Bioinformatics, 2022, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Structural bioinformatics

Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes

Wensi Zhu[1†], Aditi Shenoy[1†], Petras Kundrotas[1,2] and Arne Elofsson[1*],

[1]Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, 171 21 Solna, Sweden

[2]Center for Computational Biology, The University of Kansas, Lawrence, KS 66047, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that the first two authors should be regarded as Joint First Authors

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Despite near-experimental accuracy on single-chain predictions, there is still scope for improvement among multimeric predictions. Methods like AlphaFold-Multimer and FoldDock can accurately model dimers. However, how well these methods fare on larger complexes is still unclear. Further, evaluation methods of the quality of multimeric complexes are not well established.

Results: We analysed the performance of AlphaFold-Multimer on a homology-reduced dataset of homo- and heteromeric protein complexes. We highlight the differences between the pairwise and multi-interface evaluation of chains within a multimer. We describe why certain complexes perform well on one metric (e.g. TM-score) but poorly on another (e.g. DockQ). We propose a new score, the Predicted Multi-Interface DockQ (pDockQ$_i$), to estimate the quality of each interface in a multimer. Finally, we modelled protein complexes (from CORUM) and identified two highly confident structures that do not have sequence homology to any existing structures.

Availability: All scripts, models, and data used to perform the analysis in this study are freely available at https://gitlab.com/ElofssonLab/afm-benchmark.

Contact: arne@bioinfo.se
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1. Introduction

Most biological processes and cellular functions depend on protein structures and interactions between proteins. Understanding how proteins form three-dimensional structures can give key insights into protein function. Therefore, the determination of the 3D structure of a protein from its primary amino acid sequence has been a fundamental problem in biology (Dill et al., 2008). With the recent advancement of AlphaFold (Jumper et al., 2021), obtaining a three-dimensional protein structure with near experimental accuracy is possible for most proteins by giving only the amino acid sequence as input. AlphaFold has been trained on protein chains and has shown remarkable performance in single-domain protein chain predictions. However, inside the densely packed environment of the cell, proteins are constantly near each other and perform functions by forming contacts with other biological macromolecules (Bergendahl & Marsh, 2017). Shortly after the release of AlphaFold, the same pipeline was recognised to be useful for predicting structures of protein assemblies if the input sequences are concatenated using flexible linkers or by modifying their residue numbers (Bryant et al., 2022; Mirdita et al., 2022). Here, the generation of a "paired" alignment is crucial.

In FoldDock (Bryant et al., 2022), the multiple sequence alignments (MSA) for individual chains are joined in various ways and then submitted into the AlphaFold pipeline with templates being disabled. AF2Complex (Gao et al., 2022) can build multimeric proteins using structural templates without paired multiple sequence alignment of the individual protein chains. AlphaFold-Multimer, (Evans et al., 2022), an extension of AlphaFold for multimeric proteins, was specifically trained on multi-chains proteins.

In this study, we evaluated the performance of AlphaFold-multimer(AlphaFold multimer) predictions on a homology-reduced dataset independent from the AlphaFold multimer training set consisting of homomeric and heteromeric complexes with two to six chains. Current evaluation metrics (i.e. DockQ, TM-score) were used to estimate the quality of the interface as well as the global topology of the complexes. The overall success rate ranges from

~40-60% across all states, with a small decrease for larger heteromeric complexes. We also present a novel score, the Predicted Multi-Interface DockQ (pDockQ$_i$), which estimates the quality of the interfaces in oligomers in the real-case scenario when the native (reference) structure is unknown and can be used to identify partially correct multimeric models.

2. Methods

2.1. Benchmark Dataset

Initially, the first biological unit of all structures with two to six chains (each with at least 30 residues) released after 30/04/2018 (the last date used in the training set of AlphaFold) was downloaded from the biounits part of the Protein Data Bank (PDB). The structures were classified as homomers (all chains are 100% identical) or heteromers. Further, nine structures where at least one protein chain does not have contact with any other proteins (this may occur after DNA/RNA removal from the PDB structure) were removed from the dataset. This resulted in 23222 proteins, with 13136 dimers, 3025 trimers, 4257 tetramers, 942 pentamers and 1862 hexamers.

Similarity and homology reduction within each oligomeric state: We removed similar structures within each oligomeric state separately. This was done by aligning all-vs-all structures by MMalign (Mukherjee & Zhang, 2009) and subsequent clustering by the resulting MM-scores (TM-score (Zhang & Skolnick, 2005) calculated for all chains in the structure) utilising the Highly Connected Subgraphs (HCS) method (Hartuv & Shamir, 2000). In this study, we used the MMscore threshold of 0.6 for clustering, which roughly corresponds to the TM-score threshold for the individual proteins to have the same fold. To reduce the number of combinations, we only consider protein pairs where at least one protein is similar, using FoldSeek (van Kempen et al., 2022) with default settings. After clustering, the structure with the highest resolution was chosen as the representative structure. If two structures had the same resolution, then the one having the minimum difference between the SEQRES sequence and ATOM section of the PDB record was selected. This resulted in a dataset of 5402 proteins, with 3051 dimers, 548 trimers, 1071 tetramers, 205 pentamers and 527 hexamers.

