Improved prediction of protein-protein interactions using AlphaFold2

Patrick Bryant
https://orcid.org/0000-0003-3439-1866
Gabriele Pozzati
Stockholms Universitet
Arne Elofsson (arne@bioinfo.se)
Stockholms Universitet https://orcid.org/0000-0002-7115-9751

Article

Keywords: heterodimeric protein complexes, interacting protein chains, AlphaFold2

Posted Date: October 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-951605/v1

Version of Record: A version of this preprint was published at Nature Communications on March 10th, 2022. See the published version at https://doi.org/10.1038/s41467-022-28865-w.

Abstract

Predicting the structure of interacting protein chains is fundamental for understanding the function of proteins. Here, we examine the use of AlphaFold2 (AF2) for predicting the structure of heterodimeric protein complexes. We find that using the default AF2 protocol, 44% of the models in a test set can be predicted accurately. However, by optimising the multiple sequence alignment, we can increase the accuracy to 59%. In comparison, the alternative fold-and-dock method RoseTTAFold is only successful in 10% of the cases on this set, template-based docking 35% and traditional docking methods 22%. We can distinguish acceptable (DockQ>0.23) from incorrect models with an AUC of 0.85 on the test set by analysing the predicted interfaces. The success is higher for bacterial protein pairs, pairs with large interaction areas consisting of helices or sheets, and many homologous sequences. Further, we test the possibility to distinguish interacting from non-interacting proteins and find that by analysing the predicted interfaces, we can separate truly interacting from non-interacting proteins with an AUC of 0.82 in the ROC curve, compared to 0.76 with a recently published method. In addition, when using a more realistic negative set, including mammalian proteins, the identification rate remains (AUC=0.83), resulting in that 27% of interactions can be identified at a 1% FPR. 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 90[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].

In a Fold and Dock approach, two proteins are folded and docked simultaneously. 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]. 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.

In that study, we found that generating the optimal MSA is crucial for obtaining accurate Fold and Dock solutions, but this is not always trivial due to the necessity to identify the exact set of interacting protein 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, while too shallow alignments do not provide sufficient co-evolutionary signals.

We systematically applied the AF2 pipeline on two different datasets to Fold and Dock protein-protein pairs 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. In addition, we analyse the ability to distinguish truly interacting from non-interacting proteins using the created pipeline. AF2 outperforms a recent state-of-the-art method[26] developed using the same data at this task as well.

Material And Methods

Data

Development set

A set of heterodimeric complexes from Dockground benchmark 4[27] is used to develop the pipeline, focusing on the AF2 configuration 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. 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. 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[26] 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%).

107 of the complexes in the test set lack beta carbons (Cβs), 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.

For the mammalian proteins from Negatome, seven out of 1733 single chains were redundant according to Uniprot (C4ZQ83, I0LJR4, I0LL25, K4CRX6, P62988, Q8NI70, Q8T3B2), 34 had no matching species in the MSA pairing, 106 produced out of memory exceptions during prediction using a GPU with 40 Gb RAM, 35 gave a tensor reshape error, and 65 complexes were homodimers, leaving 1715 complexes for this set.

CASP14 set and novel protein complexes

As an additional test set, we used a set of six heterodimers from the CASP14 experiment. 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], see Table S1. We selected this small set to test the performance on data AF2 is guaranteed not to have seen.

Non-interacting proteins

Two datasets of known non-interacting proteins were used, one from the same study as the positive test set[26]. Here, all proteins are from *E.coli.* Two methods were used to identify non-interacting proteins, first a set of proteins with no reported interaction signal in Yeast Two-Hybrid Experiments[30] and secondly complexes whose individual proteins were found in different APMS benchmark complexes[31]. This dataset contains in total 3989 non-interacting pairs.

The second set contains 1964 unique mammalian protein complexes filtered against the IntAct[32] dataset from Negatome[33]. This data deemed “the manual stringent set” contains proteins annotated from the literature with experimental support describing the lack of protein interaction. Some structures in this dataset are homodimers (65) and are therefore excluded, resulting in 1705 structures. Together there are 5694 non-interacting protein complexes.

Methods to generate MSAs

AlphaFold2 default methodology

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[34] (BFD) with HHBlits[35] (from hh-suite v.3.0-beta.3 version 14/07/2017) and both MGnify v.2018_12[36] and Uniref90 v.2020_01[37] with jackhmmer from HMMER3[38]. 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.

Fused HHblits MSAs

In addition to the default AF2 MSA, we generated an additional MSA by simply “fusing” MSAs generated independently from each of the two chains. These MSAs were constructed by running HHblits[35] version 3.1.0 against uniclust30_2018_08[39] with these options:

hhblits -E 0.001 -all -oa3m -n 2

The “fusing” is done by writing gaps for the length of the interacting chain for each sequence in both individual chain MSAs.

Paired MSAs

In addition to the fused 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 organism information was, using the OX identifier, was extracted from the two HHblits MSAs[40]. Next, all hits with more than 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. Pairing the correct sequences should result in MSAs containing inter-chain co-evolutionary information[26].

