Benchmarking Peptide-Protein Docking and Interaction Prediction with AlphaFold-Multimer

Isak Johansson-Åkhe[1] Björn Wallner[1*]

[1] Division of Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden

November 16, 2021

Abstract

Protein interactions are key in vital biological process. In many cases, particularly often in regulation, this interaction is between a protein and a shorter peptide fragment. Such peptides are often part of larger disordered regions of other proteins. The flexible nature of peptides enable rapid, yet specific, regulation of important functions in the cell, such as the cell-cycle. Because of this, understanding the molecular details of these interactions are crucial to understand and alter their function, and many specialized computational methods have been developed to study them.

The recent release of AlphaFold and now AlphaFold-Multimer has caused a leap in accuracy for computational modeling of proteins. Additionally, AlphaFold has proven generalizable enough that it can be adapted to a number of specialized protein modeling challenges outside of the original single-chain protein modeling it was trained for.

In this paper, the ability of AlphaFold to predict which peptides and proteins interact as well as its accuracy in modeling the resulting interaction complexes are benchmarked against established methods in the fields of peptide-protein interaction prediction and modeling. We find that AlphaFold-Multimer consistently produces predicted interaction complexes with the best DockQ-scores, with a mean DockQ of 0.49 for all 247 complexes investigated. Additionally, it can be used to separate interacting from non-interacting pairs of peptides and proteins with ROC-AUC and PR-AUC of 0.75 and 0.54, respectively, best among the method in benchmark. However, there is still room for improvement, for a decent precision of 0.8 it only recalls 0.2 of the positive examples (FPR=0.01), which means the will miss many true interactions. By combining AlphaFold-Multimer with InterPep2 the model quality for interacting proteins is increased, but it does not improve the separation of interacting from non-interacting.

1 Introduction

Protein-protein interactions are central to all biological process, and understanding the molecular details of these interactions are crucial to understand

*To whom correspondence should be addressed (bjorn.wallner@liu.se)

and alter their function. Of protein-protein interactions, up to 40% are considered peptide-protein interactions, many of which are responsible for vital functions such as cell life-cycle regulation (Lee *et al.*, 2019; Tu *et al.*, 2015). These are interactions between a protein and a smaller peptide fragment, which is sometimes part of a disordered region of a larger protein. Because peptide fragments display a high degree of flexibility and are often disordered when unbound, investigating the molecular details of such interactions and indeed even identifying direct interaction at all has been difficult both from a computational and experimental point of view, and specialized solutions have been developed to investigate peptide-protein interactions specifically (Petsalaki and Russell, 2008; Alam *et al.*, 2017; Lei *et al.*, 2021; Öztürk *et al.*, 2018; Johansson-Åkhe *et al.*, 2021).

The unprecedented accuracy of AlphaFold (Jumper *et al.*, 2021) and the recent release of the source code have transformed the field of computational and structural biology. It is now possible to achieve highly accurate protein structure prediction on par with experimental accuracy for many proteins, e.g. for human proteins 58% can be modeled accurately (Tunyasuvunakool *et al.*, 2021), as a comparison the previously experimentally determined structures of human proteins covered 17%. This and the improved methods that will follow in the foot steps of AlphaFold will have a huge impact not only on structural biology but the whole life-science field.

Despite the fact that AlphaFold was trained on monomeric strucures, it has demonstrated an impressive ability and stability to allow manipulation of its input to predict protein complexes, i.e. using a 200 residue gap to infer a chain break (Bryant *et al.*, 2021) or using a flexibel linker (Ko and Lee, 2021; Tsaban *et al.*, 2021). However, it is clear that even tough the input-adapted versions of AlphaFold performed better than state-of-the-art, the recently released retrained AlphaFold-Multimer system (Evans *et al.*, 2021) successfully predicts the interface (DockQ (Basu and Wallner, 2016) ≥ 0.23) for 67% of the cases. and at high accuracy (DockQ ≥ 0.8) for 23% of the cases on a set of 4,433 non-redundant interfaces, an improvement over the input-adapted version with 60% and 92%, respectively (Evans *et al.*, 2021).

In this manuscript, we investigate whether these milestones are applicable not only to the prediction of complexes of globular structures, but if AlphaFold and AlphaFold-multimer can also be used to further the field of peptide-protein interaction prediction as well as peptide-protein complex modeling.

2 Methods

2.1 Dataset

A test-set was constructed for the benchmarking both of peptide-protein interaction prediction as well as peptide-protein complex modeling.

The positive samples of the test set were the same as in the state-of-the art paper for peptide-protein interaction prediction by Lei *et al.* (2021), where the positive set consists of 262 peptide-protein complexes with experimentally solved structures in the PDB (Berman *et al.*, 2000). A negative set was created in the same manner as in that manuscript; 1310 random pairings of peptides and protein receptors from the positive set was used as a negative set, discard-

ing any random pairing which paired peptides and protein receptors from the same PDB ID. Although randomly pairing proteins is no guarantee of lack of interaction, the resulting false negative rate will be statistically insignificant, especially considering the proteins hail from different species, and is as such common practice for constructing negative sets for protein-protein interaction prediction (Wang *et al.*, 2019; Guo *et al.*, 2008).

This scheme results in a dataset which can both be used to evaluate the docking performance of methods investigated on the positive set as well as be used for benchmarking the capacity to predict whether peptide-protein pairs interact or not.