Homology reduction against the AlphaFold training dataset: MMseqs2 (Steinegger & Söding, 2017) was used to remove homology between the benchmark dataset and the AlphaFold training dataset (PDB structures before 2018/04/30). A structure in our datasets was removed if at least one chain shared ≥ 30% sequence identity with any sequence in the AlphaFold training dataset. Finally, a manual examination of global stoichiometries for each oligomeric state was performed, and protein structures with conflicting stoichiometries were removed. This resulted in 1997 proteins, with 1151 dimers, 224 trimers, 397 tetramers, 70 pentamers and 155 hexamers.

2.2. CORUM Dataset

For further evaluation, we used the CORUM Version 3.0 Core Set (released in September 2018) (Giurgiu et al., 2019), containing 512 complexes with two to six chains. To identify complexes having no homology to any previous PDB structure, we ran MMseqs2 (Steinegger & Söding, 2017) between all chains of the selected CORUM complexes and PDB structures. If none of the chains in a CORUM complex has ≥ 30% sequence identity to any PDB chain, then this complex was kept for further modelling. This strict criterion reduced the number of CORUM complexes to 53, for which we ran through the AlphaFold multimer pipeline with its default settings. Out of those, 14 complexes did not produce MSAs for at least one of the chains and further 10 cases, modelling failed due to out-of-memory or out-of-time errors, leaving 29 potentially novel complexes.

2.3. Model Generation

AlphaFold multimer (Evans et al., 2022) models were generated using AlphaFold v2.2.0 and the top-ranked model by AlphaFold was used for the analysis. The default parameters for multiple sequence alignment (MSA) generation and the number of recycles (3) were used. In 119 cases, the default pipeline did not produce all MSAs due to an out-of-memory error when run on 12 cores (Intel Xeon E5-2690v4) for 60 hours. These proteins were rerun using only one HHblits (Remmert et al., 2011) iteration and the reduced database small_bfd for 80 hours, leaving 18 proteins which did not generate alignments even with those reduced settings. AlphaFold multimer was run on a single NVIDIA DGX-A100 Core GPU for 72 hours. Still, the modelling of 60 complexes failed with the above settings due to an out-of-time error; these were ignored. The final dataset consists of 1928 protein complexes (1148 dimers, 220 trimers, 367 tetramers, 62 pentamers and 131 hexamers).

2.4. Evaluation

Scores to evaluate the quality of models against native PDB: we used two scores to evaluate the quality of the models - the DockQ score (Basu & Wallner, 2016) and the TM-score produced by the MM-align program (Mukherjee & Zhang, 2009) (also referred to as MM-score). Both scores range between 0 and 1. According to the study (Basu & Wallner, 2016), if the DockQ score > 0.23, then the quality of the model is acceptable by the CAPRI criteria (Janin et al., 2003). MM-score was obtained using the default MM-align settings. It has been previously shown that a TM-score of 0.5 roughly refers to a model having the right fold. However, when considering a protein complex of two or more chains, it is possible to obtain scores higher than 0.5 if the larger chain(s) structure is correct. Still, the prediction of the quaternary structure can be wrong, i.e. it is necessary to use a higher cut-off to separate correct and incorrect multi-chain models.

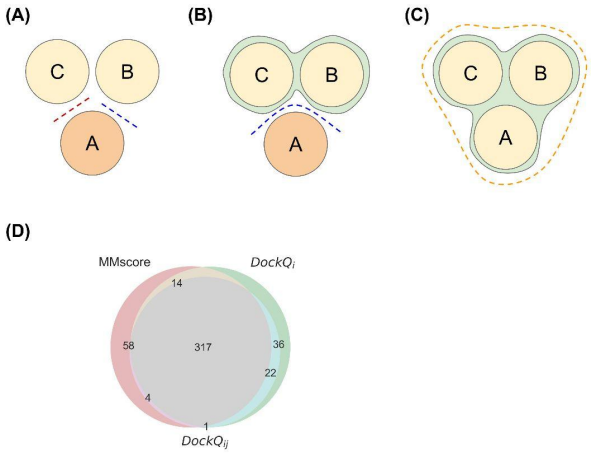

Fig 1. Schematic representation of two types of interface in an exemplary trimer (A) using pairwise interfaces (i.e. DockQ of chain A

vs chain B and DockQ of chain A vs chain C) (B) using multi-interface (i.e. DockQ when chain A is the ligand and chain B & chain C together form the receptor). Dashed lines in both panels represent interfaces with a non-zero number of contacts. (C) using overall structure as in MMscore (d) Venn diagram for number of successful docking models indicated by three different metrics (MM-score > 0.75, red circle, $DockQ_{ij}$ > 0.23, cyan circle, and $DockQ_i$ > 0.23, green circle).