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[42]. Unaligned FASTA sequences were extracted from the three AF2 default MSAs. Obtained sequences were processed with the CD-HIT software[43] version 4.7 (http://weizhong-lab.ucsd.edu/cd-hit/) using the options:

```
-c 0.62 -G 0 -n 3 -aS 0.9
```

We calculated the Neff scores separately for the paired and the AF2 MSAs.

Prediction of protein-protein complexes

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 the development of 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, therefore, 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. Further, 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[44] 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[45], was used. Here, two protein models are docked using a Fast Fourier Transform (FFT) procedure to generate 340'000 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 an easy case for GRAMM based docking, and performance drops significantly when unbound structures or models are used [46]. 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[47].

Template-based docking

For comparison, a template-based docking protocol[7] referred to as “TMdock" is also adopted. The adopted template library includes 11756 protein complexes obtained from the Dockground database[27] (release 28-10-2020). Target complexes are structurally aligned with the supplied template structures (depleted of the target structure PDB ID). TM-scores resulting from the alignment of target proteins to each template are averaged and used to score obtained docking models. Alternatively, we refer to “TMdock Interfaces” when targets are structurally aligned only to the template interfaces, defined as every residue with a Cβ atom closer than 12 Å from any Cβ atom in the other chain.

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[48] was then run to compare the predicted and actual interfaces. This program compares interfaces using a combination of three different CAPRI[49] 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 and scoring models

To analyse the ability of AF2 to distinguish correct models as well as interacting from non-interacting proteins, 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 (Cβs from different chains within 8 Å from each other), the total number of interactions between Cβs 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 scores below the threshold. From the built confusion matrix, we derive the True Positive Rate (TPR), 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 a Receiver Operating Characteristic (ROC) curve. We compute the Area Under Curve (AUC) for ROC curves obtained for each metric to compare different metrics. The AUC is defined as:

$$AUC = \int_{x=0}^{1} TPR\left(\frac{1}{FPR(x)}\right)dx$$

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

$$PPV = \frac{TP}{TP + FP}$$
$$FDR = 1 - PPV$$
$$SR = Fraction\ of\ predicted\ models\ with\ DockQ \geq 0.23$$

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[50], 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).

Computational cost

To compare the computation required for each MSA, we compared the time it took to generate MSAs for three protein pairs (PDB: 4G4S_P-O, 5XJL_A-2 and 5XJL_2-M), using either the fused or AF2 protocol. The tests were performed on a computer using 16 CPU cores from an Intel Xeon E5-2690v4.

Fusing the MSAs took 3 seconds on average per tested complex. It took 7884 seconds for generating the AF2 MSAs, the single-chain searches took 338 seconds on average and the pairing 2 seconds. The pairing and fusing are thereby negligible compared to searching, resulting in a speedup of 24 times for the hhblits searches. In comparison, folding using the m1-10-1 strategy took 191 seconds on average for these pairs.

Results And Discussion

Identifying the best AlphaFold2 model

The fraction of acceptable models (DockQ>0.23), the success rate (SR) is used to measure performance for each AF2 model using the different MSAs. The best performance is 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. The average performance of the AF2 and the paired MSAs is similar, but for individual protein pairs, frequently one of the two MSAs is superior to the other, as seen from that the Pearson correlation coefficient for the DockQ scores between AF2 vs paired MSAs is 0.48 (Table S2). Therefore, combining AF2 and paired MSAs improves the results.

Next, we compared the default AF2 model (model_1) with the fine-tuned versions of (model_1_ptm). Surprisingly, the original AF2 model_1 outperforms AF2 model_1_ptm in most cases (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.

Success rate (DockQ≥0.23) of modelled 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

Test set analysis

Test set performance

The best model and configuration for AF2 (m1-10-1) was used for further studies on the test set. The best outcome using this modelling strategy results in an SR of 55.9% (828 out of 1481 correctly modelled complexes) for the AF2+paired MSAs compared with 43.9% using the AF2 MSAs alone (Figure 2, Table S3). The results using the fused+paired MSAs are almost identical (SR=56.0%,median=0.302). Further, running five initialisations with random seeds and ranking the models using the average plDDT in the interface increases the SR to 57.8% and 58.7% for the AF2+paired and fused+paired MSAs, respectively (model variation and ranking, Figure 3). Using the combination of AF2 and paired MSAs increases performance, suggests that AF2 gains both from larger and paired MSAs, although it often can manage with less information.

What is most striking is that AF2 outperforms all other methods by a large margin.

RF is better than AF2 only for 14 pairs in the test set, while GRAMM and template-based docking (TMdock interface) outperform AF2 for 188 and 225 pairs, respectively. 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.

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 (Cβ atoms from different chains within 8 Å from each other) in the interface, (ii) the total number of interactions between Cβ atoms 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 3A). Three criteria result in very similar areas under the curve (AUC) measures. The total number of interactions between Cβs and the number of residues in the interface can separate the correct/incorrect models with an AUC of 0.86, while the average interface plDDT results in an AUC of 0.85. However, pLDDT results in higher TPRs at lower FPRs; therefore, it is better for model ranking.

