Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments

P. Bryant[1,2]*, G. Pozzati[1,2]* and A. Elofsson[1,2]
***contributed equally**
1) Science for Life Laboratory, 172 21 Solna, Sweden
2) Department of Biochemistry and Biophysics, Stockholm University, 106 91 Stockholm, Sweden
Corresponding Author: Patrick Bryant, email: patrick.bryant@scilifelab.se
Postal address: Box 1031, 171 62 Solna, Sweden
ORCID
PB: https://orcid.org/0000-0003-3439-1866
GP: https://orcid.org/0000-0002-4303-9939
AE: https://orcid.org/0000-0002-7115-9751

Abstract
Predicting the structure of single-chain proteins is now close to being a solved problem due to the recent achievement of AlphaFold2 (AF2). However, predicting the structure of interacting protein chains is still a challenge. Here, we utilise AF2 to optimise a protocol for predicting the structure of heterodimeric protein complexes using only sequence information. We find that using the default AF2 protocol, 32% of the models in the Dockground test set can be modelled accurately. By tuning the input alignment and identifying the best model, we adjusted the performance to 43%. Our protocol uses MSAs generated by AF2 and MSAs paired on the organism level generated with HHblits. In a more extensive, more realistic, independent test set, the accuracy is 59%. In comparison, the alternative fold-and-dock method RoseTTAFold is only successful in 10% of the cases on this set and traditional docking methods 22%. However, for the traditional method, the performance would be lower if the bound form of both monomers was not known. The success is higher for bacterial protein pairs, pairs with large interaction areas consisting of helices or sheets, and many homologous sequences. We can distinguish acceptable (DockQ>0.23) from incorrect models with an AUC of 0.84 on the test set by analysing the predicted interfaces. At an error rate of 1%, 13% are acceptable (at a 10% error rate, 40% of the models are acceptable). All scripts and tools to run our protocol are freely available at: https://gitlab.com/ElofssonLab/FoldDock.

Introduction

Protein-protein interactions are central mediators in biological processes. Most interactions are governed by the three-dimensional arrangement and the dynamics of the interacting proteins[1]. Such interactions vary from being permanent to transient[2,3]. Some protein-protein interactions are specific for a pair of proteins, while some proteins are promiscuous and interact with many partners. This complexity of interactions is a challenge both for experimental and computational methods.

Often, studies of protein-protein interactions can be divided into two categories, the identification of what proteins interact and the identification of how they interact. Although these problems are distinguished, some methods have been applied to both problems [4,5]. Protein docking methodologies refer to how proteins interact and can be divided into two categories; those based on shape complementarity[6] and those based on alignments (both sequence and structure) to structural templates[7]. Shape complementary approaches rely on protein structures or models of the monomers[8,9], while template-based docking needs suitable templates. However, flexibility has often to be considered in protein docking to account for interaction-induced structural rearrangements[10,11]. Therefore, flexibility limits the accuracy achievable by rigid-body docking[12], and flexible docking is traditionally too slow and inaccurate for large scale applications.

Regardless of different strategies, docking remains a challenging problem. In the CASP13-CAPRI experiments, human group predictors achieved up to 50% success rate for top-ranked docking solutions[13]. Alternatively, a recent benchmark study[8] reports success rates of different web-servers reaching up to 16% on the well known Benchmark 5 dataset[14].

Recently, in the CASP14 experiment, AlphaFold2 (AF2) reached an unprecedented performance level in structure prediction of single-chain proteins[15]. Thanks to an advanced deep learning model that efficiently utilises evolutionary and structural information, this method consistently outperformed all competitors, reaching an average GDT_TS score of 0.9[15]. Recently, RoseTTAFold was developed, trying to implement similar principles[16]. Since then, other end-to-end structure predictors have emerged using different principles such as fast MSA processing in DMPFold2[17] and language model representations[18].

As an alternative to other docking methods it is possible to utilise co-evolution to predict the interaction between two protein chains. Initially, direct coupling analysis was used to predict the interaction of bacterial two-component signalling proteins [19,20]. Later, these methods were improved using machine learning[21].

We recently developed a Fold and Dock pipeline using another distance prediction method focused on protein folding (trRosetta[22]). In this pipeline, the interaction between two chains from a heterodimeric protein complex and their structures were predicted using distance and angle constraints from trRosetta[23,24]. In a Fold and
Fock approach, the two proteins are folded and docked simultaneously. This study demonstrated that a pipeline focused on intra-chain structural feature extraction can be successfully extended to derive inter-chain features as well. Still, only 7% of the tested proteins were successfully folded and docked.

We found that generating the optimal MSA is crucial for obtaining accurate Fold and Dock solutions. However, this is not trivial due to the necessity to identify the exact set of interacting homologous pairs[25], see Figure 1. Given the existence of multiple paralogs for most eukaryotic proteins, this is difficult. We also found that this process requires an optimal MSA depth to optimise inter-chain information extraction. Too deep MSAs might contain false positives (i.e. protein pairs that interact differently), resulting in noise masking the sought after co-evolutionary signal. At the same time, too shallow ones do not provide sufficient co-evolutionary signals.

Here, we systematically applied the AF2 pipeline to Fold and Dock protein-protein pairs from two different datasets simultaneously. We explore the docking success using the MSAs generated by AF2 and combine them with MSAs paired on the organism level to study the dependence of AF2 on the input MSAs. We find that the results in terms of successful docking using AF2 are superior to all other docking methods.

Material and Methods

Data

Development set

A set of heterodimeric complexes from Dockground benchmark 4[26] is used to develop the pipeline presented here. This set contains protein pairs, with each chain having at least 50 residues, sharing less than 30% sequence identity and no crystal packing artefacts. There are 219 protein interactions for which both unbound (single-chain) and bound (interacting chains) structures are available. Unbound chains share at least 97% sequence identity with the bound counterpart and, to facilitate comparisons, non-matching residues are deleted and renumbered to become identical to the unbound counterpart. Some structures failed to be modelled for various reasons (see limitations of data generation), resulting in a total of 216 structures. The dataset consists of 54% Eukaryotic proteins, 38% Bacterial and 8% from mixed kingdoms, e.g. one bacterial protein interacting with one eukaryotic.

Test set

We used 1,661 protein complexes with known interfaces from a recent study[27] to test the developed pipeline. Here, three large biological assemblies were excluded. These complexes share less than 30% sequence identity, have a resolution between 1-5 Å and constitute unique pairs of PFAM domains (no single protein pair have PFAM domains matching that of any other pair). Some structures failed to be modelled for various reasons (see limitations of data generation), resulting in a total of 1481 structures. These proteins are mainly from H. Sapiens (25%), *S. Cerevisiae* (10%), *E.coli* (5%) and other Eukarya (30%).

CASP14 set and novel protein complexes

We used a set of six heterodimers from CASP14 to test the pipeline further. In addition, we extracted eight novel protein complexes deposited in PDB after 15 June 2021, which produced no results for at least one chain in each complex when submitted to the HHPRED web server (version 01-09-2021)[28,29]. We selected this set to ensure our pipeline is tested on data AF2 is guaranteed not to have seen, see Table S1.

MSA generation

The input to Alphafold2 (AF2) consists of several MSAs. We used the AF2 MSA generation[15], which builds three different MSAs generated by searching the Big Fantastic Database[30] (BFD) with HHBlits[31] (from hh-suite v.3.0-beta.3 version 14/07/2017) and both MGnify v.2018_12[32] and Uniref90 v.2020_01[33] with jackhmmer from HMMER3[34]. The BFD was created in the following manner: *"by clustering 2.5 billion protein sequences from Uniprot/TrEMBL+Swissprot[35], Metaclust[36] and Soil Reference Catalog Marine Eukaryotic Reference Catalog assembled by Plass[37]. We clustered sequences that could be aligned to a longer sequence with 90% of their residues and a sequence identity of 30% using Linclust/MMseqs2[38]* with the options:

```
--cov-mode 1 --min-seq-id 0.3
```

We removed all clusters with less than three sequences and turned the database into an HH-suite3[31] database using the Uniclust pipeline[39]." The AF2 MSAs were generated by supplying a concatenated protein sequence of the entire complex to the AF2 MSA generating pipeline in FASTA format. The resulting MSAs will thus mainly contain gaps for one of the two query proteins in each row, as only single chains can obtain hits in the searched databases (Figure 1). No trimming or gap removal was performed on these MSAs.

In addition to the AF2 MSAs, we used a "paired MSA", constructed using organism information, as described before[4,20,23] (Figure 1). The rationale behind using a paired MSA is to identify inter-chain coevolutionary information. An unpaired MSA has a limited inter-chain signal since the chains are treated in isolation (Figure 1).