2.5. DockQ for multimeric complexes

The DockQ score signifies the quality of an interface of a model compared with the native structure, with the larger protein acting as the receptor and the smaller protein as the ligand. In a multi-oligomeric complex, there are several ways to define an interface, e.g. (i) residues in contact between any pair of chains i and j with a complex, pairwise interface DockQ ($DockQ_{ij}$) as illustrated in Fig 1A, or (ii) residues in contact between one chain i and all the other chains, multi-interface DockQ ($DockQ_i$) as illustrated in Fig 1B. Except for homomers with identical interfaces and dimers with only one interface, most cases have $DockQ_i \neq \sum_j DockQ_{ij}$. To make sure the order of the chains in the AlphaFold multimer model is the same as the native structure, we use the MMalign output alignment mapping chains between the native and modelled structures so that the chain names are consistent. For example, for a tetramer with chains A, B, C, and D, the $DockQ_i$ was calculated as follows:

python3 DockQ.py model.pdb native.pdb -model_chain1 A -native_chain1 A -model_chain2 B C D -native_chain2 B C D

while $DockQ_{ij}$ was calculated for each pair of chains (A and B in the example) with

python3 DockQ.py model.pdb native.pdb -model_chain1 A -native_chain1 A -model_chain2 B -native_chain2 B

Note that the AlphaFold multimer models were generated for the full-length (PDB SEQRES section) sequences. Thus, to avoid the impact of residues, possibly not resolved experimentally, we calculated the DockQ score only for those residues present in both modelled and reference (PDB ATOM section) structures.

2.6. Predicting the quality of a model

AlphaFold multimer provides two intrinsic model accuracy estimates, pTM and ipTM. Both these scores estimate the average quality of the complex (or all interfaces of the complex), predicting the TM-score. However, in addition to estimating the quality of the entire predicted model, it is sometimes desirable to estimate each interface's quality within a multi-chain complex. For dimers, (Bryant et al., 2022) proposed to use a predicted DockQ score (pDockQ) calculated from the number of contacts and the average quality of the interacting residues fitted to a sigmoid function:

$$pDockQ = \frac{L}{1+exp[-k*(x-x_0)]} + b \quad (1)$$

with

$$x = < plDDT >_{int} * log(N_{int}) \quad (2)$$

where $<plDDT>_{int}$ stands for the average of plDDT (i.e. predicted lDDT score (Mariani et al., 2013)) from AlphaFold multimer) over dimer interface residues and N_{int} is the number of interface contacts. For multi-chain complexes, the pDockQ scores for chains i and j do not perform optimally (see Results). Therefore, we propose a novel multi-interface DockQ score, $pDockQ_{i,}$ for the choice of chain i as a ligand (for brevity, hereafter referred to as $pDockQ_i$ for chain i). For this score, we replace x from (2) with x_i for chain i:

$$x_i = \left(\frac{1}{1+\left(\frac{<PAE>_{int}}{d}\right)^2}\right) * < plDDT >_{int} \quad (3)$$

Here, PAE is the Predicted Aligned Error produced by AlphaFold multimer ((Evans et al., 2022)). The $<PAE>_{int}$ is the average PAE over all interfaces for chain *i*, $<plDDT>_{int}$ is the average of plDDT of that combined interface residues, and *d* is a scaling parameter optimised to 10 Å. As in pDockQ, we fit a sigmoid curve (eq. 1) (by the scipy package) to the actual $DockQ_i$ values, yielding the coefficients L= 0.0161 x0= 0.916 k= 6.4 b= 0.003

3. Results and Discussion

3.1. Interface quality in homomeric and heteromeric complexes using different measures.

How to best evaluate the quality of a multimeric protein model is not well examined as some metrics (e.g., MM-score) evaluates complex in its entirety. At the same time, other quantities provide evaluation per interface (DockQ). In this study, we investigate correlations between these types of metrics (Fig 2A). To perform per-complex comparisons, we aggregated per-interface scores ($DockQ_{ij}$ and $DockQ_i$) for the entire complex, either by using minimum or maximum scores for all interfaces.