2.2 AlphaFold versions

One input-adapted version of AlphaFold monomer and three variants of AlphaFold-Multimer were included in the benchmark, as outlined below:

- *AF-gap*: This is an input-adapted version using AlphaFold (Jumper *et al.*, 2021) monomer. The input is adapted by placing a 200 residue gap between the chains and pairing the MSAs diagonally (Bryant *et al.*, 2021). The models were ranked by the average plDDT of the peptide as in Tsaban *et al.* (2021), and rank 1 was selected.

- *AFmulti-reduced_dbs*: This is AlphaFold-Multimer (Evans *et al.*, 2021) using the reduced database setting. The models were ranked by the *ranking_confidence* score and rank 1 was selected. The *ranking_confidence* is a linear combination of the interface score, $ipTM$, and overall structural score, pTM: $0.8ipTM + 0.2pTM$.

- *AFmulti-full_dbs*: This is AlphaFold-Multimer (Evans *et al.*, 2021) using the large sequence databases, BFD and Uniclust30. The models were ranked by the *ranking_confidence* score and rank 1 was selected.

- *AFmulti-full_dbs_template*: This is AlphaFold-Multimer (Evans *et al.*, 2021) using the large sequence databases, and templates. Templates newer than 2019-10-14, E-values better than 10^{-20} to the target, or for the peptides $> 95\%$ sequence identity over the whole sequence, were filtered to ensure the target protein or close homologs could not be used. The models were ranked by the *ranking_confidence* score and rank 1 was selected.

AlphaFold was run without the final relaxation step using Amber, thus only the unrelaxed models were evaluated.

2.2.1 AlphaFold database versions

- Uniclust30 from `UniRef30_2021_06`
- Uniref90 from Aug 9, 2021.
- Uniprot, TrEMBL+SwissProt, from Nov 3, 2021.
- BFD database (Steinegger and Söding, 2018), clustering of 2.5 billion sequences from Uniprot/TrEMBL+SwissProt, Metaclust, Soil and Marine Eukaryotic Refence Catalog. Downloaded March 2019.

```
bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_cs219.ffindex
MD5 hash:  26d48869efdb50d036e2fb9056a0ae9d
```

- Mgnify, from `2018_12`.
- PDB, mmcif and seqres, from Nov 3, 2021 (restrictions where applied at run-time, see above).

2.3 InterPep2

InterPep2 (Johansson-Åkhe *et al.*, 2020) is a template-based method for peptide-protein complex structure prediction, generating multiple peptide conformations and using TMalign (Zhang and Skolnick, 2005) and InterComp (Mirabello and Wallner, 2018) for finding structural templates of interaction and evaluating them by a random forest regressor. Search models for protein receptor was created using AlphaFold monomer and InterPep2 was run with standard settings but allowed no templates deposited into the PDB later than 2019-10-14, and additionally allowed no templates matching query protein receptors with an E-value better than 10^{-20} to ensure the query protein itself or particularly close homologs could not be used as templates for prediction. The random forest predicted suitability of the top-ranking template is used as the score for interaction prediction.

InterPep2 was run without the more computationally intensive refinement protocols, generating only coarse models. This will result in a lower number of sub-Ångström decoys but will only have minor to no effect on the positioning the peptide at the correct binding site (Johansson-Åkhe *et al.*, 2020).

2.4 ZDOCK

ZDOCK is a fast and accurate FFT-based rigid-body docking method utilizing pairwise statistical potentials, and is a staple method for benchmarking protein-protein docking performance (Pierce and Weng, 2008; Pierce *et al.*, 2014; Wallner and Mirabello, 2017; Moal *et al.*, 2013; Yan and Huang, 2019). In this study, ZDOCK was used to generate 54000 decoys for each peptide conformation generated in the early steps of the InterPep2 protocol, a similar scheme as to that used in earlier ZDOCK studies and by ClusPro Peptidock or PIPER-FlexPepDock (Wallner and Mirabello, 2017; Porter *et al.*, 2017; Alam *et al.*, 2017), and the best-scoring decoy were selected as final prediction.

2.5 CAMP

The dataset of this study is constructed as in (Lei *et al.*, 2021), which presents a convolution neural network for peptide-protein interaction prediction called CAMP. As a benchmark reference in interaction prediction, the freely available model of CAMP run on all test targets.

No retraining was performed, the model was run as is. But the input since the test set is fetched from the same paper CAMP's training set was defined in.

DockQ	Model Quality
≥ 0.23	Acceptable
≥ 0.50	Medium
≥ 0.80	High

Table 1: Thresholds for DockQ measure of docked model quality.

2.6 Performance Measures

2.6.1 DockQ

To assess the quality of docked peptide-protein complexes, the DockQ program was used. DockQ assesses the quality of a docked complex with regards to ligand root mean square deviation (LRMSD), the root mean square deviation of the interface residue conformation (iRMSD), and the fraction of native contacts recalled (fnat). The quality is measured on a scale of 0.0 to 1.0, with 0.25 representing models generally considered acceptably close to native, and 0.8 or above representing models with sub-Ångström quality (Basu and Wallner, 2016), see Table 1.

2.6.2 Receiving Operand Characteristic (ROC)

A ROC curve measures a method's capacity to discover positive samples versus its number of misclassifications with a sliding score threshold, and plots true positive rate (TPR) versus false positive rate (FPR) for different threshold cut-offs. $TPR = recall = TP/P$ (TP=true positives, P=total positives in set). $FPR = FP/N$ (FP=false positives, N=total negatives in set). The ROC area under the curve (AUC) is frequently used as a measure of classification method performance (Johansson-Åkhe *et al.*, 2020; Lei *et al.*, 2021).