The paired MSA was constructed by running HHblits[31] version 3.1.0 against uniclust30_2018_08[39] with these options:

```
hhblits -E 0.001 -all -oa3m -n 2
```

Both chains were run independently. From the resulting a3m files, the organism information was extracted using the OX identifier. Next, all hits with over 90% gaps were removed. From all remaining hits in the two MSAs, the highest-ranked hit from one organism was paired with the highest-ranked hit of the interacting chain from the same organism. Here, the highest-ranked hits are assumed to be the most likely orthologs to the query proteins. Pairing the correct sequences results in paired MSAs containing co-evolutionary information across different chains and an inter-chain signal[27].

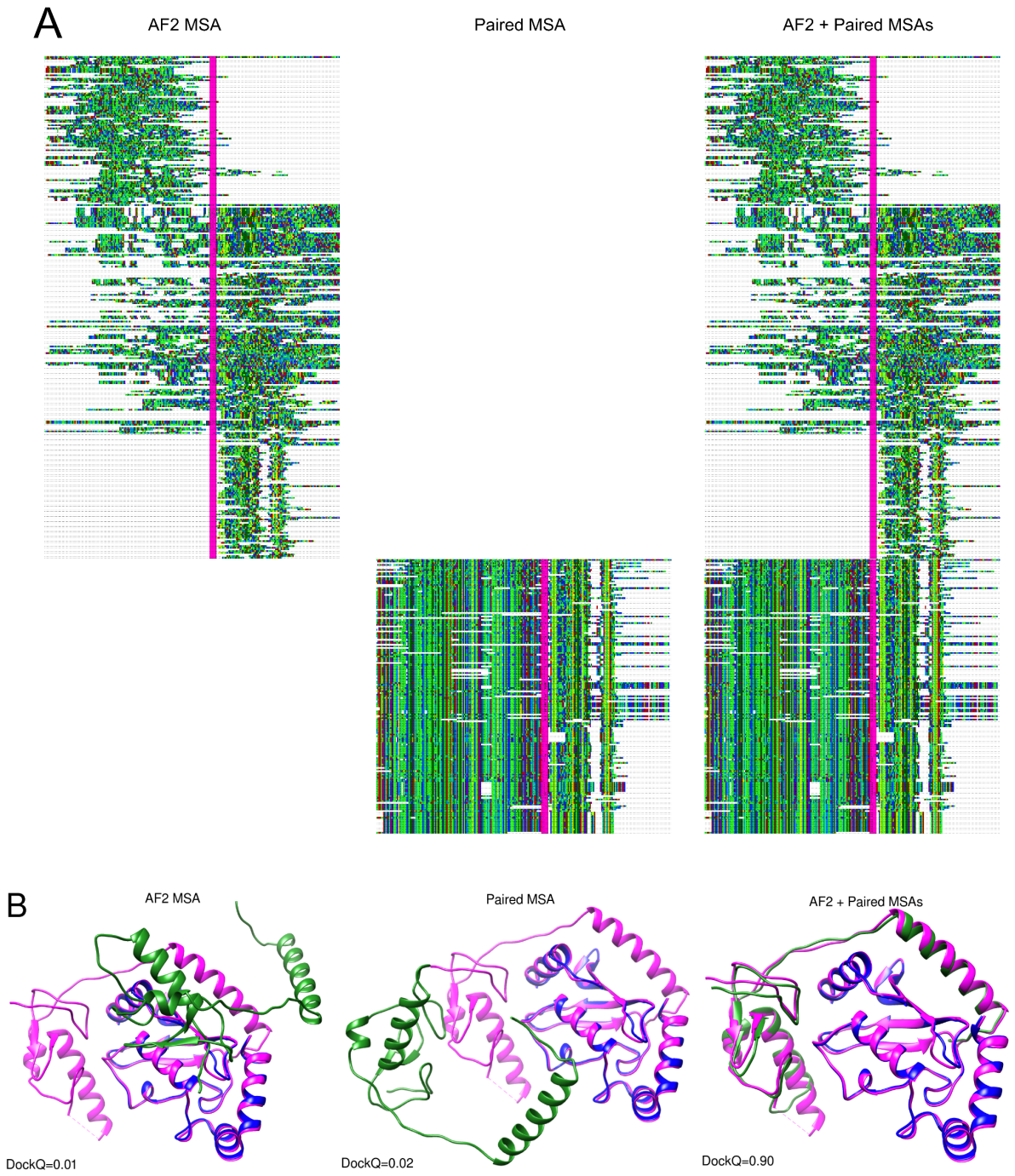

Figure 1. A) Depiction of MSAs generated by AF2 and the paired version matched using organism information. Both AF and paired representations are sections containing 10% of the sequences aligned in the original MSA. Concatenated chains are separated by a vertical line (magenta). The visualisations were made using Jalview version 2.11.1.4[40] **B)** Docking visualisations for PDB ID 5D1M chains A (blue) and B (green) using the three different MSAs in A**.** The native structure chains are displayed superimposed with the predictions in magenta. The DockQ scores are 0.01, 0.02 and 0.90 for AF2, paired, and AF2+paired MSAs, respectively.

Number of effective sequences (Neff)

To estimate the information in each MSA, we calculated the Neff score by clustering sequences at 62% identity, as used in a previous study[41]. Unaligned FASTA sequences were extracted from the three AF2 default MSAs. Obtained sequences were processed with the CD-HIT software[42] version 4.7 (http://weizhong-lab.ucsd.edu/cd-hit/) using the options -c 0.62 -G 0 -n 3 -aS 0.9. Due to the manipulation performed on paired MSAs, sequences extracted in these cases were clustered separately. In both cases, the number of obtained clusters have been adopted as Neff scores.

Structure prediction

AlphaFold2

We modelled complexes using AlphaFold2[15] (AF2) by modifying the script https://github.com/deepmind/alphafold/blob/main/run_alphafold.py to insert a chain break of 200 residues - as suggested in RoseTTAFold[16] (RF). During modelling, relaxation was turned off, and only the atoms generated in RF (N, CA, C) were used in subsequent analyses. Sidechains were thus not used to score interfaces. We note that performing model relaxation did not increase performance in the AF2 paper[15] and was ignored to save computational cost. No templates were used to build structures, as this would not assess the prediction accuracy of unknown structures or structures without sufficient matching templates. AF2 has been shown to perform well for single chains without templates and has reported higher accuracy than template-based methods even when robust templates are available[15].

We supplied three different types of MSAs to AF2: the MSAs generated by using the default AF2 settings, the top paired MSAs constructed using HHblits, described above, and finally, a concatenation of these both alignments. AF2 was run with two different network models, AF2 model_1 (used in CASP14) and AF2 model_1_ptm, for each MSA. The second model, model_1_ptm, is a fine-tuned version of model_1 that predicts the TMscore[43] and alignment errors[15]. We ran these two different models by using two different configurations. The configurations utilise a varying amount of recycles and ensemble structures. Recycle refers to the number of times iterative refinement is applied by feeding the intermediate outputs recursively back into the same neural network modules. At each recycling, the MSAs are resampled, allowing for new information to be passed through the network. The number of ensembles refers to how many times information is passed through the neural network before it is averaged[15]. The two configurations used are; the CASP14 configuration (three recycles, eight ensembles) and an increased number of recycles (ten) but only one ensemble.

Since structure prediction with AF2 is a non-deterministic process, we generate five models initiated with different seeds. To save computational cost, this was only performed for the best modelling strategy. We rank the five models for each complex by the number of residues in the interface, giving the best result.

RoseTTAFold

For comparison, the RoseTTAFold (RF) end-to-end version[16] was run using the paired MSAs with the top hits. The RoseTTAFold pipeline for complex modelling only generates MSAs for bacterial protein complexes, while the proteins in our test set are mainly Eukaryotic. Therefore, we use the paired alignments here. We compare RF with AF2 using the same inputs (the paired MSAs) for both the development and test datasets to provide a more fair comparison, as AF2 searches many different databases to obtain as much evolutionary information as possible when generating its MSAs. To predict the complexes, we use the "chain break modelling" as suggested in RF (https://github.com/RosettaCommons/RoseTTAFold/tree/main/example/complex_modeling) using the following command:

```
predict_complex.py -i msa.a3m -o complex -Ls chain1_length chain2_length
```

GRAMM

For comparison, a rigid-body docking method, GRAMM[44], was used. Here, structure pairs are processed with a Fast Fourier Transform (FFT) procedure to generate 340000 docking poses for each complex. The bound structures extracted from complexes in the test set were used as inputs. This docking generation stage mainly considers the geometric surface properties of the two interacting structures, allowing minor clashes to leave some space for conformational flexibility adjustment. As the bound form of the proteins is used, this should represent a straightforward use case for GRAMM based docking, and performance drops typically when unbound structures or models are used [45]. The atom-atom contact energy AACE18 is used to score and rank all poses, as this has been shown to provide better results than shape-complementarity alone[46]. We select the top candidates and compare them with the models generated by AF2.