For the heteromers, the spread of values for both per-interface scores tends to be larger compared to the spread for the homomers. Also, variations in $DockQ_{ij}$ values tend to be larger than in $DockQ_i$. This is related to the simple fact that most of the homomeric complexes (79.78 % for homotrimers, 91.44% for homotetramers, 73.91% for homopentamers and 89.40% for homohexamers) possess internal symmetry, making all interfaces structurally identical in the native structure and often also in the models. This, in turn, leads to the uniform quality of the predicted interfaces and, thus, to a smaller variation in per-interface quality scores compared to the heteromers, where each interface may be structurally unique. That correlates with the observation that all or no interfaces are predicted correctly in the homomers, while for the heteromers, there is a variable fraction of correctly predicted interfaces (Fig 2B). Several examples of

heteromers with partially predicted interfaces and, consequently, big differences between $DockQ_{ij}$ and $DockQ_i$ scores are shown in Figure S1. A notable example is displayed in Fig S1A, where a big difference in the scores is caused by a very low (only one in this case) number of native contacts, which was not predicted in the AlphaFold multimer model. In such cases, $DockQ_{ij}$ for that interface is very low, while contributions of that interface into all $DockQ_i$ scores are levelled out, yielding high scores (for details, see caption to Fig S1).

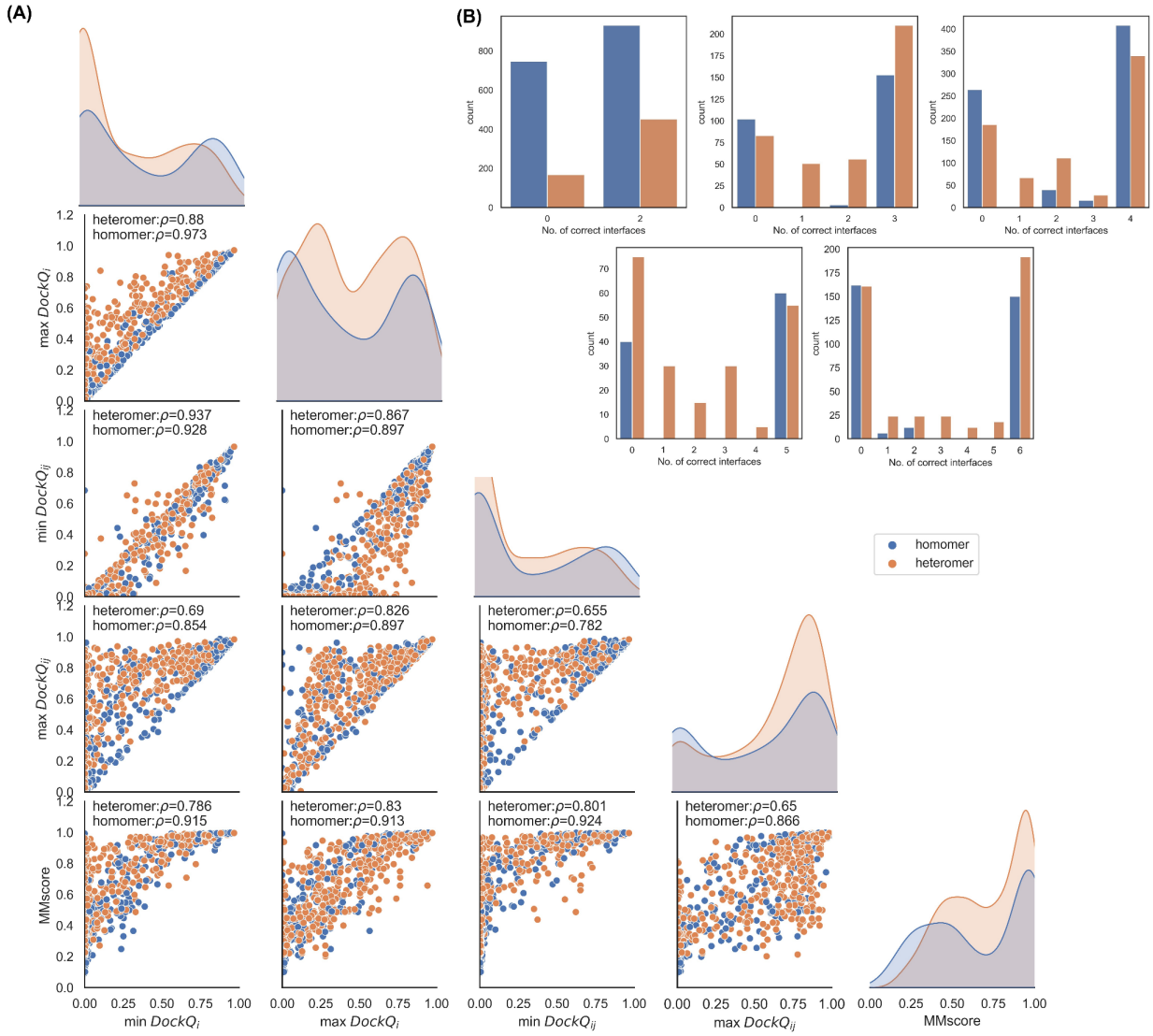

Fig 2. (A) Pairplot showing correlations between $DockQ_{ij}$ (min/max), $DockQ_i$ (min/max) and MMscore for complexes > 2 chains (B) Distribution of the number of correct multi-interfaces ($DockQ_i > 0.23$) for each oligomeric state.

