Improved prediction of protein-protein interactions using AlphaFold2

Patrick Bryant, Gabriele Pozzati & Arne Elofsson

Predicting the structure of interacting protein chains is a fundamental step towards understanding protein function. Unfortunately, no computational method can produce accurate structures of protein complexes. AlphaFold2, has shown unprecedented levels of accuracy in modelling single chain protein structures. Here, we apply AlphaFold2 for the prediction of heterodimeric protein complexes. We find that the AlphaFold2 protocol together with optimised multiple sequence alignments, generate models with acceptable quality (DockQ ≥ 0.23) for 63% of the dimers. From the predicted interfaces we create a simple function to predict the DockQ score which distinguishes acceptable from incorrect models as well as interacting from non-interacting proteins with state-of-art accuracy. We find that, using the predicted DockQ scores, we can identify 51% of all interacting pairs at 1% FPR.
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\(^2\). 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\(^3\). Protein docking methodologies refer to how proteins interact and can be divided into two categories considering proteins as rigid bodies; those based on an exhaustive search of the docking space\(^4\) and those based on alignments (both sequence and structure) to structural templates\(^5\). Exhaustive approaches rely on generating all possible configurations between protein structures or models of the monomers\(^6,7\) and selecting the correct docking through a scoring function, while template-based docking only needs suitable templates to identify a few likely candidates. However, flexibility has often to be considered in protein docking to account for interaction-induced structural rearrangements\(^8,9\). Therefore, flexibility limits the accuracy achievable by rigid-body docking\(^10\), and flexible docking is traditionally too slow for large-scale applications. A possible compromise is represented by semi-flexible docking approaches\(^11\) that are more computationally feasible and can consider flexibility to some degree during docking.

Regardless of different strategies, docking remains a challenging problem. In the CASP13–CAPRI experiments, human group predictors achieved up to 50% success rate (SR) for top-ranked docking solutions\(^12,13\). Alternatively, a recent benchmark study\(^8\) reports SRs 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 multiple sequence alignment (MSA) processing in DMPFold\(^17\) 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 (DCA) was used to predict the interaction of bacterial two-component signalling proteins\(^19\). Later, these methods were improved using machine learning\(^20\).

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\(^21\)). 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\(^22,23\). 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\(^24\). 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.

In this work, 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 AF2 pipeline in combination with different input MSAs, in order to study the relationship between the output model quality and these inputs. We also find that, by scoring multiple models of the same protein–protein interaction with a predicted DockQ score (pDockQ), we can distinguish with high confidence acceptable (DockQ ≥ 0.23) from incorrect models. The modelling success is higher for bacterial protein pairs, with large interaction areas consisting of helices or sheets, and many homologous sequences. We also test the possibility to distinguish interacting from non-interacting proteins and find that, using pDockQ, we can separate truly interacting from non-interacting proteins with consistent accuracy. We find that the results in terms of successful docking using AF2 are superior to other docking methods. AF2 clearly outperforms a recent state-of-the-art method\(^25\) and our protocol performs quite close to (63% vs 72%) the recently developed AF-multimer\(^26\), which was developed using the same data as the test set here, making a direct comparison difficult.

### Results and discussion

**Identifying the best AlphaFold2 model.** The SR, i.e., the percentage of acceptable models (DockQ > 0.23), is used to measure AF2 performance over the development set (216 proteins) using the different MSAs. The best performance is 33.3% for the AF2 MSAs and 39.4% 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 0.54 (Supplementary Table 1). 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 appear to impact the outcome the most.

| Table 1 Success rate of different modelling setups. |
|---------------------------|
| **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 |
| Setup short name | m1-10-1 | m1-3-8 | mp-10-1 | mp-3-8 |
| Parse MSA | 28.7 | 28.2 | 28.7 | 27.8 |
| AF2 MSA | 31.5 | 33.3 | 26.4 | 23.6 |
| AF2+Paired MSA | 39.4 | 38.4 | 32.4 | 31.0 |

