Bioinformatics, 2025, pp. 1–9

doi: TBD
Original Article

ORIGINAL ARTICLE

AF3Complex Yields Improved Structural Predictions of Protein Complexes

Jonathan Feldman[1, 2] and Jeffrey Skolnick[2,*]

[1]Center for the Study of Systems Biology/School of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Drive, 30332, Georgia and [2]School of Computer Science, Georgia Institute of Technology, 266 Ferst Dr, Atlanta, 30332, Georgia
[*]Corresponding author: skolnick@gatech.edu

Abstract

Motivation: Accurate structures of protein complexes are essential for understanding biological pathway function. A previous study showed how downstream modifications to AlphaFold 2 could yield AF2Complex, a model better suited for protein complexes. Here, we introduce AF3Complex, a model equipped with the same improvements as AF2Complex, along with a novel method for excluding ligands, built on AlphaFold 3.
Results: Benchmarking AF3Complex and AlphaFold 3 on a large dataset of protein complexes, it was shown that AF3Complex outperforms AlphaFold 3 to a significant degree. Moreover, by evaluating the structures generated by AF3Complex on a dataset of protein-peptide complexes and antibody-antigen complexes, it was established that AF3Complex could create high-fidelity structures for these challenging complex types. Additionally, when deployed to generate structural predictions for the two antibody-antigen and seven protein-protein complexes used in the recent CASP16 competition, AF3Complex yielded structures that would have placed it among the top models in the competition.
Availability: The AF3Complex code is freely available at https://github.com/Jfeldman34/AF3Complex.git.
Contact: Please contact skolnick@gatech.edu.

Introduction

Underpinned by a near-complete corpus of structures for single-domain proteins along with the plethora of data in protein sequence databases, AlphaFold 2, a highly accurate deep learning-based approach to predicting protein structures from amino acid sequences unprecedented in its ability to model single and multi-domain proteins with high fidelity, reinvigorated and led to the rapid progress of protein-folding research [1, 2, 3]. Of the myriad of research studies launched after the advent of AlphaFold 2, many aimed to improve upon the underlying processes and architecture of the model to enhance the accuracy of its structural predictions–advances that would prove to be immensely impactful in the model's clinical applications where precision is key [4, 5, 6]. Among those studies, one, in particular, developed a derivative of AlphaFold 2, called AF2Complex, which significantly improved upon the former model's ability to predict the structure of protein-protein complexes, which are highly relevant biologically and clinically [4]. When released, AF2Complex provided state-of-the-art performance for protein-protein complex modeling, outperforming even AlphaFold-Multimer, a model specialized for that task [4, 7].

In May 2024, the newest version of the AlphaFold suite of models, AlphaFold 3, was released. The architecture of the AlphaFold 3 model deviated from the previous models of its kind tremendously. It employed a diffusion-based architecture that imbued it with the abilities of a generative artificial intelligence model and minimized the importance of the multiple sequence alignments (MSAs) within the model; relegating it to a peripheral role in the inference pipeline [8]. This new architecture has granted AlphaFold 3 several impressive new capabilities. AlphaFold 3, unlike AlphaFold 2 or AlphaFold-Multimer, can predict the structures of proteins with bound ligands and ions, including them in the final structural prediction, and those of protein-nucleic acid complexes, which are complexes formed by one or more proteins interacting with one or more molecules of DNA or RNA [2, 8]. Additionally, AlphaFold 3 outperforms the previous generation of AlphaFold models on all modeling tasks, including protein-protein complex structure prediction [8].

It remained unclear, however, as to how well AlphaFold 3 would perform when compared to specialized models from outside of its model suite, such as the aforementioned AF2Complex model. This question is especially vital to protein complex modeling, where minute differences in the structure, especially at the interface between protein chains, may lead to drastically different behaviors.

Thus, preliminarily, AF2Complex and AlphaFold 3 were compared on a dataset of 972 protein complexes. This analysis showed that AlphaFold 3 was the clear victor in terms of accuracy, beating AF2Complex in both macroscopic structural accuracy, measured using the TM-Score [9] and lDDT [10] score, and in interfacial structural accuracy, evaluated using the DockQ score [11].

Since AlphaFold 3 was superior to AF2Complex and could model protein complexes with additional ligand or ion information, it seemed appropriate to test whether a derivative model could be developed from AlphaFold 3, as AF2Complex was built from AlphaFold 2. This model, AF3Complex, would apply the improvements of AF2Complex to the AlphaFold 3 model, taking advantage of the newer model's advances in the accuracy of structural predictions and the scope of macromolecules it can model [8].

In this study, we examined to what extent the retrofitting of AlphaFold 3 with a modified MSA generation algorithm, one which wholly omits the paired MSA, and a specialized confidence score that examines the interfacial residues in a protein complex more accurately could improve the ability of the model to predict the structure of protein complexes. Those modifications were the very same ones made to AlphaFold 2 to develop AF2Complex [4]. However, this study also explores a third change: generating two sets of structural models, one with ligands or ions, if such information is included in the model input, and one without ligands or ions, to examine which model the deep learning framework is more confident in. Together, these three downstream modifications, when applied to the AlphaFold 3 model, yielded the AF3Complex model, which was capable of outperforming the former across several metrics of total and interfacial structural accuracy and modeling challenging protein complex structures. These included protein-peptide interactions and antibody-antigen interactions, which previous models struggled to properly predict [12].