Data in Fig 2A indicate that a significant fraction of the complexes have MM-score indicating good modelling quality (> 0.75) while $min(DockQ_i)$ scores point to modelling failure. Only five cases are homomers out of these 58 complexes with $min(DockQ_i) < 0.1$ and MMscore > 0.75. Fig 3A shows a homodimeric membrane protein (7STL) with a small alpha-helical swapped between the chains in the native structure. Alphafold multimer predicts that those domains are associated with the corresponding main chain, i.e. not being domain-swapped. Fig 3B illustrates another homo-dimeric complex (6WKU) with low $min(DockQ_i)$ and high MMscore for the AlphaFold multimer model. Here, a single chain is the result of the fusion of three different proteins, and in the native structure, both chains have the same direction (clockwise) of component order. However, in the AlphaFold multimer model, those directions are different (clockwise for one chain and counterclockwise for the other). MM-score is high in this case as a structural alignment algorithm is performed without paying attention to the exact residue matching.

Finally, we see a high overlap among the good models identified by the different methods using the following cutoffs, MMscore > 0.75, $min(DockQ_{ij}) > 0.23$, and $min(DockQ_i) > 0.23$, see Fig 1C. Overall, 70% of the good models are identified by all three measures, showing that for benchmarking, it is not crucial which measure is used. Overall, minimum $DockQ_i$ shows robustness in the prediction evaluation. Therefore, all our subsequent analysis involves only the minimum $DockQ_i$ for each complex.

Table 1. Success rates (i.e. the fraction of acceptable models with min ($DockQ_i$) ≥ 0.23) for protein complex predictions by AlphaFold-Multimer

	2mer	3mer	4mer	5mer	6mer
Homomer	55.5% (n=838)	59.3% (n=86)	56.0% (n=182)	60.0% (n=20)	45.5% (n=55)
Heteromer	72.9% (n=310)	52.2% (n=134)	45.9% (n=185)	26.2% (n=42)	42.1% (n=76)
Overall	60.2% (n=1148)	55% (n=220)	50.9% (n=367)	37.1% (n=62)	43.5% (n=131)

3.2. Performance of AlphaFold-Multimer

The success rate for AlphaFold multimer for each oligomeric state, using $DockQ_i > 0.23$, is reported in Table 1. Evaluations using other measures can be found in Table S1 and S2. The performance of AlphaFold multimer does not largely vary across oligomeric states, meaning that, still, for complexes with six chains, there is a 50% chance that the predicted protein complex is correct.

Analysis of other multichain prediction methods, FoldDock (Bryant et al., 2022), Haddock (Dominguez et al., 2003) and Omegafold (Wu et al., 2022), on a subset of the benchmark dataset, shows that for dimers Haddock and FoldDock makes 40% correct models, while OmegaFold makes less than 25% of good models (Table S3). For higher order multimers, FoldDock is worse than AlphaFold multimer but still makes 30-40% correct models, OmegaFold 10-15%, while Haddock is rarely successful (2-8%).

In Fig S3, we examine the difference in quality for proteins included in the training set of AlphaFold multimer or not (deposited into PDB before 2018-04-30 (Evans et al., 2022)). The difference between homomers before and after the dataset is insignificant (P-value=0.49 obtained using Mann–Whitney T-test). However, heteromeric structures show a significant difference (P-value=1.98e-17) with newer proteins being better than the ones used for training, i.e. there seems to be no indication that AlphaFold multimer is severely overtrained.

While distinguishing good ($DockQ_i > 0.23$ and MMscore > 0.75) and bad models ($DockQ_i$ <=0.23 and MMscore < 0.75), we found that the number of effective sequences in the paired alignment does not appear to significantly affect the quality of the models (Fig S2A). Most complexes with DockQi close to 0 have a Neff score of 0, but the inverse is false. Since we observe some heteromeric predictions with large differences among chains, we check if the high variance among chain lengths within one complex might be a major factor for performance. Fig S2B shows no significant separation in chain length variance between good and bad models. There are some failed cases with long disordered regions (i.e. residues in the SEQRES sequence missing from the ATOM record). However, some successful cases also have such long disordered regions, see Fig S2C. Symmetry does not appear to influence the quality of predicted models, see Fig S2D.

