alternative idea was proposed; that the recycle function actually represented a ready-made refinement loop.

ColabFold (Mirdita *et al.*, 2022) is a free open-source tool combining the AlphaFold2 algorithm with the fast alignment package MMseqs2 and hosted on Google Colaboratory. Neither early AlphaFold2 or ColabFold versions supplied a way of manually controlling the selection of templates feeding into the system. However, after an initial experimental version developed for use with Phenix software (Terwilliger, 2022), this function was added as a "custom template" input function to the main ColabFold software as shown in Figure 2.8. Thus, it became possible to add alternative models as "templates" straight into the recycle loop and its potential as a full model refinement tool became available for investigation.

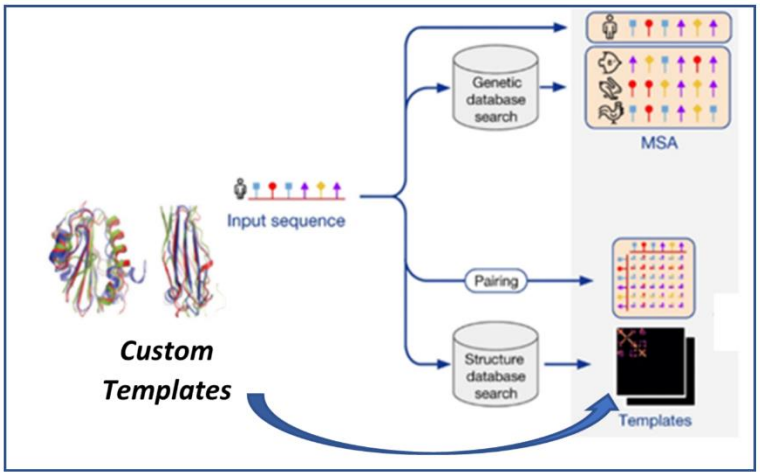


Figure 2.8. "Custom template" inputs into the AlphaFold2 architecture. Custom templates may now be manually added in addition to a template search. They are incorporated into the recycling loop shown in Figure 2.7. Image adapted from (Jumper *et al.*, 2021).

2.2 Objectives

The main hypothesis for this study was based on the supposition that full tertiary structure models of proteins could be successfully refined via the custom template option included in ColabFold by recycling through the NN architecture. The primary outcome was that repeated recycling would show improvement in these models beyond their starting quality with support for this viewpoint coming from the ColabFold team's own paper (Roney and Ovchinnikov, 2022), which postulated that the AlphaFold2 neural network had learned a potential protein folding energy function. Our primary hypothesis was:

H0: *Custom template recycling through ColabFold results in models no different in quality to the baseline models input as templates. H1: Custom template recycling results in models of higher quality than the baseline models which were input as templates.*

There were also three secondary considerations. The first of these was particularly relevant as it has been shown that the accuracy of AlphaFold2 predictions decreases markedly when it is not able to construct an MSA (Lin *et al.*, 2023;Roney and Ovchinnikov, 2022). If model improvement was seen from recycling in single sequence mode it would suggest that AF2 is using internal factors to effect those improvements. For this reason, a hypothesis was also constructed for the first of the secondary considerations.

1. Would similar improvement be seen for recycling in both single sequence and MSA modes?

H0: Recycling in single sequence mode produces models no different in quality compared to the baseline models used as templates. H1: Recycling in single sequence mode produces improvement in models similar to that seen for MSA mode.

2. Would improvement be seen in the official DeepMind AF2 competition models?
3. Would improvement be linear with recycle number and can an optimal number of recycles be determined?

If the primary outcome was proven, then the custom recycling strategy could be adopted as a key component to our CASP15 modelling pipeline, which would potentially confer an advantage over other state-of-the-art modelling software.

2.3 Materials and Methods

The goal of the project, of which this study formed the foundation part, was to improve upon the performance of AF2-Multimer quaternary structure modelling through custom template recycling. This study represented the initial proof of concept phase using tertiary structures which, it was reasoned, represented a simpler basis for testing and scoring.

To test the hypotheses, the study was designed around free modelling (FM) CASP14 tertiary structure targets to eliminate any confounding effects associated with TBM models. These were quality-assessed against their CASP reference structures to build a bank of baseline observed scores. The models were then submitted to ColabFold as custom templates for recycling through the algorithm. The resulting top-ranked models could then be rescored against the same reference models and the scores then directly compared to the baseline to assess any improvement in model quality. If successful, the study could be extended to quaternary structure models.

Two sets of CASP14 tertiary structure models were selected and the AlphaFold2 NN weights trained on pre-CASP14 data were used to recycle them. The selection of this particular dataset was important as the AlphaFold2 neural network was trained on models populating the PDB prior to CASP14 (2020), using more recent datasets risked introducing a bias into the modelling as AlphaFold2 could potentially have already encountered the structures. At the time, this represented the most suitable set of 3D models and great care was taken to ensure that the pre-CASP14 neural network weights were selected when using ColabFold as this guaranteed training had taken place on data predating these models.

The first set of models chosen were DeepMind's AlphaFold group (group 427) official CASP14 submissions, the rationale being that AF2 is essentially an FM modelling tool and according to the findings in section 2.1.5 it should be possible to refine these. Secondly, at the time these represented the best independently verified models available and so a technique able to improve these should be able to improve any other models available. The second set of models were selected from the five groups ranking immediately below group 427 in the CASP14 official rankings. These were viewed as lower-quality starting models which, again according to the findings in section 2.1.5, may allow greater potential for improvement by refinement. The argument that this procedure amounted to simple remodelling was controlled for in two ways; firstly by using the official AlphaFold group's CASP14 models with the AlphaFold2 model trained on pre-CASP14 data - the rationale being that any improvement in model quality must be due to recycling refinement, as the same software should not be able to improve upon its original model unless a different internal process is invoked. Secondly, by running parallel MSA and single sequence recycling (with all other parameters matched) any influence of an updated

MSA should be negated. Consistent improvement under these conditions would suggest that AlphaFold2 is refining the model supplied rather than ignoring the template and remodelling from scratch. Also, as a further control measure to ensure we were testing for the recycling effect only and no other refinement stages, the Amber relaxation option was not enabled.

2.3.1 Refinement of 16 CASP14 AlphaFold2 models

CASP14 rank 1 AlphaFold2 tertiary structure models were downloaded from the CASP Data Archive (https://predictioncenter.org/download_area/) along with their official results tables. Previous research has suggested that protein models created using template-based modelling (TBM) have a lower tendency for improvement compared to those created from free modelling (FM) methods (Adiyaman, 2021). Therefore, to maximise refinement potential, the 16 models submitted by the AlphaFold group in the CASP FM-only class were selected, matching those targets used in the ReFOLD4 analysis which was included a section A of the research paper (Adiyaman *et al.*, 2023).

Two structural alignment scoring methods - the TM-score (Template Modelling score) and the lDDT score (local Distance Difference Test) were used to provide performance metrics for model benchmarking. These scores, generated by downloadable versions of the TM-score (Zhang and Skolnick, 2004) and lDDT (Mariani *et al.*, 2013) methods, describe the backbone (TM-score) and local environment (lDDT) similarities of two protein models. Initially, the downloaded model and the experimentally determined native structure were compared to collect baseline TM and lDDT scores. Each AlphaFold2 model was then converted from pdb to mmCif format using https://mmcif.pdbj.org/converter which makes use of the RSCB PDB MAXIT suite of programs.

To eliminate ColabFold runtime errors the following workarounds were necessary. The template model name and job name needed to match with a maximum of 4 characters. The jobname was therefore always set as the numeric part of the CASP target, e.g. 10**. A chain identifier was required in column 22 of the PDB file prior to mmCif conversion. Also, to satisfy the AF2 algorithm's requirement for a creation date, the following information was added to the bottom of each template mmCif file:

*"loop_*
_pdbx_audit_revision_history.ordinal
_pdbx_audit_revision_history.data_content_type
_pdbx_audit_revision_history.major_revision
_pdbx_audit_revision_history.minor_revision
_pdbx_audit_revision_history.revision_date
1 'Structure model' 1 0 2020-06-17
2 'Structure model' 1 1 2021-01-20"

The converted mmCif model files were then submitted to the Google Colaboratory version of ColabFold (release 3, v1.3.0 [4-Mar-2022]) as custom templates along with their respective amino acid sequences. Each model was submitted eight separate times using the following recycle and MSA combinations; MSA: 1, 3, 6 and 12 recycles; Single sequence: 1, 3, 6 and 12 recycles. The following ColabFold settings were used.

- *Google Colab version: AlphaFold2 using MMseqs2.*
- *Template_mode: custom*
- *msa_mode: MMseqs2 (UniRef+Environmental) OR single_sequence*
- *pair_mode: unpaired+paired*
- *model-type: auto*[1]
- *num_recycles: 1, 3, 6, 12.*

The five models created by default for each individual ColabFold run were collected along with their predicted pTM and plDDT scores. Rank 1 models were then rescored with TM-score and lDDT programs in the same way as described for baseline scoring. In this way TM-score and lDDT scores obtained at baseline and for each recycle combination, along with the ColabFold-generated predicted scores (pTM and plDDT), could be directly compared.

2.3.2 Refinement of 47 CASP14 non-AlphaFold2 models

To explore the capacity for improvement of lower-scoring models, the same CASP14 targets were selected from groups making up the next five best-ranked groups beneath the AlphaFold group in the CASP14 rankings. These were (by rank): Baker (Gp.473, Av Sum Z-score=90.8), Baker-experimental (Gp.403, Av Sum Z-score=88.9), Feig-R2 (Gp.480, Av Sum Z-score=72.5), Zhang (Gp.129, Av Sum Z-score=67.9) and tFold_human (Gp.009, Av Sum Z-score=61.2). By comparison, AlphaFold2 (Gp.427) had an Av sum Z-score of 244. All groups had a total Domain count of 92 so the comparison of Sum Z-scores is valid. In addition, only models with a CASP TM-score of ≥ 0.45 were used, as those below this threshold cannot be guaranteed to have the same fundamental fold as the reference models (Xu and Zhang, 2010), so a total of 47 non-AlphaFold2 models were processed.

Models, scores and reference structures for these targets were downloaded from the CASP14 website and scored with the TM-score and lDDT algorithms in the same way as previously described in 2.3.1. ColabFold recycling using MSA was submitted to the same Google Colaboratory version of ColabFold (release 3, v1.3.0 [4-Mar-2022]) as used in 2.3.1, recycling using single sequence submissions (no MSA) was carried out using the same release (v1.3.0) of LocalColabFold (Mirdita *et al.*, 2022), which was installed on our own local server to avoid

[1] GitHub - DeepMind/AlphaFold: Open source code for AlphaFold2. ("selecting Auto from the model type monomer_ptm: This is the original CASP14 model fine-tuned with the pTM head, providing a pairwise confidence measure. It is slightly less accurate than the normal monomer model.")

Google Colab GPU restrictions adversely affecting our available modelling time. The equivalent LocalColabFold settings used were:

--num-recycle (1, 3, 6, 12) --msa-mode single_sequence --model-type auto --rank plddt --pair-mode unpaired+paired --templates --custom-template-path

LocalColabFold was therefore run with the following command format:

colabfold_batch --num-recycle 12 --msa-mode single_sequence --model-type auto --rank plddt --pair-mode unpaired+paired --templates --custom-template-path <path to mmCif files> <full path of fasta file> <full path of output directory>

Again, the five resulting models and their predicted scores for each ColabFold run were collected and rank 1 models were rescored with the TM-score and lDDT programs. The workflow for the methodology is summarised in Figure 2.9 allowing baseline and recycle TM-score and lDDT to be directly compared. Statistical analysis for all models was performed using R-studio version 1.3.1093.

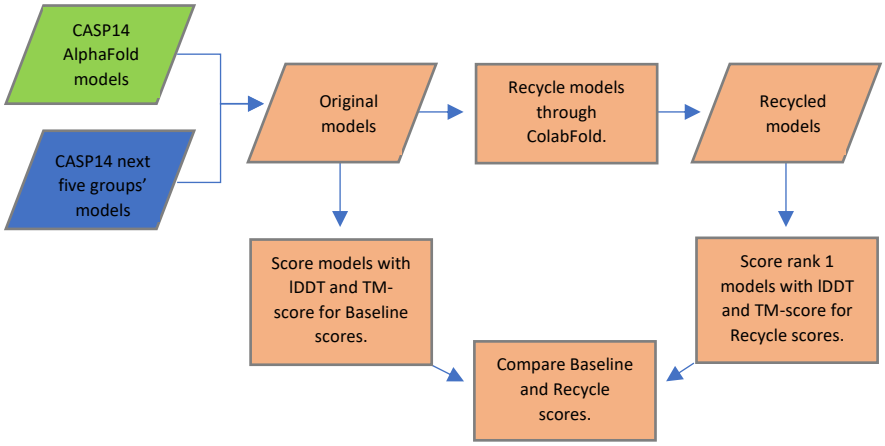


Figure 2.9. A workflow summary for the custom template recycling experiment. This shows how AF2 and other groups' models were recycled through ColabFold and assessed by comparison to baseline lDDT and TM-scores.

2.3.3 Treatment of quaternary structures

The processes described above were repeated for multimeric CASP14 targets by a co-researcher as part of the published collaborative study (Adiyaman *et al.*, 2023). This part used models from ten targets (H1045, H1065, H1072, T1032, T1054, T1070, T1073, T1078, T1083, T1084) for the top 5 performing groups in the CASP14 assembly category. These were Baker-experimental (Baek *et al.*, 2021), Venclovas (Dapkunas *et al.*, 2021), Takeda-Shitaka, Seok (Park *et al.*, 2021) and DATE. DeepMind did not submit multimeric models for CASP14 and so models were generated using AF2-Multimer (AFM) for the same targets to allow for common subset analysis. Baseline models were scored and refined using similar parameters described for monomers and observed scores were generated using MM-Align (Mukherjee and Zhang, 2009) for TM-score and OpenStructure (Biasini *et al.*, 2013) for IDDT-oligo and, additionally,

QS-score (Bertoni *et al.*, 2017). A short summary of key results for quaternary structure models will be included in the relevant results sections.

2.3.4. Study design.

This study can be categorised as having two factors; the type of recycling and the modelling software used to create the initial models. Both factors have two levels. For recycling these are either MSA or single sequence; for modelling they are either AF2 or non-AF2 models.

Table 2.6. The recycle experiment study design in terms of factors, level and treatment groups.

		Factor 1 – recycling	
	Levels	MSA	Single sequence
Factor 2 – modelling	AF2 models	MSA modelling, AF2 models	Single sequence modelling, AF2 models
	Non-AF2 models	MSA modelling, non-AF2 models	Single sequence modelling, non-AF2 models

Therefore, there are four treatment groups as shown in Table 2.6.

2.4 Results and Discussion.

2.4.1 Primary hypothesis. Repeated recycling shows improvement of models beyond their initial quality.

Both global TM-scores and lDDT scores were collected during the investigation, however, the analysis concentrated on the improvement in lDDT scores. The rationale being that, unlike TM-score, which is primarily associated with backbone configuration, lDDT is more likely to detect small changes in the local atomic arrangement which typically result from refinement.

Table 2.7 shows the significance values (p-values) obtained from the comparison between baseline and recycled lDDT scores for the 16 CASP14 AlphaFold2 (AF2) and 47 non-AlphaFold2 (non-AF2) tertiary models. P-values were calculated at the 95% confidence level using a 1-tailed Wilcoxon signed-rank test for non-parametric data between observed baseline lDDT scores (template model) and recycled lDDT scores (rank 1 output model). These have been calculated between baseline and each recycle and also between consecutive recycles. A p-value of ≤0.05 shows a significant difference between any two model populations, suggesting an improvement in quality for that number of recycles. Table 2.8a shows equivalent data for TM-scores calculated by the same method. It is worth restating here the primary hypotheses being tested:

H0. Custom template recycling through ColabFold results in models no different in quality to the baseline model input as templates. H1. Custom template recycling results in models of higher quality than the baseline models input as templates.

In Table 2.7, rows 1 and 2 indicate a significant improvement in quality for the AF2 models compared to baseline for recycle 1, 3, 6 and 12 as indicated by values in bold. Although significant improvement after 1 recycle is limited to models recycled in single sequence mode

Table 2.7. Calculated p-values for observed lDDT scores between baseline and recycled CASP14 AF2 and non-AF2 monomer models. P-values ≤0.05 are in bold.

Models	Recycle model	Base to 1 recycle	1 to 3 recycles	Base to 3 recycles	3 to 6 recycles	Base to 6 recycles	6 to 12 recycles	Base to 12 recycles.
AF2	MSA	0.187	0.756	**0.005**	0.043	**0.007**	0.351	**0.013**
	SS	**0.011**	0.954	**0.018**	0.124	0.059	0.637	**0.038**
Non-AF2	MSA	**$1.23x10^{-9}$**	**$1.21x10^{-8}$**	**$7.10x10^{-15}$**	**0.015**	**$7.10x10^{-15}$**	0.473	**$1.23x10^{-9}$**
	SS	**$1.70x10^{-9}$**	**$4.91x10^{-5}$**	**$1.50x10^{-9}$**	0.175	**$1.50x10^{-9}$**	0.587	**$1.40x10^{-9}$**

Key: Base = Baseline, SS=Single sequence. MSA=Multiple Sequence Alignment. The 1-tailed Wilcoxon signed-rank test P-values were calculated at the 95% confidence level using lDDT scores across 16 AlphaFold2 CASP14 top-ranked models (upper two rows) and 47 non-AlphaFold models from CASP14 targets (lower 2 rows).

Table 2.8a. Calculated p-values for observed TM-scores between baseline and recycled for CASP14 AF2 and non-AF2 monomer models. P-values ≤0.05 are in bold.

Models	Recycle model	Base to 1 recycle	1 to 3 recycles	Base to 3 recycles	3 to 6 recycles	Base to 6 recycles	6 to 12 recycles	Base to 12 recycles.
AF2	MSA	0.679	0.801	0.796	0.106	0.958	0.363	0.776
	SS	0.717	0.909	0.860	**0.033**	0.897	0.782	0.698
Non-AF2	MSA	**$1.42x10^{-14}$**	**$6.31x10^{-5}$**	**$4.36x10^{-12}$**	0.898	**$8.05x10^{-9}$**	0.240	**$2.40x10^{-12}$**
	SS	**$5.13x10^{-7}$**	**$7.61x10^{-5}$**	**$3.35x10^{-7}$**	0.033	**$1.98x10^{-7}$**	0.660	**$1.62x10^{-7}$**

Key: Base = Baseline, SS=Single sequence. MSA=Multiple Sequence Alignment. The 1-tailed Wilcoxon signed-rank test P-values were calculated at the 95% confidence level using TM-scores across 16 AlphaFold2 CASP14 top-ranked models (upper two rows) and 47 non-AlphaFold2 models from CASP14 targets (lower 2 rows).

and improvement after 6 recycles is limited to MSA mode, improvement after both 3 and 12 recycles is seen for both recycling modes. Similarly, rows 3 and 4 show that significant improvement in non-AF2 model quality compared to baseline occurred for both modes after all recycles. From 6 to 12 recycles there was no further significant improvement for either method. From these results the null hypothesis can be rejected for improvement in lDDT from baseline and it can be stated that recycling produces significantly higher quality models than baseline in the majority of cases (14 out of the 16 recycle phases across the two groups studied) in agreement with the alternative hypothesis.

Table 2.8a shows similar data for TM-score improvement. There are significant increases between baseline and 1 recycle and from 1 to 3 recycles for non-AF2 models, however there is no further significant improvement between 3 to 6 or 6 to 12 recycles. For the AF2 models there is no significant improvement in TM-score except for the one isolated result of 0.033 which occurred between 3 to 6 recycles in single sequence mode. From these results, with respect to TM-score, the null hypothesis must be accepted for AF2 models, but the alternative hypothesis may be accepted for non-AF2 models. This supports the rationale above that the superposition dependent TM-score based on the backbone is not sufficiently sensitive to relatively small changes in local atomic arrangement.

The data for quaternary structures shown in Table 2.8b also showed significant improvement upon recycling and, again, the improvement was greater for non-AFM than AFM models with a pattern that was not linear with recycle number. Specifically, non-AFM models showed significant improvement as measured by oligo-lDDT, TM-score and QS-score for baseline to all recycles for both MSA and single sequence recycling (with the exception of single sequence recycling measured by oligo-lDDT where significant improvement was only seen between 1 to 3 and 6 to 12 recycles). For AFM models, the best improvement was seen for TM-scores which showed significant improvement from baseline to 1 and 6 recycles for MSA recycling and for baseline to all recycles for single sequence recycling. For oligo-lDDT significant improvement was seen for MSA recycling (baseline to 16 and 12 recycles) but not to the same extent for single sequence recycling. Significant improvement by QS-score was seen in one isolated case (1 to 3 recycles) for AFM models. Despite this, absolute rates in terms of the percentage of models improved were calculated as 80% (MSA) and 30% (SS) for AFM models and 94% (MSA) and 64% (SS) for non-AFM models as measured by oligo-lDDT, 70% (MSA) and 80% (SS) for AFM models and 98% (MSA) and 82% (SS) for non-AFM models as measured by TM-score and 50% (MSA) and 30% (SS) for AFM models and 86% (MSA) and 60% (SS) for non-AFM models as measured by QS-score (Adiyaman *et al.*, 2023).

Table 2.8b Calculated P-values for observed oligo-lDDT (A), TM-score (B) and QS-score (C), for recycled AFM and non-AFM CASP14 multimer models. P-values ≤0.05 are in bold.

A								
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
AFM	MSA	$1.11x10^{-1}$	$5.20x10^{-1}$	$1.79x10^{-1}$	$7.68x10^{-2}$	**$4.16x10^{-2}$**	$1.11x10^{-1}$	**$5.15x10^{-2}$**
	SS	$9.97x10^{-1}$	**$4.16x10^{-2}$**	$9.37x10^{-1}$	$6.18x10^{-2}$	$9.37x10^{-1}$	$9.74x10^{-1}$	$9.37x10^{-1}$
non-AFM	MSA	**$3.75x10^{-3}$**	**$4.27x10^{-5}$**	**$1.40x10^{-5}$**	**$6.92x10^{-3}$**	**$1.02x10^{-6}$**	$9.56x10^{-1}$	**$4.93x10^{-7}$**
	SS	$8.49x10^{-1}$	**$1.48x10^{-2}$**	$5.12x10^{-1}$	$1.61x10^{-1}$	$4.20x10^{-1}$	**$1.01x10^{-2}$**	$3.29x10^{-1}$
B								
AFM	MSA	**$3.33x10^{-2}$**	$2.70x10^{-1}$	$6.31x10^{-2}$	$6.31x10^{-2}$	**$2.08x10^{-2}$**	$8.21x10^{-1}$	$1.54x10^{-1}$
	SS	**$5.15x10^{-2}$**	$1.79x10^{-1}$	**$2.08x10^{-2}$**	$6.20x10^{-1}$	**$1.25x10^{-2}$**	$6.20x10^{-1}$	**$2.08x10^{-2}$**
non-AFM	MSA	**$2.07x10^{-9}$**	$2.72x10^{-1}$	**$1.45x10^{-9}$**	$9.43x10^{-1}$	**$2.93x10^{-9}$**	$9.86x10^{-1}$	**$6.89x10^{-9}$**
	SS	**$3.34x10^{-3}$**	**$3.97x10^{-3}$**	**$5.52x10^{-4}$**	$7.46x10^{-1}$	**$1.37x10^{-4}$**	$3.75x10^{-1}$	**$2.95x10^{-4}$**
C								
AFM	MSA	$4.16x10^{-1}$	$5.72x10^{-1}$	1.98x10-1	$4.27x10^{-1}$	$5.00x10^{-1}$	$5.00x10^{-1}$	$5.00x10^{-1}$
	SS	$7.99x10^{-1}$	**$5.02x10^{-2}$**	5.00x10-1	$1.86x10^{-1}$	$3.42x10^{-1}$	$8.62x10^{-1}$	$3.38x10^{-1}$
non-AFM	MSA	**$1.58x10^{-7}$**	$2.27x10^{-1}$	**$2.58x10^{-7}$**	$1.58x10^{-1}$	**$1.10x10^{-7}$**	$2.33x10^{-1}$	**$6.80x10^{-8}$**
	SS	**$3.49x10^{-2}$**	**$1.12x10^{-2}$**	**4.18x10-3**	$3.08x10^{-1}$	**$2.55x10^{-3}$**	$2.41x10^{-1}$	**$4.09x10^{-3}$**

*SS=Single sequence. P-values were calculated at the 95% confidence level. The 1-tailed Wilcoxon signed-rank test P-values were calculated using oligo-lDDT scores (A), TM-scores (B) and QS-scores (C) for AFM models of 10 CASP14 targets (generated with ColabFold) and the same 10 targets for models submitted by the 5 top-ranking groups in CASP14 (non-AFM). Supporting raw data is available in Appendix 6.

2.4.2 Secondary hypothesis. Is similar improvement seen for recycling in both single sequence and MSA modes?

To further investigate the differential improvement between MSA and single sequence recycling the two methods were directly compared using a 1-tailed Wilcoxon signed-rank test to test whether lDDT scores for MSA recycling were significantly higher than those obtained for single sequence recycling. In addition, a 1-tailed Ansari-Bradley test was used to investigate any significant differences in quartiles which may be occurring in the data but that remain hidden when using tests comparing mean values. Again, it is worth restating the hypotheses being tested: *H0. Recycling in single sequence mode results in models no different in quality compared to the baseline models used as templates. H1. Recycling in single sequence mode results in improvement in models similar to that seen for MSA mode.* Table 2.9, below, shows p-values obtained for the 16 CASP14 AlphaFold2 tertiary structure models. Equivalent data for the 47 non-AlphaFold2 tertiary models is presented in Table 2.10.

Table 2.9. CASP14 AF2 model comparisons between mean lDDT scores (top) and scale parameters (bottom). Single-sequence and MSA recycling across 1, 3, 6 and 12 recycles.

P-value / Test	Recycle 1 (SS v MSA)	Recycle 3 (SS v MSA)	Recycle 6 (SS v MSA)	Recycle 12 (SS v MSA)
Wilcox signed rank	0.097	0.052	0.111	0.129
Ansari test	0.397	0.500	0.425	0.544

Key: SS=Single sequence. MSA=Multiple Sequence Alignment. The 1-tailed Wilcoxon signed-rank test (top row) and Ansari test (bottom row) P-values were calculated at the 95% confidence level (those <0.05 are in bold) using lDDT scores for the 16 AlphaFold2 CASP14 top-ranked models from CASP14 targets.

Table 2.10. CASP14 non-AF2 model comparisons between mean lDDT scores (top) and scale parameters (bottom). Single-sequence and MSA recycling across 1, 3, 6 and 12 recycles.

P-value / Test	Recycle 1 (SS v MSA)	Recycle 3 (SS v MSA)	Recycle 6 (SS v MSA)	Recycle 12 (SS v MSA)
Wilcox signed rank	**1.42×10^{-14}**	**5.34×10^{-9}**	**2.94×10^{-12}**	**7.80×10^{-9}**
Ansari test	**0.014**	**0.015**	**0.019**	**0.012**

Key: SS=Single sequence. MSA=Multiple Sequence Alignment. The 1-tailed Wilcoxon signed-rank test (top row) and Ansari test (bottom row) P-values were calculated at the 95% confidence level (those <0.05 are in bold) using lDDT scores for the 47 non-AlphaFold2 CASP14 top-ranked models from CASP14 targets.

From Tables 2.9 and 2.10 it can be seen that there is no significant difference in model quality between MSA and single sequence recycling methods for the AF2 models according to both the Wilcoxon and Ansari tests. However, there is a significant difference, detected by both tests, at every recycle for non-AF2 models. However, it is unknown whether an equivalent MSA was used to produce the non-AF2 models, and so it is possible that this difference simply highlights the power of the MSA in producing better contact and distance maps on which to base models.

Figure 2.10 shows a graphical representation of the relative improvements in lDDT score for all models on which the values in Table 2.9 and 2.10 were based. As expected from the values

in the table, MSA recycling (left plot) shows an increase in quality for non-AF2 models to a much greater extent than that for single sequence. Nevertheless, according to the values in Table 2.7 models produced by single sequence recycling were still significantly improved. In this case the null hypothesis should be rejected as single sequence recycling clearly results in model improvement which is significant for both model populations. However, it is also true that the alternative hypothesis can be applied only to the AF2 models but not the non-AF2 models as the improvement for single sequence recycling could not reasonably be said to be similar to that for MSA recycling for this model population. Therefore, a different interpretation maybe required; that single sequence recycling can be viewed as representing refinement due to AlphaFold2's learned protein folding function and the difference between the improvement seen between AF2 and non-AF2 models is the effect of additionally using a multiple sequence alignment.

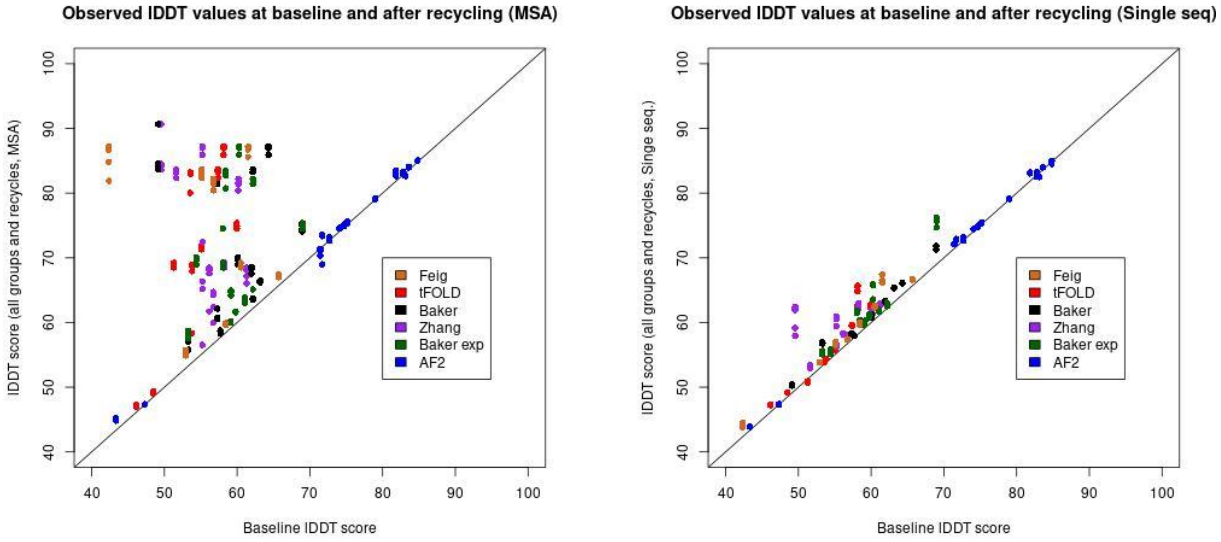


Figure 2.10. Scatter plots to show comparisons in observed lDDT scores between baseline and all recycles for all monomeric models. Left. MSA recycling. **Right.** Single sequence recycling.

During discussions for this section and 2.4.1, the secondary consideration of whether improvement would be seen in official DeepMind AF2 competition models has also been answered. The fact that this improvement was seen for both MSA and single sequence recycling using the AF2 model with pre-CASP14 weights, meaning that no new information was presented to the algorithm, is further indication that improvement is occurring via some sort of learned function within the AF2 neural network. One contextual point to note is that DeepMind entered CASP14 as a manual group meaning that changes may have been made to models which were not due entirely to the AF2 software and that this part of the experiment could have been carried out with models generated by ColabFold. Nevertheless, it remains that the CASP14 AF2 models represented the best independently benchmarked models available to us at the time.

In addition, quaternary structure model improvement was also seen for single sequence recycling as well as MSA recycling, and this improvement was also apparent for AFM models. As measured by lDDT-oligo score, 30% of AFM models were improved from baseline after single sequence recycling, compared to 80% using MSA recycling, but the percentage improvement was higher for non-AFM models where 64% improved with single sequence recycling compared to 94% for MSA. Similar levels of improvement were seen for TM-score (up to 80% of AFM models and 98% non-AFM models) and QS-score (up to 50% of AFM models and 86% non-AFM models) (Adiyaman *et al.*, 2023).

2.4.3 Is improvement linear with recycle number and can an optimal number of recycles be determined?

Finally, the secondary considerations of linearity and identification of an optimal recycle number need to be addressed. Table 2.7 shows that improvement in model quality doesn't follow a linear trend; higher recycle numbers do not consistently yield more significant improvements. For AF2 models, only two consecutive recycles (baseline to 1 recycle and 1 to 3 recycles) show a significant increase in lDDT for single sequence modelling and three (baseline to 1 recycle, 1 to 3 recycles and 3 to 6 recycles) for MSA modelling. Similarly, any improvement in score for non-AF2 models after 6 recycles (3 for single sequence) also becomes non-significant.

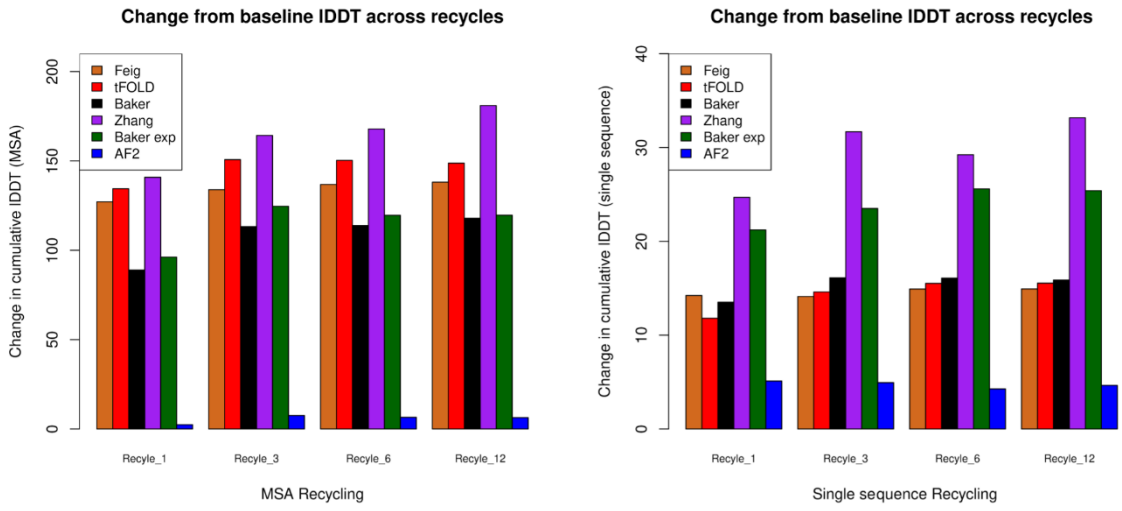


Figure 2.11. Plots to show the change from baseline in cumulative observed lDDT scores (all recycles) per modelling group. Left. MSA recycling. **Right.** Single sequence recycling. Data for all monomer models for AF2 and non-AF2 groups.

Identification of the recycle number producing the most improvement is not immediately obvious from the data in the tables. Therefore, it may be worth looking at the cumulative lDDT change from baseline for all individual groups to get a better representation of the trends as shown below in Figure 2.11. For the MSA recycling data (Figure 2.11, left plot), two groups,

Zhang and Feig, showed a slight increase in cumulative score from recycle 3 to 6 and a further increase from recycle 6 to 12, which also included a marginal increase for the Baker group. All other groups showed no further improvement after 3 recycles. For the single sequence recycling data, the Feig, tFOLD and Baker-experimental groups all showed improvement after 3 recycles with the Zhang group showing further improvement only at 12 recycles. The AF2 and Baker groups showed no further improvement after 3 recycles. Interestingly a number of groups showed a slight decrease in model quality after 3 recycles, specifically AF2, Baker-experimental and tFOLD for MSA recycling and AF2 for single sequence (with a dip from Zhang group at 6 recycles). A decrease in quality for some models is not uncommon with refinement procedures (Adiyaman and McGuffin, 2019) and in light of this, to avoid the risk of a decrease in quality during recycling, it would be prudent to suggest 3 recycles as the optimum number for tertiary structures.

2.4.4 Improvement of non-AF2 models beyond AF2 quality.

An important and unexpected effect seen when recycling non-AF2 models was the improvement of some models beyond the quality of the equivalent DeepMind AF2 competition models as measured by lDDT score. This was surprising as the full power of the DeepMind neural network and MSA search facility would have been used to create these original models whereas the quicker MMSeqs search method was used with the ColabFold method, producing a slightly different MSA, which would not necessarily be expected to out-perform the former. This enhanced improvement may, again, be indicative of a process other than simple correction of modelling inaccuracies using the information available in an MSA. Two examples of this are shown in Figure 2.12 and 2.13 below.

Figure 2.12. Images of CASP14 target T1074. Left. The Baker group's predicted model (blue, lDDT 0.491, TM-score 0.576) superposed with the native structure (purple). **Centre.** The refined model in red (lDDT 0.906, TM-score 0.959). **Right.** The refined model superposed with the native structure and showing a very close alignment.

Figure 2.12 shows the improvement seen in the Baker group's model for the CASP14 target T1074. The left-hand image shows the model (coloured blue) in superposition with the native structure, revealing a misaligned lower beta sheet and resulting in a TM-score of 0.576 and an

lDDT score of 0.491. The recycled model, centre (coloured red) and right in superposition with the native structure, shows that this misalignment has been corrected and the TM-score has improved to 0.959 with a similarly improved lDDT score of 0.906. For comparison the competition AF2 model scored a TM-score of 0.930 and an lDDT score of 0.848.

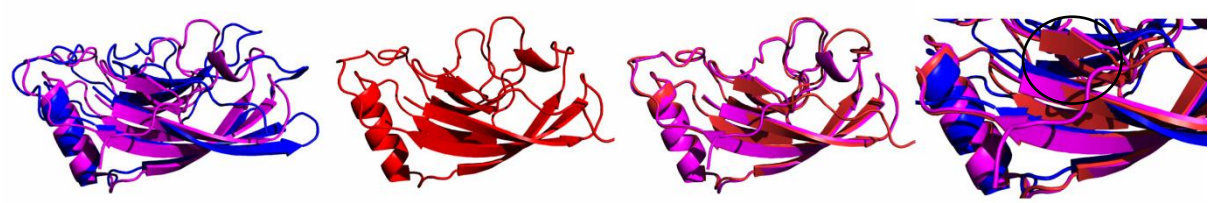


Figure 2.13. Images of CASP14 target T1049. Far left. The Zhang group's predicted model (blue, lDDT 0.552, TM-score 0.674) superposed with the native structure (purple). **Centre left.** The refined model in red (lDDT 0.872, TM-score 0.940). **Centre right.** The refined model superposed with the native structure and showing a closer alignment. **Right.** An enlargement showing a superposition of all three models and highlighting a newly formed β-strand (circled), absent in predicted model.

Similarly Figure 2.13 shows the improvements seen to the Zhang group's model for T1049. Again, the original model is coloured blue and, in the superposition with the native structure on the left, shows a number of positional inconsistencies in the beta strands as well as the loop sections. These have been corrected in the recycled (red) model and led to improvement in scores from 0.674 to 0.940 for the TM-score and from 0.552 to 0.872 for lDDT. The right-hand graphic shows an enlarged section of a superposition of all three models (predicted, recycled and native) highlighting the correct inclusion of a small beta section in the recycled model (circled in black) which was not present in the predicted model. For comparison the equivalent AF2 model scored 0.930 and 0.848 for TM-score and lDDT respectively, which was, again, lower than this recycled model.

Similar levels of improvement were seen for quaternary structures, although for these models it could not be claimed that recycling had improved DeepMind official competition models as the AFM models used in the study were specifically created for this project using ColabFold (DeepMind did not submit any predictions for multimeric targets). Nevertheless, a good example of AFM model improvement is shown in Figure 2.14 panel A in which the interface orientation of the AFM model is corrected via recycling, resulting in improvement in the lDDT-oligo, TM-score and QS-score. Further to this, Figure 2.14 panel B shows a Venclovas group model for H1045 which improved to match the scores achieved by the equivalent AFM model for TM-score, and slightly beyond that achieved by AFM for lDDT-oligo and QS-score.

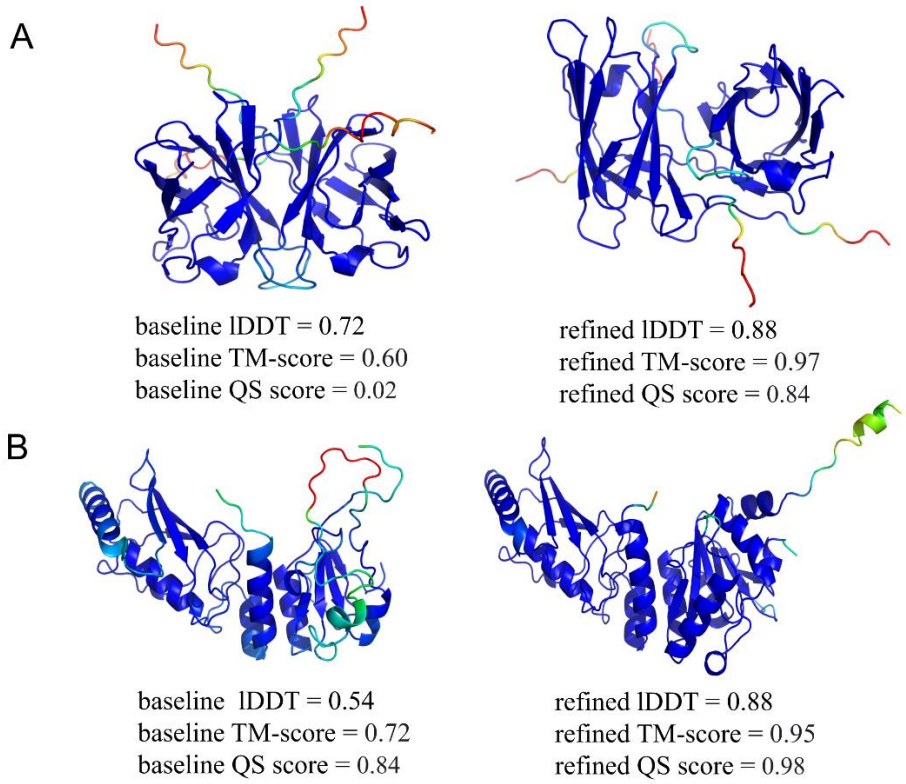


Figure 2.14. A. Images of the AFM model for CASP14 multimeric target T1078. Left. The AFM predicted model with lDDT-oligo, TM-score and QS-score values. **Right.** The refined model with equivalent scores. **B. Images of the Venclovas group model for CASP14 target H1045. Left.** The original predicted model, again with lDDT-oligo, TM-score and QS-score values. **Right.** The refined model and equivalent scores. The scores for the equivalent AFM model were: lDDT-oligo 0.87, TM-score 0.95 and QS-score 0.97. Images coloured by plDDT and adapted from (Adiyaman *et al.*, 2023).

2.5 Conclusions

The conclusion for the primary outcome is that recycling full tertiary structure protein models via the ColabFold custom template option is possible, and that it significantly improves full model structures beyond their starting quality. Conversely, significant improvements were not seen in a parallel study using more conventional molecular dynamics refinement techniques (Adiyaman *et al.*, 2023).

Findings for the three secondary considerations can be summarised as following:

Improvement for both MSA and single sequence modes.

Firstly, recycling using both MSA and single sequence modes leads to significant improvement in model quality compared to the baseline, as measured by lDDT score. It has been demonstrated that although a greater improvement in model quality occurs when the AlphaFold2 algorithm is able to perform a multiple sequence alignment (MSA) during recycling, a significant improvement in model quality is still apparent when only the amino acid sequence is supplied (single sequence modelling).

Improvement in official DeepMind AF2 competition models.

It has been shown that improvement occurs not only with non-AF2 models with a median baseline lDDT score of 0.580 but also with the official DeepMind AlphaFold2 models with a much higher median baseline lDDT of 0.751. Also, there is no significant difference in lDDT scores between MSA and single sequence recycled AF2 model populations. As it has been previously documented that AF2 performance considerably decreases when using sequence-only modelling (Roney and Ovchinnikov, 2022), this strongly suggests that model improvement is being achieved via template refinement rather than remodelling despite any slight differences in MSA construction. Indeed, as the AF2 models were originally built by the same software it should not be possible to improve them by remodelling if no further information is available to the algorithm. In seeming contradiction to this, there is a significant difference between MSA and single sequence recycling for the non-AF2 models. However, this is likely due to differences in the original modelling software used, i.e. that AF2 is finding inconsistencies between the models and its own algorithm, which it is able to better correct using the additional information provided by a new or different MSA.

Pattern of improvement and optimal recycle number.

Thirdly, that improvement in model quality is non-linear. The lack of significant improvement in consecutive recycles (evidenced in Table 2.7 and 2.8) shows that a higher recycle number does not equate to more significant improvement. Therefore, committing tertiary structures to more than 3 recycles is unlikely to further improve the model and may represent an unnecessary processing overhead along with the risk of decreasing the quality of high scoring models.

In summary, it can be concluded that recycling through the AF2 DNN, via the use of custom templates, will lead to an improvement in tertiary and quaternary structure model quality in the vast majority of cases, even for models with a very high level of initial accuracy. Using the MSA mode as a recycling option will likely lead to a certain amount of remodelling if the template is not an AF2 model and that single sequence mode therefore probably better represents “pure refinement”.

In explanation, it may be that the AF2 algorithm is using the template as an enhanced starting model in place of its usual contact matrix thus allowing modelling to start at a point deeper in the folding funnel of the energy landscape. Alternatively, it may be that the AF2 DNN has, to some degree, learnt a protein folding function (an algorithmic ability to recognise correct or incorrect folds) which can be exploited to improve models without any additional information simply by repeated iteration of the model through the network regardless of the quality of the initial model. A similar alternative may be related to the recently published work on diffusion

de-novo protein design (Watson *et al.*, 2023). Here model coordinates are deliberately obscured with a noise function prior to denoising using a specially trained version of RoseTTAFold (RFdiffusion) and one of the training strategies (self-conditioning) was cited as being ‘inspired by AF2 recycling”. It’s possible that some of the lower quality models acted as crude “noisy” structures in that they provided a rough starting point and that, through successive iterations, AF2 was able to reduce the initial 'noise' and progressively improve the structure. While the folding funnel concept focuses on the energy landscape and convergence towards the native state, diffusion suggests iterative refinement of an initially crude structure.

The results above were important because they inspired the following MultiFOLD pipeline, shown below in Figure 2.15, which was used for the CASP15 competition and also underlies the version available publicly on the IntFOLD website. Here a dual modelling process is used including two versions of the AlphaFold2 model. LocalColabFold 1.0.0 features the AF2 model used in the AF2 Advanced version which is the model trained on tertiary structures but which was used extensively, following CASP14, to model multimeric structures (Bryant *et al.*, 2022) before AF2-Multimer was released. During the process AMBER relaxation is used in 50% of models and ranking is performed by ModFOLDdockR, see Chapter 4 for details of the development of this version. Of the 20 models created by the dual pathway, the top 5 are then recycled as custom templates in a refinement stage before being pooled with the remaining population ready for final quality assessment and ranking.

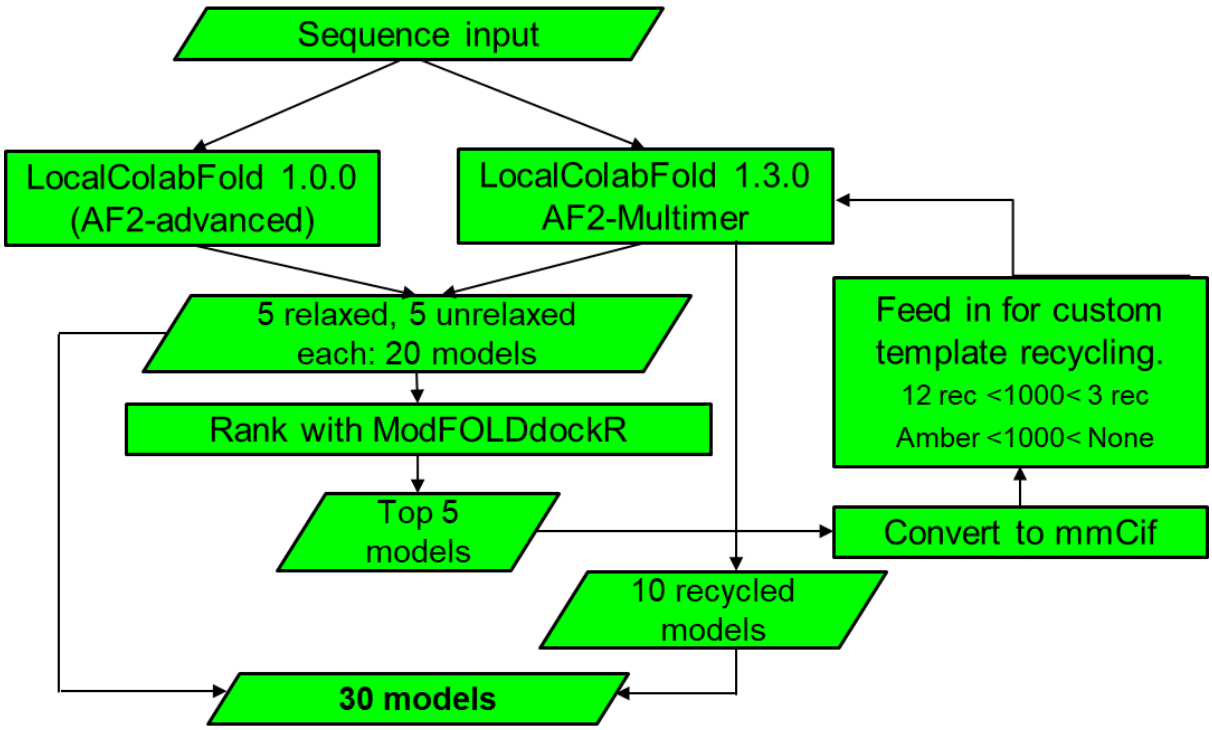


Figure 2.15. The updated MultiFOLD pipeline developed for CASP15. This configuration was inspired by the recycling results described in this chapter.

CHAPTER 3

Development of new global and local quality estimates for quaternary structure models using artificial Neural Network (NN) comparisons with CASP quality scores.

3.1 Background and historical context

Protein modelling software has historically produced large numbers of models, some of which may be native-like, while others (decoys) may be structurally different. Modelling confidence scores are designed to objectively differentiate between these two model groups (Elofsson *et al*., 2018) and can be categorised into two broad types. Accuracy self-estimate (ASE) scores usually refer to the modelling confidence scores output by the modelling software while the term estimates of model accuracy (EMA) is usually applied to confidence scores calculated by separate, independent, software. The term model quality assessment (MQA) can be considered an umbrella term covering both ASEs and EMA, however the terms EMA and MQA are often used interchangeably in the published literature.

Modelling pipelines often provide proprietary ASE scores, and this clearly presents a problem when attempting to meaningfully evaluate models from different sources by ASE score comparison alone. This was demonstrated during CASP15 where groups were required to standardise their ASE scores into predicted lDDT scores (plDDT) for the competition; some were more successful than others and the accuracy varied depending on the target protein (Gabriel Studer *et al*., 2023). Outside of the competition arena where ASE scores may remain proprietary in nature, model quality assessment via independent EMA methods remains a critical stage in selecting the most representative model from multiple modelling sources.

3.1.1 A brief history of MQA

Conceptually, model quality assessment appears a relatively straight forward problem but it has been shown to be challenging to reliably determine whether two protein structures are similar enough for one to be representative of the other (Xiao Chen *et al*., 2021). This has led to a high degree of variation in the approaches used.

Traditionally, MQA has been divided into single-model and clustering methods. In general, single-model methods employ a number of physical checks to assess each model's structural integrity. These range from residue environment compatibility, e.g. hydrophobicity and solvent accessibility to structural features, such as secondary structure compatibility and assessment of backbone torsion angles (McGuffin, 2010). Users are then presented with scores showing how well each model conforms to hypothetical 3D norms. One problem with these plausibility checks is that a model may score well because it conforms to pre-programmed ideals, whereas another, which could be closer to the native structure, may score badly due to minor structural defects (Edmunds and McGuffin, 2021). In an aim to reduce these errors, consensus and clustering approaches were developed. Single-model consensus approaches operate as described above but include a number of diverse scoring algorithms which can then be combined to create a single consensus score. In reality, most consensus approaches perform

a clustering routine (McGuffin and Roche, 2010) where models are clustered on the basis of their conformational similarities determined by distance-based pairwise measurements. Here the distances between any two amino acids in one structure are directly compared to the distances between equivalent residues in all other models across the population. Models representative of large clusters are proposed to have a higher likelihood of resembling the native structure than remote models as correct conformations should occur repeatedly while errors are deemed to occur randomly. The obvious drawbacks with clustering are that accuracy will likely diminish with a decreasing model population and that all-against-all comparisons become computationally restrictive for very large populations. As a compromise, quasi-single-model methods attempt to exploit the best of both worlds by creating a population of reference models which are then used to perform one-against-all comparisons with the target structure. These comparisons are less computationally expensive and quasi-single-model approaches such as the ModFOLD suite of programs (McGuffin, 2008) performed well in CASP tertiary structure EMA competitions (Chen and Siu, 2020) which have been running since CASP7 in 2006.

Latterly, approaches centring on the assessment of contact profile similarity, for example CAD-score, VoroMQA (Olechnovic and Venclovas, 2017) and CDA-score (Maghrabi and McGuffin, 2017), and those employing machine learning (ML) techniques have been developed. While early support vector machine (SVM) algorithms such as ProQ2 and 3 (Uziela *et al*., 2016) were successful, deep learning techniques using neural networks (NN), such as ProQ3D (Uziela *et al*., 2017), were able to flourish by using training datasets from a model pool which had significantly increased in quality following CASP13 (2018) (Chen and Siu, 2020). Finally, hybrid consensus programs such as ModFOLD7 and 8 (McGuffin *et al*., 2021), and MULTICOM (Cheng *et al*., 2023) combined single and multi-model techniques, contact information and trained neural networks to further improve performance.

The systems described above were initially developed for tertiary structure MQA and, in 2018 when this project was conceived, quaternary structure MQA was less well developed. At this time the modelling landscape was dominated by numerous docking programs scattered across a number of websites, for example FRODOCK (Garzon *et al*., 2009) was accessed from the InterEvDock webserver (Vavrusa M *et al*., 2016) and PatchDock (Duhovny *et al*., 2002) from the SymmDock (Schneidman-Duhovny *et al*., 2005) site, which required the user to supply fully modelled monomers as well as constraints and other technical data in some cases. The many programs without interactive webservers often additionally required the download and installation of stand-alone docking software. Estimations of model accuracy were mostly via docking or reranking scores like ZDOCK's ZRANK score (Pierce and Weng, 2007), although there were two early attempts at objective ASE in the form of Swiss-model's QSQE score

(Bertoni *et al*., 2017) and HADDOCK's own accuracy score (Vangone *et al*., 2017). However, independent predicted MQA programs were largely absent with the exception of ProQDock (Basu and Wallner, 2016b), available as a download from 2016. Significant barriers to accessible and accurate predicted model quality assessment for quaternary structures therefore existed at this time.

This chapter will focus on an analysis of our multimer MQA software performance over recent CASP experiments identifying both limitations in early versions and the identification of areas for development. The intention is to document our research and development leading up to the world-class performance of ModFOLDdock in the CASP15 EMA competition (2022), which is fully documented in Chapter 4.

3.1.2 Scores for calculating the observed model quality by comparison with native structures

Descriptions of commonly used scores for determining the observed quality of 3D models are given in Appendix 1, but the main features of RMSD, GDT_TS, TM-score, lDDT and QS-score will be repeated here for convenience as these scores are often referred to in this chapter. The descriptions include a three-point classification of protein model evaluation methods (Olechnovic *et al*., 2019) which categorises them as; either superposition-based or superposition-free; global or local in similarity and all atom or atom subset (e.g. Cα or Cβ atoms) in coverage.

Root Mean Square Deviation (**RMSD**) (Arun *et al*., 1987) (superposition-based, global, Cα atoms only) calculates the sum of the squares of the distances between Cα atoms of the model and native structure. This value is then divided by the total number of residues and the square root calculated to give a normalised deviation. Scores closer to 0 are better. The main drawbacks with RMSD stem from each Cα atom being treated equally. A small area of deviation within the model, often a loop or terminal section, can quite heavily penalise an otherwise representative model, also the interpretation of both an acceptable deviation distance (e.g. 5Å) as well as the length of the superposition alignment may vary for chains of different length. For example, a lower RMSD score calculated over a 50% alignment may not be better than a higher score calculated over 75%.

Global Distance Test, Total Score (**GDT_TS**) (Zemla, 2003) (superposition-based, global, Cα atoms only) represents the percentage of residues in the largest superimposable substructure falling within a predefined distance compared to the native structure. CASP uses the mean of four distances (1, 2, 4 and 8Å) to calculate the overall score on a 0-100 scale. It can be summarised as GDT TS (Mp, Mr) = (P1 + P2 + P4 + P8)/4 where Mp and Mr represent the

predicted and reference models respectively, and P1, P2, P4, and P8 represent the percentage of Cα atoms which can be superposed at each distance cut-off.

The oligo version of the GDT score is very similar and uses the above distances to construct a rotational matrix (Kabsch, 1976) which can be manipulated to find the minimum super-position RMSD before calculating the final score. The GDT attempts to improve upon RMSD by using the mean of all four cut-off distances to limit the effect of a small number of large errors however, it still suffers from length-dependent bias (Zhang and Skolnick, 2004) as a substructure alignment of only 60% within 8Å might be considered poor for a short protein of 50 residues but be more favourably viewed for one of 500 residues. This is also an issue for the **MaxSub** score (Siew *et al*., 2000) which calculates a similar alignment substructure agreement but uses only one cut-off distance (often 3.5Å).

The global score which is widely accepted to have solved the length dependence problem is **TM-score** (Zhang and Skolnick, 2004) which not only normalises by the whole length of the native structure (L_N), but also calculates a length dependent distance cut-off (d_0=1.24 $\sqrt[3]{\mathrm{LN}-15}-1.8$) meaning that TM-scores for chains of different lengths can be directly compared (N.B. the TM-score for multimers is calculated using the MM-align package). Using *di* as the distance between corresponding residues in the target and reference protein and with d_0 and L_N as defined above, the TM-score calculation can be summarised as follows.

$$\text{TM-score} = \max\left[1/L_N \sum \frac{1}{1+\left(\frac{di}{d0(LN)}\right)^2}\right]$$

The local Distance Difference Test (**lDDT**) (superposition-free, local, all atom) is a 0-1 score expressing the fraction of contacts shared or conserved between a model and its native structure regardless of orientation. lDDT-oligo is the multimer equivalent which uses the QS-score (see below) chain mapping routine to identify intra and inter-chain contacts prior to calculating the test score. The score penalises both deficiency of atoms and incorrect stoichiometry in the model structure and, while this is a good measure of, for example, domain or individual chain similarity, it gives little impression of the orientation of one domain to the next or one chain to another (for the oligo version). In some ways this could be considered an advantage given the multi-conformational nature of some proteins, but it can also be argued that it elicits limited information about the interface quality.

Quaternary Structure (**QS-score**) (Bertoni *et al*., 2017) (superposition-based, local interface, Cβ atoms). A score representing the fraction of shared interface contacts within 12Å between model and reference structure once a mapping algorithm has identified multimer symmetry and equivalent chains. A 0–1 score where 0 represents different quaternary structures and 1

suggests very similar models. Higher scores therefore represent correct stoichiometry, symmetry, and a high fraction of conserved interface contacts.

3.1.3 Scores used in the CASP13 version of ModFOLDdock

Analyses in this chapter is concerned with ModFOLDdock score optimisation and the contributing scores are described below. ModFOLDdock can compute both predicted and observed scores, the latter being calculated compared to a known reference structure. There are six predicted scores (1 single model and 5 clustering scores); consisting of two DockQ scores; ProQDock and DockQJury, IA-score (ModFOLDIA), two QS-scores; QSscoreJury and QSscoreOfficialJury and an lDDT score (lDDTOfficialJury). ProQDock is the only single model score, all scores have a 0-1 range.

The **DockQ** (Basu and Wallner, 2016a) routine creates a score based on the CAPRI (Critical Assessment of Prediction of Interactions) quality measures Fnat, LRMS and iRMS. Fnat is defined as the fraction of native interface contacts observed in the model, LRMS is the root mean square deviation (RMSD) of the chain denoted the ligand (smaller chain of a complex) after superposition of the larger chain and iRMS is the RMSD between interface residues seen in the native structure compared to the model. A 0 to 1 score, the range of DockQ scores matches the following CAPRI quality classes: < 0.23 (Incorrect), 0.23 – 0.49 (Acceptable), 0.49 – 0.8 (Medium) and > 0.8 (High). ModFOLDdock calculates two DockQ scores; **ProQDock** (Basu and Wallner, 2016b) (single-model method) and a clustering-style **DockQJury** method.

IA-score (ModFOLDIA). A proprietary score created by the McGuffin group. To calculate this score, interface residues are identified (defined as ≤5Å between non-Hydrogen atoms in different chains) and the minimum contact distance (Dmin) for each contacting residue is measured. Equivalent residues in all other models are then identified and the mean Dmin is then calculated across the sample. Si and Mean Si are then calculated as follows:
$Si = 1/(1+(Dmin /20)^2)$ and $Mean\ Si = 1/(1+(Mean\ Dmin /20)^2)$.
The IA score for each interface residue (i) is then the absolute difference of Si from the mean Si, i.e. $IA = 1-|S\ i - MeanS\ i|$

The global predicted ModFOLDIA score for a model is the sum of the residue scores normalised by the maximum mean number of interface residues across all models for the same target. Scores of <1 represent variation from the mean.

QSscoreJury (QSJ) and QSscoreOfficialJury (QSOJ) (see QS-score definition above and in Appendix 1). The difference between the two QS-scores is that QSJ uses in-house code to calculate the fraction of correctly modelled interface contacts normalised by the total predicted

contacts, whereas QSOJ employs OpenStructure (Biasini *et al*., 2013) to calculate QS-scores using the “ost compare-structures” action.

lDDTOfficialJury (again, see the lDDT definition above and in Appendix 1)
Scores suffixed ‘Jury’ are calculated by the 3D Jury method. The rationale being that the average of many low-energy conformations is closer to the native structure than the absolute lowest-scoring model. In terms of scores, this translates as pairwise comparisons between models on an all-against-all basis followed by the calculation of the mean scores. During the calculation, models are assigned a MaxSub score to calculate similarity by counting the numbers of pairs of Cα atoms that remain within 3.5A after optimal super positioning. Structures are considered dissimilar if the Cα atom count is less than 40. The final 3D-Jury score is the sum of similarity scores across model pairs divided by the number of model pairs (+1). Models representative of the mean value of the largest cluster are therefore selected.

There are also five comparative ModFOLDdock observed scores, which require a reference or native structure for score calculation. All scores are calculated as described above (without clustering) or in Section 3.1.2 and are calculated on a 1-versus-all basis rather than the all-verses-all used for predicted scores. The scores are **IA-score**, **DockQ**, **QSscore_Calc**, **QsscoreOfficial** and **lDDTOfficial**.

3.1.4 Multimer MQA lacked accuracy at CASP13 and 14

CASP13 took place in 2018 and included a quaternary or assembly modelling category which comprised 42 multimeric targets. The McGuffin group modelling methodology and subsequent performance is covered in detail in Chapter 2 along with some of the issues and shortcomings, which were revealed and targeted for improvement. This chapter will focus exclusively on the issues surrounding model quality assessment of the 30 models submitted.

Selected scores for all homomeric targets modelled by our group are shown in Table 3.1. Additionally, images of the structures along with a table of scores showing models with the largest discrepancies between submitted and observed scores can be found in Appendix 4. The CASP scores presented in Table 3.1 were chosen to represent the quality of the models by both global relatedness and interface similarity to the native structure. GDT and lDDT are the two scores contributing to the global element of the overall Z-score calculation on which CASP13 group rankings were based (see Section 2.1.2) and QS-score represents a single overall score encompassing both global stoichiometry and interface geometry accuracy (Haas *et al*., 2018). These are presented along with the ModFOLDdock predicted Consensus6 score (an unweighted mean of all constituent predicted scores) and the retrospectively calculated *Observed mean score*.

Table 3.1. Quality assessment scores (predicted and observed) for McGuffin CASP13 assembly models. Rows labelled *Submitted* represent scores for the group's top model submitted to CASP whereas those labelled *Best* were identified retrospectively by mean observed score. Most of the models labelled as *Best* were not selected for CASP submission and therefore have no accompanying CASP scores. Best model scores are not given for either T0996 as it was a manually created single model or for T1016 where the submitted model was also the best model.

Target	*Model*	*Predicted C6 score*	*Observed mean score*	*CASP scores*			
				GDT_TS	*RMSD*	*lDDT*	*QS-score*
T0960	Submitted	0.356	0.156	6.55	71.86	0.285	0.000
	Best	0.343	0.328				
T0961	Submitted	0.370	0.441	23.70	31.07	0.689	0.000
	Best	0.338	0.841				
T0963	Submitted	0.317	0.144	6.83	77.57	0.331	0.000
	Best	0.243	0.291				
T0965	Submitted	0.369	0.436	32.75	15.19	0.582	0.200
	Best	0.322	0.487				
T0966	Submitted	0.331	0.161	30.66	33.58	0.597	0.000
	Best	0.202	0.29				
T0970	Submitted	0.379	0.207	20.71	14.31	0.351	0.000
	Best	0.295	0.301				
T0973	Submitted	0.364	0.172	26.76	20.21	0.340	0.016
	Best	0.260	0.293				
T0976	Submitted	0.378	0.166	27.05	25.88	0.570	0.001
	Best	0.259	0.569				
T0977	Submitted	0.446	0.191	14.40	42.55	0.477	0.002
	Best	0.179	0.468				
T0979	Submitted	0.367	0.256	14.17	47.54	0.314	0.000
	Best	0.260	0.452				
T0981	Submitted	0.510	0.148	6.51	59.09	0.318	0.001
	Best	0.156	0.371				
T0983	Submitted	0.399	0.287	45.04	21.14	0.751	0.000
	Best	0.370	0.834				
T0984	Submitted	0.399	0.326	45.38	5.53	0.634	0.477
	Best	0.372	0.604				
T0989	Submitted	0.462	0.125	8.88	34.53	0.250	0.014
	Best	0.350	0.197				
T0991	Submitted	0.375	0.114	11.04	23.45	0.231	0.001
	Best	0.277	0.199				
T0995	Submitted	0.733	0.225	10.40	33.28	0.590	0.018
	Best	0.606	0.268				
T0996	Submitted	NA	NA	3.84	59.72	0.492	0.006
T0997	Submitted	0.321	0.179	31.10	15.38	0.494	0.000
	Best	0.273	0.261				
T0998	Submitted	0.341	0.08	8.21	29.04	0.165	0.000
	Best	0.273	0.188				
T0999	Submitted	0.242	0.198	12.80	39.41	0.691	0.005
	Best	0.173	0.274				
T1000	Submitted	0.284	0.158	23.86	23.47	0.568	0.000
	Best	0.263	0.313				
T1001	Submitted	0.384	0.169	39.03	9.17	0.669	0.036
	Best	0.291	0.263				
T1003	Submitted	0.331	0.228	42.58	27.02	0.643	0.000
	Best	0.217	0.470				

T1004	Submitted	0.378	0.246	16.56	53.19	0.527	0.003
	Best	0.272	0.366				
T1006	Submitted	0.406	0.319	49.66	14.46	0.639	0.000
	Best	0.361	0.865				
T1009	Submitted	0.285	0.270	32.39	16.37	0.575	0.004
	Best	0.253	0.409				
T1010	Submitted	0.358	0.260	26.14	10.38	0.357	0.072
	Best	0.285	0.382				
T1016	Submitted	0.458	0.667	76.73	2.50	0.686	0.693
T1018	Submitted	0.354	0.212	39.89	14.62	0.637	0.000
	Best	0.264	0.381				
T1020	Submitted	0.462	0.381	23.62	22.71	0.567	0.019
	Best	0.306	0.621				

The data in Table 3.1 show that the best available model was selected for submission on only one occasion (T1016). In all but two other cases (T0961 and T0965), the submitted models were overpredicted compared to their observed scores with a mean overprediction value of 0.146 (this difference was shown to be statistically significant using a Wilcoxon signed rank test in Chapter 2 (Table 2.2, P-value of 2.18×10^{-05}). Calculating the score difference across all best available models, on the other hand, shows a mean underprediction of -0.128. Just as importantly, the observed scores for the best available models are on average 0.18 higher than that for the models selected for submission, with a maximum difference as high as 0.546 for target T1006, showing that models closer to the native structure were clearly available in the decoy population and should have been selected. This data, which was collected for an initial exploratory study into ModFOLDdock performance, is represented graphically in both Figure 3.1, showing comparisons between ModFOLDdock predicted scores and the CASP scores listed in Table 3.1, and in Figure 3.2, showing the improved correlations obtained between equivalent ModFOLDdock observed scores and the same CASP scores.

