Bioinformatics, 2025, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Evaluating Deep Learning Based Structure Prediction Methods on Antibody-Antigen Complexes.

Samuel Fromm,[1] Marko Ludaic[1] and Arne Elofsson[1,*]

[1]Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, Solna Postcode, Sweden
[*]Corresponding author. Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, Box 1031, Solna Postcode, Sweden. arne@bioinfo.se
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation AlphaFold2 significantly improved the prediction of protein complex structures. However, its accuracy is lower for interactions without co-evolutionary signals, such as host-pathogen and antibody-antigen interactions. Two strategies have been developed to address this limitation: massive sampling and replacing the evoformer with the pairformer, which does not rely on co-evolution, as introduced in AlphaFold3.
Results In this study we benchmark structure prediction methods on unseen antibody-antigen complexes reveals that: (i) increased sampling improves the chances of generating correct models in a roughly log-linear way, (ii) for all methods, a significant challenge is identifying the correct models among those generated, (iii) AlphaFold3 outperforms AlphaFold2, Boltz-1 and Chai-1, and (iv) AlphaFold3's performance declines significantly for complexes lacking structural similarities to the training set, indicating that it has primarily learnt to detect remote structural similarities rather than to learn the physics of interactions.
Availability and Implementation All data is available from `gitlab.com/ElofssonLab/abag-benchmark-dataset` and DOI;10.5281/zenodo.15764903. Code is available at `https://github.com/samuelfromm/abag-benchmark-set`.
Supplementary information

Key words: AlphaFold, Protein Structure Predictions, Antibody-antigen interactions

Introduction

Antibody-antigen interactions play a crucial role in immunology and various biotechnological applications. Furthermore, the use of therapeutic antibodies has increased in recent decades[1]. Therefore, accurately predicting these interactions is essential. It can accelerate the discovery process by identifying potential candidates for immunotherapies and enhance our ability to design targeted treatments for various diseases, including cancers and autoimmune disorders.

The prediction of protein complexes has advanced considerably[2] with the advent of AlphaFold2[3], initially designed to predict the architectures of individual proteins. However, AlphaFold2's accuracy decreases for interactions lacking co-evolutionary signals, particularly in complex biological systems like host-pathogen[4] and antibody-antigen associations[5]. Recognizing this limitation, two primary strategies have been established to improve prediction accuracy.

The first strategy involves extensive sampling methods demonstrated during the CASP15 competition[1, 6], where multiple models were generated, increasing the likelihood of identifying accurate conformations. Several alternative methods to increase the efficiency of the sampling have been proposed, including using dropouts[1, 6, 7], masking columns in the MSA[8], and subsampling the MSA[9].

The second strategy, introduced in AlphaFold3, is to replace the evoformer with the pairformer architecture. In contrast to the evoformer, the pairformer does not reason over multiple sequence alignment. This shift facilitates a more robust computational reasoning about structural interactions without relying primarily on co-evolutionary data, enhancing model accuracy. This method is also used in Chai-1[10] and Boltz-1[11]. AlphaFold3 and Chai-1 reported an improved accuracy, specifically in predicting antibody-antigen interactions.

Several scoring functions have been developed to estimate the accuracy of predicted structures, starting with the introduction of pDockQ, used to predict the DockQ from AlphaFold2.0[2]. With the introduction of AlphaFold-multimer (AlphaFold2.1)[12], an internal score (ipTM) can be used to estimate the quality of the predicted interaction. Internally, AlphaFold2 and the other methods use the ranking confidence, a linear combination of 80% ipTM and 20% pTM. Several attempts have been made to improve the ranking. For instance, pDockQ2 (Zhu et al. 2023) can be used to predict the quality of interactions of each chain, actifpTM (Varga, Ovchinnikov, and Schueler-Furman 2024) is tuned to rank interactions involving flexible regions, and ipSAE (Dunbrack 2025) is developed to

© The Author 2025. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

better separate true from false interactions than ipTM. All of these methods might provide a somewhat limited improvement over the self-estimated AlphaFold2 (i)pTM scores. However, it has not been clearly demonstrated that they significantly improve the per-target ranking of antibody-antigen complexes.

In this study, our objective was to evaluate new structure prediction methods for the prediction of antibody-antigen complexes using a diverse benchmark. We assess whether larger sampling, model selection, and confidence metrics can improve the identification of accurate models. Additionally, we examine the effectiveness of confidence metrices in ranking predicted structures.

Material and Methods

Dataset

The dataset consists of 111 antibody-antigen structures, created using a slightly modified version of the Antibody Antigen Dataset Maker (AADaM) [13] available from `https://github.com/samuelfromm/AADaM-fork`. We initially use all structures deposited in SAbDab [14] as of 2024-10-23. These structures are then divided into two subsets: a "before" 2021-09-30 dataset and an "after" 2021-09-30 (benchmark) dataset. The chosen date 2021-09-30 represents the cutoff date for the training data used in all benchmarked methods. The "before" dataset was filtered to retain only structures in which at least one of the antigen chains is a peptide or protein.

Antibody-peptide structures were excluded from the benchmark dataset which was further filtered to include only structures determined by X-ray diffraction or electron microscopy, with a resolution of 3.5 Å or better. Structures with nonstandard residues were excluded unless at the sequence ends, in which case they were trimmed. Structures missing more than 10% of residues were also discarded. Structures are discarded from the benchmark dataset if the sequence identity between the benchmark and 'before" dataset exceeded 80%. Sequence identity between the two datasets was calculated individually for the heavy (H), light (L), and antigen chains using a global alignment. The benchmark set itself is further clustered by sequence identity using the same procedure and threshold. During clustering, the structure with the fewest missing residues in the H, L, and antigen chains is retained. If two structures have the same number of missing residues, the one with the shorter antigen sequence is selected.