3.3. $pDockQ_i$ - estimate score for multi-chain predictions

In addition to per-structure quality estimation scores (pTM and ipTM), we believe there is a need to quantify the quality of each interface in a multi-chain complex prediction. It would be possible to use pDockQ for this, but it was found that this score is not ideal, see Fig 4A. Therefore, we developed $pDockQ_i$(see methods and Fig 4B), which takes the relative position of each chain concerning the others. A comparison between the confidence scores and quality measurement metrics ($DockQ_i$ and MMscore) is shown in Fig 4A. For per-complex comparisons, we use the minimum pDockQ and $pDockQ_i$ over all interfaces for each model. As expected, pTM and ipTM are better at predicting the MMscore than pDockQ and $pDockQ_i$, while $pDockQ_i$ is better at predicting the $DockQ_i$ score for heteromers.

Although $pDockQ_i$, in general, correlates well with $DockQ_i$, for some models, the difference is large, i.e. some wrong models are predicted to be good or vice versa, and we took a further step on the possible reasons to explain these outliers. Using $pDockQ_i < 0.1$ and $DockQ_i > 0.6$ resulted in 50 interfaces. By comparing different metrics, we found that these models have a big length difference between the SEQRES sequences and the PDB sequences ($p<0.01$ obtained using Mann–Whitney T-test) (see Fig S4). An example (PDB 6XWT) is shown in Fig S5A. The total SEQRES length of this hexamer is 2514 residues, whereas the PDB sequence length is only 352 residues. AlphaFold multimer successfully predicts the part of all chains which are aligned to the native structure well, and this produces $DockQ_i$ of 0.614. However, the overall estimation of the prediction gives low confidence, although $pDockQ_i$ only utilises the residues in the interface. Further, 2 out of 6 interfaces have average interface plDDT below 50, and the average interface PAE values of these two interfaces are high, 18.4 and 23.5, respectively. Though the partial prediction might be correct, the confidence of the whole predicted model would be affected. Thus $pDockQ_i$ for this prediction is 0.012.

For extreme cases with high $pDockQ_i$ (>0.8) but low $DockQ_i$ cases (<0.1), 145 interfaces satisfy this condition. By manually checking the predicted models, the most prominent finding is that usually individual chains are predicted quite accurately, but the docking positions are wrong compared with the native structures. However, the high $pDockQ_i$ values indicate that AlphaFold multimer is quite confident regarding the interface residues. Apart from this, 78% of the pTM scores for the complexes are above 0.8. Thus, other possible explanations might be needed. For example, the number of residues in contacts at the interface in the native structure of PDB 6JBD (Fig S5B) and PDB 5XLL (Fig S5C) is far fewer than those in the AlphaFold multimer prediction. We checked the original reference paper for the structure (Kita et al., 2020). We found that the dimer was cut out of a crystallisation structure of pantoate kinase, which has twofold homodimers. The AlphaFold multimer prediction gives the homodimer interface if one cuts the tetramer along the active binding site. In other words, the biological assembly from PDB might not represent the only biological assembly. In total, we encounter 65 such cases, which might be interesting to examine in detail, Table S4.

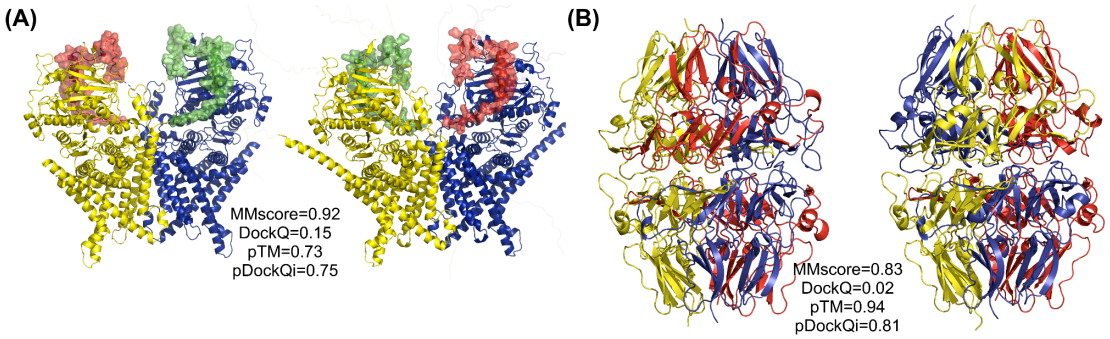

Fig 3. Different domain positions in the native (left panels) and model (right panels) structures. (A) Chains A and B of homodimer from PDB 7STL are shown as yellow and blue cartoons, correspondingly, and domains consisting of residues 868 - 914 are shown as semi-transparent green and red surfaces for chains A and B, respectively. (B) Chains A (lower structures) and B (upper structures) of a homodimer from PDB ID 6WKU. Each chain results from the fusion of three proteins with UNIPROT IDs Q01955 (residues 16 - 243, yellow cartoons), P53420 (residues 246 - 468, red cartoons), and P29400 (residues 471 - 695, blue cartoons). Disordered (flexible) parts of the model structures, which are missing in the native structures, but were modelled by AlphaFold multimer, are omitted for clarity.