Results of AF2 run on the development set (n = 216) using different MSAs and neural network configurations. Row labels in bold indicate AF2 setup features. Every column in the table refers to an overall setup and every corresponding value refers to a run of the described setup with a different input MSA. Values represent the percentage of acceptable models (DockQ ≥ 0.23) overall the development set. The highest success rates for each MSA type are highlighted in bold.
Fig. 1 DockQ scores for the test set (n = 1481 for all but RF, n = 1455). Distribution of DockQ scores as boxplots for different modelling strategies on the test set. Boxes encompass data quartiles, horizontal lines mark the medians and upper and lower whiskers indicate respectively maximum and minimum values for each distribution. All AF2 models have been run with the same neural network configuration (m1-10-1). Outlier points are not displayed here. AF2, refers to running AF2 using the default AF2 MSAs, “Paired” refers to using MSAs paired using information about species and “Block” refers to using block diagonalization MSAs.

### 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 57.8% (856 out of 1481 correctly modelled complexes) for the AF2 + paired MSAs compared with 45.0% using the AF2 MSAs alone (Fig. 1, Table 2). The results using the block diagonalization + paired MSAs are almost identical (SR = 58.4%, median = 0.363). Further, running five initialisations with random seeds and ranking the models using the predicted DockQ score (pDockQ, Fig. 2c), increases the SR to 61.7% and 62.7% for the AF2 + paired and block diagonalization + paired MSAs, respectively (model variation and ranking, Fig. 2).

Using the combination of AF2 and paired MSAs increases performance, suggesting 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 tested docking methods by a large margin (Fig. 1, Table 2). 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 best performing method in the CASP14-CAPRI experiment, MDockPP, achieves a SR of only 24.2%. The reason GRAMM, TMdock and MDockPP reach this level of 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.

The recently developed AF-multimer has the best performance (SR = 72.2%, median = 0.560, Table 2). This method was trained using the same data as the test set, which makes a direct comparison difficult. Regardless, we do believe it is likely that using AF-multimer, the performance would increase over the results of our pipeline, but it is possible the difference is less than the observed 9%.

### Distinguishing acceptable from incorrect models.
It is not only essential to obtain improved predictions, but also to be able to discriminate between acceptable and non-acceptable ones. We measure the separation between correct (DockQ ≥ 0.23) and incorrect models provided by several metrics using a receiver operating characteristic (ROC) curve. Different criteria were examined over the test set, 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 (Fig. 2a). 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 0.92 and 0.91 respectively, while the average interface pLDDT results in an AUC of 0.88. However, pLDDT results in higher TPRs at lower FPRs; therefore, we multiply the pLDDT with the logarithm of the interface contacts resulting in an AUC of 0.95.

Interestingly, the average pLDDT of the entire complex only results in an AUC of 0.66, suggesting that both single chains in a complex are often predicted very well, while their relative orientation may still be incorrect.

Figure 2b shows that increasing both the number of interface contacts and the average interface pLDDT results in higher DockQ scores for the test set. Using the combination of pLDDT with the logarithm of the interface contacts, we, therefore, fit a simple sigmoidal function to the DockQ scores (Fig. 2c), see methods. This enables the prediction of the DockQ scores.
(pDockQ) in a continuous manner with an overall average error of 0.11 on the test set. The AUC using pDockQ as a separator is identical to the combination of plDDT with the logarithm of the interface contacts, 0.95 (Fig. 2a).

**Model variation and ranking for the test set.** Five models are generated using the best strategy (m1-10-1 with AF2 + paired MSAs) with different initialisation (random seeds). The average SR (57.2% ± 0.0%) is 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 (Fig. 2d), 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 62.9%. Although this is unachievable, ranking the models using the pDockQ score results in a SR of 61.7%. The AUC using the same metric for the ranked test set is 0.93, which means that 31% of all models are acceptable at an error rate of 1% and 54% at an error rate of 10% (Supplementary Table 2).