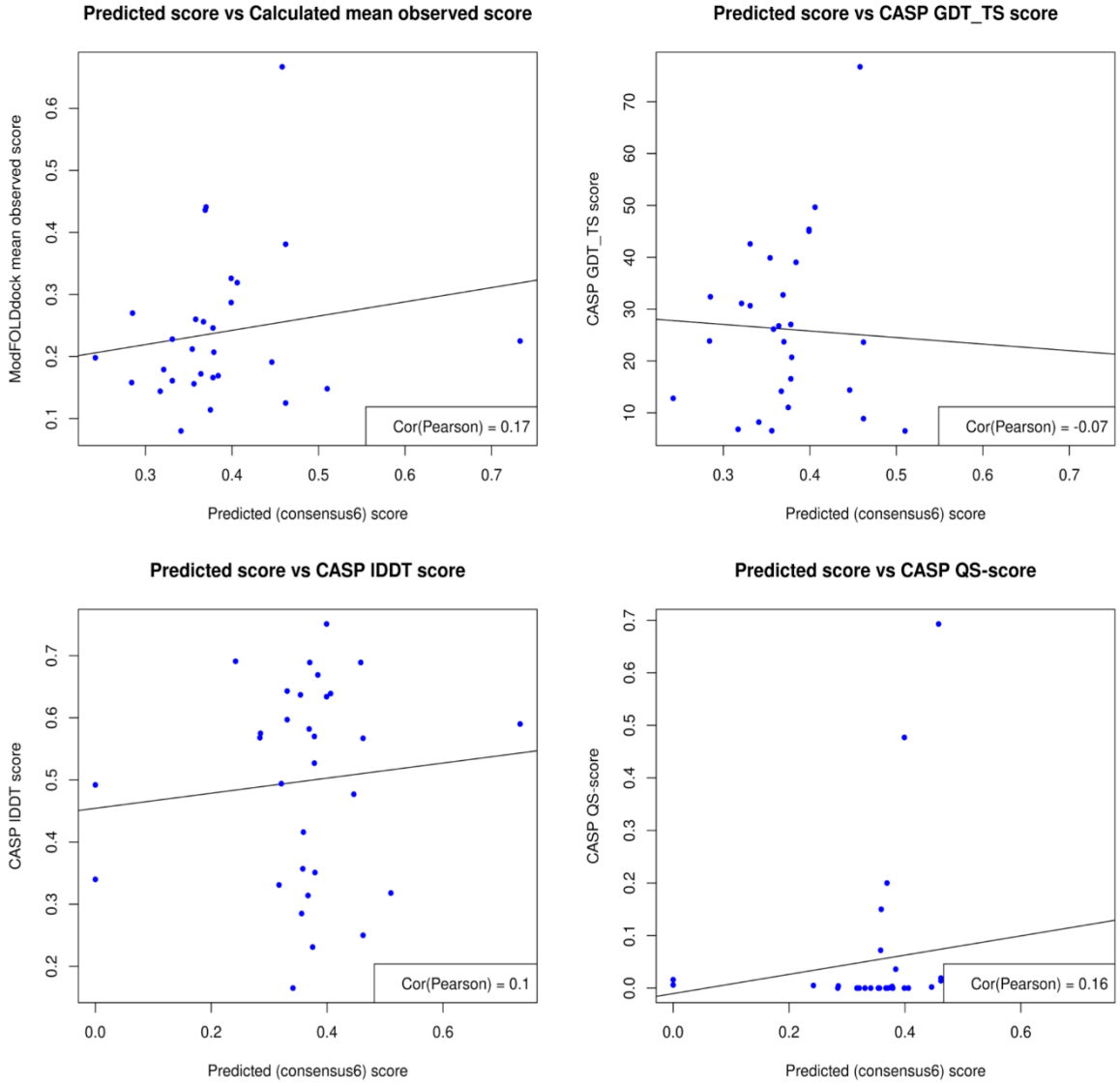


Figure 3.1. Correlation of ModFOLDdock Consensus6 score with observed scores for McGuffin group CASP13 assembly models. Top left. With observed mean score. **Top right.** With CASP13 GDT TS score. **Bottom left.** With CASP13 lDDT-oligo. **Bottom right.** With CASP13 QS-Score.

Figure 3.1 clearly shows that the ModFOLDdock predicted Consensus6 score does not correlate well with either our own observed mean score or the CASP scores shown, demonstrating that the unweighted ModFOLDdock Consensus6 score used in CASP13 was not a good model quality differentiation tool. Figure 3.2 shows much a stronger positive correlation between GDT and ModFOLDdock mean observed score, despite some spread in the scatter. Correlations with lDDT and QS-score are weaker, but still improved from their predicted counterparts. As ModFOLDdock uses the same contributing scores for both predicted and observed calculations, these differences suggest that there is potentially hidden predictive power within the ModFOLDdock score blend which could be improved with optimisation.

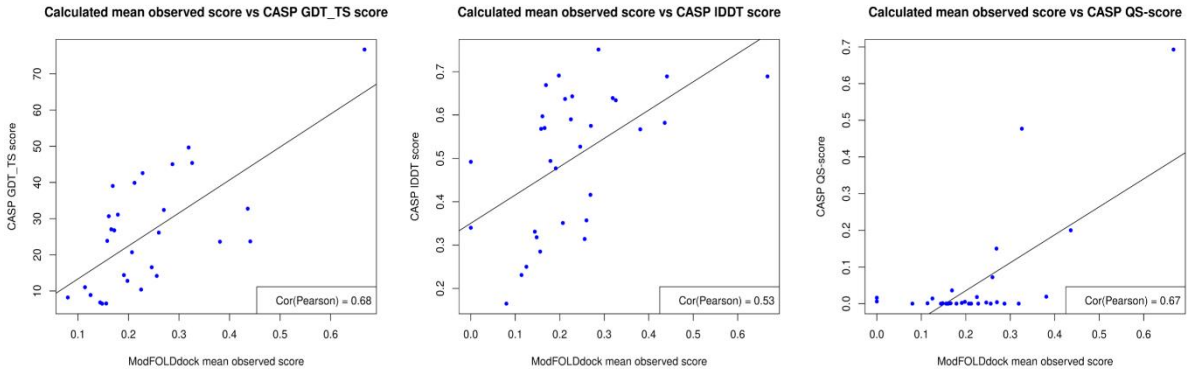


Figure 3.2. Correlation of mean observed score with CASP13 observed scores for McGuffin group CASP13 assembly models. Left. With GDT TS score. **Middle.** With lDDT score. **Right.** With QS-Score.

CASP13 analysis was not performed until the latter part of 2019 and so it had not been possible for a full optimisation programme to be implemented prior to the start of CASP14 which took place during the first half of 2020. In an effort to reduce the disparity between predicted and observed performance the McGuffin MQA pipeline was updated to include the Voronoi tessellation program VoroMQA (Olechnovic and Venclovas, 2014) alongside ModFOLDdock scores. The VoroMQA score was combined with the ModFOLDdock Consensus6 score to create a hybrid unweighted mean of both scores which was used as the primary ranking score for CASP14 (see Appendix 7 for a short analysis of VoroMQA versus ProQDock on which this decision was based). CASP14 organisers had also updated their assessor Z-score calculation, replacing GDT TS with TM-score as the score representing global quality alongside lDDT-oligo. Figures 3.3 and 3.4 reflect these changes in metrics and show plots equivalent to those in Figures 3.1 and 3.2 for McGuffin group CASP14 assembly models. Models were submitted for 14 out of 22 CASP14 targets.

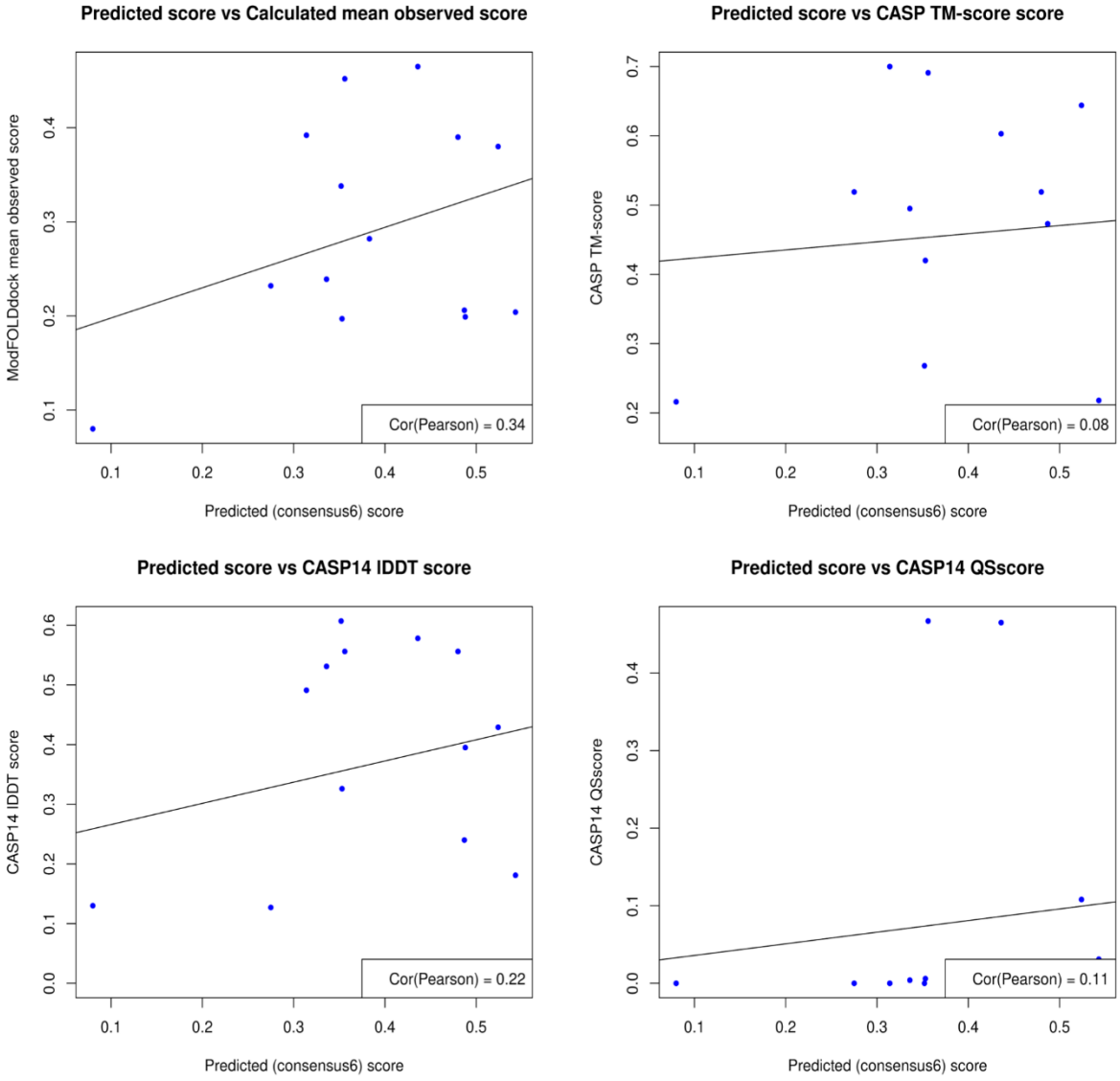


Figure 3.3. The correlation of ModFOLDdock Consensus6 score with observed scores for McGuffin group CASP14 assembly models. Top left. With observed mean score. **Top right.** With CASP14 TM-score. **Bottom left.** With CASP14 lDDT-oligo. **Bottom right.** With CASP14 QS-Score.

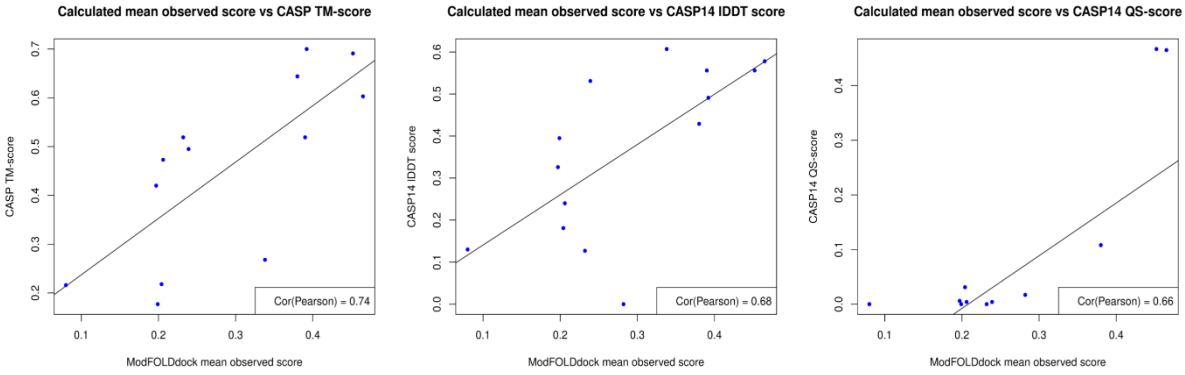


Figure 3.4. The correlation of mean observed score with CASP14 observed scores for McGuffin group CASP14 assembly models. Left. With TM-score. **Middle.** With lDDT-oligo score. **Right.** With QS-Score.

The plots in Figures 3.3 show similar trends for the CASP14 data to those seen for CASP13. ModFOLDdock predicted scores again correlated poorly with our own calculated observed score as well as CASP TM-score, lDDT-oligo and QS-score. Additionally, and again similarly to the CASP13 data, Figure 3.4 shows better correlations between the ModFOLDdock calculated observed score and the official CASP measures.

The conclusions drawn from the results were that the correlations obtained for observed scores suggested that ModFOLDdock components represented a set of promising metrics when compared to observed global superposition scores, but that there was a large accuracy gap between the observed and predicted scores. We aimed to reduce this accuracy gap via a program of optimisation of the predicted score combination. This can be a challenging process as it is unclear which aspect of model quality best represents overall accuracy, hence the multiple scores quoted by CASP and CAPRI, for example. Therefore, in order to achieve reliable optimisation, it was vital that a suitable single or very few target scores be identified.

3.1.5 Identifying a target score for optimisation is not immediately obvious.

Following the success of the first four CASP experiments which focussed mainly on tertiary structure prediction, CAPRI (**C**ritical **A**ssessment or **PR**ediction of **I**nteractions) was set up in 2001 (Lensink *et al*., 2018) as an additional experiment looking specifically at the prediction of protein interactions. Like CASP, CAPRI is a blind prediction competition using unpublished experimental structures which are supplied to participating groups. The expertise that CAPRI have accrued with quaternary structure models has led to the development of their own method for evaluating model quality compared to native structures, which relies on three related measures: Fnat, IRMS and iRMS. Fnat is the fraction of interfacial contacts expressed in the reference structure that are maintained or conserved in the model with a contact defined as any non-Hydrogen atom from either chain within 5Å. lRMS is a score representing the backbone RMSD of the (smaller) chain, deemed the ligand, within the complex upon superposition of the longer chain, deemed the receptor and iRMS is a score representing the RMSD of interfacial residues as measured by Cβ atoms with a distance cut-off of 10Å (sometimes 8Å) between the superposed native and predicted structures (Basu and Wallner, 2016a). These are used to define four quality classes as follows:

- Incorrect: Fnat < 0.1 or both lRMS > 10 and iRMS > 4.0
- Acceptable: Fnat between 0.1 and 0.3 and either LRMS ≤ 10.0 or iRMS ≤ 4.0 or Fnat ≥ 0.3 and both LRMS > 5.0 and iRMS > 2.0
- Medium: Fnat between 0.3 and 0.5 and either LRMS ≤ 5.0 or iRMS ≤ 2.0 or Fnat ≥ 0.5 and both LRMS > 1.0 and iRMS > 1.0
- High: Fnat ≥ 0.5 and either LRMS ≤ 1.0 or iRMS ≤ 1.0

There are some problems, however, with using the CAPRI scoring routine as a target score for optimisation: firstly, that the quality classifications do not easily equate to a single numerical scale but, also, that the calculations for each class rely on knowledge of both the predicted and native structure. Nevertheless, Basu and Wallner (Basu and Wallner, 2016a) were able to adapt the routine into a combined numerical predictor called DockQ (described in Section 3.1.3) which has proved to be a popular score due to its documented correlation with the CAPRI quality classifiers and has been chosen as a comparator score for a number of published studies (Johansson-Akhe and Wallner, 2022, Pozzati *et al*., 2022). DockQ was therefore initially considered as a potential target score.

In 2014 CASP held the first of two joint CASP/CAPRI experiments. This represented CASP round 11 and CAPRI round 30 and led to a vital sharing of information and technologies between the two groups. A second joint experiment was held two years later in CASP12/ CAPRI-37 and this set the precedent for the inclusion of a blinded complex modelling category in all future CASP competitions. CASP organisers have routinely ranked participant groups based on combined calculated Z-scores across a number of measures. For example, in CASP12 group rankings for TBM domains were determined using the combined Z-scores of the following scores: GDT_HA + (SG+lDDT+CAD)/3 + ASE (see Appendix 1 for score definitions). In a similar vein, CASP assessors' formula for assembly structures has been developed around the combined Z-scores of four methods and, like the CAPRI assessment, they comprise a mix of local interface and global similarity scores. The four individual scores are ICS (Interface Contact Score), often referred to as F1, IPS (Interface Patch Score), often referred to as Jaccard, lDDT-oligo (local Distance Difference Test for oligomers) and GDT_TS (up to CASP13) replaced by TM-score (CASP14 and above). See Appendix 1 for ICS and IPS definitions and calculation formulae.

Sum Z-scores for overall group rankings were calculated as an unweighted mean of:
Z-score(F1) + *Z-score(Jaccard)* + *Z-score(lDDT-oligo.)* + *Z-score(GDT_TS)* in CASP13 and
Z-score(F1) + *Z-score(Jaccard)* + *Z-score(lDDT-oligo.)* + Z-score(TM-score) in CASP14.

If a Z-score can be considered as simply a statistically normalised version of the raw score, it follows that the higher the value of the raw score the higher the value of the equivalent Z-score and therefore the higher the contribution of that Z-score to a groups' ranking position. Considering this relationship between raw Z-score and successful modelling, the magnitude of these four numerical scores was considered an important indicator of model quality. Just as importantly, though, is that these scores can be used separately or in combination to assess different aspects of model quality. These scores represented alternative target scores to

DockQ and to test whether combinations of these scores did indeed offer greater potential flexibility, three artificial scores were defined as:

- Local score: a calculated unweighted mean of F1 and Jaccard,
- Global score: a calculated unweighted mean of lDDT-oligo and GDT_TS/TM-score
- Total score: a calculated unweighted mean of all four scores.

Matrices of Pearson correlation coefficients were then produced to compare these scores with ModFOLDdock observed scores, which conveniently also included a DockQ score, using two datasets of CASP13 and 14 data (see 3.3 below for a description of the dataset).

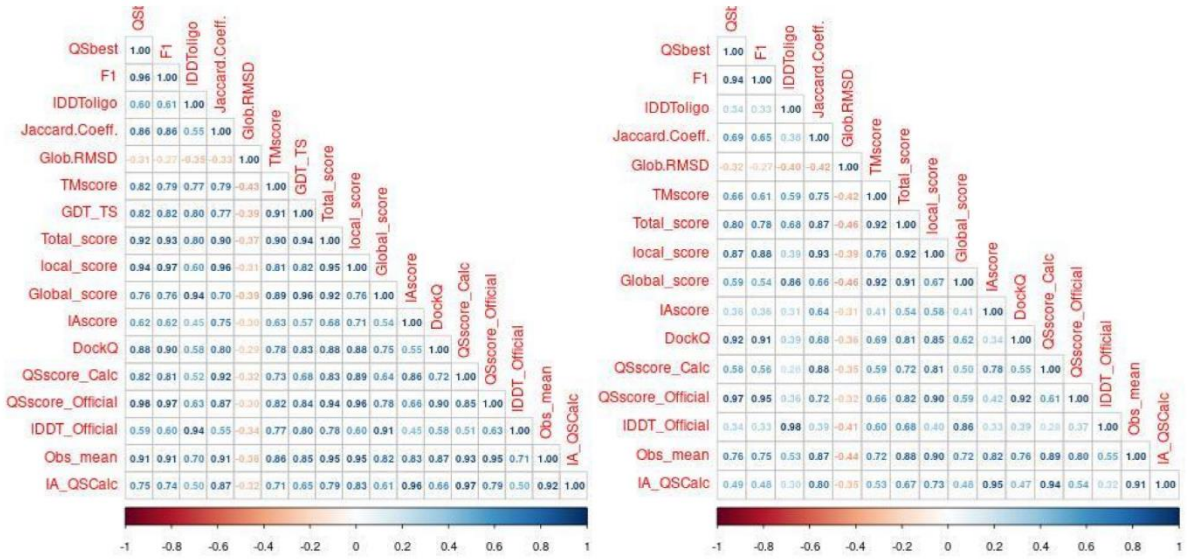


Figure 3.5. Pearson correlation matrices of CASP assessor scores with each ModFOLDdock observed score using CASP13 and 14 assembly models from all CASP groups. Left. CASP13 data**. Right.** CASP14 data.

From both matrices in Figure 3.5 it can be seen that DockQ has a strong correlation with QSscore_Official (0.90 for CASP13 and 0.92 for CASP14 data). However, the calculated Local score has a similarly strong linear relationship with both QS-scores (0.89/0.96 for CASP13 and 0.81/0.90 for CASP14 data), in addition the calculated Global score has a strong correlation with lDDT_Official (0.9 CASP13 and 0.86 CASP14), which is not seen for DockQ. As a result, it was considered that using the in-house calculated CASP scores would indeed offer a greater flexibility in assessing aspects of both local and global model quality whereas DockQ might offer a more limited assessment. Also, the complexity of the input values for DockQ score gives it a high variability of contributing factors, that is to say that when optimising to DockQ there could be uncertainty whether improvements in agreement represented those in global superposition, chain orientation or interface contacts (all of which contribute to the overall DockQ score). Using separate target scores might give a clearer signal about the individual conformational contribution to improvement and so calculated Local, Global and Total scores were selected as target scores for ModFOLDdock optimisation.

3.2 Objectives

The objective of this investigation was to identify the maximum level of agreement between the ModFOLDdock predicted scores and the three scores defined above (Local, Global and Total) with emphasis on the Local and Global scores. The primary outcome was that optimally combined ModFOLDdock scores would show improved agreement with the target scores beyond their consensus baseline level. This is described by the following two hypotheses:

1. *H0: There is no relationship between ModFOLDdock predicted scores and the combined CASP local quality measures ICS and IPS. H1. Individual ModFOLDdock predicted scores can be combined to form strong positive correlations with combined CASP local quality measures.*

2. *H0: There is no relationship between ModFOLDdock predicted scores and the combined CASP global quality measures lDDT-oligo and either GDT TS (for CASP13 data) or TM-score (for CASP14 data). H1. Individual ModFOLDdock predicted scores can be combined to form strong positive correlations with combined CASP global quality measures.*

The primary outcomes will be measured by Pearson correlation coefficient, ROC plot and associated AUC value with an additional Wilcoxon signed rank test performed on observed scores as a confirmatory measure. All of these should show improvement over consensus baseline values.

3.3 Materials and Methods

To investigate the hypotheses, a dataset was created containing all CASP13 and 14 assembly models. CASP scores for all modelling groups were taken from the CASP raw data tables available at https://predictioncenter.org/casp13/results.cgi?view=targets&trtype=multimer and https://predictioncenter.org/casp14/results.cgi?view=targets&trtype=multimer. The dataset comprised all homomeric models of varying stoichiometries and difficulty ratings along with CASP assessor scores and, in total, comprised 3282 models over 44 CASP targets (T0960 – T1087). Models were rescored with ModFOLDdock on a per target basis and both predicted and observed scores were applied. Due to the high number of models, the rescoring process required a substantial time investment and it was decided to exclude all heteromeric targets from the dataset due to the high CPU demands experienced when predicting the scores of those with higher order stoichiometry.

3.3.1 Objective data processing using an RSNNS Neural Network (NN)

An objective method of investigation, and one that offered potential for revealing hidden relationships within the ModFOLDdock score population, was the creation and supervised training of a neural network. The R Stuttgart Neural Network Simulator library (RSNNS), (Bergmeir and Benitez, 2012) was used to create a neural network with the architecture of a simple feed-forward multi-layer perceptron (MLP) with one hidden layer, an example of which is depicted in Figure 3.6. This was chosen due to the flexibility of the program described in the above article and personal familiarity with R programming. The ROCR library was used to create Receiver Operator Characteristic (ROC) plots and corresponding Area Under the Curve (AUC) metrics to measure performance of the classification by True Positive Rate (TPR) over all thresholds of False Positive Rate (FPR).

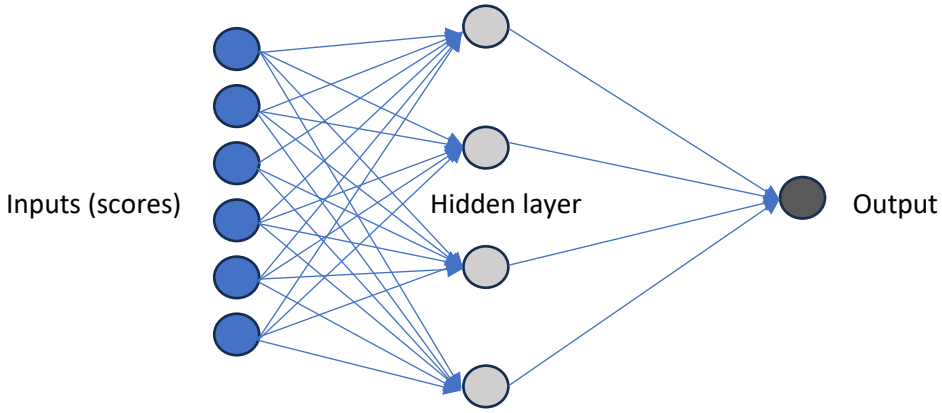


Figure 3.6. A schematic of a single hidden layer MLP NN with six inputs similar to that programmed in this study.

The main pitfalls with neural networks are accidental over- and underfitting. Overfitting results when a powerful network learns the whole training dataset rather than the trends within the

data. This is often revealed when the predicted and true-label data are plotted and characterised by a very close fit between the distributions of the two variables with a high variability in the regression line matching the distribution. Underfitting results from poor learning and is often characterised by a poor fit to the data distribution and low variability (inappropriate straight line) in a regression plot. These problems can often be limited by using a recognised supervised learning technique coupled with hyperparameter optimisation.

3.3.2 Ensuring fair score distribution - three-fold cross validation.

Supervised training, when a target training value (referred to as a true label) is supplied, was undertaken using the 3-fold cross validation method. This approach attempts to control for problems that may be encountered when using a dataset that is simply split to form a single training and testing set. The most notable issue with this simple approach is that the data may not follow a random distribution in each set, meaning that one set could have more data in the correct or incorrect class or that the numerical magnitude is substantially uneven between the sets. These differences may lead to the NN performing well on the training dataset but poorly on the testing dataset (underfitting). Cross validation allows every data point in the whole dataset to be included fairly in both training and testing stages (an example R program is included in Appendix 14).

To set up the cross-validation, CASP scores were first used to calculate the Local, Global and Total target scores as described at the end of Section 3.1.5. Three subset arrays were then defined containing models for different targets; subset1 contained 15 targets, subset2, 15 targets and subset3, 14 targets. Targets were assigned using a random generator and resulted in the following subset populations.

subset1 (T0999 T1038 T0977 T0997 T1083 T1054 T0989 T1016 T1048 T1003 T1087 T0984 T0983 T1020 T0966)

subset2 (T0963 T0995 T1001 T1018 T0976 T0998 T0965 T1061 T1062 T1010 T1070 T1078 T1000 T1080 T1084)

subset3 (T0970 T1006 T0961 T1032 T0973 T1034 T0960 T0979 T1004 T0981 T0996 T0991 T1009 T0985)

Test and training datasets were then created from these subset arrays for each of the CASP Global, Local, and Total scores. During programming, the Global score was tested first and so the process will be described for this score only but it was then repeated in exactly the same way for the Local and Total scores.

- Training_set1 comprised data from subset 2 and 3 but **no** data from subset 1.
- Training_set2 comprised data from subset 1 and 3 but **no** data from subset 2.
- Training_set3 comprised data from subset 1 and 2 but **no** data from subset 3.

This organisation is represented in Figure 3.7.

Training set 1 →	subset 1	subset 2	subset 3	→ score 1
Training set 2 →	subset 1	subset 2	subset 3	→ score 2
Training set 3 →	subset 1	subset 2	subset 3	→ score 3
	Testing set 1	Testing set 2	Testing set 3	

Figure 3.7 The model populations used for supervised MLP training. Those selected are in grey and those omitted are in white for each training and testing dataset.

Adopting this strategy ensured there was no overlap of targets in individual training and testing datasets (the full list of targets in each can be found in Appendix 8). The data within each training and testing dataset was programmatically shuffled into a random order preventing any bias in score distribution. For each master dataset, two further datasets were then created which contained only input (ModFOLDdock scores) or output (Global score) scores. A binary cut-off was created to allow the calculation of true and false predictions, which were used to populate the confusion matrices and determine the TPR and FPR for the ROC calculations. To do this the predicted scores were compared to the target Global scores to calculate a difference. The difference value was then used to ensure scores were correctly rounded to one decimal place. Finally, the scores were converted to binary values using 0.5 as the cut-off value, i.e. score > 0.5=1 and score ≤ 0.5=0.

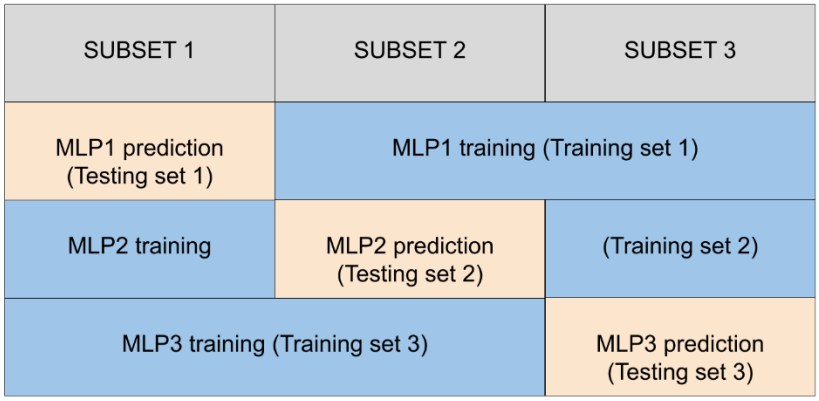


Figure 3.8. A diagram showing the training and testing subsets used in 3-fold cross validation for MLP 1, 2 and 3.

The neural networks, henceforth referred to as MLPs, were then created to predict on each testing set after supervised learning using each corresponding training set. Three separate MLPs were defined with different network weights, one for each training and testing set combination, i.e. MLP1 was trained on training set 1 and tested on testing dataset 1, MLP2 was trained on training set 2 and tested on testing set 2 and MLP3 was trained on training set 3 and tested on testing set 3 as shown in Figure 3.8.

Initially the setup was performed using default hyperparameters (row 1 of Table 3.2 below, values in red) to obtain starting point weights. Before the full cross validation was run, hyperparameter optimisation was required to find the values giving the best performance (described in full in Section 3.3.4). Once this was complete, predictions from each of the three MLPs could be run using optimised hyperparameter settings. As defined in the Objectives section, Pearson correlation coefficients, ROC plots and AUC calculations were used as the primary outcome measures for improvement over baseline data. Additional data analysis was also conducted using the LM-style measures adjusted R-squared, residual standard error and maximum standardised residual values (the latter as a measure of the residual size in standard deviation units). These metrics were used to provide additional insight into the relationships within the regression models. All comparison values were calculated with reference to calculated CASP Local, Global or Total scores.

3.3.3 Creating baseline and observed values for comparisons.

Baseline values for Pearson correlations, ROC plots and AUC values were manually calculated using two different predictor values versus each of the three target scores. To do this, testing sets 1, 2 and 3 inputs were combined with testing sets 1, 2 and 3 outputs minus the NN training and MLP prediction stage. These baseline values could then be directly compared to post-training values to quantify any improvement. The two baseline predictor values were:

1. The ModFOLDdock Consensus6 predicted score.
2. The optimal combination of one or more individual ModFOLDdock predicted scores.

A third set of values representing the optimal combination of one or more individual ModFOLDdock *observed* scores was also created. These scores would act as the theoretical maximum agreement that the predicted scores could reach.

3.3.4 Fine-tuning the RSNNS MLPs – Hyperparameter optimisation.

During setup with default hyperparameter values it was noticed that the lowest values for both individual score correlations and all score predictions was consistently achieved with comparisons to calculated Local score. This, then, appeared to be the most difficult score to predict (which seems reasonable due to the nature of predicting interfacial contacts). The Local score program was therefore used for initial hyperparameter testing.

The four hyperparameters included in test variations were *learning rate*, *maximum difference considered an error (Max, error)*, *maximum iterations (Max It.)* and *number of hidden neurones (Size)*. Ten different variations of hyperparameters were created and the results were assessed by Pearson correlation coefficient and ROC AUC for each of the three MLPs.

Table 3.2. RSNNS MLP hyperparameter testing variations and performance results for local scores. Data was collected using the combined MLP 1, 2 and 3 training datasets.

Hyperparameter	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Learning rate	0.01	0.05	0.01	0.1	0.01	0.01	0.01	0.01	0.01	0.01
Max Diff.	0.01	0.01	0.05	0.1	0.01	0.01	0.01	0.01	0.01	0.01
Max It.	100	100	100	100	200	100	100	100	200	500
Size	4	4	4	4	4	5,4	5,4,2	4,2	4,2	4
Performance	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MLP1 correlation	0.91	0.91	0.91	0.91	0.92	0.91	0.9	0.9	0.91	0.91
MLP2 correlation	0.83	0.83	0.84	0.84	0.84	0.83	0.82	0.83	0.83	0.83
MLP3 correlation	0.93	0.93	0.93	0.93	0.94	0.93	0.93	0.93	0.93	0.93
All correlation	0.87	0.87	0.87	0.88	0.89	0.87	0.87	0.87	0.87	0.87
MLP1 ROC AUC	0.988	0.989	0.988	0.988	0.989	0.989	0.988	0.989	0.989	0.989
MLP2 ROC AUC	0.943	0.94	0.943	0.94	0.941	0.939	0.94	0.94	0.942	0.94
MLP3 ROC AUC	0.989	0.989	0.99	0.99	0.99	0.99	0.989	0.99	0.989	0.989
All ROC AUC	0.97	0.972	0.973	0.972	0.973	0.972	0.972	0.973	0.973	0.972

Variation 5 (shaded grey) resulted in the overall best performance indicators (as calculated by a mean score across all performance values) and so hyperparameters were set to: learning rate = 0.01, Maximum Diff. = 0.01, Maximum It. = 200 and hidden nodes (Size) = 4 for the cross-validation process. This process was performed twice more for Global and Total target scores and very similar results were obtained. The Global target score program showed the best performance with the same hyperparameter settings listed above whereas the Total score program performed better with a Maximum It. of 100.

3.3.5 Iterative and regression errors – checking for over and underfitting.

Plots were created for iterative error with loss measured by the sum of squares error (SSE) across iterations. For these graphs the training loss is represented by the black line and the estimated validation loss, i.e. that which would be encountered on unseen data, is represented by the red line. For both, a smooth downward curve is desirable; a curve which remains high, particularly for the validation loss may represent underfitting, whereas an upward trend in either line may represent overfitting. The iterative error and regression error plots for MLP1, 2 and 3 are shown below.

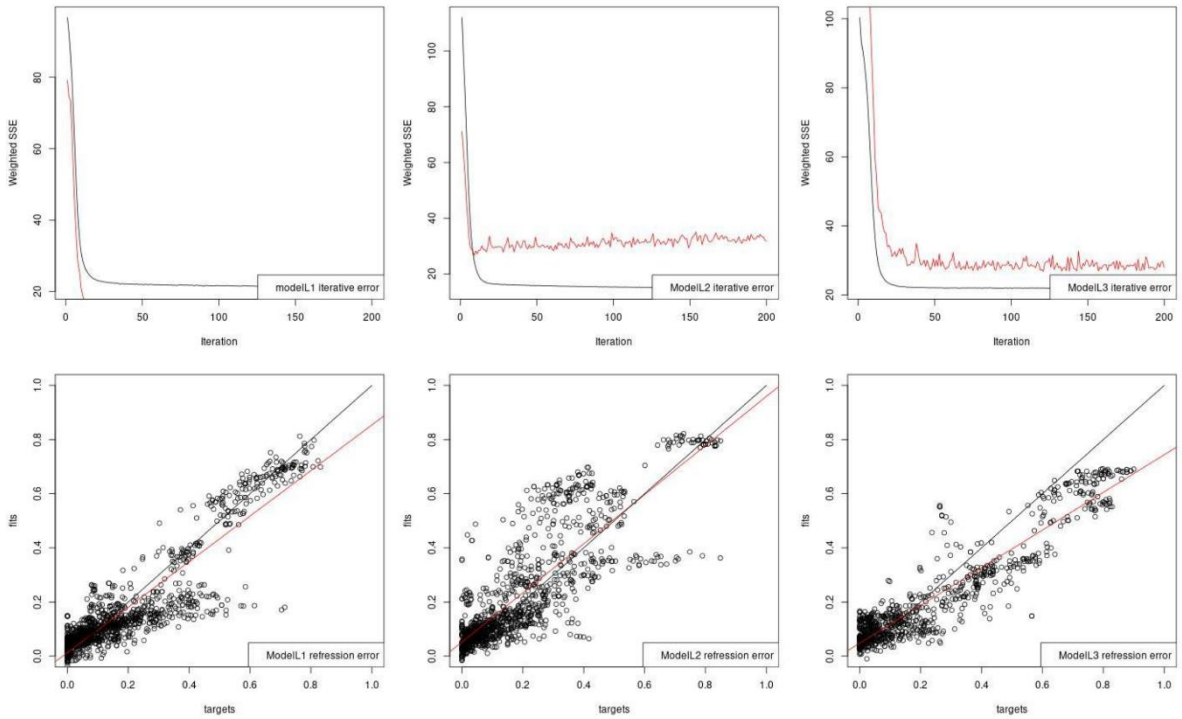


Figure 3.9 Iterative and regression error plots for the three RSNNS MLPs. Top. Iterative error for MLP1 (left), MLP2 (middle) and MLP3 (right). **Bottom.** Regression error for the same MLPs. Data was collected using the combined MLP 1, 2 and 3 training datasets.

The iterative error plots in the top row of Figure 3.9 show that the MLPs have been well trained on each respective training dataset as shown by the smooth downward curve. All three plots show little further improvement and plateau after approximately 20 iterations which is indicative of each model reaching a point where further training iterations do not significantly improve its fit to the training data. This assumption is supported by the validation error line for MLPs 1 and 3 which show decreases with no sustained increase. MLP2, however, does show evidence of underfitting with a larger difference between the training and validation data. To assess whether this was problematic, the regression error and supporting statistics for this MLP were checked. From the lower three graphs in Figure 3.9 it can be seen that MLP2 actually has the lowest deviation between ideal (black) and test (red) regression lines but that the data from training set 2 has the widest scatter with more outliers. A look at the accompanying regression statistics for this model reveals coefficients with an estimated intercept of 0.015582 and an estimate for the training set of 0.760812. This shows that when true values are 0, estimated predicted values would be 0.015582 and that each 1-unit change in the true value would lead to a 0.760812 change in predicted value. Additionally, an R-squared value of 0.6898 indicates a relatively good fit, explaining about 68.98% of the variance in the dependent variable. An F-statistic, which measures the ratio of variance explained by regression to that explained by residuals, of 2591 with a p-value of $2x10^{-16}$ is highly significant, supporting the overall

significance of the MLP scores. In conclusion, MLP2 shows some deviation from ideals but predictions remain significant compared to actual values.

3.4 Results and Discussion.

3.4.1 The baseline values.

3.4.1.1 Results for the Consensus6 predicted score.

The following regression plots, ROC plots and AUC values shown in Figure 3.10 were produced using the combined training and testing datasets defined above but with no neural network training or prediction. These show the baseline relationships between the unweighted Consensus6 score with each of the calculated Local, Global and Total target scores. Comparisons of the primary outcome measures Pearson corelation and ROC AUC as well as additional LM-style regression metrics are presented in Table 3.3.

Table 3.3. Comparisons of two primary outcome measures (Pearson coefficient and ROC AUC - in bold) and LM-style regression metrics for baseline ModFOLDdock Consensus6 scores. Comparison values are calculated with reference to calculated CASP Local, Global and Total target scores using the combined training datasets but with no MLP prediction.

Regression statistic	Local	Global	Total
Pearson coefficient	0.80	0.765	0.835
Adjusted R-squared	0.64	0.58	0.69
Max. std. residual	4.54	3.15	3.58
Residual standard error	0.13	0.12	0.104
Statistical comparisons	Local	Global	Total
ROC AUC	0.966	0.875	0.948

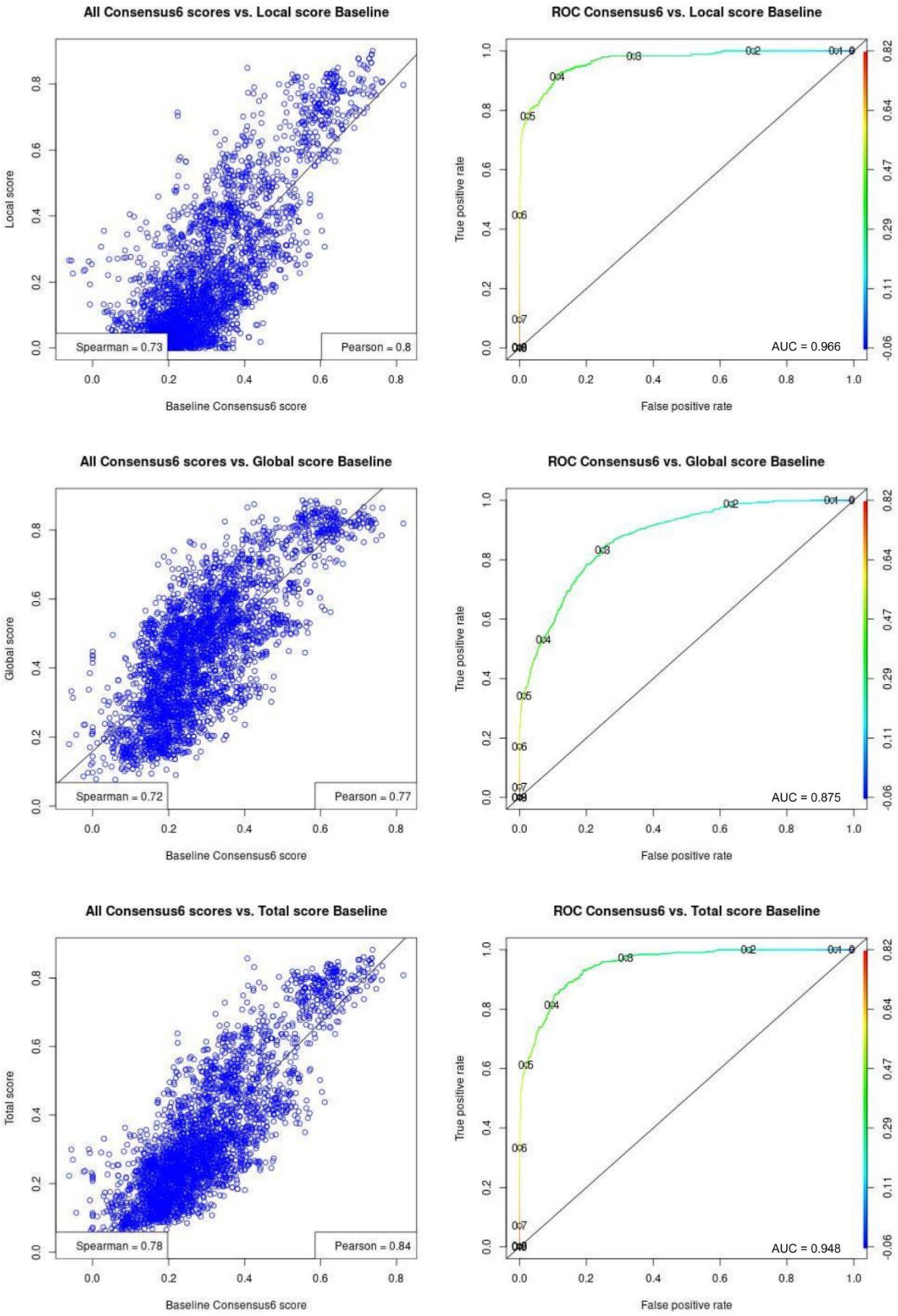


Figure 3.10 Scatter plots and ROC plots for ModFOLDdock Consensus6 score versus all target scores for the combined training and testing datasets. Top. Local score. **Middle.** Global score. **Bottom.** Total score. ROC plot right-hand axis shows AUC values, coloured blue to red for low to high values respectively. Values on the plotted line represent the thresholds used to calculate the AUC.

The results appear to show promising levels of agreement between the unweighted consensus score and the three selected target scores as measured by Pearson correlation coefficient. However, a closer look at the accompanying standardised residual values and the statistical data in Table 3.3 reveals that the spread of the data is relatively high with all three distributions having maximum standardised residuals greater than 3.0 (values greater than 3.0 are generally considered as representing outliers (Lin *et al*., 2017)) as well as R-squared values ranging from 0.58 to 0.69 showing that between 31-42% of the variation in the plots cannot be accounted for by the relationship between the scores.

3.4.1.2 The optimal combination of individual ModFOLDdock predicted scores.

Table 3.4 shows the same regression and statistical comparisons as Table 3.3 but this time for the optimal agreement between target scores and either any single or any combination of the ModFOLDdock predicted scores. Table 3.4 and the graphical data in Figure 3.11 below show that optimal agreement was seen between QSscoreOfficialJury and the Local target score, the unweighted mean of lDDTOfficialJury and QSscoreOfficialJury and the Global target score and the unweighted mean of lDDTOfficialJury, QSscoreOfficialJury and DockQJury and the Total target score.

Table 3.4. Comparisons of the two primary outcome measures (Pearson coefficient and ROC AUC - in bold) and LM-style regression metrics for baseline ModFOLDdock optimal score combinations. Comparison values are calculated with reference to CASP Local, Global and Total target scores using the combined training datasets but with no MLP prediction.

Regression statistic	Local	Global	Total
Pearson coefficient	0.89	0.88	0.90
Adjusted R-squared	0.78	0.77	0.80
Max. std. residual	5.13	4.12	4.85
Residual standard error	0.100	0.088	0.083
Statistical comparisons	Local	Global	Total
ROC AUC	0.976	0.931	0.971

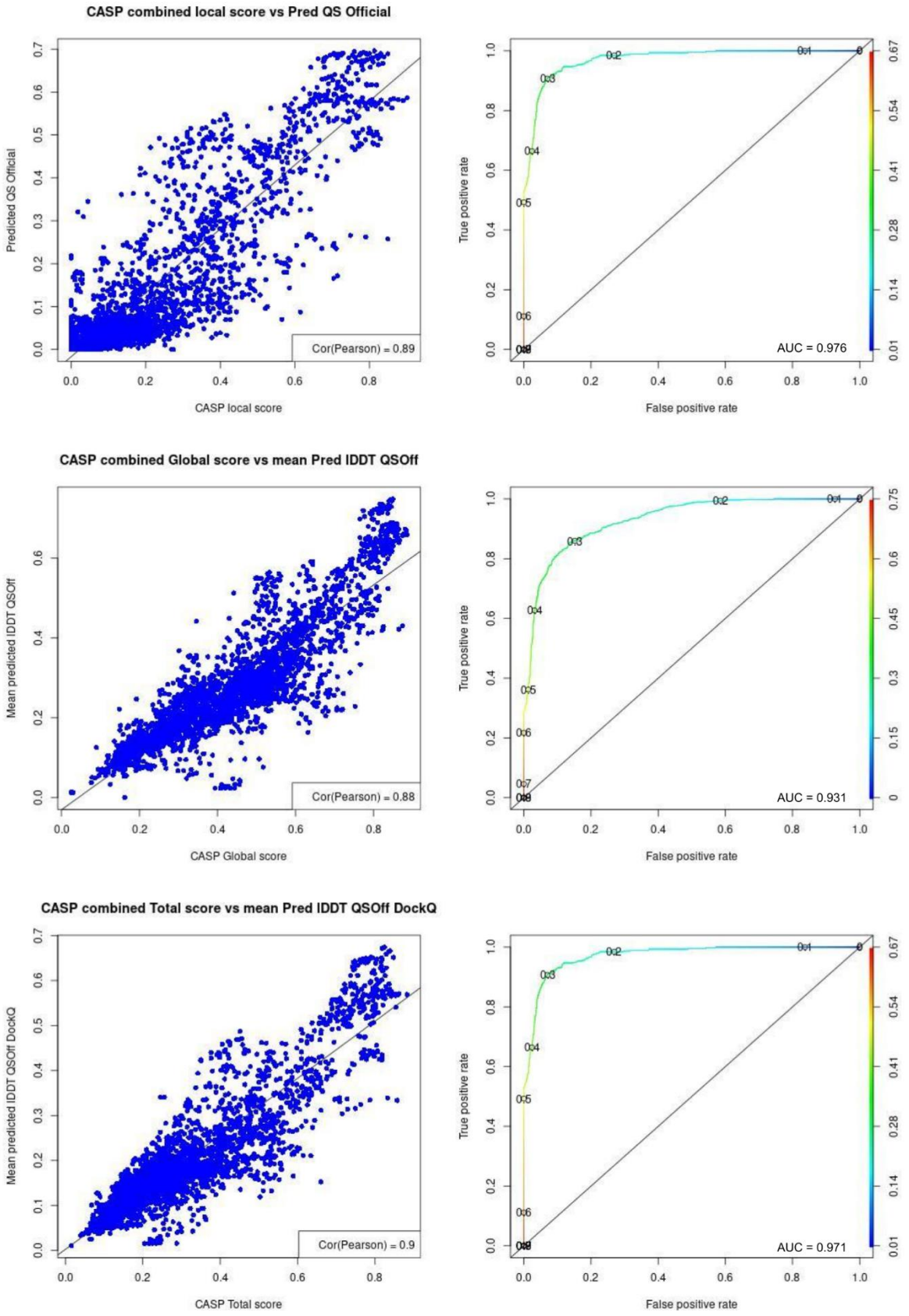


Figure 3.11 Scatter plots and ROC plots for optimal combinations of ModFOLDdock predicted scores versus all target scores for the combined training and testing datasets. Top. Local score. **Middle.** Global score. **Bottom.** Total score. ROC plot right-hand (y2) axis shows AUC values, coloured blue to red for low to high values respectively. Values on the plotted line represent the thresholds used to calculate the AUC.

Comparing this set of results to those in Table 3.3 and Figure 3.10 for the Consensus6 baseline, it can be seen that better agreements between the target scores and either single or combinations of the ModFOLDdock predicted scores are achievable. This is true for the Pearson correlation coefficients, ROC AUC values and R-squared values. However, the maximum standardised residual values have increased compared with those for Consensus6 scores suggesting that outliers may be fewer but possibly more extreme in value.

3.4.1.3 Results for optimal combination of individual observed scores.

Figure 3.12 shows the Pearson correlation coefficients obtained between the target scores and optimal combinations of observed scores similar to those in Figure 3.11. Table 3.5 summarises the baseline and observed correlation and ROC AUC primary outcome measures from sections 3.4.1.1 and 3.4.1.2 for easy comparison. Values obtained for observed scores should, in theory, represent the maximum agreement obtainable between target and predicted scores.

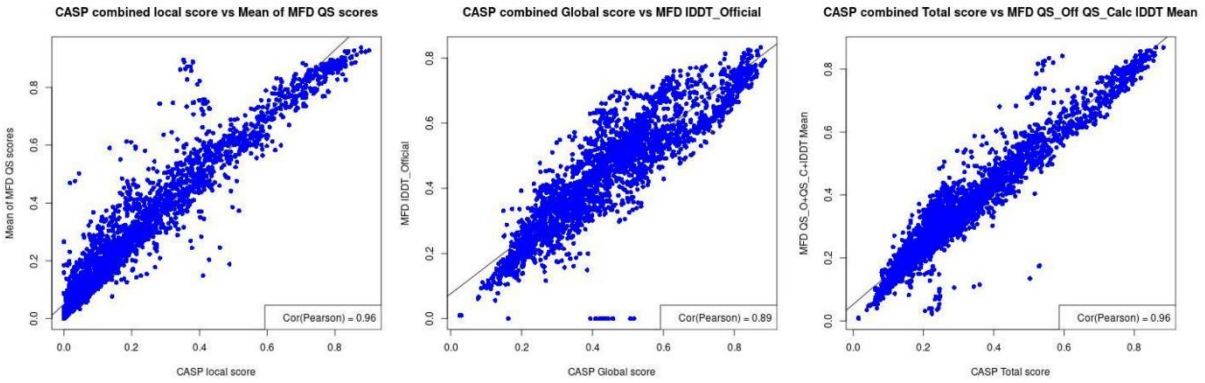


Figure 3.12 Scatter plots for optimal combinations of observed scores versus all target scores for the combined datasets. Left. Local score. **Middle.** Global score. **Right.** Total score.

Table 3.5 A comparison of Pearson correlation and ROC AUC primary outcome measures between ModFOLDdock baseline (Consensus6 and optimally combined) and observed scores and all three target scores.

Target score	ModFOLDdock predicted baseline score	Correlation coefficient	ROC AUC
Local	Consensus6	0.80	0.966
Global	Consensus6	0.77	0.875
Total	Consensus6	0.84	0.948
	Optimally combined predicted baseline score		
Local	QSscoreOfficialJury	0.89	0.977
Global	(lDDTOfficialJury + QSscoreOfficialJury) /2	0.88	0.931
Total	(lDDTOfficialJury + QSscoreOfficialJury + DockQJury) /3	0.90	0.971
	Optimally combined observed scores		
Local	(QSscoreOfficial + QSscore_Calc) /2	0.96	0.995
Global	lDDTOfficial	0.89	0.934
Total	(QSscoreOfficial + QSscore_Calc + lDDTOfficial) /3	0.96	0.988

The improvement in Pearson correlation and ROC AUC values between the Consesnsus6 scores and those for manually created optimal combinations is summarised in Table 3.5 along with further improvements seen for the observed score combinations. This data suggests that improved agreements are possible with optimal combination of the ModFOLDdock scores and that there remains room for improvement up to a ceiling shown by the observed score combinations. It is therefore reasonable to postulate that a neural network may be able to improve optimal combinations beyond that possible manually. The key outcomes of a successful neural network training and prediction process are therefore a further increase in the Pearson correlation, ROC AUC and R-squared values, with a simultaneous reduction in the magnitude of the residual standard error and maximum standardised residual values.

3.4.2 Three-fold cross validation results

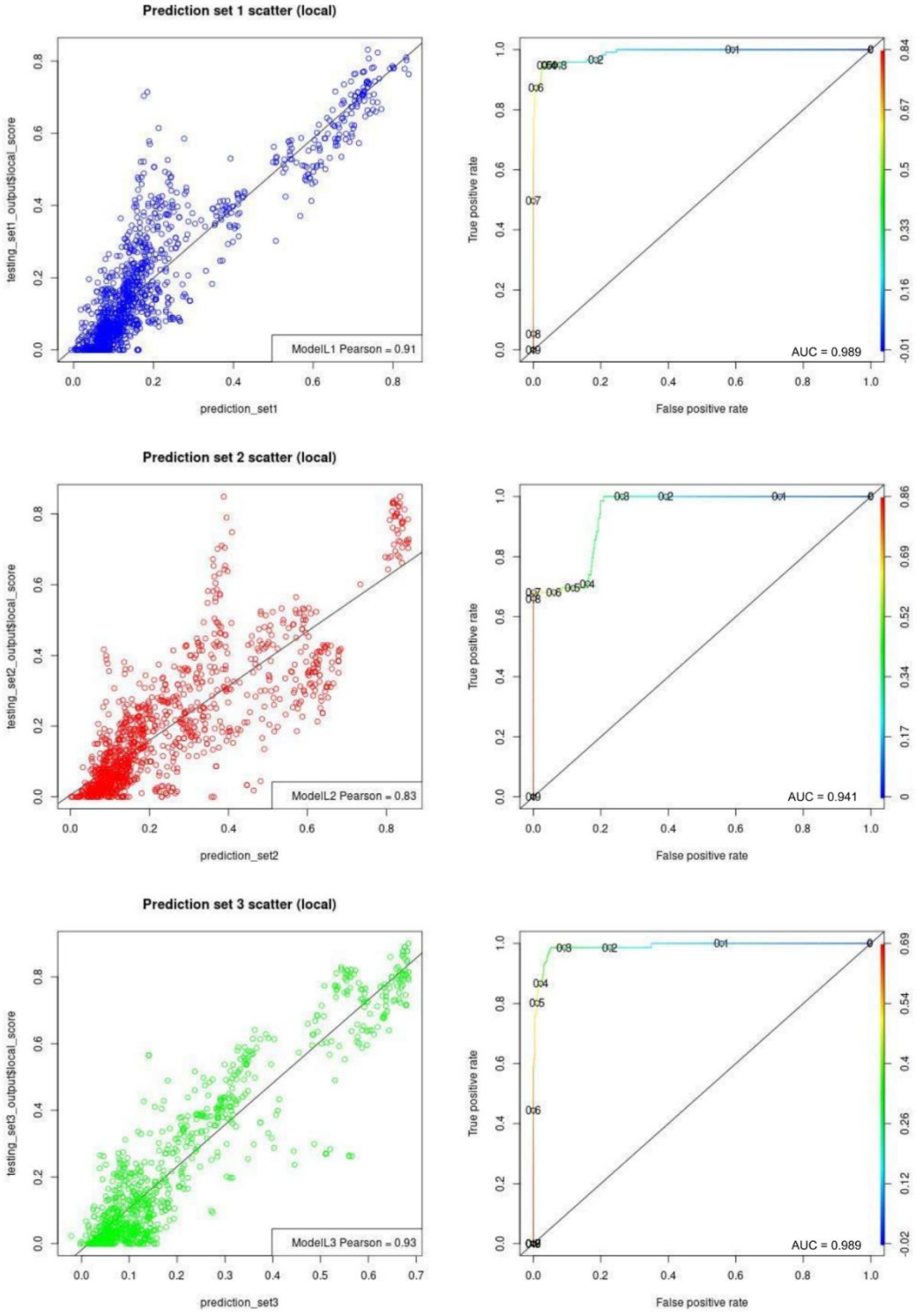


Figure 3.13 Scatter plots (left) and ROC plots for cross-validation of NN predictions of Local target score. Top. Results for MLP1. **Middle.** Results for MLP2. **Bottom.** Results for MLP3. ROC plot right-hand (y2) axis shows AUC values, coloured blue to red for low to high values respectively. Values on the plotted line represent the thresholds used to calculate the AUC.

The plots in Figure 3.13, for Local target score, show that the two primary outcome measures, Pearson coefficient and ROC AUC values, have increased beyond those achieved for both

Consensus6 and optimal baseline values for MLP1 and 3. However, for MLP2, while the Pearson coefficient has increased beyond the Consensus6 baseline value of 0.80 it has not exceeded the optimal baseline value of 0.89. The AUC has also reduced from baseline of 0.966 to 0.941.

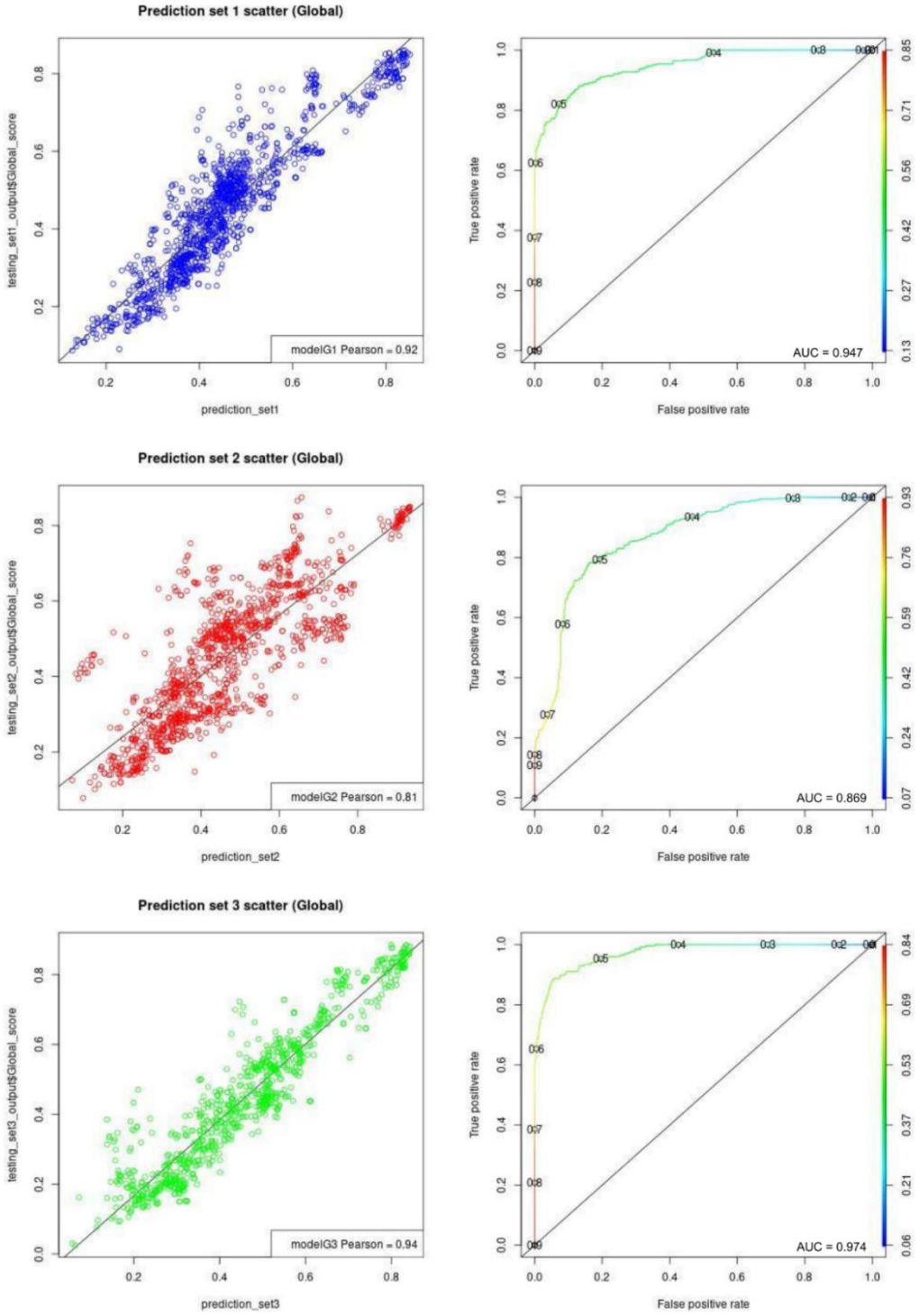


Figure 3.14 Scatter plots (left) and ROC plots (right) for cross-validation of NN predictions for Global target scores. Top. Results for MLP1. **Middle.** Results for MLP2. **Bottom.** Results for MLP3. ROC plot right-hand (y2) axis shows AUC values, coloured blue to red for low to high values respectively. Values on the plotted line represent the thresholds used to calculate the AUC.

Again, Figure 3.14 shows that the Pearson coefficient and ROC AUC score primary outcome measures for MLP1 and 3 have increased beyond baseline values for Global scores but, again the MLP2 Pearson coefficient has not exceeded the optimal combination baseline and the AUC value remains below both baseline values.

