Peer-Review

Wang, Yu-Hong A. 2025. “Future of Protein Structure Modeling.” *Journal of High School Science* 9 (2): 271–303.

This is a very well written and well presented manuscript; especially relevant to the Nobel prize in Chemistry this year. However, the manuscript presents data and analysis from the public domain and therefore does not add significantly to the existing corpus of knowledge in the field. I think - that with the excellent review that the author has put together - they should be able to make some more predictions regarding the ‘future’ of protein structure folding; even though these may be speculative in nature. Please address my comments below in the manuscript.
1.The title of your manuscript is “future of protein structure folding”, therefore, some predictions based on your review are in order. Can you speculate that there will be competing models based on AF2 in the future? [A] a model that will be based on changes to the AF2 architecture (with random weights to layers) so that the base model generalizes better, [B] ‘add-ons’ to the base AF2 architecture with different competing algorithms so as to reach greater accuracy in the the operations listed below:
p] loop structures
q]Intrinsic disordered domains
r]protein-protein complexes
s]high local disordered regions
t]low sequence conservation regions
and [c] a bias modification to the AF2 architecture depending on whether the end-user wants high bias to predict a particular (high priority) structure with high accuracy or low bias to predict a (low priority) structure with better generalizability.
Can you include some discussion in the manuscript. In other words, do you think that future predictive models will mostly involve changes to the ML architecture of AF2, mostly involve add-on’s to AF2 or let users tune the accuracy they want from AF2 by making its ‘hyperparameters’ user-defined? Any other trajectories?
2.You mention that training data - obtained by empirical means such as NMR, cryo-EM etc. is important to continue to increase the 0.08% ratio of 3D predicted to PDB sequence data. My question is: is there a need for ‘filler algorithms’ ? For example, if AF2 can predict with 99% Ground truth, the beta sheet and high sequence disorder regions of a fair number of proteins that fall into a certain category of the PDB database, then … for a protein in that category whose sequence is known, can we just perform an NMR or cry-EM on those regions to predict the structure and leave out other domains ? …. then feed this data into AF2 ? This way the predicted AF2 structure can be assumed to be 99% accurate with the actual structure (parts of which have still not been deciphered with NMR or cryoEM) . Would this be a legitimate way to increase the production of training data without compromising ground truth accuracy?
I look forward to the revised manuscript

---

1. The title of your manuscript is “future of protein structure folding”, therefore, some predictions based on your review are in order. Can you speculate that there will be competing models based on AF2 in the future?

   1.1-[A] a model that will be based on changes to the AF2 architecture (with random weights to layers) so that the base model generalizes better,

   Author’s Response

The author thanks the reviewer for catching the lack of speculation. Speculation has been made and included in the updated manuscript that there will be such competing models to AF2 in the future with random weights for better generalizability:

"The model generalizability is undoubtedly required for improving overall predictions, specifically for proteins uncommon to contemporary studies. Therefore, it is important to take in account successful de novo methods. One of which includes random perturbation, random adjustments of the model's weights, notably introduced by Johansson-Åkhe and Wallner as a prominent aspect to improving generalizability of the AlphaFold2 model." (see Page 18 of the updated manuscript)

1.1-[B] 'add-ons' to the base AF2 architecture with different competing algorithms so as to reach greater accuracy in the the operations listed below:

p]loop structures
q]Intrinsic disordered domains
r]protein-protein complexes
s]high local disordered regions
t]low sequence conservation regions

Author's Response

Competing algorithms for specific regions is also discussed below:

"The trade-off of added duration of the prediction program in turn for more accuracy as demonstrated by MULTICOM is worth so, due to the time still being in hours or days, as opposed to the expenses, months or years, or other difficulties required by experimental processes. Some of AF2's weak prediction regions like protein-protein complexes have shown to be still as or more accurate than other programs dedicated to its modeling, like traditional docking approaches. This is a typical result of supervised learning and other AI systems, where unintended outcomes are given without its specificity in the program. If specificity is provided, such as utilizing multiple algorithms for separate regions and applying new confidence evaluation metrics to compare the best models produced by either a new-algorithm-less or a new-algorithm-included process, allowing for the program to optimize and choose over the iterations the highest-accuracy model produced. In this incorporation, there should be region-specific or protein-specific algorithms for at least the following regions, as previously presented as large concerns and weaknesses in AF2's predictions in this review: protein-protein complexes, loop structures, intrinsic disordered domains, high local disordered regions, and low sequence conservation regions. Keeping these additions and the heavy use of MSAs could continue to provide super accurate predictions in more common proteins, increasing the efficiency of template and data collection. This path of highly accurate predictions for the sake of data collection could be a solution, while there is another more difficult path of finding an improved model focusing on generalization." (see Page 18-19 of the updated manuscript)

1.1-[c] a bias modification to the AF2 architecture depending on whether the end-user wants high bias to predict a particular (high priority) structure with high accuracy or low bias to predict a (low priority) structure with better generalizability.