**Bacterial complexes are predicted more accurately.** In the test set, about 60% of the complexes can be modelled correctly. We try to identify what distinguishes the successful and unsuccessful cases by analysing different subsets of the test set. First, we divide the proteins by taxa, next by interface characteristics and finally by examining the alignments. The SRs for each kingdom are; Eukarya 61%, Bacteria 73.7%, Archaea 84.5%, and Virus 60% (Supplementary Fig. 1b). Further, the SRs for *Saccharomyces cerevisiae* is better than for *Homo sapiens* (66% vs 58%, Fig. 3d). The higher performance in prokaryotes is consistent with previous observations regarding the availability of evolutionary information in prokaryotes compared to Eukarya27 (Supplementary Fig. 1). The higher
performance in *S. cerevisiae* compared to *H. sapiens* suggests a similar relationship between higher and lower order organisms within the same kingdom.

Next, we examine the interfaces. Different secondary structural content of the native interfaces is investigated (Fig. 3a). The highest SR is obtained mainly for 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 divide the dataset by interface size, and find that pairs with larger interfaces are easier to predict, as the SR increases from 47 to 74% between the smallest and biggest tertiles (Fig. 3b).

We continue to examine features of the MSAs. First, the impact of the number of non-redundant sequences (Neff) in both paired and AF2 MSAs was analysed. It is clear that the fraction of correctly modelled sequences increases with larger Neff scores (Fig. 3c). Also, paired MSA Neff (Fig. 3c) has a stronger influence on the outcome than the Neff of the AF2 MSAs (Supplementary Fig. 2a). Secondly, the MSA interface signal in the paired MSAs, measured by the fraction of correct interface contacts using DCA, was analysed. MSAs with stronger interface signals show higher SRs, even if the paired MSAs are used in combination with the AF2 MSAs (Supplementary Fig. 3). This suggests that MSA co-evolutionary signal and, thereby, correct identification of orthologous protein sequences, has a strong impact on the outcome.

**CASP14 and novel proteins without templates.** Chains derived from CASP14 heteromeric targets and chains from PDB complexes with no templates are 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 (Supplementary Fig. 4), 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 Fig. 5a (7EIV_A-C]) and B (7MEZ_A-B). More interesting, in one of the incorrect models (7NJ0_A-C], Supplementary Fig. 5), the predictions get the location of both chains correct, but their orientations wrong, resulting in DockQ scores close to 0. For 7EL1_A-E (Fig. 4c), 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), Fig. 4d, 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 pDockQ, it is possible to distinguish the 3989 non-interacting proteins from *Escherichia coli* and the 1481 truly interacting proteins from the test set with an AUC of 0.87. Another recently published method obtains AUC 0.76 on this data set27. However, these results are probably overstated since the negative set only contains bacterial proteins, while the positive set 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 studies27. 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.

There is a big difference between the performance of AF2 on the development and test sets, reporting 39.4% SR vs 57.8% for the AF2 + Paired MSAs. This discrepancy suggests that the performance is highly dependent on the specific interacting partners being predicted. It is not clear what causes this difference as the composition in terms of kingdom, found to be very important (Supplementary Fig. 1b) is similar (54% vs 60% Eukaryotic proteins), the MSAs have similar Neff scores (2699 vs. 2764 on average), the proteins are of similar sizes (222 vs. 203 AAs on average), and the number of residues in the interface is similar (139 vs 120 on average). This leads us to believe that there may be some unknown selection bias in how the sets were chosen. It can be noted that the development is much smaller than the

**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 studies27. 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.
test set though (216 vs 1481 proteins), which is why performance should be assessed on as large non-redundant datasets as possible.

Findings and future prospects. Here we show that AlphaFold216 (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 tested docking methods, including other Fold and Dock methods17,24, methods based on shape complementarity30,32 and template-based docking. Using optimised MSAs with AF2, we can accurately predict the structure of heterodimeric complexes for an unprecedented SR of 62.7% on a large test set. The SR is higher in E.coli (76.4%) than in H. sapiens or S. cerevisiae (58.1% and 66.2% respectively).