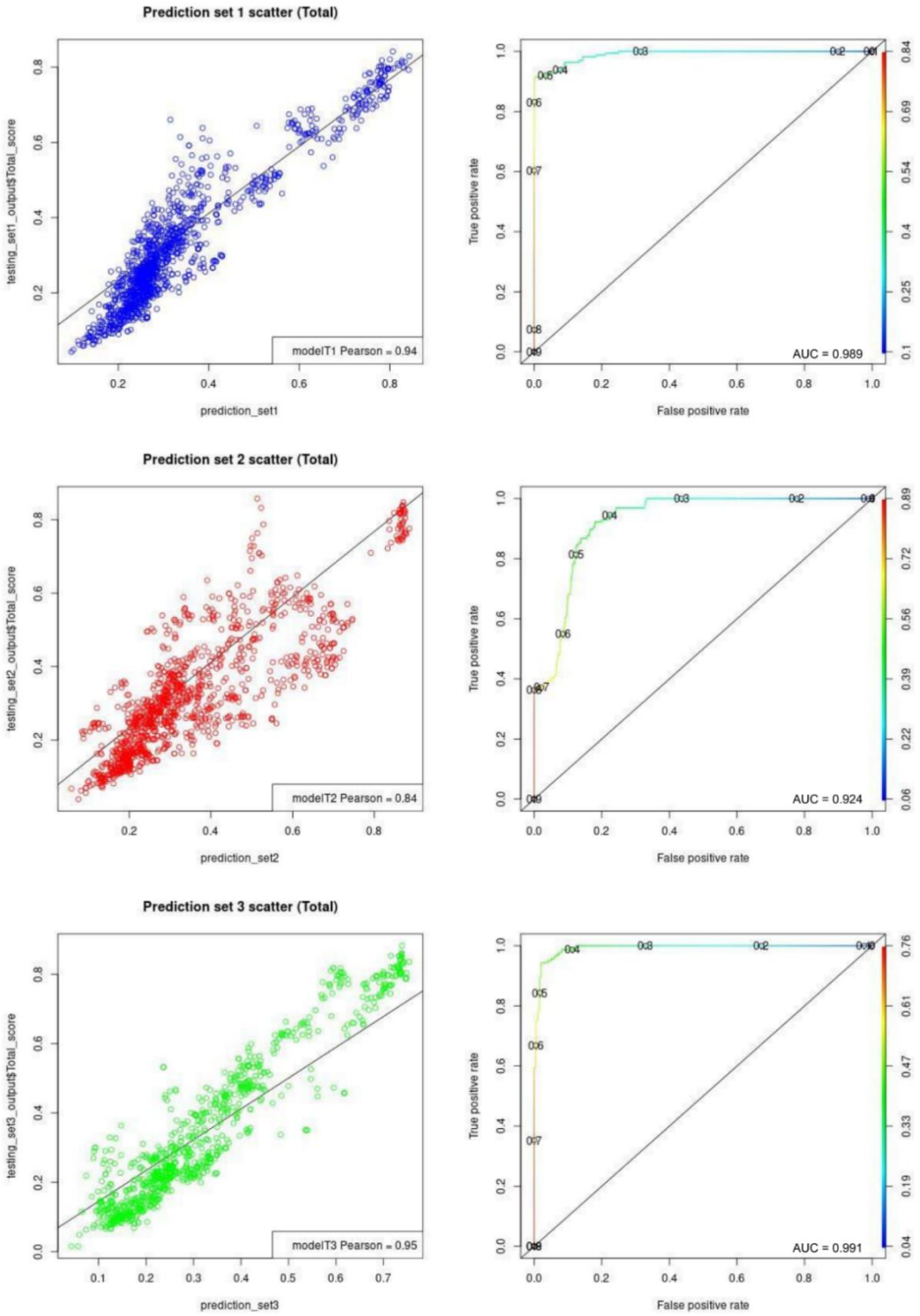


Figure 3.15 Scatter plots (left) and ROC plots (right) for cross-validation of NN predictions for Total target scores. Top. Results for MLP1. **Middle.** Results for MLP2. **Bottom.** Results for MLP3. ROC plot right-hand (y2) axis shows AUC values, coloured blue to red for low to high values respectively. Values on the plotted line represent the thresholds used to calculate the AUC.

Similarly, Figure 3.15 shows that MLP1 and 3 have again improved from baseline for Total score but MLP2 shows no improvement in either Pearson coefficient or ROC AUC value. These results suggest that the NN is training successfully on training sets 1 and 3 and predicting accurately on their respective testing sets. However, either training or testing set 2 appears to contain some data that is inhibiting successful training. Although all effort was made to ensure a random distribution of models across datasets, a closer inspection of the model populations reveals that models for targets T1061 (T5 phage tail subcomplex), T1070 (Escherichia virus CBA120) and T1080 (Bdellovibrio bacteriovorus), all of which were rated “Difficult” by CASP and mentioned as being poorly modelled by the majority of groups in the CASP14 Assembly Assessment (Karaca, 2020), are included together in testing set 2. It is possible that MLP2 has not had sufficient training on similarly poorly modelled structures in training set 2 to make accurate predictions for these structures in testing set 2.

3.4.3 Combining NN predictions to produce a final prediction result.

To create the following plots the results from MLPs 1, 2 and 3 (trained on set 1, 2 or 3) were combined to predict the final scores. As before, the predictions were compared by Pearson correlation coefficient and ROC AUC values.

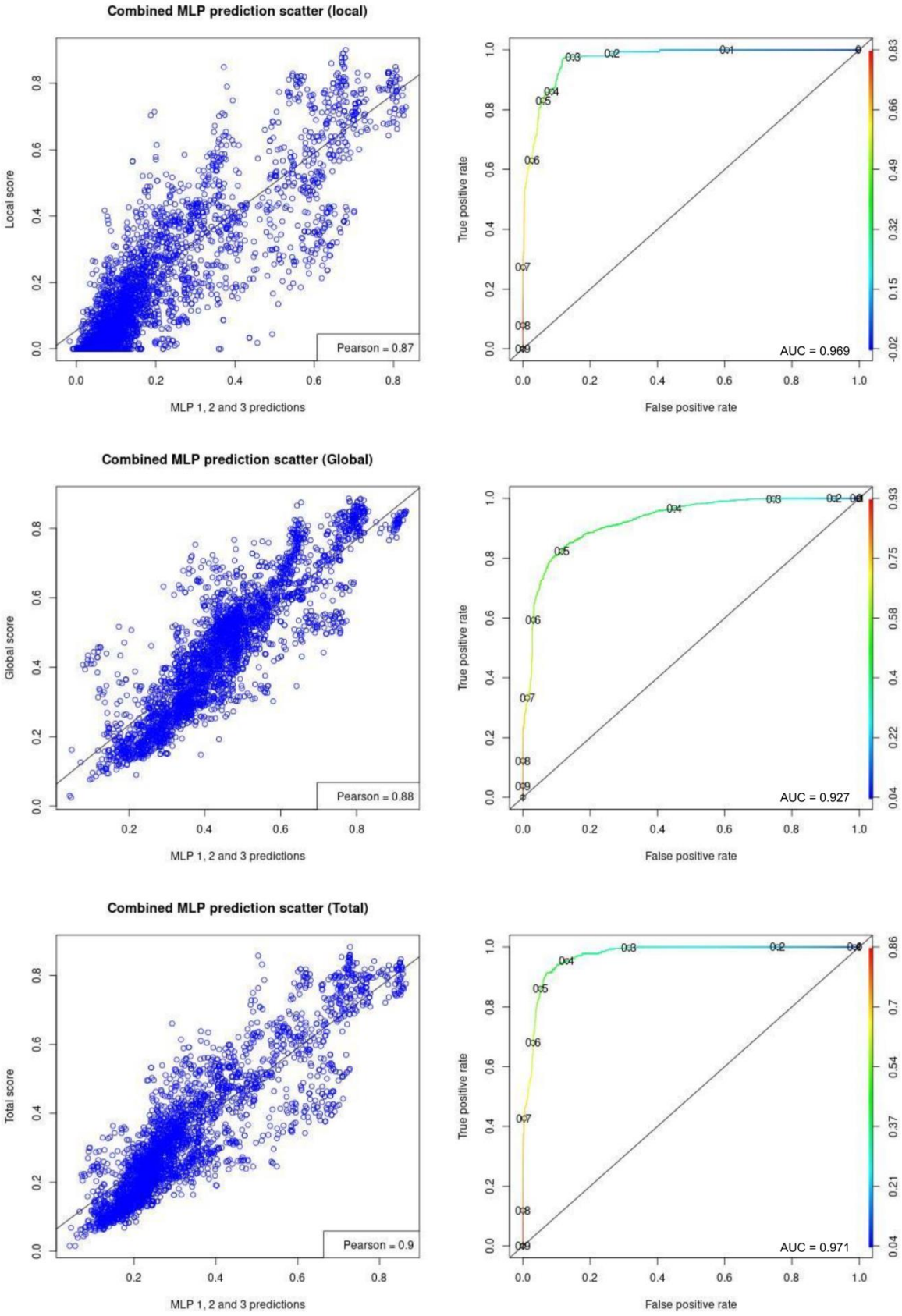


Figure 3.16 Scatter plots (left) and ROC plots (right) for predictions from the combined MLPs for each target score. Top. Local score. **Middle.** Global score. **Bottom.** Total score. ROC plot right-hand (y2) axis shows AUC values, coloured blue to red for low to high values respectively. Values on the plotted line represent the thresholds used to calculate the AUC.

To allow at-a-glance comparisons across all baseline, cross-validation and final prediction stages, the Pearson correlation coefficient and ROC AUC primary outcome measures from Figure 3.16 are collated with those from baseline and cross validation training in Table 3.6

Table 3.6. A comparison of primary outcome measures Pearson coefficient and ROC AUC values for the three combined RSNNS MLPs for all 3 target scores. C6 = Consensus6 baseline, Max=optimal combinations baseline.

Correlation score	Measure	Baseline		Cross-validation			Final Prediction
		C6	Max.	MLP 1	MLP 2	MLP 3	
Local	Pearson r	0.80	0.89	0.91	0.83	0.93	**0.87**
	ROC AUC	0.966	0.977	0.989	0.941	0.989	**0.969**
Global	Pearson r	0.77	0.88	0.92	0.81	0.94	**0.88**
	ROC AUC	0.875	0.931	0.947	0.869	0.970	**0.927**
Total	Pearson r	0.84	0.9	0.94	0.84	0.95	**0.90**
	ROC AUC	0.948	0.971	0.989	0.924	0.991	**0.971**

Table 3.6 shows that the effect of NN training for all three target scores has resulted in an increase in the primary outcome measures of Pearson coefficient and ROC AUC compared to their Consensus6 baseline values. In comparison to the optimal combinations baseline, the data are slightly less clear. NN predictions have equalled the baseline Pearson coefficient value for the Global score (0.88) and both Pearson coefficient (0.90) and ROC AUC (0.971) values for the Total score. For Local score both Pearson coefficient and ROC AUC fell slightly short of the values achieved with the optimal combination baseline, as did ROC AUC for Global score.

Table 3.7. A comparison of the LM-style regression measures for the three combined RSNNS MLPs for all 3 target scores. C6 = Consensus6 baseline, Max=optimal combinations baseline.

Correlation score	Measure	Baseline		Final Prediction
		C6	Max.	
Local	Adjusted R-squared	0.64	0.78	0.756
	Max. standardised residual	4.54	5.13	4.87
	Residual standard error	0.13	0.100	0.107
Global	Adjusted R-squared	0.58	0.77	0.781
	Max. standardised residual	3.15	4.12	4.30
	Residual standard error	0.12	0.088	0.087
Total	Adjusted R-squared	0.69	0.80	0.804
	Max. standardised residual	3.58	4.85	4.35
	Residual standard error	0.104	0.083	0.084

Table 3.7 shows the LM-style regression statistics for Consensus6 and optimal combination baselines as well as those for the final NN prediction. All R-squared values show an increase from the Consensus6 baseline suggesting a better fit to the regression line by the post-training values. R-squared values also increase over the optimal combination baseline values for

Global and Total scores. The maximum standardised residual values, however, have not returned to their Consensus6 baseline low and remain above the 3.0 outlier cut-off for all scores. The residual standard error scores have decreased from the Consensus6 baseline showing that the overall standard deviation of the regression residuals has reduced over this baseline but they are not noticeably reduced compared to the optimal combination baseline.

In summary, when measured against the primary outcomes, the results showed that the Pearson correlation coefficient and ROC AUC values had improved over the Consensus6 baseline values but were either slightly below or equal to those obtained for the optimal combination baseline. For the LM-style regression measures, the R-squared values improved beyond both baselines in two out of the three cases, and a simultaneous reduction in the magnitude of the residual standard error was seen in comparison with the Consensus6 baseline. Maximum standardised residual values, in general, did not decrease from baseline.

3.4.3.1 Results of a Wilcoxon signed rank test for significance.

There are some conflicting results in both the primary outcome measures of Pearson correlation coefficient and ROC AUC as well as the LM-style regression scores. To resolve these and determine more objectively whether there was significant improvement after neural network training, a Wilcoxon signed rank test for significance was performed between both sets of baseline scores and the final NN predictions. To do this, the top scoring model for each CASP target was determined for each of the Consensus6 and optimal combination baselines as well as the NN predicted scores. The observed scores associated with each of these top-ranked models were then compared using the Wilcoxon test. In this way a more objective measure of performance can be made by utilising observed scores which always have a higher degree of accuracy. The non-parametric Wilcoxon signed-rank test was chosen as scores were not normally distributed and a paired version was used, as each method predicts over the same target model set. The test was one-tailed to assess an increase in MPL predictions over both baselines. Top scoring models were sampled for each of the Local, Global and Total scores and the results are presented in Table 3.8.

Table 3.8. A comparison of observed scores for models ranked top (1) for each scoring method using a paired Wilcoxon signed rank test. Models were ranked by Local, Global and Total predicted scores and the equivalent associated observed scores were sampled for the test. P-values were calculated at the 95% confidence and significant values of ≤0.05 are in bold.

Score	Comparison	P-value
Local	Consensus6 baseline versus final prediction	**0.0146**
	Optimal combination baseline versus final prediction	0.7532
Global	Consensus6 baseline versus final prediction	**8.14×10^{-5}**
	Optimal combination baseline versus final prediction	0.2781
Total	Consensus6 baseline versus final prediction	**2.36×10^{-4}**
	Optimal combination baseline versus final prediction	0.6025

The results in Table 3.8 show more definitively that neural network (NN) training was able to significantly improve the prediction of all three scores when compared to the Consensus6 baseline. However, the process was not able to significantly improve upon the predictions for the optimally combined baseline scores. This was not to say that no improvement was detected - for each predicted score the sum of observed scores was always highest for the MLP prediction - just that the improvement was only significant compared to the Consensus6 baseline.

3.5 Conclusions

3.5.1 There is agreement between NN predictions and CASP assessor scores.

This study has established that there are promising levels of agreement between ModFOLDdock predicted model quality scores and the CASP official observed scores which had not been seen before. The results suggest that by referencing the CASP Z-score calculations intended to assign overall group rankings, three useful target scores representing the Local, Global and Total quality of the protein models could be determined. Furthermore, results from Pearson correlation coefficients and R-squared values along with ROC AUC values confirmed that predictions can be improved by using weighted combinations of the scores. The same measures also confirm that similar improvement in all three predicted scores can be achieved by using the target scores to train a simple multi-layer perceptron (MLP) prior to prediction. Lastly, it has been shown that, according to a Wilcoxon signed rank test, MLP training significantly improves all three scores over the original Consensus6 score.

It must also be noted that improvement over the optimal combination baseline was not consistently seen in Pearson correlation coefficient or ROC AUC values. Furthermore, one would have hoped to see a consistent improvement in the LM-style measures showing a better fit to the regression line for NN predictions, i.e. an increase in R-squared values (showing that a higher percentage of variance in one variable is explained by the other) and a decrease in both residual standard error and maximum standardised residual showing that the size of the

residuals is decreasing. Unfortunately, this was only seen in comparison to the Consensus6 baseline.

3.5.2 The lack of improvement beyond optimal combinations can be explained.

It may be possible that the failure of the MLP training to significantly improve predictions over the optimally combined baseline was the product of the relatively small size of the dataset (while there are over 3000 models in the dataset, these models represent only 44 distinct targets) coupled with two further limitations inherent in the data. The first of these limitations is to do with the design of ModFOLDdock itself. Namely that there are relatively few (six) inputs from contributing scores which may produce a narrow bandwidth of data for the MLP to interpret. That is to say that there may not be enough variation within the six scores for an anomalous result in one single score to be sufficiently outweighed by the others. The second issue is to do with the reliability of the scores within the dataset. While Global and Total scores appear to have been well predicted by the MLP, Local score prediction has been less successful. A look at the mean values reveals that the mean Local score was only 0.21 compared with 0.45 for Global score. This tendency for generally lower values throughout the dataset resulted in a more limited range of model quality for the MLP to interpret. The disparity in range between Local score (which is calculated from IPS and ICS score) and global score (calculated from TM-score and lDDT) is described in the official CASP14 assembly modelling review (Karaca, 2020) which described a low modelling success rate of 38% as measured by ICS compared to 86% as measured by TM-score (success was defined as scores >0.4). This was cited as evidence that the interface area was consistently less accurately modelled than the global fold in CASP14 models. This may explain the decreased performance of the MLP for Local score prediction.

It is also likely that increasing both the size of the training dataset and the number of data points supplied to the NN would enhance accuracy. Both of these changes would also likely allow the size of the NN to be increased without fear of overfitting which would result in a deeper NN architecture.

3.5.3 The data support the hypotheses.

Notwithstanding these issues, the results obtained do support both hypotheses outlined in the objectives, i.e. that individual ModFOLDdock scores can indeed be combined to form strong positive correlations with combined CASP Local and Global quality measures. Further to this, the increase in agreement achieved between MLP predicted and target score for all scores is statistically significant compared to the original Consensus6 score. These results are important as they reveal that the simple consensus approach used up until this point was masking potent information hidden within the ModFOLDdock constituent scores. As such ModFOLDdock now

represents a MQAP with potential to reliably distinguish between native-like and decoy models of protein multimeric complexes.

Post CASP15 (2022) the ModFOLDdock MQAP was made publicly available via the IntFOLD website (https://www.reading.ac.uk/bioinf/ModFOLDdock/).

CHAPTER 4

Independent performance benchmarking of MultiFOLD and ModFOLDdock using CASP15 data

Work presented in this chapter has been published in the following paper:

Estimation of model accuracy in CASP15 using the ModFOLDdock server. *Edmunds, NS, Alharbi, SMA, Genc, AG, Adiyaman, R & McGuffin, LJ.* Proteins. 2023.

Individual author contributions are as follows.
Edmunds NS: Development of ModFOLDdock, conceptualisation, writing & editing.
Alharbi SMA: CASP15 modelling and EMA assistance.
Genc AG: CASP15 modelling and EMA assistance.
Adiyaman R: Technical guidance.
McGuffin LJ: Conceptualisation, overview, guidance, software, review & editing.

Cited in the text as (Edmunds *et al.*, 2023).

And also contributed to this publication:

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

Individual author contributions are as follows.
Liam J. McGuffin: Conceptualisation, project administration, data curation, writing & editing.
Nicholas S. Edmunds: MultiFOLD and ModFOLDdock conceptualisation, data curation, analysis, review & editing.
Ahmet G. Genc: MultiFOLD conceptualisation, data curation and analysis, review & editing.
Shuaa M. A. Alharbi: Conceptualisation, data curation, analysis, review & editing.
Bajuna R. Salehe: Software development.
Recep Adiyaman: IntFOLD conceptualisation, data curation, analysis, review & editing.

Cited in the text as (McGuffin *et al.*, 2023).

4.1 Background

If CASP14 was notable for the unprecedented increase in tertiary structure prediction accuracy achieved by AlphaFold, CASP15 was also notable for a definite shift in emphasis toward multimeric or quaternary structure modelling. This was demonstrated, not only by an increase in the number of assembly targets, up to 41 from only 22 in CASP14 (37 of which were shared CAPRI targets), but also by the inclusion of an estimation of model accuracy (EMA) category for quaternary structures for the first time. For the EMA competition, specific score definition and submission formats called QMODE1 and QMODE2 were required and the work described in this chapter builds upon the three new ModFOLDdock consensus scores (*localscore*, *globalscore* and *totalscore*) identified in Chapter 3. Part one of the results describes the correlations and ranking agreements achieved during the QMODE2 calibration process and part two documents the successful ModFOLDdock performance using data from the CASP15 assessors' official analysis. Also considered in a post-CASP analysis is the effect of enhanced ModFOLDdock accuracy on MultiFOLD modelling performance as well as comparisons to previous CASP competitions.

4.1.1 ModFOLDdock updates

In Section 3.1.4 the inclusion of a VoroMQA score into the CASP14 ModFOLDdock pipeline and its use in calculating an extended consensus score was explained. However, VoroMQA had not formally replaced ProQDock as a single-model component method within the program code itself. With the ModFOLDdock updates that were undertaken to meet CASP15 requirements, this change was now also included. A second addition was made to the ModFOLDdock code base, and this was prompted by the positive results seen at CASP14 for the ModFOLD8 tertiary MQA server (McGuffin *et al.*, 2021) and which were, in part, due to an increased contribution of the Contact Distance Agreement (CDA) score (Maghrabi and McGuffin, 2017). It was realised that it would be possible to create a multimeric version of the CDA score by direct sampling of the AlphaFold2 contact map created during the modelling process (see section 4.3.1 for the methodology) and so work was undertaken to add a multimeric CDA score as a seventh ModFOLDdock constituent score. Due to these updates and the additional EMA requirements, a second round of finer-grained ModFOLDdock optimisation was now required in addition to that already described in Chapter 3. This round is henceforth referred to as the "QMODE2 calibration".

4.1.2 The QMODE specifications

The new EMA category was solely concerned with multimeric (assembly) models and required submissions of scores within 48 hours of the release of each model population. Competing groups were required to submit scores for all models (often in the region of 300) as they were released, on a target-by-target basis, in either of two formats: QMODE1 or QMODE2. Both

QMODE formats required a global score (SCORE), in a 0-1 range, as an estimate of the overall accuracy of the whole modelled complex. This score was mandatory. A second score (QSCORE), also with a 0-1 range, and intended to reflect the overall accuracy of the model interface, was specified for both QMODE formats, but its inclusion was optional. QMODE2 additionally required a series of individual residue-level confidence scores (again with a range of 0-1). These were to be applied to all amino acid residues located on different chains where the Cβ to Cβ (Cα for Glycine) distance was measured as ≤8Å. These were intended to reflect the likelihood of the identified interface residues in the model matching the interface residues of the native structure. Figure 4.1 below shows an example of the required QMODE2 format and all ModFOLDdock variants were programmed to submit all three scores for QMODE2. The QMODE format description is available from:
https://predictioncenter.org/casp15/index.cgi?page=format#QA.

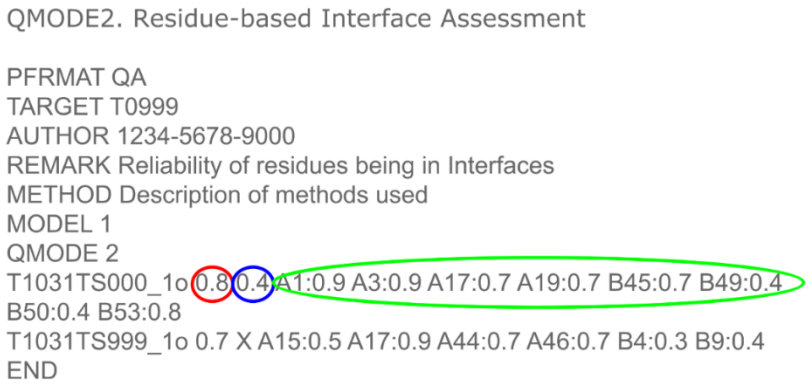


Figure 4.1. QMODE2 scoring requirements for the CASP15 EMA competition.
Ringed in red: the global score (SCORE); ringed in blue: the overall interface score (QSCORE) and ringed in green: the residue-level confidence scores. Image taken from (Edmunds *et al.*, 2023).

4.1.3 TS format updates for modelling

The only change in the modelling format was that the B-factor column, which is used as a residue accuracy measure, now needed to be populated with a predicted lDDT-like score (plDDT) with a range of 0-100 instead of a displacement estimate in Ångströms. This meant that higher scores would now signify a closer predicted agreement with the native structure rather than a more distant one.

4.2 Objectives

The previous chapter described how the CASP14 “Global” and “Local” scores used for Z-score calculations took the form of unweighted means of lDDT-oligo plus TM-score and F1 or interface contact score (ICS) plus Jaccard coefficient or interface patch score (IPS) respectively. In the context of the CASP15 EMA scores, the CASP14 “Global” score could be considered broadly comparable to the CASP15 “SCORE” and the CASP14 “Local” score to the CASP15 “QSCORE”. It was reasoned then, that considering the demonstrated relationships of predicted *localscore* and *globalscore* to their observed Local and Global score counterparts, comparable projected ModFOLDdock score combinations might be used to generate the SCORE and QSCORE for the QMODE2 files. As a result, three main objectives and one secondary consideration were established for the QMODE2 calibration.

The first objective was to identify the new maximum agreements which could be obtained between ModFOLDdock predicted scores, which now included the CDA and Voronota-JS derived VoroMQA scores as components, and the observed Global and Local scores as proxies for SCORE and QSCORE respectively.

The second objective was to optimally combine predicted scores into an individual residue confidence score.

The third objective was to modify the output of MultiFOLD and ModFOLDdock to report similarity or "plDDT" scores, scaled to 0-100, in the B-factor to conform with TS format requirements and to update all contact identification to 8Å to conform with EMA requirements, respectively.

The secondary consideration was to explore the relationship between the combinations of predicted scores optimised for either correlation or ranking, i.e., are these the same or different optimal combinations of scores?

4.3 Materials and Methods

4.3.1 Justifying a closer focus on interface contacts within ModFOLDdock

In section 4.1 the CDA score's promising contribution to ModFOLD8 performance was mentioned. There were, however, three further reasons for introducing an adaptation of this score as well as an updated version of the VoroMQA score into ModFOLDdock. Firstly, the poor ICS scores seen in CASP14 (Karaca, 2020) represent an obvious area for improvement and it was likely that a greater emphasis on interface contacts would be required to achieve modelling success in CASP15. Secondly, the CASP15 EMA criteria specifically required a residue-level confidence score for each amino acid calculated to be within the model interface. Scores which were directly based on interface or contact identification would therefore likely make valuable contributions to this score. Lastly, it was considered important to maintain some single-model methods in the MQA pipeline as they often have superior performance compared to clustering methods in cases when there are few variations between models or when only few models are considered (Elofsson *et al.*, 2018).

4.3.2 The multimer CDA score calculation

This is an adaptation of the tertiary structure Contact Distance Agreement score (Maghrabi and McGuffin, 2017) which compares contact probabilities from contact prediction software such as DeepMetaPSICOV (Kandathil *et al.*, 2019) to the Euclidean distance (measured in Ångströms) of equivalent atom pairs within a model. The CDA score is determined for any residue in the model having a Cβ-Cβ within 8Å and is calculated as:

$$(\sum p)/c$$

where p is the predicted contact probability and c is the number of residue-residue contacts in the model where p has a value. The quaternary structure version operates using a similar concept except that the contact probability values are instead supplied by the AlphaFold2 contact map which is conveniently generated during LocalColabFold (v1.0.0) modelling. The same logic is used for the multimeric calculation, and the CDA score for any residue with Cβ-Cβ $\leq$ 8Å will again be $(\sum p)/c$ where p is the LocalColabFold contact map probability and c is the number of residue-residue contacts in the model where p has a value.

4.3.3 The Voronota-js-VoroMQA score calculation

This was calculated using the Voronota-JS JavaScript expansion of the core Voronota software (Olechnovic and Venclovas, 2014) called voronota-js-voromqa. As part of this release version, it was still possible to continue to calculate the overall VoroMQA score from the core software (now referred to as the *Voro-light* score) as well as an updated set of scores known as *Voro-dark* scores. Key among these were two scores referred to as "*global*" and "*interface energy*" in the Venclovas group's description of their modelling and quality assessment

process for CASP13 (Olechnovic and Venclovas, 2017). These scores, along with an "*interface atoms"* score, are produced by the *-inter-chain* qualifier and are output as the *full_dark_score* and *sel_energy* scores which equate to the global and interface energy scores respectively. The Venclovas team used these to create a tournament scoring scenario, the format of which was to compare two models (A & B) to create a win, lose or draw result as follows: If A scores higher than B in all 3 scores, A wins; if any of the 3 scores disagree a draw is declared; if A scores lower than B in all 3 scores, A loses. The *-tour-sort* qualifier runs this function on an all-against-all basis and assigns a final rank to each model.

The Venclovas group have been consistently highly placed in CASP competitions and so it was considered a worthwhile time investment to re-score the full CASP13 and 14 dataset used in Chapter 3 with full Voro-Dark scores using the *-interchain* option as described above, as well as invoking the tournament scoring function for each individual target. Although the time-consuming tournament scoring did not produce rankings well correlated with the observed scores (maximum Pearson coefficient 0.22), promising correlations were seen between the Voro full_dark_score and CASP lDDT-oligo score (Pearson coefficients of 0.77 and 0.71, with the CASP13 and CASP14 data, respectively). From this evidence it was concluded that, while it was not worthwhile formally recreating tournament scoring as part of our QA pipeline, the underlying Voronota-JS full_dark_score likely contained important, if slightly orthogonal, information about model interface quality. It was therefore decided that the full_dark_score using the *-interchain --output-dark-scores* command switches, representing the Voronoi tessellation score for interface atoms, would represent a useful additional interface-focussed score.

4.3.4 A CASP14 dataset and manual comparisons were the best choices for the QMODE2 calibration

There is a general consensus of opinion that models from the latest CASP experiment represent the most up-to-date modelling techniques and state-of-the-art technology (Kryshtafovych *et al.*, 2019) and it is therefore preferable to use these data whenever possible. This viewpoint prompted a decision about the makeup of the dataset used for QMODE2 calibration. On one hand was the undeniable validity of the above statement, which would favour using a CASP14 only dataset. On the other, was documented analysis showing that, in general, interfaces had not been well modelled in CASP14 (as explained in Chapter 3, section 3.5.2), shown by a clear difference in mean ICS scores of 20.89 for CASP13 models compared to 6.58 for CASP14 models. In mitigation of this and in favour of CASP14 data are the three following points: firstly, the mean IPS scores were much more similar (0.31 versus 0.23) meaning that the identification of the interface patch was roughly equivalent over the two experiments but contact identification appeared to be lacking for CASP14 models. Secondly,

CASP14 had a much lower percentage of easy targets (6.9%) compared to 28.5% in CASP13, meaning that templates for the full assembly, including the interface, were less common for the later experiment. Lastly, CASP14 modelling still showed a shift towards higher scores in general compared to CASP13 (Karaca, 2020). One last but important point to consider was the likely CPU-load and time investment required to rescore all models with the updated version of ModFOLDdock. To re-score a combined CASP13/14 dataset with more than 3000 models was considered too time intensive in light of the available window until the start of CASP15. Based on these considerations a CASP14 dataset was selected for QMODE2 optimisation. The dataset comprised all models (2060) submitted by all groups for 17 CASP14 targets T1032, T1034, T1038, T1048, T1054, T1062, T1070, T1078, T1080, T1083, T1084, T1087, H1036, H1045, H1047, H1065 and H1072. This set of targets represented the population for which native structures were available from the CASP prediction centre website at the time and for which ModFOLDdock predicted scores could be generated within a 24-hour timeframe.

The manual comparison method used for this optimisation, mentioned in the section title and described in detail below, was also adopted in consideration of the time constraints. An ideal scenario would have seen the optimisation achieved by a further round of neural network training with an improved MLP design (see Chapter 6 for improvement details).

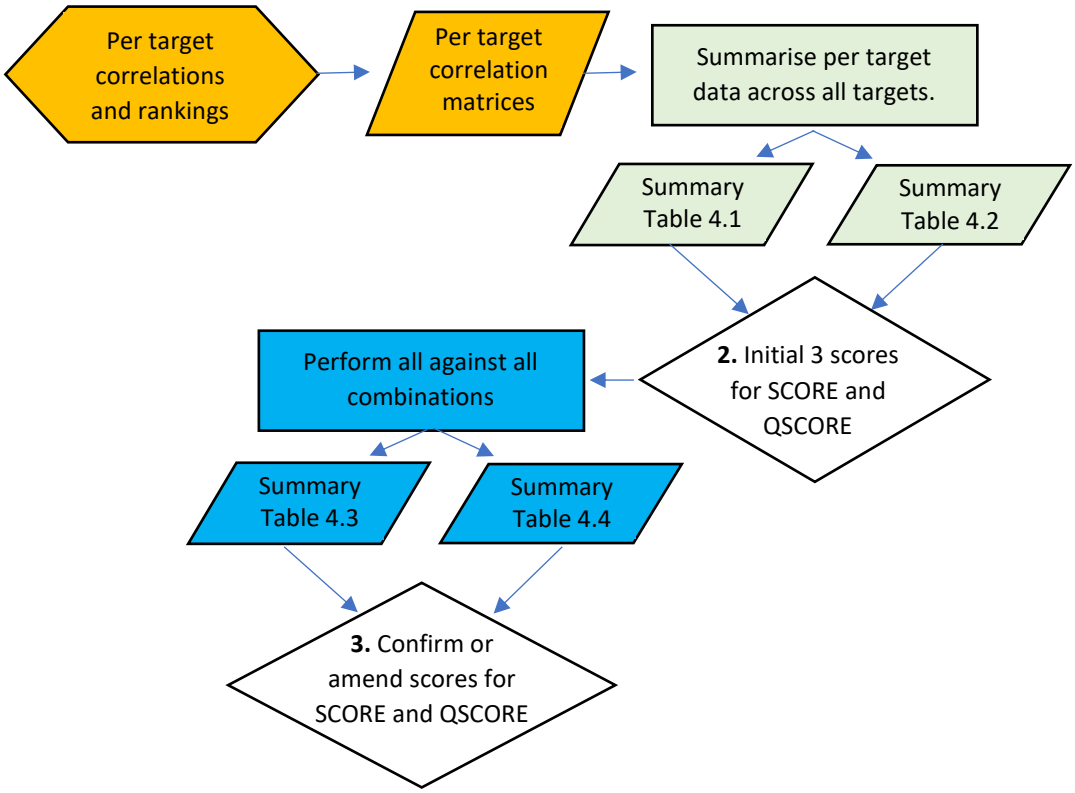


Figure 4.2. A work flowchart of the QMODE2 manual ModFOLDdock optimisation process. Stage 1 processes are coloured yellow, stage 2 are coloured green and stage 3 are coloured blue. Decision points 2 and 3 are coloured white.

Instead, the comparisons were performed on an iterative basis with each comparison stage informing decisions about the format of the next. As an objective MLP was not used, this exploratory method helped to reduce any bias resulting from the strong pre-existing relationships described in Chapter 3, notwithstanding their importance empirically or as a basis for this work. The workflow described in the following sections is summarised in the Figure 4.2 flowchart for ease of interpretation. At each stage, comparisons were considered separately, either for correlation with observed scores or agreement with observed score ranking.

4.3.5 Per target correlation comparisons (stage 1)

For each target, Pearson correlation coefficients were calculated between all ModFOLDdock predicted scores and the following observed scores: both observed QS-scores in the ModFOLDdock pipeline (*QSscore_Calc* and *QSscore_Offici*al) and the CASP observed scores *QS-glob, F1 (ICS), lDDT-oligo, Jaccard coefficient (IPS)* and *TM-score* as well as the Local, Global and Total calculated target scores. The selection of these scores was justified as follows. The QS-score was a contributing component to all three ModFOLDdock score combinations which produced maximal baseline correlations and ROC AUC values with Local, Global and Total scores in Chapter 3. Therefore, all available QS-scores were included. The CASP scores *ICS, lDDT-oligo, IPS* and *TM-score* were chosen as they are the assessor scores for Z-score rankings which were used to calculate the target Local, Global and Total scores for the MLP training in Chapter 3. Results are presented as Pearson correlation matrices for individual targets in Figure 4.3 (heteromer targets) and Figures 4.4A and B (homomer targets).

4.3.6 Per target top-rank comparisons (stage 1)

Each model quality score (lDDT, QS-score etc.) assesses quality according to an individual calculation and therefore each of the ModFOLDdock component scores may select a different top-ranked model from a model pool. In cases where observed scores are available, one way to estimate the true quality of the top-ranked models is to use the sum of the observed scores for each model as a quality metric. On a per target basis, then, models were ranked in turn by each ModFOLDdock component score and the full set of associated observed scores (*IAscore, DockQ, QSscore_Calc, QSscore_Official, lDDTOfficial, QSglob, F1, lDDT-oligo, Jaccard, TM-score, Local, Global and Total*) were summed to produce the quality metric (*obs_sum)*. The full results table is included in Appendix 9 and data from 4.3.5 and 4.3.6 were fed into stage 2 tables 4.1 and 4.2 respectively.

4.3.7 Cross target comparisons (stage 2)

To determine the overall relative strength of agreement between predicted and observed scores, the data from stage 1 processes were used to create two cross-target listings. Data from the stage 1 correlation matrices (4.3.5) was averaged across all targets to produce mean

cross-target correlation values. These were intended to display the average cross-target performance of each predicted score versus the key observed scores ICS, IPS, TM-score, lDDT-oligo, Local and Global scores in order to identify those scores likely to positively contribute to SCORE and QSCORE. Results are presented in Table 4.1 where the best average coefficients or those ≥ 0.5 are highlighted as potential contributing scores. Data from the stage 1 cumulative top rank scores (4.3.6) was summed across all targets to identify predicted scores with consistent high-ranking performance. Results are shown in Table 4.2. Highlighted data from Tables 4.1 and 4.2 were used to inform the initial score combination decisions as shown in Figure 4.2 decision box 2.

4.3.8 Final comparisons calculated against QMODE score proxies (stage 3)

For stage 3 comparisons, the target observed scores were limited to the Global and Local calculated scores. These were intended to act as proxies for the QMODE-defined SCORE (global fold) and QSCORE (global interface) scores respectively. For correlation data, all possible combinations of ModFOLDdock component scores were calculated and the mean correlation values were then compared with the two target scores. This meant that scores A to G, representing the seven component scores, were considered individually and in every combination and Peason, Spearman and Kendall correlation coefficients were calculated for each of these combinations against the two target scores. Table 4.3 shows the key results from this process. The cumulative top-rank data was treated similarly, with all possible combinations of ModFOLDdock component scores calculated but this time using the cumulative observed scores from each of the two target scores to estimate the quality in terms of global fold and global interface for each top-ranked model. Table 4.4 shows the key results.

The number of combinations considered for the stage 3 processes was defined as follows. For 7 scores there are a total of 7! =5040 permutations. However, as the order of the score combinations is unimportant the unique combinations are reduced according to the formula:

$C(n,k) = \frac{n!}{k! \times (n-k)!}$ Each term can be calculated and then summed, thus:

$C(7,1) = \frac{7!}{1! \times (7-1)!} = 7$ $C(7,2) = \frac{7!}{2! \times (7-2)!} = 21$

$C(7,3) = \frac{7!}{3! \times (7-3)!} = 35$ $C(7,4) = \frac{7!}{4! \times (7-4)!} = 35$

$C(7,5) = \frac{7!}{5! \times (7-5)!} = 21$ $C(7,6) = \frac{7!}{6! \times (7-6)!} = 7$ $C(7,7) = \frac{7!}{7! \times (7-7)!} = 1$

Adding them up: 7+21+35+35+21+7+1=127 unique combinations. Full results tables can be found in Appendix 10.

4.4 Results and Discussion

The results are presented in two parts. Part 1 describes the results obtained from the QMODE2 calibration processes just described. Part 2 uses the official independent assessment data from CASP15 to show benchmarking comparisons of ModFOLDdock and MultiFOLD performance against other state-of-the-art MQA and modelling software.

4.4.1 Part 1. Results for QMODE2 calibration

Decision point 2. Results from the correlation experiments are shown in Figures 4.3, 4.4A and 4.4B and summarised in Table 4.1. These suggested that the component scores most likely to contribute positively to the global fold score (SCORE, labelled as Global in Table 4.1) were ***lDDTOfficialJury*** and ***VoroMQA*** with Pearson correlation coefficients of 0.79 and 0.59 respectively. The next highest coefficient was for ***QSscoreOfficialJury*** with a value of 0.47 which is only slightly below the 0.5 threshold defined for a moderate correlation. These component scores are highlighted in bold in Table 4.1 and confirm the results from the MLP training and prediction process in Chapter 3 which also selected lDDTOfficialJury and QSscoreOfficialJury as global score contributors. These latest results showed that the newly added Voronoi-JS VoroMQA score should be also considered for inclusion.

Similarly, scores likely to contribute positively to the global interface score (QSCORE, labelled as Local in Table 4.1) were ***QSscoreOfficialJury***, ***DockQJury*** and ***QSscoreJury***, with coefficients of 0.58, 0.46 and 0.42 respectively (also in bold in Table 4.1). These results partially confirmed those seen in Chapter 3 where the highest correlation coefficient was achieved by QSscoreOfficialJury alone. However, these results suggested that DockQJury and QSscoreJury should also be considered at this stage.

Results from the top-rank calculations shown in Table 4.2 showed that the most likely scores contributing to ranking by global fold score (SCORE) were ***QSscoreOfficialJury***, ***VoroMQA*** and ***lDDTOfficialJury***, which were in agreement with the correlation results. However, results for the global interface score (QSCORE) agreed with only two of the three scores suggested by the correlation results. ***DockQJury*** and ***QSscoreOfficialJury*** were again selected but ***VoroMQA*** was preferred to QSscoreJury.

At this stage it was clear that the ranking results in Table 4.2 were the result of different score combinations than the correlation results in Table 4.1. In line with the secondary consideration defined in the objectives, it was decided that ranking and correlation results should be considered separately. Thus, the main decision point 2 outcome was that two versions of ModFOLDdock would be considered separately: ModFOLDdock for best correlation with observed scores and ModFOLDdockR for best agreement with observed score ranking.

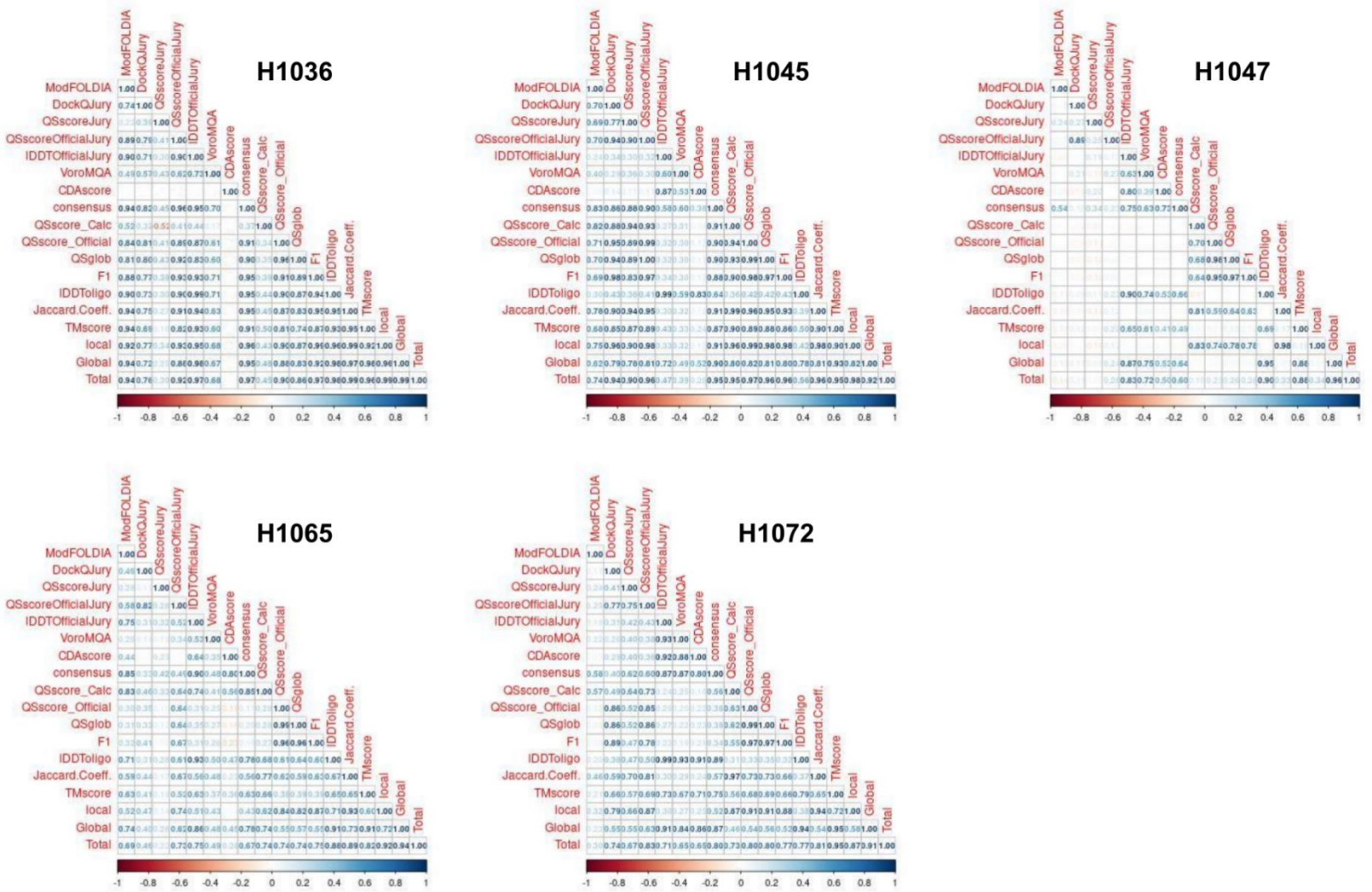


Figure 4.3. Pearson correlation matrices for CASP14 heteromer targets. From top left to right; H1036, H1045, H1047, H1065 and H1072. ModFOLDdock component scores versus single and calculated observed scores. Blank cells equate to cases where scores could not be calculated, usually due to size issues.

112

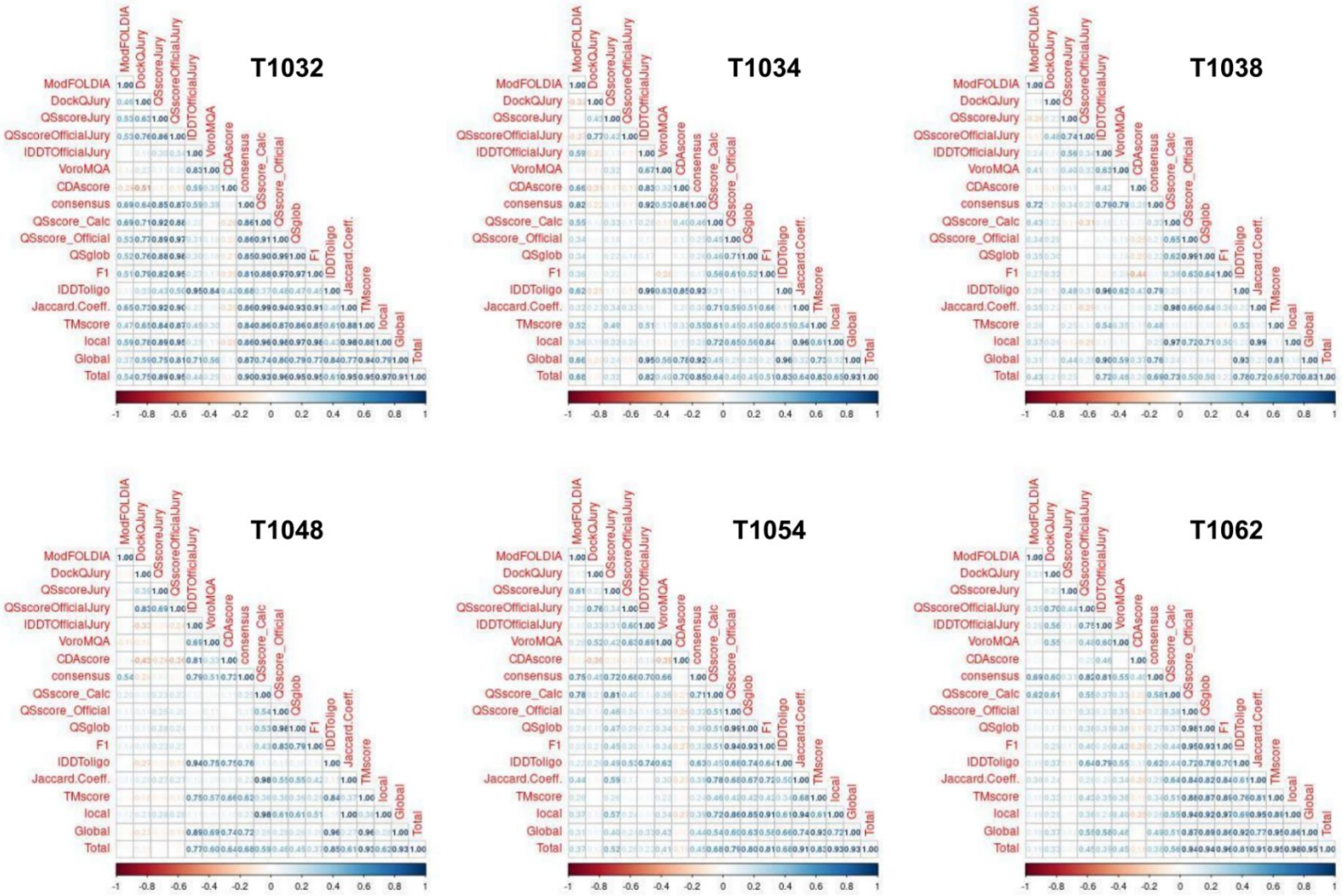


Figure 4.4A. Pearson correlation matrices for CASP14 homomer targets. From top left to right; T1032, T1034, T1038, T1048, T1054 and T1062. ModFOLDdock component scores versus single and calculated observed scores.

113

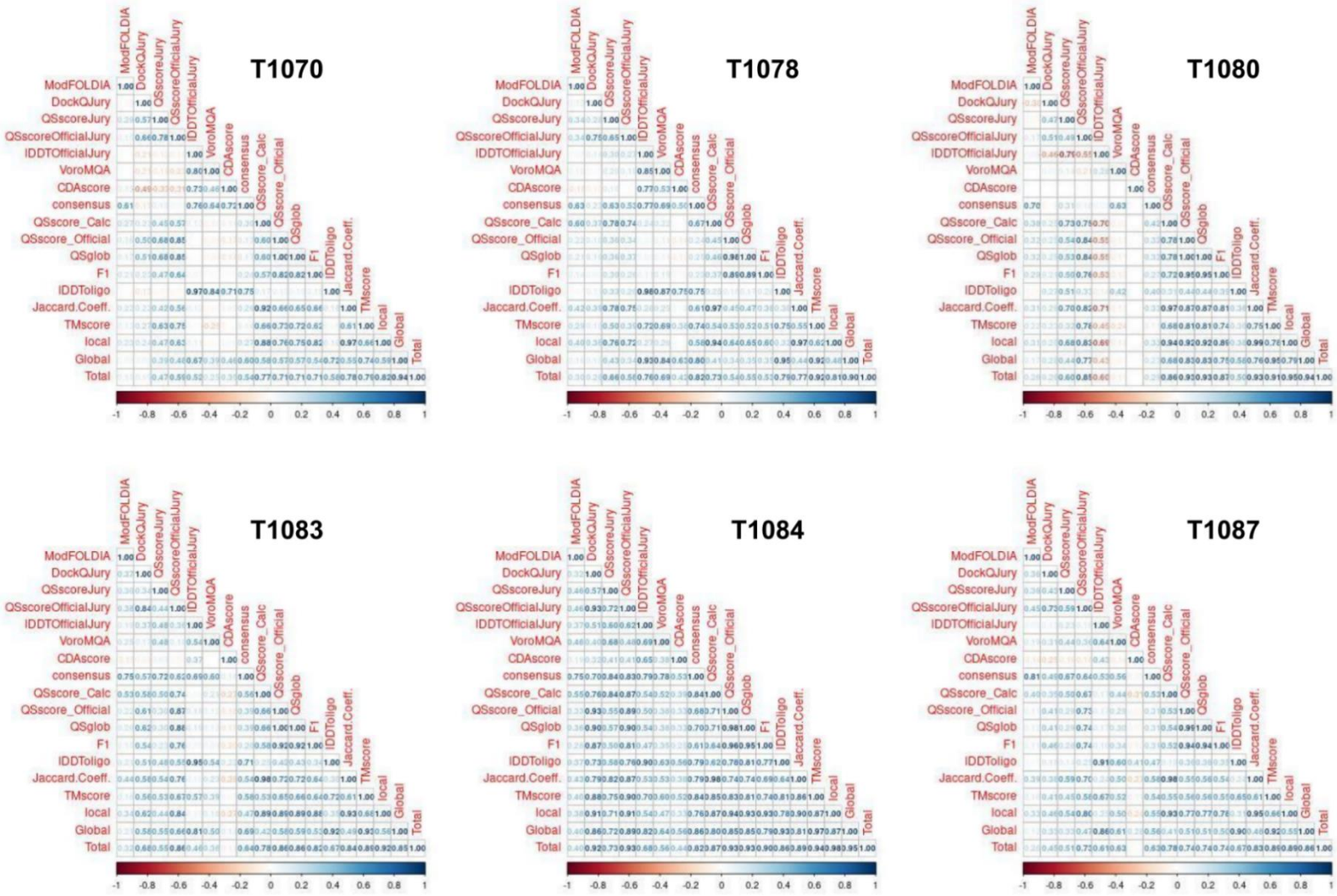


Figure 4.4B. Pearson correlation matrices for CASP14 homomer targets. From top left to right; T1070, T1078, T1080, T1083, T1084 and T1087. ModFOLDdock component scores versus single and calculated observed scores. Again, blank cells equate to uncalculated scores.

114

Ta**ble 4.1. Mean correlations for ModFOLDdock component scores (top row) versus key observed scores (left column).** Created from individual per target Pearson coefficients. Values in bold are the highest component score values achieved for observed Global and Local observed scores (also highlighted).

Score	ModFOLDIA	DockQJury	QSscoreJury	QSscoreOfficialJury	lDDTOfficialJury	VoroMQA	CDA score
QSscore_Calc	0.52375	0.4025	0.410625	0.51125	0.253125	0.1975	0.041875
QSscore_Official	0.293125	0.443125	0.366875	0.553125	0.226875	0.19938	-0.05313
QSglob	0.28875	0.44125	0.37375	0.580625	0.244375	0.21063	-0.04625
F1 (ICS)	0.275625	0.438125	0.32375	0.538125	0.21125	0.20563	-0.06938
lDDT-oligo	0.253125	0.24125	0.2825	0.376875	0.93	0.68563	0.498125
Jaccard (IPS)	0.42125	0.4225	0.43875	0.5475	0.268125	0.25375	0.01125
TM-score	0.321875	0.363125	0.411875	0.481875	0.544375	0.40813	0.223125
Local	0.38125	**0.46625**	**0.424375**	**0.585625**	0.269375	0.2525	-0.01313
Global	0.3225	0.33	0.3875	**0.47375**	**0.799375**	**0.595**	0.395
Total	0.386875	0.41875	0.446875	0.565625	0.6325	0.50313	0.2425

Table 4.2. Cumulative observed scores (top row) for models top-ranked by ModFOLDdock component scores (left column). Scores are rounded to 2 decimal places (1 for F1) for display purposes. Table is ordered by decreasing sum of all scores (obs_sum) with the top three highlighted.

Score	lAscore	DockQ	QSscore_ Calc	QSscore_ Official	lDDT_ Official	QS Glob	F1	lDDT oligo	Jaccard Coeff.	TM-score	Local	Global	Total	obs_sum
QSscoreOfficialJury	11.35	3.36	9.05	5.57	8.28	5.57	431.8	8.71	6.51	8.84	**5.41**	**8.77**	7.09	**520.32**
DockQJury	10.30	3.58	7.76	5.34	7.20	4.81	427.6	7.46	5.82	8.28	**5.05**	7.87	6.46	**507.52**
VoroMQA	10.31	2.74	7.43	4.39	8.03	4.08	381.9	8.55	5.92	8.23	**4.87**	**8.39**	6.63	**461.46**
lDDTOfficialJury	9.51	2.52	7.17	4.01	9.44	3.96	314.3	9.44	4.96	8.18	4.05	**8.81**	6.43	392.77
ModFOLDIA	11.90	1.99	8.38	3.38	7.67	3.22	274.0	7.98	5.79	7.63	4.27	7.81	6.04	350.05
QSscoreJury	5.80	1.40	4.67	2.99	7.74	2.99	222.2	7.70	4.18	7.87	3.20	7.79	5.49	284.02
CDA score	7.65	0.99	5.27	1.23	7.02	1.23	90.9	7.02	3.54	6.70	2.22	6.86	4.45	145.15

Scores are created from individual per target top-rank tables where the observed scores are collected for each of the top-ranked models for each component score, these are then summed across all targets to give rankings per observed score and an overall (summed) ranking (obs_sum).

115

Decision point 3. Tables 4.3 and 4.4 show truncated versions of the final all-against-all comparison tables described in stage 3 (again see Appendix 10 for the full version). Pearson, Spearman and Kendall correlation coefficients were calculated for the relationships, but in cases where individual coefficient scores disagreed, it was considered important to assess the data in terms of a linear relationship, taking into account proportionality of increase as well as direction and also treating outliers more strictly. It was therefore decided that the Pearson r value would be given preference over the Spearman rho or Kendall tau values when making final decisions on combinations (see Appendix 11 for the relevant coefficient formulae).

Table 4.3. Selected rows showing correlations between the observed global interface and Global fold scores and all combinations of the 7 component scores. A=ModFOLDIA, B=DockQJury, C= QSscoreJury, D=QSscoreOfficialJury, E=lDDTOfficialJury, F=voronota-js-voromqa, G=CDA-score. Top scores in each column in bold. Combinations used for ModFOLDdock scores are highlighted in green. Adapted from (Edmunds *et al.*, 2023).

Component combination	Interface (QSCORE)			Fold (SCORE)		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
B+E	0.6221383	0.4662672	0.3370294	**0.897708**	**0.8895329**	0.7178826
D+E	0.7678932	0.6149145	0.451429	0.8886437	0.8864162	**0.7204588**
B+D	**0.9005487**	0.8246907	0.6435966	0.6419381	0.5309702	0.3781203
D	0.8904282	**0.8440979**	**0.6601409**	0.6263819	0.5468863	0.389032

Table 4.4. Selected cumulative observed global interface and Global fold scores of top ranked models for every combination of the 7 component scores. (A-G are as described for table 4.3). Top scores in each column are shown in bold. ModFOLDdockR score combinations are highlighted in green. Adapted from (Edmunds *et al.*, 2023).

Component combination	Interface (QSCORE)	Fold (SCORE)
C+E+F	4.962	**9.145**
B+D+F	**5.6105**	8.479

From the ModFOLDdock global fold (SCORE) results in Table 4.3, it was clear that the combination of DockQJury (B) and lDDTOfficialJury (E) was optimal, having the highest correlation value in two out of the three correlation coefficients. The global interface (QSCORE) results were not so clear showing a disagreement between the Pearson and both the Spearman and Kendall coefficients. Nevertheless, the convention of prioritising the Pearson linear relationship was adhered to and the DockQJury (B) and QSscoreOfficialJury (D) combination was selected.

The results for maximum ranking scores in Table 4.4 were easier to interpret as there was only one top score for each category. For the final stage 3 decision, the following score combinations were selected:

ModFOLDdock: - SCORE: mean of DockQJury + lDDTOfficialJury.
QSCORE: mean of DockQJury and QSscoreOfficialJury

ModFOLDdockR: - SCORE: mean of QSscoreJury + lDDTOfficialJury + VoroMQA.
QSCORE: mean of DockQJury + QSscoreOfficialJury + VoroMQA.

As a final validation exercise, the new methods for ModFOLDdock and ModFOLDdockR were benchmarked against all component scores, with the results presented visually as bar plots in Figure 4.5. The comparative performance showed that for both the correlations in A and B and the top-rank observed totals in C and D, the combinations identified in Tables 4.3 and 4.4 outperformed all individual component scores.

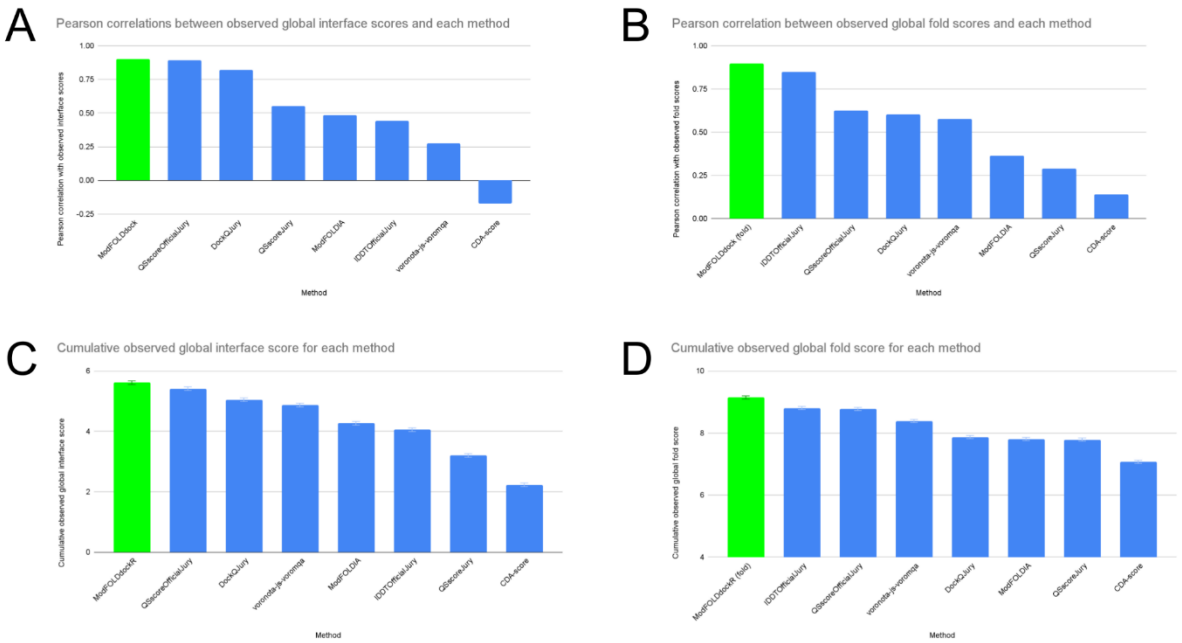


Figure 4.5. Bar plots showing benchmarking results for ModFOLDdock and ModFOLDdockR methods (in green) against all component scores (in blue). A. Pearson coefficients between calculated observed Local score and ModFOLDdock QSCORE calculated from B+D (in green) and component scores (in blue). **B.** Pearson correlations between calculated observed Global score and ModFOLDdock SCORE calculated from B+E (in green) and component scores (in blue). **C.** Cumulative observed Local score for top-ranked models identified by ModFOLDdockR calculated QSCORE (in green) and component scores (in blue). **D.** Cumulative observed Global score for top-ranked models identified by ModFOLDdockR calculated SCORE (in green) and component scores (in blue). The error bars show +/- the standard error in the observed scores of the top ranked models for each method. Reproduced from (Edmunds *et al.*, 2023).

In terms of the local residue confidence scores required for QMODE2, the most appropriate component scores to consider were ModFOLDIA, VoroMQA and CDA, each of which had been designed specifically to consider interface residues and output both global and per-residue scores as default. Unlike the two global scores (SCORE and QSCORE), a full testing programme was not undertaken for the per residue scores prior to CASP15. The main reason for this was the lack of a precedent for such scores and uncertainty over the exact method of

assessment that CASP may use to differentiate between representative and non-representative scores. Instead, the following basic logic was used to assign residue-level confidence scores. ModFOLDIA is the McGuffin group's own score designed for protein quaternary structure assessment in CASP12 with interface residue level scores, and so it was appropriate that this should form the residue-level score for the ModFOLDdock variant. Indeed ModFOLDIA was the only entry in the CASP12 Interface Accuracy (IA) assessment (https://predictioncenter.org/casp12/index.cgi?page=format), so this previous abandoned format represents the only precedent. As ModFOLDdockR was designed for ranking and the VoroMQA score had featured in both ranking scores (see Table 4.4) and considering its success in the Venclovas team model selection method mentioned in Section 4.3.1, it was decided that the ranking residue-level score should be a mean of both ModFOLDIA and VoroMQA local scores. Lastly, the other interface residue level score available was the CDA score, which was included in the ModFOLDdockS variant (see below).

The last stage of the QMODE2 optimisation stemmed from the uncertainty over the size and complexity of the models which would make up the CASP15 EMA competition. As very large models could lead to memory and CPU issues during all-against-all calculations, a quasi-single-model variant of ModFOLDdock was developed which utilised the MultiFOLD pipeline to construct 30 reference models against which comparison calculations were performed.

In summary, the ModFOLDdock global scores were optimised for positive linear Pearson correlation with observed scores, calculated using elements of the CASP14 assessors' formulae. The ModFOLDdockR global scores were optimised by rank, meaning that the predicted top-ranked model should always have the highest observed score. Finally, ModFOLDdockS used a quasi-single model approach were each model was compared to 30 reference models built using the MultiFOLD modelling pipeline. The scores contributing to the global fold score (SCORE), the overall interface accuracy score (QSCORE) and the individual residue-level confidence scores for all three variants are shown in Table 4.5 and in the organogram in Figure 4.6.

Table 4.5. **Individual ModFOLDdock component scores contributing to each CASP15 QMODE2 score for each ModFOLDdock variant.** Reproduced from (Edmunds *et al.*, 2023).