2.6.3 Precision, Recall, and F1

With an unbalanced dataset, TPR-FPR ROC curves and AUC can be used to benchmark methods against each other but fails to capture the actual absolute performance of the methods, for which Precision-Recall curves and AUCPR (AUC under such a curve, also referred to as *average precision*) should be used instead (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006). $Precision = TP/(TP + FP)$. $Recall = TPR = TP/P$. F1 is the harmonic mean between precision and recall, defined as $F1 = (2 \cdot precision \cdot recall)/(precision + recall)$.

3 Results

In this study we benchmark the quality of peptide-protein docking models, as well as the ability to distinguish interacting pairs from non-interacting pairs of proteins and peptides. The benchmark was performed using a recently published data set peptide-protein interactions (Lei *et al.*, 2021) (see Methods for details). Four versions of AlphaFold: *AF-gap*, *AFmulti-reduced_dbs*, *AFmulti-full_dbs*, and *AFmulti-full_dbs_template*; InterPep2 (Johansson-Åkhe *et al.*, 2020), and ZDOCK (Pierce *et al.*, 2011) were included in the benchmark. In addition, CAMP (Lei *et al.*, 2021), a binary peptide-protein interaction prediction was

included the interaction benchmark. All comparisons is performed on the targets for which all methods produced predictions, 247 for the positive set and 1,089 for the negative.

3.1 Model Quality

Model quality of the first ranked prediction from each method was assessed using DockQ (Basu and Wallner, 2016). The mean DockQ is 0.48 for the best AlphaFold metod, AFmulti-full_dbs_template, compared to 0.33 and 0.15 for InterPep2 and ZDOCK, respectively, see Figure 1. However, the performance without using any templates is impressive with a mean DockQ of 0.46 for AFmulti-full_dbs, and 0.46 for AFmulti-reduced_dbs using smaller versions of the sequence databases compared to full_dbs. AF-gap performs worse with a mean DockQ of 0.36, but that is expected since it was not optimized for this task. All multimer versions of AlphaFold and InterPep2 correlates well with DockQ (R$\geq$ 0.75), while AF-gap has a lower correlation (R=0.47) and ZDOCK does not correlate at all (R=0.07), see Figure 2.

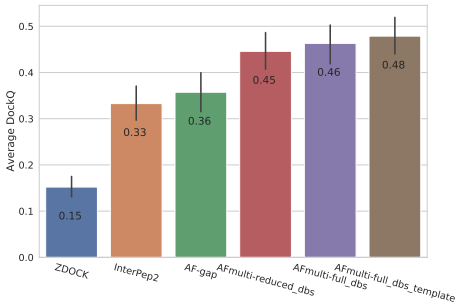
By analyzing the DockQ for the top 1 ranked models from AlphaFold-Multimer and InterPep2 it is evident that AlphaFold-Multimer and InterPep2 perform well on different targets, see Figure 3. For 70% (174/247) of the targets AlphaFold-Multimer is better than InterPep2, but for the remainder 30% (73/247) InterPep2 is better. Thus, a combination method would have the potential to improve the selection of models. In addition, since both AlphaFold-Multimer and InterPep2 have scores that correlate well with DockQ and are good at self-estimating when a prediction is good, i.e. you can trust high scores, but low scores might contain false negatives, see Figure 2. A combination method was constructed by using the scores from AlphaFold-Multimer and InterPep2 in the following way:

1. AlphaFold-Multimer models are selected if the ranking_confidence score over a threshold (>0.70).

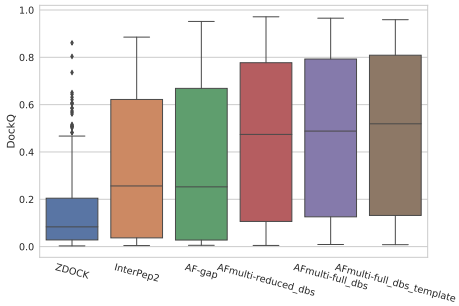
2. Otherwise InterPep2 models are selected if the InterPep2 score over a threshold. (>0.40)

3. AlphaFold-Multimer models are selected if both score are below the thresholds.

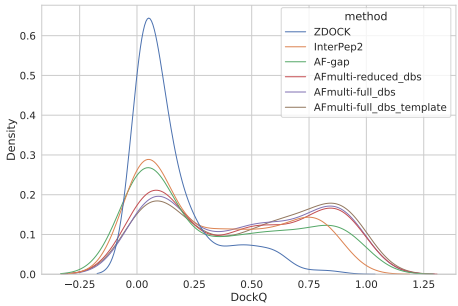
The AlphaFold-Multimer and InterPep2 thresholds were optimized, and the optimal values were determined to 0.7 and 0.4, respectively, see Figure 3. It should be noted that both these selected thresholds are exactly the default thresholds for acceptable predictions by AlphaFold-Multimer (Evans *et al.*, 2021) and InterPep2 (Johansson-Åkhe *et al.*, 2020). For 25 targets the combination method, named AFm+InterPep2, based on AFmulti-full_dbs_template , improves the selection by more than +0.2 DockQ, and only for four targets does it makes a selection that lowers the DockQ by more than -0.2, see Figure 3. The combination method improves the selection and the average DockQ is increased from 0.48 for AFmulti-full_dbs_template to 0.52 for AFm+InterPep2, see Figure 4. This might seem surprising considering the average DockQ for InterPep2 is only 0.34, but the reason for the improvement is that the InterPep2



(a) Average DockQ, error bars correspond to 95% confidence intervall.



(b) Distribution of DockQ scores



(c) Distribution of DockQ scores using kernel density estimation.

Figure 1: Model quality measured by DockQ

score is good at discriminating between at least acceptable and incorrect models, i.e. if the score is above the threshold (>0.4) it is most likely an acceptable prediction.