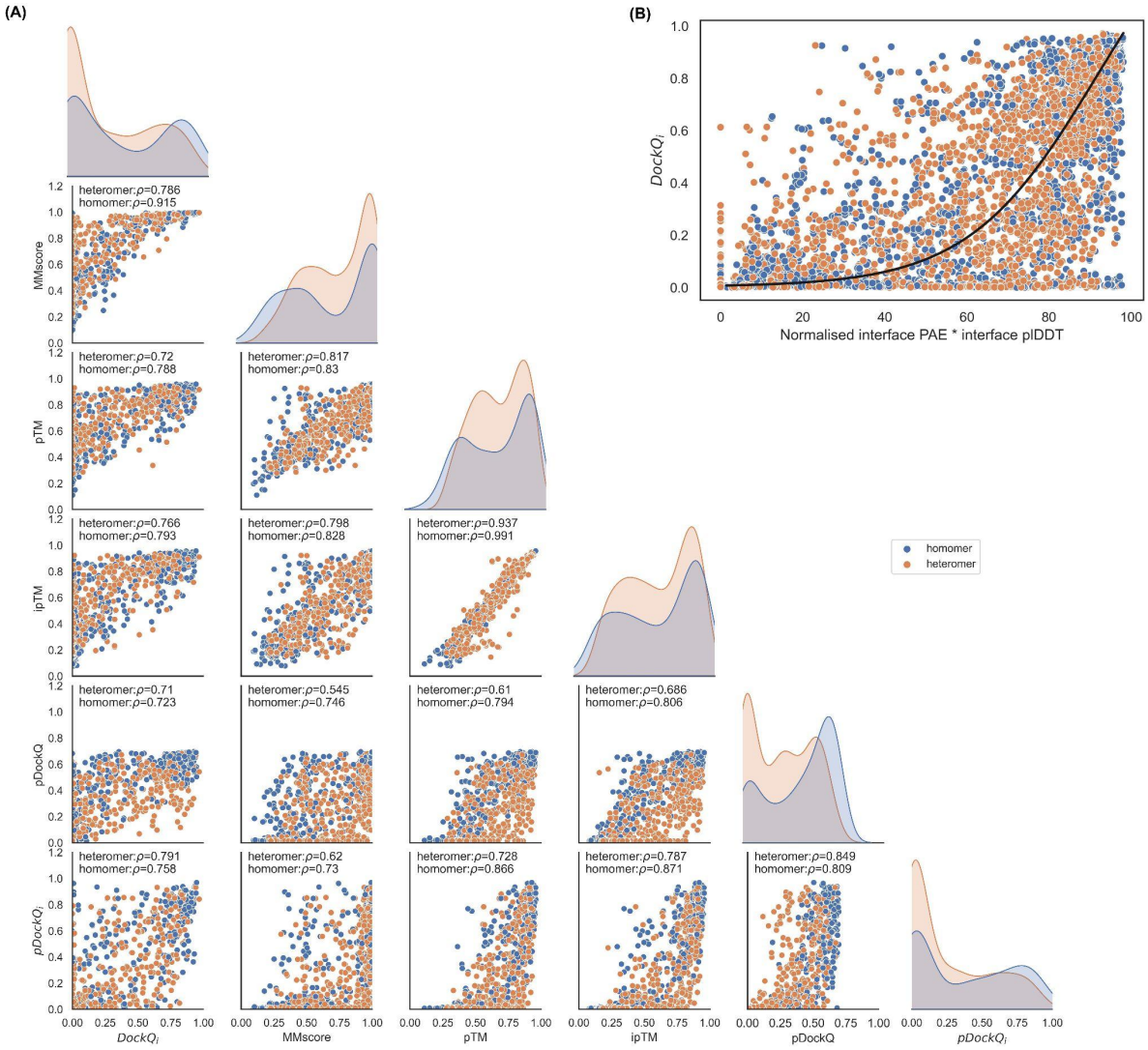

Fig 4. (A) Pairplot showing correlations between $DockQ_i$, MMscore and predicted scores (ipTM, pTM, pDockQ and $pDockQ_i$). For per-interface scores($DockQ_i$, pDockQ and $pDockQ_i$), we take the minimum value over all interfaces for each predicted complex. (B) We fit a sigmoid curve using the combining normalised interface PAE and interface plDDT to $DockQ_i$ scores on our benchmark dataset.