Interestingly, the average plDDT of the entire complex only results in an AUC of 0.68, suggesting that both single chains in a complex are often predicted very well, while their relative orientation 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 (55.2% ± 0.0%) was similar for all five runs. However, the average deviation for individual models is DockQ=0.08 when comparing the best and worst models for a target (Figure 3B), 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 61.0%. Although this is unachievable, ranking the models using the total number of interactions in the interface results in an SR of 57.8%. The AUC using the average plDDT in the interface for the ranked test set is 0.82, which means that 16% of all models are acceptable at an error rate of 1% and 37% at an error rate of 10% (Table S4).

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 what distinguishes the successful and unsuccessful cases by analysing different subsets of the test set. First, we divided the proteins by taxa, interface characteristics, and finally by examining the alignments.

The Success Rates (SRs) for each kingdom is; Eukarya 57%, Bacteria 72%, Archaea 80%, and Virus 55% (Figure S1B). Further, the SRs for *Homo Sapiens* and *S.cerevisiae* are similar (58% vs 59%). The better performance in prokaryotes is consistent with previous observations regarding the availability of evolutionary information in prokaryotes compared to Eukarya[26] (Figure S2A).

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 interfaces containing mainly sheets (59%). The loop interface SR of 53% is substantially lower than the others, suggesting that interfaces with more flexible structures are harder to predict. We divided the dataset by the size of the interface, and it is clear that pairs with larger interfaces are easier to predict, as the SR increases from 47 to 74% between the smallest and biggest tertiles (Figure 4B).

Next, we examined how the size of the MSA (both paired and AF2) influences the results. It is clear that the fraction of correctly modelled sequences increases with larger MSAs (Figure 4C), and the size of the paired MSA (Figure 4C) has a more considerable influence on the outcome than the size of the AF2 MSA (Figure S1A).

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 S3), 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. Both docking configurations (6TMM_A-B and 6TMM_A-D) put the chain in between the two binding sites. The other unsuccessful docking (6VN1_A-H) has an interface of just 19 residue pairs.

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, i.e. ranking did not improve the SR. Two acceptable models are displayed in Figures 5A (7EIV_A-C) and B (7MEZ_A-B). More interesting, in one of the incorrect models (7NJ0_A-C, Figure S4), the predictions get the location of both chains correct, but their orientations wrong, resulting in DockQ scores close to 0. For 7EL1_A-E (Figure 5C), the shorter chain E is not folded correctly, and instead of folding to a defined shape, it is stretched out and inserted within chain A. It occupies the shape of the DNA in the native structure. In the two remaining incorrect models (7LF7_A-M and 7LF7_B-M), Figure 5D, the chains only interact with a short loop of the M chain, making the docking very difficult and possibly biologically meaningless.

Identifying interacting proteins

Using the best separator from the model ranking the interface plDDT, it is possible to distinguish the 3989 non-interacting proteins from *E.coli* and the truly interacting proteins from the test set with an AUC of 0.82. Another recently published method obtains AUC 0.76 on this set[26]. However, these results are probably overstated since the negative set only contains bacterial proteins, while the positive set is mainly eukaryotic.

To obtain a more realistic estimate, we also include a set of non-interacting proteins from mammalian organisms combined with the non-interacting proteins from *E.coli*. On this set, we obtain an AUC of 0.82 for the average interface plDDT and slightly higher (0.84 and 0.85) for the number of interface contacts and residues, respectively (Figure 6A). Here, the average interface plDDT provides a better separation at low FPRs, enabling a TPR of 27% at FPR of 1% compared to 18 and 13% for the number of interface contacts and residues, respectively. At FPR 5%, the reverse is true, with the number of interface contacts and residues reporting TPRs of 49 and 42%, respectively, compared to 43% for the average interface plDDT. The distribution of the three top separators can be seen in Figure 6B.

Limitations

Here, we only consider the structures of protein complexes in their heterodimeric state, although each protein chain in these complexes may have homodimer configurations or other higher-order states. It is also possible that the complex itself exists as part of larger biological units, in potentially more complex conformations. Investigating alternative oligomeric states and larger biological assemblies is outside of the scope of this analysis and left for future work.

The study of AF2s ability to separate interacting and non-interacting proteins here contains more extensive data than recent studies[26]. However, to test this separation thoroughly, the data studied here needs to be extended to compare interactions within individual organisms. We leave this extensive analysis to further studies.

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], methods based on shape complementarity[45] and template-based docking. 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 average interface plDDT, we can separate acceptable and incorrect models with an AUC of 0.85, resulting in that 14% of the models can be called acceptable at a specificity of 99% (or 38% at 90% specificity). Interestingly, no additional constraints are implemented in AF2 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 larger interfaces are more likely to be correctly predicted.