By examining the quality in more detail it can be seen that AlphaFold-Multimer produces many more High quality models compared to the other methods: 63 compared to only 41 and 23 for AF-gap, and InterPep2, respectively, see Figure 4. However, the AFm+InterPep2 has still the most High quality models with 67. Of the different versions of AlphaFold-Multimer tested it is consistently the version with access to most information that is superior. Overall, the best AlphaFold-Multimer is able to predict at least an Acceptable model for 160/247 (64%), while the AlphaFold-Multimer with least information (AFmulti-reduced_dbs) produces 149/247 (60%) Acceptable models. AFm+InterPep2 finds most Acceptable models 179/247 (72%)

Finally, the quality of the models produced by the methods were compared head-to-head, by calculating the quality difference per target and comparing the average, see Figure 6. It was also tested if this difference was significantly different from zero using Wilcoxon signed-rank test, see right panel in Figure 6. It can be concluded that full_dbs versions of the AlphaFold versions are significantly better than the reducded_dbs and that the AFm+InterPep2 is significantly better all other methods including the best AlphaFold-Multimer version.

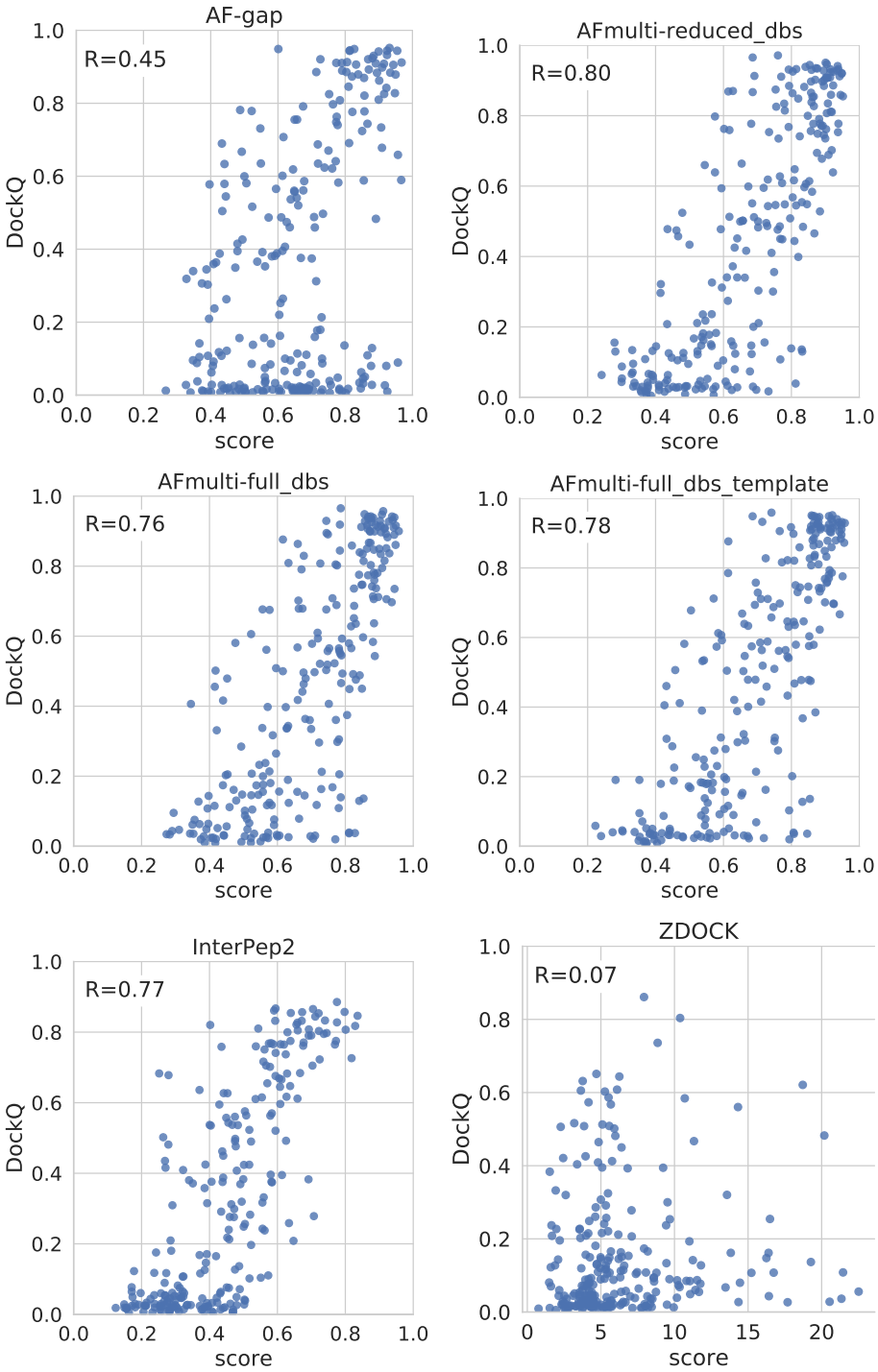
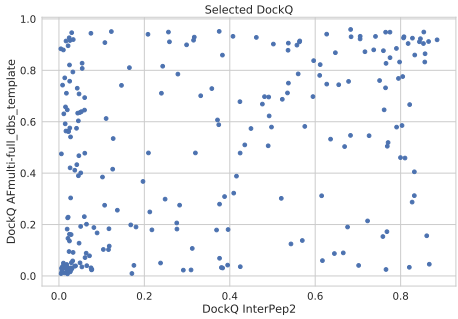
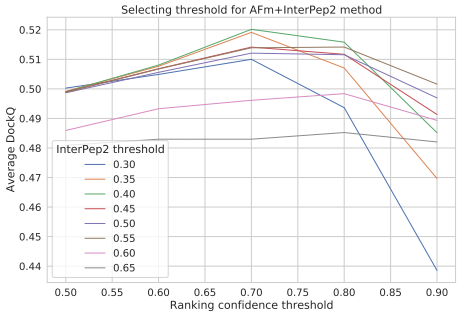