Methods

The design of the AF3Complex model pipeline is shown in Figure 1. Given input protein or nucleic acid sequences with optional ligand and ion information, the AF3Complex pipeline uses a modified MSA generation algorithm that collates only unpaired MSAs, excluding the paired MSA, which has been part of the data pipeline for all AlphaFold models [2, 8, 7]. The removal of the MSA pairing algorithm is more expedient as it omits one of the most computationally costly procedures in the pipeline and allows the model to avoid the pitfalls of generating orthologous sequences across different species–a process that can be confounded by the presence of genetic paralogs, cross-talk between protein signaling pathways, and the assimilation of pathogenic genetic material, such as that from lysogenic viruses [4, 13]. Moreover, since paralogs are often precursors to the diversification of protein function through evolutionary processes, excluding the paired MSAs imbues the model with greater flexibility, as it extricates stringent and possibly inaccurate paired structures [4]. Once the MSAs and the structural templates are generated, they are passed to the AlphaFold 3 diffusion model along with the protein and nucleic acid sequences and ligand or ion information. The diffusion model then outputs the possible protein complex structures, from which the best is chosen using the predicted Interface-Similarity score (pIS), a score that, like the pTM or ipTM metric utilized by the AlphaFold 3 model, ranges from 0 to 1, inclusive, and represents higher fidelity models at greater values [8]. Lastly, if the data input into the pipeline included ligands or ions, the model inputs the unpaired MSAs, structural templates, and protein or nucleic acid sequences anew, with the ligand or ion information omitted, thereby generating a structural prediction with those molecules excluded. The outputs–with ligands or ions and without ligands or ions–are compared, and the model with the highest pIS is chosen.

AF3Complex Pipeline Overview

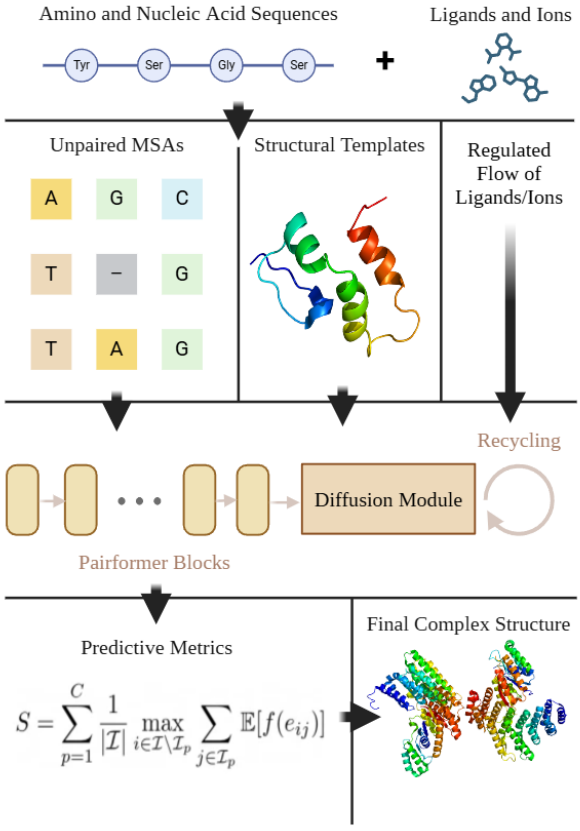

Fig. 1. A schematic overview of the AF3Complex model pipeline, including the modified MSA generation algorithm, the AlphaFold 3 diffusion model backbone, the modified scoring metric, pIS, denoted S in the graph, and the regulated flow of ligand and ion information based on model confidence.

Protein Complex Dataset Curation

The dataset of protein complexes upon which AlphaFold 3 and AF3Complex were evaluated was constructed using the Protein Data Bank (PDB) [14]. Initially, following the framework outline in the AlphaFold-Multimer paper, proteins in the PDB were filtered based on the following criteria: Filtered proteins must have less than 3,000 deposited residues, have been released after September 30th, 2021, have at least two substituent protein chains, have at most nine substituent protein chains, not be monomeric, and have an experimentally determined template structure [7]. These filters ensured that the protein complexes were within the processing capabilities and specifications of AlphaFold 3 and AF3Complex. They also ensured that the proteins were not included within the AlphaFold 3 training datasets, for which the protein-inclusion cutoff date was September 30th, 2021 [8].

This initial filtering yielded 20,839 unique complexes. Next, using the *MMseqs2* package, the dataset of 20,839 protein complexes was grouped into clusters based on a minimum

sequence identity of 40%, as in the AlphaFold-Multimer paper, which ensures that the dataset is not saturated with similar protein motifs, thereby making each protein the models are evaluated upon sufficiently distinctive [7, 15]. From these clusters, a single representative was randomly chosen, yielding 7,198 dissimilar protein complexes.

Lastly, to avoid trivializing the test by including protein structures that any of the models may have seen before, a similarity search between all 7,198 dissimilar protein complexes and all proteins in the PDB was undertaken. If any dissimilar protein complex shared at least a 40% sequence similarity with a given protein in the PDB released on or before the cutoff date of September 30th, 2021, and could thus be plausibly in the AlphaFold 3 training dataset, it was removed. This last filtering step yielded a final dataset of 972 protein complexes dissimilar to those on which AlphaFold 3 was trained and properly vetted for use in the forthcoming analysis.

Human Peptide Dataset Curation

To gauge the capacity of AF3Complex to model the interactions between proteins and peptides, which are often challenging for such models due to their diminutive size, a dataset of 91 human protein-peptide complexes was assembled from the PDB. Like the previous protein complex dataset, all of these complexes fell within the parameters of the guidelines outlined in the AlphaFold-Multimer paper and were released after the AlphaFold 3 training dataset cutoff date of September 30th, 2021 [7, 8].

Antibody-Antigen Dataset Curation

To assess AF3Complex's capability to model the interactions between antibodies and antigens, a dataset from a previous study examining how to improve antibody-antigen modeling with AF2Complex of 36 different antibodies interacting with the SARS-CoV-2 spiked RBD was obtained [12]. This dataset, like the previous datasets, followed the methodology from the AlphaFold-Multimer paper and did not include any complexes that were released on or before September 30th, 2021 [12].

CASP16 Released Protein Complex Dataset

With the exclusion of two protein complexes that were too large for AF3Complex to process on the computing resources available and the two antigen-antibody protein complexes that were analyzed separately, the dataset of seven protein-protein complexes assembled from CASP16 were chosen as they were the only proteins with publicly available experimental structures to enable model assessment. The proteins in this dataset were T1201, H1202, H1204, T1206, T1234, T1235, and H1236.

When AF3Complex was employed to predict the structures of these protein complexes, it only had the information available to the participants in CASP16: the sequence and the stoichiometry of the protein complex. No other information was given to the model. Moreover, for the purpose of ranking the structures generated by AF3Complex, the phase of competition with the models achieving the highest scores was used for comparison.

AF3Complex Predictive Metrics

A previous study introduced the Interface Similarity Score (IS-score) metric to more accurately and sensitively measure the structural similarity between protein complex interfaces [16]. This score was developed to ameliorate the shortcomings associated with the TM-Score, which did not properly focus on the structural similarity of interfacial residues [9, 16]. In that study, a modified version of the IS-score was employed to function as a predictive metric within the model, thereby facilitating the model's ability to choose the best structure from those generated.

We define the pIS score by first introducing an intermediate metric, the predicted interface TM-Score (piTM), which is defined as follows:

$$piTM = \max_{i \in \mathscr{I}} \frac{1}{I} \sum_{j \in \mathscr{I}} \frac{1}{1 + \left(\frac{\langle e_{ij} \rangle}{d_0(I)}\right)^2} \tag{1}$$

where $\mathscr{I}$ is the set of interfacial residues within the predicted protein complex structure, and the cardinality of $\mathscr{I}$ is the total number of residues [4]. Using the AlphaFold 3 confidence head, we can arrive at an estimate for the distance $\langle e_{ij} \rangle$ between the interfacial residue and its assumed position in the experimental structure [8]. The piTM score optimizes the position of the proteins in the complex, and $d_0(I)$ is the normalization factor that is defined as:

$$d_0(I) = \begin{cases} 1.24\sqrt[3]{I - 15} - 1.8, & \text{if } I \geq 22 \\ 0.02I, & \text{if } I < 22 \end{cases} \tag{2}$$

We derive the pIS from the piTM as follows:

$$pIS = \sum_{p=1}^{C} \frac{1}{I} \max_{i \in \mathscr{I} \setminus \mathscr{I}_p} \sum_{j \in \mathscr{I}_p} \frac{1}{1 + \left(\frac{\langle e_{ij} \rangle}{d_0(I)}\right)^2} \tag{3}$$

where we calculate the piTM for each protein chain p within the complex independently, adding together the values. In each chain, p, there are a certain number of interfacial residues, $\mathscr{I}_p$, and $\mathscr{I}$ represents the union of $\mathscr{I}_p$.

The major difference between the pIS metric and the ipTM metric, the default metric for protein complexes in AlphaFold 3, is that the former focuses solely on interfacial residues rather than entire chains, which is preferable when trying to improve protein complex structure prediction [4, 7, 8].

Note that in the AF3Complex pipeline, the pIS metric is used to rank the structures generated to choose the best one, but this ranking metric is augmented by a clash penalty to ensure that the model never outputs structures having overlapping chains or residues.

Model Testing Procedure

For each protein in the protein complex dataset, AlphaFold 3 and AF3Complex were provided with the same model seed, thereby keeping the internal workings of both models constant and allowing variation to stem only from the modifications made to the MSA pairing algorithm, the prediction metric calculation, and the ligand and ion data processing [8]. Both models were also given the same number of recycles, ten, and the same number of structural models to generate, five.

Similarly, when the AlphaFold 3 and AF3Complex models were fed data with ligands and ions purposefully excluded to examine to what degree such data impacts the accuracy of either model, the same MSAs and templates were employed as used in the inference pipeline for the models with ligand and ion data. This is permissible since ligand information is separate from the MSAs and template structures for a given protein [8].

Evaluative Metrics

All protein evaluation metrics were calculated using the open-source computational structural biology framework *Open Structure* [17]. All metrics, except for DockQ, were calculated by *Open Structure* precisely as in the papers or package where they were introduced.

In this study, to calculate the DockQ metric for protein complexes, we employed version two of the DockQ package, which maintains scoring compatibility with previous versions but provides greater computational efficiency and more features than the first version [11].

Statistical Tests

To evaluate the validity of the hypothesis that the model quality is improved using a certain scoring metric–the DockQ score, for example–we utilized a Wilcoxon signed-rank test, a non-parametric statistical test for distributions that do not follow a normal distribution, as is the case with the protein complexes [4]. Each of the tests was paired, as all the models make predictions on the dataset of protein complexes, and each of the tests was one-tailed.

Results

Model Comparison on Protein Complex Dataset

We first tested the performance of the AlphaFold 3 and AF3Complex models on the 972 multimeric proteins within the novel protein complex dataset assembled in this study. For each protein complex, the models were given the same input–the amino or nucleic acid sequence and the ligand or ion information, if there was any–and were tasked with producing the best structural prediction possible.

Overall, the interfacial structural predictions of the AF3Complex model outperformed those of the AlphaFold 3 model, with a mean and median DockQ of 0.558 and 0.684 for AF3Complex and 0.542 and 0.665 for AlphaFold 3. A one-tailed Wilcoxon's signed rank test on the DockQ distributions yielded a P-value of 1.6×10^{-3}, indicating a statistically significant difference in the performance of the two models in terms of interfacial structural accuracy. The relative performance of the AF3Complex model relative to AlphaFold 3 based on DockQ can be viewed in Figure 2A.

Likewise, AF3Complex outperformed AlphaFold 3 in terms of overall structural accuracy, albeit to a lesser degree, with a mean and median TM-Score of 0.789 and 0.935 for AF3Complex and 0.782 and 0.926 for AlphaFold 3, and a P-value of 1.3×10^{-2} from an identical one-tailed Wilcoxon's signed rank test as the one above.

Ablation Study With Ligand and Ion Data

It is important to note that both models utilized information about the number and types of ligands or ions in the protein complex they were tasked with modeling. In reality, one often does not know in advance the quantities and varieties of ligands or ions bound to a protein complex, thus making the gauging of the model without this additional data valuable to understand their performance in practical application [5]. Hence, both models were reevaluated on the 972-constituent protein complex dataset but, this time, with all ligand and ion data excluded from inputs.

The results of this evaluation showed the same trend as the previous: AF3Complex outperformed AlphaFold 3 in terms of interfacial structural accuracy, achieving a mean DockQ of 0.546 and a median of 0.664, while the latter model scored a mean DockQ of 0.534 and a median of 0.649. The distributions of scores achieved a P-value of 8.9×10^{-4} according to the same one-tailed Wilcoxon's signed rank test as conducted earlier in the study, demonstrating that AF3Complex outperforms AlphaFold 3 even without the added benefit of ligand information. The relative performance of each model according to DockQ in this analysis can be viewed in Figure 2B.

Impressively, even without any ligand or ion information, AF3Complex still manages to perform equivalently to the AlphaFold 3 model with that information available to it. No statistically significant difference between the two models' outputs for the DockQ and TM-Score metrics was observed. Thus, even at a significant informational disadvantage, AF3Complex can rely upon its modified MSA pairing algorithm and improved predictive metric to provide an effective improvement. The relative performance of each model in this analysis can be viewed in Figure 2C.

The AF3Complex model with ligand and ion information does, however, eclipse both these models in performance, with a P-value of 2.2×10^{-2} when compared to its counterpart with ligand and ion data excluded, emphasizing the improvements yielded by the algorithm tasked with discerning whether ligand and ion information degrades the integrity of the predicted structural model. The relative performance of the AF3Complex model with ligand and ion data compared to that of itself without that data can be viewed in Figure 2D.

Analyzing Purposeful Ligand and Ion Data Omission in AF3Complex

Having established that the inclusion of ligand and ion data improves the performance of the AF3Complex model, a natural question arises: why does AF3Complex purposefully exclude this critical data from some of its protein complex inputs? The initial intuition that spurred the addition of this feature was that in protein complexes, ligand and ion data may provide unnecessary information that may confound the model, reducing the quality of its structures as it not only has to contend with protein and nucleic acid structure modeling but also proper ligand docking.

To evaluate the validity of this assumption, we examined the 98 protein complexes from the 972-constituent protein complex dataset that had ligand and ion data purposefully excluded by the model. It was found that of the 98 protein complex structures generated by AF3Complex without ligand or ion data, 61 of them were superior to their counterparts with ligand data, meaning that the model correctly chooses which model to retain almost two-thirds of the time. Moreover, if one sets a baseline threshold of 0.05 DockQ points for the absolute difference between the two types of models–with ligands and ions and without–then of the 21 generated structures that meet this criterion, 18 of them were better without the ligand or ion information, compared to only three structures that improved with that information. From this data, the adversarial approach used by the AF3Complex model to determine the necessity of ligand and ion information seems justified.

It is interesting to note that of the 98 protein complexes for which ligand and ion data were excluded, most were smaller protein complexes with only two chains and fewer than 1000 residues. We speculate that the reason for this may be that for smaller structural models, the structural effects of the ligands