Scoring

The backbone atoms (N, CA and C) were extracted from the predicted AF2 structures (as these are the only predicted atoms in the end-to-end version of RF). The interface scoring program DockQ[47] was then run to compare the predicted and true interfaces. This program compares interfaces using a combination of three different CAPRI[48] quality measures (F_{nat}, LRMS, and iRMS) converted to a continuous scale, where an acceptable model comprises a DockQ score of at least 0.23.

Ranking models

To analyse the possibility of determining when a predicted model is acceptable (DockQ≥0.23) in AF2, we analyse the separation between acceptable and incorrect models as a function of different metrics on the development set: the number of unique interacting residues (CBs from different chains within 8 Å from each other), the total number of interactions between CBs from different chains, average predicted lDDT (plDDT) score from

AF2 for the interface, the minimum of the average plDDT for both chains and the average plDDT over the whole heterodimer.

We use these metrics as a threshold to build a confusion matrix, where True/False Positives (TP and FP respectively) are correct/incorrect docking models which places above the threshold and False/True Negatives (FN and TN respectively) are correct/incorrect docking models which places below the threshold. From the built confusion matrix we derive the True Positive Rate (TPR) and False Positive Rate (FPR) defined as:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Then, we calculate TPR and FPR for each possible value assumed by the set of dockings given a single metric and plot TPR as a function of FPR in order to obtain the Receiver Operating Characteristic (ROC) curve. To compare different metrics, we compute the Area Under Curve (AUC) for ROC curves obtained for each metric.

The TPR and FPR for different thresholds are used to calculate the fraction of models that can be called correct out of all models, as well as the Positive Predictive Value (PPV). The fraction of acceptable and incorrect models are obtained by multiplying the TPR and FPR with the success rate. Multiplying the FPR with the success rate results in the FDR and the PPV can be calculated by dividing the fraction of acceptable models with the sum of the acceptable and incorrect models (Table S2).

Analysis of models

To analyse the possibility of determining when AF2 can model a complex correctly, we analyse the structures and the multiple sequence alignments. We investigated: the Number of effective sequences (Neff)**,** the secondary structure in the interface annotated using DSSP[49], the length of the shortest chain, the number of residues in the interface and the number of contacts in the interface.

DSSP was run on the entire complexes, and the resulting annotations were grouped into three categories; helix (3-turn helix (3_{10} helix), 4-turn helix (α helix) and 5-turn helix (π helix)), sheet (extended strand in parallel or antiparallel β-sheet conformation and residues in isolated β-bridges) and loop (residues which are not in any known conformation).

Limitations of data generation

For the development set, AF2 MSAs could not be generated for three of the complexes due to memory limitations (1gg2, 2nqd and 2xwb) using a computational node with 128 Gb RAM for the MSA generation and were thus disregarded, resulting in a total of 216 complexes.

107 of the complexes in the test set lack CBs, and 50 have overlapping PDB codes with the development set and were therefore excluded. In the MSA generation from AF2, 20 MSAs report MergeMasterSlave errors regarding discrepancies in the number of match states,

resulting in a total of 1484 AF2 MSAs. When folding, three of these (5AWF_D-5AWF_B, 2ZXE_B-2ZXE_A and 2ZXE_A-2ZXE_G) report “ValueError: Cannot create a tensor proto whose content is larger than 2GB”, leading to a final set of 1481 complexes. DSSP could only be run successfully for 1391 out of the 1481 protein complexes, and we ignored the rest in the analysis.

For RF, 26 complexes produced out of memory exceptions during prediction using a GPU with 40 Gb RAM and were excluded from the RF analyses, leaving 1455 complexes.

Results and Discussion

Performance on the development set

The fraction of acceptable models (DockQ>0.23), the so-called success rate (SR), is used to measure performance for each combination of MSA and AF2 model. The best performance is 26.4% for the paired MSAs, 32.4% for the AF2 MSAs and 38.4% for the AF2+paired MSAs (Table 1). It is thereby evident that combining both paired and AF2 MSAs is superior to using them separately. Combining AF2 and paired MSAs improves the results, likely related to the frequent complementarity observed when comparing the predicted DockQ scores from the paired MSAs with those from the AF2 MSAs (Table S3). It is clear that the average performance is similar, but for individual pairs, one of the two MSAs is superior to the other. The Pearson correlation coefficient for the DockQ scores between AF2 vs paired MSAs is 0.48, while 0.63 comparing AF2 with AF2+paired (Table S3).

Next, we compared the two AF2 network models and the fine-tuned versions of the AF2 model (model_1 vs AF2 model_1_ptm). The original AF2 model_1 outperforms AF2 model_1_ptm in all cases but for the paired MSAs, where ten recycles in AF2 model_1_ptm performs equally well as model 1 (Table 1). Further, the difference between 10 recycles-one ensemble and three recycles-eight ensembles is minor across all MSAs and AF2 models. Therefore, the input information and the AF2 model appears to impact the outcome the most.

Table 1. Results from AF2 using different MSAs and neural network configurations. Paired MSAs are alignments produced relying on organism identifier information (Uniprot OX codes, detailed description in Methods), for which pairs of homologs likely to interact on their own are aligned to query chains. AF2 refers to the default method procedure, i.e. concatenating results from three different databases. Finally, combining the two MSAs was tested (AF2+Paired). The best neural network configuration for each MSA is marked in bold. In all cases, does the AF2+paired MSA option outperforms the others?

Success rate (DockQ≥0.23) of modeled structures (%)				
Neural network configuration				
NN model	model_1	model_1	model_1_ptm	model_1_ptm
Recycles	10	3	10	3

Ensembles	1	8	1	8
MSA \ short name	m1-10-1	m1-3-8	mp-10-1	mp-3-8
Paired	**26.4**	**26.4**	**26.4**	24.5
AF2	31.0	**32.4**	24.1	23.1
AF2+Paired	**38.4**	36.6	30.6	30.1

Distinguishing acceptable from incorrect models

It is not only essential to obtain improved predictions but also to be able to identify acceptable predictions. We measure the separation between acceptable and incorrect models using a receiver operating characteristic (ROC) curve. Different criteria were examined, including (i) the number of unique interacting residues (CBs from different chains within 8 Å from each other) in the interface, (ii) the total number of interactions between CBs in the interface, (iii) the average plDDT for the interface, (iv) the lowest plDDT of each single chain average, and (v) the average plDDT over the whole protein heterodimer (Figure 2A). The best separation is obtained when using the total number of interactions between CBs in the interface, resulting in an area under the curve (AUC) of 0.91. Interestingly, the plDDT of the single chains only results in an AUC of 0.63, suggesting that the single chains in a complex are often predicted very well. At the same time, their relative orientation to the partner is wrong.

Model variation and ranking

Five models were generated using the best strategy (m1-10-1 with AF2+paired MSAs) with different initialisation random seeds. The average SR (36.9% ± 1.3%) was similar for all five runs. However, the average deviation for individual models is DockQ=0.13 when comparing the best and worst models for a target (Figure 2B), i.e. there is some randomness to the success for an individual pair. If the maximal DockQ score across all models is used, the SR would be 46.3%. Although this is unachievable, ranking the models using the total number of interactions in the interface results in an SR of 43.1%, an increase of 4.7% compared to no model ranking.