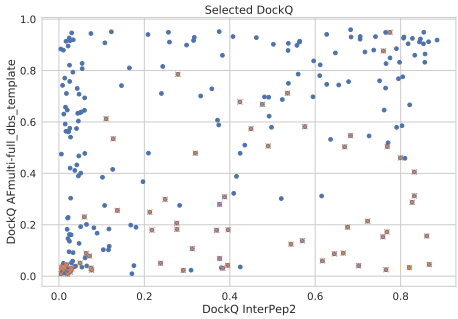
Figure 2: Correlation between method score and DockQ. For AFmulti the score is the *ranking confidence*, for AF-gap the score is the mean plDDT for the peptide. ZDOCK score is normalized by the total number of residues.



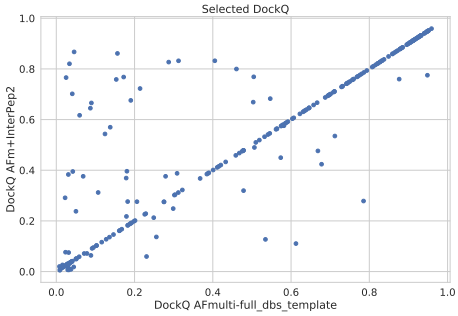
(a) Comparing selected models by AFm and InterPep2.



(b) Selecting threshold on AFm and InterPep2 to improve selection.

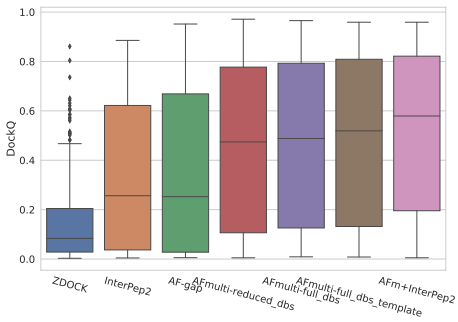


(c) Selection for the combination method AFm+InterPep2: blue points are from AFm and orange points are from InterPep2

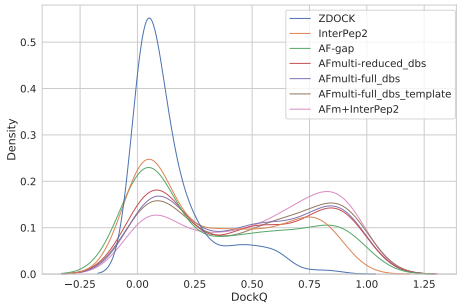


(d) Comparing AFm+InterPep2 to AFm, only in four cases the choice is suboptimal.

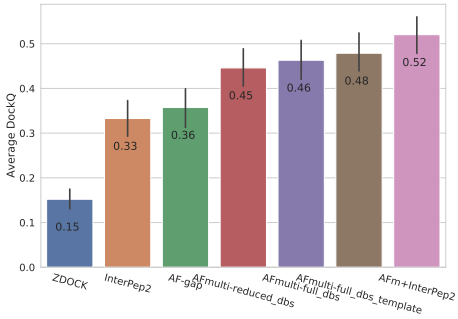
Figure 3: Combination of AlphaFold and InterPep2



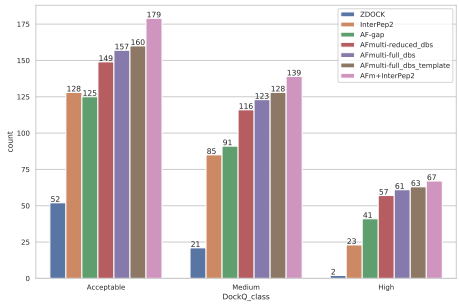
(a) Distribution of DockQ scores



(b) Distribution of DockQ scores using kernel density estimation.



(c) Average DockQ scores



(d) Selection of Acceptble, Medium and High quality models.

Figure 4: Distribution of model quality.