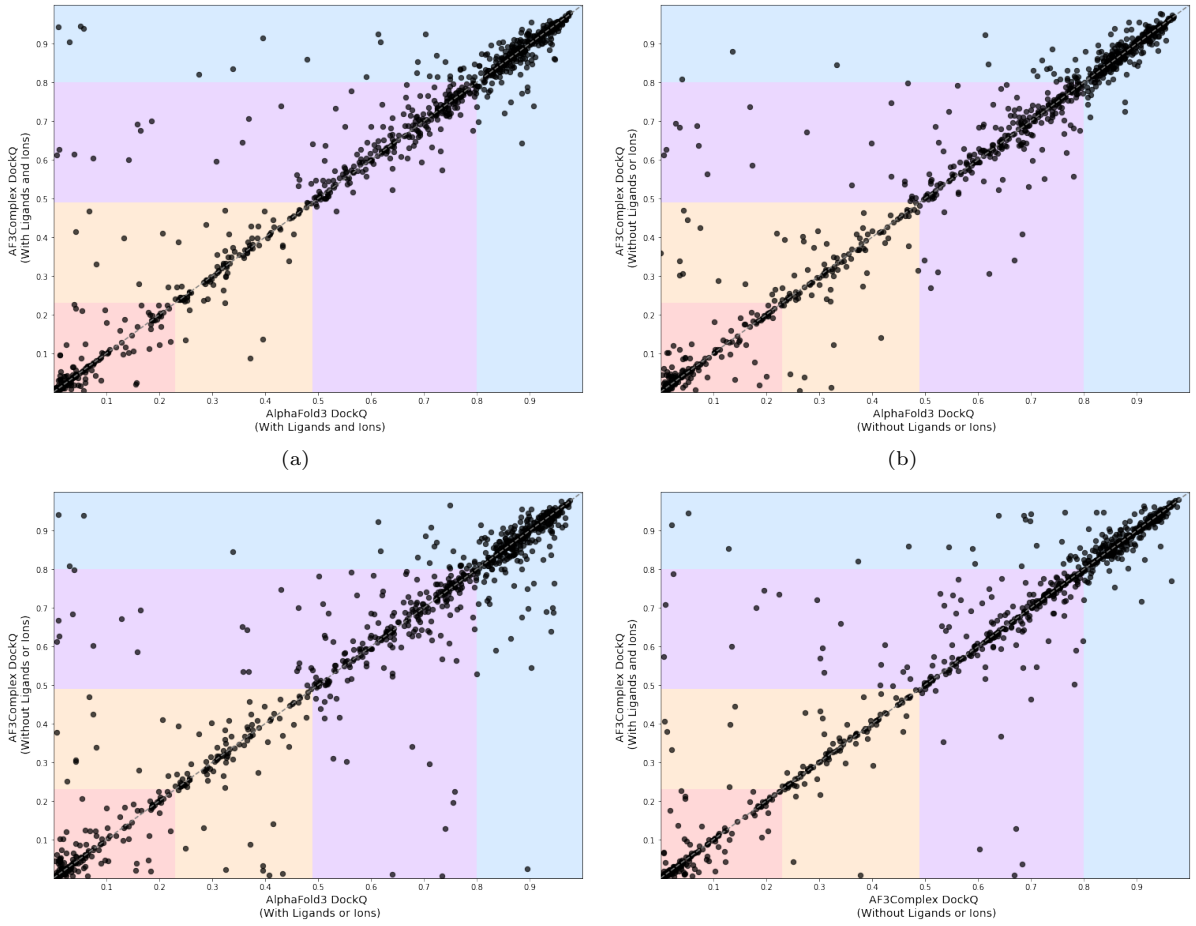

Fig. 2. The figures show the relative DockQ scores of variations of AF3Complex and AlphaFold 3, where each dot is a single protein complex and the x-axis is the AlphaFold 3 model and the y-axis is the AF3Complex model. Each colored region represents the DockQ score distributions: Red [0.00, 0.23) is the incorrect region, Orange [0.23, 0.49) is acceptable accuracy, Purple [0.49, 0.80) is medium accuracy, and Blue [0.80, 1.00] is high accuracy. Figure (a) shows the performance when both models use ligand and ion data. Figure (b) shows the performance without this data. Figure (c) shows the performance when AF3Complex has no ligand data, but AlphaFold 3 does. Figure (d) shows the performance when one AF3Complex model has that information and one does not.

and ions are less pronounced, making their inclusion only an additional complexity the AF3Complex model must overcome. Larger protein complexes, however, because of their increased size, truly benefit from the aid of ligands, as they reduce the propagation of small errors over their greater surface area and improve the conformation of interfacial regions.

Examination of the pIS Score

In addition to understanding how the purposeful exclusion of ligand and ion data affects the accuracy of the AF3Complex model, we also wanted to examine to what degree the pIS scoring metric improves the accuracy of the structural models. Though this scoring metric was shown previously to improve the outputs of AF2Complex, since the AlphaFold 3 model from which the predicted aligned error matrices are derived is different architecturally and in terms of training, it is necessary to examine the impacts of this metric specifically for the AF3Complex model, which is reliant upon AlphaFold 3 [8, 4].

For each protein complex in the 972-constituent protein complex dataset, we examined whether the best model chosen by the two metrics used in AlphaFold 3, ipTM and pTM, and the metric employed by AF3Complex, pIS, correctly reflected the best model based on the DockQ score. Of the 972 protein complexes, we found that the pIS had the best performance, choosing the best model 239 times, while the ipTM came in second, choosing the best model 227 times, and the pTM came in last, choosing the best model only 221 times.

The improvement yielded by the modified predictive metric is fundamentally limited by the training of the model upon which it operates. Though it better reflects the protein complex structures it aims to predict, it is still wholly reliant on the outputs of the AlphaFold 3 confidence head, which cannot be modified without changing the model's weights.

AF3Complex Evaluation on Human Peptide Dataset

A challenging yet vital form of protein complex prediction is the situation when a small sequence of amino acids, a peptide, interacts with one or more larger proteins.

AF3Complex was employed to predict the structure of 91 different human protein-peptide complexes to assess its ability to model this non-standard protein complex. The distribution of the DockQ values for these 91 structural models can be seen

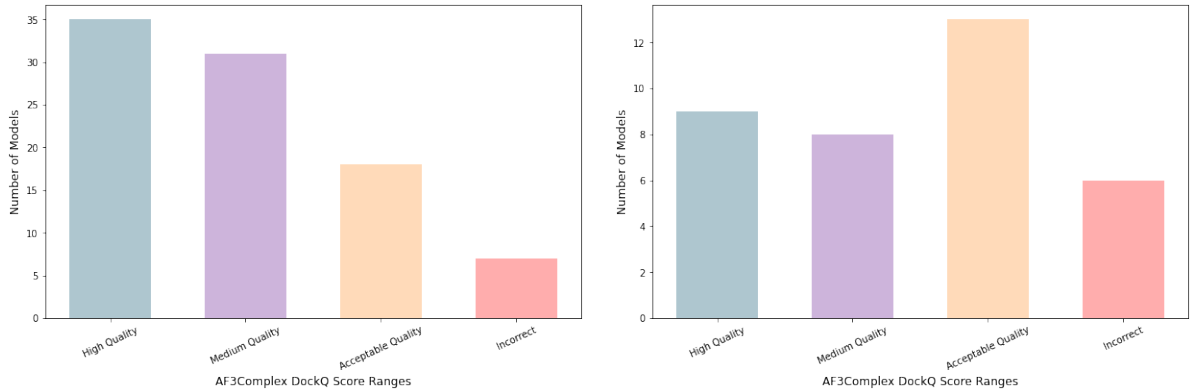

Fig. 3. The figures display the number of AF3Complex generated structures that fall into the four aforementioned DockQ score regions for the 91-constituent protein-peptide complex dataset and the 36-constituent antibody-antigen dataset. Figure (a) displays the quantities for the protein-peptide dataset, while figure (b) displays the quantities for the antibody-antigen dataset.

in Figure 3A. When compared to the protein complex templates for the dataset, the structural predictions generated by the AF3Complex proved to be highly accurate, with a median DockQ score of 0.714 and a median TM-score of 0.955 across all models, indicating that the models were similar to their native structure in terms of macroscopic and interfacial structures. Moreover, of the 91 structures generated, only 7 of the models were incorrect according to the DockQ value. More than two-thirds of the models were of medium quality or higher, with 35 models of high quality and 31 models of medium quality [11].

Overall, the AF3Complex's performance on the protein-peptide complex dataset shows a prodigious understanding of the structures, both globally and in their interface of such complexes.

AF3Complex Evaluation on Antibody-Antigen Dataset

The prediction of antibody-antigen structure is another challenging protein complex type for AlphaFold-like models. In a previous study, AF2Complex was retrofitted with modifications to better its ability to generate antibody-antigen structures. These modifications proved successful and produced many high-fidelity results on a dataset of 36 different antibodies interacting with the SARS-CoV-2 Spike RBD protein [12]. Using AF3Complex, we generated new structural models for all 36 of these complexes. The distribution of the DockQ values for the 36 structural models generated by AF3Complex can be seen in Figure 3B.

The structures generated by AF3Complex, which, unlike the modified AF2Complex model used in the aforementioned study, is a general model rather than a model meant to specialize in antibody-antigen complex modeling, proved to be accurate when compared to their respective template structures, scoring a mean and median DockQ of 0.519 and 0.420 respectively. Though the DockQ scores are not as high as those achieved by the model for the previous cases of general protein complexes and protein-peptide complexes, they represent a marked improvement upon the specialized AF2Complex model's performance on the very same dataset, as the model achieved a mean and median DockQ score of 0.366 and 0.2625, respectively, which denotes an outstanding average improvement of 0.153 points [12].

AF3Complex Evaluation on CASP16 Antibody-Antigen Complexes

In the recently concluded CASP16 competition, there were two antibody-antigen protein complex targets: H1232 and H1233. AF3Complex, provided with the same amino acid sequence information given to the other contestants in the competition, was tasked with predicting the structure of these two complexes. The structural predictions generated by AF3Complex on these two complexes can be seen in Figure 4.

For H1232, AF3Complex yielded a protein structure with a DockQ score, using the weighted average of all the chains, as in the CASP16 competition calculations, of 0.527 and a TM-Score of 0.592. These scores place it second amongst the submissions generated without human intervention for this protein according to both DockQ and TM-Score.

For the H1233 antibody-antigen complex, AF3Complex generated a protein structure with a weighted average DockQ score of 0.886 and a TM-Score of 0.819. This DockQ score places it second amongst the models generated without human intervention, trailing behind the best prediction by only a fractional value of 0.004, according to the DockQ metric.

Altogether, the accuracy of AF3Complex for these two CASP16 protein complexes is excellent, exceeding tens of other models in the competition with DockQ and TM-Scores that would solidify its place in the top ranks for both protein complexes, though it is only a general model with no special modifications to improve its performance on antibody-antigen complexes particularly.

AF3Complex Evaluation on Other CASP16 Protein Complexes

Following AF3Complex's impressive performance when modeling the two aforesaid antibody-antigen complexes, it seemed appropriate to examine how AF3Complex would perform relative to the other competitors in CASP16 on the other multimeric proteins. Hence, AF3Complex was used to model seven protein complexes from the CASP16 competition with publicly available template structures.

Overall, AF3Complex performed very well when predicting the structure of these protein complexes, generating two protein structures that proved better than any structures generated in the competition without human intervention according to

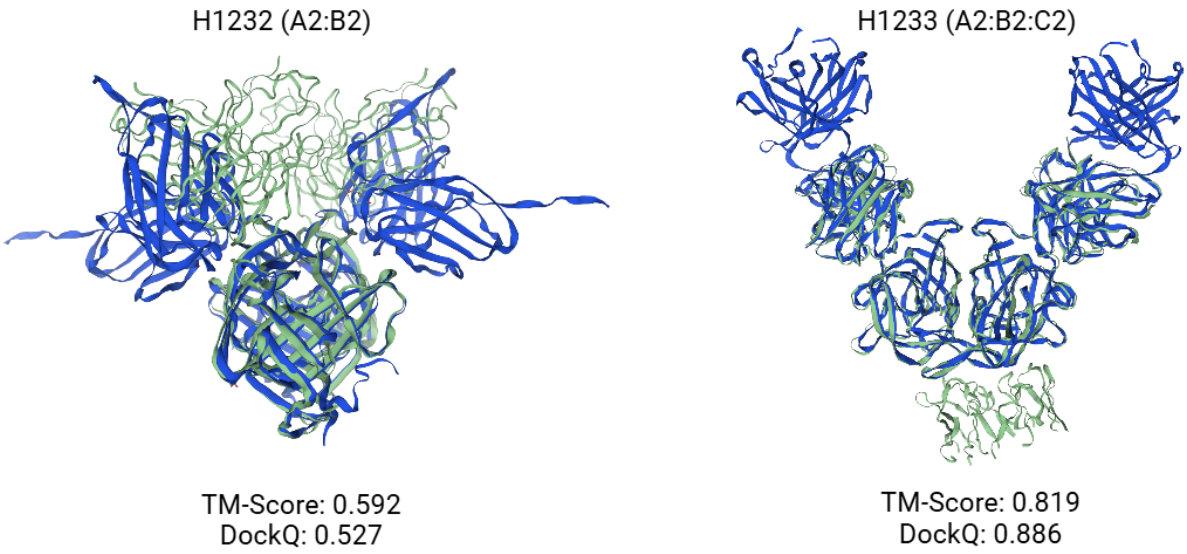

Fig. 4. The structural predictions generated by AF3Complex on the two antibody-antigen complexes from CASP16 overlayed on their native structures with the associated TM-Score and DockQ metric results. The models generated by AF3Complex are colored blue while the native models are green.

the weighted average of the DockQ metric. AF3Complex's structures were consistently ranked within the top eight structures generated without human intervention by weighted average DockQ, except for one prediction on the T1235 protein complex, a highly challenging protein target, that was ranked eleventh.

When averaged, the individual rankings by weighted average DockQ for each of the protein complex structures generated by AF3Complex, including those of the antibody-antigen complexes, yield a mean ranking of approximately 4.2. Comparing this to the average ranking of the best model for multimers in the competition not reliant on human intervention, Yang-Multimer [18], which had an approximate average ranking of 6.9 on these protein complexes, one can see that AF3Complex provides a significant increase in accuracy relative to the models used in CASP16.

Discussion

The findings of this study clearly demonstrate that the downstream modification of AlphaFold 3 can lead to significant improvement in the model's protein complex modeling capabilities. Two of the three modifications, the altered MSA generation algorithm and the use of the pIS score, were re-implementations of the modifications previously made to AlphaFold 2 to create AF2Complex [4]. The final modification, however, was novel to this study and took advantage of the AlphaFold 3 model's unique ability to model ligands and ions by providing AF3Complex with two different inputs: one with the ligand and ion information and one without [8]. This approach grants the model the freedom to determine, with the help of the pIS confidence metric, whether the additional ligand or ion information improved the structural integrity of the model or detracted from it, thereby minimizing irregularities in the interfacial structures.

These three modifications, as shown through rigorous benchmarking of AF3Complex and AlphaFold 3 on different types of protein complexes, demonstrated that AF3Complex is at the state-of-the-art. This improvement occurs throughout the structure of the entire protein complex but is especially pronounced in the interface between protein chains. [4].

All three metrics individually contribute to the increase in performance we see in AF3Complex relative to AlphaFold 3. The modified MSA generation algorithm, as a previous study employing AF2Complex showed, omits paired MSAs, which allows the model more freedom to explore the structural and energetic relationships between amino acids in a protein structure and avoids many of the pitfalls of examining evolutionarily orthologous sequences. Thus, it improves the structures generated by the models [4]. It is important to note, however, that the AlphaFold 3 model architecture, which underpins AF3Complex, employs the information provided by MSAs to a much smaller degree than AlphaFold 2, on which AF2Complex is based [4, 8]. This architectural change reduced the impact of excluding paired MSAs, though it did not wholly do so.