3.4. High confidence CORUM predictions

Now we asked if AlphaFold multimer can be used to predict truly novel structures by turning to the CORUM database. Out of the 29 potentially novel complexes (Table S5) that were predicted by AlphaFold multimer (see methods), 8 had pTM > 0.5 and $pDockQ_i$ > 0.23, see Fig S6. 2 highly confident complexes (pTM > 0.75 and $pDockQ_i$ > 0.23) were obtained. Monomers of these complexes were then run with FoldSeek (van Kempen et al., 2022) against PDB. The PAC1-PAC2 complex (CORUM Complex ID: 3034) contains two chains, with chain A having a hit to 3GAA and chain B having a hit to 7LS6. However, the sequence identity is below 30% (see Fig 5A). The Metaxin complex (Mtx1, Mtx2) complex (CORUM Complex ID:3094) is also a dimer (Fig 5B) and is found in mice. Foldseek found a match to 6WUM with chain A and 6WUT to chain B. However, these hits (6WUM and 6WUT) are membrane proteins and part of the mitochondrial SAM complexes in Thermothelomyces thermophilus.

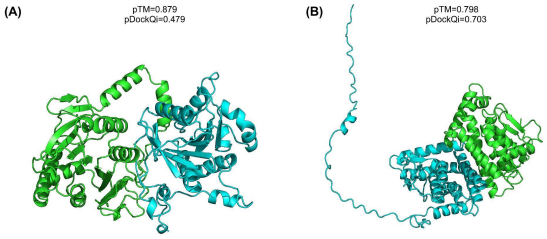

Fig 5. (A) Predicted structure of CORUM ID 3034 (B) Predicted structure for CORUM ID 3094

4. Conclusion

By preparing an independent homology-reduced dataset for benchmarking the performance of protein complex predictors, we have shown that taking the minimum $DockQ_i$ over all interfaces is a useful way to evaluate the quality of the multimeric complex. Also, the performance of AlphaFold multimer (Evans et al., 2022) slightly decreases as the size of the complex increases, i.e. the number of chains in the complex increases, and consistent with the AlphaFold multimer study (Evans et al., 2022), homomeric complex prediction outperforms heteromeric complex prediction. By assessing the quality of the models with DockQ (Basu & Wallner, 2016) and MMscore (Mukherjee & Zhang, 2009), we show that for homomeric models, they are almost exclusively either completely correct or completely wrong, while for heteromeric complexes, there are cases where one or a few of the chains are incorrectly placed while the larger part of the complex is correct. We also provide a modified version of pDockQ, the $pDockQ_i$ score for estimating the quality of an individual chain in a predicted multimer model. Lastly, we evaluate the chain-level predictions for highly confident structures using $pDockQ_i$ obtained from CORUM using AlphaFold multimer. Here we present a structure of the Metaxin complex (Mtx1, Mtx2) complex having no detectable homology to any PDB structure.

Acknowledgements

We thank Gabriele Pozzatti and Patrick Bryant for valuable discussions.

Funding

AE was funded by the Vetenskapsrådet Grant No. 2021-03979 and Knut and Alice Wallenberg foundation. The computations/data handling was enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation and SNIC, grant No: SNIC 2021/5-297 and Berzelius-2021-29.
Conflict of Interest: none declared.

References

Basu, S., & Wallner, B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. *PloS One*, *11*(8), e0161879.

Bergendahl, L. T., & Marsh, J. A. (2017). Functional determinants of protein assembly into homomeric complexes. *Scientific Reports*, *7*(1), 4932.

Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), 1265.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The Protein Folding Problem. *Annual Review of Biophysics*, *37*, 289–316.

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

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

Gao, M., Nakajima An, D., Parks, J. M., & Skolnick, J. (2022). AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nature Communications*, *13*(1), 1744.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., & Ruepp, A. (2019). CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*, *47*(D1), D559–D563.

Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters*, *76*(4), 175–181.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I., Wodak, S. J., & Critical Assessment of Predicted Interactions. (2003). CAPRI: a Critical Assessment of Predicted Interactions. *Proteins*, *52*(1), 2–9.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 1–7.

Kita, A., Kishimoto, A., Shimosaka, T., Tomita, H., Yokooji, Y., Imanaka, T., Atomi, H., & Miki, K. (2020). Crystal structure of pantoate kinase from Thermococcus kodakarensis. *Proteins*, *88*(5), 718–724.

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

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682.

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

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175.

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Söding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. In *bioRxiv*. https://doi.org/10.1101/2022.02.07.479398

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022). High-resolution de novo structure prediction from primary sequence. In *bioRxiv*. https://doi.org/10.1101/2022.07.21.500999

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–230