Model Generation

AlphaFold2.3 methods to enhance sampling

To increase sampling diversity we examined previously described methods [9, 6, 8] by modifying the AlphaFold2.3.2 code. All methods were run without relaxation. The maximum template date is set to 2021-09-30. We utilize the latest version of the weights (v3) and, for each of the five available models, we run 40 predictions, resulting in a total of 200 models per method.

In total, five different AlphaFold2-based methods were evaluated:

- **AlphaFold2.3**: AF2.3 with default parameters.
- **AF2.3-SubSampl-32-64**: AF2.3 with MSA subsampling using `num_msa=32` and `num_extra_msa=64` compared to the default values of 508 and 2048, respectively.
- **AF2.3-SubSampl-16-32**: AF2.3 with `num_msa=16` and `num_extra_msa=32`.
- **AF2.3-Dropout**: Using the ideas from AFsample [6], i.e. we turn on dropout during interference.
- **AF2.3-ColMask**: Using the ideas from AFsample2 [8], i.e. we randomly mask 15% of the columns in the msa.

The code for our implementation of the above methods is available at `https://github.com/samuelfromm/alphafold-clone`.

AlphaFold 3, Chai-1 and Boltz-1 models

For each complex in the dataset, 40 different random seeds (1 to 40) were used to generate five models each, resulting in a total of 200 models per complex.

MSA generation

The multiple sequence alignments (MSAs) were generated using the AlphaFold 2.3.2 pipeline (specifically commit `f251de6`). One of the 111 complexes failed due to an HHblits-related issue and was excluded. For running AF3 and Boltz-1, we specifically used `bfd_uniref_hits.a3m` file produced by this pipeline, while for Chai-1, MSAs were generated using the `--use-msa-server` option via the ColabFold [15] server.

Evaluation

Scoring

In order to calculate scores, we created a snakemake workflow. For each prediction, the corresponding model (query) and the PDB ground truth structure of the corresponding PDBID (reference) were used as inputs.

AADaM, the pipeline used to generate our dataset, outputs a FASTA sequence for each antibody-antigen structure. However, these structures do not always fully align with their corresponding PDB entries. Since AADaM focuses on the antibody-antigen interface, some chains present in the original PDB file may be excluded from the interface structures used in our dataset. Additionally, the sequence present in reference structure from PDB is often shorter than the SEQRES sequence used to generate model structures. To address both these issues, we first run MMalign [16] to calculate the alignment of the chains in the query and reference structure. We then take each pair of aligned chains and calculate a sequence alignment of those chains. Based on this sequence alignment, we select the residues present in both the query and reference structure.

Using the cropped query and reference structures as input, we calculate several metrics:

1. We run MMalign to calculate the TMscore.
2. We calculate the DockQ score [17, 18] for the antibody-antigen interface ("merging" multi-chain antigens into a "single chain"). This results in a single DockQ score for each complex.
3. We calculate pDockQ version 2.
4. We extract the confidence scores from the predictions as output by the respective method: `ranking_confidence` for AlphaFold-based methods, `confidence_score` for Boltz-1, and `aggregate_score` for Chai-1.
5. We calculate the aligned error between the cropped query and reference structure as well as the aligned error ranking confidence (see section *Aligned Error*).

The workflow is the same for all samples and methods with the exception of the last step: The aligned error analysis was only performed for a selection of methods. The complete workflow is available as part of the provided code.

In some cases, MMalign cannot align the chains properly, indicating that the query and reference structures are dissimilar and the predicted structure is likely of very low quality. We still run the rest of the workflow, even though the number of aligned residues can be extremely low. When the number of aligned residues, normalized by the query length, is below 0.9, the resulting DockQ scores typically fall between 0 and 0.05. We consider these low scores to accurately reflect the poor quality of the predicted interfaces in such cases.

Sampling

For each method, we generate up to 200 models and treat them as independent models, although in AF3's inference each random seed generates 5 different models. These models use the same information from the pairformer, but differ by the built-in randomness of the diffusion process used to create the models. To assess the impact of sampling, we randomly select k models with replacement, and identify the best (highest DockQ) and top-ranked (highest confidence) models, recording their DockQ scores. For each sampling size k, we average the recorded DockQ score over 50 iterations. We then average the resulting score over all IDs and sample sizes.

Structural Comparison

We used global and interface similarity to investigate whether structural similarity exists between the models in the test and training sets. First, we employed FoldSeek-Multimer [19, 20] to identify global similarity between a complex in the test set and any protein present in PDB prior to the split date (2021-09-30). For each complex, the maximum TM-score was normalized by the length of the query protein. Next, we created a smaller dataset to examine the interface similarity using USalign [21]. This dataset included all 1998 PDB entries with a TM-score > 0.5 to any complex in the test set. We first extracted all interface residues from the proteins in the test set to evaluate interface similarity. Interface residues were defined as any antigen residue where any atom was within 12 Å of the antibody or vice versa. Next, the interface was aligned to the 1998 PDB entries mentioned above using USalign with the `-mm 1` flag, and the maximum TM-score to a protein deposited in PDB before the cutoff date was recorded.