Variant	Fold	Interface	Residue
ModFOLDdock	DockQJury, lDDTOfficialJury	DockQJury, QSscoreOfficialJury	ModFOLDIA
ModFOLDdockR	QSscoreJury, lDDTOfficialJury, voronota-js-voromqa	DockQJury, QSscoreOfficialJury, voronota-js-voromqa	voronota-js-voromqa, ModFOLDIA
ModFOLDdockS	DockQJury, lDDTOfficialJury	DockQJury, QSscoreOfficialJury	CDA, voronota-js-voromqa, ModFOLDIA

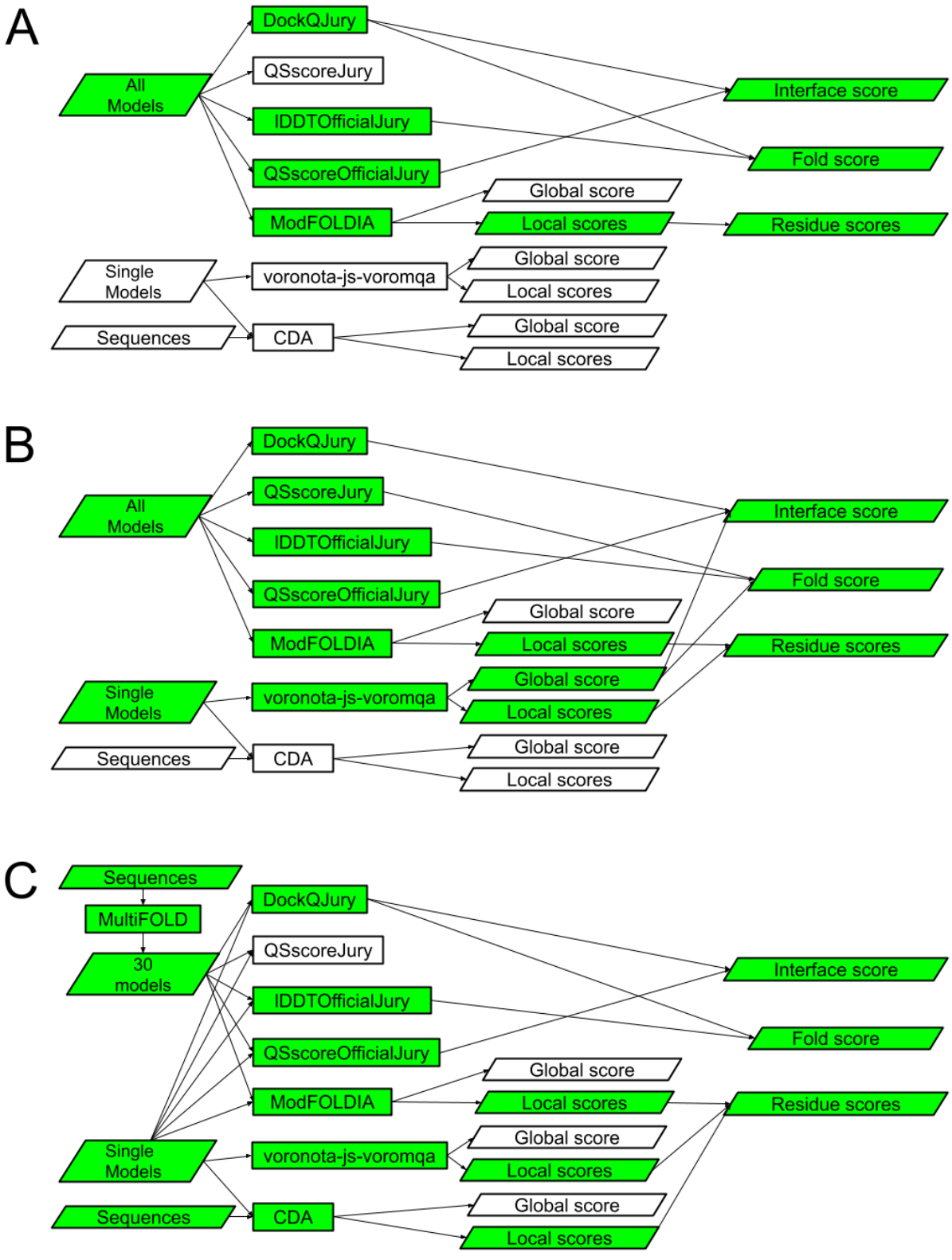


Figure 4.6. **A flowchart showing the constituent component methods and their contributions to the consensus and residue confidence scores for the three ModFOLDdock variants.**
A. ModFOLDdock, **B.** ModFOLDdockR, **C.** ModFOLDdockS. Green coloured boxes indicate the scores that contribute directly to the overall global fold (SCORE), overall interface (QSCORE) and individual residue confidence scores. Reproduced from (Edmunds *et al.*, 2023).

4.4.2 Part 2. CASP15 official rankings and results

4.4.2.1 ModFOLDdock achieved peak performance across EMA categories

All three ModFOLDdock variants were successful in submitting predictions across *all* CASP targets for *all* three scores in the QMODE2 category. This is shown by the three bar charts in Figure 4.7 which display a 100% prediction rate for ModFOLDdock variants alongside the rates achieved by other EMA software. Only those meeting the 80% threshold were considered successful EMA predictors and included in further CASP analysis. In the local residue confidence score category (right-hand plot in Figure 4.7), ModFOLDdock variants were notable as the only methods to continue to achieve a 100% prediction rate, showing a reliability and consistency across the full range of targets, not achieved by any other method.

The bar plots in Figure 4.8 display the official CASP15 rankings of EMA software meeting the 80% threshold using official assessor quality measures. These plots show that assessors placed at least one of the ModFOLDdock variants first or second within each of the three EMA categories (disregarding the placing of the CASP assembly consensus (AC) method). For ease of interpretation, the ranks achieved by all ModFOLDdock variants for each QMODE2 score shown in Figure 4.8, are summarised in Table 4.6.

For reference, the assembly consensus benchmark score (AC) is an all-against-all predicted accuracy score (S) calculated for each residue (*i)* in each model (x). It is the average per-residue score (f) calculated using all models (*y*) in the target population (N) as reference. For SCORE, f is the oligo-GDT TS score and for QSCORE, f is the QS-score (calculated using the QS-align tool25).

$$S(x) = \frac{1}{(N-1)} \sum_{y!=x} f(xi, y)$$

Table 4.6. **A summary of ModFOLDdock variant rankings in CASP15 QMODE2 EMA categories.** Numbers with an asterix signify rankings with the assembly consensus (AC) disregarded (to convert these to the actual ranks shown in Figure 4.8 add 1 to the score shown in the table).

Variant	Rank /23 (SCORE)	Rank /18 (QSCORE)	Rank /13 (residue)
ModFOLDdock	2	2*	6
ModFOLDdockR	4*	1*	2
ModFOLDdockS	12*	5*	3

This independently verified performance (SCORE rank 2, QSCORE rank 1 and Local residue rank 2) showed that the ModFOLDdock methods were among the top few EMA programs at CASP15 (arguably the best overall if ranks are averaged over the three categories, which

would be legitimate as each variant is available to users from the ModFOLDdock server webpage). On this basis the McGuffin group was invited to present the ModFOLDdock method at the CASP15 conference and also to publish the work described in this chapter in the Proteins 2023 special edition, as listed on the chapter title page.

4.4.2.2 ModFOLDdock local per-residue scores showed unique qualities

The local residue scores in the right-hand bar plot in Figure 4.8 were calculated using a combination of per-residue lDDT, CAD, PatchQS and PatchDockQ scores (definitions of the patch scores and the local residue Z-score calculations can be found in Appendix 12). Of the four scores, lDDT and CAD were used to assess accuracy in terms of relative neighbourhood atom positions, while PatchQS and PatchDockQ were primarily used to assess inter-chain positioning (Studer, Tauriello and Schwede, 2023), meaning that these latter two scores were important in correctly identifying native-like patches of interface residues. Figure 4.8 shows that GuijinLab-RocketX out-performed the ModFOLDdockR and S variants (second and third places respectively) according to the calculated summed per-residue score with ModFOLDdock ranked in only sixth place. However, a closer look at the contributing scores shown in the Figure 4.8 plot shows that almost all of the ModFOLDdock score is composed of the two patch scores suggesting a particular sensitivity to interface patch identification. Indeed, when the emphasis of the analysis was changed to focus on the recognition of native interface residues by averaged ROC AUC scores, the ModFOLDdock variant moved from sixth to first place. The results of this aspect of the CASP analysis are presented in Figure 4.9 showing the recalculated ranks with ModFOLDdock at the top.

Further to this, a final piece of CASP analysis focussed specifically on the antibody-antigen binding interactions described by heteromers H1166, H1167 and H1168. Results for this analysis are shown in Figure 4.10 and, again, showed that ModFOLDdock variants performed well in the overall ranking derived from all four lDDT, CAD, PatchQS and PatchDockQ scores (shown in panel A), where they were once again second only to GuijinLab-RocketX. Again, and in line with the reranking described above, when the ROC AUC scores were considered in isolation, all ModFOLDdock variants were shown to out-perform all other methods (panel B).

The ModFOLDdock methods, therefore, seem particularly well suited to the task of identifying patches of native-like interface residues and it appears that this ability becomes enhanced when applied to antibody-antigen interactions. This could be a unique property of the ModFOLDdock method.

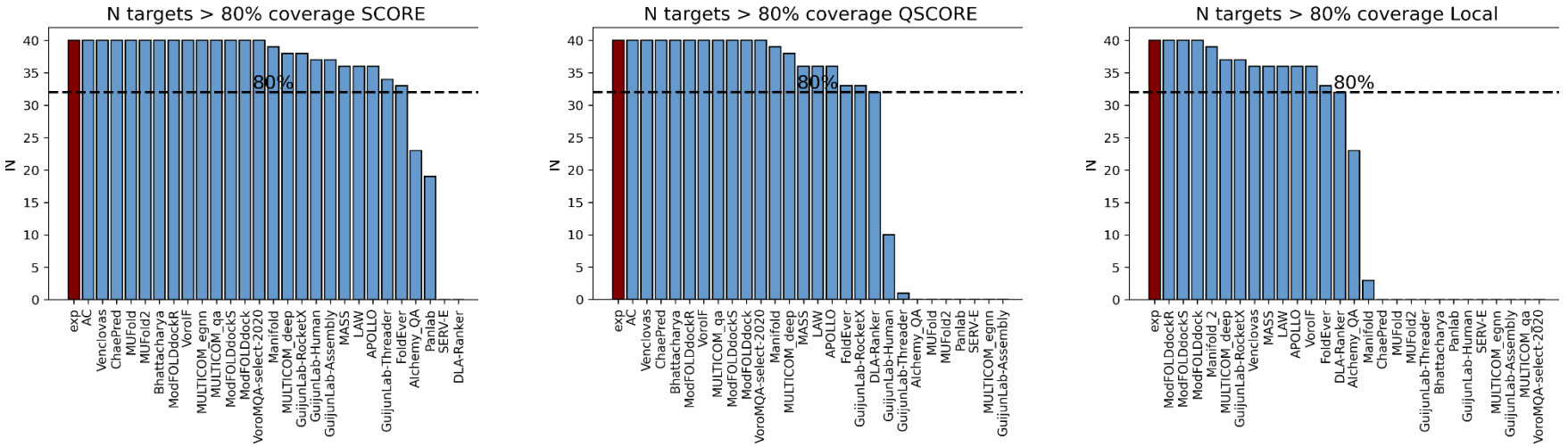


Figure 4.7. CASP15 EMA software meeting the 80% threshold. Left. For global fold SCORE. Middle. For global interface QSCORE. Right. For local residue confidence scores. AC is the assembly consensus baseline (described in Section 4.4.2.1). Reproduced from (Studer *et al.*, 2023).

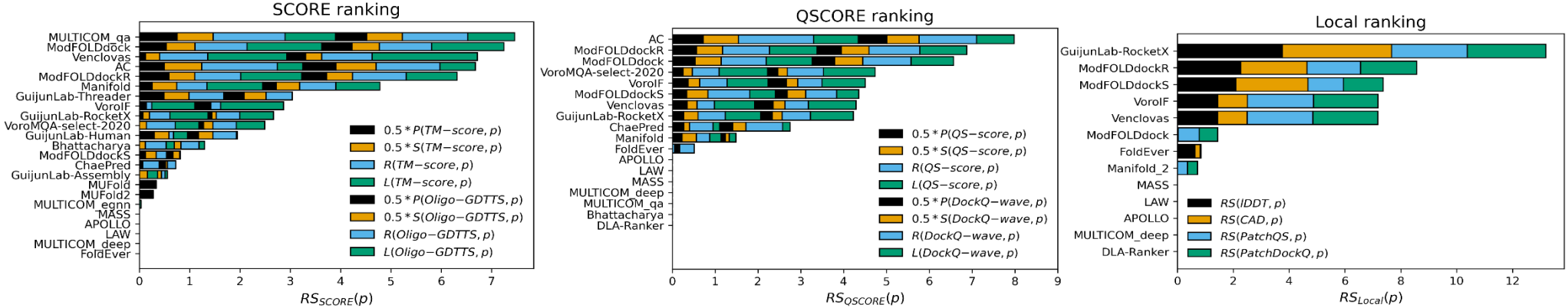


Figure 4.8. CASP15 EMA rankings. Left. Overall ranking by Z-score for global fold SCORE where ranking score (RSscore) = RS(TM-score) + RS(Oligo-GDTTS). Middle. Similar Z-score rankings for global interface QSCORE, RSqscore=RS(QS-score)+RS(DockQ-wave). Right. Local interface accuracy based on Z-scores where RSLocal=RS(lDDT)+RS(CAD)+RS(PatchQS)+RS(PatchDockQ). AC is the assembly consensus baseline. For SCORE and QSCORE, P= Pearson r, S=Spearman rho, R=ROC AUC and L=Loss. DockQ-wave is the DockQ weighted average used to score higher-order complexes. Reproduced from (Studer *et al.*, 2023).

122

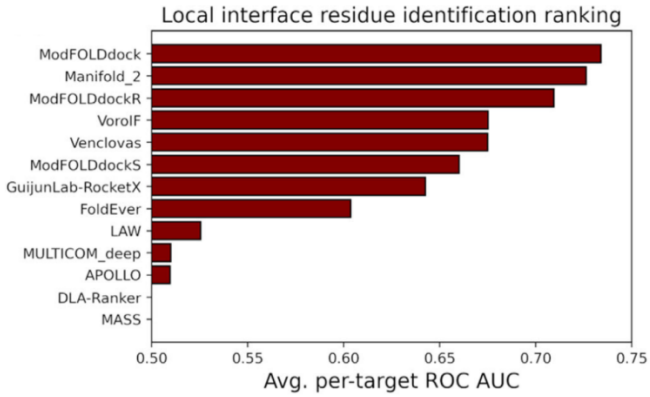


Figure 4.9. CASP15 EMA local interface residue identification ranking calculated by averaged ROC AUC scores. Showing identification of model interface residues matching those in the native structure. Reproduced from (Studer *et al.*, 2023).

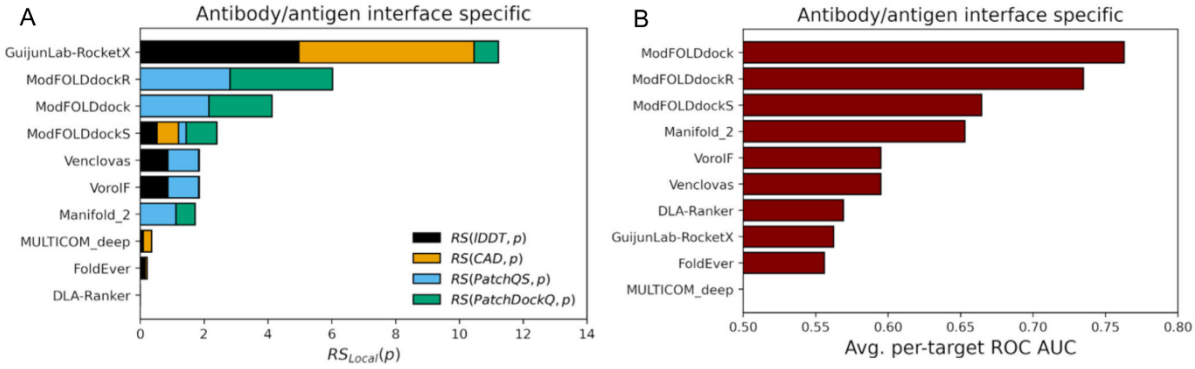


Figure 4.10. CASP15 EMA antibody/antigen local score evaluation. A. A similar analysis to Figure 4.8 (right hand graph for local) but for the antibody-antigen targets H1166, H1167 and H1168 only. **B.** Identification of interface residues similar to Figure 4.9 but, again, only for the three antibody/antigen targets. Reproduced from (Studer *et al.*, 2023).

4.4.2.3 Multimer modelling analysis.

A brief analysis of modelling performance is included here as a comparison with the analysis described in Chapter 2 (and briefly in Chapter 3) for CASP13 and 14 modelling. It is also pertinent to ModFOLDdock performance due to the inclusion of the method within the CASP15 modelling pipeline as described in Figure 2.15 (Chapter 2). Table 4.7 shows selected CASP15 modelling group rankings by sum Z-score (which continues to be calculated as Z-score(ICS) + Z-score(IPS) + Z-score(lDDT-oligo) + Z-score(TM-score)).

As can be seen from Table 4.7, both the McGuffin (manual) and MultiFOLD (server) groups (both of which used the MultiFOLD/ModFOLDdock pipeline) were placed above the naïve NBIS-AF2-Multimer group, which acted as the AlphaFold2-Multimer modelling baseline, as well as the ColabFold group (which used the same base software), in all categories with the exception of TBM/FM for MultiFOLD. This is reflected in the CASP15 official assembly results (Burcu Ozden *et al.*, 2023) and supports the two hypotheses from Chapter 2 that the MultiFOLD pipeline, in general, added value to the baseline modelling capabilities of AlphaFold2-Multimer.

Table 4.7. CASP15 assembly group rankings (Sum Z-score >0.0, for rank1 models) by category. Groups selected are unmodified AFM/ColabFold users or those with the highest prediction accuracy (Yang) or top TBM method (PEZY). Model total is given in column headings, Z-scores in brackets. AF2 baseline (NBIS-AF2-Multimer) is shaded. Data is for multimers, sorted by overall rank and taken from the CASP results page (https://predictioncenter.org/casp15/).

Category / Group	TBM Rank /82	TBM/FM Rank /85	FM Rank /68	Overall Rank /87
Yang (439)	7 (7.35)	5 (14.25)	11 (2.56)	5 (24.17)
McGuffin (manual)	15 (6.01)	10 (11.01)	8 (2.86)	9 (19.89)
PEZY Foldings (278)	4 (7.73)	27 (7.96)	16 (2.24)	13 (17.94)
MultiFOLD (server)	12 (6.22)	37 (5.71)	3 (3.29)	23 (15.23)
ColabFold (446)	30 (4.73)	33 (6.15)	18 (1.87)	29 (12.79)
NBIS-AF2-Multimer	20 (5.37)	32 (6.24)	38 (0.64)	30 (12.27)
Maximum Z-score	11.63	21.28	4.93	35.29

Across the modelling categories, it can be seen that both the McGuffin and MultiFOLD groups fared roughly equally for TBM models (Z-scores of 6.01 and 6.22 respectively), whereas human processing appeared to have a large positive effect on TBM/FM models (Z-score of 11.01 compared to 5.71 for the server models). However, this effect was reversed for FM models where the MultiFOLD server was more accurate (Z-score of 3.29 compared to 2.86 for the McGuffin group). As the base models would have been very similar, this suggests that the objective model selection process carried out by the server version was superior to human interpretation for FM models.

Results for the same categories were also available for the groups' best-scoring models rather than models designated model 1. The data for groups' best models is shown in Table 4.8.

Table 4.8. Selected CASP15 assembly group rankings (Sum Z-score >0.0, for models rated best) by category. Equivalent to the data shown in Table 4.7 but for groups' best-rated models. Data is, again, for multimers, sorted by overall rank and taken from the CASP results page (https://predictioncenter.org/casp15/).

Category / Group	TBM Rank /82	TBM/FM Rank /85	FM Rank /68	Overall Rank /87
PEZY Foldings (278)	1 (17.20)	17 (10.97)	20 (2.63)	4 (30.07)
Yang (439)	6 (8.90)	5 (16.92)	14 (3.88)	6 (28.60)
McGuffin (manual)	18 (6.25)	10 (12.79)	15 (3.69)	13 (22.73)
ColabFold (446)	17 (6.57)	20 (10.60)	23 (2.44)	19 (18.38)
MultiFOLD (server)	21 (6.00)	37 (7.33)	10 (4.11)	26 (17.42)
NBIS-AF2-Multimer	24 (5.94)	33 (8.64)	48 (1.04)	30 (14.89)
Maximum Z-score	17.20	28.41	6.49	41.80

An interesting trend was noticed on comparison of the data across the two tables. Both the McGuffin and MultiFOLD groups were ranked higher for rank 1 models than for their best models, except in the TBM/FM category where there was no difference in rank. This is best illustrated by the MultiFOLD comparative ranks; **12**/21 (TBM), **37**/37 (TBM/FM), **3**/10 (FM) and **23**/26 (overall) with rank 1 model ranks shown in bold. This effect could be observed for the

NBIS-AF2-Multimer group and Yang group (included as the group having the highest prediction accuracy (Studer *et al.*, 2023) and also as a AF2-Multimer user) but arguably not as strongly as for MultiFOLD. Notably, ColabFold, which was rated highest for self-evaluation metrics in the same CASP publication by Studer *et al*., did not show this effect.

Group 278 (PEZY Foldings) was included in the tables as it was the top modelling group in the TBM category when the best model is selected. This group exemplified the expected trend in rank performance across the two tables, that at least some ranks would improve when a group's best models are considered. This is because the chances of detecting a good model increases with a widened model population, i.e. groups will not always select the best model as their rank 1 model. Rather than being penalised by limiting assessment to rank 1 models, the McGuffin and particularly the MultiFOLD group benefitted from this. It follows that the rate of identification of the best model as the rank 1 model must have been better than average for McGuffin and MultiFOLD groups. It is possible that this is an effect of the AlphaFold2-Multimer ranking methods (plDDT and pTM) due to data from the NBIS-AF2-Multimer group but, if this were the cause, is it curious that the effect was not also seen for the ColabFold group. It is possible, therefore, that one strength of the MultiFOLD modelling pipeline was the selection of the best model by ModFOLDdock variants. If correct, this would represent significant progress in addressing the issues described in previous chapters surrounding CASP13 and 14 model selection.

4.4.2.4 Comparative analysis across CASP competitions.

This section attempts to mirror the analysis caried out for CASP13 and 14 modelling in Chapter 3 by listing ModFOLDdockR predicted scores (this variant was used as it was the primary ranking tool in the MultiFOLD pipeline) alongside CASP assessor scores and an observed ModFOLDdock score as an additional a measure of predicted score accuracy. The column titled "*Difference between rank 1 and this model"* was calculated as an absolute difference between the calculated observed score for the "best" model and the predicted rank 1 model; this demonstrates the high performance of model ranking and selection.

Table 4.9. **A summary of group 462 (MultiFOLD) CASP15 multimer models rated as "best models" in the CASP results tables.** CASP Global and Local scores have been artificially calculated to give comparisons with ModFOLDdockR predicted scores. The difference is calculated as an absolute difference between Total score for the model scored "best" by CASP and the rank 1 model from the MultiFOLD pipeline (a score of 0.0 denotes the best model was selected as rank 1).

Target	Stoichiometry	MFDR Predicted		CASP calculated		Difference between rank 1 and this model
		Global	Local	Global	Local	
H1106	A1B1	0.7073	0.5700	0.831	0.663	0.005
H1111	A9B9C9	0.4857	0.4191	0.077	0.0395	0.0
H1114	A4B8C8	0.6020	0.4664	0.2285	0.178	NA
H1129	A1B1	0.6008	0.1978	0.65	0.0315	0.0

H1134	A1B1	0.6893	0.5769	0.9295	0.797	0.286
H1135	A9B3	0.7166	0.3899	0.59	0.347	0.0
H1137	A1-I1	0.6638	0.5112	0.6475	0.735	0.0
H1140	A1B1	0.6438	0.2373	0.681	0.2485	0.0
H1141	A1B1	0.6996	0.4355	0.7395	0.1655	0.054
H1142	A1B1	0.6815	0.2876	0.6985	0.07	0.005
H1143	A1B1	0.7691	0.5741	0.9045	0.776	0.008
H1144	A1B1	0.7057	0.4187	0.7545	0.3255	0.028
H1151	A1B1	0.7668	0.6466	0.8985	0.8065	0.0003
H1157	A1B1	0.7929	0.6413	0.756	0.7	0.0004
H1166	A1B1C1	0.8066	0.5322	0.7795	0.5845	0.0
H1167	A1B1C1	0.8106	0.5527	0.772	0.6185	0.0
H1168	A1B1C1	0.8317	0.7195	0.901	0.828	0.0037
H1171	A6B1	0.4235	0.2972	0.66	0.6085	0.0001
H1172	A6B2	0.4337	0.2847	0.8395	0.5345	0.006
H1185	A1B1C1D1	0.8186	0.7384	0.8825	0.664	NA
T1109	A2	0.8502	0.7849	0.8705	0.4765	0.0
T1110	A2	0.8653	0.8295	0.955	0.931	0.004
T1113	A2	0.8133	0.7309	0.9105	0.886	0.005
T1115	A16	0.4882	0.2907	0.064	0.0385	0.0
T1121	A2	0.8183	0.6560	0.6565	0.3425	0.006
T1123	A2	0.3052	0.1794	0.203	0.02	0.002
T1124	A2	0.8345	0.7252	0.9155	0.865	0.0
T1127	A2	0.8506	0.8335	0.932	0.8865	0.001
T1132	A6	0.7771	0.4746	0.9645	0.879	0.0
T1153	A2	0.7833	0.6394	0.8945	0.748	0.019
T1160	A2	0.8341	0.7794	0.3945	0.336	0.0
T1161	A2	0.7572	0.6363	0.5615	0.563	0.01
T1170	A6	0.6788	0.2878	0.8665	0.6335	0.002
T1173	A3	0.6393	0.3666	0.4065	0.4265	0.017
T1174	A3	0.7355	0.4521	0.6725	0.724	0.0
T1176	A8	0.5245	0.2210	0.298	0.015	0.01
T1178	A2	0.1749	0	0.4325	0	0.0007
T1179	A2	0.3732	0.1820	0.188	0.0595	0.009
T1181	A3	0.8269	0.5660	0.805	0.5125	0.03
T1187	A2	0.6035	0.2767	0.8985	0.855	0.0
T1192	A10	0.6966	0.3120	0.7655	0.553	NA

Chapter 3 scores for CASP13 modelling (Table 3.1) showed that the best model was selected as the rank 1 only once (1/30 or 3.3%). This number has increased to 14/42 (or 33.3%) for CASP15 modelling (shown as a difference of 0.0) – a tenfold increase. Further to this, the difference in observed scores between best and rank 1 models reduced from an average of 0.18 and a maximum of 0.546 for CASP13 data to an average only 0.013 and a maximum of 0.286 for CASP15 data. However, it must be pointed out that a maximum difference of this magnitude was only seen for one target with the next highest value of 0.054, an order of magnitude lower. The scatter plots in Figure 4.11 include Pearson correlation coefficients between ModFOLDdockR predicted and calculated CASP Local, Global, and Total scores for the models in Table 4.9. For comparison with Chapter 3, Figure 3.1, TM-score, QS-score and IDDT-oligo correlations are also included in Figure 4.12. The plots in both figures, although

comparative rather than exact duplicates, show an increase in accuracy in CASP15 compared to CASP13 modelling. Specifically, the correlation coefficients between ModFOLDdock predicted scores and CASP assessor scores, shown in Figure 4.12, have increased from -0.07, 0.1 and 0.16 seen in Figure 3.1, for GDT TS, lDDT-oligo and QS-score respectively, to 0.63, 0.64 and 0.61 for the equivalent TM-score, lDDT-oligo and QS-score respectively.

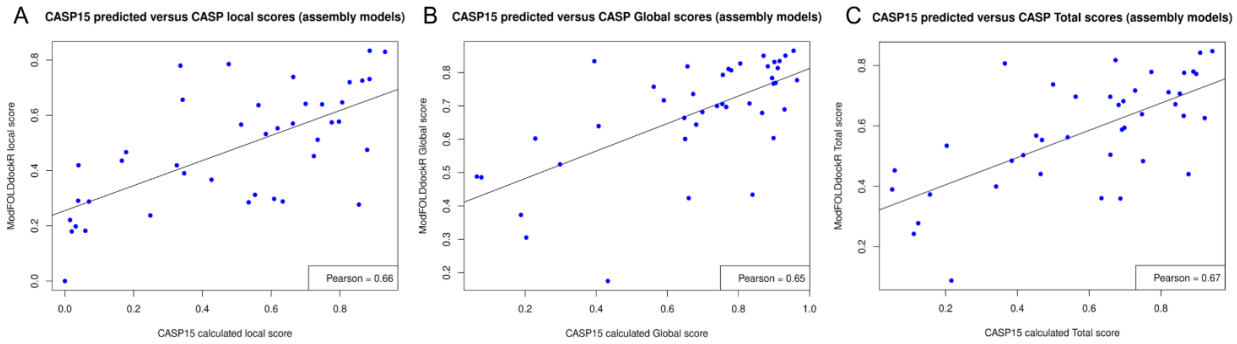


Figure 4.11. Pearson correlations for ModFOLDdockR predicted scores and equivalents calculated from CASP15 scores for group 462 (MultiFOLD) multimer models. A. ModFOLDdockR calculated Local score versus a Local score calculated from CASP15 ICS and IPS scores. **B.** ModFOLDdockR calculated Global score versus a Global score calculated from CASP15 TM-score and lDDT-oligo score. **C.** ModFOLDdockR calculated Total score versus an equivalent score calculated from all 4 CASP15 scores.

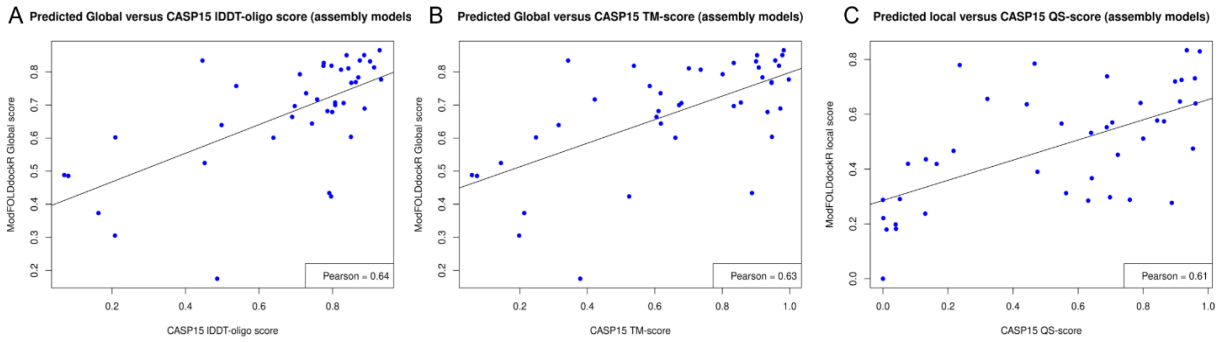


Figure 4.12. Pearson correlations between ModFOLDdockR predicted scores and individual CASP15 scores for group 462 (MultiFOLD) multimer models. A. ModFOLDdockR calculated Global score versus CASP15 lDDT-oligo score. **B.** ModFOLDdockR calculated Global score versus CASP15 TM-score. **C.** ModFOLDdockR calculated Local score versus CASP15 QS-score.

Figure 4.13A shows similar plots for the ModFOLDdockR variant but extending the data to include models from all CASP15 groups. In these plots homomer targets T1160 and T1161 and heteromer targets H1171 and H1172 have been excluded. These form two pairings of the five alternative ensemble structures (T1109-T1110, T1158 series, T1160-T1161, H1171-T1172 and T1195-T1197) which were added to the CASP15 experiment as specific modelling challenges. T1160 and T1161 represent two different conformations of very similar sequences resulting from the effect of five mutations and different crystallisation conditions, whereas H1171 and H1172 are two alternative functional conformations of the Holiday junction complex (Kryshtafovych *et al.*, 2023). When included, these targets produced clear outliers affecting the

correlations, although the T1109-T1110 pair appeared to be well scored (native structures for T1195-T1197 were not available at the time of analysis).

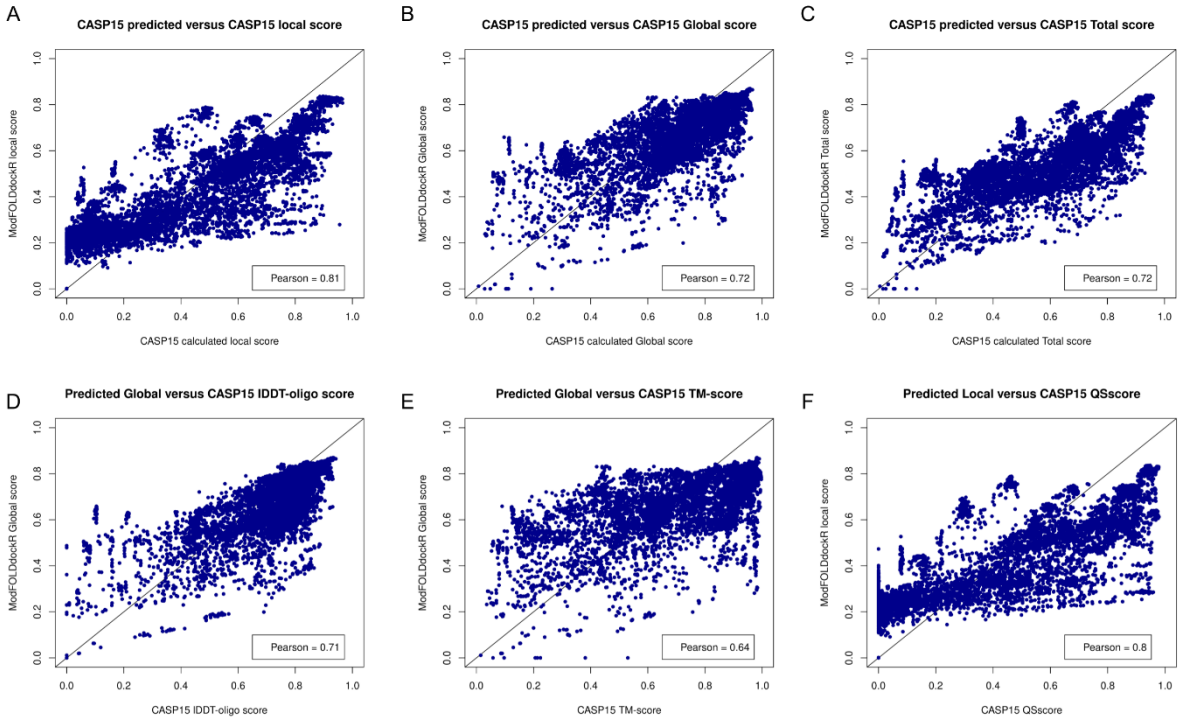


Figure 4.13A. Scatter plots with Pearson correlations for ModFOLDdockR predicted scores and equivalents calculated from CASP15 scores for all group models. A. ModFOLDdockR calculated Local score versus a Local score calculated from CASP15 ICS and IPS scores. **B.** ModFOLDdockR calculated Global score versus a Global score calculated from CASP15 TM-score and lDDT-oligo score. **C.** ModFOLDdockR calculated Total score versus an equivalent score calculated from all 4 CASP15 scores. **D.** ModFOLDdockR calculated Global score versus CASP15 lDDT-oligo score. **E.** ModFOLDdockR calculated Global score versus CASP15 TM-score. **F.** ModFOLDdockR calculated Local score versus CASP15 QS-score.

However, upon removal, strong Pearson correlation coefficients of 0.81 and 0.72 with calculated Local and Global observed scores respectively were revealed, increased from 0.66 and 0.65 obtained with the group 462 (MultiFOLD) models alone. Slightly better coefficients of 0.81 and 0.76 for the same scores are also shown for the ModFOLDdock variant in Figure 4.13B, with increases likely due to this variant's development for correlation with observed scores.

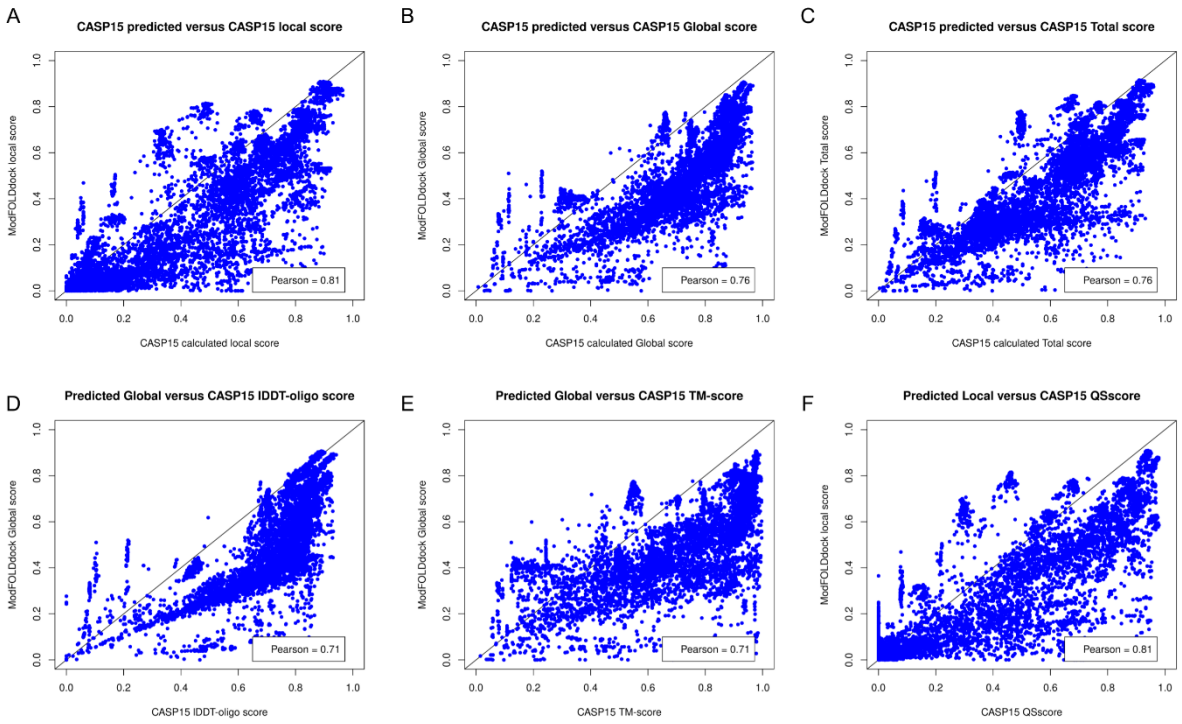


Figure 4.13B. Scatter plots with Pearson correlations for ModFOLDdock predicted scores and equivalents calculated from CASP15 scores (all groups' models). A. ModFOLDdock calculated Local score versus a Local score calculated from CASP15 ICS and IPS scores. **B.** ModFOLDdock calculated Global score versus a Global score calculated from CASP15 TM-score and lDDT-oligo score. **C.** ModFOLDdock calculated Total score versus an equivalent score calculated from all 4 CASP15 scores. **D.** ModFOLDdock calculated Global score versus CASP15 lDDT-oligo score. **E.** ModFOLDdock calculated Global score versus CASP15 TM-score. **F.** ModFOLDdock calculated Local score versus CASP15 QS-score.

ModFOLDdockS scores are added for comparison in Figure 4.13C and show slightly weaker correlations, likely due to the modelling challenges of some larger targets which would impact on the quality and variety of decoy structures used for the clustering algorithms.

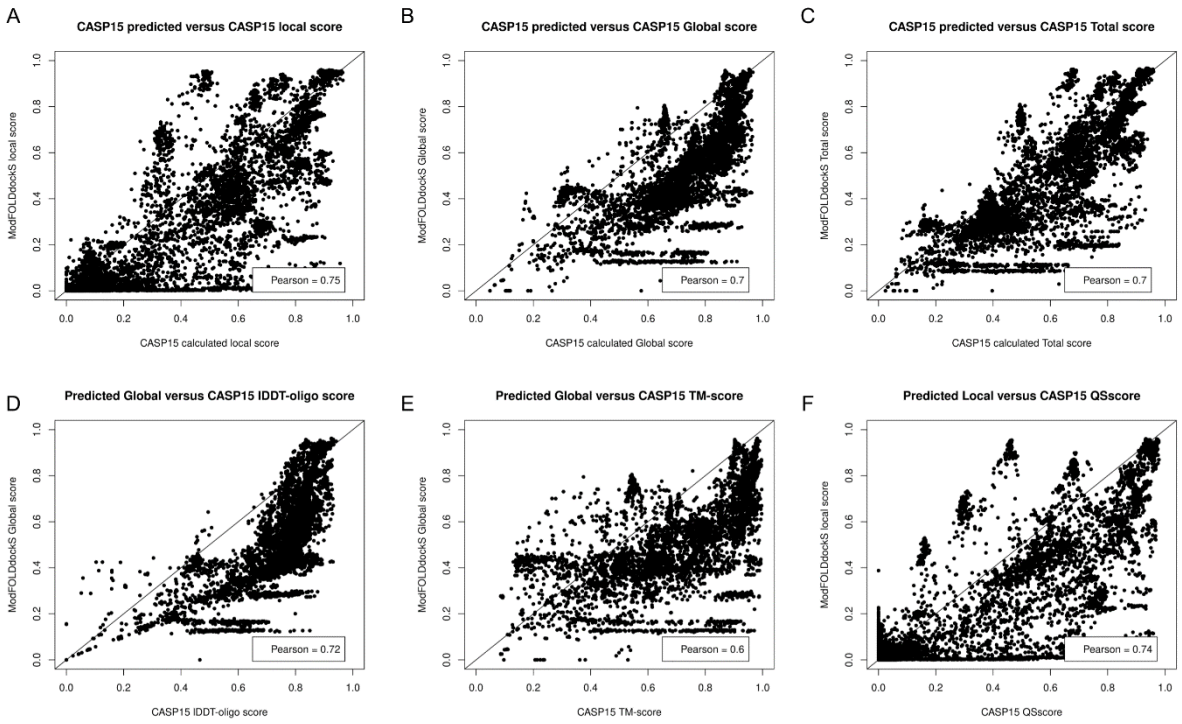


Figure 4.13C. Scatter plots with Pearson correlations for ModFOLDdockS predicted scores and equivalents calculated from CASP15 scores (all groups' models). A. ModFOLDdockS calculated Local score versus a Local score calculated from CASP15 ICS and IPS scores. **B.** ModFOLDdockS calculated Global score versus a Global score calculated from CASP15 TM-score and lDDT-oligo score. **C.** ModFOLDdockS calculated Total score versus an equivalent score calculated from all 4 CASP15 scores. **D.** ModFOLDdockS calculated Global score versus CASP15 lDDT-oligo score. **E.** ModFOLDdockS calculated Global score versus CASP15 TM-score. **F.** ModFOLDdockS calculated Local score versus CASP15 QS-score.

Finally Figure 4.14 shows similar data to that in Figures 4.13A and B but differentiated into separate homomer and heteromer plots. This shows that all ModFOLDdock variants maintain comparative performance across protein targets with differing stoichiometry and symmetry. Targets T1160, T1161, H1171 and H1172 were again omitted as explained previously.

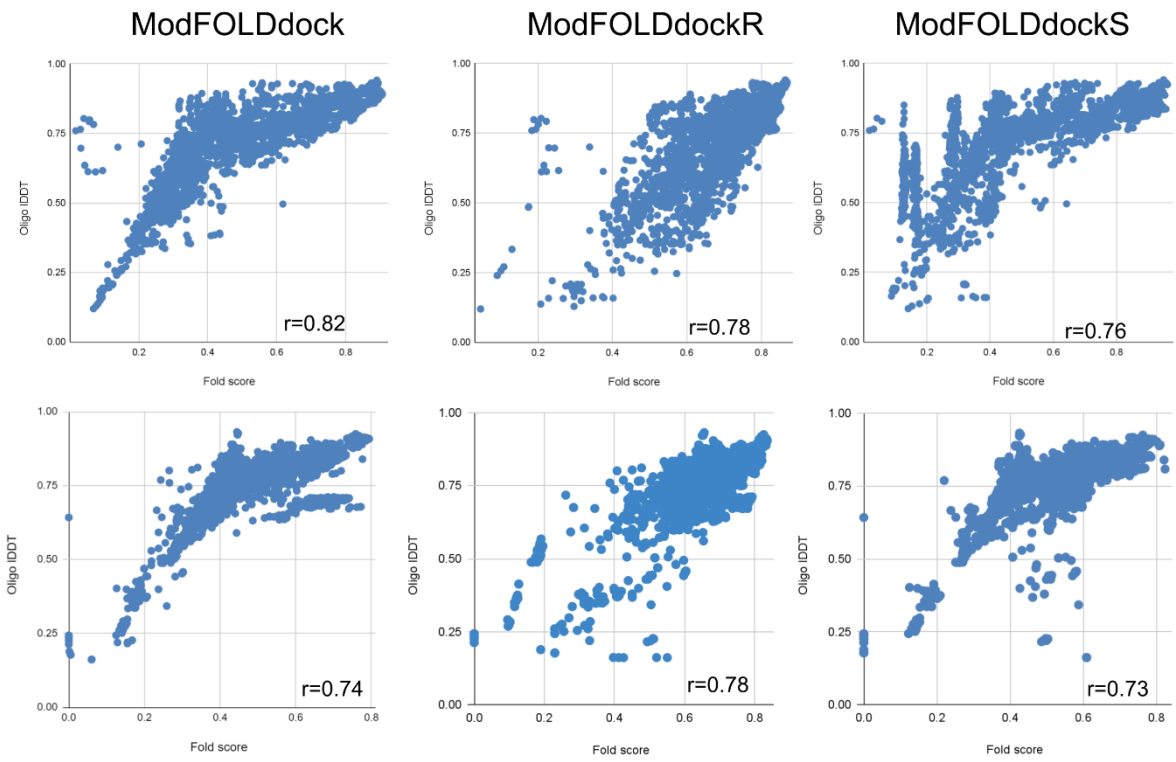


Figure 4.14. Scatter plots with Pearson R value between predicted Global (fold) score and observed lDDT-oligo for all ModFOLDdock variants for CASP15 models from all groups. Top. Plots for homomeric targets for ModFOLDdock (left), ModFOLDdockR (middle) and ModFOLDdockS (right). **Bottom.** The plots between the same variables in the same left to right variant order for all heteromeric targets. Image taken from (Edmunds *et al.*, 2023).

4.5 Conclusions

This chapter described the QMODE2 optimisation process of the hybrid consensus MQA programs ModFOLDdock, ModFOLDdockR and ModFOLDdockS. For ModFOLDdock, component quality scores were combined to achieve optimal correlations with observed target scores. For ModFOLDdockR, the quality scores were combined for optimal ranking, meaning that the models with the highest observed scores were ranked top. ModFOLDdockS was additionally developed to address the limitations of clustering-based systems by employing a quasi-single model approach using MultiFOLD reference models. In all cases the target observed scores used for optimisation were those identified in Chapter 3. The data in this chapter resulted from blind independent benchmarking of the ModFOLDdock MQA method and showed that its performance was competitive with the best methods available in 2022.

The official results from the CASP15 competition suggest that the three ModFOLDdock variants could reasonably be described as having performed better than any other single method in the new multimer EMA category. Specifically, being alone in achieving a 100% prediction rate across all three EMA categories as well as achieving best rankings of 2nd place in the SCORE category (ModFOLDdock), 1st place in the QSCORE category

(ModFOLDdockR) and 2nd place in the local interface residue category (ModFOLDdockR). Additionally, all three ModFOLDdock variants showed superior interface patch identification abilities as measured by the PatchQS and PatchDockQ scores, which was stronger still for antibody-antigen binding interactions. Later analysis also showed an increase in correlations between ModFOLDdock variants' predicted scores and CASP observed scores from a maximum of 0.16 seen in CASP13 to a maximum of 0.64 when using MultiFOLD group data with a further increase to a maximum of 0.81 when all CASP15 data was considered. This represents at least a 4-fold increase in accuracy as measured by Pearson correlation and this was maintained across homo and heteromer model populations.

In terms of multimer modelling, MultiFOLD out-performed both the NBIS-AF2-Multimer and the ColabFold groups which represent the baseline modelling performance using the AF2-Multimer and ColabFold software respectively. This success appeared, at least in part, to be due to the ability of ModFOLDdockR to rank the best model at the top of a decoy population resulting in MultiFOLD's competition ranking being higher for rank 1 models than for the CASP-selected best models. Later analysis showed that the observed best model was correctly identified as the rank 1 model in 33% of cases, a 10-fold increase in the same metric seen in CASP13 and that the average difference or loss between predicted and observed scores reduced from 0.18 seen at CASP13 to 0.013, as measured by an average of global fold and global interface score.

CHAPTER 5

Benchmarking of AlphaFold2 accuracy self-estimates as empirical quality measures and model ranking indicators and their comparison with independent model quality assessment programs.

Work presented in this chapter is currently available in bioRXiv preprint format:

Benchmarking of AlphaFold2 accuracy self-estimates as empirical quality measures and model ranking indicators and their comparison with independent model quality assessment programs.

Nicholas S. Edmunds, Ahmet G. Genc, Liam J. McGuffin

bioRxiv 2023.12.15.571846

The same work is currently accepted for publication in the Oxford University Press (OUP) Bioinformatics journal, subject to successful review.

5.1 Background

Since the success of AlphaFold2 (Jumper *et al.*, 2021) at CASP14 in 2020 many articles have detailed the methodology by which AF2 achieved its level of accuracy, most notably by the DeepMind group itself (Evans *et al.*, 2022) as well a group led by Jeffrey Skolnick (Skolnick *et al.*, 2021) and the group who pioneered the development of the ColabFold adaptation of the software (Mirdita *et al.*, 2022). It is usual for protein modelling software to provide accuracy self-estimate scores to accompany their models (Varadi *et al.*, 2022) and while competition modellers are concerned with correlation agreements and statistical measures of significance across large datasets, the accuracy and usefulness of a single predicted score for one or only a few models may be more important to the general biologist. AlphaFold2's state-of-the-art predicted models are increasingly relied upon and so it is vital that their accuracy is independently verified. In straightforward tertiary structure modelling AlphaFold2's predicted lDDT score (plDDT) has been considered a useful indicator of quality (Takei and Ishida, 2022), but it is unclear whether this reliability transfers to quaternary structure modelling and whether there are any occasions when the accuracy of these scores should be questioned.

5.1.1 AlphaFold2 predictions of model accuracy (plDDT, PAE and pTM)

plDDT is based on the local distance difference test (lDDT) (Mariani *et al.*, 2013) which compares distances between individual atoms to estimate confidence in the arrangement of amino acid residues in the local environment (for a full description see Appendix 1). It is useful for assessing the local accuracy of domains, for example, as it will not penalise incorrect relative orientations of domains within a model of a multi-domain protein if there is a good match between the inter-atomic distance matrices. AF2 provides local plDDT per-residue scores in the B-factor column of a model's coordinates file and a global per-model score which is output in the modelling log file.

The plDDT score itself is derived from the lDDT-Cα score (Tunyasuvunakool *et al.*, 2021) which considers only the backbone Cα atoms in the distance calculation rather than the full all-atom lDDT score. It has a range of 0-100 (but lDDT values are also sometimes quoted as decimals in the 0-1 range), with high scores indicating higher confidence (Jumper *et al.*, 2021). In general, plDDT values ≥ 90 equate to high confidence, those between 90 and 70 as confident, from 70 to 50 as low confidence and <50 as very low confidence with a tendency for disorder (Varadi *et al.*, 2022). These confidence levels mean that plDDT scores are somewhat different to regular all-atom lDDT scores. Pfam (Stroe, 2021), for example, considers lDDT scores of ≥0.6 as representing reasonable models, 0.7 as good quality models and those above 0.8 as great models.

PAE represents the Predicted Aligned Error for residue backbone atoms, measured in Ångströms and calculated for each residue. Values are designed to measure the confidence in the predicted super-position of any two residues within the model and the native structure and it can be used to compare the residue confidence scores within a domain to those between domains. Lower scores represent low predicted error and therefore higher confidence, and higher scores (capped at 31.75) (Varadi *et al.*, 2022) represent higher predicted error and therefore lower confidence. PAE is output as a colour-coded image mapping the areas of high and low confidence and also as machine-readable Json-formatted individual residue scores.

pTM is based on the topological similarity score TM-score (Zhang and Skolnick, 2004) and is calculated from the PAE matrix (Wallner, 2023). In later AlphaFold2 versions this is also output in the modelling log file and has a range of 0-1. No published confidence boundaries could be found for pTM but, traditionally, a TM-score of 1.0 would suggest a perfect match between a model and its native structure, a score greater than 0.5 is mostly interpreted as representing the same globular fold and scores below 0.17 are associated with unrelated proteins (Zhang and Skolnick, 2004). However, Jumper et al. (2021) described a relationship between pTM and TM-score as TM-score = 0.98 × pTM + 0.07 and so it may be appropriate to artificially construct pTM confidence boundaries using this relationship, if desired.

This study will concentrate on plDDT and pTM only for three simple reasons; PAE is not automatically normalised into an overall value meaning plDDT and pTM are the most often quoted AlphaFold2 confidence metrics; AF2 models are ranked by plDDT and AF2-Multimer models are ranked by pTM (Evans *et al.*, 2022) (see footnote[1] for Evans' description and ColabFold versions to which it applies), and that these scores are familiar and directly measurable against their observed counterparts, lDDT and TM-score.

5.1.2 Documented descriptions of AlphaFold2 predicted scores

One of the strengths of the AF2 algorithm has been described as its ability to recognise low accuracy local areas (Shao *et al.*, 2022) or indeed whole models and apply confidence scores appropriately. As stated above, linear relationships have been described (Jumper *et al.*, 2021) for lDDT-Cα as 0.997 × plDDT − 1.17 and TM-score as 0.98 × pTM + 0.07. While these relationships acknowledge a tendency for some over-prediction with plDDT, the suggestion is that both scores are consistently applied across the scoring range. However, at CASP15 (2022), despite plDDT and pTM scores from AF2 successfully contributing to many groups'

[1] a weighted combination of pTM and interface ipTM, calculated as (0.8 x ipTM + 0.2 x pTM). ColabFold v1.5.0 (Jan-2022 onwards) used the weighted ipTM-pTM score to rank multimers when using the AlphaFold2_mmseqs2, AlphaFold2_batch and colabfold_batch variants.

model-selection algorithms, it was noticed that there was a variability in these scores, particularly for multimer models of very similar quality. One group (Wallner, 2023) reported that up to one-third of models with a ranking confidence of pTM > 0.8 actually had the wrong domain orientation and our own experiences during CASP15 modelling revealed an increase in plDDT as high as 40 points during model refinement, which would suggest an overprediction of model quality improvement.

5.1.3 Wider uses of AlphaFold2 rely on accurate predicted quality

Since the CASP14 success detailed above, AlphaFold2 has been used in a DeepMind-EMBL collaboration to create the AlphaFold Protein Structure Database https://alphafold.ebi.ac.uk (Tunyasuvunakool *et al.*, 2021). This is aimed at creating a community resource allowing easy access to protein structures which remain unsolved by traditional experimental methods. With the growing profile of in-silico modelling against the backdrop of a growing community investment in artificial intelligence (AI), databases such as this are likely to increase in popularity along with a greater reliance on computational modelling software. Although, for now, the database is limited to tertiary structures, it might, nevertheless, be prudent to examine whether AlphaFold2's confidence metrics can be relied upon to rate and rank models accurately across the whole quality range.

Further to this, at least three published works describe using the AlphaFold derivative ColabFold to input models as custom templates. One group (Terwilliger *et al.*, 2022) input electron density maps during model generation from experimental data, another (Adiyaman *et al.*, 2023) described a procedure for model improvement using custom template recycling as a refinement strategy, and a third (Roney and Ovchinnikov, 2022) described a method for using AlphaFold2 as a quality assessment tool. The latter study suggested that AlphaFold2 has the ability to quality-rank sidechain-masked custom templates with state-of-the-art accuracy and that the results provide evidence for a neural network-learned protein folding energy function which AlphaFold2 is able to apply without external co-evolutionary data.

It is clear, then, that significant reliance is being placed on plDDT and pTM scores and this study aims to assess the performance of these scores in both monomer and multimer model populations in comparison to their observed lDDT and TM-score counterparts. Within the model populations, models will be generated both with and without custom template recycling to evaluate whether there is a variation in predictive performance with this single variable. In addition plDDT and pTM will be compared to quality scores generated by the independent quality assessment programs ModFOLD9 (tertiary structures) and ModFOLDdock (quaternary structures) (McGuffin *et al.*, 2023).

5.2 Objectives

Using blind modelling and assessment data from CASP15, the relationship between AlphaFold2 predicted scores and their observed counterparts, the lDDT score (including lDDT-Cα and oligo-lDDT) and the TM-score, will be examined. First, the analysis will attempt to objectively assess the scores' accuracy at describing tertiary and quaternary structures in terms of both global model quality and ranking agreement with observed scores. Second, blind prediction scores used for the CASP15 EMA competition will then be used to examine the comparative performance between ModFOLDdock and AF2-Multimer scores for quaternary structures. Similarly, ModFOLD9 predictions, which were also blind and run in house prior to the release of the CASP15 experimental structures, will be used to examine the performance between AlphaFold2 and ModFOLD9 scores for tertiary structures. Finally, the effect of using custom template recycling is examined in terms of the accuracy of the AlphaFold2 and AF2-Multimer scores. For this part, a CASP14 dataset similar to that described in Chapter 2 is used.

The study is designed around four primary and one secondary consideration. The following four hypotheses address the primary considerations.

1. Allowing for the published modest overestimation in plDDT, are AF2 predicted scores higher than the equivalent observed scores?
   H0. There is no increase in magnitude between the AF2 predicted and equivalent observed scores. H1. The magnitude of the AF2 predicted scores is higher than the equivalent observed scores.
2. Is AlphaFold2 model ranking reliable compared to ranking by observed scores, as measured by association between model rank categories?
   H0. There is no association between the AF2 predicted and observed score ranking categories. H1. There is an association between the AF2 predicted and observed score ranking categories.
3. Can model ranking accuracy be improved by independent MQA programs?
   H0. There is no difference between the independent QA and AF2 rankings as measured by the association between model rank categories. H1. Independent QA and observed score model rank are more closely associated than AF2 and observed score model ranks.
4. Is the accuracy of predicted scores affected by custom template recycling?
   H0. There is no difference between AF2 regular modelling and custom template modelling predicted scores, when compared to equivalent observed scores. H1. AF2 predicted scores following custom template modelling show greater variation than scores from regular modelling, when compared to equivalent observed scores.

Secondary consideration. Do the results support the notion by Roney and Ovchinnikov that AlphaFold2 can be successfully repurposed as a general model quality assessment tool?

5.3 Materials and Methods

5.3.1 Selection of models to test the hypotheses

Four individual datasets were used for this study.

Population A (CASP15 monomers) comprised the McGuffin group's tertiary structure submissions for CASP15. Population B (CASP15 multimers) was composed of both the McGuffin group's (MultiFOLD, group 462) and the ColabFold group's (group 446) multimer submissions for CASP15. Group 446 submissions are publicly available from https://casp15.colabfold.com/). Population C (recycled monomers) is a superset of the AF2 and non-AF2 models used in the custom-template recycling experiment described in Chapter 2. The original model population was fixed at 16 CASP14 targets to form a common subset with the ReFOLD4 molecular dynamics analysis which used only the FM targets submitted by the AlphaFold group. The emphasis of this experiment has shifted from measuring model improvement to measuring model quality overprediction and so it was felt that the inclusion of four additional FM/TBM targets, for which scores had already been collected, was justified to increase the model population without significantly altering the difficulty of the models. This increased the total target number to 20. Population D (recycled multimers) is the same multimer population used in the custom-template recycling experiment also in Chapter 2.

5.3.2 The Population A dataset – CASP15 monomers

This consisted of all McGuffin group's blind model submissions for CASP15 regular tertiary structure targets for which ModFOLD9 scores and a reference native structure were available. The dataset comprised a total of 26 targets: T1104, T1112, T1120, T1122, T1125, T1130, T1131, T1133, T1139, T1145, T1146, T1147, T1150, T1154, T1155, T1158, T1159, T1162, T1163, T1175, T1177, T1180, T1182, T1183, T1188 & T1194.

Our group's modelling algorithm used two separate rounds as shown in Chapter 2, Figure 2.15. Round 1 used regular modelling only, with no refinement process, whereas round 2 included refinement by custom template recycling. Models were therefore split into two sub-populations; Population A1 represented the round 1 models (regular modelling) and these were created with a default of 12 recycles and both with and without AMBER relaxation, resulting in 20 models per target (5 unrelaxed AF2, 5 relaxed AF2, 5 unrelaxed AFM and 5 relaxed AFM). For a small minority of large targets memory constraints meant relaxation was not always possible resulting in fewer models. Population A2 represented the round 2 models (denoted by the addition of R to the model's name, e.g., AFMR) which were subject to custom template recycling and resulted in 10 models per target. Again, 5 of these underwent AMBER relaxation while the other 5 remained unrelaxed. In this way a maximum of 30 models were created per target. Predicted plDDT and pTM scores were harvested directly from the server for both sub populations and predicted ModFOLD9 scores were collected from the original cached datasets

used during CASP15. Observed lDDT and TM-scores were generated using the downloadable versions of TM-score (Zhang and Skolnick, 2004) and lDDT score (Mariani *et al.*, 2013) to compare models for each target with the CASP observed structures. A total of 735 models were analysed; consisting of 490 round 1 and 245 round 2 models.

5.3.3 The Population B dataset – CASP15 multimers

This model population comprised all blind multimer (assembly) CASP15 model submissions for both the MultiFOLD (462) and ColabFold (446) group servers. These two sets of models were chosen because they were created using the same base ColabFold software (although exact versions may differ slightly) but differed by the use of custom template recycling in the MultiFOLD pipeline. The rationale was that the ColabFold models could be used to asses AF2-Multimer score overprediction when only regular modelling was used and, that by comparing the ColabFold and MultiFOLD populations, the effect of the additional custom template recycling on predicted scores could be assessed.

The ColabFold group multimers are named Population B1. For these, custom template recycling and AMBER relaxation were not used and 12 recycles was used as default (Ovchinnikov *et al.*, 2022). The predicted scores, plDDT, pTM (and iPTM where available) were harvested directly from the server website (see 5.3.1 for the URL).

The MultiFOLD group models are named Population B2. The same pathway as outlined in 5.3.2 above, including custom template recycling, was used to create these models. Only the final 5 models submitted to CASP were used for analysis and again predicted scores were collected directly from the server.

For comparisons with observed scores, the official CASP15 assessor oligo-lDDT and TM-scores were downloaded from the CASP15 prediction centre results webpage (https://predictioncenter.org/casp15/results.cgi?view=targets&tr_type=multimer). As the ModFOLDdock server participated in the CASP15 EMA experiment, predicted ModFOLDdock and ModFOLDdockR scores were also readily available for both sub populations of models.

Scores for rank 1 to 5 models were collected for all multimer models for which data were available, resulting in 395 individual models across the following 41 targets (the ColabFold group submitted no models for three targets making a total of 38); H1106, H1111, H1114, H1129, H1134, H1135, H1137 (MultiFOLD only), H1140, H1141, H1142, H1143, H1144, H1151, H1157, H1166, H1167, H1168, H1171, H1172, H1185, T1109, T1110, T1113, T1115 (MultiFOLD only), T1121, T1123, T1124, T1127, T1132, T1153, T1160, T1161, T1170, T1173, T1174, T1176, T1178, T1179, T1181, T1187 and T1192 (MultiFOLD only). In total the Population B dataset consisted of 395 multimer scores.

5.3.4 The Population C dataset – recycled monomers

This dataset consisted of the custom template recycled AlphaFold2 and non-AlphaFold2 tertiary models used in Chapter 2, with the addition of four extra targets as explained in section 5.3.1 above. There were minor processing differences when creating the recycled AF2 and non-AF2 models which are explained below.

The AlphaFold2 Rank 1 models were downloaded from the CASP14 website for the following 20 CASP14 FM targets: T1027, T1029, T1031, T1033, T1037, T1039, T1040, T1041, T1042, T1043, T1047s1, T1047s2, T1055, T1058, T1064, T1074, T1090, T1093, T1094, T1096. Again, as described in section 5.3.2, observed quality assessment scores were generated using the downloadable versions of TM-score and lDDT score. To affect the recycling, model PDB files were converted to mmCIF format using the RSCB PDB MAXIT suite of programs (https://mmcif.pdbj.org/converter). These were then submitted to the Google Colaboratory hosted ColabFold (release 3, v1.3.0 [4-Mar-2022]) as custom templates along with their respective amino acid sequences. ColabFold was run twice per model (both MSA and single-sequence modes), and, within each mode, the model was submitted four times for 1, 3, 6 and 12 recycles. ColabFold settings used were: Template_mode: custom; msa_mode: MMseqs2 (UniRef+Environmental) OR single sequence; pair_mode: unpaired+paired; model-type: auto; num_recycles: 1, 3, 6, 12 (selecting “auto” from the model type defaulted to the original pre-CASP14 AF2 model). Amber relaxation was not enabled. Models created for each ColabFold run were collected along with their predicted pTM and plDDT scores and then rescored with TM-score and lDDT as described above. The process, illustrated below in Figure 5.1, created 800 individual scores from 8 sets of scores per model across 5 models per target for 20 targets.

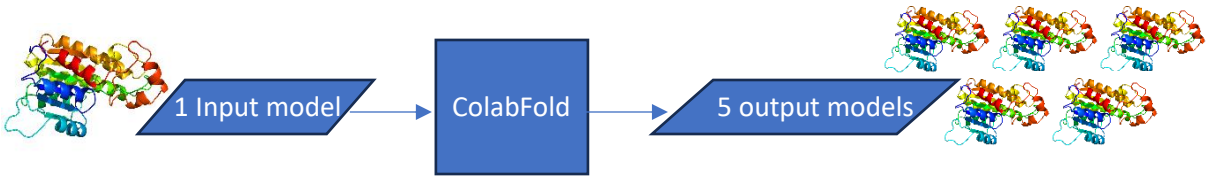


Figure 5.1 An illustration of input and output models during ColabFold custom template recycling. Each model (template) was input into ColabFold eight times using different recycling modes (MSA and single sequence) and produced five new models by default each time.

The same logic was employed for non-AF2 CASP14 models. These were selected from the same 20 FM targets for the next five best-ranked groups beneath AlphaFold2 at CASP14. These were Baker (473), Baker-experimental (403), Feig-R2 (480), Zhang (129) and tFold_human (009). To ensure consistency in terms of globular fold similarity, only models with a TM-score ≥0.45 were selected and this resulted in a total of 47 individual models.
The full list of models used is:

T1029TS009_1-D1, T1031TS009_1-D1, T1033TS009_1-D1, T1037TS009_1-D1, T1041TS009_1-D1, T1042TS009_1-D1, T1043TS009_1-D1, T1049TS009_1-D1, T1090TS009_1-D1, T1031TS129_1-D1, T1037TS129_1-D1, T1040TS129_1-D1, T1041TS129_1-D1, T1042TS129_1-D1, T1049TS129_1-D1, T1074TS129_1-D1, T1090TS129_1-D1, T1096TS129_1, T1027TS403_1-D1, T1031TS403_1-D1, T1033TS403_1-D1, T1037TS403_1-D1, T1039TS403_1-D1, T1041TS403_1-D1, T1042TS403_1-D1, T1043TS403_1-D1, T1049TS403_1-D1, T1090TS403_1-D1, T1096TS403_1, T1031TS473_1-D1, T1033TS473_1-D1, T1037TS473_1-D1, T1039TS473_1-D1, T1041TS473_1-D1, T1042TS473_1-D1, T1043TS473_1-D1, T1049TS473_1-D1, T1074TS473_1-D1, T1090TS473_1-D1, T1031TS480_1-D1, T1037TS480_1-D1, T1041TS480_1-D1, T1042TS480_1-D1, T1049TS480_1-D1, T1074TS480_1-D1, T1090TS480_1-D1, T1096TS480_1.

Models were downloaded from the CASP14 website, scored with TM-score and lDDT and recycled as templates with the MSA option in the same way as described for AF2 models. Single sequence recycling was carried out using release v1.3.0 of LocalColabFold (Mirdita *et al.*, 2022) installed on our own server to overcome the Google Colaboratory GPU restrictions in the time available. The equivalent LocalColabFold settings were used: msa-mode: single_sequence; model-type: auto; rank: plddt; pair-mode: unpaired+paired; templates: --custom-template-path. The resulting rank 1-5 models were collected along with their plDDT and pTM scores and rescored against the native structure to produce a set of observed lDDT and TM-scores. This again resulted in eight sets of five models per input model creating a total of 1880 individual model scores.