Further, by analysing the predicted interfaces, we can predict the DockQ score33 (pDockQ) with an average error of 0.1, resulting in the separation of acceptable and incorrect models with an AUC of 0.95. This means that 31% of the models can be called acceptable at a specificity of 99% (or 54% at 90% specificity). Interestingly, no additional constraints are implemented in AF2 to pull two chains in contact, meaning that 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 (performance increase of 1% SR) by simply fusing MSAs from two HHBlits34 runs on Uniclust35 instead of using the MSAs from AF2. Fast MSA generation circumvents the main computational bottleneck in the pipeline. Using pDockQ makes it possible to separate truly interacting from non-interacting proteins with an AUC of 0.87, resulting in the separation of acceptable and incorrect models at an error rate of 1%, this results in the structure of 19,842 human heterodimeric protein structures. The computational cost to run all of this is ~5 days on an Nvidia A100 system and has since the development of the pipeline presented here, deemed FoldDock, been applied37.

Methods
Development set. A set of heterodimeric complexes from Dockground benchmark40 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 <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 34% Eukaryotic protein pairs, 38% Bacterial and 8% from mixed kingdoms, e.g., one bacterial protein interacting with one eukaryotic.

Test set. We used 1661 protein complexes with known interfaces from a recent study27 to test the developed pipeline. Here, three large biological assemblies were excluded. These complexes share <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 1481 AF2 MSAs. When folding, three of these (5AWE_D-5AWE_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 143 complexes.

For the mammalian proteins from Negatome, seven out of 1733 single chains were redundant according to Uniprot (C42Q38, 10L1J4, 10L25, K4CRX6, P62988, Q8N70, Q8T3B2). 34 had no matching species in the MSA pairing, 106 produced against the IntAct dataset from Negatome. This data deemed benchmark complexes. This dataset contains in total 3989 non-interacting pairs.

Non-interacting proteins. Two datasets of known non-interacting proteins were used, one from the same study as the positive test set. 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 and secondly complexes whose individual proteins were found in different APMS.

MSA block diagonalization. In addition to the default AF2 MSA, we generated an additional MSA by simply concatenating diagonally MSAs generated independently from each of the two chains. These MSAs were constructed by running HHblits v.3.1.0 against uniclust30_2018_08 with these options:

```
hhblits -E 0.001 -a all -a3m -n 2
```

The concatenation is done by joining side-by-side the two input chains; then sequences from one MSA are added, aligned to the corresponding input chain. Each sequence in the MSA is then elongated with gaps (on the right side if it is the left sequence MSA or the other way around), to reach the length of the two concatenated input chains. The process is then repeated for the other input chain MSA to complete the block diagonalization.

**Number of effective sequences (Neff).** To estimate the information in each MSA, we clustered sequences at 62% identity, as described in a previous study. The number of clusters obtained in this way has been used to indicate a Neff value for each MSA.

Unaligned FASTA sequences were extracted from the three AF2 default MSAs. Obtained sequences were processed with the CD-HIT software using the following command:

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

No optimisation of the RF protocol was made here.

**Template-based docking.** For comparison, a template-based docking protocol was run using the provided webserver (https://zougrouptoolkit.missouri.edu/MDockPP). This docking algorithm is based on fast Fourier transform (FFT). The docking results are assessed using the “in-house” scoring function iTScorePP.

**GRAMM.** For comparison, a rigid-body docking method, GRAMM, was used. Here, two protein models are docking using a 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. The atom-atom contact energy AAC138 is used to score and rank all poses, as this has been shown to provide better results than shape-complementarity alone.

**Template-based docking.** For comparison, a template-based docking protocol referred to as “TMdock” is also adopted. The adopted template library includes 11756 protein complexes obtained from the Dockground database (release 28-10-2020). Monomers from target complexes are structurally aligned with complexes in the supplied libraries (depleted of the target structure PDB ID) in order to identify the best available template structure. The bound form of the template structures was used. 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 “MDock 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.
AlphaFold-multimer. The simultaneous fold-and-dock program based on the same principles as AF2, AlphaFold-multimer\textsuperscript{28}, was run with the default settings. These entail creating four different MSAs. Three different MSAs are created by searching Uniref90 v.2020\textunderscore{}01\textsuperscript{46}, Uniprot v.2021\textunderscore{}04\textsuperscript{48} and MGnify v.2018\textunderscore{}12\textsuperscript{45} with jackhmmer from HMMER\textsuperscript{347} and one joint is created by searching the Big Fantastic Database\textsuperscript{44} (BFD) and uniclust30\textunderscore{}2018\textunderscore{}08\textsuperscript{35} with HHBlits\textsuperscript{33} (from hh-suite v.3.0\textunderscore{}beta.3 version 14/07/2017).