Similarly, the revised predictive metric used in AF3Complex, pIS, was shown to improve the accuracy of structural predictions in AF2Complex relative to AlphaFold-Multimer [4]. A similar result holds on comparing AlphaFold 3 with AF3Complex by comparison of the two confidence metrics employed by AlphaFold 3, ipTM and pTM, as well as our metric, pIS. This analysis showed that the pIS score more frequently chose the best structures from among those generated by AF3Complex. Ultimately, the pIS metric is constrained by the predicted aligned error matrices output by the internal AlphaFold 3 model's confidence head, which is dictated by the model's training, but, despite this, measurably improves the accuracy of the model's predictions [4, 8].

Lastly, through an ablation study, we saw that the inclusion of ligand and ion data significantly improved the accuracy of

the AF3Complex model, allowing it to outperform AlphaFold 3. Additionally, even without the ligand and ion data, the other modifications made to AF3Complex allow it to perform just as well as AlphaFold 3 with that data given to it. Moreover, we found that when AF3Complex chooses to omit ligand and ion data from the model inputs based on its predictive metrics, it overwhelmingly yields a better structure for the protein complex in question than an AF3Complex model that always employs ligand and ion data. This conscious omission of ligand and ion data was largely limited to smaller proteins of fewer than a thousand residues, for which we speculate the inclusion of ligand and ion data does less to stabilize the overall structure as compared to larger multimeric proteins—where small errors can propagate over a large surface area—and may even go as far as to confound the model by introducing the superfluous task of optimally placing the ligand.

Furthermore, we found that the excellent performance of AF3Complex was not limited to standard multimeric protein complexes. On a dataset of 91 protein-peptide complexes, AF3Complex yielded structures with impressively high interfacial and global structural accuracy. On a dataset of 36 antibodies interacting with the SARS-CoV-2 Spike Protein RBD from a previous study that employed a modified AF2Complex model adapted for antibody-antigen modeling, AF3Complex generally outperformed the AF2Complex-based methodology, despite being a general model suited to all types of protein complexes [12].

Ultimately, AF3Complex is a derivative of AlphaFold 3 that builds upon that model's ability to model protein complexes with nucleic acids, ligands, and ions. Often, but not always, it can provide highly accurate protein complex structures that exceed those of the model from which it was derived. This tool should have widespread applications in both research and clinical science.

References

1. Jeffrey Skolnick, Mu Gao, Hongyi Zhou, and Suresh Singh. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *Journal of Chemical Information and Modeling*, 61(10):4827–4831, October 2021.
2. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.
3. Zhenyu Yang, Xiaoxi Zeng, Yi Zhao, and Runsheng Chen. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduction and Targeted Therapy*, 8(1):1–14, March 2023. Publisher: Nature Publishing Group.
4. Mu Gao, Davi Nakajima An, Jerry M. Parks, and Jeffrey Skolnick. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nature Communications*, 13(1):1744, April 2022.
5. Maarten L. Hekkelman, Ida de Vries, Robbie P. Joosten, and Anastassis Perrakis. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, February 2023. Publisher: Nature Publishing Group.
6. Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, August 2021.
7. Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with AlphaFold-Multimer, March 2022. Pages: 2021.10.04.463034 Section: New Results.
8. Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024.
9. Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.
10. Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, November 2013.
11. Claudio Mirabello and Björn Wallner. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586, October 2024.
12. Mu Gao and Jeffrey Skolnick. Improved deep learning prediction of antigen-antibody interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 121(41):e2410529121, October 2024.
13. Mu Gao, Davi Nakajima An, and Jeffrey Skolnick. Deep learning-driven insights into super protein complexes for outer membrane protein biogenesis in bacteria. *eLife*,

11:e82885, December 2022. Publisher: eLife Sciences Publications, Ltd.

14. Stephen K Burley, Helen M Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Ken Dalenberg, Jose M Duarte, Shuchismita Dutta, Zukang Feng, Sutapa Ghosh, David S Goodsell, Rachel K Green, Vladimir Guranović, Dmytro Guzenko, Brian P Hudson, Tara Kalro, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Periskova, Andreas Prlić, Chris Randle, Alexander Rose, Peter Rose, Raul Sala, Monica Sekharan, Chenghua Shao, Lihua Tan, Yi-Ping Tao, Yana Valasatava, Maria Voigt, John Westbrook, Jesse Woo, Huanwang Yang, Jasmine Young, Marina Zhuravleva, and Christine Zardecki. Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, 47(D1):D464–D474, 10 2018.
15. Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017.
16. Mu Gao and Jeffrey Skolnick. New benchmark metrics for protein-protein docking methods. *Proteins: Structure, Function, and Bioinformatics*, 79(5):1623–1634, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22987.
17. M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):701–709, May 2013. Publisher: International Union of Crystallography.
18. Zhenling Peng, Wenkai Wang, Hong Wei, Xiaoge Li, and Jianyi Yang. Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. *Proteins*, 91(12):1704–1711, December 2023.

Competing interests

No competing interest is declared.

Author contributions statement

J.F. and J.S. designed the research, J.F. wrote the source code, J.F. performed research and analyzed the data, J.F. prepared the first draft of the manuscript, J.F. and J.S. revised the manuscript, and all authors proofread the manuscript.

Acknowledgments

This research was supported in part by a grant GM 118039 from the Division of General Medical Sciences of the National Institutes of Health. A gift from the Ovarian Cancer Institute is gratefully acknowledged. The authors thank Google DeepMind for their work on AlphaFold 3, without which AF3Complex would not have been developed. The authors would also like to thank Dr. Mu Gao, Davi Nakajima An, and Dr. Jerry M. Parks for their work developing AF2Complex, from which inspiration for AF3Complex was derived.