5.3.5 The Population D dataset – recycled multimer models

This dataset consisted of the custom template recycled multimer models used in Chapter 2. As the AlphaFold2 group did not submit multimer (assembly) models at CASP14, models for this dataset were selected from the CASP14 top five ranked groups. According to official results tables, these were Baker, Venclovas, Takeda-Shitaka, Seok and DATE. Some of the multimer targets were too large to recycle through AF2-Multimer (training was limited to models up to 1536 residues and the algorithm can experience memory issues with models of more than a few thousand residues (Bryant *et al.*, 2022)) and therefore the targets used in this set were limited by size to: H1045, H1065, H1072, T1032, T1054, T1070, T1073, T1078, T1083, T1084. Again, top-ranked models were used as the custom templates and were subjected to recycling (1, 3, 6 and 12) using ColabFold (MSA and SS modes) in the same way as described for the monomer structures above. The resulting 50 rank 1-5 models were then collected along with their plDDT and pTM scores. Observed scores were obtained by assessing each model against their relevant native structures using the OpenStructure and MM-Align (Mukherjee and

Zhang, 2009) programs to obtain observed oligo-lDDT and TM-scores respectively. Using the same calculation as above, this resulted in eight sets of scores for each of the five models per individual group-target combination, a total of 2000 individual scores. The processing of Population D described above was conducted by Ahmet Gurkan Genc and kindly shared with me as part of the joint experiment on recycling described in Chapter 2. An overall total of 5,810 model scores were collected across the whole study.

Table 5.1. A summary of the different model populations used in the study.

Model population	Type and modelling software	Stoichiometry and type of modelling
Population A1	CASP15, MultiFOLD round 1	Monomer, regular modelling.
Population A2	CASP15, MultiFOLD round 2	Monomer, custom recycling included.
Population B1	CASP15, ColabFold	Multimer, regular modelling.
Population B2	CASP15, MultiFOLD	Multimer, custom recycling included.
Population C	CASP14, AF2 and non-AF2	Monomer, custom recycling included.
Population D	CASP14, top 5 groups.	Multimer, custom recycling included.

5.3.6 Handling of Multimer pTM scores and the procedure for model ranking

Multimer models created by AlphaFold2 variants are, by default, ranked by pTM rather than plDDT. As stated in the introduction there is a slight difference in the calculation of the pTM-based ranking between versions of ColabFold. In AF2-Multimer and in later versions of ColabFold (v1.5.0) ranking is calculated based on a ratio of 0.8 * ipTM + 0.2 * pTM (Evans *et al.*, 2022), whereas in earlier versions, ranking is calculated on pTM score alone. As some multimer models in this population were created with ColabFold v1.3 and some with v1.5 there was potentially heterogeneous ranking across the model population, and it was necessary to allow for this when comparing ranks. To this end, multimer models were routinely re-ranked by pTM score before comparison with observed rankings. The procedure for deriving model ranks in R consisted of ranking each individual set of 5 related models, i.e., models output from a single run of AlphaFold2 modelling, using the statement *rank <score>, ties.method =”random”* where <score> can be replaced with any of the predicted or observed scores as necessary. This was applied to observed score ranking but also to ranking by pTM for the reasons explained above. In this way any differences in the way the AlphaFold2 algorithm originally ranked the data were negated and the ranks were assigned uniformly across all populations.

Multi-factor contingency tables to display ranking comparisons were created in R using the caret package with the confusionMatrix() command and four further statistical measures were used to assess relatedness. Sensitivity, specificity, precision, and accuracy were calculated for individual rank classes (1, 2, 3, 4 and 5) and, to construct meaningful comparisons between the contingency tables, macro-averaged versions of these statistics were calculated as mean values across all categories for each table. Individual metrics were calculated as follows:

Sensitivity = TP / (TP + FN),

Specificity = TN / (TN + FP),

Precision = TP / (TP + FP) and

Accuracy = (TP + TN) / (TP + FN + TN + FN).

(TP=true positive, TN=true negatives, FP=false positive and FN=false negatives, see Appendix 13 for a more comprehensive description of these metrics)

As ranking data is categorical, it is possible to assess the association between the predicted model ranks and observed model ranks using the Chi-squared and Fisher's exact tests, where P-values <0.05 would suggest relatedness between distributions. Fisher's exact test is often used for smaller sample sizes (single contingency table cells of less than 5) or where independence of observations cannot be guaranteed and, while the concept of independence holds for the assignment of ranks based on predicted and observed scores, some tables do have low figures in individual cells. As regards multi-contingency tables (larger than 2x2), no clear distinction between the two tests could be found other than Chi-squared may run into problems with very sparse data and Fisher's can become computationally intensive for large tables. It was decided that both tests would be run as a check for each other, i.e., agreement between the two tests would confer confidence in the result. A Monte Carlo resampling method (simulate.p.value) with default simulations of 2000 was used for the Fisher's exact test to allow a more robust estimate of the p-value and prevent any computational overheads which can occur when this test is applied to larger contingency tables. This procedure generates 2000 random datasets and computes the test statistic for each one. The sample test statistic is then compared to the distribution of simulated test statistics to estimate the p-value whilst avoiding exhaustive calculations (Crawley, 2015). Analysis was performed using R version 3.6.3 running in R-studio.

5.4 Results and Discussion

Results will be considered in relation to the four hypotheses in the objectives.

5.4.1 Hypothesis 1. Are AF2 predicted scores higher than the equivalent observed scores?

In this study hypothesis four deals specifically with the effects of custom template recycling on predicted score reliability. Therefore, in order to focus on one independent variable at a time, the simpler question of whether predicted scores are good quality indicators must be answered using only models which have *not* undergone custom template recycling. For monomers, this is population A1 (round 1 models) and for multimers this is population B1 (ColabFold multimers). Population A1 will be considered first.

5.4.1.1. Part 1. Monomer data; Population A1, (round 1)

Monomers are ranked by default by plDDT scores and so monomer results will focus on plDDT/lDDT similarity.

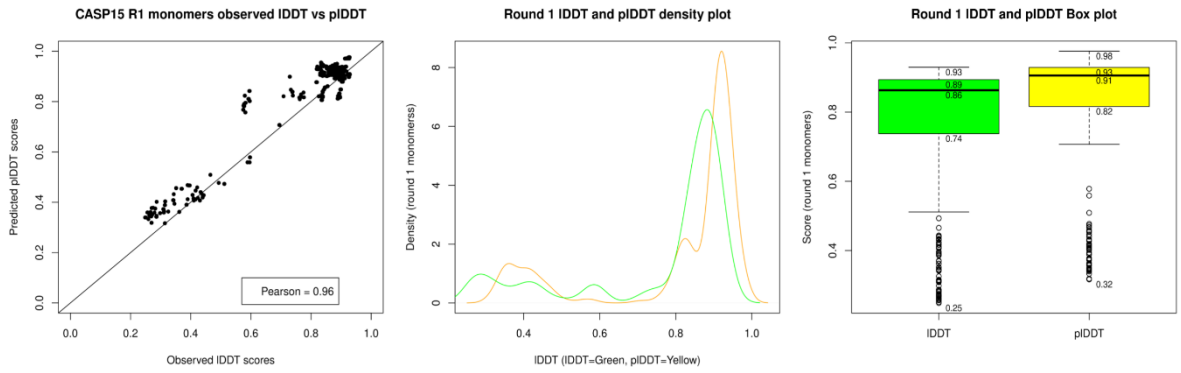


Figure 5.2. Plots of plDDT versus observed lDDT for round 1 monomers in population A1. Left. A scatter plot. **Middle.** A density plot. **Right.** A boxplot for the same population. For all plots plDDT has been rescaled to fit the 0-1 lDDT range.

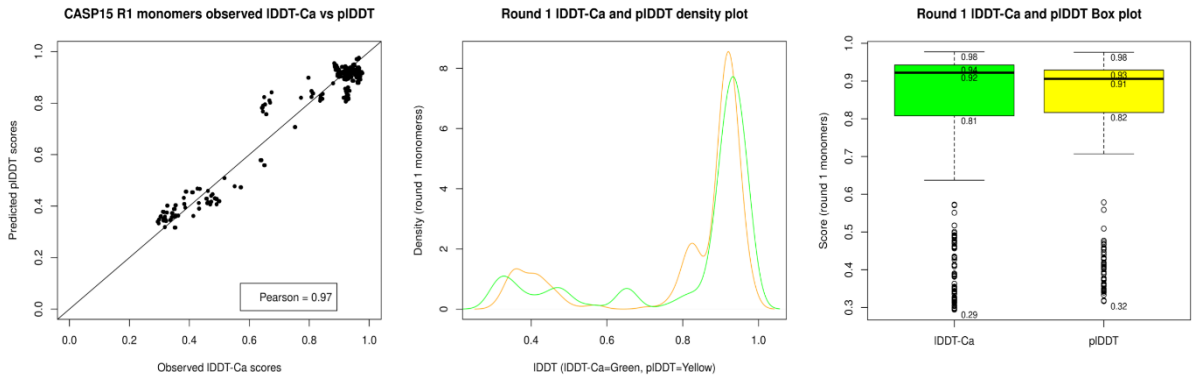


Figure 5.3. Plots of plDDT versus observed lDDT-Cα for round 1 monomers population A1. Left. A scatter plot. **Middle.** A density plot. **Right.** A boxplot for the same population. Again, plDDT has been rescaled to the 0-1 range.

The plots in Figure 5.2 show that plDDT scores are slightly elevated compared to the all-atom lDDT scores. However, when plDDT scores are considered with reference to lDDT-Cα scores

(Jumper *et al.*, 2021;Tunyasuvunakool *et al.*, 2021) in Figure 5.3, there is no evidence of plDDT over-prediction, in-fact the boxplot in Figure 5.3 shows a slightly lower median score for plDDT. It should also be possible to check whether the plDDT values in this sample are in line with the published linear relationship described in section 5.1.2 (lDDT-Cα=0.997×plDDT− 1.17). If the median plDDT value of 0.91 from the Figure 5.3 boxplot is considered as a convenient example, the median lDDT-Cα score can be calculated from plDDT in three simple steps:

1. Convert plDDT back to its 0-100 range: 0.91 x 100 = 91.0
2. Calculate lDDT-Cα from the relationship: 0.997 x 91 - 1.17 = 89.56
3. Convert lDDT-Cα back to the 0-1 range: 89.56/100 = 0.8956 or 0.90 to 2.d.p.

From Figure 5.3 it can be seen that the actual lDDT-Cα is 0.92, meaning that rather than being overpredicted, plDDT has in fact been slightly underpredicted for this sample of models.

To formally test this data against hypothesis 1, a Wilcoxon signed rank test for non-parametric paired data was carried out to test significance. The following results were obtained (a Shapiro test for normality gave p-values of <0.05 for all three (plDDT, lDDT and lDDT-Cα) scores, showing the distributions to be non-normal in all cases).

Table 5.2. Calculated p-values from a Wilcoxon signed rank test for population A1, round 1 monomers. P-values ≤0.05 are in bold.

Scores compared	***Independence and distribution symmetry***	***p-value***
plDDT versus lDDT	Paired; 2-sided test.	**$2.2x10^{-16}$**
plDDT versus lDDT	Paired; 1-sided (plDDT > lDDT)	**$2.2x10^{-16}$**
plDDT versus lDDT-Cα	Paired; 2-sided test.	**$9.69x10^{-6}$**
plDDT versus lDDT-Cα	Paired, 1-sided (plDDT < lDDT-Cα)	**$4.81x10^{-6}$**

Wilcoxon signed-rank test P-values were calculated at the 95% confidence level using plDDT and lDDT or lDDT-Cα scores.

From the results in Table 5.2, it can be concluded that, for this sample of monomers, there is a significant difference between predicted and observed lDDT scores as shown by the P-values of $2.2x10^{-16}$ and $9.69x10^{-6}$ for the 2-sided Wilcoxon tests for all atom lDDT and lDDT-Cα respectively. However, there is disagreement between the two scores, with row 2 of the table showing that according to a 1-sided test, plDDT values are greater than those for all atom lDDT (p-value of $2.2x10^{-16}$) while row 4 shows the opposite, that lDDT-Cα values are actually significantly higher than plDDT values (p-value of $4.81x10^{-6}$). Considering the published works cited earlier confirming that plDDT is based on lDDT-Cα it would be more appropriate to accept the null hypothesis in this case. Therefore, for monomers constructed from regular straight-forward AF2 modelling and assessed by lDDT-Cα: *There is no increase in magnitude between the AF2 predicted and equivalent observed scores*.

5.4.1.2 Part 2. Multimer data; Population B1 (ColabFold multimers).

For multimers pTM is the default ranking metric, however plDDT was used in early versions of ColabFold and so both scores are considered here.

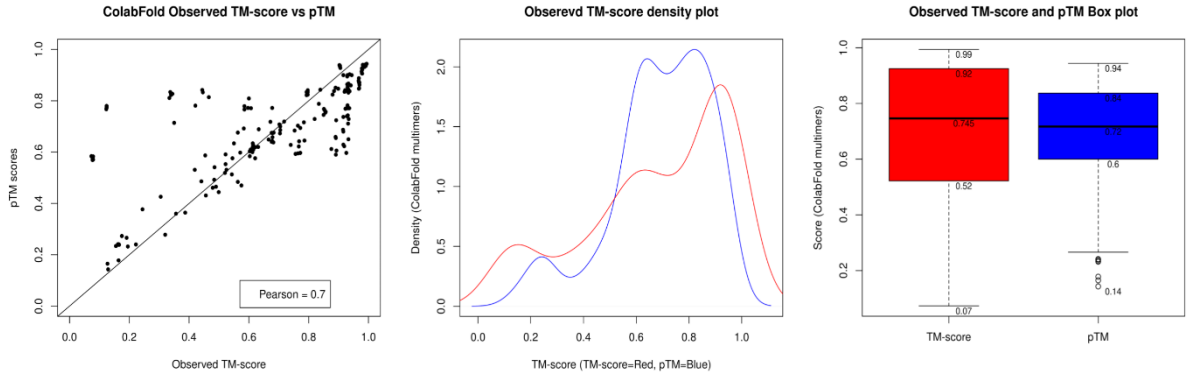


Figure 5.4. Plots of pTM score versus observed TM-score for Population B1 (ColabFold multimers). Left. A scatter plot. **Middle.** A density plot. **Right.** A boxplot.

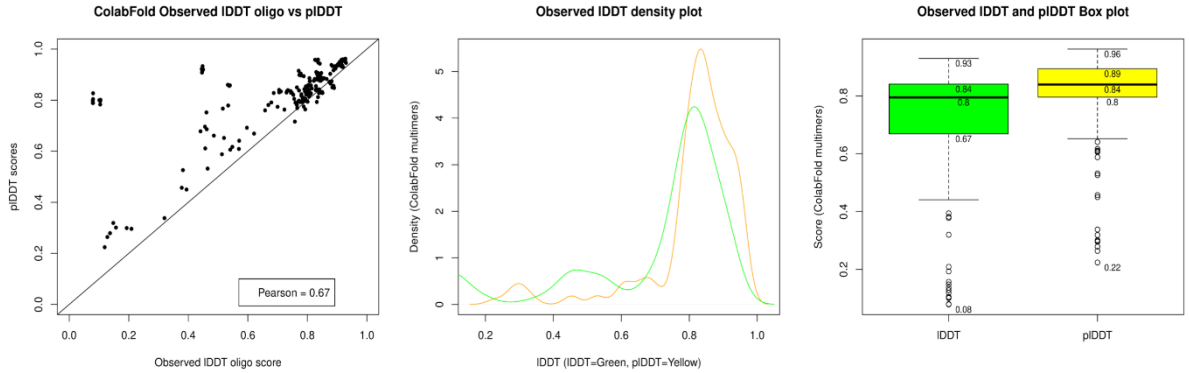


Figure 5.5. Plots of plDDT score versus observed CASP oligo-lDDT for Population B1 (ColabFold multimers). Left. A scatter plot. **Middle.** A density plot. **Right.** A boxplot. plDDT has been rescaled to 0-1.

Both the scatter and density plots in Figure 5.4 appear to show an under-estimation of pTM score for higher quality multimer models but a relatively large overestimation for some lower-quality models. For Figure 5.5, plDDT appears to be over-estimated across the quality range which may be accounted for by the use of an all-atom observed oligo-lDDT score. However, as with pTM scores, there is a more pronounced overestimation for some models in the lower quality range. The Shapiro test for normality (all scores were non-normal) and Wilcoxon signed rank test for significance were executed in the same way as described for monomer data.

Table 5.3. Calculated p-values from a Wilcoxon signed rank test for population B1, ColabFold multimers. P-values ≤0.05 are in bold.

Scores compared	***Independence and distribution symmetry***	***p-value***
plDDT versus oligo-lDDT	Paired; 1-sided test, plDDT > oligo-lDDT	**2.2×10^{-16}**
pTM versus TM-score	Paired, 2-sided	**0.038**
pTM versus TM-score	Paired; 1-sided test, pTM> TM-score	0.980
pTM versus TM-score	Paired; 1-sided test, pTM< TM-score	**0.0192**

Wilcoxon signed-rank test P-values were calculated at the 95% confidence level using plDDT and oligo-lDDT or pTM and TM-scores.

Table 5.3 shows that there is a significant difference between predicted plDDT and observed oligo-lDDT scores and that plDDT values are significantly higher than oligo-lDDT as shown by the p-value of 2.2×10^{-16}. For hypothesis 1, with respect to lDDT, the alternative hypothesis can therefore be accepted for ColabFold multimers, i.e., *The magnitude of the AF2 predicted scores is higher than the equivalent observed scores.*

The data are not so clear for TM scores. There is a significant difference between pTM and TM-score but rather than pTM being the greater of the two (p-value of 0.980), TM-score may in fact be greater than pTM (p-value of 0.019). To reveal more information about the relationship between pTM and TM-score, a further investigation into the variation in the two scores is described in Figure 5.6 below.

Figure 5.6. Two plots showing the difference between predicted and observed scores for population B1 (ColabFold multimers). The line at 0.0 represents the observed score; predicted scores are represented as points. **Left.** pTM versus TM-score. **Right.** plDDT (rescaled to 0-1) versus oligo-lDDT score. Numbers on the x-axis are the models in the population, ordered from low to high observed score.

The relationships suggested in Figure 5.5 and Table 5.3 are more clearly shown by the two plots in Figure 5.6. Both plots show that an overestimation of predicted scores is more likely for lower quality models with a maximum difference of +0.65 for pTM and +0.74 for plDDT. Again, a tendency for underestimation of pTM in higher quality models is apparent with a maximum difference of -0.32. This explains the Wilcoxon test results for pTM; there is both over and under-estimation occurring which is quality-related and which, to some extent, cancel each other out. While there is an allusion to minor pTM underprediction in the mathematical relationships described in section 5.1.2 (Jumper *et al.*, 2021), no documentation relating to an overprediction for lower quality models could be found. A similar pattern of underestimation is not seen for plDDT.

For hypothesis 1, with respect to TM-score, the null hypothesis must be accepted for ColabFold multimers, i.e., *There is no increase in magnitude between the AF2 predicted and equivalent observed scores.* However, a caveat must be added to this last statement, that, for this population (intended to represent regular multimer modelling), while a significant increase in predicted TM-score could not be detected in the overall population, overprediction was observed in models of lower observed quality.

5.4.2 Hypothesis 2. Is AlphaFold2 model ranking reliable compared to ranking by observed scores, as measured by association between model rank categories?

Again, to answer this question fairly, models which have not undergone custom template recycling must be used. Therefore, this analysis will use the same data as used in 5.4.1 - Population A1 (round 1 models) and Population B1 (ColabFold multimers). Ranking values and statistics were calculated as described in section 5.3.6.

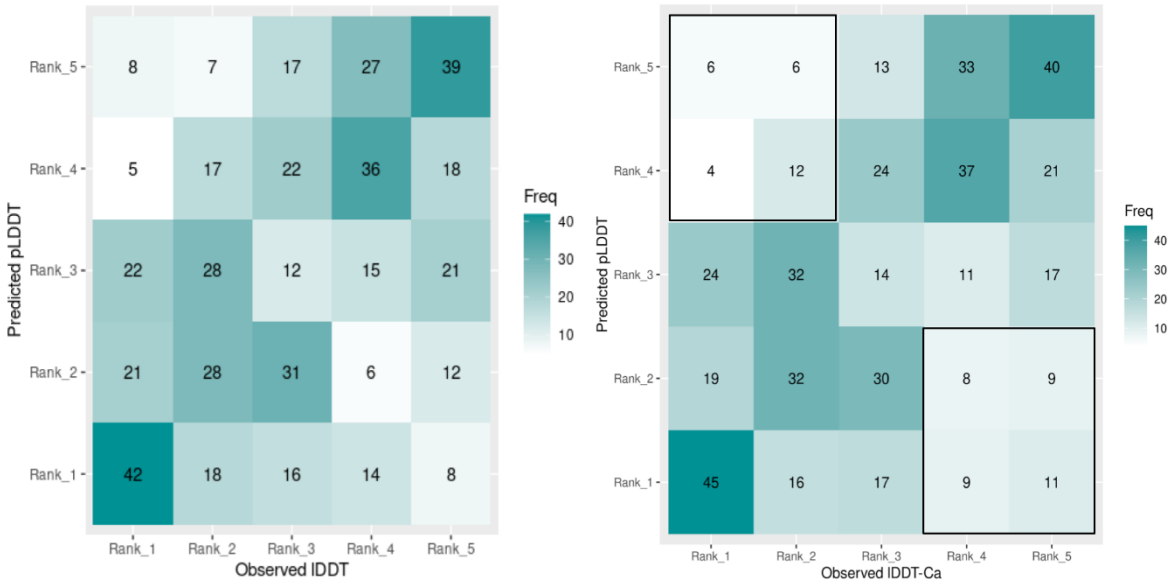


Figure 5.7. Contingency tables showing the rank agreement between observed lDDT and plDDT values for Population A1 (round 1 monomers). Left. For all-atom observed lDDT scores. **Right.** For observed lDDT-Cα scores. Accompanying table of calculated statistics below.

Table 5.4. Summary statistics, including four macro-averaged test characteristics, Fisher's exact test and Chi-squared test for population A1 (round 1 monomers) ranking agreement between predicted and observed ranks.

Test	*lDDT result*	*lDDT-Cα result*
Macro-Sensitivity (TPR)	0.3204	0.3428
Macro-Specificity	0.8301	0.8357
Macro-Precision	0.3204	0.3428
Macro-Accuracy	0.7281	0.7371
Fisher's Exact (p-value)	< 0.001	< 0.001
Chi-squared (χ^2; p-value)	128.27; 2.2×10^{-16}	167.35; 2.2×10^{-16}

P-values were calculated at the 95% confidence level meaning those ≤0.05 are considered significant.

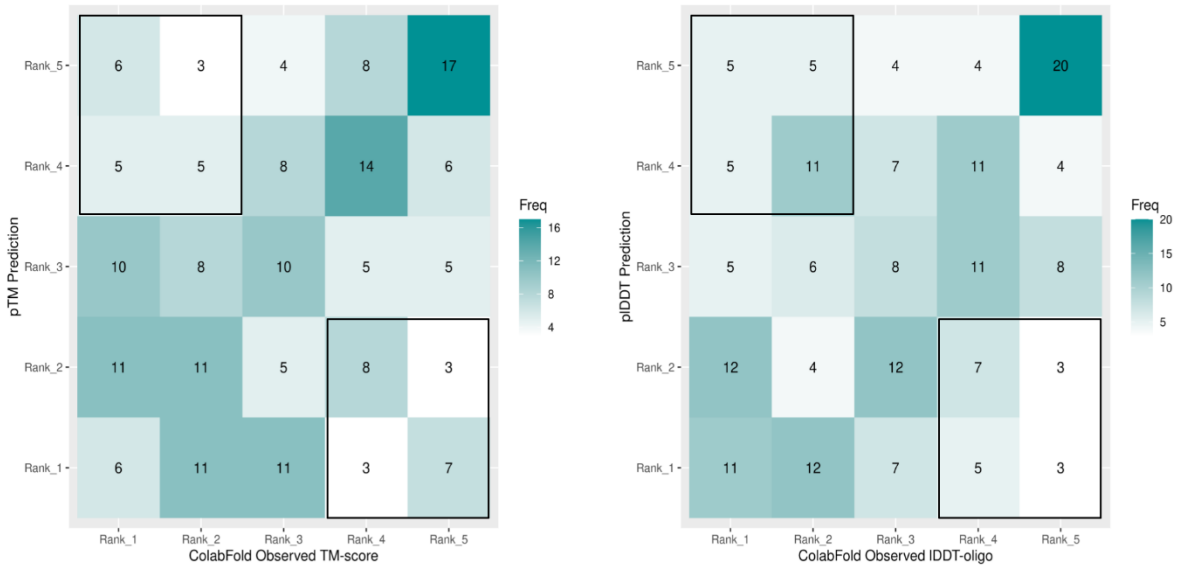


Figure 5.8. Contingency tables showing rank agreement for Multimers in Population B1 (ColabFold multimers). Left. Observed TM-scores versus pTM. **Right.** Observed oligo-lDDT versus plDDT scores. Accompanying table of calculated statistics below.

Table 5.5. Summary statistics, including four macro-averaged test characteristics, Fisher's exact test and Chi-squared test for population B1 (ColabFold multimers) ranking agreement between predicted and observed ranks.

Test	***pTM result***	***plDDT result***
Macro-Sensitivity (TPR)	0.3052	0.2842
Macro-Specificity	0.8263	0.8210
Macro-Precision	0.3052	0.2842
Macro-Accuracy	0.7221	0.7136
Fisher's Exact (p-value)	< 0.001	< 0.001
Chi-squared (χ^2, p-value)	40.26; 0.0007	51.31; 1.41×10^{-5}

P-values were calculated at the 95% confidence level meaning those ≤0.05 are considered significant.

The contingency tables in Figure 5.7 show strong agreement for monomer data between observed lDDT-derived ranks and plDDT predicted ranks with a slightly stronger agreement when lDDT-Cα is used as the observed measure. The level of agreement for rank 1 and rank 5 data shown in the contingency tables is supported by mean true positive rates (TPR) of 32.04% and 34.28% for lDDT and lDDT-Cα respectively. In addition, the Fisher's exact tests return p-values well below the significance level of 0.05 and the Chi-squared tests return values of 128.27 (lDDT) and 167.35 (lDDT- Cα) with very small p-values. These data provide robust evidence that this distribution was unlikely to occur by chance and that there is a significant positive relationship between the predicted and observed scores.

For the multimer population represented by Figure 5.8, the agreement looks appreciably less certain for both pTM and plDDT scores. The summary statistics show a reduction in mean TPR to 30.5% for pTM and 28.4% for plDDT. Both Fisher's exact and Chi-squared p-values, however, remain significant suggesting a relationship between the two rank sets, although it is

notable that the magnitude of the χ^2 statistic has decreased for both pTM and plDDT suggesting a weaker association between predicted and observed ranks.

For hypothesis 2, these results suggest that there is significant association between the distribution of predicted and observed ranks for both monomer and multimer model populations created via regular modelling and the alternative hypothesis can be accepted, i.e., *There is an association between the AF2 predicted and observed score ranking categories.* Similarly, to section 5.4.1, though, a qualifying statement may be appropriate here to add that the association appears more robust for tertiary structure ranking by plDDT than for multimer ranking by either plDDT or pTM.

5.4.3 Hypothesis 3. Can model ranking accuracy be improved by independent MQA?

The individual rank-agreement and TPR values described above for monomer and multimer models need to be contextualised by comparison to another leading QA method. This section presents identical analysis for ranking based on predicted scores from the independent QA programs ModFOLD9 (monomer data) and ModFOLDdock (multimer data).

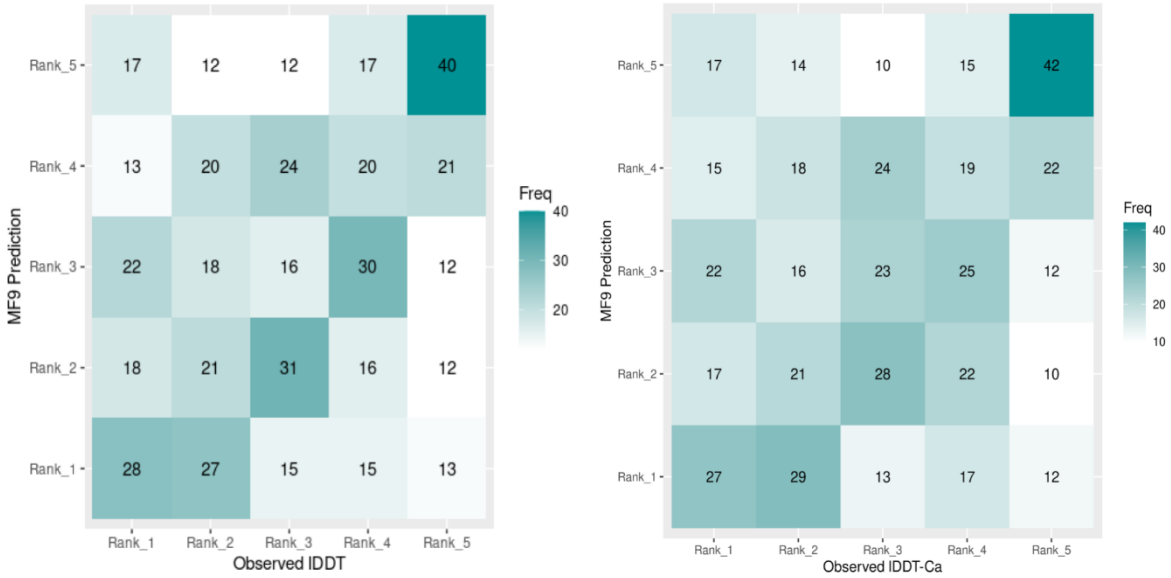


Figure 5.9. Contingency tables showing the rank agreement between observed lDDT and ModFOLD9 values for Population A1 (round 1 monomers). Left. Using all-atom lDDT scores. **Right.** Using observed lDDT-Cα scores. Accompanying table of calculated statistics below.

Table 5.6. Summary statistics for population A1 (round 1 monomers) ranking agreement between predicted ModFOLD9 and lDDT observed ranks.

Test	*lDDT result*	*lDDT-Ca result*
Macro-Sensitivity (TPR)	0.2551	0.2693
Macro-Specificity	0.8137	0.8173
Macro-Precision	0.2551	0.2693
Macro-Accuracy	0.7020	0.7077
Fisher's Exact (p-value)	< 0.001	< 0.001
Chi-squared (χ^2; p-value)	61.93; 2.5×10^{-7}	63.67; 1.24×10^{-7}

P-values were calculated at the 95% confidence level meaning those ≤0.05 are considered significant.

Visual comparison of the data in Figure 5.9 to those in Figure 5.7 shows that ModFOLD9 has been unable to improve upon the ranking agreement between plDDT and lDDT scores for monomers. TPR is reduced from 34.2% to 26.9% (lDDT-Cα) and all other macro-averaged statistics are lower than previously reported. Although the Fisher's exact and Chi-squared tests continue to return significant p-values, the χ^2 values, in agreement the TPR scores, have reduced suggesting a weaker overall association between the ranks.

Therefore, the closeness of the relationship has not been improved by ModFOLD9 and for hypothesis 3, in respect to ModFOLD9, the null hypotheses must be accepted, i.e., *There is no difference between the independent QA and AF2 rankings as measured by the association between model rank categories.*

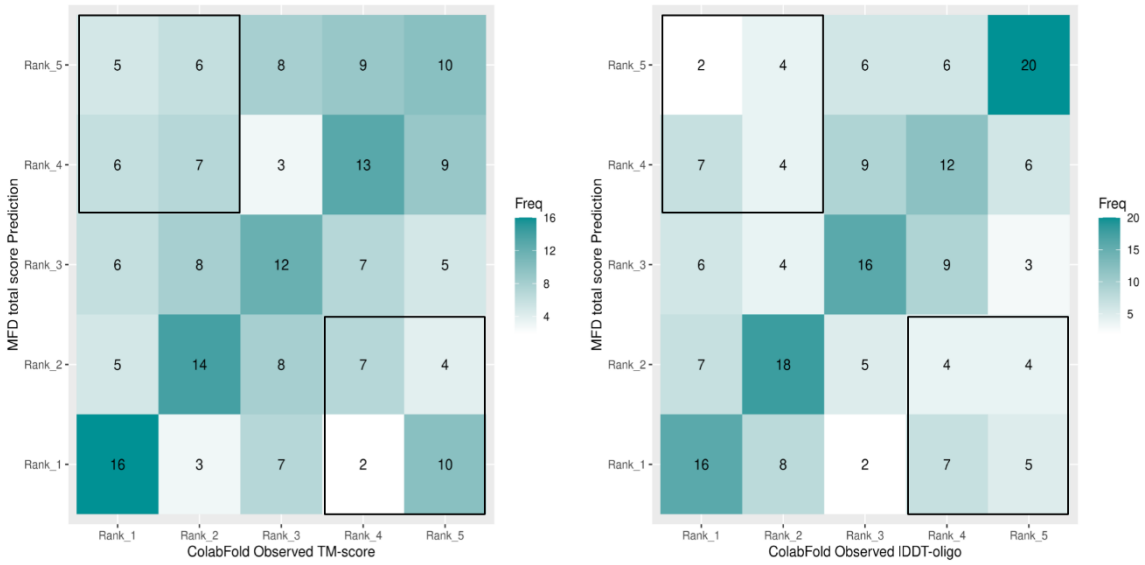


Figure 5.10. Contingency tables showing rank agreement for Population B1 (ColabFold multimers). Left. Between observed TM-scores and ModFOLDdock score. **Right.** Between observed oligo-lDDT and ModFOLDdock score. Accompanying table of calculated statistics below.

Table 5.7. Summary statistics for population B1 (ColabFold multimers) ranking agreement between predicted ModFOLDdock and observed oligo-lDDT ranks.

Test	*TM-score Result*	*lDDT result*
Macro-Sensitivity (TPR)	0.3421	0.4315
Macro-Specificity	0.8355	0.8578
Macro-Precision	0.3421	0.4315
Macro-Accuracy	0.7368	0.7726
Fisher's Exact (p-value)	< 0.001	< 0.001
X-squared (χ^2; p-value)	38.42; 0.0013	78.94; 2.57×10^{-10}

P-values were calculated at the 95% confidence level meaning those ≤0.05 are considered significant.

In contrast, a visual comparison of the data in Figure 5.10 with those from Figure 5.8 suggests ranking agreement for multimers is stronger for ModFOLDdock scores, particularly for lDDT rank agreement. This is supported by the data in Table 5.7 where the TPR has increased from 30.5% in Table 5.5 to 34.2% in Table 5.7 for TM-score and more appreciably from 28.4% to

43.1% for oligo-lDDT score. The Chi squared values have remained similar for TM-score across the two tables, however Table 5.7 shows an increase in the χ^2 statistic from 51.31 to 78.94 for oligo-lDDT ranking. This increase, along with the increased TPR values, is strongly suggestive of a closer positive association between ModFOLDdock and oligo-lDDT ranking.

For hypothesis 3, then, with respect to multimer ranking by TM-score, there is insufficient evidence to reject the null hypothesis. *There is no difference between the independent QA and AF2 rankings as measured by the association between model rank categories.*
However, with respect to multimer ranking by oligo-lDDT, considering the increases in scores described above, there may be sufficient evidence to accept the alternative hypothesis, i.e., *Independent QA and observed score model rank are more closely associated than AF2 and observed score model ranks.*

5.4.4 Hypothesis 4. Is the accuracy of predicted scores affected by custom template recycling?

To answer this question data is presented from the four model populations which underwent custom template recycling. For monomers this is Population A2 (CASP15 round 2 monomers) and Population C (recycled monomers), for multimers it is Population B2 (CASP15 MultiFOLD group multimers) and Population D (recycled multimers). It would be logical to start with the data for populations A2 and B2 because these two groups can be directly compared to their unrecycled counterparts, i.e. Population A2, the CASP15 round 2 monomers (recycled) can be directly compared with the Population A1 CASP15 round 1 monomers (unrecycled) which were discussed in section 5.4.1.1 and Population B2, the MultiFOLD group multimers (recycled) can be directly compared to the Population B1 ColabFold group multimers (unrecycled) which were discussed in section 5.4.1.2. Populations C and D have no direct comparisons and so will be discussed last to provide support of the population A and B data.

5.4.4.1 Population A2 (CASP15 round 2 monomers)

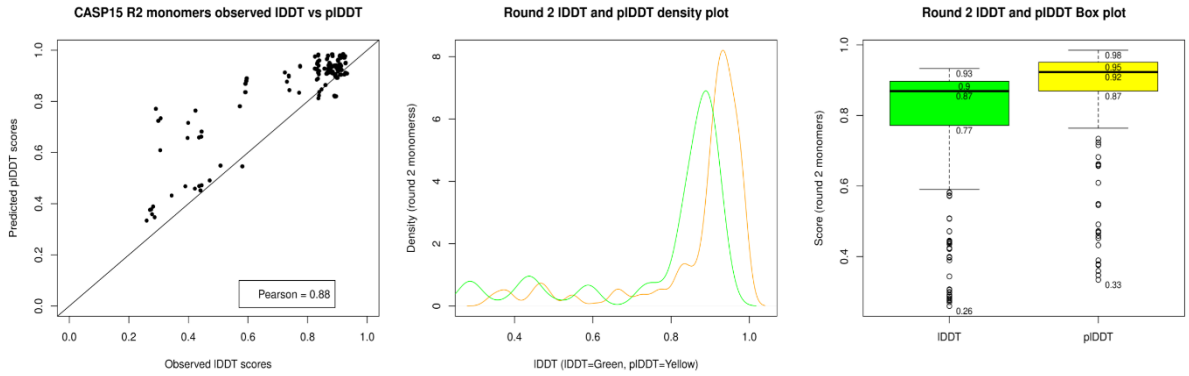


Figure 5.11. Plots for plDDT versus observed lDDT for Population A2 (CASP15 round 2 monomers). **Left.** A scatter plot. **Middle.** A density plot. **Right.** A boxplot. For all plots plDDT has been rescaled to fit the 0-1 lDDT range.

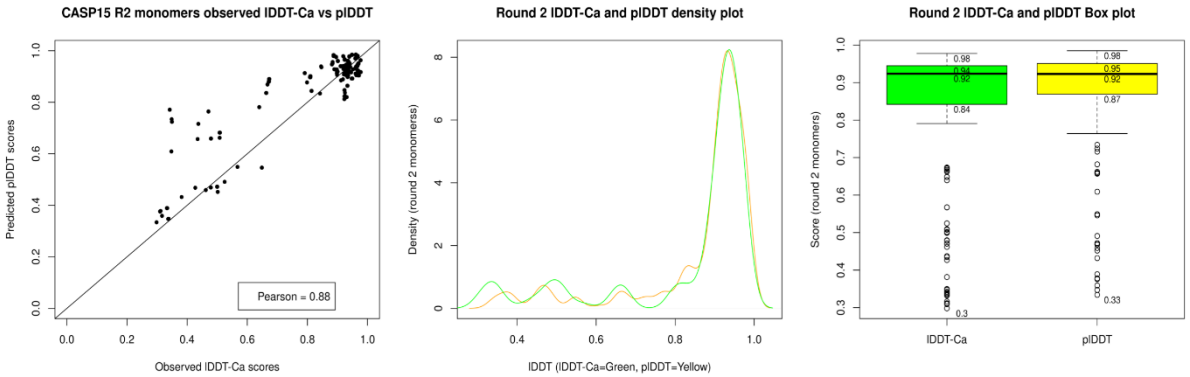


Figure 5.12. Plots for plDDT versus observed lDDT-Cα for Population A2 (CASP15 round 2 monomers). **Left.** A scatter plot. **Middle.** A density plot. **Right.** A boxplot. For all plots plDDT has been rescaled to fit the 0-1 lDDT range.

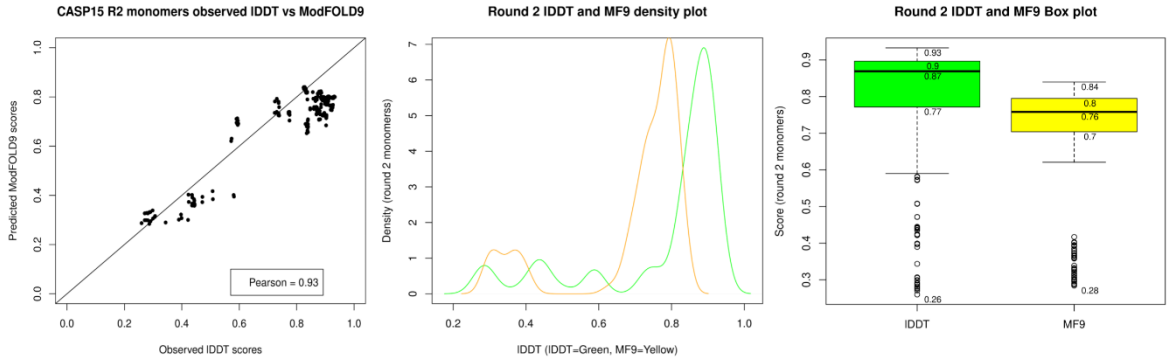


Figure 5.13. Equivalent plots of ModFOLD9 score versus observed lDDT for Population A2 (CASP15 round 2 monomers). Left. A scatter plot. **Middle.** A density plot. **Right.** A boxplot. For all plots plDDT has been rescaled to fit the 0-1 lDDT range.

Comparing the data from Figure 5.11 directly with that for the round 1 monomers presented in Figure 5.2 (section 5.4.1.1), it is apparent that there is a wider spread of data in the scatter plot in Figure 5.11 with an increase in plDDT scores, which are reflected in the density plot and the boxplot. Figure 5.12, for lDDT-Cα scores, shows a similar spread in the scatter plot but accompanied by a less noticeable difference between the plDDT and lDDT-Cα distributions in the density and boxplot. Wilcoxon signed rank tests for significance in Table 5.8 (below), however, reveal that the difference between the plDDT and lDDT-Cα score is significant as is the difference between the round 2 monomer plDDT scores and their round 1 counterparts.

Table 5.8. Calculated p-values from Wilcoxon signed tests for population A2, round 2 monomers. P-values ≤0.05 are in bold.

Row	*Scores compared*	*Independence and distribution symmetry*	*p-value*
1	*R2 plDDT versus lDDT-Cα*	Paired; 2-sided test	**0.0001**
2	*R2 plDDT versus lDDT-Cα*	Paired; 1-sided test, plDDT > lDDT-Cα	**5.83×10^{-5}**
3	*R2 plDDT versus R1 plDDT*	Unpaired; 2-sided	**1.293×10^{-9}**
4	*R2 plDDT versus R1 plDDT*	Unpaired; 1-sided, R2 > R1	**6.465×10^{-10}**
5	*R2 lDDT-Cα versus R1 lDDT-Cα*	Unpaired; 2-sided	0.1255

Wilcoxon test P-values were calculated at the 95% confidence level using and those ≤0.05 are considered significant.

Table 5.8, row 1, shows that according to a paired 2-sided Wilcoxon test there is a significant difference between plDDT and lDDT-Cα observed scores for round 2 monomers and, further to this, the results of a paired 1-sided test in row 2 show that that plDDT scores are significantly higher. These findings agree with the scatter plots in Figures 5.11 and 5.12 showing over-prediction in mid-quality models which was not present in the round 1 data. Notably, the over-prediction is also absent from the equivalent round 2 ModFOLD9 scatter plot shown in Figure 5.13. This is good evidence that overprediction of plDDT occurs in monomer models with custom template recycling.

To further test this, a 2-sided Wilcoxon test was used to directly compare round 1 and round 2 monomer plDDT scores (row 3) and this showed a significant difference between the two scores, evidenced by a p-value of 1.293×10^{-9}. Further, it was established that the round 2 monomer scores were significantly higher than those for round 1, evidenced by a p-value of 6.465×10^{-10} from the 1-sided test in row 4. Importantly, there was no such difference between the equivalent round 1 and 2 monomer observed lDDT-Cα scores as shown by the p-value of 0.1255 (row 5 of the table) meaning that round 1 and 2 monomer models were not significantly different in quality.

It is therefore reasonable to conclude that these prediction errors have been introduced by custom template recycling and, for hypothesis 4 in respect to monomer models, the alternative hypothesis can be accepted, i.e., *AF2 predicted scores following custom template modelling show greater variation than scores from regular modelling, when compared to equivalent observed scores.*

5.4.4.2 Population B2 (CASP15 MultiFOLD multimers)

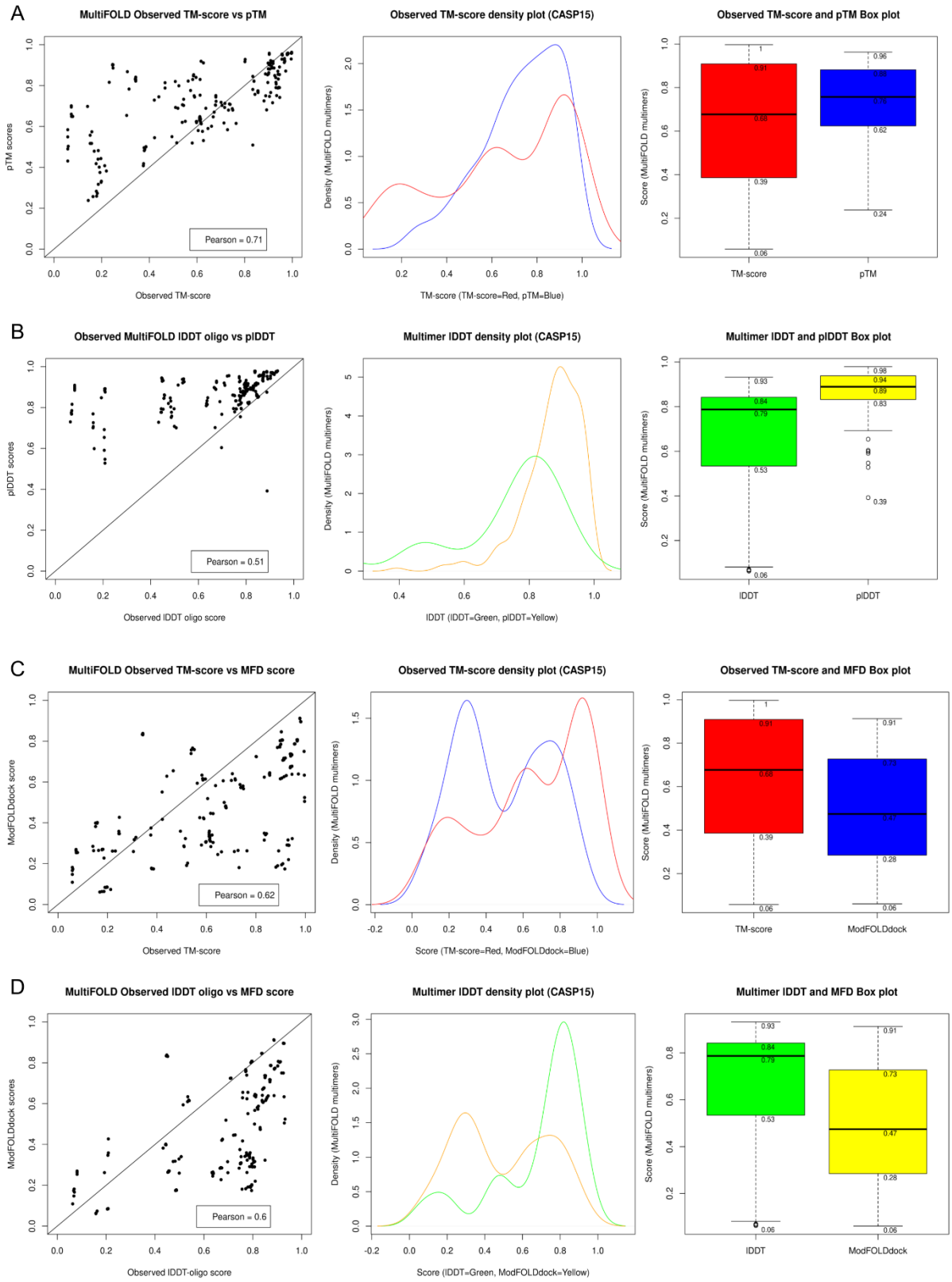


Figure 5.14. Plots for Population B2 (MultiFOLD multimers). Scatter plots (left), density plots (middle) and box plots (right) for; A. pTM versus observed TM-score. **B.** plDDT versus observed CASP oligo-lDDT. **C.** Comparison plots for ModFOLDdock score versus TM-score. **D.** Comparison plots for ModFOLDdock versus oligo-lDDT. plDDT figures are rescaled to 0-1.

The plots in Figure 5.14, panels A and B, can be directly compared to Figures 5.4 and 5.5 for ColabFold multimers in section 5.4.1.2. Considering the plots in panel A for TM-scores, the spread of points in the scatter plot is noticeably greater than that shown in Figure 5.4. Further, although the mean *observed* TM-score in the boxplots *reduces* from 0.745 (Figure 5.4) to 0.68 across the two populations, the equivalent mean pTM rises from 0.72 to 0.76. Secondly, considering panel B for lDDT scores in a similar way, the scatter plot again shows an increase in the spread of data compared to its equivalent in Figure 5.5 and there is also a marked shift to the right in plDDT when comparing the density plots, and a corresponding increase in mean plDDT score shown in the boxplot. These changes suggest a similar overprediction to that seen for monomers is also occurring for multimers which have been subject to custom template recycling. For comparison, the scatter plots in panels C and D showing ModFOLDdock scores versus both observed TM-score (C) and oligo-lDDT scores (D) for the same population, show little evidence of sustained overprediction. If anything, ModFOLDdock appears to suffer from a tendency for under-prediction of these models.

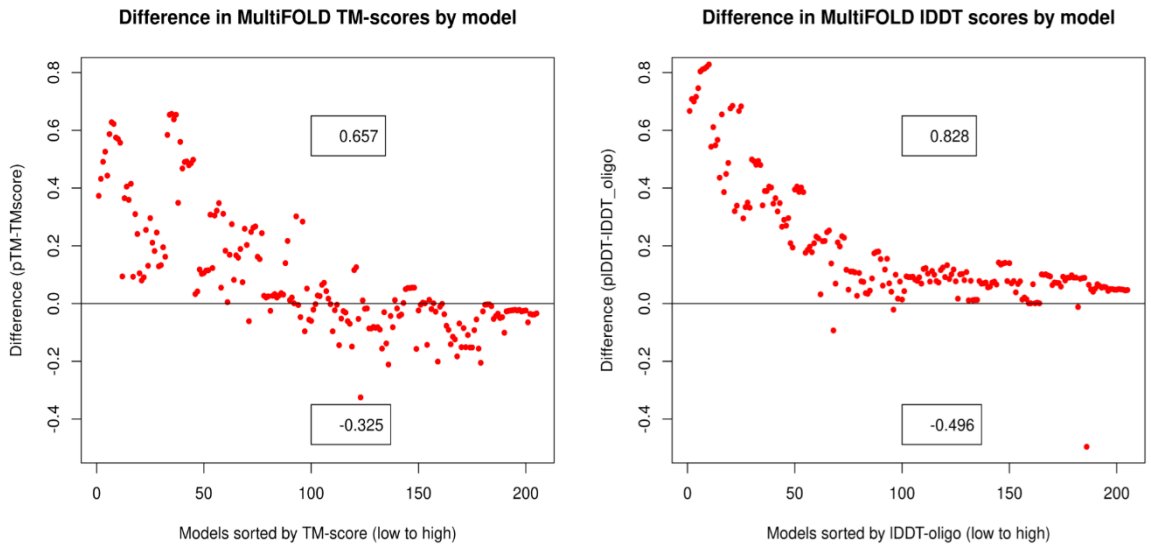


Figure 5.15. Plots to show variation between predicted and observed scores for Population B2 (MultiFOLD multimers). Left. pTM versus TM-score. **Right.** plDDT versus oligo-lDDT. Plots are equivalent to those in Figure 5.6 for ColabFold multimers. plDDT figures are rescaled to 0-1.

The relationships suggested in Figure 5.14, panels A and B, are more clearly shown by the two variation plots in Figure 5.15. In agreement with Figure 5.6, both plots show overprediction of scores for lower quality models with maximum and minimum differences of +0.657 and -0.325 respectively for pTM score and a maximum difference of +0.828 for plDDT score. Although the maximum and minimum deviation in the data for pTM score are almost identical to those from Figure 5.6, the maximum deviation in plDDT scores has increased from 0.747 to 0.828. Also, upon visual comparison of the two pairs of plots it is clear that the number of

models in the over-predicted regions in Figure 5.15 has increased over those in Figure 5.6 despite a similar number of models overall (205 and 190 respectively). Wilcoxon signed rank tests were again used to quantify these differences in terms of significance and the results are presented in Table 5.9 below.

Table 5.9. Wilcoxon tests for Population B2 MultiFOLD multimers and Population B1 ColabFold multimers. P-values ≤0.05 are in bold.

Row	*Scores compared*	*Independence and distribution symmetry*	*p-value*
1	*MultiFOLD plDDT versus oligo-lDDT*	Paired; 1-sided test, plDDT > oligo-lDDT	**2.20×10^{-16}**
2	*MultiFOLD pTM versus TM-score*	Paired; 1-sided test, pTM > TM-score	**1.46×10^{-5}**
3	*MultiFOLD versus ColabFold plDDT*	Unpaired; 1-sided, MultiFOLD > ColabFold	**7.193×10^{-8}**
4	*MultiFOLD versus ColabFold oligo-lDDT*	Unpaired; 2-sided.	0.283
5	*MultiFOLD versus ColabFold pTM*	Unpaired; 1-sided; MultiFOLD > ColabFold	**0.014**
6	*MultiFOLD versus ColabFold TM-score*	Unpaired; 2-sided.	0.252

Wilcoxon test P-values were calculated at the 95% confidence level using and those ≤0.05 are considered significant.

Table 5.9, rows 1 and 2 confirm that both predicted plDDT and pTM scores are significantly greater than their observed counterparts (oligo-lDDT and TM-score) for MultiFOLD multimers as evidenced by p-values of 2.20×10^{-16} for plDDT versus oligo-lDDT and 1.46×10^{-5} for pTM versus TM-score. Furthermore, there is confirmation that the plDDT (row 3) and pTM (row 5) scores are significantly greater than the equivalent predicted scores for ColabFold multimers but, importantly, there is no significant difference between the equivalent two sets of observed scores (row 4 for oligo-lDDT and row 6 for TM-score). This again shows that, for a similar set of models based on the same CASP targets, both sets of observed scores are similar but both predicted pTM and plDDT scores are significantly different and are higher in both cases for the group subject to custom template recycling.

Therefore, with respect to multimers, the alternative hypothesis must again be accepted, i.e., *AF2 predicted scores following custom template modelling show greater variation than scores from regular modelling, when compared to equivalent observed scores.*

5.4.4.3 Population C (recycled monomers).

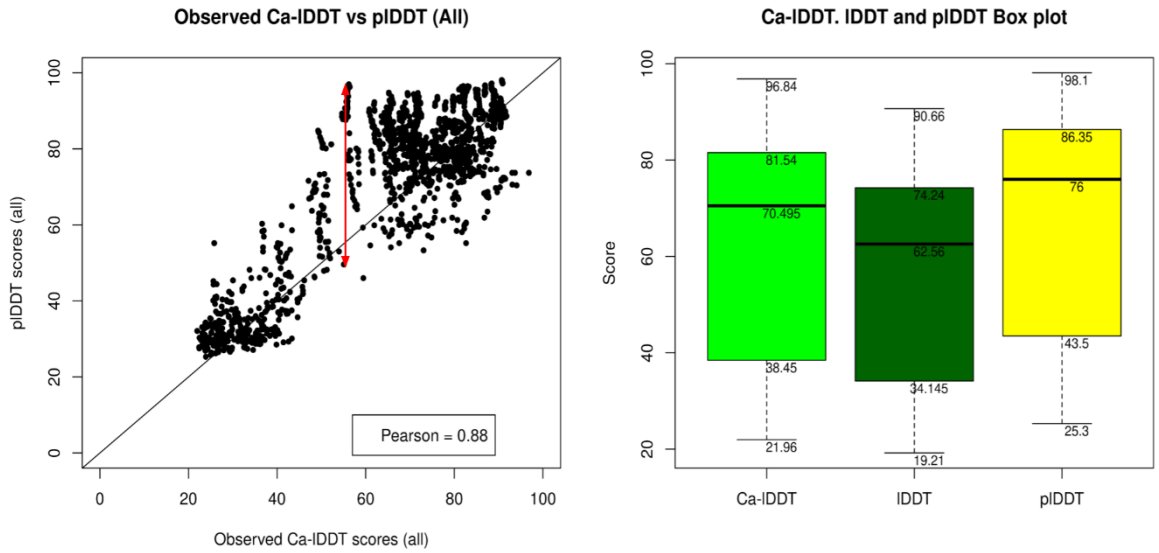


Figure 5.16. Plots for plDDT versus observed lDDT-Cα for population C (recycled monomers). Left. A scatter plot showing the spread of data. **Right.** A boxplot comparing the distribution of lDDT-Cα, lDDT and plDDT scores for the same population. lDDT and lDDT-Cα have been rescaled to the 0-100 range. Arrow (in red) on the scatter plot shows the potential degree of variation in predicted scores for models with similar observed scores.

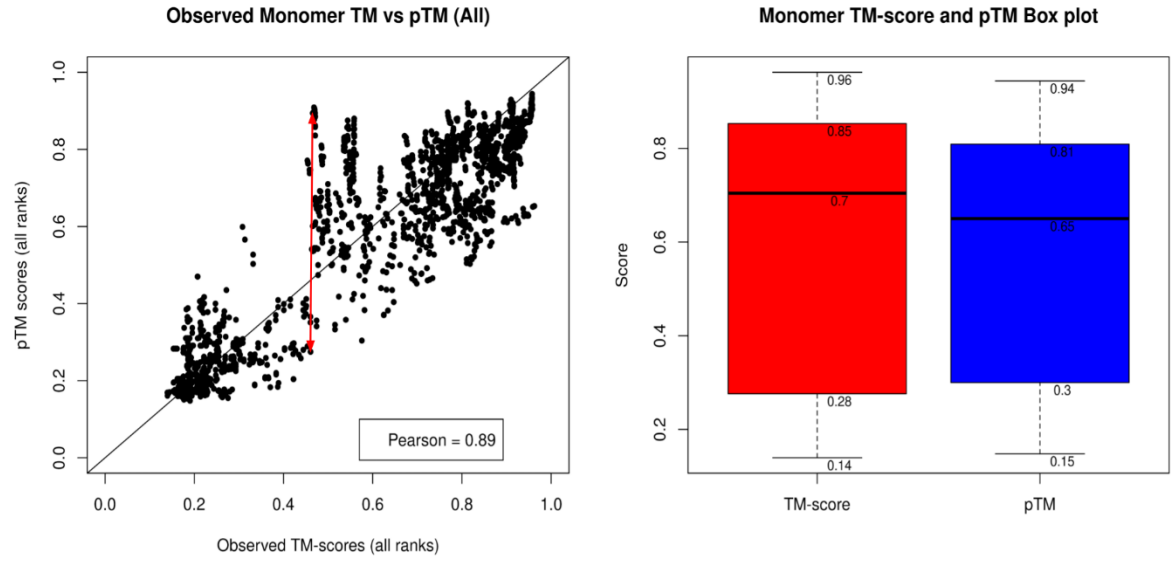


Figure 5.17. Plots for pTM versus observed TM-score for population C (recycled monomers). Left. A scatter plot showing the spread of data. **Right.** A boxplot for both scores from the same population. Arrow (in red) on the scatter plot shows the potential degree of variation in predicted scores for models with similar observed scores.

Finally, the purely recycled models (population C and D) are considered. The scatter plots in Figures 5.16 and 5.17 show Pearson correlation coefficients of 0.87 and 0.89 respectively between predicted and observed scores. Although these correlations appear very respectable, both plots show a pronounced spread in the data with a high proportion of outliers. The red

bars on each scatter plot show the potential degree of variation in predicted scores for models with similar observed scores. For an observed score of approximately 0.5, predicted plDDT scores range from approximately 0.5 to 0.9 (Figure 5.16) and pTM scores range from approximately 0.3 to 0.9 (Figure 5.17). These results strongly support the hypothesis that using custom template recycling appears to produce a much higher degree of variability both plDDT and pTM scores.

5.4.4.4 Population D (recycled multimers).

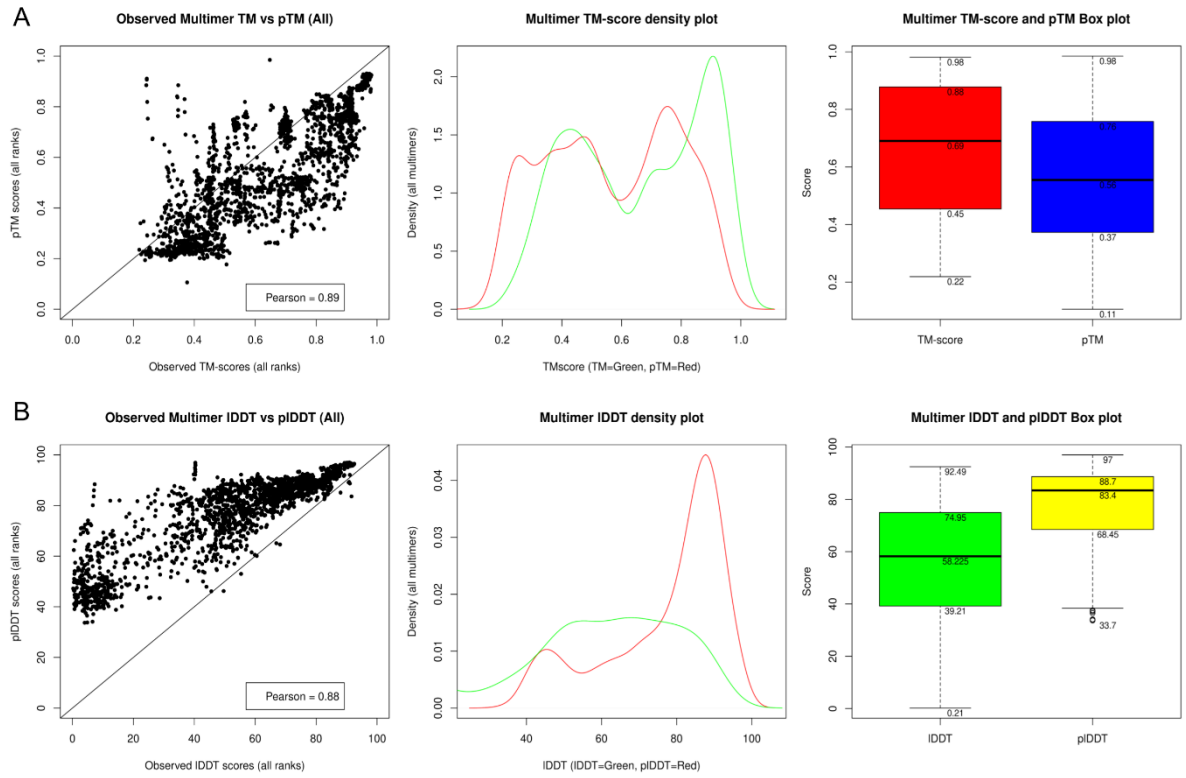


Figure 5.18. Plots for Population D (Recycled multimers). Scatter plots (left), density plots (middle) and box plots (right). A. pTM versus observed TM-score. **B.** plDDT versus observed CASP oligo-lDDT. plDDT values are rescaled to 0-1.

Figure 5.18 shows a similar spread of data to that seen in Figures 5.16 and 5.17. Panel A again shows a tendency for multimer pTM over and under-prediction meaning a high variation in predicted pTM score for models with similar observed scores. In panel B, all three plots demonstrate a high tendency for plDDT over-prediction and again, this is more pronounced for mid to lower quality models.

As both population C and D were subject to up to 12 recycles and were entirely created via custom template recycling, these results support the hypotheses drawn above for population A and B, that using custom template recycling produces a higher degree of variability in AlphaFold2 predicted scores for both monomer and multimer models and that this effect is more pronounced for multimers.

5.5 Conclusions

Throughout, data in this chapter has been orientated toward answering four primary questions concerning the accuracy of the often-quoted AlphaFold2 predicted scores plDDT and pTM, both as empirical descriptors of model quality and as reliable ranking scores. Further to this, there remains the more challenging secondary consideration of whether the AlphaFold2 neural network has learnt a useful energy function which can be applied to extend its use to general model quality assessment.

plDDT is a reliable indicator of tertiary structure (monomer) model quality and ranking.
From the data presented in section 5.4.1.1, plDDT was shown to be a reliable indicator of tertiary structure model quality when straightforward regular modelling was used and showed impressive Pearson correlation coefficients with both observed lDDT and lDDT-Cα scores which the independent QA method, ModFOLD9, was unable to improve upon. plDDT prediction accuracy appeared to be maintained across the scoring range and any over-prediction may be potentially explained by the published linear relationship with lDDT-Cα.

Ranking of the same tertiary model population also showed an agreement between plDDT and lDDT-Cα assigned ranks, which ModFOLD9 was, again, unable to improve. For straightforward regular tertiary modelling and it can be concluded that plDDT appears to be a reliable quality descriptor and ranking tool for tertiary structure models.

Both pTM and plDDT show variability as indicators of multimer model quality and ranking.
The reliability, however, was not maintained for all multimers. As shown by the plots in section 5.4.1.2, both pTM and plDDT showed variability for models of similar quality with pTM showing a tendency for underestimation for higher quality models and both scores showing overestimation for some lower quality models. The overestimation was more pronounced for plDDT.

This variability also affected ranking accuracy, with both pTM and plDDT showing a lower association with observed score ranking than was seen for monomers. Of the two scores the association was less strong for plDDT-assigned ranks. ModFOLDdock, which did not show over-prediction to the same degree, was able to improve upon the rank agreement of plDDT although there was insufficient evidence to draw the same conclusion for pTM. Nevertheless, there remains some unreliability in the ability of pTM and plDDT to differentiate between some high and low quality multimer models created by regular modelling and ModFOLDdock scores represent a more reliable method for ranking multimer models.

Greater variability of AF2 predicted scores is seen if custom template recycling is used. Finally, convincing evidence is presented in section 5.4.4 that using the custom template option to recycle models through the AlphaFold2 algorithm results in a much greater variability in predicted scores for both tertiary structures and multimers and that the variability is more extreme for multimer models. This provides cautionary evidence that the use of AF2 and AF2-Multimer outside of their intended end-to-end operation could result in mis-scoring and mis-ranking of models.

Independent MQA is essential but AF2 is an unlikely MQA program in its current form. In light of these results, while recycling custom templates through AlphaFold2 improves model quality (Adiyaman *et al.*, 2023) the accuracy of the accompanying predicted scores will be severely affected in some cases. For this reason, it is considered unlikely that AlphaFold2 would be a useful tool for accurate quality assessment (QA) of whole models as the only way to achieve this would be via the custom template route.