Aligned Error

We follow the notation in [3] [§1.9.7]. Assume that we are given two structures - a reference and query structure - for the same sequence. Let $Y = \{\vec{y}_j\}$ and $X = \{\vec{x}_j\}$ denote the C^α-positions of the reference and query structure, respectively, let $T_{Y,i}$ and $T_{X,i}$ be the corresponding backbone frames, and let $e_{ij}(X,Y) = \left\| T_{X,i}^{-1} \circ \vec{x}_j - T_{Y,i}^{-1} \circ \vec{y}_j \right\|$ be the error in the Cα-position of residue j when the query and reference structures are aligned using the backbone frame of residue i. The non-symmetric matrix $(e_{ij})_{i,j=1,\ldots,N_{\text{res}}}$, where N_{res} is the number of residues, is called the aligned error (AE) between the two structures. Let

$$f(d,n) = \frac{1}{1 + \left(\frac{d}{d_0(n)}\right)^2}, \quad d_0(n) = 1.24\sqrt[3]{\max(n,19) - 15} - 1.8.$$

We define the aligned error TM-score (aeTM) and aligned error interface TM-score (aeiTM) as

$$\text{aeTM}(X,Y) := \max_{i\in[1,\ldots,N_{\text{res}}]} \frac{1}{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} f(e_{ij}(X,Y), N_{\text{res}}) \quad (1)$$

and

$$\text{aeiTM}(X,Y) := \max_{i\in[1,\ldots,N_{\text{res}}]} \frac{1}{|\mathfrak{D}_i|} \sum_{j=1}^{N_{\text{res}}} f(e_{ij}(X,Y), |\mathfrak{D}_i|), \quad (2)$$

respectively. Here $\mathfrak{D}_i$ denotes the set of all residues except those of the chain of residue i.

Note that the aeTM score and the aeiTM score can be viewed as idealized versions of the predicted TM (pTM) score and interface predicted TM (ipTM) score, respectively. Indeed, let us replace X with a predicted structure and assume that the ground truth structure X^{true} is unknown. Assume we have a random variable $Y(\zeta)$, sampling structures from a distribution of probable ground truth structures. Then the pTM and ipTM scores are given by

$$\text{pTM}(X) = \max_{i\in[1,\ldots,N_{\text{res}}]} \frac{1}{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} \mathbb{E}\left[f(e_{ij}(X,Y(\zeta)), N_{\text{res}})\right] \quad (3)$$

and

$$\text{ipTM}(X) = \max_{i\in[1,\ldots,N_{\text{res}}]} \frac{1}{|\mathfrak{D}_i|} \sum_{j=1}^{N_{\text{res}}} \mathbb{E}\left[f(e_{ij}(X,Y(\zeta)), |\mathfrak{D}_i|)\right], \quad (4)$$

respectively (see AF2, §1.9.7 and AF2-multimer, §7.9). Here the expectation is taken over the probability distribution of $e_{ij}(X,Y(\zeta))$. Now, in an ideal situation, $\mathbb{P}(Y(\zeta) = X^{\text{true}}) = 1$, in other words the random variable Y takes on the values X^{true} with probability 1, in which case $\mathbb{E}\left[f\left(e_{ij}(X,Y(\zeta))\right)\right] = f\left(e_{ij}(X,X^{\text{true}})\right)$ so that $\text{pTM}(X) = \text{aeTM}(X,X^{\text{true}})$ and similarly $\text{ipTM}(X) = \text{aeiTM}(X,X^{\text{true}})$.

In practice, AlphaFold learns to predict the probability distribution of the elements of the aligned error $e_{ij}(X,Y(\zeta))$. To this end, the distribution of e_{ij} is discretized into 64 bins, ranging from 0 to 31.5 Å with a bin width of 0.5 Å, where the final bin captures any errors larger than 31.5 Å. When calculating the aligned error between two structures X and Y, we do however allow unlimited error values.

While the pTM score is a predicted metric, the aeTM takes as input any two structures of the same sequence, similarly to the TM-score. Note also that the aeTM score is bounded from above by the TM-score (see AF2, §1.9.7).

Finally, recall that AlphaFold's ranking confidence score is defined as $0.2 \cdot \text{pTM} + 0.8 \cdot \text{ipTM}$. In a similar fashion we define the aligned error (ae) ranking confidence as $0.2 \cdot \text{aeTM} + 0.8 \cdot \text{aeiTM}$.

Results and Discussion

Here, we evaluated the performance of AlphaFold2.3, AlphaFold3, Boltz-1, and Chai-1 on a dataset of 110 unseen antibody-antigen sets, given a different number of generated models. For AlphaFold2.3, we also examined other strategies that have been reported to increase the sampling efficiency.

It has been reported in several studies[6, 7, 22] that generating more samples improves the prediction quality of AlphaFold2. Further, in AlphaFold3[23] paper, it is claimed that generating up to 1000 samples improved the predictions of antibody-antigen complexes. We generated 200 models for each method and examined their performance, measured by the DockQ score[17, 18] between the antigen and the antibody. The performance was measured for each sampling size by using the mean DockQ of the best model for each target and the mean DockQ of the top-ranked models, see methods for details. A summary of the results is presented in Figure 1.

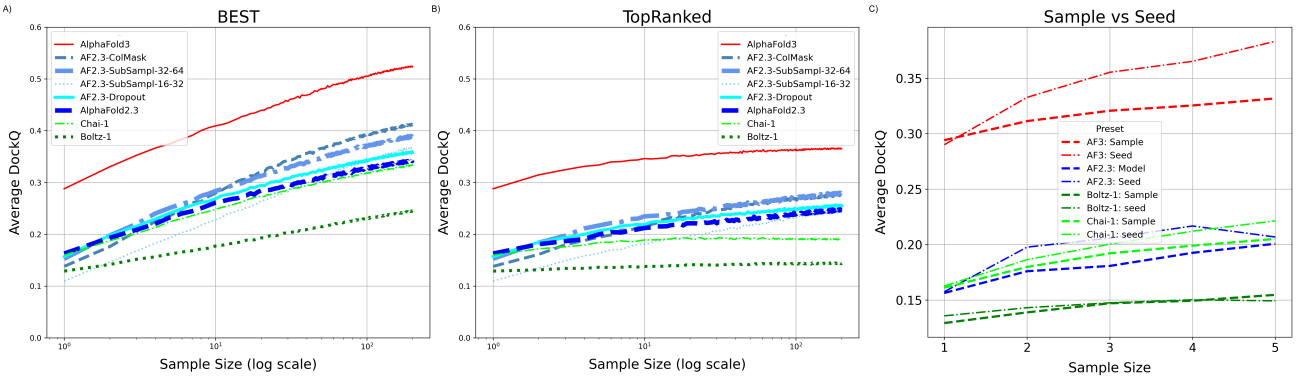

Fig. 1. For each method and sample size n we randomly select n models out of the 200 structures we created (with replacement). Out of these n selected models, we select the model with the highest DockQ score (A) and highest ranking confidence (B) respectively, and record its DockQ score on the y-axis. C) Seed vs Sample for five models