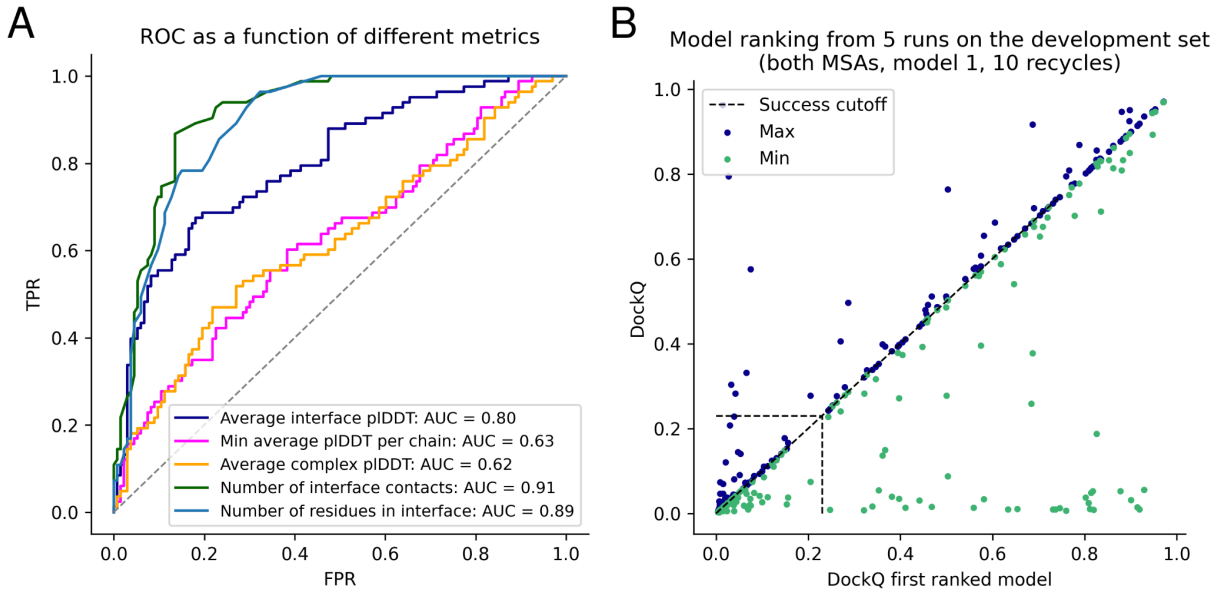

Figure 2. A) ROC curve as a function of different metrics for the development dataset (first run). CBs within 8 Å from each other from different chains are used to define the interface. **B)** Impact of different initialisations on the modelling outcome in terms of DockQ score on the development dataset. The maximal and minimal scores are plotted against the top-ranked models using the total number of interactions in the interface.

While estimating docking quality, the pIDDT metric produced by the AF2 predictor does not provide the best separation between acceptable and incorrect docking models. However, this is an excellent estimator for single-chain accuracy. This is likely a consequence of the original AF2 focus on single-chain folding[15]. As mentioned above, the separation of acceptable and incorrect models is higher when considering predicted interfaces' features and not the single chains (Figure 2A). Interestingly, no additional constraints are implemented to pull two chains in contact, meaning chain interactions (and subsequently interface sizes) are exclusively determined by the amount of inter-chain signals extracted by the predictor. Assuming that all residues in an interface contribute to the interaction energy could explain why more extensive interfaces are more likely to be correctly predicted.

Performance on the test set

The best strategy on the test set results in an SR of 55.9% (828 out of 1481 correctly modelled complexes) compared with 43.9% using the AF2 MSAs (Figure 3, Table S4). Further, running five initialisations with random seeds and ranking the models using the total number of interactions in the interface increases the SR to 59.0% (Figure S1). In contrast, the paired MSAs (48.1% SR) outperform the AF2 MSAs (43.9%) in the development set. The AUC using the total number of interactions between CBs in the interface for the ranked test set is 0.84, which means that 13% of all models can be called acceptable at an error rate of 1% and 48% at an error rate of 10% (Table S2).

Using both AF2 and paired MSAs increase performance suggests that AF2 gains both from larger and from paired MSAs containing co-evolutionary information across different chains, although it often can manage with only one type of alignment. The performance of the paired

alignments on the test set is slightly higher with the paired MSAs (43.9% vs 48.1%), concluding that AF2 is indeed more likely to be acceptable if the information supplied is more ordered.

RF outperforms AF2 only for 14 pairs in the test set, while GRAMM outperforms AF2 for 188 pairs. The reason for GRAMM's good performance is likely due to the use of the bound form of the proteins, resulting in very high shape complementarity and therefore having the "answer" provided in a way.

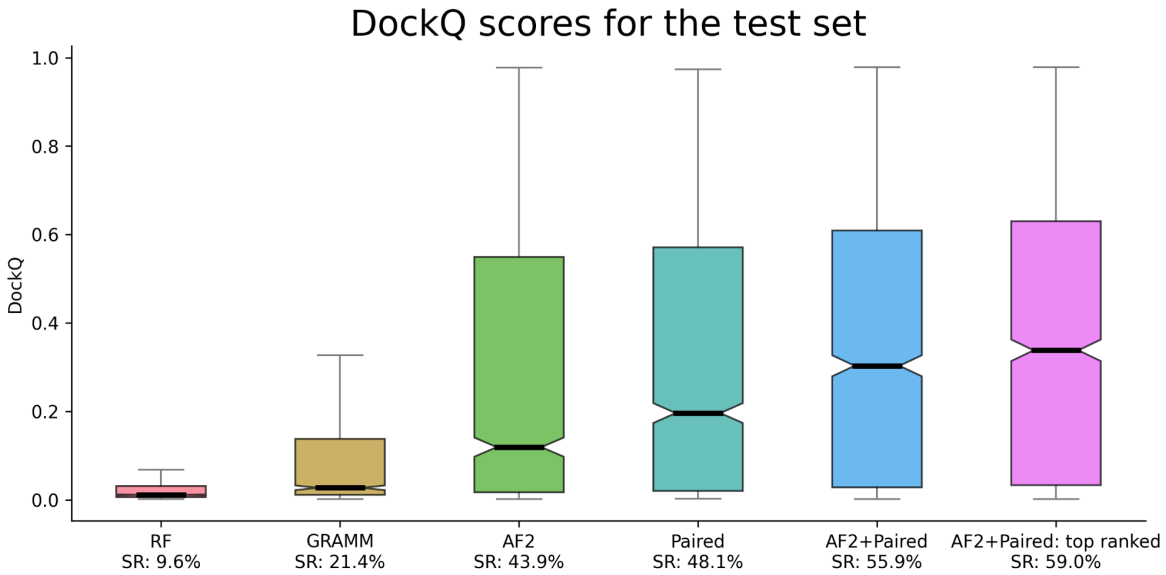

Figure 3. Distribution of DockQ scores as boxplots for different modelling strategies on the test set. The boxes encompass the quartiles of the data, while the notches and horizontal lines mark the medians. The success rates (SR) are reported below the name of each method. All AF2 models have been run with the same neural network configuration (one ensemble, ten recycles). Outlier points are not displayed here.

Bacterial protein pairs with large interfaces and many homologs are easier to predict

In the test set, about 60% of the complexes can be modelled correctly. We tried to answer why this is by analysing different subsets of the test set. First, we divided the set by taxa, interface characteristics, and finally by examining the alignments.

When comparing the performance per organism on the test set, the SRs for *Homo Sapiens* and *S.cerevisiae* (both Eukarya) are similar (58% vs 59%), while the performance is higher in *E.coli* (75%, part of the kingdom Bacteria). The SRs per kingdom follow the same pattern as the organisms, with Eukarya having 57% SR, Bacteria 72 %, Archaea 80% and Virus 55% (Figure S3 B). This is consistent with previous observations regarding the availability of evolutionary information in the more ancient kingdoms of Bacteria and Archaea compared to Eukarya. It has been observed that it is easier to obtain both deeper and more accurate

MSAs[27] for bacteria than Eukarya. As a result, Bacteria and Archaea also have higher Neffs than Eukarya (Figure S3A).

Next, we examined the interfaces. First, different secondary structural content of the native interfaces was investigated (Figure 4A). The highest SR is obtained for mainly helix interfaces (62%), followed by mainly sheet interfaces (59%). The loop interface SR of 53% is substantially lower than the others, suggesting that interfaces with more flexible structures are harder to predict. To compare the importance of the size of the interface, these were divided into tertiles, and the SR was computed for each subgroup (Figure 4B). Here, it is clear that pairs with more extensive interfaces are easier to predict, as the SR increases from 47 to 74% between the bottom and top tertiles.

Next, we examined how the size of the MSA (both paired and AF2) influences the results. It seems as the Neff of the paired MSA (Figure 4C) has a more significant influence on the outcome than the size of the AF2 MSA (Figure S2A). Although some proteins with few sequences in the MSA (Neff<64) can be modelled correctly, it is also clear that the fraction of correctly modelled sequences increases with larger MSAs (Figure 4C).