Further to this, where the custom template option is used for tertiary structure modelling, it is essential that an independent QA program such as ModFOLD9 is used to ensure accuracy in predicted scoring and model ranking. For any multimer modelling, whether straightforward or via custom template recycling, an independent QA program such as ModFOLDdock should also be used for the same reasons. MQA programs not only offer an alternative opinion on quality but also enable models from different software to be objectively compared.

CHAPTER 6

Synthesis, conclusion and next directions

6.1 Synopsis of studies

This thesis describes a body of work completed over a 5 year period (2018–23) with two main aims; one, to identify methods for the improvement of predicted protein quaternary structure modelling over that achievable by docking technology; two, to develop a method for accurate and independent quaternary structure predictive model quality assessment, a technology that was largely missing from the modelling toolkit at the time. A tacit third aim was the symbiosis of these two developments to drive improvement in quaternary structure model quality.

6.1.1 Analysis of MultiFOLD performance and incorporating AF2 recycling

An extensive analysis of quaternary structure modelling performance at CASP13 was carried out as a baseline for development. In this analysis, and based on a similar analysis by the Venclovas group (Dapkunas *et al*., 2019) successful modelling was defined as models having a QS-score > 0.1. It was found that, even with this fairly low threshold, the early hybrid docking/TBM version of the MultiFOLD pipeline achieved only 10% success rate (3/30 models).

Following the success of AF2 at CASP14 and the subsequent code release by the DeepMind group leading to the development of ColabFold, we were able to explore the possibility of model refinement and improvement via the custom template recycling option with the intention that this could provide a unique advantage in an updated version of MultiFOLD.

It was established that this novel use of recycling using full structural coordinate files was possible, and that it significantly improved models beyond their starting quality as measured by comparison with their native structures using the lDDT score. This improvement was not seen in a parallel MD refinement study (Adiyaman *et al.*, 2023). Further to this it was shown that a significant improvement in model quality was achievable without the need for an MSA and that official DeepMind AlphaFold2 models were also slightly but significantly improved by this process. Success in this initial study underpinned the application of this technique to quaternary structure modelling where similar improvements were seen, including improvements measured by QS-score, suggesting improvements to multimer interfaces.

Documented evidence (Roney and Ovchinnikov, 2022) showed that AF2 performance decreases considerably without an MSA, meaning that the improvements we obtained provide some evidence for either a learnt protein folding function in the AF2 DNN or that template information can be used to avoid local minima within the folding funnel energy landscape. Benchmarking at CASP15 resulted in MultiFOLD outperforming both the naïve NBIS-AF2-Multimer and the ColabFold groups (Burcu Ozden *et al.*, 2023) showing that the unique combination of AF2 features with a blend of existing and proprietary EMA scores added value beyond the baseline modelling capabilities of AlphaFold2-Multimer.

6.1.2 Developing new quality estimates and optimisation of artificial Neural Network (NN) correlations for CASP15

The main aim of this part of the study was the development of an independent, publicly available model quality assessment program (MQAP) to predict the quality of quaternary structure models. Although many MQAP options existed for tertiary structures, very few resources existed for the quality comparison of multimer models built using different modelling software. If the life sciences community was to accept multimeric models, a reliable method for predicting model quality was vital. To achieve this, the first part of the study was dedicated to finding a route for the improvement of the unpublished MQAP ModFOLDdock, which had been used during the CASP13 assembly modelling competition but which had shown inaccuracies compared with observed scores of up to 0.546 (0-1 scale) and had a success rate in selecting the best model from a decoy set of 1/30 or 3.3%.

Initial regression analysis was performed with observed scores and this found that there was useful information contained within the six ModFOLDdock predicted scores which had been previously masked by the calculation of an overall consensus score. A novel use of the CASP assessor scores as Local, Global and Total target scores revealed improved correlations. These scores were then used as target scores to train a simple MLP by supervised learning using three-part cross-validation. The resulting unique combination of the six distance-based scores was designed to differentiate between models on the basis of global fold, interface quality and overall similarity.

When ModFOLDdock was optimised for CASP15 it was found that the Local and Global scores defined previously fitted the definitions of the QMODE2 SCORE (global fold) and QSCORE (global interface) categories specified for the new assembly EMA competition. During further optimisation three variants of ModFOLDdock were defined to produce quality scores according to user requirements, these were; ModFOLDdock – optimised for correlation with observed scores and likely to provide a good estimate of empirical quality; ModFOLDdockR – optimised for ranking, more likely to differentiate between decoy models for top model selection, and ModFOLDdockS – designed as a quasi-single model method to allow reliable quality assessment of a single or only few models.

ModFOLDdock variants achieved a 100% prediction rate (i.e., they generated scores for all targets) in all three CASP15 EMA categories and were ranked in 2nd place for (global) SCORE (ModFOLDdock), 1st place for (global interface) QSCORE (ModFOLDdockR) and 2nd place for local interface residue score (ModFOLDdockR). Additionally, ModFOLDdock variants rated highly when identifying the interface patch in antibody-antigen interactions. Overall, ModFOLDdock variants improved the Pearson correlation between predicted and CASP

observed scores from 0.16 seen in CASP13 to a maximum of 0.81 when all CASP15 data was considered. Further to this, the increase in modelling success described in 6.1.1 was partially attributable to model selection by ModFOLDdockR and provides good evidence of model quality assessment improving modelling quality.

6.1.3 Comparison of AF2 accuracy estimates with ModFOLDdock MQA scores

As the life sciences community becomes more used to AlphaFold modelling, there is likely to be more reliance on the AF2 predicted quality measures, which have shown great reliability for tertiary structure models. It had yet to be established if this reliability extends to quaternary structure modelling.

In this part of the study, the AlphaFold2 predicted quality measures, plDDT and pTM were investigated as reliable descriptors of both model quality and as ranking measurements for both tertiary and quaternary structure models. Their performance was compared to both ModFOLD9 for tertiary structures and ModFOLDdock for quaternary structures.

To the best of our knowledge, this work showed for the first time the pattern of over and under estimation of these scores for quaternary structures, as applied by AF2-Multimer. It also shows that the variation is exacerbated by the use of custom template recycling, suggesting that that AFM quality scores may be a product of MSA strength and when the custom template option is used, the accuracy is somewhat overwritten. It was demonstrated that ModFOLDdock, as an independent MQA method, did not show the same pattern of variation with changing modelling conditions and statistical analysis suggested that ModFOLDdock represented a more accurate ranking tool than either plDDT or pTM for multimeric models.

6.2 Conclusions

6.2.1 Quaternary structure modelling

The recycling experiment showed that significant improvements, as measured by native-dependent quality scores, can be made to both tertiary and quaternary structure models at a low computational cost by this process. In some cases, non-AF2 tertiary structure models were improved beyond the quality of the equivalent AF2 model, and quaternary structure model improvement was evident from increases in both TM-score and QS-score. Furthermore, improvement was significant in the absence of an MSA, even for some high quality models and, considering documented evidence for a decrease in AF2 modelling quality without an MSA (Roney and Ovchinnikov, 2022), these improvements are suggestive of a learned function within the AF2 neural network. Recycling was subsequently included in the MultiFOLD pipeline and blind benchmarking at CASP15 showed that MultiFOLD out-performed the baseline NBIS-AF2-Multimer and the ColabFold group in assembly modelling.

A parallel study using the same dataset as the recycling experiment showed that the molecular dynamics refinement program ReFOLD4 (Adiyaman *et al.*, 2023), improved the geometry-based MolProbity scores (Chen *et al.*, 2010) rather than native-dependent scores. This means that creating models via MultiFOLD, which includes recycling, followed by further refinement with ReFOLD4 could improve both the backbone and atomic positioning within 3D models with a low computational overhead, considering ReFOLD4's targeted constraint approach which will not attempt to refine residues with high plDDT scores. This could be important in moving closer to models accurate enough for medical or drug interaction studies, indeed Section 6.2.4 includes references to some studies where MultiFOLD has been chosen specifically for these advantages.

6.2.2 Quaternary model quality assessment

The identification of novel target scores for the combination of ModFOLDdock predicted quality measures and the comparison of these to CASP assessor scores led to prediction accuracy increases of 8.75%, 14% and 7% for Local, Global and Total scores respectively. The relationships defined by this process and confirmed by supervised NN training were then used in defining and optimising three ModFOLDdock variants for the CASP15 EMA competition as well as for an additional ranking tool within the MultiFOLD pipeline.

ModFOLDdock variants were highly placed in all three EMA categories making ModFOLDdock arguably the most successful EMA method at CASP15. ModFOLDdockR additionally demonstrated its ranking ability within the MultiFOLD modelling pipeline by identifying the best model from a decoy population 33% of the time, a 10-fold increase from CASP13 performance. Additionally, ModFOLDdock variants showed especially good interface patch identification for antibody-antigen binding interactions which may be an important aspect to advertise to the biological modelling community. Finally, it was demonstrated that ModFOLDdock was able to outperform plDDT and pTM as a ranking tool for AFM models as it did not show variability or a tendency for overprediction in the same way as the AFM quality measures.

These achievements show that correctly optimised distance-based scoring algorithms can compete with machine learning (ML) systems which can suffer from accuracy issues when there is a lack of high-quality models for training datasets. Until multimer modelling reaches the levels of quality associated with AF2 tertiary models, traditional quality measures could play an important role in driving the development of quaternary structure modelling in the short to medium term. The final point on ModFOLDdock versus plDDT and pTM scores provides evidence that adapting AF2 and AFM for use outside of their end-to-end operation could result in model mis-scoring and mis-ranking.

6.2.3 Continued benchmarking of MultiFOLD

The CASP15 success achieved in 2022 has been further validated by encouraging results for the MultiFOLD server in the ongoing CAMEO BETA modelling of structures and complexes community project (Haas *et al.*, 2019). The data presented in Figure 6.1 show that MultiFOLD (Server 1) outperforms the other three currently registered servers as measured by both lDDT-oligo and QS-score. The values plotted are cumulative scores normalised to the 0-1 range to compensate for the different number of targets modelled by each server. Servers 1 and 76 modelled 127 targets, Server 4, 80 targets and Server 2 only 40 targets, therefore normalised values were created by calculating cumulative Server 1 scores for sets of targets matching each other server. The normalised score is then calculated as

$$\frac{Score}{Server\ 1\ score}$$

Further analysis (Genc, 2024) has shown that the MultiFOLD server maintains its performance advantage over the other servers and, in all except the homomer category, the advantage is significant at the 95% confidence level. This slight performance dip for homomers is due to the difficulty with stoichiometry determination which is unclear from the single sequence provided for homomers versus the three which would be provided for a trimer, for example. Homomer stoichiometry must therefore be inferred from templates which can be problematic for previously uncharacterised proteins.

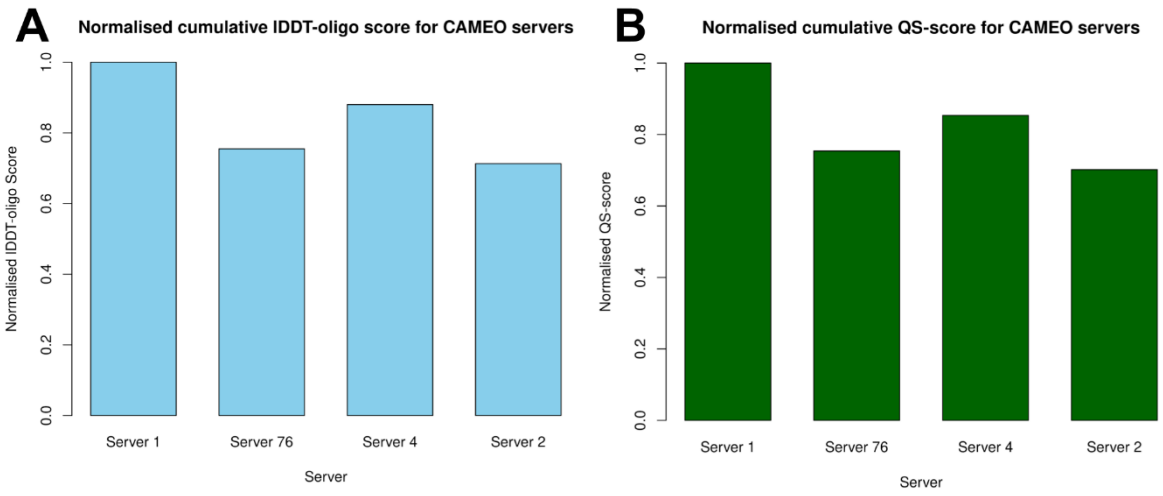


Figure 6.1 Relative performance of MultiFOLD (Server 1) and the other servers competing in CAMEO BETA modelling. Data is for combined heteromer and homomer models. **A.** lDDT-oligo scores normalised as described above for matching target populations modelled by each server (common subsets). **B.** Similarly normalised data for QS-scores. Data was collected between January 2023 and March 2024 and kindly provided by Ahmet G Genc (Genc, 2024). Server identities are hidden to all except CAMEO organisers.

6.2.4 Impact of the MultiFOLD and ModFOLDdock servers

A number of groups have published papers citing both MultiFOLD and ModFOLDdock as integral parts of their research. Brief descriptions of four example studies are given below.

1. Diverse genetic contexts of HicA toxin domains propose a role in anti-phage defense, (Gerdes, 2024). In this study dimers of the PaV-LD phage class 1 HicAB and the *Campylobacter* class 2 HicAB were modelled using MultiFOLD and quality assessed using ModFOLDdock. This was a bioinformatics examination of the role of the HicA domain in the toxin–antitoxin (TA) system as an anti-phage defence mechanism. The elucidation of the interaction was described as advancing the understanding of the TA system functionality within the microbial world.

2. Disabling spidroin N-terminal homologs' reverse reaction unveils why its intermolecular disulfide bonds have not evolved for 380 million years, (Mi *et al.*, 2023). This study cited MultiFOLD's independently validated improved performance over AlphaFold2 and used the server to predict the NT and CT self-assembly spidroin domains.

3. Are the integrin binding motifs within SARS CoV-2 spike protein and MHC class II alleles playing the key role in COVID-19? (Gerencer and McGuffin, 2023). This study used MultiFOLD models to predict the visibility of the integrin-binding ECD (Glu-Cys-Asp) and LDI (Leu-Asp-Ile) motifs on the S (spike) protein.

4. In Silico Evaluation, Phylogenetic Analysis, and Structural Modeling of the Class II Hydrophobin Family from Different Fungal Phytopathogens, (Bouqellah and Farag, 2023). This study used MultiFOLD as well as AF2 and trRosetta to model HFBII structures and found that "MultiFOLD showed a higher modelling precision than the other [Alphafold2 and trRossetta] tools, by pTM and plDDT". This was verified by observed TM-scores of 7.1 (MultiFOLD), 0.69 (AF2) and 0.62 (trRosetta) when compared to the experimental HFBII structure (PDB: 4AOG).

Of these, the HicAB (1) and the SARS-CoV-2 binding motifs (3) study exemplify how understanding protein binding and complex formation can be integral to the furtherance of biomedical research. The HicA proteins are small bacterial proteins with a domain responsible for their toxic activity. Production of HicA is often increased at times of bacterial stress, inhibiting cellular function and leading to dormancy. This is thought to be a potential route for antibiotic resistance allowing the bacteria to lie dormant until levels of antibiotic are reduced. The HicB proteins bind to the HicA domain and prevent its exposure. Understanding the dimerisation could lead to the development of drug-induced HicA binding, preventing dormancy and therefore reducing resistance.

Both the ECD and LDI motifs lie within the SARS-CoV-2 receptor binding domain (RBD) and have been implicated in integrin (ECD) and angiotensin-converting enzyme 2 (ACE2) binding (LDI). Studying the interactions of these domains may enhance understanding of the ability of the SARS-CoV-2 virus to infect a diverse range of cells causing the severe viral loads which were seen in some cases during the pandemic. The ECD is of particular interest as integrin can be activated via cytokinin release (Liu *et al.*, 2022) leading to enhanced integrin-mediated cell entry following initial ACE2-mediated entry. A vaccine developed specifically against this domain may limit the ability of Covid to cause serious disease.

The publications resulting from the work in this thesis have been useful to the general community. According to a Google Scholar search on 11/4/24, Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers (McGuffin *et al.*, 2023) has received 21 citations; Estimation of model accuracy in CASP15 using the ModFOLDdock server (Edmunds *et al.*, 2023) has received 10 citations and Improvement of protein tertiary and quaternary structure predictions using the ReFOLD refinement method and the AlphaFold2 recycling process (Adiyaman *et al.*, 2023) has received 7 citations.

6.3 Future directions

6.3.1 Short term developments

The first and most pressing task is the development of version 2 of ModFOLDdock to further improve performance and maintain competitiveness at CASP16. To achieve this, a modified version of the neural network which was proposed in Chapter 3 will be integrated into the source code with the function of optimising an enhanced set of component scores into a consensus score. At the time of writing this is currently in the pretraining test stage for the residue-level confidence score, a representation of which is shown in Figure 6.2. As suggested in Chapter 3, one way to increase the predictive power of a neural network while avoiding the overfitting problem is to increase the number of inputs available for consideration. The logic adopted for ModFOLDdock version 2 is to firstly identify interface residues as those within 8Å of a residue, in the hypothetical example below this is residue “A11”. The five closest neighbouring residues are then calculated by the shortest Euclidean distance from the target residue. In Figure 6.2 this is A10 and A15 on the same chain and B11, B6 and B25 on the complementary chain, as shown by the yellow and cyan graphic on top the left. The values (0-1) for eight quality scores for each of the residues identified by the measures above are then used as input to the neural network, making a total of 48 input scores per residue. From these an optimal consensus score describing the modelling accuracy of each interface residue is calculated. The structural image top right shows a target residue (yellow) in the middle of an interface, the bottom right image shows a target residue (again in yellow) on the edge of an

interface, in cases where edge residues have less than five contacting residues (within 8Å) one or more of the scores will be set to 0 for padding.

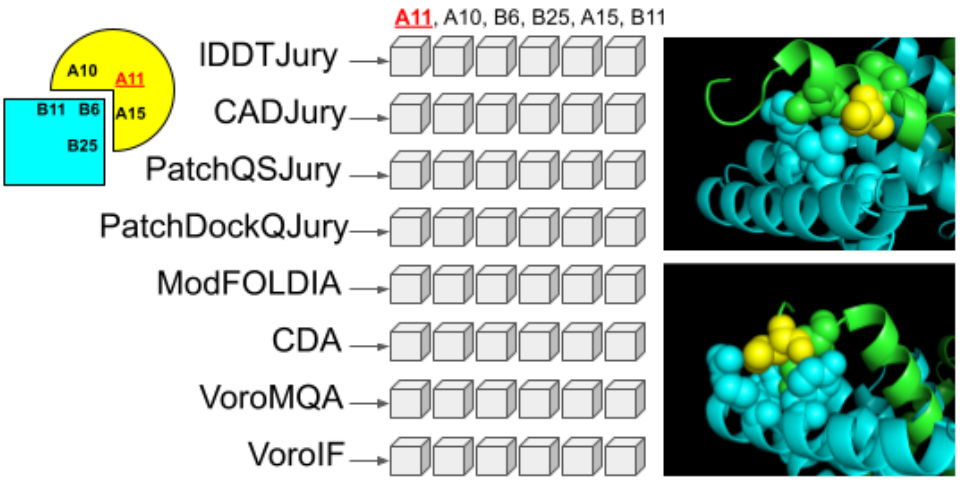


Figure 6.2 The proposed format for the version 2 ModFOLDdock MLP used to calculate optimal residue level confidence scores. Each residue, identified in the left-hand circular graphic and the right-hand structural image, is scored on the basis of eight different scoring methods for itself and the closest five residues, as measured by Euclidean distance.

The CASP15 results, documented in Chapter 4, highlighted that ModFOLDdock had a particular affinity for interface patch identification, which was particularly strong for antibody-antigen binding interactions. Understanding what characterises a protein segment as a potential antibody interaction patch could be important in the design of vaccines or autoimmune treatment (Guarra and Colombo, 2023). This affinity could be specifically explored and developed by extended testing on antibody-antigen targets to determine the power of patch detection.

The use of DNNs in MQA programs has been shown to increase performance over previous version of the program, for example VoroIF out-performed VoroMQA in testing (Olechnovic and Venclovas, 2023). The idea underpinning the Bonvin group's DeepRank was to use the flexible programming language PyTorch to create a trainable NN which could be used out of the box or retrained to the users' specifications. The aim was to predict the quality of protein multimers on the basis of the similarity of their interfaces with experimentally derived proteins. In 2021, when this was investigated for possible integration into the ModFOLDdock pipeline, there appeared to be difficulties with interpretation of the output of the program, however continued development and a greater time period to fully investigate the flexibility of the NN may make this a viable direction. This is possibly preferable to suggestions to adapt the AF2 NN for this purpose (Roney and Ovchinnikov, 2022) as evidence from Chapter 5 suggests a perturbation in quality score reliability when AF2 is used in this way.

Also of value would be further investigation into the score profile of the two AF2 model accuracy measures plDDT and pTM. It was established in Chapter 5 that there is score overprediction for some low and medium quality models, a situation which becomes more significant if custom templates are used during modelling. It has been proposed that the AF2 NN has learned a protein folding function, but this is clearly incomplete or inaccurate in some aspects – as AF2 continues to rely on MSA data to build accurate models. Understanding exactly which models are prone to overprediction and which conditions exacerbate overprediction may help to uncover some of the inaccuracies in this proposed folding function, presenting an opportunity for targeted improvement.

6.3.2 Longer term developments

For MultiFOLD modelling there are two potential directions to improve model quality. These were highlighted in the analysis of the AFsample method (Wallner, 2023) during CASP15 which suggested that where the evolutionary signal from an MSA is weak, improvements to model quality can be made by either augmenting the MSA or performing neural network dropout to increase the diversity of models sampled. There is evidence for and against adding a custom, paired MSA to AlphaFold2. Work on AF2 PPI prediction (Bryant *et al.*, 2022) found that this significantly improved model quality while similar work on AlphaFastPPI (Yin *et al.*, 2022) suggested that pairing was not important. Despite this disagreement, as MultiFOLD runs both AF2 and AFM in tandem, it would likely be worthwhile investigating the effect of a paired MSA on the AF2 arm of the pipeline as well as the effect of MSAs constructed on structural similarity (rather than sequence) on the AFM arm.

The Wallner group were able to program a dropout rate into the AF2 neural network for their AF2sample pipeline, meaning that some of the weights in the network were randomly set to zero. During training, this is often employed to prevent overfitting and allows the network to learn different solutions to the same problem by sampling a greater diversity of models. According to the CASP15 results page, AF2sample was officially ranked in third place for multimeric modelling and so this approach has proven efficacy and it would be a viable research method for inclusion into the MultiFOLD pipeline.

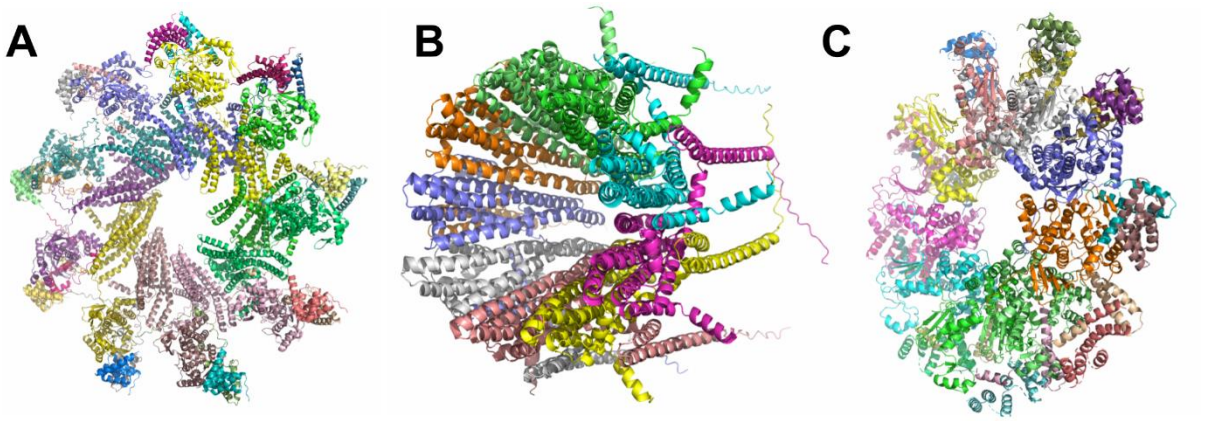


Figure 6.3 Two proposed structures for CASP15 target H1111. A. The McGuffin model as a cyclic nonomer of ABC trimers. **B.** The same target as a cyclic nonomer of AB chains with a chain C transmembrane tail. **C.** The CASP nonomer reference structure. Structures are coloured by chain.

One frustrating problem which arose during both CASP14 and CASP15 modelling, and which applies equally to MQA, is that of stoichiometry or symmetry. For homomers, it is not always obvious whether individual chains form a dimer or higher association and for heteromers, if the stoichiometry is known, it is not always clear how the different chains repeat and fit together. While the former problem is addressed by the new stoichiometry prediction protein language model QUEEN (QUaternary state prediction using dEEp learning) (Avraham *et al.*, 2023) with some encouraging results, a good example of the latter problem is the heteromeric CASP15 target H1111 with a A9B9C9 stoichiometry, shown in Figure 6.3. It was not clear whether this would result in a circular structure of all three chains in one plane (a polo style shape), represented by panel A or whether two chains formed the circular pore with the third forming a trans-membrane tail section, represented by panel B. The CASP native structure, in panel C, shows that the former idea was closer to the truth. In addition, and as described in the ReFOLD4 refinement work (Adiyaman *et al.*, 2023), research is increasingly considering a protein conformational landscape as a more important concept than a single correct or incorrect model, a concept first proposed by Alexei Kurakin (Kurakin, 2009).

To address these issues of arrangement and flexibility, a pragmatic approach would be to continue with small percentage gains in modelling, producing an ever-improving population of quaternary structure models. These in-turn will act as an improved quality training dataset for ML approaches to both modelling and MQA, leading to gradual improvement in ML learning and the development of a true fold and dock approach. The neural network dropout approach described above may have a role in addressing the flexibility issue specifically. If dropout produces an increased variety of models for a neural network to sample, it may be that these intermediates actually represent different but valid conformations of the model. Instead of allowing the network to simply assess these during the creation of a single final model, it may be useful to output these alongside the final model as a representation of the conformational

landscape of the protein. This may lead to deeper understanding of the flexibility inherent in certain structures as well as creating an increased diversity of models for subsequent training datasets.

References

Adiyaman, R. 2021. *Improvement of MD-Based Protocols for the Refinement of 3D Protein Models.* PhD, Reading.

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. *Bioinform Adv,* 3**,** vbad078.

Adiyaman, R. & Mcguffin, L. J. 2019. Methods for the Refinement of Protein Structure 3D Models. *Int J Mol Sci,* 20.

Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J. & Al-Khabori, M. 2021. Proteomics: Concepts and applications in human medicine. *World J Biol Chem,* 12**,** 57-69.

Alsaadi, E. & Jones, I. 2019. Membrane binding proteins of coronaviruses. *Future Virology,* 14.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

Anfinsen, C. & Scheraga, H. 1975. Experimental and Theoretical Aspects of Protein Folding. *In:* C.B. ANFINSEN, J. T. E., FREDERIC M. RICHARDS, (ed.) *Advances in Protein Chemistry.* Academic Press.

Arun, K. S., Huang, T. S. & Blostein, S. D. 1987. Least-squares fitting of two 3-D point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*

Ashmore, J., Carragher, B., Rosenthal, P. B. & Weis, W. 2021. A resolution record for cryoEM. *Fac Rev,* 10**,** 64.

Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L. & Schueler-Furman, O. 2023. Protein language models can capture protein quaternary state. *BMC Bioinformatics,* 24**,** 433.

Baek, M., Anishchenko, I., Humphreys, I., Cong, Q., Baker, D. & Dimaio, F. 2023. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv*.

Baek, M., Anishchenko, I., Park, H., Humphreys, I. R. & Baker, D. 2021a. Protein oligomer modeling guided by predicted interchain contacts in CASP14. *Proteins,* 89**,** 1824-1833.

Baek, M., Dimaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millan, C., Park, H., Adams, C., Glassman, C. R., Degiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. 2021b. Accurate prediction of protein structures and interactions using a three-track neural network. *Science,* 373**,** 871-876.

Bansal, P., Kumar, R., Singh, J. & Dhanda, S. 2021. In silico molecular docking of SARS-CoV-2 surface proteins with microbial non-ribosomal peptides: identification of potential drugs. *J Proteins Proteom,* 12**,** 177-184.

Barozet, A., Chacon, P. & Cortes, J. 2021. Current approaches to flexible loop modeling. *Curr Res Struct Biol,* 3**,** 187-191.

Basu, S. & Wallner, B. 2016a. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One,* 11**,** e0161879.

Basu, S. & Wallner, B. 2016b. Finding correct protein-protein docking models using ProQDock. *Bioinformatics,* 32**,** i262-i270.

Benjin, X. & Ling, L. 2020. Developments, applications, and prospects of cryo-electron microscopy. *Protein Sci,* 29**,** 872-882.

Bergmeir, C. & Benitez, J. 2012. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software,* 46.

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. *Sci Rep,* 7**,** 10480.

Bhattacharya, D. & Cheng, J. 2013. 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins,* 81**,** 119-31.

Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A. & Schwede, T. 2013. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr D Biol Crystallogr,* 69**,** 701-9.

Bolten, E., Schliep, A., Schneckener, S., Schomburg, D. & Schrader, R. 2001. Clustering protein sequences--structure prediction by transitive homology. *Bioinformatics,* 17**,** 935-41.

Bonvin, A. M. 2006. Flexible protein-protein docking. *Curr Opin Struct Biol,* 16**,** 194-200.

Bouqellah, N. A. & Farag, P. F. 2023. In Silico Evaluation, Phylogenetic Analysis, and Structural Modeling of the Class II Hydrophobin Family from Different Fungal Phytopathogens. *Microorganisms,* 11.

Bowie, J. U., Luthy, R. & Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* 253**,** 164–70.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. 1983. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.,* 4**,** 187-217.

Bryant, Pozzati & Elofsson 2022a. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications,* 13.

Bryant, P., Pozzati, G., Zhu, W., Shenoy, A., Kundrotas, P. & Elofsson, A. 2022b. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat Commun,* 13**,** 6028.

Buenavista, M. T., Roche, D. B. & Mcguffin, L. J. 2012. Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics,* 28**,** 1851-7.

Burcu Ozden, Andriy Kryshtafovych & Karaca, E. 2023. The Impact of AI-Based Modeling on the Accuracy of Protein Assembly Prediction: Insights from CASP15. *BioRxiv*.

Callaway, E. 2020. Revolutionary cryo-EM is taking over structural biology. *Nature,* 578**,** 201.

Chen, H. & Skolnick, J. 2008. M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys J,* 94**,** 918-28.

Chen, J. & Siu, S. W. I. 2020. Machine Learning Approaches for Quality Assessment of Protein Structures. *Biomolecules,* 10.

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. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr,* 66**,** 12-21.

Chen, X., Morehead, A., Liu, J. & Cheng, J. 2023. A gated graph transformer for protein complex structure quality assessment and its performance in CASP15. *Bioinformatics,* 39**,** i308-i317.

Cheng, J., Roy, R. S., Liu, J., Giri, N. & Guo, Z. 2023. Combining pairwise structural similarity and deep learning interface contact prediction to estimate protein complex model accuracy in CASP15. *bioRxiv*.

Chothia, C. & Lesk, A. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.,* 5**,** 823-6.

Chou, P. Y. & Fasman, G. D. 1974. Prediction of protein conformation. *Biochemistry,* 13**,** 222-45.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J. & Kollman, P. 1996. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc.,* 118**,** 2309–2309.

Crawley, M. J. 2015. *Statistics. An introduction using R.*, Wiley.

Croll, T. I., Sammito, M. D., Kryshtafovych, A. & Read, R. J. 2019. Evaluation of template-based modeling in CASP13. *Proteins,* 87**,** 1113-1127.

D'aguanno, S. & Del Bufalo, D. 2020. Inhibition of Anti-Apoptotic Bcl-2 Proteins in Preclinical and Clinical Studies: Current Overview in Cancer. *Cells,* 9.

Dapkunas, J., Olechnovic, K. & Venclovas, C. 2019. Structural modeling of protein complexes: Current capabilities and challenges. *Proteins,* 87**,** 1222-1232.

Dapkunas, J., Olechnovic, K. & Venclovas, C. 2021. Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction. *Proteins,* 89**,** 1834-1843.

De Juan, D., Pazos, F. & Valencia, A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet,* 14**,** 249-61.

Deyaert, E., Leemans, M., Singh, R. K., Gallardo, R., Steyaert, J., Kortholt, A., Lauer, J. & Versees, W. 2019. Structure and nucleotide-induced conformational dynamics of the Chlorobium tepidum Roco protein. *Biochem J,* 476**,** 51-66.

Djinovic-Carugo, K. & Carugo, O. 2015. Missing strings of residues in protein crystal structures. *Intrinsically Disord Proteins,* 3**,** e1095697.

Duarte, J. & Guzenko, D. CASP 13 Assembly assessment. 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, 2018.

Duhovny, D., Nussinov, R. & Wolfson, H. 2002. Efficient Unbound Docking of Rigid Molecules. *Algorithms in Bioinformatics*.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

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**,** 1871-1878.

Edmunds, N. S. & Mcguffin, L. J. 2021. Computational Methods for the Elucidation of Protein Structure and Interactions. *In:* OWENS, R. J. (ed.) *Methods in Molecular Biology.* Springer Nature.

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**,** 1922-1939.

Elofsson, A., Joo, K., Keasar, C., Lee, J., Maghrabi, A. H. A., Manavalan, B., Mcguffin, L. J., Menendez Hurtado, D., Mirabello, C., Pilstal, R., Sidi, T., Uziela, K. & Wallner, B. 2018. Methods for estimation of model accuracy in CASP12. *Proteins,* 86 Suppl 1**,** 361-373.

Englander, S. W. & Mayne, L. 2014. The nature of protein folding pathways. *Proc Natl Acad Sci U S A,* 111**,** 15873-80.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. & Sali, A. 2006. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics,* Chapter 5**,** Unit-5 6.

Evans, R., O'neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. 2022. Protein complex prediction with AlphaFold-Multimer. *bioRxiv***,** 2021.10.04.463034.

Fan, H. & Mark, A. E. 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci,* 13**,** 211-20.

Feig, M. 2017. Computational protein structure refinement: Almost there, yet still so far to go. *wiley interdiscip. rev. comput. mol. sci,* 7.

Fischer, D. 2003. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins,* 51**,** 434-41.

Fiser, A. 2010. Template-based protein structure modeling. *Methods Mol Biol,* 673**,** 73-94.

Fiser, A. & Sali, A. 2003. ModLoop: automated modeling of loops in protein structures. *Bioinformatics,* 19**,** 2500-1.

Gajria, D. & Chandarlapaty, S. 2011. HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. *Expert Rev Anticancer Ther,* 11**,** 263-75.

Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. 2022. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat Commun,* 13**,** 1744.

Garzon, J. I., Lopez-Blanco, J. R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J. & Chacon, P. 2009. FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics,* 25**,** 2544-51.

Genc, A. G. 2024. Improvement of refinement techniques for protein quaternary structures.: University of Reading.

Gerdes, K. 2024. Diverse genetic contexts of HicA toxin domains propose a role in anti-phage defense. *mBio,* 15**,** e0329323.

Gerencer, M. & Mcguffin, L. J. 2023. Are the integrin binding motifs within SARS CoV-2 spike protein and MHC class II alleles playing the key role in COVID-19? *Front Immunol,* 14**,** 1177691.

Ghani, U., Desta, I., Jindal, A., Khan, O., Jones, G., Hashemi, N., Kotelnikov, S., Padhorny, D., Vajda, S. & Kozakov, D. 2022. Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro. *bioRxiv*.

Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. 2022. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol,* 23**,** 40-55.

Greer, J. 1980. Model for haptoglobin heavy chain based upon structural homology. *Proc Natl Acad Sci USA.,* 77**,** 3393-7.

Guarra, F. & Colombo, G. 2023. Computational Methods in Immunology and Vaccinology: Design and Development of Antibodies and Immunogens. *J Chem Theory Comput,* 19**,** 5315-5333.

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**,** 387-398.

Han, Y., He, F., Chen, Y., Qin, W., Yu, H. & Xu, D. 2021. Quality Assessment of Protein Docking Models Based on Graph Neural Network. *Front Bioinform,* 1**,** 693211.

Henderson, R. 1995. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys,* 28**,** 171-93.

Henikoff, S. & Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A,* 89**,** 10915-9.

Heo, L., Lee, H. & Seok, C. 2016. GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. *Sci Rep,* 6**,** 32153.

Hu, Y., Cheng, K., He, L., Zhang, X., Jiang, B., Jiang, L., Li, C., Wang, G., Yang, Y. & Liu, M. 2021. NMR-Based Methods for Protein Analysis. *Anal Chem,* 93**,** 1866-1879.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., Wodak, S. J. & Critical Assessment Of, P. I. 2003. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins,* 52**,** 2-9.

Johansson-Akhe, I. & Wallner, B. 2022. InterPepScore: a deep learning score for improving the FlexPepDock refinement protocol. *Bioinformatics,* 38**,** 3209-3215.

Jones, D. T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol,* 292**,** 195-202.

Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics,* 28**,** 184-90.

Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. 2015. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics,* 31**,** 999-1006.

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., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M.,

Pacholska, M., Berghammer, T., Silver, D., Vinyals, O., Senior, A., Kavukcuoglu, K., Kohli, P. & Hassabis, D. 2021a. Applying and improving AlphaFold at CASP14. *Proteins,* 89.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. a. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. 2021b. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596**,** 583-589.

Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*.

Kabsch, W. & Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* 22**,** 2577-637.

Kamisetty, H., Ovchinnikov, S. & Baker, D. 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A,* 110**,** 15674-9.

Kandathil, S. M., Greener, J. G. & Jones, D. T. 2019. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins,* 87**,** 1092-1099.

Karaca, E. CASP14 Assembly Assessment. 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, 2020.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A,* 89**,** 2195-9.

Kinch, L. N., Pei, J., Kryshtafovych, A., Schaeffer, R. D. & Grishin, N. V. 2021. Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction (CASP14). *Proteins,* 89**,** 1673-1686.

Klukowski, P., Riek, R. & Guntert, P. 2022. Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat Commun,* 13**,** 6151.

Krissinel, E. 2010. Crystal contacts as nature's docking solutions. *J Comput Chem.,* 31**,** 133-143.

Krissinel, E. & Henrick, K. 2007. Inference of macromolecular assemblies from crystalline state. *J Mol Biol,* 372**,** 774-97.

Kryshtafovych, Schwede, Topf, Fidelis & Moult 2019. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins,* 87**,** 1011-1020.

Kryshtafovych, A. & Fidelis, K. 2009. Protein structure prediction and model quality assessment. *Drug Discov Today,* 14**,** 386-93.

Kryshtafovych, A., Monastyrskyy, B. & Fidelis, K. 2014. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins,* 82 Suppl 2**,** 7-13.

Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Moult, J., Schwede, T. & Tramontano, A. 2018. Evaluation of the template-based modeling in CASP12. *Proteins,* 86 Suppl 1**,** 321-334.

Kryshtafovych, A., Montelione, G. T., Rigden, D. J., Mesdaghi, S., Karaca, E. & Moult, J. 2023. Breaking the conformational ensemble barrier: Ensemble structure modeling challenges in CASP15. *Proteins,* 91**,** 1903-1911.

Kurakin, A. 2009. Scale-free flow of life: on the biology, economics, and physics of the cell. *Theor Biol Med Model,* 6**,** 6.

Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F. & Mackay, J. P. 2011. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J,* 278**,** 687-703.

Kwon, S., Won, J., Kryshtafovych, A. & Seok, C. 2021. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. *Proteins,* 89**,** 1940-1948.

Lafita, A., Bliven, S., Kryshtafovych, A., Bertoni, M., Monastyrskyy, B., Duarte, J. M., Schwede, T. & Capitani, G. 2018. Assessment of protein assembly prediction in CASP12. *Proteins,* 86 Suppl 1**,** 247-256.

Lassmann, T. & Sonnhammer, E. L. 2005. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics,* 6**,** 298.

Lee, H., Baek, M., Lee, G. R., S, P. & Seok, C. 2017. Template-based modelling and ab initio refinement of protein oligomer structures using GALAXY in CAPRI round 30. *Proteins,* 85**,** 399-407.

Lensink, M., Velankar, S. & Wodak, S. 2017. Modelling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins,* 85**,** 359-377.

Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C. & Wodak, S. J. 2018. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins,* 86 Suppl 1**,** 257-273.

Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S. Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., Grudinin, S., Popov, P., Neveu, E., Lee, H., Baek, M., Park, S., Heo, L., Rie Lee, G., Seok, C., Qin, S., Zhou, H. X., Ritchie, D. W., Maigret, B., Devignes, M. D., Ghoorah, A., Torchala, M., Chaleil, R. A., Bates, P. A., Ben-Zeev, E., Eisenstein, M., Negi, S. S., Weng, Z., Vreven, T., Pierce, B. G., Borrman, T. M., Yu, J., Ochsenbein, F., Guerois, R., Vangone, A., Rodrigues, J. P., Van Zundert, G., Nellen, M., Xue, L., Karaca, E., Melquiond, A. S., Visscher, K., Kastritis, P. L., Bonvin, A. M., Xu, X., Qiu, L., Yan, C., Li, J., Ma, Z., Cheng, J., Zou, X., Shen, Y., Peterson, L. X., Kim, H. R., Roy, A., Han, X., Esquivel-Rodriguez, J., Kihara, D., Yu, X., Bruce, N. J., Fuller, J. C., Wade, R. C., Anishchenko, I., Kundrotas, P. J., Vakser, I. A., Imai, K., Yamada, K., Oda, T., Nakamura, T., Tomii, K., Pallara, C., Romero-Durana, M., Jimenez-Garcia, B., Moal, I. H., Fernandez-Recio, J., Joung, J. Y., Kim, J. Y., Joo, K., Lee, J., Kozakov, D., Vajda, S., Mottarella, S., Hall, D. R., Beglov, D., Mamonov, A., Xia, B., Bohnuud, T., Del Carpio, C. A., Ichiishi, E., Marze, N., Kuroda, D., Roy Burman, S. S., Gray, J. J., Chermak, E., Cavallo, L., Oliva, R.*, et al.* 2016. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins,* 84 Suppl 1**,** 323-48.

Li, D. 2022. Understanding the significance and architecture of AlphaFold.

Li, J., Hou, C., Ma, X., Guo, S., Zhang, H., Shi, L., Liao, C., Zheng, B., Ye, L., Yang, L. & He, X. 2021. Entropy-Enthalpy Compensations Fold Proteins in Precise Ways. *Int J Mol Sci,* 22.

Li, Y., Zhang, C., Bell, E. W., Yu, D. J. & Zhang, Y. 2019. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins,* 87**,** 1082-1091.

Lin, L., Chu, H. & Hodges, J. S. 2017. Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. *Biometrics,* 73**,** 156-166.

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**,** 1123-1130.

Liu, J., Liu, D., He, G. & Zhang, G. 2023. Estimating protein complex model accuracy based on ultrafast shape recognition and deep learning in CASP15. *Proteins,* 91**,** 1861-1870.

Liu, J., Lu, F., Chen, Y., Plow, E. & Qin, J. 2022. Integrin mediates cell entry of the SARS-CoV-2 virus independent of cellular receptor ACE2. *J Biol Chem,* 298**,** 101710.

Lundström, J., Rychlewski, L., Bujnicki, J. & Elofsson, A. 2001. Pcons: a neural-network–based consensus predictor that improves fold recognition. *Protein Sci.,* 10**,** 2354–2362.

Lupo, U., Sgarbossa, D. & Bitbol, A. F. 2022. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat Commun,* 13**,** 6298.

Lv, Z., Chu, Y. & Wang, Y. 2015. HIV protease inhibitors: a review of molecular selectivity and toxicity. *HIV AIDS (Auckl),* 7**,** 95-104.

Ma, P., Li, D. W. & Bruschweiler, R. 2023. Predicting protein flexibility with AlphaFold. *Proteins,* 91**,** 847-855.

Maghrabi, A. H. A. & Mcguffin, L. J. 2017. ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. *Nucleic Acids Res,* 45**,** W416-W421.

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**,** 2722-8.

Marze, N. A., Roy Burman, S. S., Sheffler, W. & Gray, J. J. 2018. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics,* 34**,** 3461-3469.

Masahito Ohue, Takehiro Shimoda, Shuji Suzuki, Yuri Matsuzaki, Takashi Ishida & Akiyama, Y. 2014. MEGADOCK 4.0: an ultra-high-performance protein-protein docking software for heterogeneous supercomputers. *Bioinformatics,* 30**,** 3281-3283

Mashiach-Farkash, E., Nussinov, R. & Wolfson, H. J. 2011. SymmRef: a flexible refinement method for symmetric multimers. *Proteins,* 79**,** 2607-23.

Mcguffin, L., Adiyaman, R., Brakenridge, D. A., Nealon, J., Philomina, L. & Shuid, A. 2018. Manual Prediction of Protein Tertiary and Quaternary Structures and 3D Model Refinement. CASP13, 2018 Riviera Maya, Mexico. 106.

Mcguffin, L., Adiyaman, R., Maghrabi, A., Shuid, A., Brackenridge, D., Nealon, J. & Philomina, L. 2019. IntFOLD: an integrated web resource for high performance protein structure and function prediction. Nucleic Acids Res. *Nucleic Acids Res,* 47.

Mcguffin, L. J. 2008. The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics,* 24**,** 586-7.

Mcguffin, L. J. 2010. Model Quality Prediction. *In:* KARYPIS, R. A. (ed.) *Introduction to protein structure prediction: Methods and Algorithms.*: John Wiley & Sons.

Mcguffin, L. J., Adiyaman, R., Brakenridge, D. A., Edmunds, N. S. & Philomina, L. S. Manual Prediction of Protein Tertiary and Quaternary Structures and 3D Model Refinement. CASP14, 2020.

Mcguffin, L. J., Aldowsari, F. M. F., Alharbi, S. M. A. & Adiyaman, R. 2021. ModFOLD8: accurate global and local quality estimates for 3D protein models. *Nucleic Acids Res,* 49**,** W425-W430.

Mcguffin, L. J., Buenavista, M. T. & Roche, D. B. 2013. The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Res,* 41**,** W368-72.

Mcguffin, L. J., Edmunds, N. S., Genc, A. G., Alharbi, S. M. A., Salehe, B. R. & Adiyaman, R. 2023. Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. *Nucleic Acids Res,* 51**,** W274-W280.

Mcguffin, L. J. & Roche, D. B. 2010. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics,* 2**6,** 182 - 188.

Mi, J., Zhou, X., Sun, R. & Han, J. 2023. Disabling spidroin N-terminal homologs' reverse reaction unveils why its intermolecular disulfide bonds have not evolved for 380 million years. *Int J Biol Macromol,* 249**,** 125974.

Mirdita, M., Schutze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods,* 19**,** 679-682.

Mirdita, M., Von Den Driesch, L., Galiez, C., Martin, M. J., Soding, J. & Steinegger, M. 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res,* 45**,** D170-D176.

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**,** D570-D578.

Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol,* 15**,** 285-9.

Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. 2011. Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins,* 79 Suppl 10**,** 1-5.

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins,* 23**,** ii-v.

Mu, X., Gillman, C., Nguyen, C. & Gonen, T. 2021. An Overview of Microcrystal Electron Diffraction (MicroED). *Annu Rev Biochem,* 90**,** 431-450.

Mukherjee, S. & Zhang, Y. 2009. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res,* 37**,** e83.

Mulvaney, T., Kretsch, R., Elliott, L., Beton, J., Kryshtafovych, A., Rigden, D., Das, R. & Topf, M. 2023. CASP15 cryoEM protein and RNA targets: refinement and analysis using experimental maps. *bioRxiv*.

Murata, K. & Wolf, M. 2018. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim Biophys Acta Gen Subj,* 1862**,** 324-334.

Nakane T, K. A., Sente a, Mcmullan G, Masiulis S, Brown Pmge, Grigoras It, Malinauskaite L, Malinauskas T, Miehling J, Uchański T, Yu L, Karia D, Pechnikova Ev, De Jong E, Keizer J, Bischoff M, Mccormack J, Tiemeijer P, Hardwick Sw, Chirgadze Dy, Murshudov G, Aricescu Ar, Scheres Shw : 2020. Single-particle cryo-EM at atomic resolution. *Nature,* 587**,** 152 – 6.

Nealon, J. O., Philomina, L. S. & Mcguffin, L. J. 2017. Predictive and Experimental Approaches for Elucidating Protein-Protein Interactions and Quaternary Structures. *Int J Mol Sci,* 18.

Needleman, S. B. & Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol,* 48**,** 443-53.

Olechnovic, K., Kulberkyte, E. & Venclovas, C. 2013. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins,* 81**,** 149-62.

Olechnovic, K., Monastyrskyy, B., Kryshtafovych, A. & Venclovas, C. 2019. Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics,* 35**,** 937-944.

Olechnovic, K. & Venclovas, C. 2014. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *J Comput Chem,* 35**,** 672-81.

Olechnovic, K. & Venclovas, C. 2017. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins,* 85**,** 1131-1145.

Olechnovic, K. & Venclovas, C. 2023. VoroIF-GNN: Voronoi tessellation-derived protein-protein interface assessment using a graph neural network. *Proteins,* 91**,** 1879-1888.

Opella, S. J. & Marassi, F. M. 2017. Applications of NMR to membrane proteins. *Arch Biochem Biophys,* 628**,** 92-101.

Outeiral, C., Nissley, D. A. & Deane, C. M. 2022. Current structure predictors are not learning the physics of protein folding. *Bioinformatics,* 38**,** 1881-1887.

Ovchinnikov, S., Steinegger, M. & Mirdita, M. 2022. Benchmarking ColabFold in CASP15. *CASP15 Abstracts***,** 50.

Ozden, B., Kryshtafovych, A. & Karaca, E. 2023. The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15. *Proteins,* 91**,** 1636-1657.

Pages, G., Charmettant, B. & Grudinin, S. 2019. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics,* 35**,** 3313-3319.

Parasa, N. A., Namgiri, J. V., Mohanty, S. N. & Dash, J. K. 2021. Introduction to Unsupervised Learning in Bioinformatics. *In:* SATPATHY, R., CHOUDHURY, T., SATPATHY, S., MOHANTY, S. & ZHANG, X. (eds.) *Data Analytics in Bioinformatics: A machine learning perspective.*

Park, T., Woo, H., Yang, J., Kwon, S., Won, J. & Seok, C. 2021. Protein oligomer structure prediction using GALAXY in CASP14. *Proteins,* 89**,** 1844-1851.

Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M. & Lupas, A. N. 2021. High-accuracy protein structure prediction in CASP14. *Proteins,* 89**,** 1687-1699.

Pierce, B. & Weng, Z. 2007. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins,* 67**,** 1078-86.
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**,** 1771-3.
Pozzati, G., Kundrotas, P. & Elofsson, A. 2022. Scoring of protein-protein docking models utilizing predicted interface residues. *Proteins,* 90**,** 1493-1505.
Rangwala, H. & Karypis, G. 2011. *Introduction to Protein Structure Prediction: Methods and Algorithms*, John Wiley & Sons.
Remmert, M., Biegert, A., Hauser, A. & Söding, J. 2012. Hhblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods,* 9**,** 173 -175.
Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Reau, M. F., Bonvin, A. & Xue, L. C. 2021. DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. *Nat Commun,* 12**,** 7068.
Roney, J. P. & Ovchinnikov, S. 2022. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys Rev Lett,* 129**,** 238101.
Rykunov, D. & Fiser, A. 2010. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics,* 11**,** 128.
Sailbil, H. 2010. *Principles of Protein Science* [Online]. Birkbeck, University of London. Available: cryst.bbk.ac.uk/PPS95 [Accessed 2024].
Sanchez Rodriguez, F., Chojnowski, G., Keegan, R. M. & Rigden, D. J. 2022. Using deep-learning predictions of inter-residue distances for model validation. *Acta Crystallogr D Struct Biol,* 78**,** 1412-1427.
Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B. W., Pruitt, K. D. & Sherry, S. T. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Res,* 50**,** D20-D26.
Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res,* 33**,** W363-7.
Schrödinger, L. 2018. The PyMOL Molecular Graphics System. Version 2.0 ed.
Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, 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. 2019. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins,* 87**,** 1141-1148.
Shao, C., Bittrich, S., Wang, S. & Burley, S. K. 2022. Assessing PDB macromolecular crystal structure confidence at the individual amino acid residue level. *Structure,* 30**,** 1385-1394.
Shaw Stewart, P. & Mueller-Dieckmann, J. 2014. Automation in biological crystallization. *Acta Cryst.,* F70**,** 686-696.
Shuid, A. N., Kempster, R. & Mcguffin, L. J. 2017. ReFOLD: a server for the refinement of 3D models of proteins guided by accurate quality estimates. *Nucleic Acids Research,* 45.
Sievers, F. & Higgins, D. 2014. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol.,* 1079**,** 105-16.
Siew, N., Elofsson, A. & Rychlewski, L. 2000. Maxsub: an automated measure for the assessment of protein structure prediction quality. . *Bioinformatics.* .
Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol,* 268**,** 209-25.
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**,** 4827-4831.

Smith, T. & Waterman, M. 1981a. Improved tools for biological sequence comparison. . *Proceedings of the National Academy of Sciences of the United States of America (PNAS),* 78**,** 3824-3828.
Smith, T. F. & Waterman, M. S. 1981b. Identification of common molecular subsequences. *J Mol Biol,* 147**,** 195-7.
Soding, J., Biegert, A. & Lupas, A. N. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res,* 33**,** W244-8.
Soleimani Zakeri, N. S., Pashazadeh, S. & Motieghader, H. 2021. Drug Repurposing for Alzheimer's Disease Based on Protein-Protein Interaction Network. *Biomed Res Int,* 2021**,** 1280237.
Sousa, S. F., Ribeiro, A. J., Coimbra, J. T., Neves, R. P., Martins, S. A., Moorthy, N. S., Fernandes, P. A. & Ramos, M. J. 2013. Protein-ligand docking in the new millennium--a retrospective of 10 years in the field. *Curr Med Chem,* 20**,** 2296-314.
Sowmya, G., Breen, E. J. & Ranganathan, S. 2015. Linking structural features of protein complexes and biological function. *Protein Sci,* 24**,** 1486-94.
Steinegger, M. & Soding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol,* 35**,** 1026-1028.
Strasser, B. J. 2010. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. *J Hist Biol,* 43**,** 623-60.
Stroe, O. 2021. *Pfam releases structures for every protein family* [Online]. Available: https://www.ebi.ac.uk/about/news/announcements/Pfam-protein-structures/ [Accessed 24/11/23].
Studer, G., Tauriello, G. & T, S. 2023. Assessment of the assessment—All about complexes. *Proteins*.
Suzek, B. E., Wang, Y., Huang, H., Mcgarvey, P. B., Wu, C. H. & Uniprot, C. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics,* 31**,** 926-32.
Takei, Y. & Ishida, T. 2022. How to select the best model from AlphaFold2 structures? *bioRxiv***,** 2022.04.05.487218.
Terashi, G. & Kihara, D. 2018. Proteins. *Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent.,* 86.
Terwilliger 2022. Improving AlphaFold modeling using implicit information from experimental density maps. *BioRxiv*.
Terwilliger, T. C., Poon, B. K., Afonine, P. V., Schlicksup, C. J., Croll, T. I., Millan, C., Richardson, J. S., Read, R. J. & Adams, P. D. 2022. Improved AlphaFold modeling with implicit experimental information. *Nat Methods,* 19**,** 1376-1382.
The-Uniprot-Consortium 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research,* 49**,** D480-D489.
Thomas, J. M. 2012. Centenary: The birth of X-ray crystallography. *Nature,* 491**,** 186-7.
Tonges, U., Perrey, S. W., Stoye, J. & Dress, A. W. 1996. A general method for fast multiple sequence alignment. *Gene,* 172**,** GC33-41.
Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, 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., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. & Hassabis, D. 2021. Highly accurate protein structure prediction for the human proteome. *Nature,* 596**,** 590-596.
Uziela, K., Menendez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. 2017. ProQ3D: improved model quality assessments using deep learning. *Bioinformatics,* 33**,** 1578-1580.
Uziela, K., Shu, N., Wallner, B. & Elofsson, A. 2016. ProQ3: Improved model quality assessments using Rosetta energy terms. *Sci Rep,* 6**,** 33509.

Van Dijk, A. D., Boelens, R. & Bonvin, A. M. 2005. Data-driven docking for the study of biomolecular complexes. *FEBS J,* 272**,** 293-312.

Vangone, A., Rodrigues, J. P., Xue, L. C., Van Zundert, G. C., Geng, C., Kurkcuoglu, Z., Nellen, M., Narasimhan, S., Karaca, E., Van Dijk, M., Melquiond, A. S., Visscher, K. M., Trellet, M., Kastritis, P. L. & Bonvin, A. M. 2017. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins,* 85**,** 417-423.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. 2022a. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res,* 50**,** D439-D444.

Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., Anyango, S., Bienert, S., Borges, C., Deshpande, M., Green T, Hassabis D, Hatos A, Hegedus T, Hekkelman M, Joosten, R., Jumper, J., Laydon, A., Molodenskiy, D., Piovesan, D., Salladini, E., Salzberg S, Sommer, M., Steinegger, M., Suhajda, E., Svergun, D., Tenorio-Ku, L., Tosatto, S., Tunyasuvunakool, K., Waterhouse, A., Žídek, A., Schwede, T., Orengo, C. & Velankar, S. 2022b. 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. *GigaScience,* 11.

Vasker, I. 2014. Protein-Protein Docking: From Interaction to Interactome. *Biophys J,* 21**,** 1785-1793.

Vavrusa M, Andreani J, Rey J, Tuffery P & R, G. 2016. InterEvDock: A docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res*.

Venkatraman, V., Yang, Y. D., Sael, L. & Kihara, D. 2009. Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics,* 10**,** 407.

Wallner, B. 2023a. AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *bioRxiv***,** 2022.12.20.521205.

Wallner, B. 2023b. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics,* 39.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De Beer, T. a. P., Rempfer, C., Bordoli, L., Lepore, R. & Schwede, T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res,* 46**,** W296-W303.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., Dimaio, F., Baek, M. & Baker, D. 2023. De novo design of protein structure and function with RFdiffusion. *Nature,* 620**,** 1089-1100.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., 3rd, Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. 2018. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci,* 27**,** 293-315.

Wu, S. & Zhang, Y. 2007. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res,* 35**,** 3375-82.

Xiao Chen, Jian Liu, Zhiye Guo, Tianqi Wu, Jie Hou & Cheng, J. 2021. Protein model accuracy estimation empowered by deep learning and inter-residue distance prediction in CASP14. *Sci Rep,* 11.

Xu & Zhang 2010. How significant is a protein structure similarity with TM-score=0.5? *Bioinformatics***,** 889-895.

Xu, Q. & Dunbrack, R. L., Jr. 2020. ProtCID: a data resource for structural information on protein interactions. *Nat Commun,* 11**,** 711.
Xue, L. C., Dobbs, D., Bonvin, A. M. & Honavar, V. 2015. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett,* 589**,** 3516-26.
Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. & Baker, D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A,* 117**,** 1496-1503.
Yang, J. & Zhang, Y. 2015. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res,* 43**,** W174-81.
Yang, Z., Zeng, X., Zhao, Y. & Chen, R. 2023. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Target Ther,* 8**,** 115.
Yin, C. & Yau, S. S. 2017. A coevolution analysis for identifying protein-protein interactions by Fourier transform. *PLoS One,* 12**,** e0174862.
Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. 2022. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci,* 31**,** e4379.
Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. 2020. Atomic-resolution protein structure determination by cryo-EM. *Nature,* 587**,** 157-161.
Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.***,** 3370–3374.
Zhang, C., Liu, S. & Zhou, Y. 2004. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Sci,* 13**,** 391-9.
Zhang, X. & Kortholt, A. 2023. LRRK2 Structure-Based Activation Mechanism and Pathogenesis. *Biomolecules,* 13.
Zhang, Y. & Skolnick, J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins,* 57**,** 702-10.
Zhang, Y. & Skolnick, J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res,* 33**,** 2302-9.
Zhong, Q., Xiao, X., Qiu, Y., Xu, Z., Chen, C., Chong, B., Zhao, X., Hai, S., Li, S., An, Z. & Dai, L. 2023. Protein posttranslational modifications in health and diseases: Functions, regulatory mechanisms, and therapeutic implications. *MedComm (2020),* 4**,** e261.
Zwanzig, R., Szabo, A. & Bagchi, B. 1992. Levinthal's paradox. *Proc Natl Acad Sci U S A,* 89**,** 20-2.

Data availability. Unless otherwise stated, all data is open access and derived from the Critical Assessment of Techniques for Protein Structure Prediction (CASP) community resource; CASP13 data: (https://predictioncenter.org/download_area/CASP13/) and CASP14 data (https://predictioncenter.org/download_area/CASP14). Data for Chapters 2 and 5 has been substantially processed as so has been made available in the form used in this study at https://GitHub.com/NickEdmunds /recycle_data/

Appendices

Appendix 1

Definitions of key quality scoring routines used in this study.

GDT_TS. Global Distance Score (Total Score) is a common CASP score and represents the number of model residues which fall into a predefined distance constraint when compared to the native structure. The score is expressed as a percentage and so the higher the score, the greater the percentage of residues found within this distance. Higher scores are better with 100 representing the perfect fit. CASP uses the mean sum of four constraint distances (1, 2, 4 and 8A), i.e. GDT_TS = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4.

RMSD. Root Mean Square Deviation. This considers the distance in 3-D space (x,y,z) between two sets of coordinates (the model (r) and native structure (r')) for C-alpha atoms. The squares of each distance (rx – r'x, ry - r'y, rz - r'z etc.) are summed and divided by the total number of residues considered. RMSD is the square root of this value (closer to 0 the better).

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{rx} - \mathrm{r'x})^2 + (\mathrm{ry} - \mathrm{r'y})^2 + (\mathrm{rz} - \mathrm{r'z})^2}$$

TM-Score. Template Modelling Score which is traditionally used to assess the similarity between the tertiary structures of two proteins. A 0-1 score with >0.5 considered to generally represent the same globular fold. It is essentially the reciprocals of target sequence length multiplied by the sum of the distance of each aligned residue divided by the modified cubed root of the aligned length (d0).

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

lDDT. Local Distance Difference Test expressed as a score between 0 and 1. The score is designed to be super-position-independent and expresses the fraction of contacts shared between a model and native structure regardless of any difference in the actual orientation. lDDT is calculated for all pairs of atoms present in the native structure within an inclusion radius (often 5-15Å) and lDDT scores are calculated as the fraction of preserved contacts where a preserved contact is determined as being within 0.5Å, 1Å, 2Å and 4Å. Total lDDT score is an average of the fraction of contacts preserved over the four distances. If one of the atoms in a pair is missing the distance is considered non-conserved. A score of 0 represents no conserved contacts and 1 represents a perfect match. In reality these extremes are rarely seen, and scores tend to be in the range 0.25 – 0.6.

QS-score. Quaternary Structure score. It expresses the fraction of shared interface contacts within 12A. A 0–1 score with 0 representing a radically different quaternary structures and 1 suggesting very similar models. QS-score is Calculated as follows:

Identify equivalent chains by sequence alignment. Calculate symmetry of the complex and create symmetry groups from chains which can reproduce the full structure. Use superposition to map the chains of two identical symmetry groups from different models. For each symmetry group, consider all possible pairings using one symmetry group as a base to superpose complexes, the lowest global RMSD considered the correct mapping. Identify “mapped” residues as those equivalent by sequence alignment between models. Identify contacts as Cβ atoms (Cα for Glycine) of residues from different chains within 12Å. Identify “shared” residues as those mapped and that form a contact in both models. Non-shared residues are those that either form contacts but are not “mapped” or that are “mapped” but form contacts only in one model. Dapkūnas, Olechnovič and Venclovas (Dapkunas *et al.*, 2019) ,in an analysis of their CASP13 performance, defined categories as; high > 0.7; medium 0.3 to <0.7; low >0.1 to 0.3; and incorrect as ≤ 0.1.

Jaccard or Interface Patch Similarity (IPS). A 0-1 score calculated using the number of interface residue contacts that are present in ***both*** the model (A) and the target (B) divided by the interface residues in the target (B) but ***not*** in the model (A) + those in the model (A) but ***not*** in the target (B). Often written as **J(A, B) =** |A∩B| / |A∪B| (Lafita *et al.*, 2018).

F1 or Interface Contact Similarity (ICS). A 0-1 score equivalent to the F1 score divided by 100. It can be calculated as the combination of precision (P) and recall (R) of contact predictions where contacts are defined as non-Hydrogen atoms from residues on different chains within 5Å of each other. Distances below 3Å are treated as clashes. ICS is calculated as:

$$\text{ICS or F1(P,R)} = 2 \times \frac{\mathrm{P(M,T)\ x\ R(M,T)}}{\mathrm{P(M,T)+R(M,T)}}$$

where M is the model contact set and T is the target contact set (Lafita *et al.*, 2018).
Definitions of precision and recall are covered in Section 5.1, but briefly, Precision is TP/TP+FP and Recall (sensitivity) is TP/TP+FN. An F1 score can be calculated as the harmonic mean of Precision and Recall (i.e. the reciprocal of reciprocal values, e.g. 2/(1/prec + 1/recall)). Recall is calculated using the number of correct interface residues in the model divided by the number of all native interface residues in the target (x100) and Precision is a similar score to recall but this time calculated by the number of correct interface residues in the model divided by the sum of the correct and incorrect interface residues in the model.

GDT HA is the Global Distance Test, High Accuracy score. This is calculated by the same method as GDT TS but uses stricter distance cut-offs: (0.5Å + 1Å + 2Å + 4Å)/4

SG score is the Sphere Grinder score (https://predictioncenter.org/casp12/doc/help.html).

The Sphere Grinder score is calculated using two parameters: a sphere of fixed radius and two RMSD cutoff values of 2Å and 4Å. For each residue, the RMSD is calculated between the model and the target using only the atoms falling inside a sphere of 6Å which centres around the Cα atom. The global Sphere Grinder Score (SG) is then calculated as the percentage of residues with RMSD under each of the 2Å and 4Å cutoff values.

CAD score is the Contact Area Difference score (Olechnovic *et al.*, 2013) For this score, the contact area for each pair of residues with a nonzero contact in the target structure is calculated along with the equivalent residue contact area in the model. For every residue pair the contact area difference is then the absolute difference of contact areas between residues in the target and in model. Additional residues in the model not present in the target are excluded and residues missing from the model have their contact areas set to zero.