We find that the MSA generation process can be sped up substantially at no performance loss by simply fusing MSAs from two HHblits runs on Uniclust30 instead of using the MSAs from AF2. Fast MSA generation circumvents the main computational bottleneck in the pipeline. Analysing the interfaces of predicted complexes makes it possible to separate truly interacting from non-interacting proteins with an AUC of 0.82, making it possible to identify 27% of interacting proteins at an error rate of 1%. Features of the predicted interfaces discriminate between model quality and binary interactions. Therefore the same pipeline can identify if two proteins interact and the accuracy of their structure. 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[51]. If 14% of these can be predicted at an error rate of 1%, this results in the structure of 1.5 million human heterodimeric protein structures. The computational cost to run all of this would take approximately three months on an Nvidia A100 system.

Declarations

Authorship

PB and GP performed the studies; all authors contributed to the analysis. PB wrote the first draft of the manuscript; all authors contributed to the final version. AE obtained funding.

The authors claim no conflicts of interest.

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.

Acknowledgements

We thank Petras Kundrotas for supplying the new heterodimeric proteins without templates in the PDB.

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. 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).

27. Kundrotas, P. J. et al. Dockground: A comprehensive data resource for modeling of protein complexes. Protein Sci. 27, 172–181 (2018).

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. Rajagopala, S. V. et al. The binary protein-protein interaction landscape of Escherichia coli. Nat. Biotechnol. 32, 285–290 (2014).

31. Kuhlbrandt, W. The Resolution Revolution. Science vol. 343 1443–1444 (2014).

32. Orchard, S. et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42, D358–63 (2014).

33. Blohm, P. et al. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Research vol. 42 D396–D400 (2014).

34. BFD. https://bfd.mmseqs.com/.

35. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473 (2019).

36. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 48, D570–D578 (2020).

37. 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).

38. Eddy, S. R. Accelerated Profile HMM Searches. PLoS Computational Biology vol. 7 e1002195 (2011).

39. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170–D176 (2017).

40. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489 (2021).

41. 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).

42. Kosciolek, T. & Jones, D. T. Accurate contact predictions using covariation techniques and machine learning. Proteins 84 Suppl 1, 145–151 (2016).

43. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283 (2001).

44. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. Proteins 57, (2004).

45. Vakser, I. A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins Suppl 1, 226–230 (1997).

46. 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).

47. 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).

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

49. Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. Proteins 78, 3073–3084 (2010).

50. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983).

51. 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).