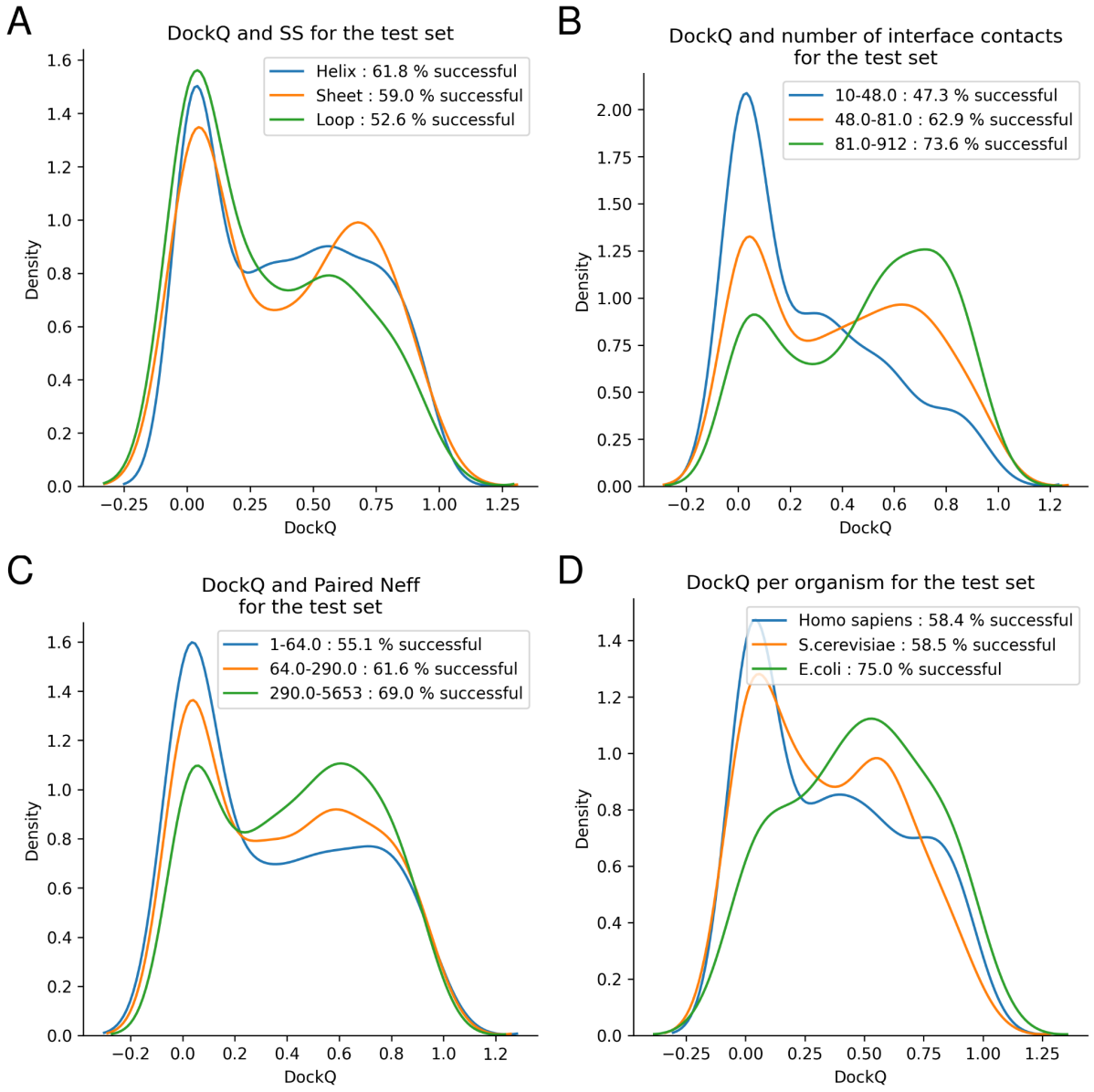

Figure 4. A) Distribution of DockQ scores for three sets of interfaces with the majority of Helix, Sheet and Coil secondary structures. While Helix and Sheet obtain comparable results, loop interfaces seem to be harder to predict. **B)** Distribution of DockQ scores for tertiles derived from the distribution of contact counts in docking model interfaces. Larger interfaces indicate consistently higher chances of success (73.6%). **C)** Distribution of DockQ scores for tertiles derived from the distribution of Paired MSAs Neff scores. Deeper alignments succeed in more cases than smaller alignments. **D)** Distribution of DockQ scores for the top three organisms *Homo Sapiens*, *S. cerevisiae* and *E. coli*. The success rate of *Homo Sapiens* and *S. cerevisiae* are comparable (58.4 and 58.5 %), while that of *E. coli* is much higher (75.0 %).

Success rate discrepancy between dev and test sets

The higher SR on the test set compared to the development set suggests that the test set contains proteins which structures are easier to predict. To analyse why this is, we compare the features that discriminate between acceptable and incorrect models. None of the features that discriminate between acceptable and incorrect models on the test set account for the SR discrepancy between the development and test sets.

The development set contains fewer eukaryotic sequences (43 vs 59%) but an almost equal proportion of bacterial sequences (22 vs 25%). There are no substantial differences in the number of interface contacts, chain lengths or Neffs either (Figure S4). Instead, there is a slightly higher number of interface contacts in the development set. Since none of the discriminators can explain the difference in SR, it may be that the development set is biased in its selection in some way. The development set is much smaller than the test set and is, therefore, less likely to encompass the whole protein variability found in nature.

CASP14 and novel proteins without templates

Chains derived from CASP14 heteromeric targets and chains from PDB complexes with no templates have been folded in pairs using the presented AF2 pipeline (default AF2+paired MSAs, ten recycles, m1-10-1 and five differently seeded runs).

For the CASP14 chains, four out of six pairs display a DockQ score larger than 0.23 (SR of 67%). No ranking is necessary in this case, given that all produced docking models for the same chain pair are very similar (the average standard deviation is 0.01 between each set of DockQ scores). An interesting unsuccessful docking is obtained modelling chains from the complex with PDB ID 6TMM (Figure S5), which are known to form a heterotetramer. In this structure, each chain A is in contact with its partner chain B at two different sites. Comparing our docking with both configurations (6TMM_A-6TMM_B and 6TMM_A-6TMM_D) the AF pipeline docked a structure between the two different binding sites. The other unsuccessful docking (6VN1_A-6VN1_H) is recognisable as incorrect for most possible interface size thresholds, given that the interface is composed of just 19 pairs of CBs closer than 8A.

The SR for docking the proteins without templates is 50%. Between the five different initialisations, the average difference in the DockQ score is 0.03, and there is no deviation in SR. This means that ranking did not improve the SR here either. Two acceptable models are displayed in Figures 5 A and B (7EIV_A-7EIV_C and 7MEZ_A-7MEZ_B). 7EIV_A-7EIV_C has a higher DockQ score than 7MEZ_A-7MEZ_B (0.76 and 0.53, respectively), although the overall similarity of 7MEZ_A-7MEZ_B is closer to the native structure. This provides a scenario where the interacting portion of a chain may be correct, although the structure in its entirety is less so.

In one of the incorrect models (7NJ0_A-7NJ0_C, Figure S6), all predictions get the location of both chains correct, but their orientations wrong, resulting in scores close to 0. For 7EL1_A-7EL1_E (Figure 5C), the shorter chain E is entirely wrong, and instead of folding to a defined shape, it is stretched out and inserted within chain A. This is likely to stabilise the structure as it contains a large nucleic acid which chain A wraps around. In the two

remaining incorrect models (7LF7_A-7LF7_M and 7LF7_B-7LF7_M, Figure 5D), the chains A and B only interact with a short loop segment from the M chain. Similarly to CASP derived unsuccessful dockings, these interactions appear not to be meaningful without considering the biological unit.