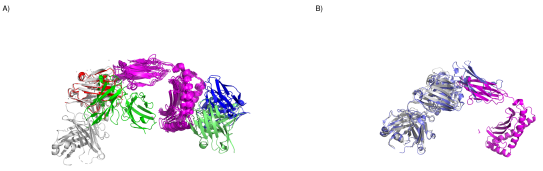

Fig. 2. A) Example (8tq7) of a prediction where AlphaFold3 is superior to all other methods. The antigen is shown in Magenta, Grey - PDB structure, Red AF3, Green Boltz-1, Lime Chai1, Blue AF2.3. B) Superposition of 8tq7 with the most similar model in the training set (6ddm)

AlphaFold3 makes better predictions than the other methods.

First, we will discuss the ability to generate a good model, while assuming that an oracle to identify the best model among all the models generated by a method exists. Figure 1A shows that AlphaFold3 outperforms all other methods by some margin, and that AlphaFold2 performs better than Chai-1 and Boltz-1. The unexpectedly poor performance of Chai-1 and Boltz-1 was surprising and led us to investigate potential errors, including rerunning a subset of targets using the web server. However, this did not yield any significant improvement. One example where AF3 performed considerably better than other methods is target 8tq7, see Figure 2A. Here, the prediction of AF3 is spot on, while all other methods predicted binding to different areas of the antigen.

Secondly, it can be observed that all methods can generate better models when sampling is increased. A roughly linear increase can be seen with the logarithmic number of samples up to about 100 samples, and then the linear trend stops for AF3. This means that the average DockQ for AF3 of the best model increases from below 0.3 to above 0.5 when 200 samples are generated. A similar increase is seen for AlphaFold2, Chai-1, and Boltz-1.

In AlphaFold3, additional models can be generated, either by rerunning the entire pipeline with a different random seed, or by letting the diffusion model produce more samples. Figure 1C shows that using different seeds yields greater model diversity than generating more samples from the diffusion model alone.

AF2.3-ColMask is the best strategy to increase sampling.

As shown before, it is possible to increase the diversity of the generated models by several different strategies. We examined these strategies by modifying AlphaFold2.3, as it is unclear if the AlphaFold3's license would allow this modification. Also, this study was initiated before AlphaFold3 was released. In CASP15, it was shown that increased sampling and using dropouts increased the accuracy of the predictions, particularly for Antigen-antibodies[6]. This method was called AFsample. However, we refer to it here as AF2.3 dropout, as the implementation is not identical to the one used in CASP15. More recently, the Wallner group has reported that an even more aggressive method to add noise by masking entire columns in the MSA can generate alternative conformations[8]. Alternatively, it is possible to create alternative conformations by subsampling the MSA[9]. Here, we used two different settings to generate these models.

Figure 1B shows that all methods, on average, generate worse models than the default AlphaFold2.3 setting when only one model is generated, see Sample Size=1(10^0). The difference is negligible for all methods except for AF2.3-ColMask (which is very similar to AFsample2), which produces significantly worse average models, DockQ of 0.138 vs 0.163. However, when mode models are generated, all these methods produce better models than using AlphaFold2.3 natively. Overall, AF2.3-ColMask generated most models, but it is possible that if even more than 200 models are generated, AF2.3SubSample-16-32 would outperform AF2.3-ColMask, as it seems to improve more with additional samples. It should be noted that, except for SubSample, we used the default parameters for these methods. Other hyperparameters, such as dropout rates, might yield even better models.

A large gap between the best and top-ranked models.

It is necessary to identify good models in addition to generating them. Figure 1B shows the average DockQ score change with sample size. For all methods, the increase in average DockQ with sample size is much smaller than what is observed for the best model (Figure 1A). This means that the internal ranking of the models cannot identify the best generated model. For AlphaFold3, it means that the average DockQ only increases from 0.29 to 0.37 instead of 0.52 if the best models had been identified. For Boltz-1 and Chai-1, there is hardly any