Figures

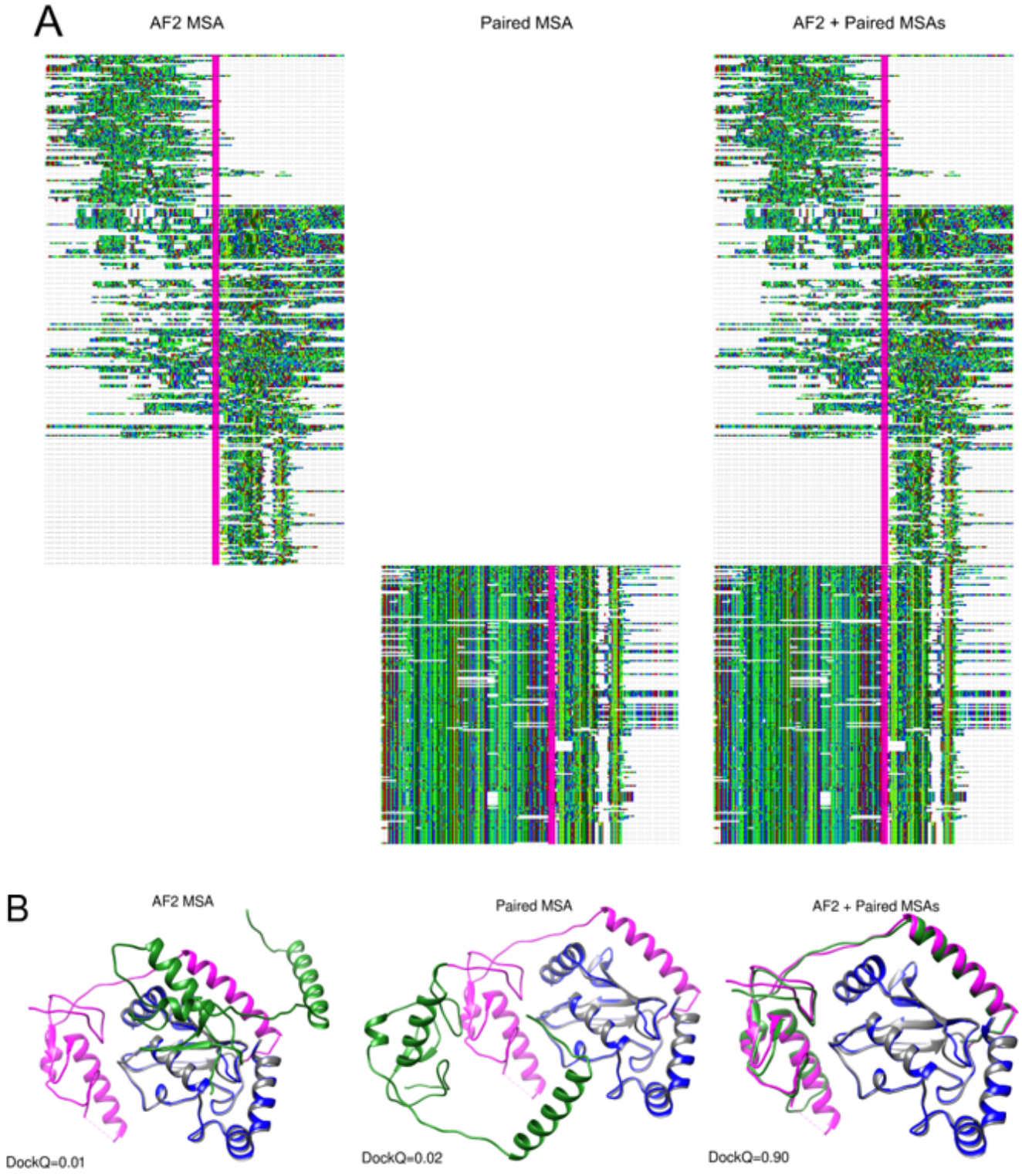

Figure 1

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 2.11.1.441 B) Docking visualisations for PDB ID 5D1M with the model/native

chains A in blue/grey and B in green/magenta using the three different MSAs in A. The DockQ scores are 0.01, 0.02 and 0.90 for AF2, paired, and AF2+paired MSAs, respectively.

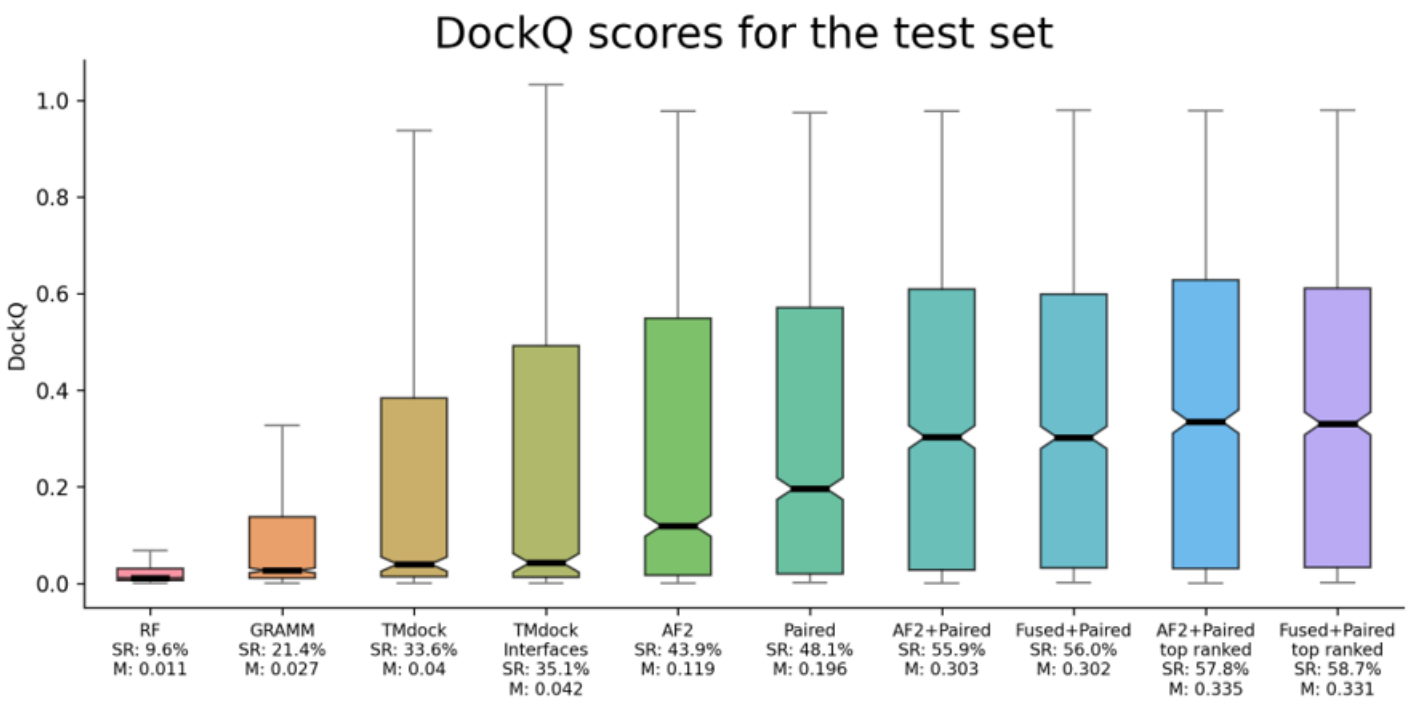

Figure 2

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) and medians (M) are reported below the name of each method. All AF2 models have been run with the same neural network configuration (m1-10-1). Outlier points are not displayed here.