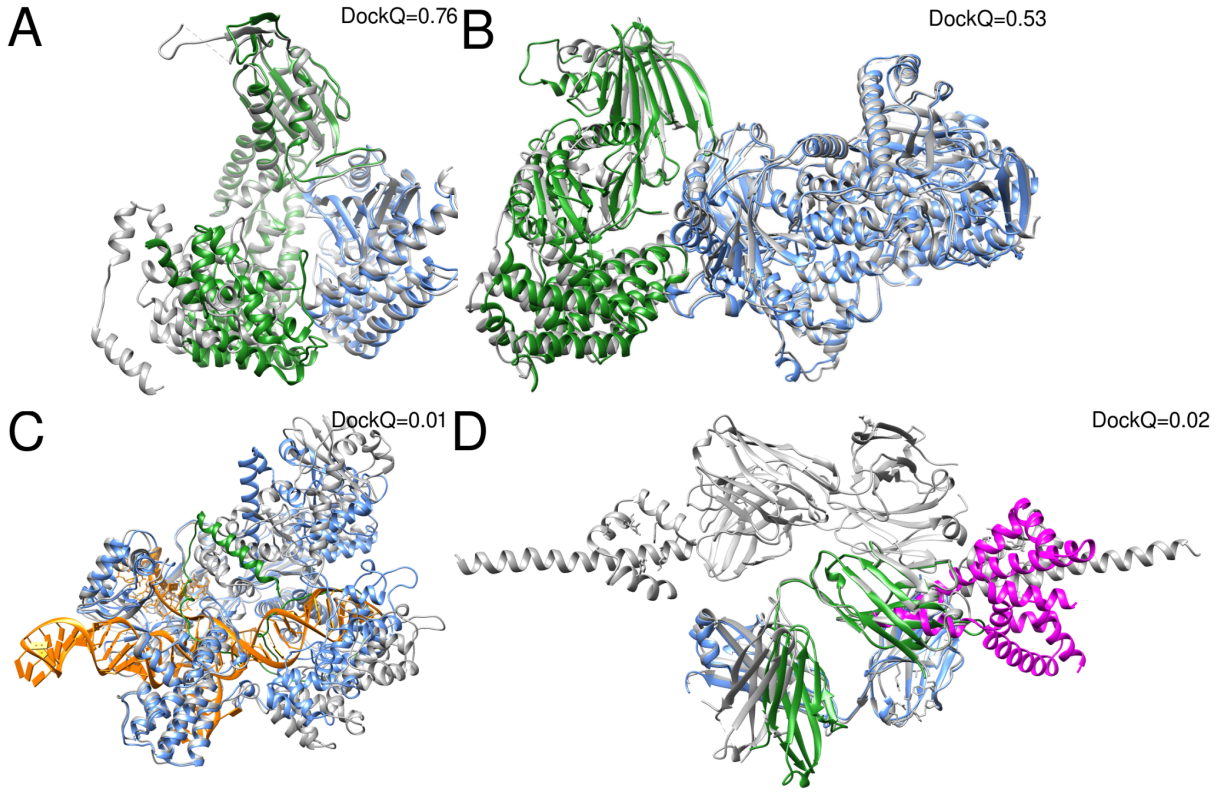

Figure 5. Predicted and native structures from the set of novel proteins without templates. The native structures are represented as grey ribbons **A)** Successful docking of 7EIV chains A (blue) and C (green) (DockQ=0.76). **B)** Successful docking of 7MEZ chains A (blue) and B (green) (DockQ=0.53). **C)** Prediction of structure 7EL1 chains A (blue) and E (green) (DockQ=0.01). The DNA going through chain A is coloured in orange. In this model, chain E occupies the location of the DNA and, therefore, does not fold into a well-defined structure. It does not appear to be possible to fold this complex without the nucleic acid present, but AF2 cannot handle ligands at the moment. **D)** 7LF7 chains A (blue) and M (magenta) (DockQ=0.02) and chains B (green) and M (magenta) (DockQ=0.02). As can be seen in the upper structure, the biological assembly of this protein complex is a hexamer (two mirrored trimers). Chain M only interacts with chains A and B with a short loop segment, suggesting that these interactions are only possible as viewed in the full biological assembly.

Limitations to the current study

Here, we have only studied the possibility to predict the structure of known interacting proteins and leave the task of identifying if a protein interacts with other proteins (both experimental and computational). We only consider the structures in their heterodimeric state, although they may have homodimer configurations or other higher-order states in their potential biological units. Investigating alternative oligomeric states is outside of the scope of this analysis and left for future work.

Conclusions

Here we show that AlphaFold2 (AF2) can predict the structure of many heterodimeric protein complexes, although it is trained to predict the structure of individual protein chains. Even using the default settings, it is clear that AF2 is superior to all other docking methods, including other Fold and Dock methods[16,23], and methods based on shape complementarity[44]. Using optimised multiple sequence alignments with AF2, we can accurately predict the structure of heterodimeric complexes for an unprecedented success rate of 59.0% on a large test set. The success rate is higher in E.Coli (75%) than in *Homo Sapiens* or *S. cerevisiae* (58 %). Further, by examining the interface size, we can separate acceptable and incorrect models with an AUC of 0.84, resulting in that 13% of the models can be called acceptable at a specificity of 99% (or 40% at 90% specificity). Never before has the potential for expanding the known structural understanding of protein interactions been this large, at such a small cost. There are currently 11.9 million pairwise human protein interactions in the String DB[50]. If 13% of these can be predicted at an error rate of 1%, this results in the structure of 1.5 million human heterodimeric protein structures.

Funding

Financial support: Swedish Research Council for Natural Science, grant No. VR-2016-06301 and Swedish E-science Research Center. Computational resources: Swedish National Infrastructure for Computing, grants: SNIC 2021/5-297, SNIC 2021/6-197 and Berzelius-2021-29.

References

1. Liddington, R. C. Structural Basis of Protein–Protein Interactions. *Protein-Protein Interactions* 003–014 doi:10.1385/1-59259-762-9:003.
2. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* **108**, 1225–1244 (2008).
3. Nooren, I. M. A. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. *The EMBO Journal* vol. 22 3486–3492 (2003).
4. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
5. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
6. Marshall, G. R. & Vakser, I. A. Protein-Protein Docking Methods. *Proteomics and Protein-Protein Interactions* 115–146 doi:10.1007/0-387-24532-4_6.
7. Kundrotas, P. J., Zhu, Z., Janin, J. & Vakser, I. A. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 9438–9441 (2012).
8. Porter, K. A., Desta, I., Kozakov, D. & Vajda, S. What method to use for protein–protein docking? *Current Opinion in Structural Biology* vol. 55 1–7 (2019).
9. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443 (2002).
10. Clarke, J. Mechanisms of Folding upon Binding. *The FASEB Journal* vol. 29 (2015).
11. Eginton, C., Naganathan, S. & Beckett, D. Sequence-function relationships in folding upon binding. *Protein Science* vol. 24 200–211 (2015).
12. Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H. J. Principles of flexible protein-protein docking. *Proteins* **73**, 271–289 (2008).
13. Lensink, M. F. *et al.* Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* **87**, 1200–1221 (2019).
14. Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **427**, 3031–3041 (2015).
15. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
16. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
17. Kandathil, S. M., Greener, J. G., Lau, A. M. & Jones, D. T. Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterised proteins. doi:10.1101/2020.11.27.401232.
18. Chowdhury, R. *et al.* Single-sequence protein structure prediction using language models from deep learning. *bioRxiv* 2021.08.02.454840 (2021) doi:10.1101/2021.08.02.454840.
19. Procaccini, A., Lunt, B., Szurmant, H., Hwa, T. & Weigt, M. Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. *PLoS ONE* vol. 6 e19729 (2011).
20. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 67–72 (2009).
21. Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* vol. 34 i802–i810 (2018).
22. Yang, J. *et al.* Improved protein structure prediction using predicted inter-residue orientations. doi:10.1101/846279.
23. Pozzati, G. *et al.* Limits and potential of combined folding and docking using

PconsDock. doi:10.1101/2021.06.04.446442.
24. Lamb, J. & Elofsson, A. pyconsFold: a fast and easy tool for modelling and docking using distance predictions. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab353.
25. Szurmant, H. & Weigt, M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr. Opin. Struct. Biol.* **50**, 26–32 (2018).
26. Kundrotas, P. J. *et al.* Dockground: A comprehensive data resource for modeling of protein complexes. *Protein Sci.* **27**, 172–181 (2018).
27. Green, A. G. *et al.* Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1–12 (2021).
28. Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinformatics* **72**, e108 (2020).
29. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
30. BFD. https://bfd.mmseqs.com/.
31. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
32. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
33. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
34. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* vol. 7 e1002195 (2011).
35. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
36. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* vol. 9 (2018).
37. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).
38. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
39. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
40. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
41. Kosciolek, T. & Jones, D. T. Accurate contact predictions using covariation techniques and machine learning. *Proteins* **84 Suppl 1**, 145–151 (2016).
42. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
43. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, (2004).
44. Vakser, I. A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* **Suppl 1**, 226–230 (1997).
45. Singh, A., Dauzhenka, T., Kundrotas, P. J., Sternberg, M. J. E. & Vakser, I. A. Application of docking methodologies to modeled proteins. *Proteins* **88**, 1180–1188 (2020).
46. Anishchenko, I., Kundrotas, P. J. & Vakser, I. A. Contact Potential for Structure Prediction of Proteins and Protein Complexes from Potts Model. *Biophys. J.* **115**, 809–821 (2018).
47. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **11**, e0161879 (2016).
48. Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins* **78**, 3073–3084 (2010).

49. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
50. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2020).

Supplementary Material

Tables

Table S1. PDB codes and interacting chains for the CASP14 set and set with missing templates. The two interactions involving 6TMM are different possible configurations of the same two chains.

PDB	Chain 1	Chain 2	Top-ranked DockQ score
CASP 14 heterodimers			
7M5F	A	C	0.85
6XOD	A	B	0.94
6VN1	A	H	0.01
6VN1	H	L	0.45
6R17	A	C	0.54
6TMM*	A	B	0.03
6TMM*	A	D	0.04
Heterodimers without templates			
7EIV	A	C	0.76
7EL1	A	E	0.01
7K01	1	6	0.36
7LDG	A	B	0.45
7LF7	A	M	0.02
7LF7	B	M	0.02
7MEZ	A	B	0.53
7NJ0	A	C	0.04

Table S2. False Positive Rates (FPR) and True Positive Rates (TPR) for the test set using the number of interface contacts in the obtained docking models as a threshold. Additionally, the fraction of acceptable and incorrect models and the Positive Predictive Value (PPV) are reported. The fraction of acceptable and incorrect models are obtained by multiplying the TPR and FPR with the success rate (SR=0.59) and the PPV by dividing the TPR with TPR+FPR. For all rates and thresholds, see: https://gitlab.com/ElofssonLab/FoldDock/-/blob/main/analysis/plots/roc_df_marks.csv.

