

The latest AI breakthroughs in structural biology: protein binder design and conformational state prediction

Corresponding Author: Dr Luciano Abriata

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Abriata provides a nice Commentary on potential impact of evolving methods for modeling alternative conformational states and conformational landscapes of proteins, and the rapidly evolving platforms for de novo protein binder design. I agree that these emerging breakthroughs - although still under development - have the potential to significantly expand our understanding of the nature of proteins and their roles in biology. I think the Commentary is timely and accurate.

Realizing that this is a Commentary and not a Review - I appreciate that it need not address every issue relevant to these emerging technologies. But I feel that a few key issues are missing or need to be better addressed in order provide a honest assessment to the community.

Minor points to be addressed:

1. The issue of training memorization in ML is particularly relevant for modeling multiple conformational states of proteins - where the modeling can be biased toward the conformational states present in the training data. Memorization in AI-based modeling is mentioned on line 34 ("excessive dependence on memorization and involving wrong physics (refs 12,14–16)), but memorization needs to be mentioned specifically in the context of modeling alternative conformational and ensembles (lines 49 - 94). The papers for memorization bias cited in line 34 really do not address this important issue of memorization impact on modeling alternative conformational states and conformational landscapes. However - it is well documented in a few recent papers which could be cited (e.g. doi's 10.1038/s41467-024-51801-z, 10.1038/s41467-024-51507-2; 10.1073/pnas.2412719121; 10.1371/journal.pcbi.1013590). Does bioemu overcome the impact of conformational state memorization? I do not think so.

2. Since much of the Commentary regarding modeling of conformational landscapes focuses on bioemu, it should be noted that bioemu models do not include sidechains. A related approach AF-RAVE (doi 10.7554/eLife.99702.2) may have advantages over bioemu for this problem.

3. lines 54-55 - regarding "Special Interest Group on Modeling Ensembles and Alternative Conformations of Proteins". This should cite the corresponding publication (doi 10.1002/prot.26833).

4. The discussion of recent progress in binder design is valuable - but does not address the very important (and not yet solved) challenge of in silico scoring of predicting binders. Related to this is the challenge of accurately predicting binding affinities. And of course also small molecule binding affinities. These challenges should also be mentioned, briefly discussed, and appropriately cited.

5. The OpenFold project is mentioned in passing. This is a significant innovation in providing easily accessible path to retraining an AF-like AI with target-specific data. It should be discussed briefly (1 sentence may suffice) and cited.

6. lines 185 -187. The CASP requirement for atom-specific confidence scores on models motivated also the pLDDT score in AF - this point should be made here.

7. Fig 1. Panels C and D are interesting - but this aspect of AI is not discussed much in the main text. Is this part of the "next AF moment?". If you want to make this point - pls add a bit more discussion to the text.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Answers to the reviewer's comments
(original comments in plain format, answers in bold italics)

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

Abriata provides a nice Commentary on potential impact of evolving methods for modeling alternative conformational states and conformational landscapes of proteins, and the rapidly evolving platforms for de novo protein binder design. I agree that these emerging breakthroughs - although still under development - have the potential to significantly expand our understanding of the nature of proteins and their roles in biology. I think the Commentary is timely and accurate.

Realizing that this is a Commentary and not a Review - I appreciate that it need not address every issue relevant to these emerging technologies. But I feel that a few key issues are missing or need to be better addressed in order provide a honest assessment to the community.

I would like to thank the reviewer for the very constructive comments, and I am happy that he/she appreciates the relevance and timeliness of the topics discussed in the manuscript.

Minor points to be addressed:

1. The issue of training memorization in ML is particularly relevant for modeling multiple conformational states of proteins - where the modeling can be biased toward the conformational states present in the training data. Memorization in AI-based modeling is mentioned on line 34 ("excessive dependence on memorization and involving wrong physics (refs 12,14–16)), but memorization needs to be mentioned specifically in the context of modeling alternative conformational and ensembles (lines 49 - 94). The papers for memorization bias cited in line 34 really do not address this important issue of memorization impact on modeling alternative conformational states and conformational landscapes. However - it is well documented in a few recent papers which could be cited (e.g. doi's 10.1038/s41467-024-51801-z, 10.1038/s41467-024-51507-2; 10.1073/pnas.2412719121; 10.1371/journal.pcbi.1013590). Does bioemu overcome the impact of conformational state memorization? I do not think so.

Excellent point, as indeed the literature I cited in the original submission did not properly cover the problem of memorization. I have now cited these papers, with a corresponding, small amendment in the text. This amendment includes the clarification that BioEmu has not been proven to not suffer from these problems.

2. Since much of the Commentary regarding modeling of conformational landscapes focuses on bioemu, it should be noted that bioemu models do not include sidechains. A related approach AF-RAVE (doi 10.7554/eLife.99702.2) may have advantages over bioemu for this problem.

This is correct and extends to other AI models, so I have made a small amendment also citing papers for other models doing conformational sampling.

2. lines 54-55 - regarding "Special Interest Group on Modeling Ensembles and Alternative Conformations of Proteins". This should cite the corresponding publication (doi 10.1002/prot.26833).

Fair point. Done.

3. The discussion of recent progress in binder design is valuable - but does not address the very important (and not yet solved) challenge of in silico scoring of predicting binders. Related to this is the challenge of accurately predicting binding affinities. And of course also small molecule binding affinities. These challenges should also be mentioned, briefly discussed, and appropriately cited.

I now briefly comment on this important challenge, clarifying that affinity prediction is even harder than simply designing the binders. I have also cited some key relevant works regarding affinity estimation (so far available for protein-small molecule binding, as far as I am aware of – but I will be happy to incorporate citations discussing the challenges in protein-protein affinity estimation if they exist).

4. The OpenFold project is mentioned in passing. This is a significant innovation in providing easily accessible path to retraining an AF-like AI with target-specific data. It should be discussed briefly (1 sentence may suffice) and cited.

Fair point; I have now added a single-sentence comment explaining why such open models are valuable.

5. lines 185 -187. The CASP requirement for atom-specific confidence scores on models motivated also the pLDDT score in AF - this point should be made here.

Point made.

6. Fig 1. Panels C and D are interesting - but this aspect of AI is not discussed much in the main text. Is this part of the "next AF moment?". If you want to make this point - pls add a bit more discussion to the text.

This remains more anecdotal now, and with so much “hype” in the field of automation (and without full expertise from my side in this domain) it is difficult to advance how much this will really be the next AF moment or not. However, I see it very motivating and hence I would like to keep this section; for this, I have added some more discussion in the text as suggested.