Author's Response

The updated manuscript has also noted the possibility of a bias modification option for better end-user interactions:

"Next, it is possible to combine the strengths of template-based and free modeling through an automatic bias modification switch where there is a specific threshold which, through multiple accurate metrics of determining bias, chooses either a free modeling or a mixed approach depending on the commonality of the protein. This would allow for scientists to interact with the end result of the program, depending on their scientific wants and research goals. An improvement of overall metrics for the model's confidence levels (either replacing or in addition to pLDDT) for particular regions could be beneficial, paired with an iterative deduction of the best generated model. The MULTICOM program utilizes a blend of aforementioned approaches, providing a good direction to study in the upcoming years. This improvement would allow for better generalizability as a result of a more accurate overall and specific evaluations of the protein models in an unbiased manner." (see Page 18 of the updated manuscript)

1.2 Can you include some discussion in the manuscript. In other words, do you think that future predictive models will mostly involve changes to the ML architecture of AF2, mostly involve add-on's to AF2 or let users tune the accuracy they want from AF2 by making its 'hyperparameters' user-defined? Any other trajectories?

Author's Response

An additional section "overarching trends" has been added as an introduction to three other newly added sections "Building a Final Model", "The Core Model", and "Other Improvements" which compile and speculate on the promising directions suggested by the rest of the review:

"With its unique iterative system and quality metric system, MULTICOM has shown a promising direction for an AF2-based model, in its evidence of the improvement from AF2-based in de novo predictions. During the 2024 CASP16, MULTICOM proved its successful result, ranking first in the average protein complex prediction without stoichiometry information (TM-score of 0.752). Most models were AF2-based, with MULTICOM scoring 88.69% average median predictions (and other top models 88-90%). On the other hand, the ROSETTA model is also controversial in its de novo modeling approach through a focus on optimization. Since most top-scoring models retain AF2's main properties, and due to AF2's scale of breakthrough in CASP14, it is likely in the upcoming years for there to arise several significant competing models to the AF2 program. As AF2 has brought the computational biology field to an unprecedented stage leaving only about 10% accuracy of predictions to be corrected, AF2-based models are increasingly trendy in adapting the path AF2 has paved. In addition, AF2's breakthrough alongside the CASP contest have inspired many scientists to become intrigued in these directions, forming them into an improved new state-of-the-art focus. This result will help speedily advance these prediction programs as they have already been with several stemming approaches to the program. Different strategies discussed in this review may be utilized to construct a model evolving AF2 into a deeper and more comprehensive model.

While the trend of the computational biology field heads towards a final model, other paths to innovation like computer technology advancement in the traditional structure determination methods such as x-ray crystallography and cryo-EM become temporary fixes. With a significant lack of diverse protein structure data, the demand for generalization is growing. While an increased use of MSAs have overarchingly shown to lead to better predictions (especially in the transition from the introduction of MSAs in AlphaFold1 to the deep MSA usage in AF2), the bias must be accounted for, through a series of changes, whether it be to the architecture, algorithm-level mechanics, or the incorporation of other methods to increase the generalizability of the model." (see Page 17-18 of the updated manuscript)

Furthermore, the updated manuscript identifies that changes to the core architecture will be the main course of improvement for the movement towards a final model:

"Considering AF2's unintended result and power to generalize and the insurmountable variety of proteins possible, the former would be only a temporary solution to increase its capabilities, while the main focus remains in altering other central aspects of the program, particularly its core architecture.

…Today's technology paired with the improvements seen from AF2 and previous models, it is possible to achieve a generalized model without bias while retaining a high prediction accuracy all-around. This goal requires the creation of another breakthrough of a purely de novo model, by changing the core system. While the modification or addition of new hyperparameters will improve the model's effectiveness and usability within specific scientific contexts, the decisive change seems to be stemming from changing the transformers, which AF2 has shown by bringing in its Evoformer transformers, the pair and MSA transformer, as well as its structure module. As fewer templates were used by AF2 than AF1, the transformers should function without the need of templates in order to create a viable de novo prediction system.

It is plausible to go the direction to alter the transformer modules such that the model's attention mechanism can improve. Since the flow of information between AF2's transformer modules is largely responsible for its predictions, there can be an introduction of a more closely connected model with extra layers while being efficient, for better back-and-forth communication extrapolation of information, perhaps done by adding a third transformer addressing the bias brought by MSAs and slightly adjusting current transformers, to avoid the risk of removing successful aspects of AF2." (Page 19 of the updated manuscript)

There has also been added examples of such changes in "4.5 The Core Model" of the updated manuscript:

"For example, an algorithm found by Han and Lu in 2016 on an alternating back-propagation algorithm (88) demonstrates a possible transformer add-on feature that could introduce many benefits. The algorithm offers a framework to the generator network model, utilizing a convolutional neural network to map latent factors to observed data like images, video, and sound. Through its back-propagation, the program "alternates" between inferring latent factors by Langevin dynamics or gradient descent and updating parameters given the inferred factors. The program is particularly effective in its ability to generalize in the advent of incomplete training data, which is prevalent in the PDB. With proper implementation and re-adjustments to structure modeling, this change could allow for coarse predictions to be refined into higher resolution, cover missing data scenarios such as low confidence regions, and therefore increase the range of conformational variability accounted by the program's modeling.