FPR	TPR	FDR: Fraction FP of all	Fraction TP of all models	PPV	Number of interface contacts

		models			
0.000	0.000	0.000	0.000	0.000	801
0.002	0.011	0.001	0.007	0.874	458
0.005	0.027	0.003	0.016	0.847	380
0.005	0.041	0.003	0.024	0.893	348
0.005	0.058	0.003	0.034	0.922	326
0.005	0.078	0.003	0.046	0.940	294
0.008	0.093	0.005	0.055	0.918	270
0.008	0.106	0.005	0.063	0.928	256
0.010	0.126	0.006	0.074	0.927	242
0.012	0.148	0.007	0.087	0.928	226
0.012	0.161	0.007	0.095	0.933	214
0.012	0.178	0.007	0.105	0.939	202
0.020	0.193	0.012	0.114	0.907	192
0.025	0.219	0.015	0.129	0.898	182
0.036	0.261	0.021	0.154	0.878	170
0.048	0.294	0.028	0.174	0.860	160
0.063	0.331	0.037	0.195	0.841	150
0.084	0.381	0.050	0.225	0.819	140
0.096	0.441	0.056	0.260	0.822	128
0.105	0.487	0.062	0.288	0.822	118
0.122	0.546	0.072	0.322	0.817	108
0.150	0.605	0.088	0.357	0.801	98
0.171	0.674	0.101	0.398	0.797	88
0.208	0.743	0.123	0.438	0.782	78
0.265	0.802	0.157	0.473	0.751	68
0.334	0.857	0.197	0.506	0.719	58
0.409	0.908	0.241	0.536	0.690	48
0.512	0.941	0.302	0.555	0.647	38

0.605	0.962	0.357	0.568	0.614	28
0.720	0.977	0.425	0.577	0.576	18
0.825	0.987	0.487	0.583	0.545	8

Table S3. Pearson correlations between DockQ scores of the different modelling runs. The strongest correlations are found using the same MSAs.

Model	**model 1**						**model 1 ptm**					
MSA	hhblits		AF2		AF2+paired		hhblits		AF2		AF2+paired	
Strategy	ens 8	rec 10	ens 8	rec 10	ens 8	rec 10	ens 8	rec 10	ens 8	rec 10	ens 8	rec 10
	1	2	3	4	5	6	7	8	9	10	11	12
1	0,00	0,93	0,55	0,48	0,79	0,72	0,98	0,96	0,50	0,49	0,78	0,71
2	0,93	0,00	0,60	0,48	0,74	0,70	0,92	0,90	0,55	0,54	0,72	0,66
3	0,55	0,60	0,00	0,87	0,70	0,66	0,56	0,56	0,82	0,80	0,65	0,60
4	0,48	0,48	0,87	0,00	0,65	0,63	0,49	0,49	0,79	0,73	0,63	0,61
5	0,79	0,74	0,70	0,65	0,00	0,92	0,80	0,79	0,63	0,65	0,86	0,82
6	0,72	0,70	0,66	0,63	0,92	0,00	0,73	0,72	0,63	0,62	0,84	0,82
7	0,98	0,92	0,56	0,49	0,80	0,73	0,00	0,96	0,50	0,51	0,77	0,72
8	0,96	0,90	0,56	0,49	0,79	0,72	0,96	0,00	0,52	0,50	0,76	0,70
9	0,50	0,55	0,82	0,79	0,63	0,63	0,50	0,52	0,00	0,89	0,71	0,66
10	0,49	0,54	0,80	0,73	0,65	0,62	0,51	0,50	0,89	0,00	0,63	0,62
11	0,78	0,72	0,65	0,63	0,86	0,84	0,77	0,76	0,71	0,63	0,00	0,87
12	0,71	0,66	0,60	0,61	0,82	0,82	0,72	0,70	0,66	0,62	0,87	0,00

Table S4. Success rate and median DockQ score for the different modelling strategies on the test set.

Modelling strategy	**Success rate**	**Median DockQ score**
RoseTTAFold	0.096	0.011
GRAMM	0.214	0.027
AF2 MSA, m1-10-1	0.439	0.119
Paired MSA, m1-10-1	0.481	0.196
AF2+Paired MSAs, m1-10-1	0.559	0.303
AF2+Paired MSAs, top ranked from five initializations, m1-10-1	0.590	0.338

Figures

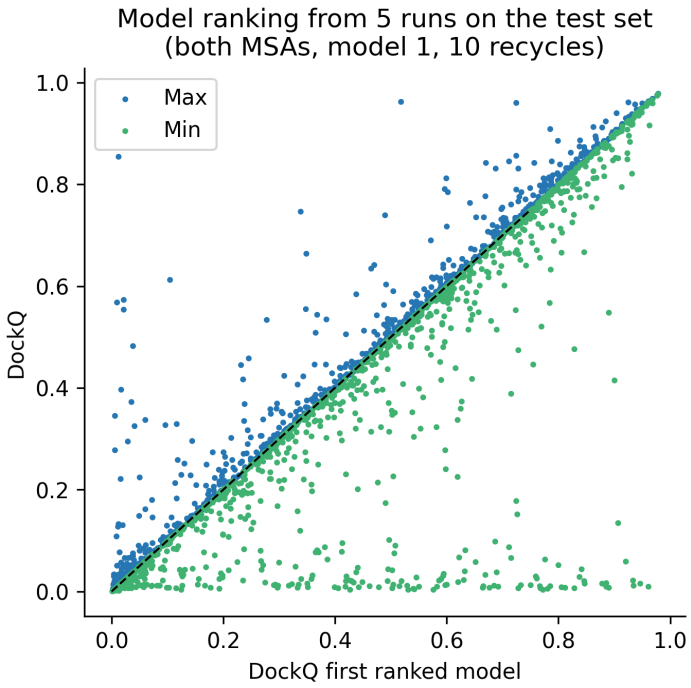

Figure S1. Impact of different initialisations on the modelling outcome in terms of DockQ score on the test dataset. The maximal and minimal scores are plotted against the top-ranked score using the interface's total number of interactions between CBs.

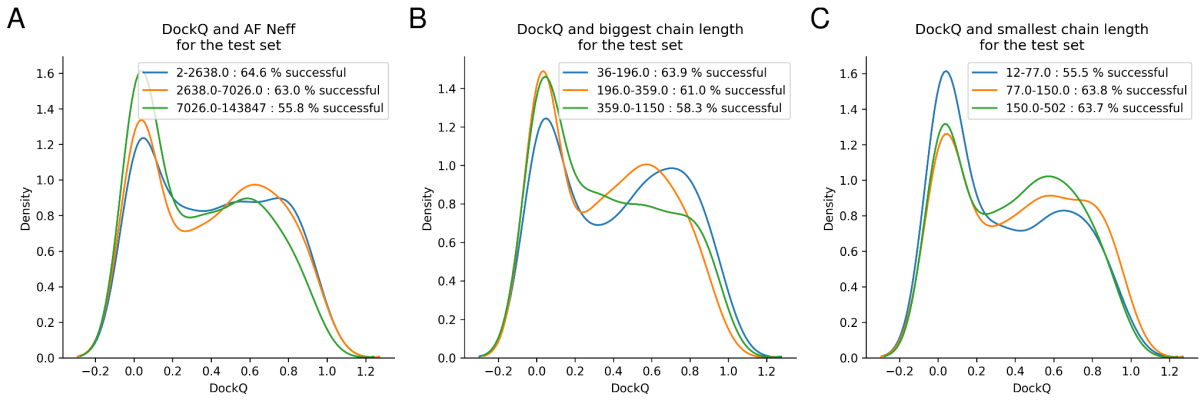

Figure S2. The distribution of DockQ scores for tertiles is derived from **A)** AF Neff **B)** biggest chain length **C)** smallest chain length. The separation between the tertiles is low for all features displaying similar success rates.