Appendix 2

Data from the CASP13 competition.

Table S2.1 Definitions of CASP multimer target difficulty categories.

Category	Description
Easy	A template exists for sub-unit and assembly.
Medium	A partial template exists for sub-unit or assembly.
Difficult	No template exist for either sub-unit or assembly.

Table S2.2 List of individual targets and scores for CASP13 assembly models submitted by the McGuffin group along with ModFOLDdock and CASP scores. (Target colour key: Hard, Med, Easy)

Target	Type	Submitted model name	ModFOLDdock scores		CASP scores			
			Consensus6	Observed Mean	GDT_TS	RMSD	IDDT (olig)	QS (best)
T0960	trimer	T0960-zdock.2.pdb	0.356	0.156	6.55	71.86	0.285	0.000
T0961	tetramer	T0961_Refine1_assembly1_4y9j.ent	0.370	0.441	23.70	31.07	0.689	0.000
T0963	trimer	T0963-zdock.5.pdb	0.317	0.144	6.83	77.57	0.331	0.000
T0965	dimer	T0965_Refold8_assembly1_4zrm.ent	0.369	0.436	32.75	15.19	0.582	0.200
T0966	dimer	T0966_Refold9_assembly1_5t09.ent	0.331	0.161	30.66	33.58	0.597	0.000
T0970	dimer	T0970-zdock-complex.7.pdb	0.379	0.207	20.71	14.31	0.351	0.000
T0973	dimer	T0970-zdock-complex.15.pdb	0.364	0.172	26.76	20.21	0.340	0.016
T0976	dimer	Frodock-T0976_25.pdb	0.378	0.166	27.05	25.88	0.570	0.001
T0977	trimer	T0977-zdock-complex.4.pdb	0.446	0.191	14.40	42.55	0.477	0.002
T0979	trimer	T0979-mzdock-complex.1.pdb	0.367	0.256	14.17	47.54	0.314	0.000
T0981	trimer	zdock-T0981-complex.5.pdb	0.510	0.148	6.51	59.09	0.318	0.001
T0983	dimer	T0983-patchdock-output.txt.15.pdb	0.399	0.287	45.04	21.14	0.751	0.000
T0984	dimer	patchdock-T0984-output.txt.5.pdb	0.399	0.326	45.38	5.53	0.634	0.477
T0985*	dimer	zdock-T0985-complex.3.pdb	0.359	0.269	34.37	9.02	0.416	0.150
T0989	trimer	megadock-T0989-ABC_11.pdb	0.462	0.125	8.88	34.53	0.250	0.014
T0991	dimer	megadock-T0991_23.pdb	0.375	0.114	11.04	23.45	0.231	0.001

T0995	octamer	zdock-T0995-ABCDEFGH-4.pdb	0.733	0.225	10.40	33.28	0.590	0.018
T0996	hexamer	Manually constructed from dimer	NA	NA	3.84	59.72	0.492	0.006
T0997	dimer	Frodock-T0997_6.pdb	0.321	0.179	31.10	15.38	0.494	0.000
T0998	dimer	zdock-T0998-14.pdb	0.341	0.08	8.21	29.04	0.165	0.000
T0999	dimer	Frodock-T0999_3.pdb	0.242	0.198	12.80	39.41	0.691	0.005
T1000	dimer	megadock-T1000_22.pdb	0.284	0.158	23.86	23.47	0.568	0.000
T1001	dimer	megadock-T1001_6.pdb	0.384	0.169	39.03	9.17	0.669	0.036
T1003	dimer	zdock-T1003-AB-2.pdb	0.331	0.228	42.58	27.02	0.643	0.000
T1004	trimer	mzdock-T1004-ABC.7.pdb	0.378	0.246	16.56	53.19	0.527	0.003
T1006	dimer	Frodock-T1006-AB_11.pdb	0.406	0.319	49.66	14.46	0.639	0.000
T1009	dimer	zdock-T1009-AB-22.pdb	0.285	0.270	32.39	16.37	0.575	0.004
T1010	dimer	Frodock-T1010-AB_3.pdb	0.358	0.260	26.14	10.38	0.357	0.072
T1016	dimer	T1016_Refold8_assembly1_4ij5.ent	0.458	0.667	76.73	2.50	0.689	0.693
T1018	dimer	Frodock-T1018-AB_1.pdb	0.354	0.212	39.89	14.62	0.637	0.000
T1020	timer	zdock-T1020-ABC-5.pdb	0.462	0.381	23.62	22.71	0.567	0.019

** Originally released as A1 although has A2 structure – excluded from any analyses.*

Appendix 3

Individual CASP13 target performance by lDDT and QS scores.

Figure S3.1 (below) shows CASP13 oligo-lDDT and QS-scores for submitted structures. Nine models (T0961, T0982, T0984, T0999, T1001, T1003, T1006, T1016 and T1018) scored above 0.6 for lDDT. Less impressive is the spread of QS-score which considers the interface and therefore implicitly the relative orientations of the monomers.

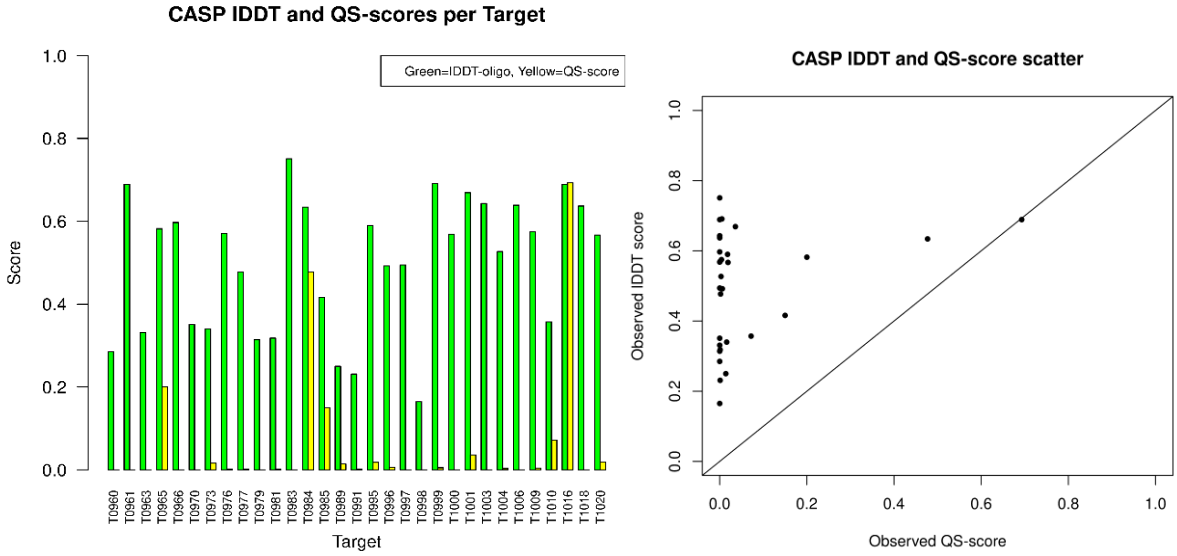


Figure S3.1 Individual CASP13 target performance by lDDT and QS scores. Left, a bar graph to show comparative score magnitude, right a scatter graph of the same data showing that MultiFOLD CASP13 models rated more highly with lDDT than QS-score.

Appendix 4

Model images from the CASP13 competition

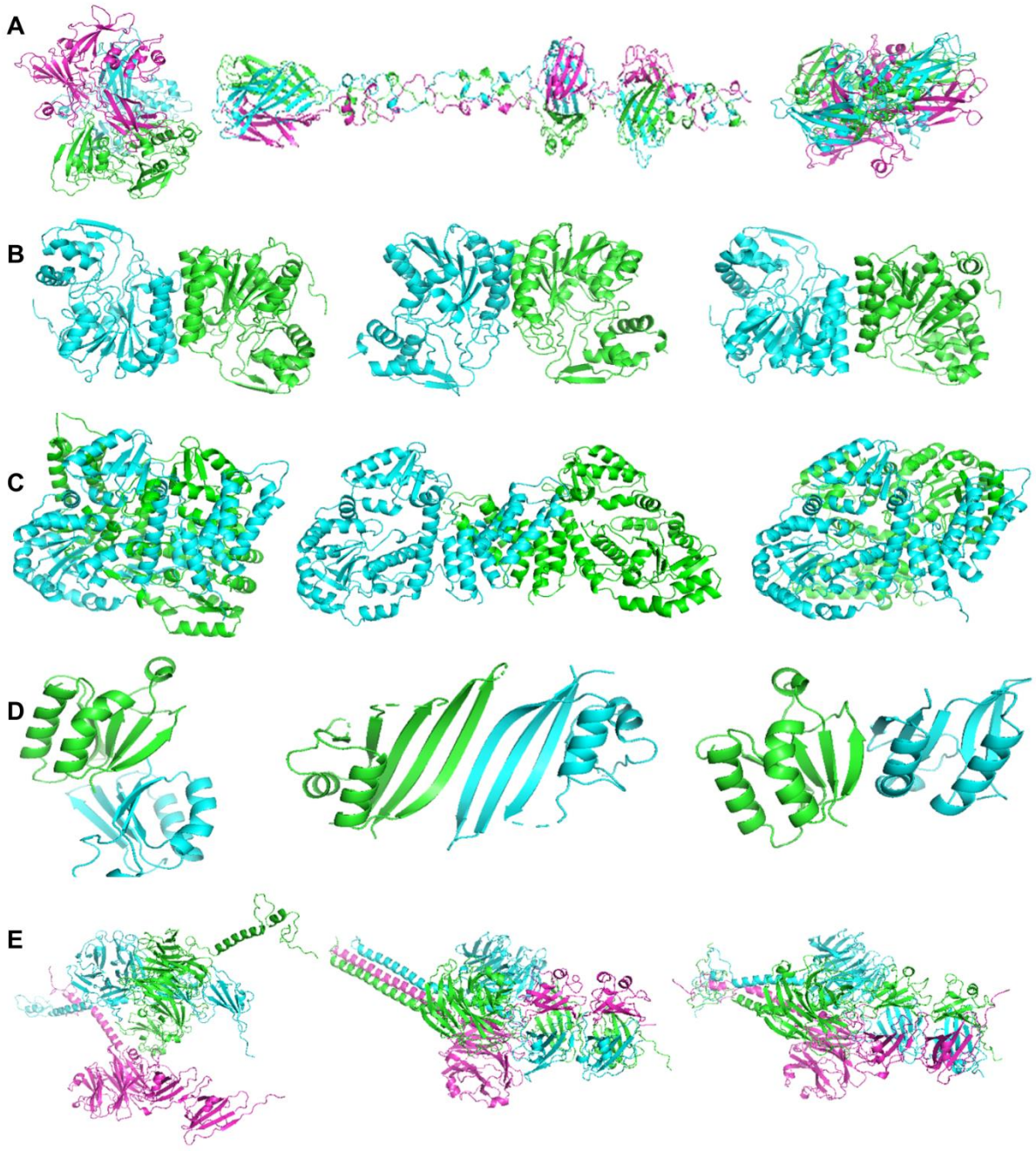


Figure S4.1 McGuffin group submitted, best and native CASP13 assembly structures. A. T0960, B. T0965, C. T0966, D. T0970, E. T0977. For each row, the submitted model is on the left, the CASP reference structure is central and the best McGuffin group model by mean observed score is on the right.

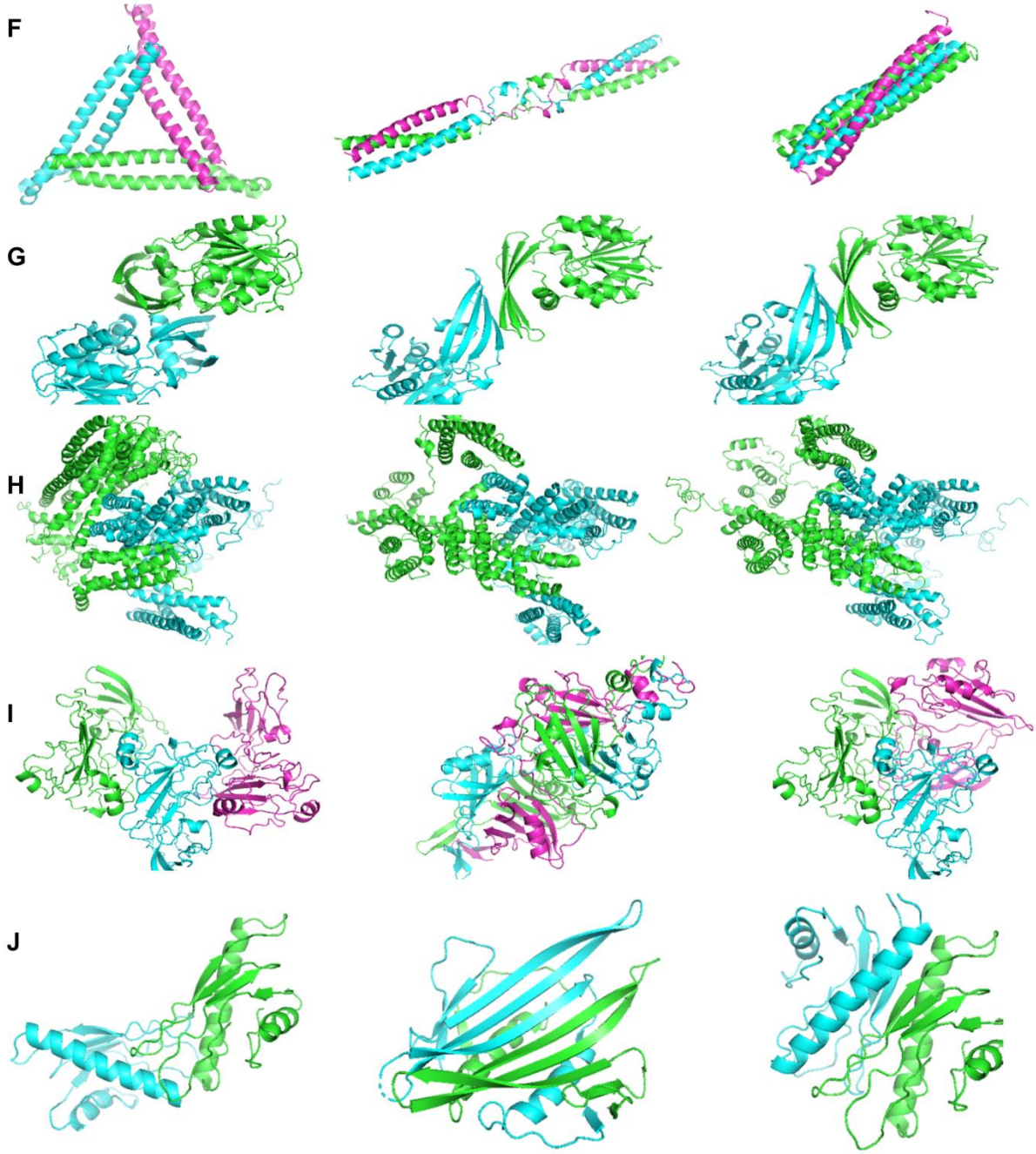


Figure S4.2 McGuffin group submitted, best and native CASP13 assembly structures. F. T0979, G. T0983, H. T0984, I. T0989, J. T0991. For each row, the submitted model is on the left, the CASP reference structure is central and the best McGuffin group model by mean observed score is on the right.

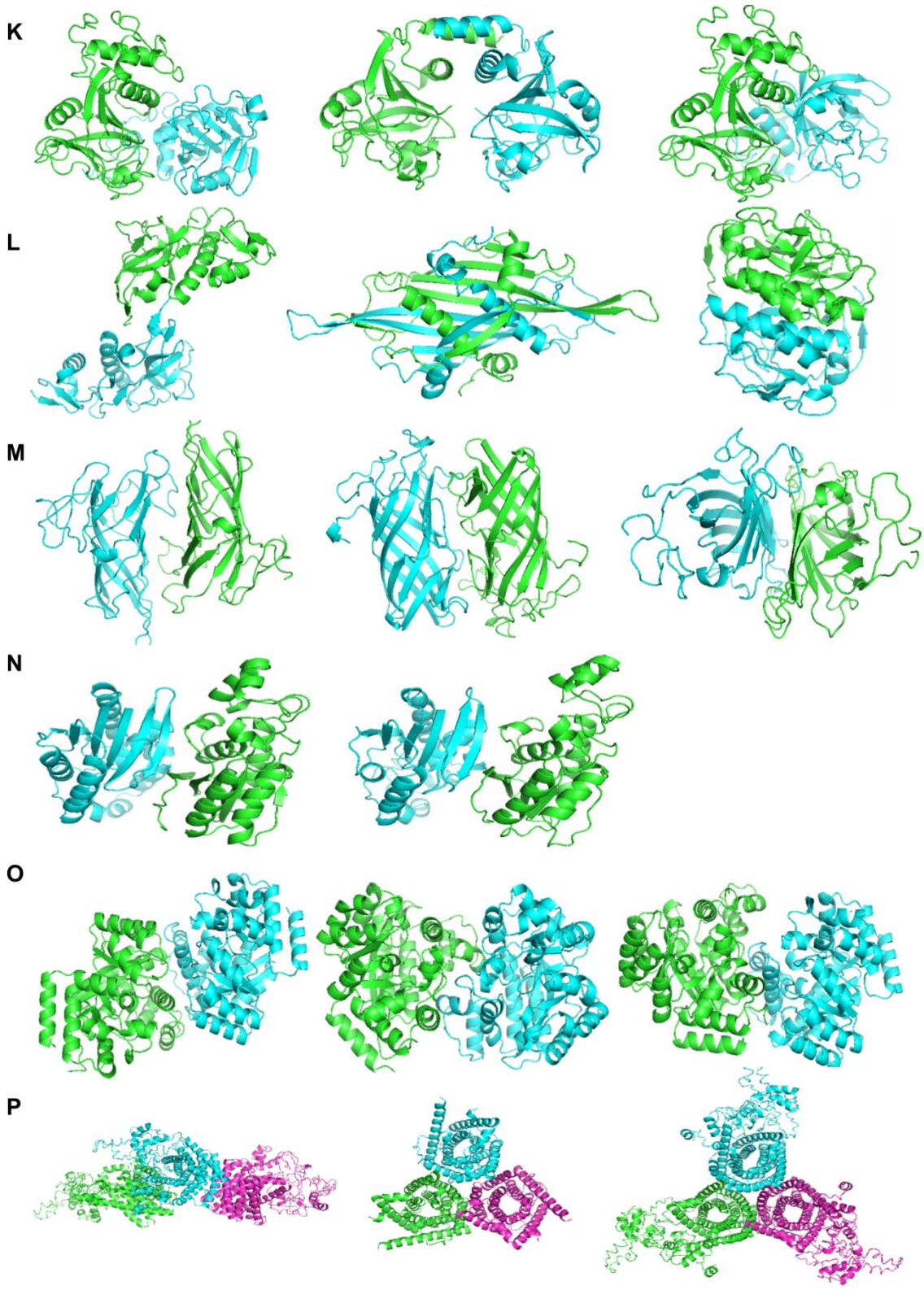


Figure S4.3 McGuffin group submitted, best and native CASP13 assembly structures. K. T0997, L. T0998, M. T1010, N. T1016, O. T1018, P. T1020. For each row, the submitted model is on the left, the CASP reference structure is central and the best McGuffin group model by mean observed score is on the right. For T1016, the submitted model was also the best model.

Table S4.1 McGuffin group submitted CASP13 assembly structures. For each row, the predicted score is the ModFOLDdock Consensus6 score and the observed score is an observed mean score. For T1016, the submitted model was also the best model.

Target	Submitted model		Best model	
	Predicted score	Observed score	Predicted score	Observed score
T0960	0.356	0.156	0.343	0.328
T0965	0.369	0.436	0.322	0.487
T0966	0.331	0.161	0.202	0.259
T0970	0.379	0.207	0.295	0.301
T0977	0.446	0.191	0.179	0.468
T0979	0.367	0.256	0.260	0.452
T0983	0.399	0.287	0.370	0.834
T0984	0.399	0.326	0.372	0.604
T0989	0.462	0.125	0.350	0.197
T0991	0.375	0.114	0.277	0.199
T0997	0.321	0.179	0.273	0.261
T0998	0.341	0.08	0.273	0.188
T1010	0.358	0.260	0.285	0.382
T1016	0.458	0.667		
T1018	0.354	0.212	0.264	0.381
T1020	0.462	0.381	0.306	0.621

Appendix 5

Data from the CASP14 competition.

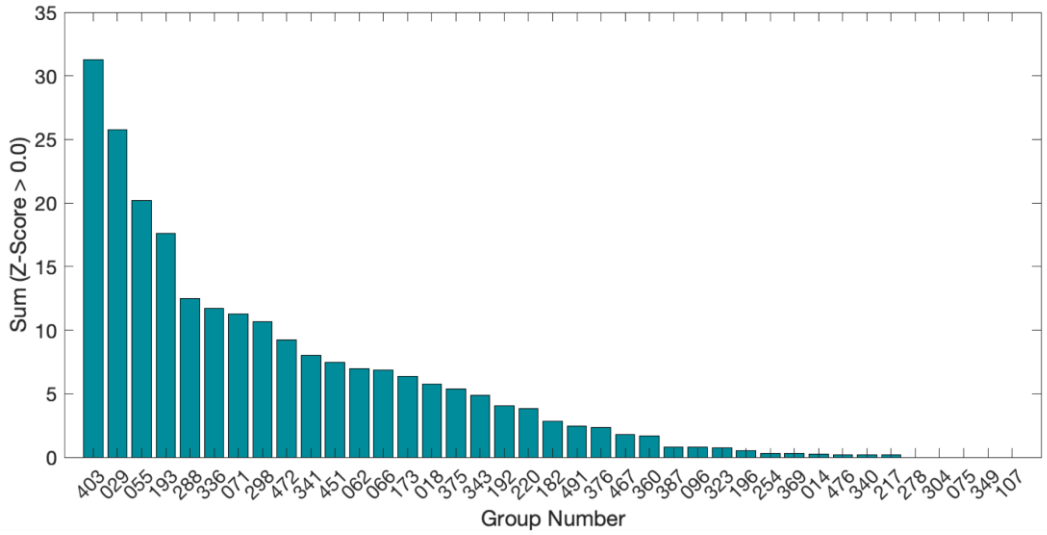


Figure S5.1 CASP14 final group rankings for assembly structures by summed Z-score. McGuffin (19th) is Group 220. (Image from https://predictioncenter.org/casp14/zscores_multimer.cgi).

Table S5.1 Full list of McGuffin group CASP14 assembly models. ModFOLDdock predicted scores, a calculated observed score and CASP official scores are also listed.

Target (Difficult Med Easy)	Type	Submitted model name	ModFOLDdock scores		CASP official scores			
			Calculated predicted (Consensus6) score.	Calculated Observed Mean	Global		Local	
					TM-score	lDDT (oligo)	ICS (F1)	IPS (Jacc)
T1032	Dimer (A2)	Yang_FM_TS3_pdb1gxj.pdb	0.524	0.380	0.644	0.429	26.0	0.28
T1034	Tetramer (D2)	Complex1.pdb (intertwined monomer)	0.352	0.338	0.268	0.607	0.0	0.06
T1038	Dimer (A2)	Model01.pdb	0.08	0.08	0.216	0.130	0.0	0.12
T1048	Tetramer (D2)	T1048_ReFOLD_pdb5k7b.pdb	0.275	0.232	0.519	0.127	1.3	0.20
T1052	Trimer (C3)	T1052_ReFOLD_pdb6f7d.pdb	0.356	0.452	0.691	0.556	33.2	0.45
T1054	Dimer (C2)	Complex5.pdb	0.336	0.239	0.495	0.531	0.0	0.19
T1061	Trimer (C3)	T1061_ReFOLD_10_pdb3cdd.pdb	0.487	0.206	0.473	0.240	0.7	0.16

T1062	Trimer	Part of H1060 T5 bacteriophage tail. Cancelled. 1700_TR1062_pdb3cop.pdb	0.383	0.282	-	-	-	-
T1070	Trimer (C3)	Complex4.pdb (intertwined monomer)	0.488	0.199	0.177	0.395	0.0	0.04
T1073	Tetramer	Cancelled	-	-	-	-	-	-
T1078	Dimer (A2)	Decoy3.pdb	0.480	0.390	0.519	0.556	0.0	0.39
T1080	Trimer (C3)	Complex3.pdb (intertwined monomer)	0.543	0.204	0.218	0.181	3.3	0.15
T1083	Dimer (C2)	Decoy9.pdb	0.436	0.465	0.603	0.578	30.6	0.44
T1084	Dimer (C2)	Decoy1.pdb	0.314	0.392	0.700	0.491	0.0	0.50
T1087	Dimer (C2)	Complex9.pdb	0.353	0.197	0.420	0.326	0.0	0.25

Appendix 6

Data for recycling models.

Table S6.1 Raw oligo-lDDT, TM-score and QS-score values for non-AF2 multimeric templates and recycled models. Values for baseline and MSA recycling up to 6 recycles.

Model	Base lDDT	Base Tmscore	Base QS	R1M lDDT	R1M TM	R1M QS	R3M lDDT	R3M TM	R3M QS	R6M lDDT	R6M TM	R6M QS
H1045TS403_1	0.6941	0.8705	0.55	0.8348	0.91572	0.97	0.8411	0.91612	0.97	0.8406	0.91485	0.97
H1065TS403_1	0.6964	0.79217	0.54	0.8489	0.96635	0.8	0.8995	0.96694	0.92	0.904	0.96781	0.91
H1072TS403_1	0.3793	0.39764	0.04	0.7326	0.90913	0.74	0.7153	0.65021	0.6	0.71	0.65284	0.6
T1032TS403_1	0.5428	0.69464	0.54	0.548	0.71818	0.64	0.5477	0.71465	0.64	0.6385	0.71047	0.65
T1054TS403_1	0.6073	0.4408	0	0.6058	0.52535	0	0.6179	0.51768	0	0.6065	0.50974	0
T1070TS403_1	0.4009	0.35172	0.04	0.3508	0.3485	0.05	0.4022	0.34343	0.04	0.4034	0.24387	0.04
T1073TS403_1	0.5496	0.36837	0	0.3769	0.43095	0	0.549	0.31806	0	0.5548	0.36847	0
T1078TS403_1	0.4989	0.5622	0.03	0.5485	0.79588	0.38	0.7814	0.90926	0.41	0.7865	0.91857	0.42
T1083TS403_1	0.6092	0.66167	0.38	0.7143	0.83212	0.74	0.6865	0.84397	0.79	0.7117	0.84599	0.77
T1084TS403_1	0.8318	0.917	0.89	0.8367	0.91925	0.9	0.8664	0.91632	0.91	0.8652	0.91512	0.91
H1045TS029_1	0.5402	0.72313	0.84	0.8803	0.95618	0.98	0.8723	0.94974	0.97	0.8802	0.95326	0.98
H1065TS029_1	0.6243	0.61334	0.1	0.9038	0.97124	0.91	0.9082	0.97181	0.92	0.9242	0.98032	0.92
H1072TS029_1	0.4639	0.54438	0.27	0.7726	0.89189	0.8	0.7581	0.86708	0.75	0.7986	0.86773	0.75
T1032TS029_1	0.4168	0.62816	0.49	0.6819	0.68005	0.78	0.6933	0.70035	0.81	0.6956	0.69974	0.82
T1054TS029_1	0.5231	0.34135	0.05	0.5323	0.31614	0.05	0.586	0.46527	0	0.6881	0.46935	0
T1070TS029_1	0.4061	0.40359	0.17	0.0307	0.43896	0	0.1387	0.43547	0	0.3773	0.57464	0.08
T1073TS029_1	0.5097	0.30683	0	0.1387	0.37109	0	0.4726	0.38088	0	0.4952	0.32283	0
T1078TS029_1	0.5525	0.5011	0.16	0.7092	0.92204	0.78	0.9106	0.97969	0.87	0.8927	0.9788	0.82
T1083TS029_1	0.4843	0.63454	0.36	0.7831	0.88441	0.88	0.7844	0.88121	0.86	0.7388	0.81028	0.75
T1084TS029_1	0.7564	0.89285	0.86	0.8538	0.91678	0.92	0.8688	0.91706	0.91	0.8657	0.91591	0.91
H1045TS055_1	0.7309	0.86583	0.9	0.8573	0.94246	0.97	0.8633	0.94865	0.97	0.8693	0.94736	0.97
H1065TS055_1	0.5651	0.46602	0	0.8802	0.96372	0.91	0.8861	0.96441	0.86	0.9036	0.96838	0.91
H1072TS055_1	0.4558	0.52742	0.26	0.7715	0.75995	0.82	0.7876	0.87109	0.83	0.772	0.78825	0.81
T1032TS055_1	0.521	0.6318	0.41	0.6607	0.70659	0.8	0.6765	0.70192	0.73	0.6712	0.69893	0.73

T1054TS055_1	0.5277	0.47971	0.02	0.5993	0.45661	0	0.5945	0.4637	0	0.6702	0.4623	0
T1070TS055_1	0.3082	0.39635	0.11	0.0433	0.4698	0	0.1257	0.45214	0.03	0.1363	0.4689	0.04
T1073TS055_1	0.5624	0.37717	0.01	0.116	0.34005	0	0.0897	0.40542	0	0.343	0.37925	0
T1078TS055_1	0.5535	0.50414	0	0.6595	0.84231	0.23	0.6929	0.60142	0.03	0.7239	0.57459	0.06
T1083TS055_1	0.5196	0.46014	0.03	0.4958	0.54648	0.14	0.4564	0.51314	0	0.4755	0.50738	0
T1084TS055_1	0.4965	0.628	0.08	0.8376	0.90441	0.91	0.8276	0.89667	0.9	0.8417	0.89508	0.91
H1045TS193_1	0.644	0.73828	0.7	0.8364	0.93322	0.97	0.8415	0.93789	0.97	0.8542	0.94305	0.97
H1065TS193_1	0.5917	0.49039	0	0.8803	0.96799	0.92	0.9067	0.97095	0.92	0.912	0.97238	0.92
H1072TS193_1	0.3669	0.41473	0	0.7469	0.91312	0.76	0.7838	0.91805	0.84	0.7826	0.92704	0.83
T1032TS193_1	0.5245	0.67494	0.32	0.5731	0.70659	0.71	0.6769	0.70668	0.82	0.6884	0.70504	0.82
T1054TS193_1	0.5808	0.4326	0	0.396	0.4494	0	0.6301	0.46481	0	0.5134	0.46279	0
T1070TS193_1	0.3493	0.20958	0.08	0.2219	0.45695	0.03	0.3996	0.60863	0.14	0.4164	0.59298	0.15
T1073TS193_1	0.577	0.264	0	0.3073	0.3604	0	0.4706	0.32679	0	0.4833	0.33302	0
T1078TS193_1	0.5525	0.52102	0.03	0.7608	0.89011	0.41	0.7898	0.91225	0.48	0.7857	0.89763	0.46
T1083TS193_1	0.5077	0.68924	0.37	0.6348	0.82861	0.65	0.6393	0.81673	0.69	0.6759	0.78969	0.71
T1084TS193_1	0.5926	0.83961	0.26	0.8194	0.91798	0.87	0.8307	0.90978	0.87	0.8317	0.90766	0.87
H1045TS288_1	0.6941	0.89242	0.55	0.8607	0.94987	0.98	0.8741	0.95031	0.97	0.8715	0.94876	0.97
H1065TS288_1	0.5922	0.54885	0.08	0.8679	0.95833	0.91	0.9009	0.96687	0.91	0.9036	0.96805	0.91
H1072TS288_1	0.1569	0.35328	0.01	0.8312	0.7735	0.84	0.8329	0.7667	0.8	0.7714	0.7646	0.79
T1032TS288_1	0.4333	0.62798	0.38	0.6672	0.67896	0.79	0.6908	0.6917	0.8	0.6931	0.69671	0.81
T1054TS288_1	0.4423	0.37302	0	0.5719	0.52735	0	0.4375	0.47961	0	0.5818	0.46331	0
T1070TS288_1	0.3645	0.46944	0.1	0.0527	0.46773	0	0.0383	0.46671	0	0.2113	0.4758	0.13
T1073TS288_1	0.6012	0.28678	0	0.1389	0.38875	0	0.442	0.44164	0	0.2923	0.42993	0
T1078TS288_1	0.5488	0.5262	0.14	0.7619	0.91001	0.37	0.7882	0.90636	0.47	0.7851	0.89773	0.46
T1083TS288_1	0.4159	0.39931	0	0.6654	0.83358	0.71	0.7379	0.8411	0.74	0.7232	0.83325	0.76
T1084TS288_1	0.4961	0.60716	0	0.9162	0.7732	0.84	0.9108	0.8217	0.87	0.9084	0.8138	0.89

Table S6.2 Raw oligo-lDDT, TM-score and QS-score values for non-AF2 multimeric templates and recycled models. Values for single sequence recycling from 1 to 6 recycles and MSA recycling for 12 recycles.

Model	R12M lDDT	R12M TM	R12M QS	R1S lDDT	R1S TM	R3S QS	R3S lDDT	R3S TM	R1S QS	R6S lDDT	R6S TM	R6S QS
H1045TS403_1	0.8394	0.91424	0.97	0.7583	0.87639	0.88	0.7704	0.88566	0.88	0.7704	0.89318	0.88
H1065TS403_1	0.893	0.95498	0.9	0.7451	0.86745	0.61	0.7674	0.90588	0.68	0.7516	0.9282	0.6
H1072TS403_1	0.6996	0.65885	0.6	0.476	0.41265	0.2	0.4481	0.41551	0.17	0.4706	0.45148	0.24
T1032TS403_1	0.639	0.70573	0.69	0.4789	0.3871	0	0.4766	0.38763	0	0.5031	0.38598	0
T1054TS403_1	0.6173	0.44815	0	0.6039	0.54798	0	0.5962	0.56297	0	0.574	0.56699	0
T1070TS403_1	0.4034	0.2439	0.04	0.0888	0.29817	0	0.0837	0.30546	0	0.0043	0.29691	0
T1073TS403_1	0.5536	0.36797	0	0.0717	0.36162	0	0.0123	0.38273	0	0.0189	0.3786	0
T1078TS403_1	0.789	0.9003	0.45	0.4854	0.6527	0.08	0.514	0.6042	0.05	0.5169	0.62909	0.08
T1083TS403_1	0.6864	0.84445	0.77	0.6148	0.81168	0.64	0.6915	0.81831	0.78	0.7219	0.82512	0.78
T1084TS403_1	0.865	0.91401	0.91	0.8235	0.9172	0.9	0.8518	0.91599	0.91	0.8563	0.91489	0.91
H1045TS029_1	0.8624	0.94753	0.96	0.5484	0.64799	0.33	0.6003	0.77669	0.81	0.489	0.76988	0.67
H1065TS029_1	0.9142	0.97721	0.92	0.6427	0.84457	0.31	0.7001	0.87	0.51	0.6591	0.8674	0.36
H1072TS029_1	0.7794	0.89112	0.82	0.4822	0.40123	0.24	0.4954	0.41017	0.25	0.493	0.40702	0.27
T1032TS029_1	0.6946	0.70124	0.82	0.3401	0.41985	0	0.1423	0.42742	0	0.0071	0.47176	0
T1054TS029_1	0.6425	0.46563	0	0.3669	0.34013	0.04	0.3592	0.35223	0.04	0.3566	0.34413	0.04
T1070TS029_1	0.3721	0.53891	0.05	0.029	0.37798	0	0.1281	0.34773	0	0.1223	0.3379	0
T1073TS029_1	0.5689	0.3067	0	0.1081	0.35169	0	0.1453	0.38384	0	0.1347	0.38853	0
T1078TS029_1	0.891	0.97861	0.83	0.5279	0.76103	0.31	0.5821	0.76482	0.35	0.7277	0.7576	0.34
T1083TS029_1	0.7404	0.80122	0.74	0.5478	0.53528	0.33	0.5206	0.53174	0.33	0.6914	0.79302	0.74
T1084TS029_1	0.8682	0.91559	0.91	0.7723	0.88271	0.84	0.8512	0.91348	0.91	0.848	0.91499	0.93

H1045TS055_1	0.8698	0.94592	0.97	0.7694	0.90183	0.9	0.8051	0.90341	0.96	0.8054	0.90536	0.96
H1065TS055_1	0.903	0.96607	0.91	0.5997	0.74899	0.44	0.6449	0.78525	0.47	0.6412	0.78592	0.48
H1072TS055_1	0.7706	0.78414	0.82	0.4273	0.62623	0.23	0.7204	0.86813	0.74	0.7247	0.84546	0.76
T1032TS055_1	0.6698	0.69709	0.72	0.494	0.37884	0	0.4227	0.37165	0	0.5105	0.37817	0
T1054TS055_1	0.6356	0.46229	0	0.472	0.5202	0.01	0.5171	0.51574	0.01	0.54	0.52285	0.01
T1070TS055_1	0.2424	0.4737	0.14	0.0609	0.25857	0	0.0607	0.25334	0	0.05	0.25729	0
T1073TS055_1	0.2358	0.32047	0	0.0946	0.49086	0	0.0876	0.36534	0	0.1153	0.40977	0
T1078TS055_1	0.7158	0.56991	0.02	0.5523	0.846	0.24	0.5417	0.84281	0.13	0.576	0.76093	0.2
T1083TS055_1	0.4627	0.50962	0	0.5442	0.56705	0.33	0.522	0.53965	0.3	0.5211	0.54316	0.32
T1084TS055_1	0.837	0.89466	0.91	0.8277	0.91436	0.9	0.8247	0.90588	0.9	0.8283	0.90297	0.9
H1045TS193_1	0.8336	0.93511	0.96	0.6704	0.81332	0.76	0.6576	0.82093	0.67	0.6646	0.82424	0.67
H1065TS193_1	0.9011	0.96569	0.91	0.7534	0.89832	0.69	0.7763	0.91781	0.7	0.7683	0.91117	0.69
H1072TS193_1	0.7833	0.93163	0.83	0.4916	0.40838	0.25	0.4884	0.40688	0.26	0.7311	0.91764	0.77
T1032TS193_1	0.692	0.7063	0.82	0.4953	0.68727	0.23	0.4947	0.67843	0.33	0.4816	0.67847	0.32
T1054TS193_1	0.6462	0.45169	0	0.4724	0.43312	0	0.4642	0.43973	0	0.4869	0.44032	0
T1070TS193_1	0.437	0.58796	0.19	0.1562	0.34389	0	0.0329	0.35657	0	0.0247	0.33723	0
T1073TS193_1	0.4945	0.32866	0	0.073	0.35206	0	0.0726	0.38864	0	0.0712	0.36999	0
T1078TS193_1	0.7828	0.89139	0.46	0.5209	0.53855	0.05	0.5242	0.55523	0.07	0.5248	0.54367	0.06
T1083TS193_1	0.681	0.83552	0.73	0.6211	0.78945	0.65	0.6241	0.7965	0.66	0.6443	0.80441	0.65
T1084TS193_1	0.8294	0.90684	0.87	0.8011	0.91514	0.86	0.8282	0.91594	0.87	0.8195	0.91123	0.87
H1045TS288_1	0.8718	0.9463	0.97	0.7079	0.87398	0.79	0.746	0.88394	0.86	0.7449	0.88577	0.86
H1065TS288_1	0.9128	0.97535	0.91	0.6588	0.83271	0.48	0.7016	0.84097	0.55	0.6975	0.8443	0.49
H1072TS288_1	0.7555	0.7713	0.82	0.1626	0.31025	0.09	0.2681	0.37219	0.23	0.382	0.3448	0.11

T1032TS288_1	0.6936	0.69949	0.82	0.4256	0.72135	0.24	0.3926	0.72378	0.2	0.3806	0.71892	0.16
T1054TS288_1	0.5554	0.46384	0	0.4876	0.67178	0.03	0.4986	0.66589	0.02	0.4949	0.65563	0.03
T1070TS288_1	0.3229	0.55227	0.15	0.0264	0.29077	0	0.0164	0.285	0	0.0336	0.25303	0
T1073TS288_1	0.255	0.41348	0	0.0379	0.40448	0	0.0615	0.42795	0	0.0794	0.39785	0
T1078TS288_1	0.7863	0.89944	0.46	0.554	0.83157	0.33	0.5756	0.79542	0.24	0.6191	0.77877	0.28
T1083TS288_1	0.7051	0.82316	0.7	0.3505	0.43179	0	0.5192	0.5407	0.29	0.5186	0.53282	0.33
T1084TS288_1	0.90696	0.8185	0.89	0.783	0.9149	0.81	0.8037	0.91367	0.81	0.801	0.91192	0.81

Table S6.3 Raw oligo-lDDT, TM-score and QS-score values for non-AF2 multimeric templates and recycled models. Values for single sequence recycling for 12 recycles.

Model	R12S lDDT	R12S TM	R12S QS	Model	R12S lDDT	R12S TM	R12S QS	Model	R12S lDDT	R12S TM	R12S QS
H1045TS403_1	0.774	0.89081	0.78	T1078TS029_1	0.726	0.75277	0.29	T1054TS193_1	0.4676	0.43913	0
H1065TS403_1	0.7816	0.93318	0.67	T1083TS029_1	0.6994	0.79688	0.73	T1070TS193_1	0.0331	0.31707	0
H1072TS403_1	0.4658	0.44995	0.24	T1084TS029_1	0.8498	0.91575	0.93	T1073TS193_1	0.0936	0.37456	0
T1032TS403_1	0.5104	0.38489	0	H1045TS055_1	0.8049	0.90574	0.96	T1078TS193_1	0.5207	0.55806	0.06
T1054TS403_1	0.5687	0.57783	0	H1065TS055_1	0.6666	0.78877	0.59	T1083TS193_1	0.6277	0.81837	0.63
T1070TS403_1	0.0837	0.28378	0	H1072TS055_1	0.7245	0.85944	0.76	T1084TS193_1	0.8214	0.9116	0.87
T1073TS403_1	0.008	0.3491	0	T1032TS055_1	0.5127	0.37748	0	H1045TS288_1	0.7572	0.88694	0.96
T1078TS403_1	0.5105	0.63449	0.09	T1054TS055_1	0.5683	0.51446	0.01	H1065TS288_1	0.6977	0.84587	0.49
T1083TS403_1	0.7188	0.82794	0.81	T1070TS055_1	0.0494	0.24598	0	H1072TS288_1	0.3814	0.36069	0.11
T1084TS403_1	0.8571	0.91484	0.91	T1073TS055_1	0.0982	0.33388	0	T1032TS288_1	0.38	0.71991	0.14
H1045TS029_1	0.5571	0.77194	0.43	T1078TS055_1	0.5936	0.75993	0.17	T1054TS288_1	0.4829	0.66127	0.03
H1065TS029_1	0.7202	0.85967	0.5	T1083TS055_1	0.5217	0.54049	0.32	T1070TS288_1	0.109	0.2726	0
H1072TS029_1	0.4954	0.39909	0.25	T1084TS055_1	0.8302	0.90327	0.9	T1073TS288_1	0.0871	0.38828	0
T1032TS029_1	0.0083	0.47871	0	H1045TS193_1	0.6608	0.82552	0.6	T1078TS288_1	0.6903	0.74729	0.34
T1054TS029_1	0.3677	0.34346	0.04	H1065TS193_1	0.8132	0.94236	0.8	T1083TS288_1	0.5352	0.52642	0.36
T1070TS029_1	0.1203	0.31612	0	H1072TS193_1	0.7627	0.92985	0.82	T1084TS288_1	0.8006	0.91217	0.81
T1073TS029_1	0.0826	0.39551	0	T1032TS193_1	0.4876	0.67988	0.32				

Table S6.4 Raw oligo-lDDT, TM-score and QS-score values for AF2 generated multimeric templates and recycled models. Values for all recycles.

Model	Group	Base_lDDT	Base_Tm	Base_QS	R1_lDDT	R1_TM	R1_QS	R3_lDDT	R3_TM	R3_QS	R6_lDDT	R6_TM	R6_QS	R12_lDDT	R12_TM-	R12_QS
H1045	AF2M-MSA	0.8742	0.94927	0.97	0.8815	0.94909	0.97	0.8806	0.94912	0.97	0.8797	0.9486	0.97	0.8804	0.94779	0.97
H1065	AF2M-MSA	0.9114	0.97252	0.92	0.9153	0.97046	0.91	0.9159	0.97184	0.91	0.916	0.97196	0.92	0.9161	0.9721	0.92
H1072	AF2M-MSA	0.7616	0.78298	0.79	0.7617	0.80674	0.82	0.7534	0.77881	0.78	0.7542	0.78988	0.78	0.7548	0.78678	0.78
T1032	AF2M-MSA	0.6903	0.69637	0.82	0.6719	0.7038	0.82	0.6714	0.70364	0.82	0.6708	0.70611	0.82	0.6709	0.70384	0.82
T1054	AF2M-MSA	0.5338	0.46892	0	0.6729	0.46242	0	0.6663	0.46288	0	0.6709	0.46291	0	0.672	0.46312	0
T1070	AF2M-MSA	0.5618	0.54822	0.09	0.5794	0.54973	0.1	0.5823	0.54936	0.11	0.7577	0.54926	0.12	0.5794	0.54927	0.11
T1073	AF2M-MSA	0.5986	0.30144	0	0.0899	0.30779	0	0.2908	0.30818	0	0.5978	0.36487	0	0.5869	0.27121	0
T1078	AF2M-MSA	0.7149	0.59815	0.02	0.8734	0.96779	0.84	0.8778	0.97032	0.83	0.8781	0.97006	0.84	0.8797	0.97156	0.84
T1083	AF2M-MSA	0.8241	0.90093	0.89	0.8436	0.90928	0.88	0.843	0.90799	0.88	0.8437	0.90834	0.88	0.8429	0.90764	0.88
T1084	AF2M-MSA	0.5843	0.6453	0.06	0.5059	0.8423	0.03	0.5061	0.84365	0.08	0.2059	0.852	0.01	0.5885	0.85356	0.02
H1045	AF2M-SingleSeq	0.2428	0.28591	0.07	0.1315	0.27903	0	0.2133	0.26856	0	0.2199	0.27281	0	0.2178	0.24783	0
H1065	AF2M-SingleSeq	0.3764	0.49303	0.02	0.3483	0.48146	0.02	0.3504	0.48756	0.02	0.3504	0.48924	0.02	0.3448	0.48944	0.01
H1072	AF2M-SingleSeq	0.2373	0.33465	0.17	0.4529	0.39146	0.24	0.7298	0.88703	0.75	0.7404	0.88638	0.76	0.7168	0.88807	0.74
T1032	AF2M-SingleSeq	0.2065	0.16496	0	0.1709	0.26459	0	0.071	0.26805	0	0.0708	0.26749	0	0.0684	0.26338	0
T1054	AF2M-SingleSeq	0.2815	0.25325	0.01	0.088	0.31607	0	0.1841	0.32831	0	0.2588	0.28748	0.03	0.2487	0.28768	0.03
T1070	AF2M-SingleSeq	0.222	0.1463	0	0.0357	0.2046	0	0.0586	0.19999	0	0.0561	0.1956	0	0.0566	0.20109	0

T1073	AF2M-SingleSeq	0.4184	0.2514	0	0.0019	0.41818	0	0.0002	0.32082	0	0.0006	0.33985	0	0.0005	0.35228	0
T1078	AF2M-SingleSeq	0.2119	0.38906	0.03	0.0726	0.40806	0		0.41676		0.1526	0.41072	0.01	0.1503	0.39622	0.02
T1083	AF2M-SingleSeq	0.5246	0.52756	0.31	0.4608	0.46824	0		0.526		0.5295	0.53289	0.31	0.5304	0.53116	0.31
T1084	AF2M-SingleSeq	0.824	0.90504	0.86	0.825	0.91626	0.9		0.91644		0.8425	0.9131	0.91	0.8441	0.91369	0.91

Appendix 7

A short analysis of ProQDock versus VoroMQA as a single model method.

The Voronoi tessellation-based model quality assessment program VoroMQA produces a single score between 0-1 with the following rating categories: <0.3 poor, 0.3 – 0.399 variable and >0.4 good with 5.5 likely to represent a native structure. The following is an investigation into the potential for VoroMQA to replace ProQDock as a single model prediction tool to improve the ModFOLDdock predicted consensus score.

There are two reasons for selecting VoroMQA and both centre around the fact that ProQDock and VoroMQA are single model scores. Firstly, this means they are easily comparable across different runs of the program without requiring recreation of the same model population every time (as is necessary with a clustering score) and, secondly, that in cases where only a few models exist and clustering routines understandably drop in accuracy accordingly, it is vital that a single model approach be retained as its accuracy should be maintained. VoroMQA would therefore represent a like-for-like replacement for ProQDock in this sense. The reason for the replacement of ProQDock is that, despite being a 0-1 score, there are occasions where the scores have ranged either greater than 1.0 or less than 0.0. Anecdotally it has also been noticed that ProQDock scores tend to be more extreme than others within ModFOLDdock and therefore may be disproportionally influencing the final Consensus score.

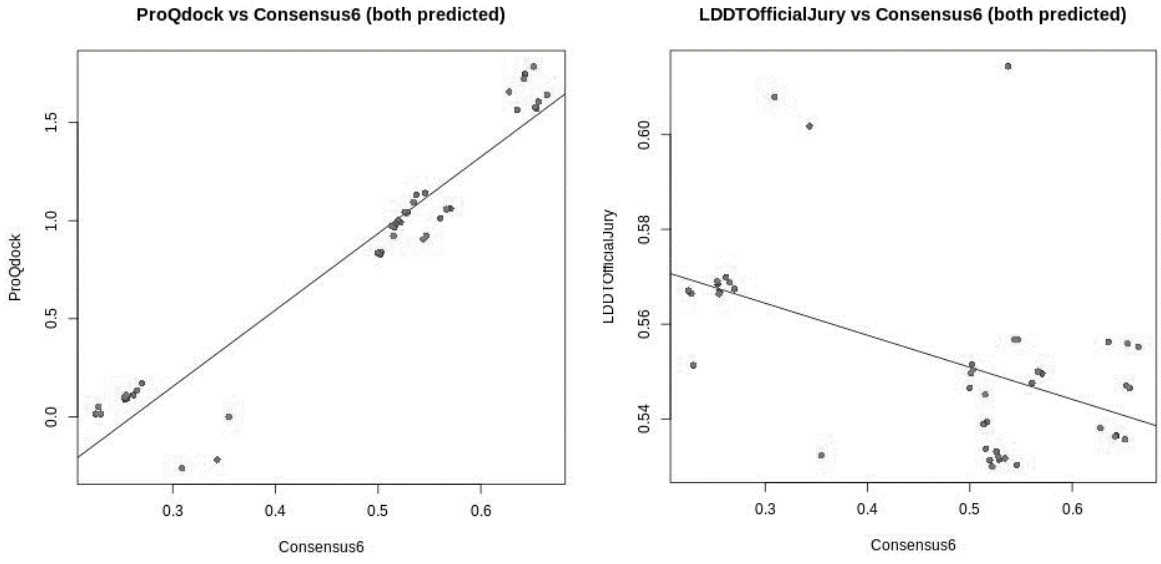


Figure S7.1. Scatter plots of an unweighted ModFOLDdock predicted consensus score versus a predicted ProQDock score (left) and a predicted lDDT score (right) for three randomly selected targets (T0965, T0966 and T1016).

To illustrate this point the left hand graph in Figure S7.1 shows how the predicted ProQDock values for three randomly selected targets correlate well with the predicted consensus score and appear to outweigh the contribution of lDDT score (right), for example. To further

investigate the ProQDock influence a range of partial consensus scores were calculated and compared to the full Consensus6 score (the mean of all six ModFOLDdock scores) and a calculated Consensus5 score (omitting ProQDock). These were:

Consensus5 – all ModFOLDdock scores, omitting only ProQDock;

Consensus4 - omitting both ProQDock and DockQJury;

Consensus3a - ModFOLDIA, QSscoreOfficialJury and LDDT score only;

Consensus3b - ModFOLDIA, QSscoreJury and LDDT score only;

Consensus2a - QSscoreJury and LDDT score only;

Consensus2b – ModFOLDIA score and LDDT score only.

Table S7.1 shows the Pearson and Spearman correlation coefficients calculated between partial consensus scores and the full consensus6 score. Table S7.2 shows similar data calculated with respect to the consensus5 score.

Table S7.1 Pearson and Spearman-rank correlation coefficients calculated between the consensus6 score and all other consensus scores for the three chosen targets.

Score (x).	Score (y).	Pearson correlation	Spearman correlation
Consensus6	Consensus5	0.874	0.693
Consensus6	Consensus4	0.864	0.669
Consensus6	Consensus3a	0.868	0.661
Consensus6	Consensus3b	0.853	0.660
Consensus6	Consensus2a	0.678	0.604
Consensus6	Consensus2b	0.854	0.654

Table S7.2 Pearson and Spearman-rank correlations calculated with respect to the consensus5 score (ProQDock removed) using the same targets as Table S7.1.

Score (x).	Score (y).	Pearson correlation	Spearman correlation
Consensus5	Consensus4	0.998	0.983
Consensus5	Consensus3a	0.994	0.953
Consensus5	Consensus3b	0.996	0.972
Consensus5	Consensus2a	0.888	0.898
Consensus5	Consensus2b	0.989	0.936

Tables S7.1 and S7.2 show that both Pearson and Spearman coefficients improve when ProQDock is removed, suggesting that a better agreement between all other individual predicted scores exists, giving an initial rationale for further investigation.

To assess the relative agreement between predicted scores, ProQDock and VoroMQA can be compared to a calculated mean observed score as this is likely to represent true model quality more accuracy. For this analysis a total of 96 models across the 16 CASP13 targets listed in Figures 4.1, 4.2 and 4.3 were used.

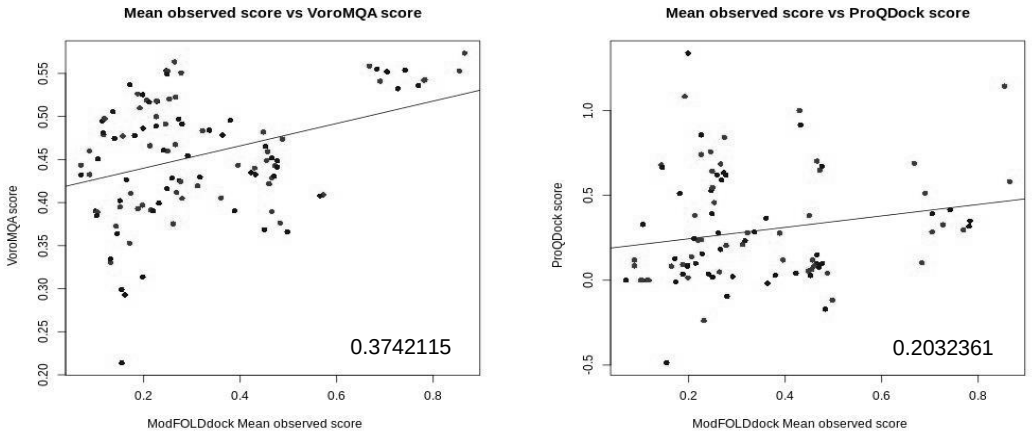


Figure S7.2 Scatter plots between calculated observed mean and VoroMQA score (left) and calculated observed mean and ProQDock score (right). Values shown are Pearson coefficients.

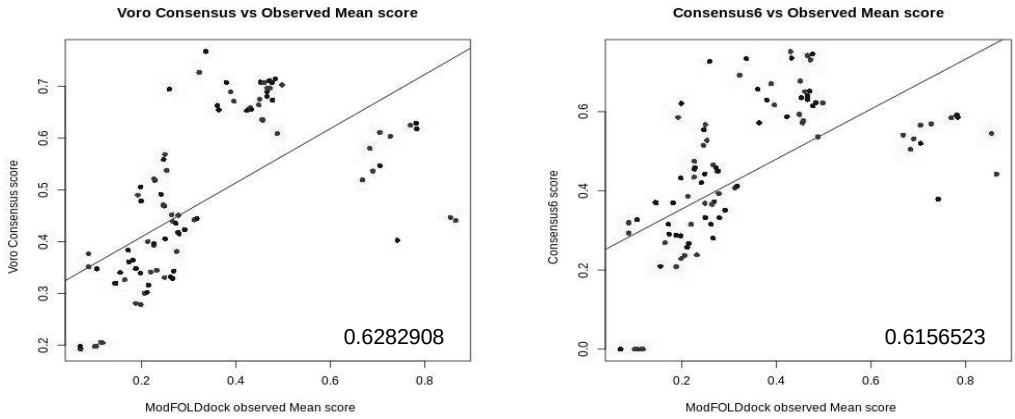


Figure S7.3 Scatter plots between observed mean score versus the consensus6 score calculated with VoroMQA score (left) and the consensus6 score calculated with ProQDock score (right). Values shown are Pearson coefficients.

Table S7.3 Pearson correlations coefficients between individual observed scores and predicted VoroMQA score and ProQDock scores.

Score	VoroMQA correlation	ProQDock correlation
Mean	0.37	0.20
IA score	0.21	0.04
DockQ	0.34	0.27
QS Score	0.16	0.11
QS Official	0.35	0.20
lDDT	0.61	0.26

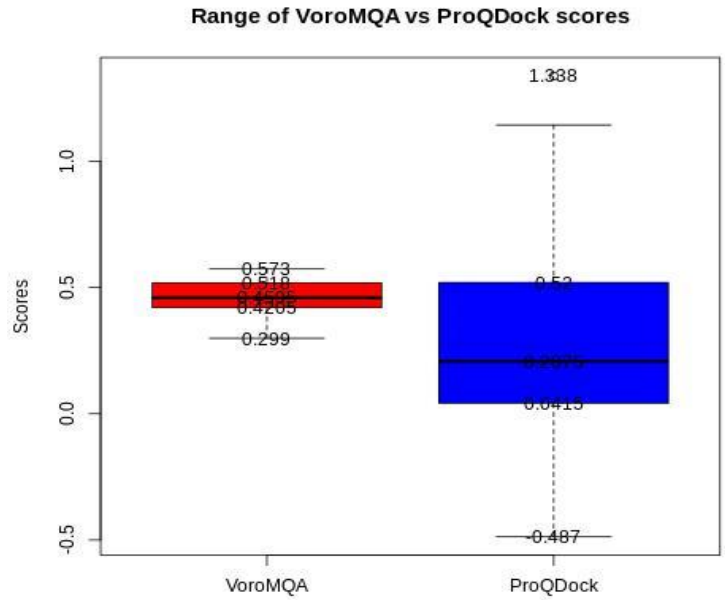


Figure S7.4 A box plot of predicted VoroMQA scores (left) and ProQDock scores (right) for CASP13 multimers. The minimum ProQDock score is -0.487 and the maximum score is 1.388, which are both outside of the 0-1 range.

Figure S7.2 shows no discernible difference between the correlations achieved with VoroMQA and ProQDock against the mean observed score. Similarly, the data in Figure S7.3 show that there is no clear difference between a consensus score calculated with VoroMQA and one calculated with ProQDock when plotted against the observed mean score. However, the data in Table 7.3 show that the VoroMQA score is more closely correlated with individual observed scores than is ProQDock. Additionally, Figure S7.4 shows that ProQDock produces a score with a much larger range (-0.487 to +1.388) than VoroMQA (0.299 to 0.573), meaning that the ProQDock contribution to the consensus score is likely to be both greater and more variable than other scores.

In conclusion, although VoroMQA score has not been clearly demonstrated to be a more accurate single-model score than ProQDock with this dataset, the lower variability in range suggests that it is likely to be a more reliable contributor to a consensus score. Additionally, VoroMQA score correlates slightly better with individual observed scores and is at least an order of magnitude quicker at calculating the score than ProQDock. From this initial data, the best conclusion that can be drawn is that VoroMQA is unlikely to lead to a decrease in accuracy of the calculated consensus score. A larger study with increased numbers and variability in models may produce more informative data.

Appendix 8

Full list of targets in each neural network training and testing dataset.

Training_set1 is T0960 T0961 T0963 T0965 T0970 T0973 T0976 T0979 T0981 T0983 T0984 T0985 T0995 T0996 T0998 T1000 T1001 T1004 T1006 T1010 T1018 T1032 T1034 T1061 T1062 T1070 T1078 T1080 T1084. Testing_set1 is T0966 T0977 T0989 T0991 T0997 T0999 T1003 T1009 T1016 T1020 T1038 T1048 T1054 T1083 T1087.

Training_set2 is T0960 T0961 T0966 T0970 T0973 T0977 T0981 T0985 T0989 T0991 T0996 T0997 T0999 T1000 T1003 T1004 T1006 T1009 T1010 T1016 T1020 T1032 T1034 T1038 T1048 T1054 T1080 T1083 T1087. Testing_set2 is T0963 T0965 T0976 T0979 T0983 T0984 T0995 T0998 T1001 T1018 T1061 T1062 T1070 T1078 T1084.

Training_set3 is T0963 T0965 T0966 T0976 T0977 T0979 T0983 T0984 T0989 T0991 T0995 T0997 T0998 T0999 T1001 T1003 T1009 T1016 T1018 T1020 T1038 T1048 T1054 T1061 T1062 T1070 T1078 T1083 T1084 T1087. Testing_set3 is T0960 T0961 T0970 T0973 T0981 T0985 T0996 T1000 T1004 T1006 T1010 T1032 T1034 T1080,

Appendix 9

Per-target top-rank comparisons by summed observed score.

Table S9.1 Per-target top-rank comparisons by summed observed scores. Used to create Chapter 4, Table 4.2. Cumulative observed scores for models top-ranked by ModFOLDdock component scores.