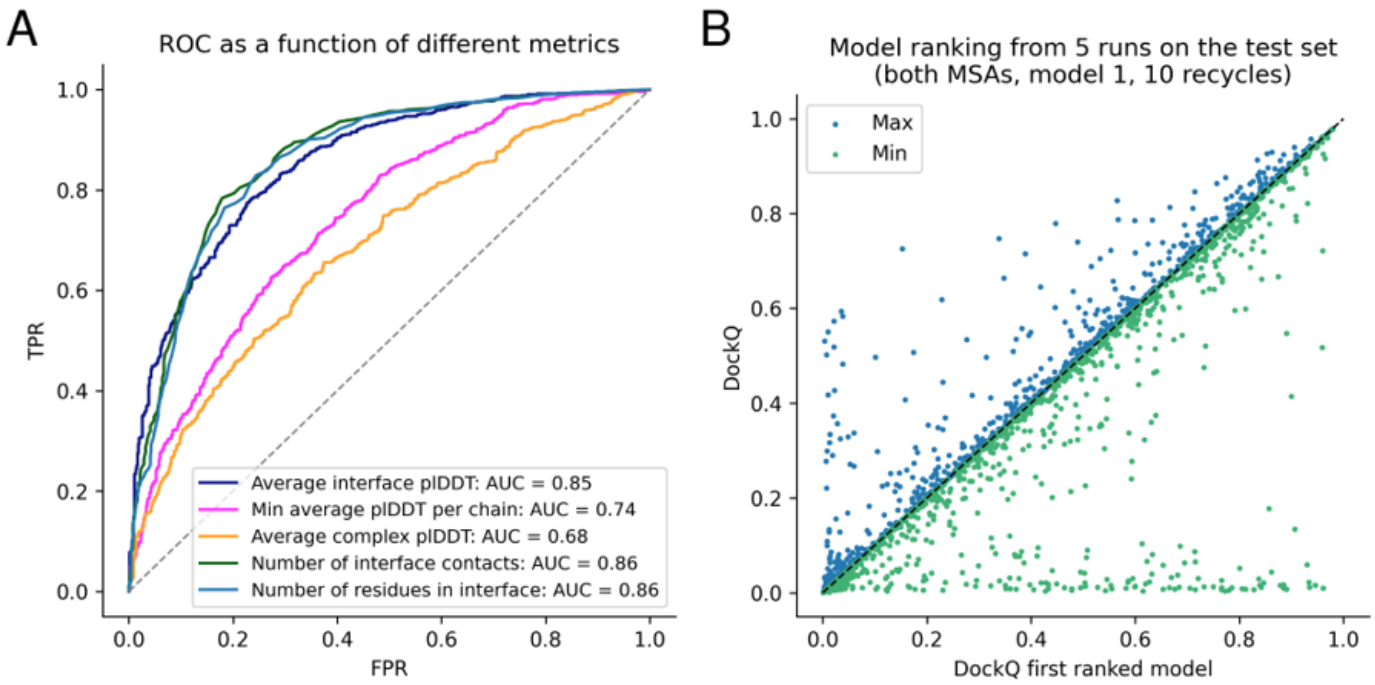

Figure 3

A) ROC curve as a function of different metrics for the development dataset (first run). Cβs 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 average plDDT in the interface for the AF2+paired MSAs, m1-10-1.