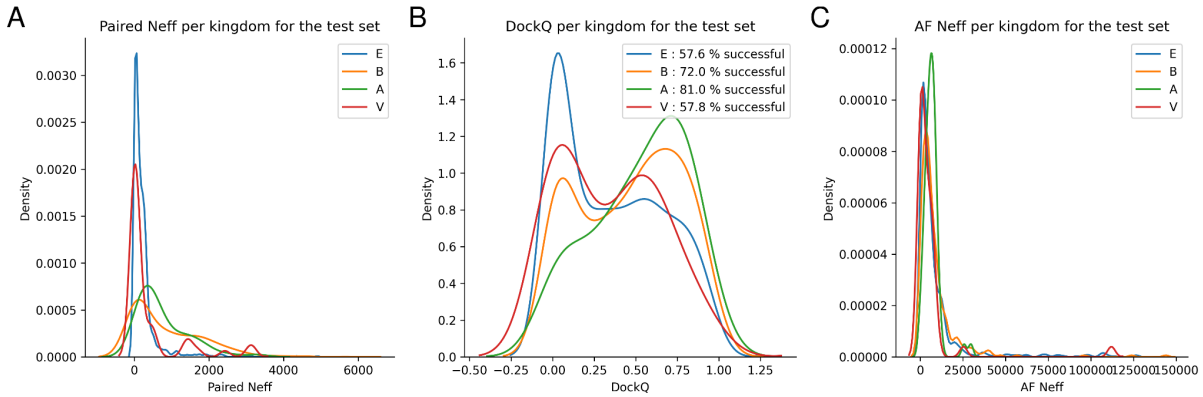

Figure S3. Distribution per kingdom (E=Eukarya, B=Bacteria, A=Archaea, V=Virus) of **A)** Paired Neff, **B)** DockQ scores, and **C)** AF Neff.

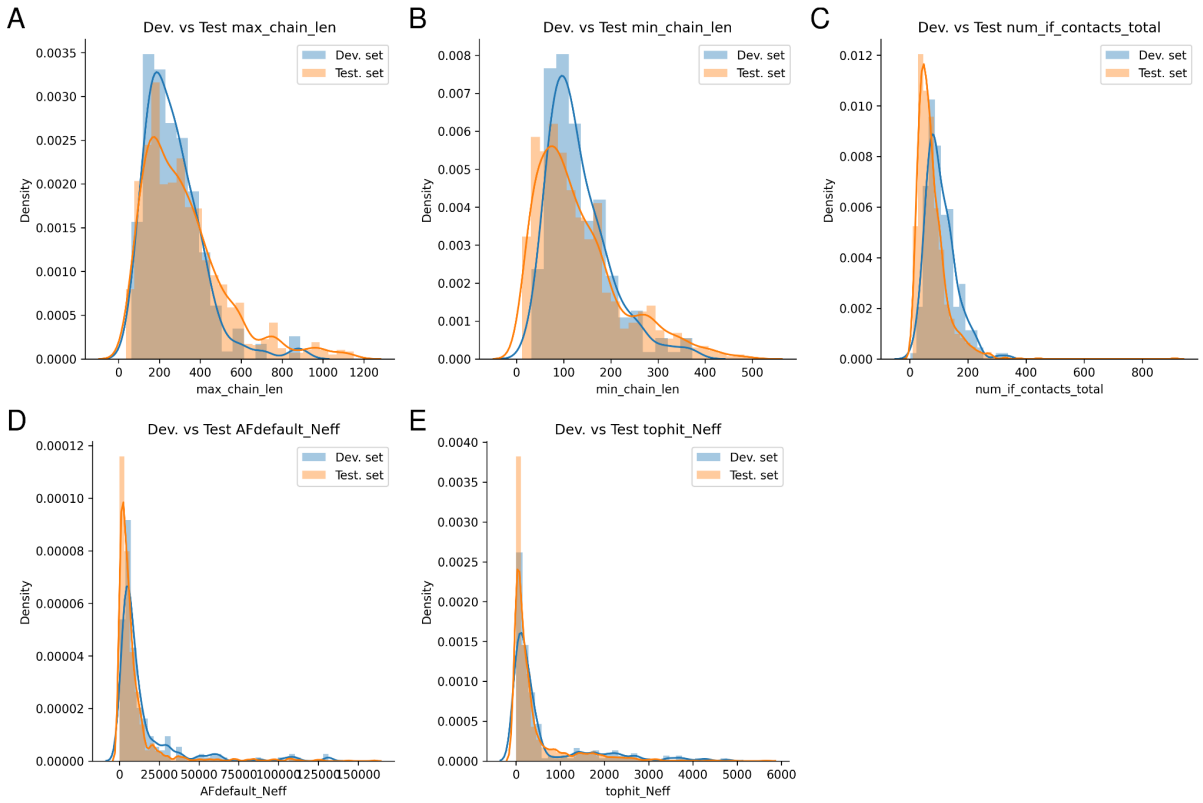

Figure S4. Comparison of development and test sets over several structural features (complex longest and shortest chain lengths, interface contacts) and evolutionary features (MSAs Neff scores).

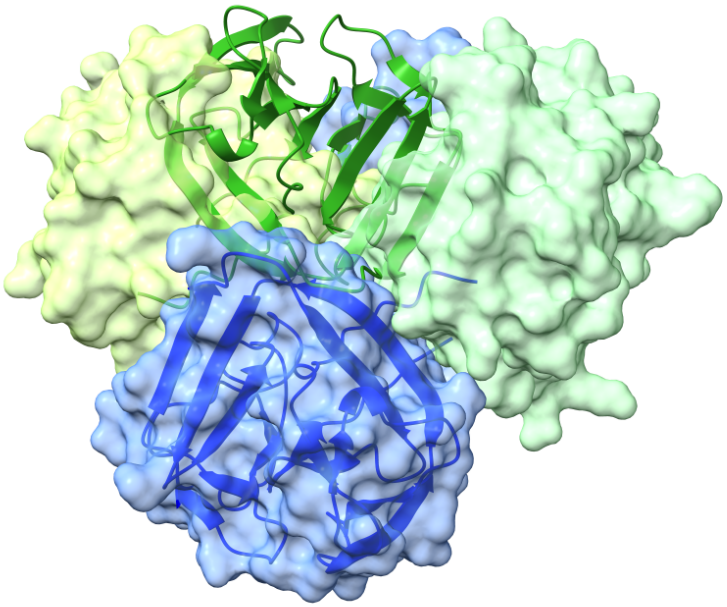

Figure S5. Docking model of CASP14 heterodimer 6TMM (ribbons) superposed to the native heterocomplex (surfaces). The docking model smaller chain (green) is positioned halfway between the two alternative binding sites between blue and green surfaces.

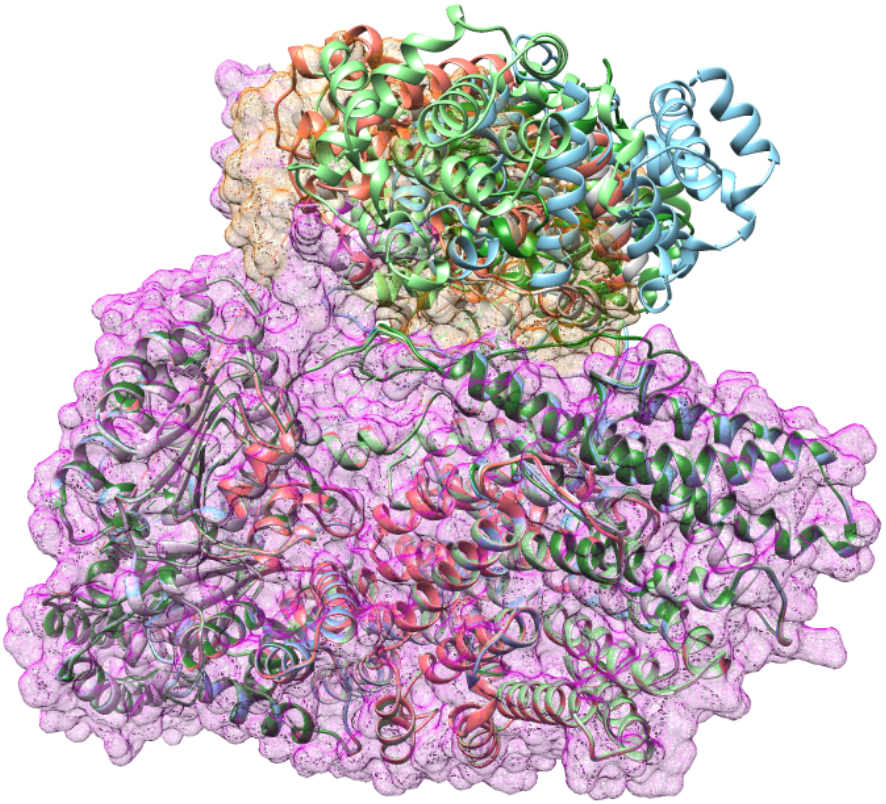

Figure S6. Prediction of 7NJ0_A-7NJ0_C with the native structure represented as a mesh surface (orange and magenta). All predictions (ribbons) get the location of the chains correct, but the interface and orientations are slightly wrong, resulting in DockQ scores f close to 0.