Model	method	predicted score	IAscore	DockQ	QSscore Calc	QSscore Official	lDDT Official	QS-glob	F1	oligo-lDDT	Jaccard Coeff.	TM-score	local	Global	Total	Obs sum
H1036TS403_4	QSscoreOfficialJury	0.436142	0.888154	0.315667	0.804396	0.711567	0.755689	0.712	68.3	0.756	0.7	0.712	0.6915	0.734	0.71275	76.79372
H1036TS403_4	lDDTOfficialJury	0.585221	0.888154	0.315667	0.804396	0.711567	0.755689	0.712	68.3	0.756	0.7	0.712	0.6915	0.734	0.71275	76.79372
H1036TS191_3	consensus	0.544371	0.375422	0.001	0.285714	0.702524	0.750596	0.703	68.9	0.751	0.72	0.702	0.7045	0.7265	0.7155	76.03776
H1036TS336_2	VoroMQA	0.674968	0.377688	0.001833	0.278119	0.635775	0.737535	0.642	68.2	0.737	0.7	0.702	0.691	0.7195	0.70525	75.1277
H1036TS018_4	QSscoreJury	0.590578	0.326421	0.000917	0.259341	0.659856	0.689083	0.66	61.5	0.689	0.65	0.701	0.6325	0.695	0.66375	68.12687
H1036TS221_1	ModFOLDIA	0.936138	0.376363	0.000917	0.274424	0.595587	0.599698	0.541	53.5	0.582	0.64	0.634	0.5875	0.608	0.59775	59.53724
H1036TS221_2	DockQJury	0.195891	0.39396	0.000917	0.285088	0.590039	0.601442	0.51	53.2	0.579	0.64	0.638	0.586	0.6085	0.59725	59.2302
H1045TS288_3	ModFOLDIA	0.827475	0.935895	0.551	0.8	0.906336	0.733067	0.818	71.5	0.706	0.69	0.835	0.7025	0.7705	0.7365	80.6848
H1045TS288_2	DockQJury	0.259323	0.918805	0.695	0.847826	0.929725	0.738588	0.624	78.4	0.713	0.78	0.889	0.782	0.801	0.7915	87.91044
H1045TS177_3	QSscoreJury	0.547336	0.483834	0.143	0.465116	0.510624	0.645535	0.511	29.3	0.646	0.45	0.676	0.3715	0.661	0.51625	35.37986
H1045TS298_4	QSscoreOfficialJury	0.368645	0.94704	0.622	0.911111	0.904321	0.719356	0.904	80.4	0.719	0.87	0.869	0.837	0.794	0.8155	90.31233
H1045TS217_5	lDDTOfficialJury	0.683542	0.070839	0.005	0	0	0.773455	0	0	0.773	0	0.482	0	0.6275	0.31375	3.045544
H1045TS288_2	VoroMQA	0.668059	0.918805	0.695	0.847826	0.929725	0.738588	0.624	78.4	0.713	0.78	0.889	0.782	0.801	0.7915	87.91044
H1045TS477_4	CDAscore	0.882939	0.874816	0.554	0.744186	0.8871	0.68006	0.887	68.7	0.68	0.65	0.869	0.6685	0.7745	0.7215	77.69066
H1045TS298_3	consensus	0.582653	0.967714	0.58	0.886364	0.906374	0.717168	0.906	76.1	0.717	0.81	0.862	0.7855	0.7895	0.7875	85.81512
H1047TS062_3	ModFOLDIA	0.605809	0.00287	0.003	0.007622	0	0.161766	0	0	0.606	0.01	0.155	0.005	0.3805	0.19275	1.524508
H1047TS323_1	DockQJury	0.03125	0.005518	0.005	0.028286	0	0.155111	0	0	0.59	0.04	0.328	0.02	0.459	0.2395	1.870416
H1047TS217_3	QSscoreJury	0.235788	0.030685	0.002	0	0	0.651188	0	0	0.65	0	0.152	0	0.401	0.2005	2.087373
H1047TS323_1	QSscoreOfficialJury	0.035118	0.005518	0.005	0.028286	0	0.155111	0	0	0.59	0.04	0.328	0.02	0.459	0.2395	1.870416
H1047TS298_3	lDDTOfficialJury	0.478137	0.082832	0.004	0.012048	0	0.648291	0	0	0.648	0.01	0.373	0.005	0.5105	0.25775	2.551421
H1047TS029_5	VoroMQA	0.627563	0.000124	0.003	0.002292	0	0.030649	0	0	0.611	0	0.372	0	0.4915	0.24575	1.756315
H1047TS018_3	CDAscore	0.813968	0.217757	0.007	0.019608	0	0.615743	0	0	0.616	0.01	0.351	0.005	0.4835	0.24425	2.569858
H1047TS323_1	consensus	0.380102	0.005518	0.005	0.028286	0	0.155111	0	0	0.59	0.04	0.328	0.02	0.459	0.2395	1.870416
H1065TS029_1	ModFOLDIA	0.867387	0.817291	0.046	0.666667	0.103135	0.624287	0.103	4.1	0.624	0.57	0.611	0.3055	0.6175	0.4615	9.649879
H1065TS192_1	DockQJury	0.043515	0.852361	0.497	0.75	0.71669	0.692479	0.632	47.7	0.672	0.65	0.867	0.5635	0.7695	0.6665	56.02903
H1065TS403_1	QSscoreJury	0.45057	0.429595	0.328	0.433333	0.540531	0.696362	0.541	39.7	0.696	0.43	0.792	0.4135	0.744	0.57875	46.32307
H1065TS403_1	QSscoreOfficialJury	0.093168	0.429595	0.328	0.433333	0.540531	0.696362	0.541	39.7	0.696	0.43	0.792	0.4135	0.744	0.57875	46.32307
H1065TS375_2	lDDTOfficialJury	0.577558	0.801337	0.399	0.65	0.618291	0.686356	0.562	40.7	0.679	0.56	0.8	0.4835	0.7395	0.6115	48.29048
H1065TS193_2	VoroMQA	0.685827	0.836294	0.08	0.433333	0.008418	0.632559	0.008	0	0.633	0.3	0.584	0.15	0.6085	0.37925	4.653354
H1065TS018_1	CDAscore	0.909294	0.52779	0.041	0.416667	0.125663	0.607921	0.126	7.6	0.608	0.35	0.536	0.213	0.572	0.3925	12.11654
H1065TS375_2	consensus	0.491271	0.801337	0.399	0.65	0.618291	0.686356	0.562	40.7	0.679	0.56	0.8	0.4835	0.7395	0.6115	48.29048
H1072TS029_4	ModFOLDIA	0.96409	0.840724	0.025833	0.663158	0.012158	0.393542	0.012	4.2	0.394	0.21	0.325	0.126	0.3595	0.24275	7.804665
H1072TS055_3	DockQJury	0.052395	0.880054	0.009667	0.631579	0.288027	0.481315	0.288	22.1	0.481	0.31	0.408	0.2655	0.4445	0.355	26.94264

H1072TS451_5	QSscoreJury	0.656695	0.368273	0.0125	0.321053	0.008352	0.350496	0.008	1.7	0.35	0.11	0.291	0.0635	0.3205	0.192	4.095673
H1072TS055_3	QSscoreOfficialJury	0.099584	0.880054	0.009667	0.631579	0.288027	0.481315	0.288	22.1	0.481	0.31	0.408	0.2655	0.4445	0.355	26.94264
H1072TS336_2	lDDTOfficialJury	0.40573	0.887155	0.0345	0.788945	0.270489	0.490034	0.27	20	0.49	0.36	0.391	0.28	0.4405	0.36025	25.06287
H1072TS403_5	VoroMQA	0.631099	0.179369	0.002667	0.364407	0.015288	0.257962	0.015	2	0.258	0.24	0.348	0.13	0.303	0.2165	4.330192
H1072TS451_5	CDAscore	0.979508	0.368273	0.0125	0.321053	0.008352	0.350496	0.008	1.7	0.35	0.11	0.291	0.0635	0.3205	0.192	4.095673
H1072TS029_1	consensus	0.509477	0.852835	0.011667	0.689474	0.273271	0.471468	0.273	21.5	0.464	0.36	0.377	0.2875	0.4205	0.354	26.33471
T1032TS018_2o	ModFOLDIA	0.877046	0.733537	0.266	0.576923	0.448143	0.468454	0.448	39.6	0.468	0.48	0.606	0.438	0.537	0.4875	45.55756
T1032TS029_1o	DockQJury	0.140868	0.866015	0.319	0.662651	0.490257	0.416794	0.49	47.5	0.417	0.51	0.624	0.4925	0.5205	0.5065	53.81522
T1032TS403_4o	QSscoreJury	0.607873	0.625015	0.289	0.512195	0.595944	0.548577	0.596	39.5	0.549	0.46	0.676	0.4275	0.6125	0.52	45.91173
T1032TS055_5o	QSscoreOfficialJury	0.346661	0.869346	0.311	0.695122	0.617453	0.459678	0.617	49.8	0.458	0.58	0.663	0.539	0.5605	0.54975	56.71985
T1032TS403_1o	lDDTOfficialJury	0.527605	0.807074	0.29	0.573171	0.537606	0.542758	0.538	37.6	0.543	0.46	0.688	0.418	0.6155	0.51675	44.12986
T1032TS193_1o	VoroMQA	0.629243	0.185468	0.389	0.30198	0.323403	0.524503	0.323	34.1	0.525	0.27	0.668	0.3055	0.5965	0.451	38.96335
T1032TS062_2o	CDAscore	0.874542	0.239736	0.064	0.207317	0	0.456007	0	0	0.447	0.12	0.424	0.06	0.4355	0.24775	2.701311
T1032TS055_2o	consensus	0.508739	0.734103	0.239	0.55914	0.396982	0.523653	0.397	30.2	0.524	0.42	0.626	0.361	0.575	0.468	36.02388
T1034TS298_4o	ModFOLDIA	0.877023	0.781745	0.014	0.365714	0	0.628997	0	0	0.629	0.15	0.255	0.075	0.442	0.2585	3.599956
T1034TS278_3o	DockQJury	0.062049	0.266635	0.005333	0.173913	0.002725	0.271955	0.003	0	0.272	0.08	0.251	0.04	0.2615	0.15075	1.778811
T1034TS278_1o	QSscoreJury	0.420891	0.18021	0.004333	0.118012	0.002134	0.282398	0.002	0	0.282	0.07	0.252	0.035	0.267	0.151	1.646088
T1034TS336_4o	QSscoreOfficialJury	0.062853	0.856235	0.0405	0.440994	0.038755	0.576263	0.039	5.5	0.57	0.16	0.289	0.1075	0.4295	0.2685	9.316246
T1034TS403_4o	lDDTOfficialJury	0.566801	0.602784	0.007	0.130435	0	0.63489	0	0	0.635	0.03	0.229	0.015	0.432	0.2235	2.939609
T1034TS403_5o	VoroMQA	0.722309	0.539521	0.005667	0.136646	0	0.62443	0	0	0.624	0.01	0.229	0.005	0.4265	0.21575	2.816514
T1034TS298_1o	CDAscore	0.873061	0.496648	0.006333	0.006211	0	0.608733	0	0	0.609	0	0.235	0	0.422	0.211	2.594926
T1034TS298_4o	consensus	0.474863	0.781745	0.014	0.365714	0	0.628997	0	0	0.629	0.15	0.255	0.075	0.442	0.2585	3.599956
T1038TS288_2o	ModFOLDIA	0.753016	0.691331	0.043	0.34375	0.059225	0.350848	0.059	5.1	0.351	0.24	0.165	0.1455	0.258	0.20175	8.008405
T1038TS055_3o	DockQJury	0.017068	0.564309	0.015	0.081633	0	0.345438	0	0	0.345	0.04	0.244	0.02	0.2945	0.15725	2.10713
T1038TS173_2o	QSscoreJury	0.272994	0.154355	0.017	0	0	0.352876	0	0	0.353	0	0.244	0	0.2985	0.14925	1.56898
T1038TS055_3o	QSscoreOfficialJury	0.072302	0.564309	0.015	0.081633	0	0.345438	0	0	0.345	0.04	0.244	0.02	0.2945	0.15725	2.10713
T1038TS173_2o	lDDTOfficialJury	0.415824	0.154355	0.017	0	0	0.352876	0	0	0.353	0	0.244	0	0.2985	0.14925	1.56898
T1038TS029_2o	VoroMQA	0.619228	0.456601	0.008	0	0	0.21605	0	0	0.216	0	0.175	0	0.1955	0.09775	1.364901
T1038TS491_4o	CDAscore	0.727472	0.228225	0.005	0	0	0.142094	0	0	0.142	0	0.139	0	0.1405	0.07025	0.86707
T1038TS029_1o	consensus	0.370974	0.670027	0.015	0.196721	0.005929	0.347871	0.006	1.4	0.348	0.12	0.248	0.067	0.298	0.1825	3.905047
T1048TS491_4o	ModFOLDIA	0.916972	0.696915	0.0365	0.656863	0.037263	0.277298	0.037	4.1	0.277	0.25	0.411	0.1455	0.344	0.24475	7.514089
T1048TS029_3o	DockQJury	0.126001	0.951356	0.142333	0.722222	0.111374	0.340394	0.114	11.5	0.325	0.25	0.371	0.1825	0.348	0.26525	15.62343
T1048TS403_5o	QSscoreJury	0.684588	0.915932	0.213	0.849162	0.310906	0.465348	0.311	24.2	0.465	0.61	0.729	0.426	0.597	0.5115	30.60385
T1048TS336_1o	QSscoreOfficialJury	0.183577	0.586138	0.125333	0.577586	0.086821	0.378914	0.087	9.2	0.379	0.18	0.416	0.136	0.3975	0.26675	12.81704
T1048TS336_1o	lDDTOfficialJury	0.449982	0.586138	0.125333	0.577586	0.086821	0.378914	0.087	9.2	0.379	0.18	0.416	0.136	0.3975	0.26675	12.81704
T1048TS029_3o	VoroMQA	0.60206	0.951356	0.142333	0.722222	0.111374	0.340394	0.114	11.5	0.325	0.25	0.371	0.1825	0.348	0.26525	15.62343
T1048TS491_4o	CDAscore	0.928728	0.696915	0.0365	0.656863	0.037263	0.277298	0.037	4.1	0.277	0.25	0.411	0.1455	0.344	0.24475	7.514089
T1048TS336_1o	consensus	0.496773	0.586138	0.125333	0.577586	0.086821	0.378914	0.087	9.2	0.379	0.18	0.416	0.136	0.3975	0.26675	12.81704
T1054TS071_1o	ModFOLDIA	0.800832	0.634481	0.03	0.261364	0.019021	0.478512	0.019	2.2	0.479	0.18	0.306	0.101	0.3925	0.24675	5.347627
T1054TS477_2o	DockQJury	0.016371	0.50209	0.016	0.25	0.017895	0.238881	0.018	1.1	0.239	0.17	0.225	0.0905	0.232	0.16125	3.260616
T1054TS071_4o	QSscoreJury	0.286667	0.299806	0.01	0.045455	0.00134	0.459168	0.001	0	0.459	0.03	0.261	0.015	0.36	0.1875	2.129268
T1054TS155_3o	QSscoreOfficialJury	0.060389	0.428122	0.018	0.170455	0.03022	0.34375	0.03	1.3	0.344	0.13	0.28	0.0715	0.312	0.19175	3.649795

T1054TS193_5o	lDDTOfficialJury	0.517932	0.44815	0.013	0.159091	0	0.591884	0	0	0.592	0.12	0.443	0.06	0.5175	0.28875	3.233375
T1054TS029_1o	VoroMQA	0.683178	0.838984	0.03	0.386364	0.045859	0.523084	0.046	1.9	0.523	0.25	0.34	0.1345	0.4315	0.283	5.732291
T1054TS343_5o	CDAscore	0.896283	0.431243	0.019	0.068182	0	0.552377	0	0	0.552	0.04	0.425	0.02	0.4885	0.25425	2.850552
T1054TS403_1o	consensus	0.428236	0.585018	0.013	0.022727	0	0.607308	0	0	0.607	0.01	0.44	0.005	0.5235	0.26425	3.077803
T1062TS451_2o	ModFOLDIA	0.966394	0.796361	0.06	0.692308	0.088247	0.366607	0.088	12.5	0.367	0.26	0.309	0.1925	0.338	0.26525	16.32327
T1062TS375_5o	DockQJury	0.05188	0.867152	0.093333	0.769231	0.09625	0.379098	0.057	10.7	0.298	0.27	0.306	0.1885	0.302	0.24525	14.57181
T1062TS029_5o	QSscoreJury	0.806536	0.114141	0.038	0.115385	0.065213	0.155578	0.065	6.8	0.156	0.07	0.19	0.069	0.173	0.121	8.132317
T1062TS403_1o	QSscoreOfficialJury	0.236498	0.872957	0.083333	0.807692	0.097339	0.382572	0.097	9.4	0.383	0.26	0.308	0.177	0.3455	0.26125	13.47564
T1062TS062_5o	lDDTOfficialJury	0.477273	0.865886	0.096	0.730769	0.078981	0.376321	0.079	11.9	0.376	0.25	0.317	0.1845	0.3465	0.2655	15.86646
T1062TS029_3o	VoroMQA	0.698997	0.833768	0.152	0.730769	0.535384	0.419598	0.535	43.3	0.42	0.55	0.478	0.4915	0.449	0.47025	49.36527
T1062TS288_4o	CDAscore	0.901135	0.376197	0.033333	0.333333	0.03688	0.360429	0.037	0	0.36	0.16	0.309	0.08	0.3345	0.20725	2.627923
T1062TS451_2o	consensus	0.543647	0.796361	0.06	0.692308	0.088247	0.366607	0.088	12.5	0.367	0.26	0.309	0.1925	0.338	0.26525	16.32327
T1070TS360_2o	ModFOLDIA	0.763751	0.472349	0.007667	0.128415	0.002273	0.251581	0.002	0	0.251	0.03	0.243	0.015	0.247	0.131	1.781285
T1070TS155_4o	DockQJury	0.034772	0.207268	0.032	0.150273	0.055088	0.104761	0.055	1.8	0.105	0.06	0.204	0.039	0.1545	0.09675	3.06364
T1070TS155_2o	QSscoreJury	0.348167	0.232155	0.058333	0.112022	0.061075	0.182231	0.061	2	0.182	0.07	0.29	0.045	0.236	0.1405	3.670316
T1070TS173_4o	QSscoreOfficialJury	0.086956	0.516053	0.023667	0.352459	0.103594	0.210385	0.104	2.5	0.21	0.17	0.287	0.0975	0.2485	0.173	4.996159
T1070TS062_2o	lDDTOfficialJury	0.399771	0.140748	0.006667	0.098361	0	0.405659	0	0	0.406	0.02	0.163	0.01	0.2845	0.14725	1.682184
T1070TS193_1o	VoroMQA	0.664489	0.750588	0.007667	0.401639	0.082869	0.349272	0.083	5.8	0.349	0.18	0.177	0.119	0.263	0.191	8.754036
T1070TS099_3o	CDAscore	0.820657	0.096606	0.007333	0.027322	0	0.423508	0	0	0.424	0.01	0.181	0.005	0.3025	0.15375	1.631019
T1070TS221_1o	consensus	0.405337	0.418661	0.016667	0.237705	0.104024	0.435398	0.104	3.1	0.435	0.12	0.32	0.0755	0.3775	0.2265	5.970954
T1078TS343_2o	ModFOLDIA	0.884971	0.879712	0.029	0.549296	0.003849	0.542363	0.004	0	0.503	0.38	0.509	0.19	0.506	0.348	4.44422
T1078TS155_4o	DockQJury	0.028052	0.69152	0.042	0.366197	0.105355	0.400917	0.105	1.3	0.4	0.26	0.424	0.1365	0.412	0.27425	4.91774
T1078TS341_2o	QSscoreJury	0.492857	0.275209	0.021	0.225352	0	0.455837	0	0	0.448	0.21	0.458	0.105	0.453	0.279	2.930398
T1078TS029_1o	QSscoreOfficialJury	0.083671	0.542543	0.078	0.514563	0.161571	0.552705	0.162	6	0.552	0.44	0.496	0.25	0.524	0.387	10.66038
T1078TS451_2o	lDDTOfficialJury	0.495588	0.650297	0.024	0.408451	0.000326	0.555001	0	0	0.555	0.3	0.491	0.15	0.523	0.3365	3.993575
T1078TS029_4o	VoroMQA	0.709846	0.681954	0.145	0.422535	0.16387	0.558115	0.15	12.4	0.537	0.33	0.629	0.227	0.583	0.405	17.23247
T1078TS099_1o	CDAscore	0.862483	0.389903	0.029	0.366197	0	0.54563	0	0	0.545	0.3	0.529	0.15	0.537	0.3435	3.73523
T1078TS029_2o	consensus	0.472829	0.735909	0.057	0.43662	0.061399	0.54002	0.061	1.1	0.54	0.28	0.486	0.1455	0.513	0.32925	5.285698
T1083TS343_5o	ModFOLDIA	0.916158	0.708569	0.079	0.448718	0.028797	0.505065	0.016	1	0.46	0.33	0.453	0.17	0.4565	0.31325	4.9689
T1083TS029_4o	DockQJury	0.067	0.660058	0.531	0.628205	0.638186	0.552899	0.638	57	0.553	0.6	0.697	0.585	0.625	0.605	64.31335
T1083TS062_4o	QSscoreJury	0.593162	0.363217	0.018	0.320513	0	0.503681	0	0	0.504	0.3	0.533	0.15	0.5185	0.33425	3.545161
T1083TS403_2o	QSscoreOfficialJury	0.155181	0.854141	0.232	0.761905	0.46968	0.617292	0.47	15.1	0.617	0.65	0.59	0.4005	0.6035	0.502	21.86802
T1083TS403_5o	lDDTOfficialJury	0.566593	0.897379	0.149	0.679487	0.371674	0.605726	0.372	9.4	0.606	0.52	0.53	0.307	0.568	0.4375	15.44377
T1083TS403_3o	VoroMQA	0.592349	0.767266	0.228	0.755319	0.323609	0.60834	0.324	18.3	0.608	0.7	0.566	0.4415	0.587	0.51425	24.72328
T1083TS491_1o	CDAscore	0.94229	0.308305	0.015	0.205128	0	0.274788	0	0	0.275	0.18	0.369	0.09	0.322	0.206	2.245221
T1083TS071_2o	consensus	0.509299	0.752382	0.296	0.653846	0.549226	0.587014	0.549	31.9	0.586	0.58	0.655	0.4495	0.6205	0.535	38.71347
T1084TS055_2o	ModFOLDIA	0.946689	0.975269	0.462	0.690909	0.685708	0.648825	0.686	45.4	0.649	0.53	0.79	0.492	0.7195	0.60575	53.33496
T1084TS029_1o	DockQJury	0.164807	0.748833	0.751	0.680556	0.875724	0.765828	0.857	63.3	0.756	0.63	0.893	0.6315	0.8245	0.728	72.44194
T1084TS288_1o	QSscoreJury	0.620482	0.35643	0.058	0.290909	0	0.502626	0	0	0.496	0.27	0.596	0.135	0.546	0.3405	3.591465
T1084TS298_5o	QSscoreOfficialJury	0.239461	0.86197	0.604	0.781818	0.8707	0.754856	0.871	77.7	0.755	0.72	0.896	0.7485	0.8255	0.787	87.17634
T1084TS403_1o	lDDTOfficialJury	0.606947	0.84711	0.635	0.822581	0.892888	0.831829	0.893	81	0.832	0.77	0.917	0.79	0.8745	0.83225	90.93816
T1084TS403_2o	VoroMQA	0.650103	0.825112	0.507	0.745455	0.854123	0.758857	0.854	73	0.759	0.68	0.884	0.705	0.8215	0.76325	82.1573

T1084TS071_5o	CDAscore	0.938053	0.837262	0.089	0.654545	0.028895	0.496844	0.029	0	0.497	0.54	0.677	0.27	0.587	0.4285	5.135046
T1084TS298_5o	consensus	0.560929	0.86197	0.604	0.781818	0.8707	0.754856	0.871	77.7	0.755	0.72	0.896	0.7485	0.8255	0.787	87.17634
T1087TS177_4o	ModFOLDIA	0.926989	0.637445	0.293	0.542169	0.325078	0.479538	0.325	24.4	0.48	0.49	0.541	0.367	0.5105	0.43875	29.82948
T1087TS177_1o	DockQJury	0.06075	0.658947	0.406	0.518072	0.405164	0.526152	0.405	31.5	0.526	0.45	0.649	0.3825	0.5875	0.485	37.49933
T1087TS173_2o	QSscoreJury	0.558957	0.141323	0.073	0.13253	0.071864	0.477447	0.072	5.4	0.456	0.13	0.487	0.092	0.4715	0.28175	8.286413
T1087TS177_1o	QSscoreOfficialJury	0.127095	0.658947	0.406	0.518072	0.405164	0.526152	0.405	31.5	0.526	0.45	0.649	0.3825	0.5875	0.485	37.49933
T1087TS403_1o	lDDTOfficialJury	0.513801	0.780363	0.393	0.735294	0.444684	0.629435	0.445	36.2	0.629	0.68	0.786	0.521	0.7075	0.61425	43.56553
T1087TS029_2o	VoroMQA	0.636739	0.790714	0.322	0.566265	0.320614	0.505094	0.321	30.4	0.505	0.45	0.598	0.377	0.5515	0.46425	36.17144
T1087TS066_3o	CDAscore	0.922755	0.639529	0.029	0.530121	0.03695	0.475315	0.037	2.4	0.475	0.47	0.465	0.247	0.47	0.3585	6.633415
T1087TS193_1o	consensus	0.498275	0.647962	0.093	0.530121	0.293467	0.525697	0.293	21.2	0.526	0.47	0.536	0.341	0.531	0.436	26.42325

Appendix 10

Full versions of final all-against-all comparison tables described in stage 3.

Table S10.1 Data for Chapter 4, Table 4.3. Correlations between the observed global interface and fold scores and every combination of the 7 component scores, based on the CASP14 multimer data: A=ModFOLDIA, B=DockQJury, C=QSscoreJury, D=QSscoreOfficialJury, E=lDDTOfficialJury, F=voronota-js-voromqa, G=CDA-score. The top scores in each column are shown in bold. The combinations used for the ModFOLDdock fold and interface scores are highlighted in green.

Method combination	Interface			Fold		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
B+E	0.6221383	0.4662672	0.3370294	**0.897708**	**0.8895329**	0.7178826
D+E	0.7678932	0.6149145	0.451429	0.8886437	0.8864162	**0.7204588**
B+D+E+F	0.7370915	0.5618972	0.4084465	0.8755656	0.8648914	0.6910571
D+E+F	0.6796071	0.5390013	0.3894662	0.8748695	0.8658919	0.6912109
B+D+E	0.8155852	0.6325805	0.4671395	0.8738063	0.8812126	0.7138623
B+E+F	0.5398861	0.4028433	0.2887446	0.8507348	0.8403956	0.6561161
E	0.4398352	0.3730243	0.2678815	0.8503973	0.8587005	0.6726669
E+F	0.4053162	0.3287048	0.2352872	0.8024292	0.8084877	0.6153324
B+C+D+E+F	0.7941014	0.7417475	0.552793	0.7869698	0.7413835	0.564248
C+D+E+F	0.7561131	0.7344757	0.5432355	0.7773438	0.740087	0.5629397
A+B+D+E+F	0.7440054	0.7063684	0.5241025	0.7740935	0.7298914	0.5474411
B+C+E+F	0.6886405	0.682026	0.4928813	0.77257	0.7296966	0.5545109
B+D+F	0.82149	0.6114097	0.4479444	0.7606479	0.7381083	0.5590382
D+F	0.7698032	0.5881448	0.4268164	0.7599284	0.7393382	0.5587745
A+D+E+F	0.7021574	0.6944457	0.511267	0.7595768	0.7250281	0.5418011
B+C+D+E	0.8339838	0.7859983	0.592779	0.7462676	0.696275	0.520593
C+E+F	0.6134071	0.6547062	0.4648403	0.7417553	0.7143769	0.538481
A+B+E+F	0.6275222	0.629427	0.4553829	0.7412427	0.7066384	0.5248789
A+B+D+E	0.7705623	0.7432268	0.5556862	0.7353772	0.6776837	0.4981892
C+D+E	0.7973255	0.779981	0.5836362	0.7337487	0.6933543	0.5177601
B+C+E	0.7368397	0.7398483	0.5406193	0.7320624	0.6813849	0.5099464
A+D+E	0.7265591	0.7328193	0.5432948	0.7169642	0.671624	0.4915085
B+F	0.5660863	0.3635316	0.2592923	0.7137864	0.6852296	0.5128599
B+D+E+F+G	0.404772	0.3780916	0.2710916	0.7101008	0.7061886	0.553011
A+B+C+D+E+F	0.7635909	0.7739465	0.581026	0.7083849	0.6634288	0.4891847
A+E+F	0.5560334	0.5970161	0.4261682	0.7073137	0.6875881	0.5044643
A+B+E	0.6486699	0.6680471	0.4850158	0.6952704	0.6487127	0.4725049
B+C+D+E+F+G	0.5244234	0.5830563	0.412383	0.693544	0.6456508	0.4798443
A+C+D+E+F	0.7308335	0.7665216	0.5709852	0.6924358	0.6571411	0.4831878

A+B+D+E+F+G	0.5098543	0.5772571	0.4069296	0.6923436	0.6586154	0.4819218
C+E	0.651986	0.7111715	0.5072599	0.6921837	0.6585876	0.4896367
A+B+C+E+F	0.6820997	0.7336256	0.5382033	0.6785479	0.6422252	0.4708995
D+E+F+G	0.322943	0.336453	0.2426468	0.6655271	0.6797333	0.5371782
A+B+C+D+E	0.7767018	0.7949579	0.6008529	0.6636642	0.6161361	0.4468922
A+B+C+D+E+F+G	0.5752266	0.6696365	0.4772083	0.6628724	0.61441	0.4467918
A+D+E+F+G	0.4534166	0.5418677	0.3798655	0.6615978	0.6340081	0.4614429
C+D+E+F+G	0.4632582	0.5481547	0.3875863	0.6606479	0.6218252	0.4615751
B+D+E+G	0.3875972	0.3909207	0.2861465	0.657835	0.6568015	0.5137236
A+E	0.5664097	0.6327542	0.4510429	0.6517214	0.623084	0.4474086
A+B+D+F	0.7623274	0.7355312	0.5461741	0.6514716	0.5946109	0.4284181
A+C+E+F	0.6316307	0.7140192	0.5164196	0.650924	0.6241992	0.4542625
A+B+D+E+G	0.5031556	0.5929763	0.4203144	0.6474392	0.6030616	0.4331001
B+C+D+E+G	0.5184236	0.5958047	0.4257621	0.6438089	0.5959577	0.4371545
A+C+D+E	0.7419265	0.78715	0.5899385	0.6436479	0.6069291	0.4387054
B+D	**0.9005487**	0.8246907	0.6435966	0.6419381	0.5309702	0.3781203
A+C+D+E+F+G	0.5309632	0.6464253	0.4565751	0.6379261	0.5954371	0.4301802
B+C+D+F	0.8175272	0.7770315	0.5889212	0.637794	0.5740778	0.4180225
D	0.8904282	**0.8440979**	**0.6601409**	0.6263819	0.5468863	0.389032
A+B+C+E	0.692102	0.7551501	0.5565514	0.6252666	0.5871694	0.4235903
A+D+F	0.7117394	0.7213149	0.5301804	0.622492	0.5825022	0.4172562
A+B+C+D+E+G	0.5709968	0.6804326	0.4864792	0.6196486	0.5683172	0.406208
A+B+E+F+G	0.3599946	0.4367296	0.3031905	0.6129679	0.5757836	0.417113
A+D+E+G	0.4407902	0.5547524	0.3917973	0.6112724	0.5745229	0.4106077
C+D+F	0.7728359	0.7688163	0.5767856	0.6092083	0.5646974	0.4102714
B+C+E+F+G	0.3607219	0.4482979	0.3162082	0.6082477	0.5674795	0.4255933
C+D+E+G	0.4501244	0.5582474	0.4005675	0.6043174	0.5679783	0.4172322
A+B+C+E+F+G	0.4609644	0.5808734	0.4031719	0.6038874	0.5565473	0.3965046
D+E+G	0.2933906	0.3537209	0.2610584	0.6029764	0.6337003	0.5030273
B	0.8191334	0.6607223	0.508491	0.6028232	0.4526431	0.3333887
A+B+D+F+G	0.4890414	0.5849852	0.4139539	0.5929576	0.5377722	0.3810692
A+C+E	0.6362988	0.7327199	0.5304996	0.5909937	0.5620175	0.402464
A+C+D+E+G	0.5230199	0.6551545	0.4642964	0.5908642	0.5464175	0.3876725
B+D+F+G	0.354711	0.3933636	0.2905136	0.590493	0.5839326	0.4563542
A+B+C+D+F	0.7620261	0.790109	0.5947494	0.5876953	0.5479381	0.3910265
B+E+F+G	0.1851618	0.2087258	0.1507441	0.5831396	0.6225731	0.4928905
A+B+F	0.618582	0.6458525	0.4640224	0.5803044	0.5419518	0.3857262

B+C+D+F+G	0.5012339	0.5883603	0.4232738	0.5783218	0.5269261	0.3826374
F	0.2763438	0.1935796	0.1387928	0.5760358	0.5914943	0.4220089
B+C+F	0.693108	0.7096416	0.5203948	0.5736845	0.5263652	0.383877
A+C+E+F+G	0.4048945	0.5446632	0.3756973	0.5690548	0.5268085	0.3730967
A+E+F+G	0.2891742	0.391625	0.271775	0.5683542	0.5419194	0.3920983
A+B+C+D+F+G	0.5594844	0.674622	0.482897	0.5664158	0.5174481	0.3660899
A+B+D	0.768613	0.7525183	0.558776	0.5610124	0.4917038	0.3451411
A+C+D+F	0.7226791	0.7797477	0.5814859	0.5604365	0.5346216	0.380508
C+E+F+G	0.2814387	0.4114624	0.2892574	0.5582347	0.5390696	0.4061662
A+B+E+G	0.3378762	0.4456156	0.3136935	0.5549658	0.511629	0.3660532
A+B+C+E+G	0.4475483	0.5869245	0.410364	0.5516895	0.5016167	0.3515447
A+D+F+G	0.4182374	0.5407973	0.3815873	0.5494807	0.5019142	0.3543656
B+C+E+G	0.3358963	0.4615773	0.3322916	0.5424231	0.5047689	0.379338
A+C+D+F+G	0.5065592	0.6472278	0.4586581	0.5327191	0.491706	0.3455039
C+D+F+G	0.4227912	0.5436172	0.3943254	0.5296769	0.4914799	0.359338
A+B+C+F	0.6645752	0.741465	0.5427793	0.5288693	0.5045896	0.3578339
B+C+D	0.8269361	0.8091608	0.6240246	0.5266822	0.4474464	0.3143817
D+F+G	0.2409491	0.3563423	0.2642462	0.5200666	0.5632211	0.4457454
A+D	0.7113254	0.7359118	0.5398023	0.5183876	0.472099	0.3289444
A+F	0.5191631	0.5945452	0.416958	0.5173752	0.4959425	0.3466355
E+F+G	0.08656197	0.1623278	0.1190342	0.516321	0.5948857	0.4702613
A+B+D+G	0.4680578	0.5842217	0.4176302	0.5161636	0.4451738	0.3088823
A+C+E+G	0.3863691	0.5462825	0.3798923	0.5119406	0.4660067	0.3253892
A+B+C+D	0.7580323	0.7981737	0.6005767	0.5110592	0.4718043	0.3301004
B+E+G	0.1361889	0.2239605	0.1650852	0.5046949	0.5940821	0.4727034
A+E+G	0.2593274	0.3971522	0.2793416	0.5032234	0.4725844	0.338792
A+B+C+D+G	0.5443057	0.6752321	0.4848112	0.5008184	0.4516605	0.3133573
C+F	0.5830781	0.6636917	0.4721933	0.5005905	0.4783326	0.347197
B+C+D+G	0.4768018	0.5843225	0.4297474	0.4884084	0.4384612	0.3118733
A+B+C+F+G	0.4226695	0.5722815	0.4006841	0.4856252	0.4384016	0.3041032
A+C+F	0.600404	0.7118308	0.5111797	0.4840006	0.4705569	0.331071
C+E+G	0.2468374	0.4239733	0.303394	0.483393	0.4721439	0.3579832
A+B+F+G	0.3005459	0.423398	0.2983356	0.4808792	0.4279865	0.3033854
C+D	0.776152	0.8001199	0.6106079	0.4783409	0.430884	0.301257
B+D+G	0.3094441	0.4280989	0.3231873	0.4782347	0.476807	0.362241
A+C+D	0.7141838	0.7852169	0.5846252	0.4757161	0.4529065	0.3157709
A+D+G	0.3891704	0.5360537	0.3833224	0.4642765	0.4031501	0.2789366

A+C+D+G	0.4865639	0.6439374	0.4579553	0.46168	0.4198982	0.2893849
A+B	0.6071178	0.6558811	0.4686983	0.4534903	0.4125144	0.285917
B+C+F+G	0.2905314	0.4458545	0.3235596	0.4523077	0.4241992	0.318057
A+C+F+G	0.3548218	0.5249093	0.3653983	0.439687	0.3948	0.2728716
A+B+C	0.650793	0.7447487	0.5436193	0.4311374	0.4121513	0.2869405
C+D+G	0.3883645	0.5329598	0.3977753	0.429081	0.3939667	0.2836269
E+G	0.02468937	0.1761515	0.1302842	0.4251367	0.5724286	0.4517576
A+F+G	0.2113104	0.3688911	0.2583106	0.4196708	0.3812478	0.2700233
A+B+C+G	0.3957246	0.5623156	0.3984426	0.4068951	0.3560086	0.2435724
B+C	0.6909383	0.7484381	0.5628754	0.4061099	0.3721297	0.2628535
B+F+G	0.05367426	0.1978878	0.1437184	0.3952618	0.5245636	0.4094789
D+G	0.1774821	0.4080011	0.306333	0.3898023	0.4662864	0.3551107
A+B+G	0.2594555	0.4228055	0.3027759	0.3843279	0.32582	0.2272073
C+F+G	0.1884297	0.405049	0.2890516	0.3814986	0.385034	0.291479
A+C	0.5777453	0.7109297	0.5076395	0.3745237	0.3699924	0.2545747
A	0.4867195	0.5868364	0.4057687	0.3654596	0.3366946	0.2293649
A+C+G	0.3220312	0.5120821	0.3609473	0.3550122	0.307833	0.2099138
B+C+G	0.241169	0.4606827	0.3420289	0.3371482	0.3392478	0.250484
A+G	0.1615911	0.3670324	0.2590266	0.3145923	0.2748015	0.1898057
F+G	-0.07501003	0.1258351	0.0909146	0.2987003	0.4859893	0.3728236
C	0.5505007	0.6904874	0.5030818	0.2886947	0.2966009	0.2073917
C+G	0.128631	0.4234011	0.3054489	0.2560357	0.2986689	0.2202136
B+G	-0.03138041	0.2517043	0.181093	0.2454451	0.4397153	0.3312597
G	-0.1693019	0.1327867	0.09281258	0.1382419	0.3581771	0.2661346

Table S10.2 Data used for Chapter 4, Table 4.4. Cumulative observed global interface and fold scores of the top ranked models for every combination of the 7 component scores based on the CASP14 multimer data: A=ModFOLDIA, B=DockQJury, C=QSscoreJury, D=QSscoreOfficialJury, E=lDDTOfficialJury, F=voronota-js-voromqa, G=CDA-score. The top scores in each column are shown in bold. The ModFOLDdockR fold and interface score combinations are highlighted in green.

Method combination	Interface	Fold
C+E+F	4.962	**9.145**
B+D+G	5.2505	9.097
E+F	5.04	9.091
B+E+F	5.4545	9.0885
D+E+F	5.117	9.064
B+E+F+G	5.136	9.0625
C+D+E+G	5.006	9.0485

B+E	5.167	9.01
D+E	5.3215	9.003
B+C+D+E+G	5.196	8.9935
B+C+D+E	5.2155	8.985
A+B+C+D+E+F+G	5.0855	8.956
B+D+E	5.345	8.948
B+C+D+G	5.126	8.9285
B+D+E+F	5.1455	8.913
D+E+G	4.8725	8.9055
A+C+F	5.286	8.883
D+E+F+G	4.6635	8.8575
C+D+E	4.919	8.856
C+D+G	4.6205	8.8535
B+D+E+G	4.912	8.812
A+B+D+E+F+G	4.944	8.8085
E	4.0515	8.805
B+D+E+F+G	4.657	8.802
A+B+D+E+F	5.0655	8.797
A+B+C+D+F+G	5.2595	8.785
A+B+C+D+E+G	4.931	8.7825
A+C+D+E+G	4.931	8.7825
B+C+E+F	4.7155	8.7825
A+B+C+E	5.3425	8.7795
D	5.414	8.7745
A+B+C+D+E+F	5.408	8.7745
C+D+E+F	4.6355	8.774
D+G	4.6135	8.7625
A+C+D+E+F+G	4.9075	8.7585
A+B+D+F	5.3465	8.7565
A+B+D+F+G	4.905	8.7535
A+B+C+E+F	4.8465	8.7495
A+B+C+E+G	4.757	8.7145
A+B+C+E+F+G	4.7275	8.7075
B+D+F+G	4.8505	8.707
B+C+D+E+F	4.943	8.704
A+B+F+G	4.6465	8.6965
C+E+F+G	4.3265	8.6965

A+C+D+E	5.114	8.68
B+E+G	4.3635	8.6705
A+D+E+F	4.7285	8.67
B+C+E+G	4.339	8.67
E+F+G	4.349	8.6685
A+B+C+F+G	4.4715	8.6675
A+B+C+D+F	5.335	8.666
A+D+E+F+G	4.6845	8.6655
A+C+D+E+F	4.7645	8.6605
A+B+E+F+G	4.646	8.6535
A+B+E+F	4.5695	8.6525
A+C+F+G	4.7275	8.648
A+B+C+F	5.057	8.6475
A+D+F	5.207	8.6465
C+D+F+G	4.5735	8.646
A+B+C+D+E	5.1365	8.644
B+C+D+F+G	4.796	8.6435
A+C+E+F	4.6035	8.6425
C+D+E+F+G	4.5855	8.6385
B+C+D+F	5.4965	8.638
C+D+F	5.185	8.638
B+C+D+E+F+G	4.808	8.636
A+B+D+G	5.238	8.6305
A+D+E+G	4.7865	8.6305
A+B+D+E+G	4.776	8.6185
A+F+G	4.4245	8.615
A+C+E	4.87	8.6145
B+F+G	4.852	8.602
A+C+E+F+G	4.558	8.5985
C+E	4.2275	8.596
A+D+F+G	4.7225	8.5825
A+C+D	5.2855	8.5745
C+F+G	3.9805	8.574
B+C+E+F+G	4.559	8.572
A+D+E	4.9065	8.57
B+C+F+G	4.549	8.5665
D+F	5.4435	8.564

D+F+G	4.687	8.563
C+E+G	4.036	8.555
A+C+D+F+G	4.9095	8.5535
A+D+G	5.1505	8.5475
B+D	5.412	8.547
B+C+F	5.3875	8.5365
A+C	5.247	8.5345
B+C+G	4.3575	8.5335
A+B+E+G	4.365	8.5065
A+B+D+E	4.8725	8.493
A+B+C+D+G	5.4715	8.488
A+B+E	4.6945	8.4835
A+C+D+G	5.283	8.483
B+D+F	**5.6105**	8.479
A+C+E+G	4.394	8.4625
B+C+E	4.4885	8.4585
A+D	5.206	8.448
A+B+C+D	5.478	8.4325
A+C+D+F	4.8075	8.42
A+C+G	4.776	8.418
B+C+D	5.2495	8.4175
A+B+G	4.775	8.4065
A+E	4.6635	8.405
C+F	4.6335	8.403
F	4.8695	8.3865
A+B+F	4.592	8.3765
A+B+C	5.026	8.367
A+B+D	5.286	8.3585
A+F	4.522	8.3505
A+E+G	4.146	8.3485
A+E+F+G	4.2285	8.3415
A+B+C+G	5.02	8.3305
C+G	3.3645	8.315
C+D	4.6965	8.3065
B+F	5.449	8.2965
A+G	4.562	8.2915
A+E+F	4.0765	8.207

E+G	2.579	8.166
F+G	3.4355	8.112
A+B	4.8615	8.089
B+G	3.818	8.055
B+C	4.4975	7.964
B	5.048	7.8685
A	4.265	7.808
C	3.201	7.788
G	2.2295	7.0715

Appendix 11

Definitions of Pearson, Spearman and Kendall correlation coefficients.

The Pearson correlation coefficient **r** is a normalised version of covariance, where the output is always between -1 and 1. The formula, for two variables X and Y is:

$$r = \frac{cov(X,Y)}{\sigma X\ \sigma Y}$$

where σ is the standard deviation (SD) of each variable.

In explanation, variance (σ^2) is a measure of how much a set of data points differ from their mean. It is calculated as the average of the squared differences between each data point and the mean, i.e. sum of $(x-\bar{x})^2/n$. Covariance measures the degree to which two variables change together, i.e. sum of $(x-\bar{x})(y-\bar{y})/n$. Standard deviation (the square root of variance) is $\sqrt{(x-\bar{x})^2/n}$.

Spearman, ρ (rho), and Kendall, τ (tau), rank correlation coefficients are both non-parametric measure using the rank variable of the data. Spearman assesses how well the relationship between two variables can be described using a monotonic function using the formula:

$$\rho = 1 - \frac{6\ (sum\ di^2)}{n(n^2-1)}$$

Where di is the difference between the ranks of corresponding pairs of observations and n is the number of observations. Spearman correlation coefficient can be calculated by assigning ranks to the values for each variable, calculating the differences between ranks for variable pairs and then squaring the differences. Finally these values are summed. Kendall measures the similarity of the orderings of the data when ranked by each of the variables and can be described as:

$$\tau = \frac{C-D}{n(n-1)/2}$$

where C is the number of concordant pairs (where the ranks agree), D is the number of discordant pairs and n is the number of observations. Kendall is calculated by counting the number of concordant and discordant pairs of observations and then using these counts to compute the correlation coefficient.

The choice between Pearson, Spearman, or Kendall correlation coefficients may alluded to in Chapter 4 can depend upon the perceived importance of: Linear Relationship: Pearson correlation coefficient is specifically designed to measure the strength and direction of a linear relationship between two variables. Proportionality of Increase: Pearson correlation coefficient

considers the proportionality of increase or decrease in the variables. It reflects the degree to which a change in one variable is associated with a proportional change in the other variable. Outlier Treatment: Pearson correlation is sensitive to outliers, meaning that extreme values can significantly affect the correlation coefficient.

Appendix 12

Definitions of CASP15 PatchQS and PatchDockQ scores and the local Z-score calculation.

PatchQS and PatchDockQ are two reference scores created specifically for the CASP15 EMA competition and based on the QS-score and DockQ methods (Studer *et al.*, 2023). They are designed to assess the quality of interchain interactions and sample each model interface residue. For an interface residue r in chain A of the model two interface patches are defined for Cβ atoms as;

Patch one = (chain A and 8Å <> r) and (12Å <> chain != A). Meaning that patch one consists of all residues in chain A within 8Å of residue r that are also within 12Å of any other chain.

Patch two: (chain != A and 8Å <> r min) and (12Å <> A). Patch two uses r min as a reference point. It consists of all residues of any chain within 8Å of r min that are also within 12Å to chain A, where r min is defined as the closest residue to r in any chain which is not chain A (!= A). (<> means within that distance)

1. ROC AUC values.

Firstly the whole dataset of observed scores is sampled to calculate the 75th quartile value. This is then used as the threshold value against which the binary variable is calculated. A ROC AUC value is then calculated (in R this can be calculated using the pROC package). AUC values less than 0.5 are considered worse than a random selection and so the minimum AUC value was set to 0.5. These values were calculated using the lDDT, CAD, PatchQS and PatchDockQ as the target (observed) value.

2. Pearson r values.

This is a straight-forward Pearson correlation value, again calculated using the lDDT, CAD, PatchQS and PatchDockQ as the target (observed) value.

3. Spearman rho values.

As above, these are straight forward Spearman correlation values calculated using the lDDT, CAD, PatchQS and PatchDockQ as the target (observed) value.

4. Calculation of Z scores.

Z-scores are calculated (e.g. using the scale(value) operator in R) for each score above. So a Z-score value will be calculated for:

ROC_AUC_lDDT	ROC_AUC_CAD	ROC_AUC_PatchQS	ROC_AUC_PatchDockQ
Pearson_lDDT	Pearson_CAD	Pearson_PatchQS	Pearson_PatchDockQ
Spearman_lDDT	Spearman_CAD	Spearman_PatchQS	Spearman_PatchDockQ

5. Summation of like scores.

Overall scores are calculated as follows, where <score> is each of lDDT, CAD, PatchQS and PatchDockQ:

Z_<score>= Z_ROC_AUC_<score> + 0.5Z_Pearson_<score> + 0.5Z_Spearman_<score>

6. Final Z score.

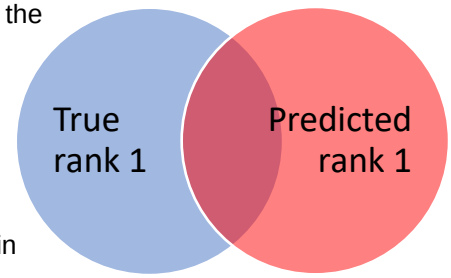
A final Z-score is calculated as a simple addition of Z_lDDT + Z_CAD + Z_PatchQS + Z_Patch_DockQ.

Appendix 13

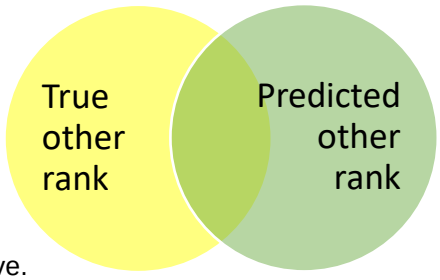
Definitions of Sensitivity, Specificity, Precision and Accuracy.

Each metric is best explained with reference to one class only, in this case rank 1 models.

Sensitivity (recall or TPR) = TP / (TP + FN). Meaning: of all the cases where models were ranked 1 by observed score, how many were also ranked 1 by predicted score. The intersection of the blue and pink circles represents the true positives (TP) and the rest of the blue circle represents the false negatives (FN). TPR is therefore the number of models in the intersect divided by the total number of models in the blue circle. Summary: Percent or fraction of True Positives.



Specificity = TN / (TN + FP). Meaning: of all the models that are not ranked 1 by observed score, how many were also not ranked 1 by predicted score. In the example, true negative (TN) is represented by the intersect between yellow and green circles and false positives (FP) are represented by the portion of the pink which does not intersect with the blue circle above. Summary: Percent or fraction of True Negatives.



Precision = TP / (TP + FP). Meaning: of all the cases that were predicted as rank 1 how many actually were rank 1. In the example diagrams, this is again the intersect of the pink and blue circles but this time divided by the total number of models in the pink circle. Summary: Percent or fraction of positives that were truly positives.

Accuracy = (TP + TN) / (TP + TN + FP + FN). Meaning: of all models in the population, how many were correctly predicted. In the example diagrams this would be both intersections added together divided by this number plus the portion of the pink circle which does not intersect with the blue and the portion of the blue circle which does not intersect with the pink. Summary: Percent or fraction correctly classified.

Appendix 14

Example R script (Global score) for MLP three-fold cross validation.

```
library(RSNNS)
library(data.table)
library(ROCR)
library(ggplot2)
#-------------------------------------------------------------------------------
# Name: Glob_RSNNS_3fold_X_Val.R Version: 1.4 Date: 29-Sep-21  Author: Nick Edmunds
# Revision history (add details of any revisions since Date above):
# 1-Oct-21 (NE): added summary stats for each baseline graph.
# 1-Oct-21 (NE): added model-checking statistics to check fit of lm to each
baseline graph and to each prediction model.
# 1-Oct-21 (NE): removed consensus6 from testing and training sets1,2 and 3 and
created a separate baseline_test1,2 and 3.
# 6-Oct-21 (NE): created a binary variable in testing_all_outputs dataset and added
ROC plots and AUC calcs for individual scores.
# 13-Oct-21 (NE): Added baseline correlation, ROC/AUC plots for Observed scores.
# 15-Oct-21 (NE): Added testing_all_sets$Global_predictions <- predictions to add
predictions to a permanent dataset created in loc program.
# Function:
# A 3-Fold cross validation for the NN prediction of Global Score from all 6
ModFOLDdock predicted scores.
# Defines 3 training sets containing models from different CASP13 and 14 targets.
Training sets are balanced to include roughly
# the same number of targets from ech CASP competition and also by number of easy,
medium and difficult rated targets.
# Creates correlation plots for each individual (of the 6) ModFOLDdock scores
against a calculated CASP Global score plus the
# same for a mean consensus 6 score - these act as baseline correlations for
comparison.
# Thereafter, each training set is predicted separately, the model saved and then
reloaded and simple correlations, regression
# errors, iterative errors as well as confusion matrices and ROC plots with AUC
calculations are output.
# Edit this to direct graph output to the desired directory>>
#
setwd('/home/nick/Post_confirmation_projects/New_Scoring/All_CASP_models/NN_work/Hy
perparam_testing_graphs/6_10_21_Glob_and_Tot_output')
#-------------------------------------------------------------------------------
# Define 3 subset arrays containing different Target ids (sub_set1=15, subset2=15,
sub_set3=14 targets).
sub_set1 <- c("T1016", "T1003", "T1020", "T0977", "T0999", "T0997", "T1083",
"T1048", "T1087", "T0966", "T0991", "T1009", "T1038", "T1054", "T0989")
sub_set2 <- c("T0995", "T0979", "T0984", "T0983", "T0963", "T1018", "T0976",
"T0998", "T0965", "T1062", "T1078", "T1084", "T1001", "T1061", "T1070")
sub_set3 <- c("T1006", "T0961", "T1032", "T0973", "T0970", "T1034", "T0960",
"T1004", "T0981", "T0996", "T0985", "T1000", "T1080", "T1010")

# Define training datasets so that training_set1 contains NO rows from sub_set1,
training_set2 contains NO rows from sub_set2 and training_set3 contains NO rows
from sub_set3.
# So training_set1 will only have rows from sub_set2 and sub_set3
training_set1 <- subset(CASP_combined, Target!="T1016" & Target!="T1003" &
Target!="T1020" & Target!="T0977" & Target!="T0999" & Target!="T0997" &
Target!="T1083" & Target!="T1048" & Target!="T1087" & Target!="T0966" &
Target!="T0991" & Target!="T1009" & Target!="T1038" & Target!="T1054" &
Target!="T0989")
```

```
# training_set2 will only have rows from sub_set1 and sub_set3
training_set2 <- subset(CASP_combined, Target!="T0995" & Target!="T0979" &
Target!="T0984" & Target!="T0983" & Target!="T0963" & Target!="T1018" &
Target!="T0976" & Target!="T0998" & Target!="T0965" & Target!="T1062" &
Target!="T1078" & Target!="T1084" & Target!="T1001" & Target!="T1061" &
Target!="T1070")
# training_set3 will only have rows from sub_set1 and sub_set2
training_set3 <- subset(CASP_combined, Target!="T1006" & Target!="T0961" &
Target!="T1032" & Target!="T0973" & Target!="T0970" & Target!="T1034" &
Target!="T0960" & Target!="T1004" & Target!="T0981" & Target!="T0996" &
Target!="T0985" & Target!="T1000" & Target!="T1080" & Target!="T1010")

# Define testing datasets so that testing_set1 contains ONLY rows for sub_set1,
testing_set2 contains ONLY rows for sub_set2 and training set3 contains ONLY rows
for sub_set3.
testing_set1 <- subset(CASP_combined,
Target=="T1016"|Target=="T1003"|Target=="T1020"|Target=="T0977"|Target=="T0999"|Tar
get=="T0997"|Target=="T1083"|Target=="T1048"|Target=="T1087"|Target=="T0966"|Target
=="T0991"|Target=="T1009"|Target=="T1038"|Target=="T1054"|Target=="T0989")
testing_set2 <- subset(CASP_combined,
Target=="T0995"|Target=="T0979"|Target=="T0984"|Target=="T0983"|Target=="T0963"|Tar
get=="T1018"|Target=="T0976"|Target=="T0998"|Target=="T0965"|Target=="T1062"|Target
=="T1078"|Target=="T1084"|Target=="T1001"|Target=="T1061"|Target=="T1070")
testing_set3 <- subset(CASP_combined,
Target=="T1006"|Target=="T0961"|Target=="T1032"|Target=="T0973"|Target=="T0970"|Tar
get=="T1034"|Target=="T0960"|Target=="T1004"|Target=="T0981"|Target=="T0996"|Target
=="T0985"|Target=="T1000"|Target=="T1080"|Target=="T1010")

# Shuffle training_set1 (data from sub_set2 and sub_set3) into a random order and
split into inputs and output datasets
training_set1_shuffle <- training_set1[sample(1:nrow(training_set1),
length(1:nrow(training_set1))),1:ncol(training_set1)]
training_set1_inputs  <- training_set1_shuffle[c(28,29,30,31,32,33)] # just the 6
MFD scores as inputs
training_set1_output  <- training_set1_shuffle[c(42)]  # just Global_score as the
output
# Same for testing_set1 (data from sub_set1) minus the random order
baseline_set1_inputs<- testing_set1[c(28,29,30,31,32,33,34)]  # MFD scores incl.
consensus6 for baseline correlation calculations.
testing_set1_inputs <- testing_set1[c(28,29,30,31,32,33)] # just the 6 MFD scores
as inputs
testing_set1_output <- testing_set1[c(42)]  # just Global_score as the output_input

# Shuffle training_set2 (data from sub_set1 and sub_set1) into a random order and
split into inputs and output datasets
training_set2_shuffle <- training_set2[sample(1:nrow(training_set2),
length(1:nrow(training_set2))),1:ncol(training_set2)]
training_set2_inputs  <- training_set2_shuffle[c(28,29,30,31,32,33)] # just the 6
MFD scores as inputs
training_set2_output  <- training_set2_shuffle[c(42)]    # just Global_score as the
output
# Same for testing_set2 (data from sub_set2) minus the random order
baseline_set2_inputs<- testing_set2[c(28,29,30,31,32,33,34)] # MFD scores incl.
consensus6 for baseline correlation calculations.
testing_set2_inputs <- testing_set2[c(28,29,30,31,32,33)]  # just the 6 MFD scores
as inputs
testing_set2_output <- testing_set2[c(42)]  # just Global_score as the output_input
```

```
# Shuffle training_set3 (data from sub_set1 and sub_set1) into a random order and
split into inputs and output datasets
training_set3_shuffle <- training_set3[sample(1:nrow(training_set3),
length(1:nrow(training_set3))),1:ncol(training_set3)]
training_set3_inputs  <- training_set3_shuffle[c(28,29,30,31,32,33)] # just 6 MFD
scores input
training_set3_output  <- training_set3_shuffle[c(42)] #just Global_score as output
# Same for testing_set3 (data from sub_set3) minus the random order
baseline_set3_inputs<- testing_set3[c(28,29,30,31,32,33,34)]  # MFD scores incl.
consensus6 for baseline correlation calculations.
testing_set3_inputs <- testing_set3[c(28,29,30,31,32,33)] # just 6 MFD scores input
testing_set3_output <- testing_set3[c(42)]  # just Global_score as the input

# ######################### Create the models for the predictions #################
# Add the general working directory so that all the models get saved to a single
directory
setwd('/home/nick/Post_confirmation_projects/New_Scoring/All_CASP_models/NN_work/')
# The model for training on training_set1 and predicting on testing_set1 (learning
rate 0.01, max difference 0.01)
modelG1 <- mlp(training_set1_inputs, training_set1_output, size = 4,
               learnFuncParams = c(0.01, 0.01), maxit = 200, inputsTest =
testing_set1_inputs, targetsTest = testing_set1_output,
               learnFunc = "BackpropMomentum", linOut=TRUE)
save(modelG1, file="modelG1_set1.RData")
rm(modelG1)
load("modelG1_set1.RData")
prediction_set1 <- predict(modelG1, testing_set1_inputs)
compare_set1 <- data.frame(testing_set1_output, prediction_set1)

# The model for training on training_set2 and predicting on testing_set2
modelG2 <- mlp(training_set2_inputs, training_set2_output, size = 4,
               learnFuncParams = c(0.01, 0.01), maxit = 200, inputsTest =
testing_set2_inputs, targetsTest = testing_set2_output,
               learnFunc = "BackpropMomentum", linOut=TRUE)
save(modelG2, file="modelG2_set2.RData")
rm(modelG2)
load("modelG2_set2.RData")
prediction_set2 <- predict(modelG2, testing_set2_inputs)
compare_set2 <- data.frame(testing_set2_output, prediction_set2)

# The model for training on training_set3 and predicting on testing_set3
modelG3 <- mlp(training_set3_inputs, training_set3_output, size = 4,
               learnFuncParams = c(0.01, 0.01), maxit = 200, inputsTest =
testing_set3_inputs, targetsTest = testing_set3_output,
               learnFunc = "BackpropMomentum", linOut=TRUE)
save(modelG3, file="modelG3_set3.RData")
rm(modelG3)
load("modelG3_set3.RData")
prediction_set3 <- predict(modelG3, testing_set3_inputs)
compare_set3 <- data.frame(testing_set3_output, prediction_set3)

############################-Processing results for set1-#########################
# Edit this again to direct the rest of the graph output to the same directory as
at the start>>
setwd('/home/nick/Post_confirmation_projects/New_Scoring/All_CASP_models/NN_work/Hy
perparam_testing_graphs/6_10_21_Glob_and_Tot_output')

# Iterative error for prediction set1 vs testing_set1
jpeg("ModelG1_IterativeError.jpg")
```

```
plotIterativeError(modelG1)
legend(x='bottomright', "modelG1 iterative error")
dev.off()
#-------------------------------------------------------------------------------
# Regression error for prediction set1 vs testing_set1
jpeg("ModelG1_RegressionError.jpg")
plotRegressionError(testing_set1_output$Global_score, prediction_set1)
legend(x='bottomright', "modelG1 refression error")
dev.off()
#-------------------------------------------------------------------------------
# Simple set1 scatter plot with regression line and correlation value
jpeg("Set1_prediction_V_Global_score_scatter.jpg")
plot(main='Prediction set 1 scatter (Global)',prediction_set1,
testing_set1_output$Global_score,
col=c("blue"),abline(lm(testing_set1_output$Global_score ~ prediction_set1)))
legend(x='bottomright', legend=paste('modelG1 Pearson =',round(cor(prediction_set1,
testing_set1_output$Global_score),2)))
dev.off()
#-------------------------------------------------------------------------------
# Summary of model1 to give R-squared values and model checking graphs.
set1_modelG <-lm(testing_set1_output$Global_score ~ prediction_set1)
summary(set1_modelG)
par(mfrow=c(2,2))
plot(main='Predict set1 versus Global score', set1_modelG)
par(mfrow=c(1,1))
#-------------------------------------------------------------------------------
# Confusion table, TPR, FPR, ROC plot and AUC
# Convert the results as binaries for creation of a confusion table so that TPR and
FPR can be clearly seen and calculated manually.
# Firstly, compare the actual and predicted value to get an absolute difference.
compare_set1$diff <- abs(compare_set1$Global_score - compare_set1$prediction_set1)
# Next, if difference is within 0.06, it can be considered correct so is set to the
SAME value as actual, if greater that 0.06 it remains
# as the predicted value.
compare_set1$bin <- ifelse(compare_set1$diff < 0.06, compare_set1$Global_score,
compare_set1$prediction_set1)
# Now, when rounded, they should have the appropriate values - prevents close
scores like 5.4 and 5.6 being rounded to different numbers.
compare_set1$Global_Rscore <- round(as.numeric(testing_set1_output$Global_score),1)
compare_set1$R_prediction  <- round(as.numeric(compare_set1$bin),1)
# Now make two binary variables. Above 0.5 =1 below = 0.
compare_set1$bin_P <- ifelse(compare_set1$R_prediction > 0.5, 1, 0) # for the
predicted value
compare_set1$bin_A <- ifelse(compare_set1$Global_Rscore > 0.5, 1, 0) # for the
actual value
# Now make a confusion matrix with the Actual Global score on the left and the
predictions as columns across the top.
sink('ModelG1_confusion_matrix.txt')
confusionMatrix(compare_set1$R_prediction, compare_set1$Global_Rscore)
sink()
# Also, just for reference, a simple binary confusion matrix
confusionMatrix(compare_set1$bin_P, compare_set1$bin_A)
# Now a ROC plot using the unaltered predicted values and the binary actual
(Global_score) values (0.5 cut-off)
pred1 <- prediction(compare_set1$prediction_set1, compare_set1$bin_A)
perf1 <- performance(pred1, "tpr", "fpr")
jpeg("ModelG1_ROC_plot.jpg")
plot(perf1, colorize=TRUE, print.cutoffs.at=seq(0,1,0.1))
abline(a=0, b=1)
```

```
dev.off()
# Calculate AUC
sink('ModelG1_AUC.txt')
auc.perf1 <-performance(pred1, measure="auc")
auc.perf1@y.values
sink()
############################-Processing results for set2-#########################
# Iterative error for prediction set2 vs testing_set2
jpeg("ModelG2_IterativeError.jpg")
plotIterativeError(modelG2)
legend(x='bottomright', "modelG2 iterative error")
dev.off()
#-------------------------------------------------------------------------------
# Regression error for prediction set2 vs testing_set2
jpeg("ModelG2_RegressionError.jpg")
plotRegressionError(prediction_set2,testing_set2_output$Global_score)
legend(x='bottomright', "modelG2 refression error")
dev.off()
#-------------------------------------------------------------------------------
# Simple set2 scatter plot with regression line and correlation value
jpeg("Set2_prediction_V_Global_score_scatter.jpg")
plot(main='Prediction set 2 scatter
(Global)',prediction_set2,testing_set2_output$Global_score,
col=c("red"),abline(lm(testing_set2_output$Global_score ~ prediction_set2)))
legend(x='bottomright', legend=paste('modelG2 Pearson =',round(cor(prediction_set2,
testing_set2_output$Global_score),2)))
dev.off()
#-------------------------------------------------------------------------------
# Summary of model2 to give R-squared values and model checking graphs.
set2_modelG <-lm(testing_set2_output$Global_score ~ prediction_set2)
summary(set2_modelG)
par(mfrow=c(2,2))
plot(main='Predict set2 versus Global score', set2_modelG)
par(mfrow=c(1,1))
#-------------------------------------------------------------------------------
# Confusion table, TPR, FPR, ROC plot and AUC
# Convert the results as binaries for creation of a confusion table so that TPR and
FPR can be clearly seen and calculated manually.
# Firstly, compare the actual and predicted value to get an absolute difference.
compare_set2$diff <- abs(compare_set2$Global_score - compare_set2$prediction_set2)
# Next, if difference is within 0.06, it can be considered correct so is set to the
SAME value as actual, if greater that 0.06 it remains
# as the predicted value.
compare_set2$bin <- ifelse(compare_set2$diff < 0.06, compare_set2$Global_score,
compare_set2$prediction_set2)
# Now, when rounded, they should have the appropriate values - prevents close
scores like 5.4 and 5.6 being rounded to different numbers.
compare_set2$Global_Rscore <- round(as.numeric(testing_set2_output$Global_score),1)
compare_set2$R_prediction  <- round(as.numeric(compare_set2$bin),1)
# Now make two binary variables. Above 0.5 =1 below = 0.
compare_set2$bin_P <- ifelse(compare_set2$R_prediction > 0.5, 1, 0) # for the
predicted value
compare_set2$bin_A <- ifelse(compare_set2$Global_Rscore > 0.5, 1, 0) # for the
actual value
# Now make a confusion matrix with the Actual Global score on the left and the
predictions as columns across the top.
sink('ModelG2_confusion_matrix.txt')
confusionMatrix(compare_set2$R_prediction, compare_set2$Global_Rscore)
sink()
```

```
# Also, just for reference,a simple binary confusion matrix
confusionMatrix(compare_set2$bin_P, compare_set2$bin_A)
# Now a ROC plot using the unaltered predicted values and the binary actual
(Global_score) values (0.5 cut-off)
pred2 <- prediction(compare_set2$prediction_set2, compare_set2$bin_A)
perf2 <- performance(pred2, "tpr", "fpr")
jpeg("ModelG2_ROC_plot.jpg")
plot(perf2, colorize=TRUE, print.cutoffs.at=seq(0,1,0.1))
abline(a=0, b=1)
dev.off()
# Calculate AUC
sink('ModelG2_AUC.txt')
auc.perf2 <-performance(pred2, measure="auc")
auc.perf2@y.values
sink()
############################-Processing results for set3-#########################
# Iterative error for prediction set3 vs testing_set3
jpeg("ModelG3_IterativeError.jpg")
plotIterativeError(modelG3)
legend(x='bottomright', "modelG3 iterative error")
dev.off()
#-------------------------------------------------------------------------------
# Regression error for prediction set3 vs testing_set3
jpeg("ModelG3_RegressionError.jpg")
plotRegressionError(prediction_set3,testing_set3_output$Global_score)
legend(x='bottomright', "modelG3 refression error")
dev.off()
#-------------------------------------------------------------------------------
# Simple set3 scatter plot with regression line and correlation value
jpeg("Set3_prediction_V_Global_score_scatter.jpg")
plot(main='Prediction set 3 scatter
(Global)',prediction_set3,testing_set3_output$Global_score,
col=c("green"),abline(lm(testing_set3_output$Global_score ~ prediction_set3)))
legend(x='bottomright', legend=paste('modelG3 Pearson =',round(cor(prediction_set3,
testing_set3_output$Global_score),2)))
dev.off()
#-------------------------------------------------------------------------------
# Summary of model3 to give R-squared values and model checking graphs.
set3_modelG <-lm(testing_set3_output$Global_score ~ prediction_set3)
summary(set3_modelG)
par(mfrow=c(2,2))
plot(main='Predict set3 versus Global score', set3_modelG)
par(mfrow=c(1,1))
#-------------------------------------------------------------------------------
# Confusion table, TPR, FPR, ROC plot and AUC
# Convert the results as binaries for creation of a confusion table so that TPR and
FPR can be clearly seen and calculated manually.
# Firstly, compare the actual and predicted value to get an absolute difference.
compare_set3$diff <- abs(compare_set3$Global_score - compare_set3$prediction_set3)
# Next, if difference is within 0.06, it can be considered correct so is set to the
SAME value as actual, if greater that 0.06 it remains
# as the predicted value.
compare_set3$bin <- ifelse(compare_set3$diff < 0.06, compare_set3$Global_score,
compare_set3$prediction_set3)
# Now, when rounded, they should have the appropriate values - prevents close
scores like 5.4 and 5.6 being rounded to different numbers.
compare_set3$Global_Rscore <- round(as.numeric(testing_set3_output$Global_score),1)
compare_set3$R_prediction  <- round(as.numeric(compare_set3$bin),1)
# Now make two binary variables. Above 0.5 =1 below = 0.
```

```
compare_set3$bin_P <- ifelse(compare_set3$R_prediction > 0.5, 1, 0) # for the
predicted value
compare_set3$bin_A <- ifelse(compare_set3$Global_Rscore > 0.5, 1, 0) # for the
actual value
# Now make a confusion matrix with the Actual Global score on the left and the
predictions as columns across the top.
sink('ModelG3_confusion_matrix.txt')
confusionMatrix(compare_set3$R_prediction, compare_set3$Global_Rscore)
sink()
# Also, just for reference,a simple binary confusion matrix
confusionMatrix(compare_set3$bin_P, compare_set3$bin_A)
# Now a ROC plot using the unaltered predicted values and the binary actual
(Global_score) values (0.5 cut-off)
pred3 <- prediction(compare_set3$prediction_set3, compare_set3$bin_A)
perf3 <- performance(pred3, "tpr", "fpr")
jpeg("ModelG3_ROC_plot.jpg")
plot(perf3, colorize=TRUE, print.cutoffs.at=seq(0,1,0.1))
abline(a=0, b=1)
dev.off()
# Calculate AUC
sink('ModelG3_AUC.txt')
auc.perf3 <-performance(pred3, measure="auc")
auc.perf3@y.values
sink()
```