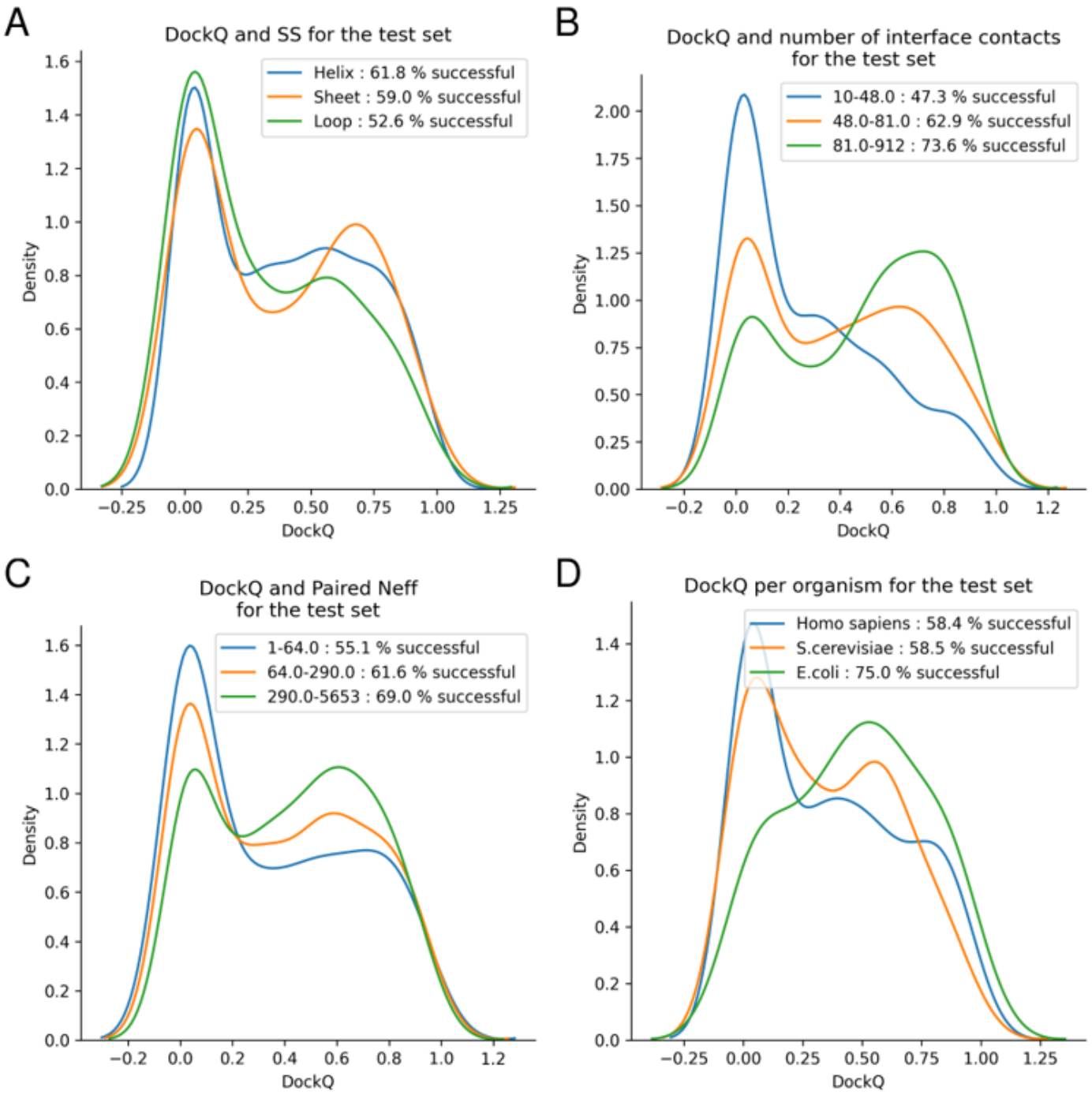

Figure 4

A) Distribution of DockQ scores for three sets of interfaces with the majority of Helix, Sheet and Coil secondary structures. B) Distribution of DockQ scores for tertiles derived from the distribution of contact counts in docking model interfaces. C) Distribution of DockQ scores for tertiles derived from the distribution of Paired MSAs Neff scores. D) Distribution of DockQ scores for the top three organisms Homo Sapiens, S. cerevisiae and E. coli.