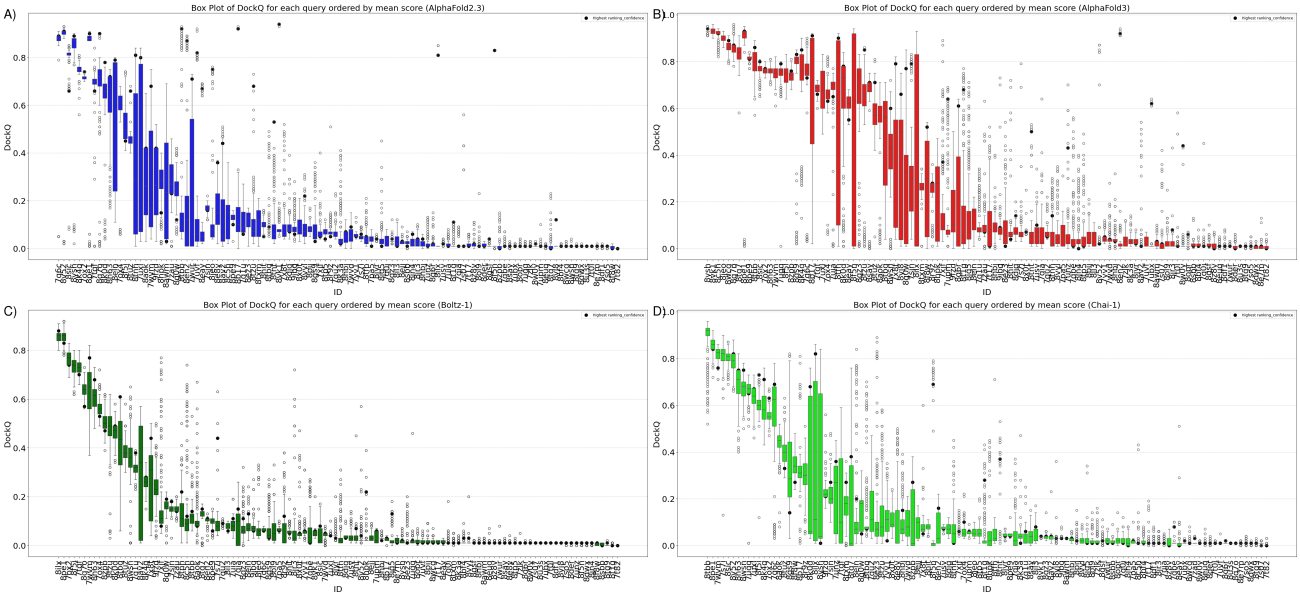

Fig. 3. Each figure shows a box plot for the DockQ score for each target ID, sorted by mean DockQ score. The highest-ranked model is highlighted with a thick dot. A) AlphaFold2.3, B) AlphaFold3, C) Boltz-1 and D) Chai-1.

improvement with larger sample sizes (from 0.12 to 0.14). For the AlphaFold2.3-based models, a similar trend as seen for AlphaFold3 can be observed. However, AF2.3-SubSample performs as well as AF2.3-ColMask when 30 or more models are generated.

In CASP, up to five models were allowed, and it is often good to submit a diverse set of models (Elofsson 2025), so if such a strategy were applied here, the difference to the best possible model would decrease. However, the best approach for selecting five models has not been well studied, and therefore, we have only focused on choosing the top-ranked model. Anyhow, it is clear that better detection of top-ranked models would significantly improve the prediction of antibody-antigen structures.

In Figure 3, we take a more detailed look at the distribution of the scores for each target. The Figure shows a box plot for the DockQ score for each target, sorted by mean DockQ score, with the highest ranked model shown as a thick dot. It can be seen that for all methods, the targets can, in general, be divided into three types. To the left in the plots are the easy-to-predict targets where most models are correct, but a few incorrect models also exist. To the right in the plots are challenging targets, where most models are wrong, but a few good models exist for some targets. In the middle are models with many good and bad models. Further, the main difference between the four methods is the number of targets within each category, so AlphaFold3 has many more easy-to-predict targets. This indicates that performing downstream analysis using the mean DockQ score of all 200 targets for each model is less noisy than using the maximum or first-ranked models.

In Figure 4, it can be seen that for both top-ranked and best models, some targets where AlphaFold2.3 outperformed AlphaFold3 exist. This indicates that there is no subset of easier models; instead, there exists a subset of models where AlphaFold3, for some reason, makes significantly better (average) models than the other methods. Next, we will try to examine the reason for this.

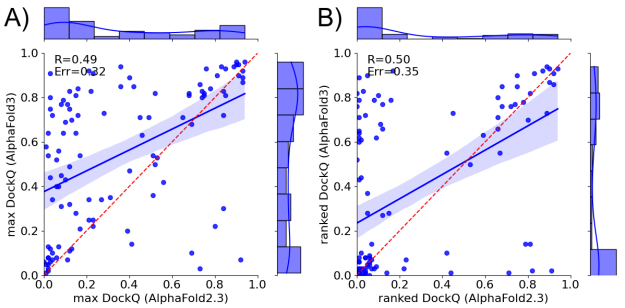

Fig. 4. DockQ scores for AlphaFold3 vs AlphaFold2.3 for best models (A) and top-ranked models (b)

Has AlphaFold3 learned to predict structure beyond its training set?

Next, we asked how AlphaFold3 can predict many targets almost perfectly while other methods failed for the same targets. There should be no co-evolutionary signal between the antibody and the antigen so that AlphaFold3 may have learned something about the physics governing the interactions. Alternatively, it may have merely learned to find patterns in the database that other methods did not see. To answer this, we compared each target in our test set with the training data, i.e., all of PDB until Sep 2021. The search was made either by examining the entire complex (Figure 5B) or just the interface (Figure 5A). For each complex in our test set, we noted the maximum similarity to any complex in the training set, either by using FoldSeek for the entire complex or using USalign for the extracted interface region, see methods for details.

As shown in Figure 5, the mean DockQ score significantly increases when a structurally similar antibody-antigen complex exists in the training set of AlphaFold3, and to a lesser extent for Chai-1, but not for other methods, as they are generally worse. However, it should be noted that well-predicted models

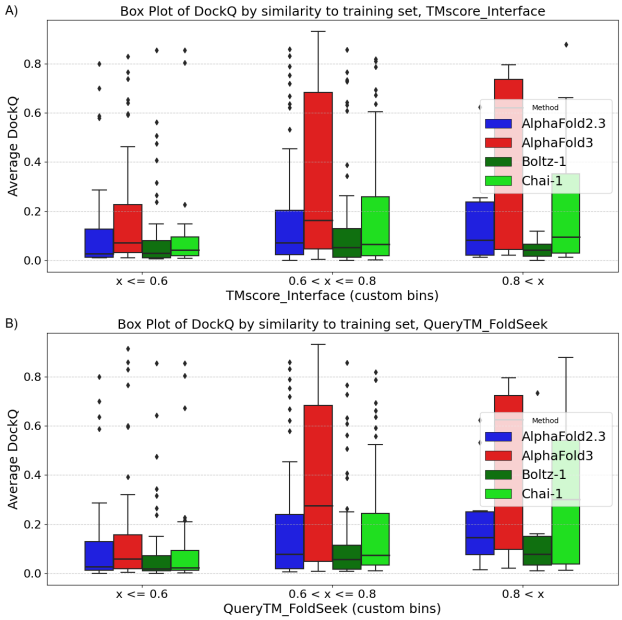

Fig. 5. Prediction performance vs. similarity to the training data for AlphaFold2.3 and AlphaFold2. FoldSeek query TM score. Mean score (max gives vastly different plot, first very similar)

exist where no structural similarity can be detected. The prediction is often dire, even if a structurally similar complex exists in the training set. In general, the correlation between the AbAG-DockQ score and the similarity to the dataset is low (Cc=0.32 for complex search and 0.17 for interface), further supporting the idea that similarity with the training set is neither a requirement nor a guarantee for a correct prediction.

We also noted that it is not trivial to understand what AlphaFold3 might have used from the training set to make excellent predictions. Figure 2B shows an example of the most similar hit in the training set to a model that AlphaFold3 (but not other methods) predicted well (8tq7). It can be seen (in magenta) that the best hit (6ddm) has a similar beta-sheet domain that binds to the antibody. However, it is also clear that the interface is not identical (interface TM-score is 0.63), indicating that, for this case, AlphaFold3 may have learned the orientation of the domains rather than the exact packing of the interface.

Ranking of AlphaFold3 models

Figure 1 clearly shows that an improved ranking of the models for each target would significantly improve antibody-antigen predictions. The average DockQ score would increase from 0.37 to 0.52. Although the correlation between ranking confidence (or other scoring method) for all targets is not too bad (Cc=0.80 for ranking confidence), it is quite bad for each target (mean CC=0.28) as can be seen in Figure 6. Although some targets correlate with ranking confidence and DockQ, many do not. Some targets, such as 8jq5 (Figure 6E), have almost identical ranking confidences. At the same time, the DockQ varies widely, while others, such as 7wvm (Figure 6B), have virtually identical (bad) DockQ scores, while the ranking confidence differs widely. Alternative measures, such as pDockQ, which are also based on the predicted qualities (pLDDT and PAEs), show very similar behaviour; data not shown. Therefore, it is likely necessary to include other types of data if this per-target correlation should increase.

Ranking analysis

One of the most innovative designs of AlphaFold2 was its ability to accurately estimate the quality of the generated models. As described in the methods section the estimates stem from two predicted features, the prediction of the accuracy of an individual residue (pLDDT) and the Predicted Aligned Error (PAE), estimating the error in the distance between pairs of residues. Since the introduction of AlphaFold-multimer these estimated distances have been used to calculate pTM, to estimate the interaction of an entire protein/protein complex, and ipTM, estimating the accuracy specifically of the interactions between chains.

Attempts to improve these scores have been proposed since the introduction of pDockQ[2] designed to predict the DockQ score predicted by AlphaFold2.0. pDockQ bases its predictions on the number of interacting residues and their pLDDT values. Several other methods based on similar ideas have been introduced later. Recently, there have been attempts to improve AlphaFolds's ipTM and pTM scores (and thus the ranking confidence) by changing the way these values are calculated from the PAE values [24, 25]). These methods rely on modifying the functions (**??**) and (**??**) that are used to calculate the pTM and ipTM scores, respectively, while still relying on the same features that are used to calculate the ipTM and pTM scores.

To demonstrate the importance of the accurate predictions of PAEs we have also included these measures calculated from perfect PAE values. One example of how this affects the predictions can be seen in Figure 7A-D. The two models are of 8hit, one is almost perfect (DockQ: 0.690), while the other is rather bad (DockQ: 0.040. However, the PAE matrices for the two cases are almost ideal.

To better understand the impact of the quality of the features versus the accuracy of the functions, we investigate how well the aligned error based scores (1), (2), as well as the aligned error Ranking Confidence perform. We compared these measures on their ability to identify a top-ranked model, as well as their correlation to DockQ, both overall and per target, see Figure 7E and Table1. Here, it is clear that independently on what measure based on the predicted PAE values is used (i) the average DockQ score of the top-ranked model is significantly lower than the best possible DockQ (0.35, vs 0.54), (ii) the overall correlation with DockQ is acceptable (> 0.7), while (iii) the per-target correlation is much worse (< 0.30). At the same time using any of the measures with idealized PAE predictions solves these problems.

Figure 7 F) shows the average DockQ score for each ranking method. As before, selecting by DockQ score yields the best possible improvement. Selecting by aeiTM yields results much closer to the best possible outcome compared to selecting by ipTM, indicating that the limiting factor when it comes to the ipTM score, at least for antibody-antigen complexes, is the accuracy of the predicted aligned error (and not necessarily a problem with the way the ipTM score is calculated itself).

Conclusion

In this study, we benchmarked different sampling and ranking strategies against a selected ensemble of unseen antibody-antigen complexes.

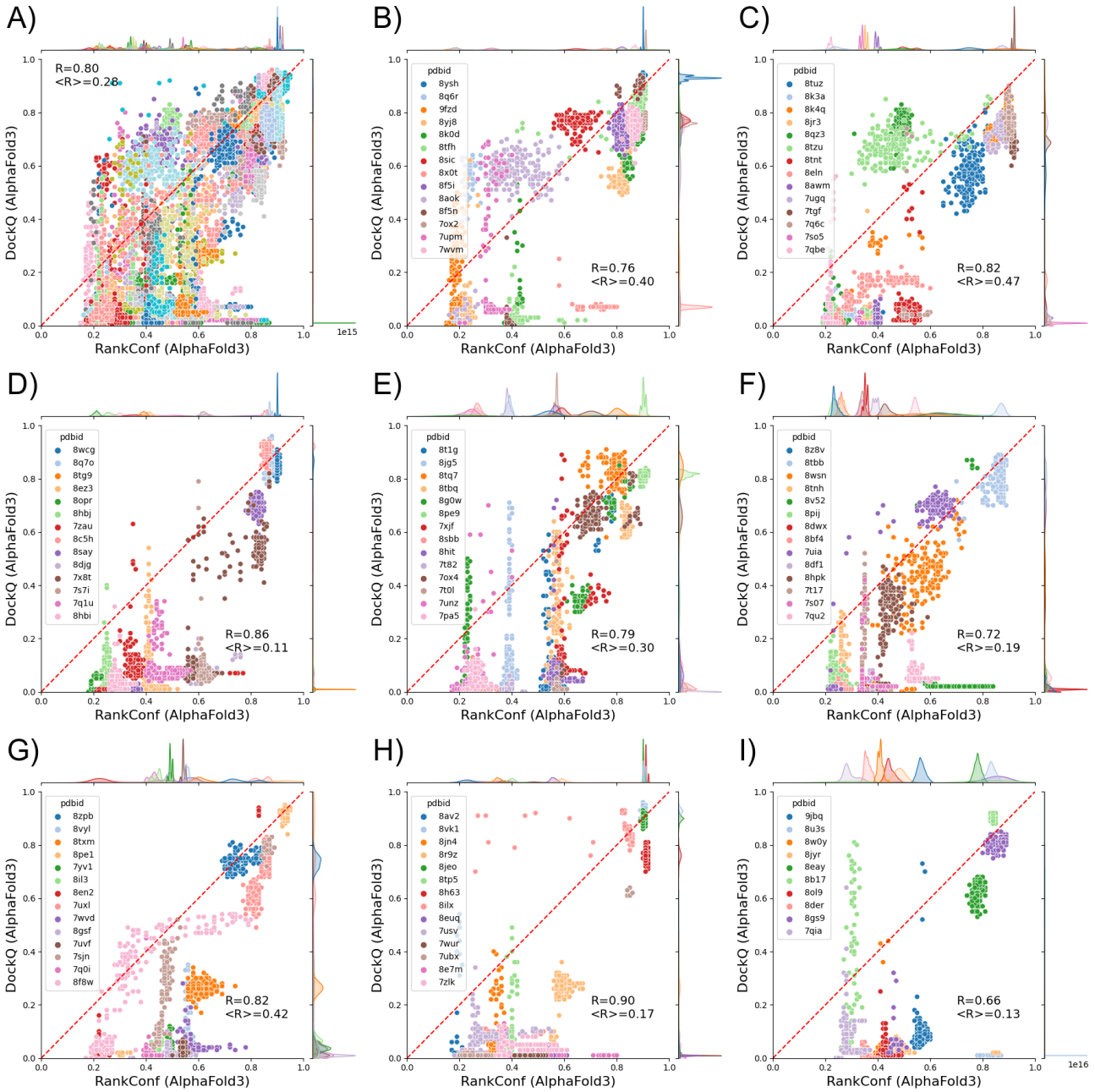

Fig. 6. DockQ scores vs Ranking Confidence for all AlphaFold3 models. A) shows all models one color for each target. The overall correlation coefficient is 0.8 while the mean correlation coefficient per target ($< R >$) is only 0.28. B-I) For better overview the targets are spread over 8 panels.

Our findings reveal that (i) AlphaFold3, but not Chai-1 and Boltz-1, consistently offers improved predictions over AlphaFold2, (ii) larger sampling sizes correlate with an increased probability of generating at least one correct model, (iii) a crucial bottleneck persists in accurately identifying the best model among the generated models, and (iv) AlphaFold3's performance drops considerably for complexes without prior structural similarities to the training dataset, underscoring the model's dependence on previously encountered data for accurate predictions.

Furthermore, to analyse the confidence metric used by AlphaFold (ipTM score), we modified the formula used to calculate ipTM to improve its accuracy, but only obtained limited improvement. In contrast, if the underlying estimates of accuracies (the predicted aligned error) are sufficiently accurate, we show that using the default ipTM formula is sufficient to achieve close to ideal per-target ranking and thereby significantly increase the DockQ score of the top-ranked model. This indicates that the limitation lies not in the ipTM formula itself but in AlphaFold's ability to accurately predict aligned errors.