The results from the Uniprot search are used for MSA pairing and the results from the remaining searches are used to create a block-diagonalized MSA, similar to the procedures described above. All four MSAs are then used to fold a protein complex. Some complexes failed due to computational limitations, resulting in 1458 out of 1481 complexes successfully folded.

Scoring models. 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\textsuperscript{33} was then run (without any special settings) to compare the predicted and actual interfaces. This program

![Comparison of different MSAs](image-url)
computes interfaces using a combination of three different CAPRI quality measures (Fm, 1RM, and 1RMS) converted to a continuous scale, where an acceptable model comprises a DockQ score of at least 0.23.

**Ranking models.** To analyze the ability of AF2 to distinguish correct models as well as interacting from non-interacting proteins, we analyze the separation between acceptable and incorrect models as a function of different metrics on the development set: the number of unique interacting residues (CIs from different chains within 8 Å from each other), the total number of interactions between CIs from different chains (referred to as the number of interface contacts), average predicted IDDT (pDDT) score from AF2 for the interface, the minimum of the average pDDT for both chains and the average pDDT 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:

\[
\text{TPR} = \frac{TP}{TP + FN} \\
\text{FPR} = \frac{FP}{FP + TN}
\]

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 an 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:

\[
\text{AUC} = \int_{0}^{1} \frac{TPR(x)}{FPR(x)} \, dx
\]

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 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:

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

**pDockQ.** As it is not only desirable to know when a model is accurate but also how accurate this model is, we developed a predicted DockQ score, pDockQ. This score is created by fitting a sigmoidal curve (Fig. 2c) using “curve_fit” from SciPy v1.1.1, to the DockQ scores using the average interface pDDT multiplied with the logarithm of the number of interface contacts, with the following sigmoidal equation:

\[
pDockQ = \frac{L}{1 + e^{-k \cdot (x - b)}}
\]

where \( x = \) average interface pDDT \cdot log(number of interface contacts) \( L = 0.724, x_0 = 152.611, k = 0.052 \) and \( b = 0.018 \).

**Analysis of models.** To analyze the possibility of determining when AF2 can model a complex correctly, we analyse the structures and the MSAs. We investigated: the number of effective sequences (Neff), the secondary structure in the interface annotated using DSSP, 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, 3-helix), 4-turn helix (4 helix) and 5-turn helix (5 helix)), sheet (extended strand in parallel or antiparallel \( \beta \)-sheet conformation and residues in isolated \( \beta \)-bridges) and loop (residues which are in no known conformation).

In addition, we assess the PPV of the top N interface DCA signals using the paired MSAs. Here, \( N \) is the number of true interface contacts (CIs from different chains within 8 Å from each other). The PPV is therefore the fraction of the top N DCA signals in the interface that are true contacts. The DCA signals are computed using GaussDCA.

\[
\text{Interface PPV} = \frac{\text{Number of correct contacts among top N interface DCA signals}}{N}
\]

**Computational cost.** To compare the computational required for each MSA, we compared the time it took to generate MSAs for three protein pairs (PDB: 4CA5, P-O, 5X1L-A and 2X1L-2_M) using either the block orthogonalization or AF2 protocol. The tests were performed on a computer using 16 GPU cores from an Intel Xeon E5-2690v4.

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

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The raw data used in this study, including multiple sequence alignments and predicted PDB files, are available in the figshare from Science for Life laboratory under accession code 16866202.v1. All other data supporting the findings of this study are available within the article and its supplementary information files. The results used to produce all figures can be found in the supplementary information. Additional information and relevant data will be available from the corresponding author upon reasonable request.

**Code availability**

All code to run FoldDock and reproduce the analysis here can be obtained here https://github.com/ElofssonLab/FoldDock (commit 2e4c96aa352338976260e0c064ceaa75392dec) under the Apache License, Version 2.0.