While it is of importance to discover new transformers for the system, some, and if not many, features are to remain. One of which is the analysis of residue on the importance of the analysis of both local and non-local residues (53), where there has been success by adding the analysis of non-local residues in contrast to traditional uses of just the local residue in the model building process. This technique will serve to ensure more accuracy, traditionally coupled with physics and mathematical principles like the triangle inequality as thresholds for the purpose of demonstrating a swift move towards the correct positioning and angles of the predicted residues." (see Page 19-20 of the updated manuscript)

Additional suggested improvements to the AF2 model are included in "4.6 Other Improvements" of the updated manuscript:

"The program's data weights should be noted to test whether the model's average prediction accuracies should improve through increasing representation of low-representation protein families in the training data, altering weights and adding new layers and iterations that allows the program to focus on more unique proteins and therefore adapt the defect problems that the current biased model is experiencing, again alike the incorporations within MULTICOM's system. Thereby this could fix the certain regions to which are poorly represented by the model.

Other adjustments may include AlphaFold2's self-supervised learning from AlphaFoldDB, where the model improves its predictions based on the collection of previous predictions. MSA is undoubtedly needed in the final model as a prediction aspect, but there may be more to it. The success of MSAs comes from its ability to garner co-evolutionary information from the primary sequence, for it has been largely successful in its dominant popularization and usage among many models adapting this approach in the early 2010s. Thus, there is likely another method of analyzation for a different aspect of the protein sequence for which its intricacies are able to be exploited for co-evolutionary inferences, with a notable example being direct coupling analysis, as examined in a AF2 study conducted by Caredda and Pagnani in 2024 (89), which may soon find its success in the protein structure prediction world. It is notable to include in the de novo prediction system that parts of ROSETTA's optimization architecture and its use of gradient descent could also be inspired from in the making of the final model, as it has shown to perform well in free modeling settings (90)." (see Page 20 of the updated manuscript)

2. You mention that training data - obtained by empirical means such as NMR, cryo-EM etc. is important to continue to increase the 0.08% ratio of 3D predicted to PDB sequence data. My question is: is there a need for 'filler algorithms'? For example, if AF2 can predict with 99% Ground truth, the beta sheet and high sequence disorder regions of a fair number of proteins that fall into a certain category of the PDB database, then … for a protein in that category whose sequence is known, can we just perform an NMR or cry-EM on those regions to predict the structure and leave out other domains ?…. then feed this data into AF2 ? This way the predicted AF2 structure can be assumed to be 99% accurate with the actual structure (parts of which have still not been deciphered with NMR or cryoEM). Would this be a legitimate way to increase the production of training data without compromising ground truth accuracy?

Author's Response

The author is unclear about the reviewer's comment. This method has been suggested to be a legitimate method of data collection in sections 4.1 and 4.2 discussing the use of the hybrid approach. Nevertheless, a comment about this hybrid approach has been added in "4.2 Emerging Approaches to Improve Predictive Modeling", summarizing clearly the possibility of such utilization of AF2's bias and its poorly predicted regions through experimental processes: "Therefore, for the known low-confidence regions of proteins of interest, going forward, it is possible to utilize a hybrid approach of applying cryo-EM testing of those specific low-confidence regions (such as mutations and loop structures) and leaving the rest to be quickly yet accurately predicted by AF2. With all the hybrid transformations occurring in the biocomputational field, this process could be an efficient solution that utilizes both methods in their accuracies and efficiencies, allowing experimental methods to aid AF2 in its path to collecting more training data and subsequently obtaining 100% ground-truth prediction accuracy. Note, AF2 would be required to be highly trained and biased, carefully ensuring its ability to consistently predict the non low-confidence regions to the best of its ability—near ground-truth accuracies." (see Page 17 of the updated manuscript)

Thank you for addressing my comments. Accepted.

However, please address the following formatting comments IN THE ATTACHED DOC and attach to the discussion thread when completed to enable continuation of copyediting.

1.References must contain 6 authors followed by an et al. where the number of authors is more than 6. Include a live DOI link for all references. ALL references must fllow the same format.

2.I have added a paragraph before the conclusion section. Please ensure that you agree with the content.

3.There is a sentence missing in 3.2. Please add.

DO NOT CHANGE FORMATTING OF THE ATTACHED DOC WHEN MAKING THE CHANGES/CORRECTIONS.