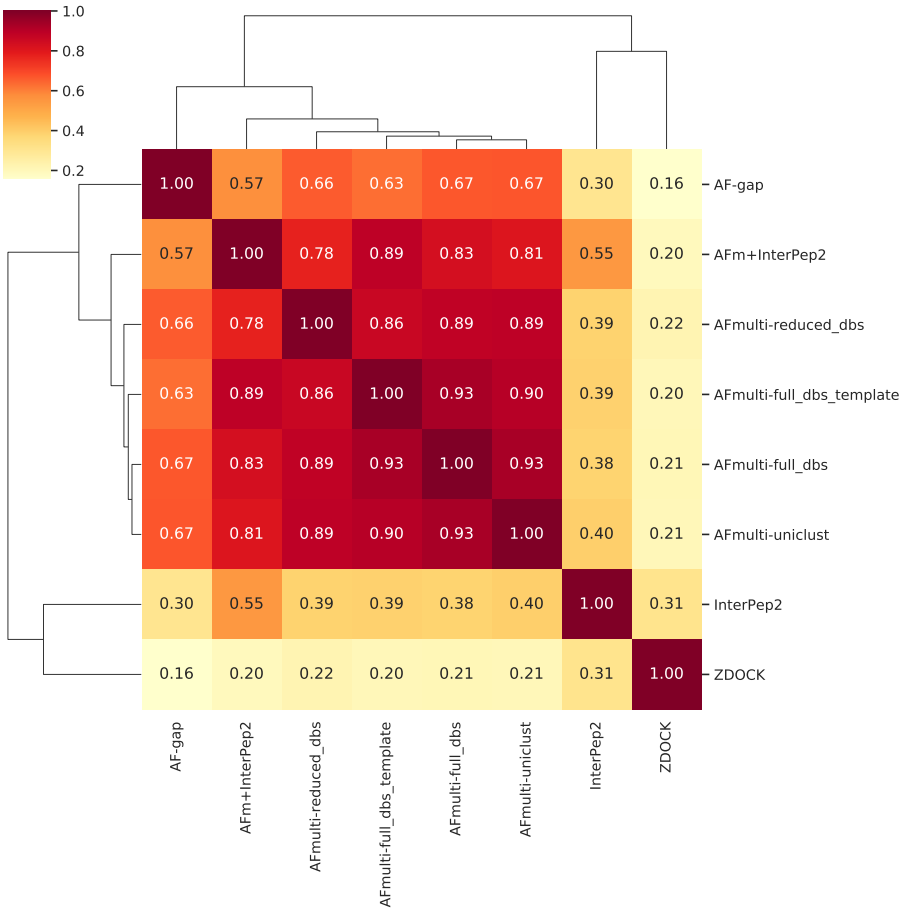
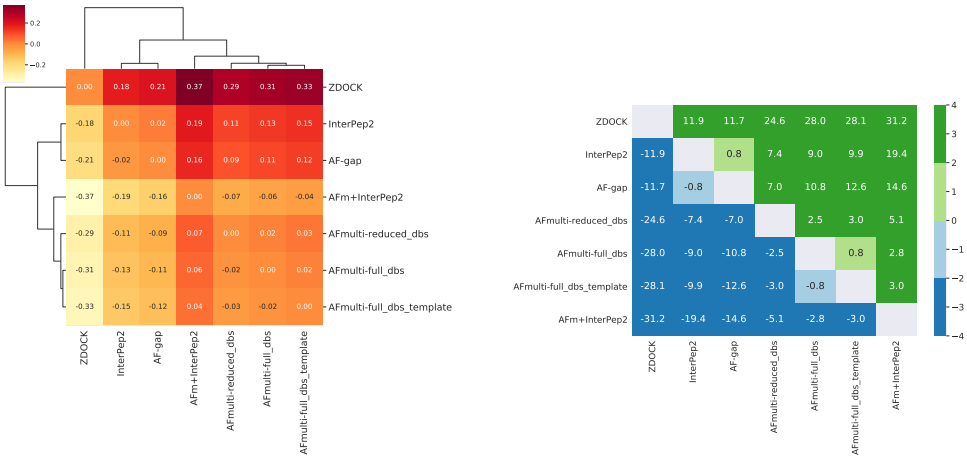


Figure 5: Clustering of model quality selected by the methods.



(a) Difference in the mean of selected models.

(b) Signed P-values for the statistical significance on the difference.

Figure 6: (a) Difference in the mean of selected models. The difference is taken as the quality method on (X-Y), a positive score means the method on the X-axis selects better models than the method on the Y-axis. (b) The signed P-values for Wilcoxon signed-rank test for the difference. The P-values are transformed using -log10(P-value)*sign(diff), i.e. for signed P-values, positive values indicates that the P-value is for a positive improvement and vice versa for negative. The scale is chosen so darkgreen corresponds to a signed P<0.01, i.e. for darkgreen boxes the methods on X-axis are significantly better than the method on the Y-axis.

3.2 Interaction prediction

The positive set used to assess model quality above was merged with the negative set containing peptide-protein pairs that do not bind (see Methods). The scores for binding vs. non-binding pairs are shown in Figure 7. In general, pairs that binds is scored higher than the negative set across all methods. The ROC-AUC curve is 0.80 for AFm+InterPep2 compared to 0.75 for AlphaFold-Multimer. Performance for interaction prediction needs to be assessed at low FPR to enable all-vs-all comparisons with high confidence predictions without too many false postives. At FPR=0.01 they both recall around 0.2 (TPR) of the positive examples, see Figure 8. From the Precision-Recall curve 0.2 recall corresponds to 0.8 precision for the best methods; corresponding to the score thresholds 0.85, 0.60, and 0.85 for AlphaFold-Multimer, InterPep2, and AFm+InterPep2, respectively, see Figure 9. At the threshold for low confidence prediction in AlphaFold (0.70) the precision is 0.5, meaning that a prediction over that threshold has a fifty-fifty chance of being correctly predicted as interacting. The low confidence threshold for InterPep2 is 0.4 and corresponds to precision 0.3 (Figure 9).

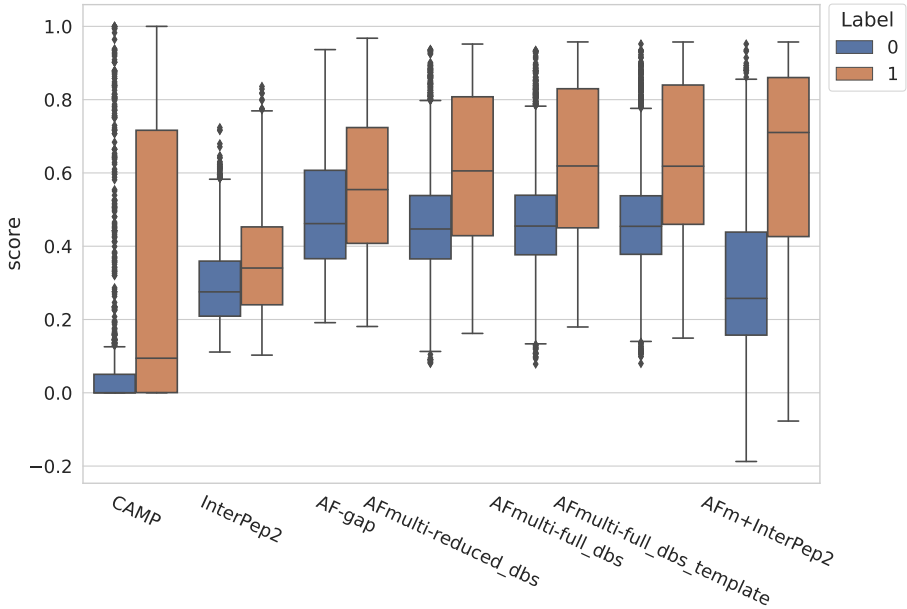
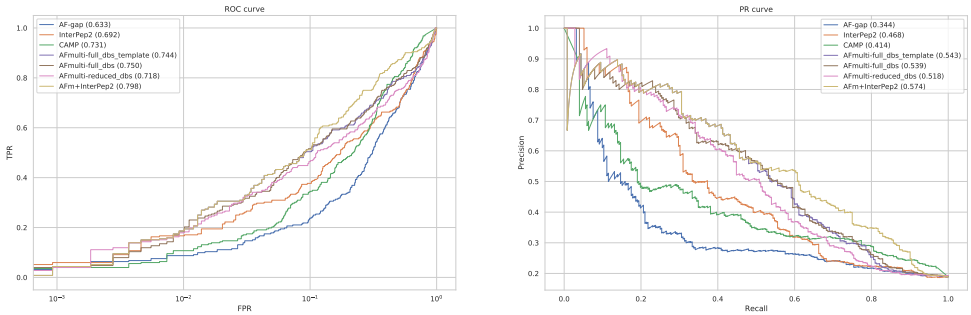


Figure 7: Score distribution for binding (label: 1)/non-binidng (label:0)



(a) ROC curve with the x-axis in log-scale. The values in the legend correspons to ROC-AUC

(b) Precision recall curve. The values in the legend correspons to PR-AUC. V

Figure 8: ROC and Precision-Recall curves for binding/non-binding

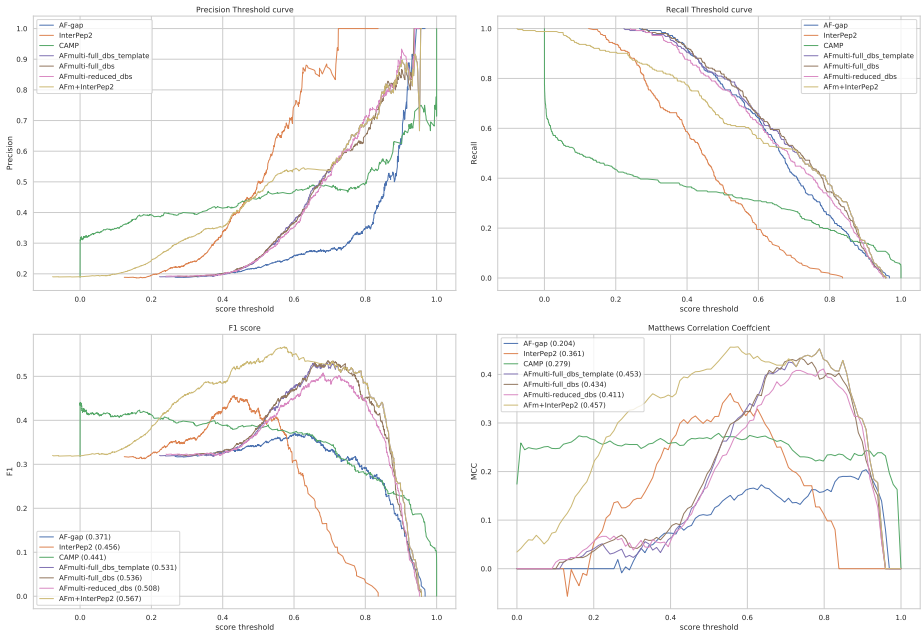


Figure 9: Score threshold vs performance. Useful to define thresholds depending on use-case.

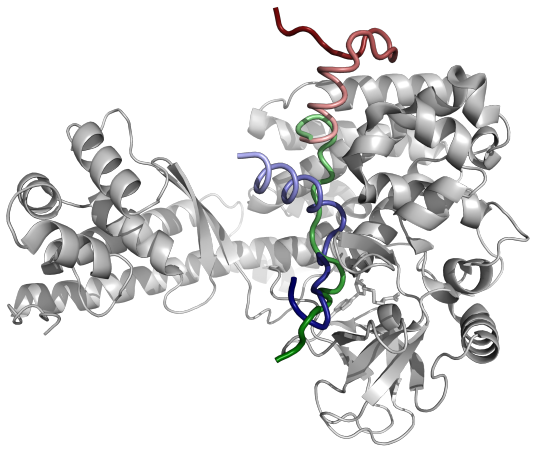


Figure 10: An example from the test set where the AFm+InterPep2 improves the prediction. The predictions are of a cytoplasmic actin peptide binding to actin-histidine N-methyltranferase (PDB ID 6v63). The prediction by AlphaFold-Multimer (red) positions the very end of the C-terminal in the correct site (native peptide colored green), but positions the peptide in the opposite direction of the actual fold, leading to most of it binding outside the pocket. The prediction by InterPep2 (blue) is coarse with some minor clashes, but has positioned the peptide inside the correct binding site. Note how both AlphaFold and InterPep2 have adopted similar folds for the peptide.

3.3 Example - Improvements for AlphaFold

To understand why AlphaFold fails at certain samples as well as why the template-based method InterPep2 can help with prediction quality in these cases, cases where AlphaFold failed were examined. One such example can be seen in Figure 10, where AlphaFold has positioned part of the peptide correctly, but has flipped the orientation of the peptide and positioned the rest of it outside its binding pocket, as if to continue the peptide chain that direction. In fact, in 55% of cases where the AFm+InterPep2 method improved an AlphaFold-Multimer model from a failed docking to at least acceptable quality, the AlphaFold model was bound close to or at the correct binding site but in the wrong direction.

In 36% of such improved cases, AlphaFold positioned the peptide at the binding site of a non-peptide binding partner, either for a co-factor or another protein chain (dimerization site, for instance). AlphaFold seems able to model the receptor protein regardless of the presence of any other binding chains or co-factors (Jumper *et al.*, 2021; Tunyasuvunakool *et al.*, 2021). However, as is evident from these examples, the absence of other binding partners and co-factors from the AlphaFold-Multimer modeling seems to result in a negative bias on the docking performance. Co-factors and other binders being absent from a complex seems to have less effect on the performance of the purely template-based method InterPep2, perhaps explaining in part why it is a good complement to AlphaFold-Multimer.

Acknowledgments

This work was supported by a Swedish Research Council grant, 2016-05369 and 2020-03352, The Swedish e-Science Research Center, and Carl Tryggers stiftelse för Vetenskaplig Forskning, 20:453. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC), Knut and Alice Wallenberg (KAW), and Linköing University (LiU) at the National Supercomputer Center (NSC) in Linköping.

References

Alam, N., Goldstein, O., Xia, B., Porter, K. A., Kozakov, D., and Schueler-Furman, O. (2017). High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS computational biology*, **13**(12), e1005905.

Basu, S. and Wallner, B. (2016). Dockq: a quality measure for protein-protein docking models. *PloS one*, **11**(8), e0161879.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, **28**(1), 235–242.

Bryant, P., Pozzati, G., and Elofsson, A. (2021). Improved prediction of protein-protein interactions using AlphaFold2. *bioRxiv*, page 2021.09.15.460468.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, page 2021.10.04.463034.

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, **36**(9), 3025–3030.

Johansson-Åkhe, I., Mirabello, C., and Wallner, B. (2020). InterPep2: Global Peptide-Protein Docking using Interaction Surface Templates. *Bioinformatics (Oxford, England)*.

Johansson-Åkhe, I., Mirabello, C., and Wallner, B. (2021). Interpeprank: Assessment of docked peptide conformations by a deep graph network. *Frontiers in Bioinformatics*, **1**, 60.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, pages 1–11.

Ko, J. and Lee, J. (2021). Can AlphaFold2 predict protein-peptide complex structures accurately? *bioRxiv*, page 2021.07.27.453972.

Lee, A. C.-L., Harris, J. L., Khanna, K. K., and Hong, J.-H. (2019). A comprehensive review on current advances in peptide drug development and design. *International journal of molecular sciences*, **20**(10), 2383.

Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D., and Zeng, J. (2021). A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature Communications*, **12**(1), 5465.

Mirabello, C. and Wallner, B. (2018). Topology independent structural matching discovers novel templates for protein interfaces. *Bioinformatics*, **34**(17), i787–i794.

Moal, I. H., Torchala, M., Bates, P. A., and Fernández-Recio, J. (2013). The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics*, **14**(1), 286.

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, **34**(17), i821–i829.

Petsalaki, E. and Russell, R. B. (2008). Peptide-mediated interactions in biological systems: new discoveries and applications. *Current opinion in biotechnology*, **19**(4), 344–350.

Pierce, B. and Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance. *Proteins: Structure, Function, and Bioinformatics*, **72**(1), 270–279.

Pierce, B. G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one*, **6**(9), e24657.

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., and Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*, **30**(12), 1771–1773.

Porter, K. A., Xia, B., Beglov, D., Bohnuud, T., Alam, N., Schueler-Furman, O., and Kozakov, D. (2017). Cluspro peptidock: efficient global docking of peptide recognition motifs using fft. *Bioinformatics*, **33**(20), 3299–3301.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, **10**(3), e0118432.

Steinegger, M. and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, **9**(1), 2542.

Tsaban, T., Varga, J., Avraham, O., Ben-Aharon, Z., Khramushin, A., and Schueler-Furman, O. (2021). Harnessing protein folding neural networks for peptide-protein docking.

Tu, W. B., Helander, S., Pilstål, R., Hickman, K. A., Lourenco, C., Jurisica, I., Raught, B., Wallner, B., Sunnerhagen, M., and Penn, L. Z. (2015). Myc and its interactors take shape. *Biochimica et Biophysica Acta*, **1849**(5), 469–483.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, **596**(7873), 590–596.

Wallner, B. and Mirabello, C. (2017). Interpred: A pipeline to identify and model protein-protein interactions. *Proteins: Structure, Function, and Bioinformatics*, **85**(6), 1159–1170.

Wang, L., Wang, H.-F., Liu, S.-R., Yan, X., and Song, K.-J. (2019). Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Scientific reports*, **9**(1), 1–12.

Yan, Y. and Huang, S.-Y. (2019). Pushing the accuracy limit of shape complementarity for protein-protein docking. *BMC bioinformatics*, **20**(25), 696.

Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, **33**(7), 2302–2309.