Received: 30 September 2021; Accepted: 11 February 2022; Published online: 10 March 2022

**References**

1. Liddington, R. C. Structural Basis of Protein–Protein Interactions. *Protein–Protein Interactions* 261, 3–14 https://doi.org/10.1385/1-9529-762-9:003 (2004).
2. Keshkin, 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. *EMBO J.* 22, 3486–3492 (2003).
4. Cong, Q., Anishchenko, I. V. Chivkinnov, 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. In *Proteins and Protein–Protein Interactions* (ed. Wakaman, G.) 115–146 (Springer, 2005).
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. USA* 109, 9438–9441 (2012).
8. Porter, K. A., Desta, I., Kozakov, D. & Vajda, S. What method to use for protein–protein docking? *Curr. Opin. Struct. Biol.* 55, 1–7 (2019).
9. Halperin, I., Ma, R., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 47, 409–443 (2002).
10. Shammas, S. L. et al. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *J. Biol. Chem.* 291, 6689–6695 (2016).
11. Eginton, C., Naganathan, S. & Beckett, D. Sequence-function relationships in folding upon binding. *Protein Sci.* 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. Kuruczuglu, Z. & Bonvin, A. M. J. Pre- and post-docking sampling of conformational changes using ClusTENM and HADDOCK for protein–protein and protein–DNA systems. *Proteins* 88, 292–306 (2020).
14. Lensink, M. F. et al. Blind prediction of homo- and hetero-protein complexes: The CASP13–CAPRI experiment. *Proteins* 87, 1200–1221 (2019).
15. 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).
16. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
17. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876 (2021).
18. 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. USA* 119, e2113348119 (2022).
19. Chowdhury, R. et al. Single-sequence protein structure prediction using language models from deep learning. Preprint at bioRxiv https://doi.org/10.1101/2021.08.02.454840 (2021).

20. 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 6, e19729 (2011).

21. 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. USA 106, 67–72 (2009).

22. Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics 34, i802–i810 (2018).

23. Yang, J. et al. Improved protein structure prediction using predicted inter-residue orientations. Preprint at bioRxiv https://doi.org/10.1101/846279 (2019).

24. Pozzati, G. et al. Limits and potential of combined folding and docking using PconsDock. Bioinformatics 38, 954–961 (2021).

25. Lamb, J. & Elofsson, A. pyconsFold: a fast and easy tool for modelling and docking using distance predictions. Bioinformatics https://doi.org/10.1093/bioinformatics/btaa353 (2021).

26. Szurmant, H. & Weigt, M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. Curr. Opin. Struct. Biol. 50, 26–32 (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. Rozov, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at bioRxiv https://doi.org/10.1101/2021.04.04.465034 (2021).

29. Lensink, M. F. et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. Proteins https://doi.org/10.1002/pro.26222 (2021).

30. Huang, S.-Y. & Zou, X. MDockPP: a hierarchical approach for protein-protein docking and its application to CAPRI rounds 15–19. Proteins 78, 3096–3103 (2010).

31. 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 Res. 42, D396–D400 (2014).

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

33. Basu, S. & Wallner, B. DockQ: a quality measure for protein-protein docking models. PLoS ONE 11, e0161879 (2016).

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

35. Mirdita, M. et al. UniProt families: predicting residue contacts and protein-interaction partners. PLoS ONE 9, e92721 (2014).

Acknowledgements

We thank Petras Kundrotas for supplying the new heterodimeric proteins without templates in the PDB. We also thank Liming Qiu and Xiaoxin Zou for their help with running their docking program MDockPP in a timely manner. 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. All financial support and computational resources were received by A.E.

Author contributions

P.B. and G.P. performed the studies; all authors contributed to the analysis. P.B. wrote the final version. A.E. obtained funding.

Funding

Open access funding provided by Stockholm University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28865-w.

Correspondence and requests for materials should be addressed to Patrick Bryant or Arne Elofsson.

Peer review information Nature Communications thanks Rodrigo Honorarto and Shoshana Wodak for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.