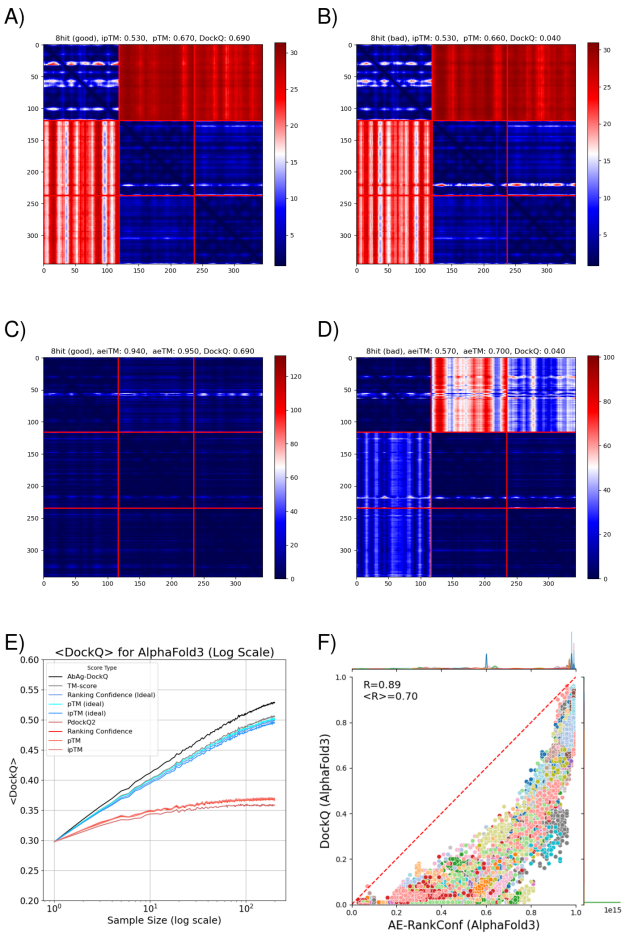

Fig. 7. A) and B) PAE plots for two models of 8hit. The chain breakers are marked in red. The mean absolute error (mae) between the two PAE matrices is 1.03 across the whole matrix and 1.37 across the inter-chain region. C) and D) AE plots for the same two models 8hit. The mae between the two AE matrices is 13.24 across the whole matrix and 19.53 across the inter-chain region. E) Ranking using different ways to use the PAEs to calculate the scores, in red methods using predicted PAEs and in blue using AE (i.e. idealized PAE) scores. The primary issue is that the ground-truth aligned error values are not sufficiently accurate, resulting in a significant performance gap. F) the aligned error ranking confidence values (compare with Figure 7A).

Data availability

All data is freely available. The benchmark dataset can be found at `https://gitlab.com/ElofssonLab/abag-benchmark-dataset`. The raw data and processed data is available at `https:???`. The code is available at `https://github.com/samuelfromm/abag-benchmark-set`.

Author contributions

S.F. prepared the dataset. S.F. and M.L. generated the models. A.E. and S.F. performed the analysis. A.E. wrote the initial draft of the manuscript. All authors contributed to the final version of the manuscript.

Competing interest

The authors declare no competing interests.

Table 1. Summary for different methods to rank AlphaFold3 models, <DockQ> and Pearson/Spearman correlations

Method	$< DockQ >$	R	$< R >$
DockQ	0.544	1.000	1.000
pTM (ideal)	0.516	0.814	0.665
ipTM (ideal)	0.508	0.873	0.562
RankConf (ideal)	**0.375**	**0.711**	0.214
ipTM	0.370	0.707	**0.220**
pTM	0.369	0.670	0.200
PAE-iTM	0.368	0.642	0.206
PAE RankiConf	0.365	0.644	0.213
PAE-TM	0.358	0.617	0.186
PdockQ2	0.352	0.603	0.184

Achnowledgement

AE was funded by Vetenskapsrådet, Grant No. 2021-03979, and the Knut and Alice Wallenberg Foundation, Grant No. 2022.0032. The computations and data handling were enabled by the supercomputing resource Berzelius, provided by the National Supercomputer Centre at Linköping University, the Knut and Alice Wallenberg Foundation, and SNIC, grant numbers SNIC 2021/5-297 and Berzelius-2021-29.

References

1. Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science*, 27(1):1–30, January 2020.
2. Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun*, 13(1):1265, March 2022.
3. 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.
4. Nicholas J Talbot. *Plant-pathogen Interactions*. CRC Press, 2004.
5. Rui Yin and Brian G Pierce. Evaluation of AlphaFold antibody-antigen modeling with implications for improving predictive accuracy. *Protein Sci*, 33(1):e4865, January 2024.
6. Björn Wallner. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*, 39(9), September 2023.
7. Nessim Raouraoua, Claudio Mirabello, Thibaut Véry, Christophe Blanchet, Björn Wallner, Marc F Lensink, and Guillaume Brysbaert. MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. *Nat Comput Sci*, 4(11):824–828, November 2024.
8. Yogesh Kalakoti and Björn Wallner. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Commun Biol*, 8(1):373, March 2025.
9. Diego Del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife*, 11, March 2022.
10. Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life.

bioRxiv, page 2024.10.10.615955, October 2024.

11. Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, page 2024.11.19.624167, May 2025.
12. 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. *bioRxiv*, page 2021.10.04.463034, March 2022.
13. Katherine Maia McCoy, Margaret E Ackerman, and Gevorg Grigoryan. A comparison of antibody-antigen complex sequence-to-structure prediction methods and their systematic biases. *Protein Sci*, 33(9):e5127, September 2024.
14. Constantin Schneider, Matthew I J Raybould, and Charlotte M Deane. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res*, 50(D1):D1368–D1372, January 2022.
15. Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022.
16. Srayanta Mukherjee and Yang Zhang. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res*, 37(11):e83, June 2009.
17. Claudio Mirabello and Björn Wallner. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10), October 2024.
18. Sankar Basu and Björn Wallner. DockQ: A quality measure for protein-protein docking models. *PLoS One*, 11(8):e0161879, August 2016.
19. Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nat Biotechnol*, 42(2):243–246, February 2024.
20. Woosub Kim, Milot Mirdita, Eli Levy Karin, Cameron L M Gilchrist, Hugo Schweke, Johannes Söding, Emmanuel D Levy, and Martin Steinegger. Rapid and sensitive protein complex alignment with foldseek-multimer. *Nat Methods*, February 2025.
21. Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods*, 19(9):1109–1115, September 2022.
22. Nessim Raouraoua, Marc F Lensink, and Guillaume Brysbaert. Massive sampling strategy for antibody-antigen targets in CAPRI round 55 with MassiveFold. *Proteins*, January 2025.
23. 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. Addendum: Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 636(8042):E4, December 2024.
24. Roland L Dunbrack. Rēs ipSAE loquunt: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv*, page 2025.02.10.637595, February 2025.
25. Julia K Varga, Sergey Ovchinnikov, and Ora Schueler-Furman. actifpTM: a refined confidence metric of AlphaFold2 predictions involving flexible regions. *arXiv*, page 2412.15970, December 2024.