Dear Professor Elofsson,

Your manuscript entitled "Improved prediction of protein-protein interactions using AlphaFold2" has now been seen again by our referees, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Nature Communications under the open access CC BY license (Creative Commons Attribution 4.0 International License).

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and our editorial requests in the attached documents. At the same time we ask that you edit your manuscript to comply with our policies and formatting requirements and to maximise the accessibility and therefore the impact of your work.

Please see the attached documents, listing a number of points that must be addressed. Failure to comply with our editorial requests will cause delays in accepting your manuscript. Please also see the *Nature Communications* formatting instructions for further information.

- If you wish, an interesting image (but not an illustration or schematic) for consideration as a **Featured Image on the Nature Communications homepage.** The file should be 1200x675 pixels in RGB format and should be uploaded as a Related Manuscript File. In addition to our home page, we may also use this image (with credit) in other journal-specific promotional material.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

No further comments, my previous points have been properly addressed and the manuscript changed to reflect the changes.

We are delighted our changes were satisfactory and thank you for your excellent comments throughout the revision process.

Reviewer #2 (Remarks to the Author):

The revised version is largely improved.

The comparison with the performance of another performant docking method MDockPP, adds credibility to the study.

We are happy this addition was satisfactory and thank you for this suggestion.

Particularly welcome is the development of the continuous model quality assessment criterion pDockQ, shown here to effectively outperform several versions of the AF2 pLDDT

reliability criterion and other simple measures for ranking models and segregating interacting from non-interacting protein pairs.

We thank you for this kind acknowledgement and agree this score is useful for both structure quality assessment and separating interacting from non-interacting proteins.

Also interesting is the deeper analysis of the MSA features contributing to successful prediction, which suggests that interface evolutionary signals as measured by the fraction of interface contacts recalled by DCA, have a strong impact on the prediction results.

Indeed it appears so, that the evolutionary signal in the interface - here measured using DCA - has a strong impact on the outcome, opening up for future improvements in docking by improving this signal.

Evaluating the performance of AlphaFold-Multimer is a marginal addition given that the corresponding DL model was in fact trained on the dataset used here as the test set.

This is correct and we are in full agreement.

I have only a few outstanding comments

Results section:

-Development versus test set performance: The significant discrepancy between the performance of AF2 on development set (33.3% - 39.4%) and the test set (57.8% - 58,4%), is striking. This suggests that AF2 performance is dataset dependent. While this is not unexpected, it begs for a comment. It doesn't seem to result from an organism bias, since the AF2 performance on the smaller development dataset is much lower even though it features a higher proportion of bacterial proteins, for which the authors observe a higher SR level in AF2 predictions.

We thank you for this comment. We agree there is a big difference and have added a comment about this under limitations, where we suggest that performance should be assessed on as large non-redundant datasets as possible to ensure any selection bias does not impact the results. We do not know exactly the origin of the difference. We tried to examine the most obvious differences between the two sets (protein size, species, size of MSAs etc) but did not find anything obvious that separated the sets. Unfortunately, trying to pinpoint the origin of the difference is beyond the goals of this study.

Lines 95-97, 139-140
…'protocol performs quite close to (63% vs 72%) the recently developed AF-Multimer which was developed using the same data as the test set here, making a direct comparison difficult.
'was developed' should be replaced by 'was trained'.'

We have changed this phrasing as suggested to clarify that it was indeed trained.

Lines :132-134

The sentence is misleading since the performance of the 3 docking methods is clearly not good.
Suggestion:
Replace “The reason for GRAMM’s, TMdock’s and MDockPP’s good performance is likely due” by “The reason GRAMM’s, TMdock’s and MDockPP’s reach this performance level is likely due”

We thank you for this suggestion and have changed the phrasing as suggested.

Lines 362-364
Projecting the fraction of human heterodimers predicted at the current 1% error rate, on the basis of the number of pairwise human PPI in the String DB (11.9 million) is overdoing it, since it is well known that a sizable fraction of the interactions in String are non-physical. The paragraph should be rephrased accordingly.

We have changed this statement to reflect a more realistic modeling scenario, which we have also applied in practice (https://www.biorxiv.org/content/10.1101/2021.11.08.467664v1).