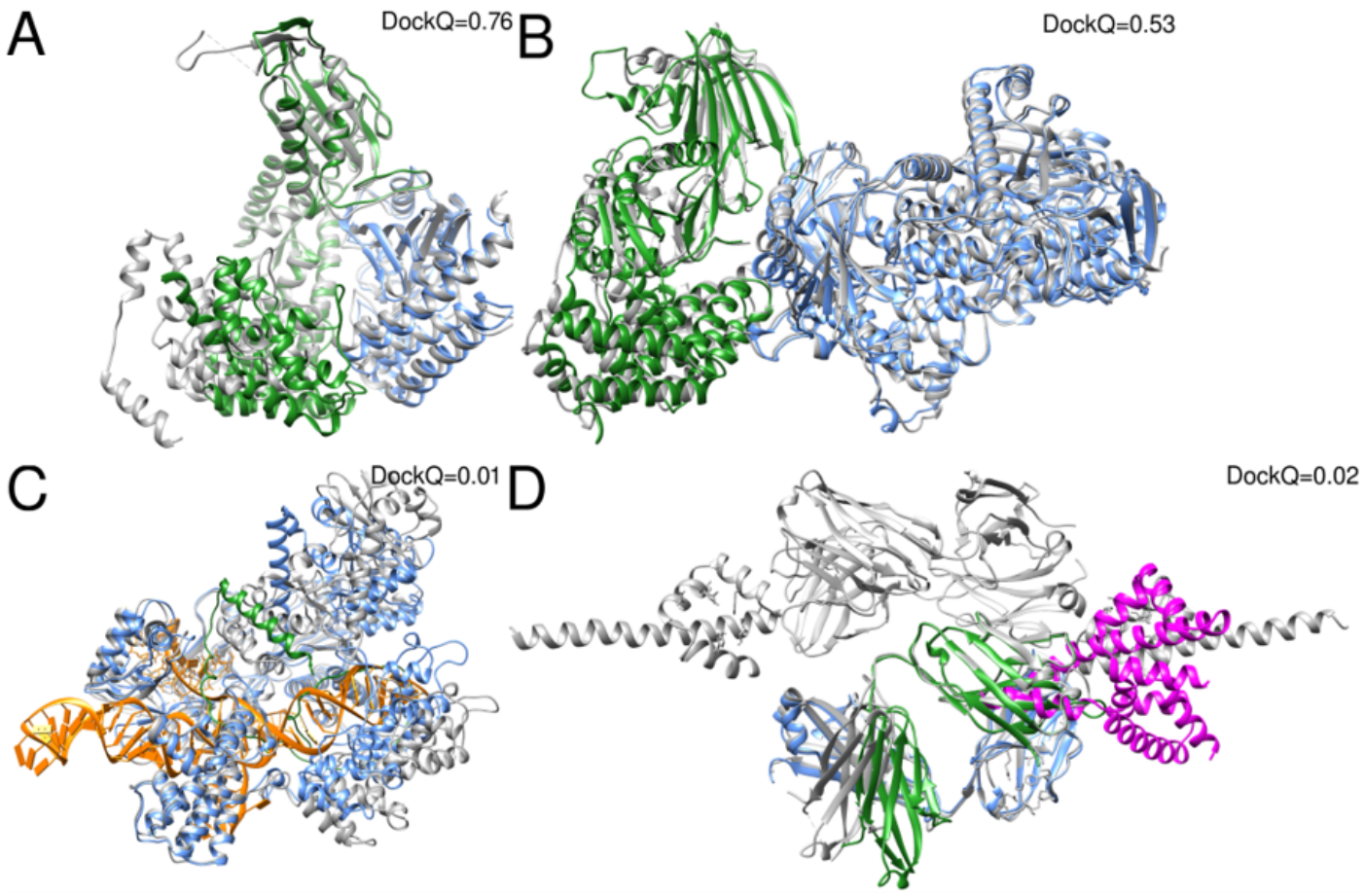

Figure 5

Predicted and native structures from the set of novel proteins without templates. The native structures are represented as grey ribbons A) Docking of 7EIV chains A (blue) and C (green) (DockQ=0.76). B) 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. D) Docking of 7LF7 chains A (blue) and M (magenta) (DockQ=0.02) and chains B (green) and M (magenta) (DockQ=0.02).

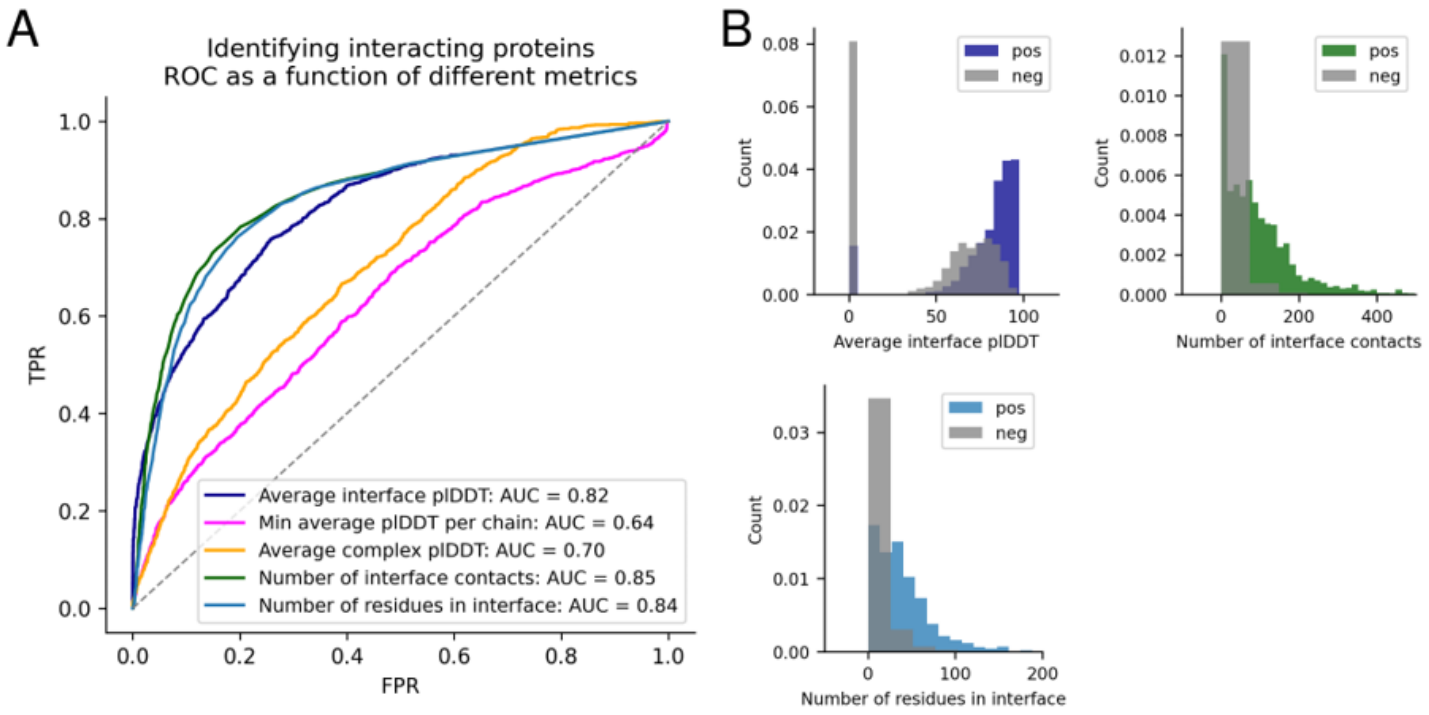

Figure 6

A) The ROC curve as a function of different metrics for discriminating between interacting and non-interacting proteins. B) Distribution of the top three discriminating features for interacting (coloured) and non-interacting proteins (grey).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx