Improved Multimer Prediction using Massive Sampling with AlphaFold in CASP15

Björn Wallner[1]

[1]Linkoping University Electronic Press

April 17, 2023

Abstract

AlphaFold has transformed structure prediction by enabling highly accurate predictions on par with experimentally determined structures. Still, for difficult cases, in particular, multimers, there is still room for improvement. Important for the success of AlphaFold is its ability to assess its own predictions. The basic idea for the Wallner group in CASP15 was to exploit the excellent ranking score in AlphaFold by massive sampling. To this end, we ran AlphaFold using six different settings, with and without templates, and with an increased number of recycles using both multimer v1 and v2 weights. In all cases, the dropout layers were enabled at inference to sample the uncertainty and increase the diversity of the generated models. A median of 4,810 models per target was generated and almost all (35/38) received a ranking_confidence >0.7. Compared to other groups in CASP15, Wallner obtained the highest sum of Z-scores based on the DockQ score, 40.8 compared to 26.3 for the second highest, much higher than -0.2 achieved by the AlphaFold baseline method, NBIS-AF2-multimer. The improvement over the baseline is substantial with the mean DockQ increasing from 0.43 to 0.56, with several targets showing a DockQ score increase by +0.6 units. Remarkable, considering Wallner and NBIS-AF2-multimer were using identical input data. The reason for the success can be attributed to the diversified sampling using dropout with different settings and, in particular, the use of multimer v1, which seems to be much more susceptible to sampling compared to v2. The method is available here: http://wallnerlab.org/AFsample/.

Improved Multimer Prediction using Massive Sampling with AlphaFold in CASP15

Björn Wallner[1]*

[1] Division of Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden

April 14, 2023

Abstract

AlphaFold has transformed structure prediction by enabling highly accurate predictions on par with experimentally determined structures. Still, for difficult cases, in particular, multimers, there is still room for improvement. Important for the success of AlphaFold is its ability to assess its own predictions. The basic idea for the Wallner group in CASP15 was to exploit the excellent ranking score in AlphaFold by massive sampling. To this end, we ran AlphaFold using six different settings, with and without templates, and with an increased number of recycles using both multimer v1 and v2 weights. In all cases, the dropout layers were enabled at inference to sample the uncertainty and increase the diversity of the generated models. A median of 4,810 models per target was generated and almost all (35/38) received a ranking_confidence>0.7. Compared to other groups in CASP15, Wallner obtained the highest sum of Z-scores based on the DockQ score, 40.8 compared to 26.3 for the second highest, much higher than -0.2 achieved by the AlphaFold baseline method, NBIS-AF2-multimer. The improvement over the baseline is substantial with the mean DockQ increasing from 0.43 to 0.56, with several targets showing a DockQ score increase by +0.6 units. Remarkable, considering Wallner and NBIS-AF2-multimer were using identical input data. The reason for the success can be attributed to the diversified sampling using dropout with different settings and, in particular, the use of multimer v1, which seems to be much more susceptible to sampling compared to v2. The method is available here: `http://wallnerlab.org/AFsample/`.

*Corresponding author: bjorn.wallner@liu.se

1 Introduction

The remarkable precision of AlphaFold (Jumper *et al.*, 2021) has ushered in a new era in the field of computational and structural biology, enabling highly accurate predictions that rival experimentally determined structures. AlphaFold has rapidly emerged as the preferred method for protein structure prediction (Cramer, 2021).

The success of AlphaFold can be attributed to its capacity to evaluate the accuracy of its own predictions. This involves estimating per-residue accuracy via the predicted LDDT (Mariani *et al.*, 2013) (pLDDT), as well as predicting the TMscore (Zhang and Skolnick, 2004) (pTM), and the predicted aligned error (PAE) between all pairs of residues (Jumper *et al.*, 2021) with high precision. The correlation coefficients for pLDDT and pTM with their actual values are 0.76 and 0.85, respectively (Jumper *et al.*, 2021), and crucially this correlation remains strong even for high-quality predictions. Furthermore, for multimer prediction, AlphaFold computes an inter-chain predicted TMscore (ipTM) for the inter-chain distances, which is also very accurate (Jumper *et al.*, 2021).

AlphaFold is capable of achieving highly accurate monomer predictions even without relying on structural templates, provided it has access to sufficient evolutionary-related sequences (Jumper *et al.*, 2021). However, this is not necessarily the case for multimers, where the evolutionary signal constraining the prediction is much weaker (Bryant *et al.*, 2022), and thus, more sampling may be necessary to improve the prediction. To address this issue, the default number of sampled structural models in AlphaFold-multimer was increased from 1 in version 1 (v1) to 5 in version 2 (v2) per neural network model. In addition, predicting transient interactions or interactions with flexible binding partners requires even more sampling to achieve optimal performance (Johansson-Åkhe and Wallner, 2022).

In cases where the evolutionary constraints have trapped the prediction in a local minimum in the conformational landscape or the evolutionary constraints are weak, simply increasing the number of sampled models may not be sufficient (Roney and Ovchinnikov, 2022). Alternative methods to achieve greater diversity among generated models include increasing the number of times the prediction is recycled in the network (Mirdita *et al.*, 2022), randomly perturbing (Alamo *et al.*, 2022), or altering the input MSA (Wayment-Steele *et al.*, 2022).

Alternatively, enabling the dropout layers in the neural network can also enhance diversity among generated models (Johansson-Åkhe and Wallner, 2022; Mirdita *et al.*, 2022). Dropout layers are typically utilized only during training to encourage neural networks to learn multiple redundant solutions to the same problem by stochastically dropping some of their weights. The AlphaFold

Weights	Dropout	Templates	Recycles	Names	
v1	Yes	Yes	3	v1-templates	weights:v1
v1	Yes*	No	3	v1-notemplates	
v1	Yes*	No	21	v1-recycles	
v2	Yes	Yes	3	v2-templates	weights:v2
v2	Yes*	No	3	v2-notemplates	
v2	Yes*	No	9	v2-recycles	

Table 1: The six different settings of AlphaFold used in by the Wallner group **Weights** refers to version of the multimer neural network weights, **Dropout** refers to if dropout was enabled, **Templates** refers to if structural templates were used or not, **Recycles** refers to how many recycles was used (default is 3), **Names** refers to what the setting or combination of settings are referred to in this study.
*No dropout in structural module

network has dropout rates of 0.1-0.25, depending on the network module. Activating these layers during inference allows the network to naturally sample the uncertainties prediction (Gal and Ghahramani, 2016), thereby increasing the structural diversity of the generated models.

2 Methods

The basic idea for the Wallner group in CASP15 was to exploit the excellent ranking score in AlphaFold by massive sampling. To this end, we ran AlphaFold using six different settings, see Table 1, involving both version 1 (v1) and version 2 (v2) multimer weight sets, templates or no templates, as well as an increased number of recycles. In all cases, the dropout layers were activated at inference, however for the cases with no templates and the increased recycles, the dropout rate in the structural module was set to 0, to disable dropout in the structural module. In a previous study, we noticed a slight increase in the correlation between ranking confidence and actual structural quality when not using dropout in the structural module (Johansson-Åkhe and Wallner, 2022).

2.1 AlphaFold Sampling

The aim was to generate 1,000 models per setting for a total of 6,000 per target. The number of models actually generated is shown in Figure S1. The median number of models is 4,810, but for some large targets, only 13 models were generated and for some other targets, 30,000 models were generated. In addition, to save computational time if a `ranking_confidence>0.7` was obtained, no further models were generated. The latter was achieved for all but three targets.

2.2 Model selection

Models were ranked according to the `ranking_confidence` reported by AlphaFold. This score is a linear combination of the interface predicted TMscore (ipTM) and the overall predicted TMscore (pTM):

$$\text{ranking\_confidence} = 0.8\text{ipTM} + 0.2\text{pTM}$$

The difference between pTM and ipTM is that pTM assesses the errors *within* each chain, while ipTM assesses the error *between* chains.

The model ranked highest was submitted as the first prediction. To avoid submitting identical predictions a filter was added to make sure submitted predictions were not more similar than TMscore>0.8 using MM-align (Mukherjee and Zhang, 2009).

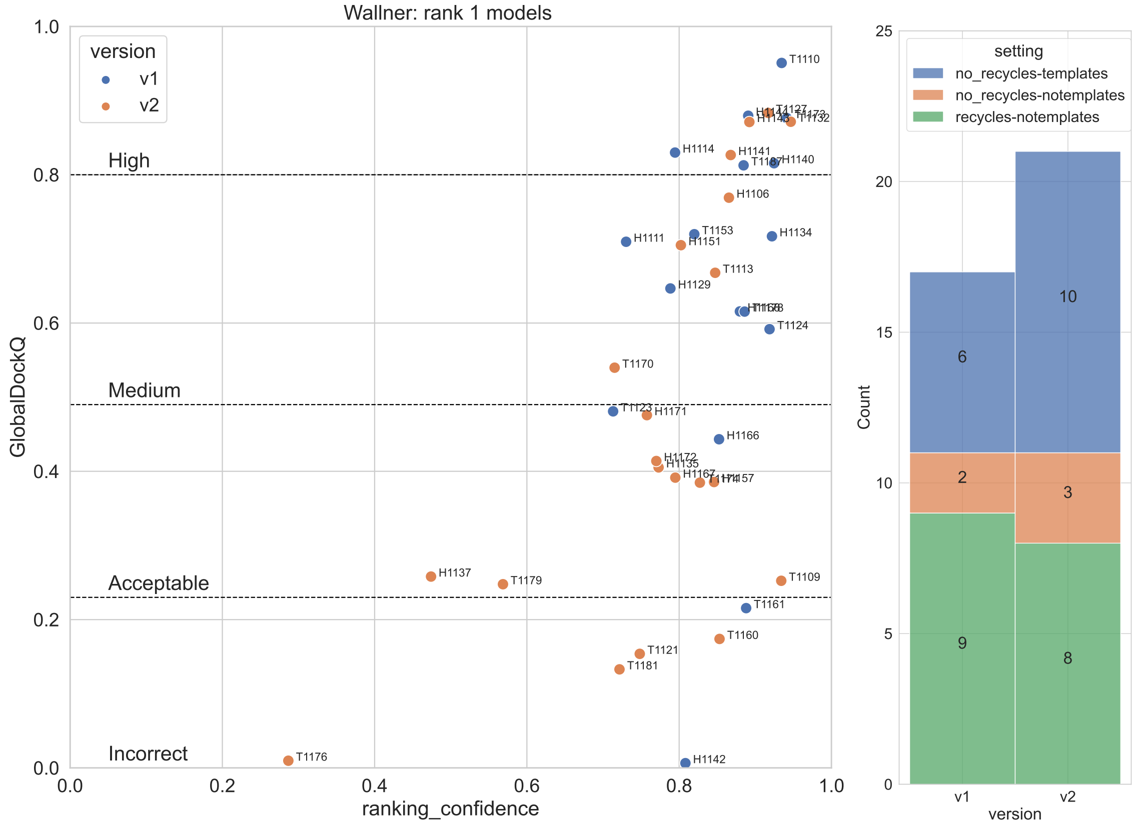
2.3 Multiple Sequence Alignment

The input multiple sequence alignments and template search were generated by the baseline method *NBIS-AF2-multimer*, and were used as is, to allow a direct comparison of the added value of the sampling approach. The input data was made available during CASP15, and are still available, at the following url: `http://bioinfo.ifm.liu.se/casp15/`. The sequence searches were made using the `--db_preset full_dbs` flag with the following databases:

- Uniclust30 (Mirdita *et al.*, 2017) version: `UniRef30_2021_03`
- Uniref90 (Suzek *et al.*, 2015) from April 22, 2022.
- Uniprot, TrEMBL, SwissProt, from April 22, 2022.
- BFD database (Steinegger and Söding, 2018)
  `*.ffindex MD5: 26d48869efdb50d036e2fb9056a0ae9d`
- Mgnify version: `2018_12`
- PDB from May 2, 2022.

3 Results and Discussion

To analyze our performance in CASP15, we used an updated version of DockQ (Mirabello and Wallner, 2023), that given a chain mapping, calculates a global DockQ score by averaging the DockQ (Basu and Wallner, 2016) score for each interface weighted by the size of the interface. This strategy was also employed by the CASP15 assessors (Studer, personal communication). The chain mapping routine in QS-score (Bertoni *et al.*, 2017) was used to determine the optimal chain



(a) rank 1

(b) origin

Figure 1: Quality and score for rank 1 models by the Wallner group

mapping. Compared to other scores to assess performance like TMscore from MMalign (Mukherjee and Zhang, 2009), DockQ focuses more on the interfaces and is stricter in penalizing incorrect interfaces. In addition, if a model is wrong the DockQ score will be close to zero, while TMscore can have 0.5 if one subunit in a dimer is correct.

The quality as measured by global DockQ, as well as the corresponding `ranking_confidence` for our first ranked CASP15 predictions for each target, are shown in Figure 1a. The average DockQ score is 0.55. Out of the 38 multimer targets, 10 were of high quality, 11 of medium quality, and 11 of acceptable, and only six were incorrect, out of which four are borderline to acceptable. This a remarkable result considering the difficulty of the targets and something one could only dream about a year ago. However, the correlation between the `ranking_confidence` and actual quality is only 0.57, which could indicate that there is room for improvement in terms of quality assessment. But an alternative explanation could also be that our current assessment scheme using one reference native state could be questioned. It is clear that the reference is *one* state, but it is not guaranteed that it is the *only* state. There are several cases of this in this CASP, T1109, and T1110 are two states, where one is the WT and the other is a single point mutation that alters the conformation, T1121 is a DNA nuclease that has at least an open and closed conformation (the reference structure in CASP15).

For the 38 multimer targets, 21 targets originate from v2 and 17 from v1, see Figure 1b. In terms of the different settings, 16 targets are from using templates, 17 from the increased recycles without templates, and 5 from no templates with default recycles. Interestingly, despite the larger number of targets with rank 1 models originating from v2, the number of medium and high-quality models are clearly over-represented by models that originate from v1, 13, and 8, for v1 and v2, respectively. In fact, only two models from v1 are deemed incorrect.

3.1 Comparison to other CASP15 groups

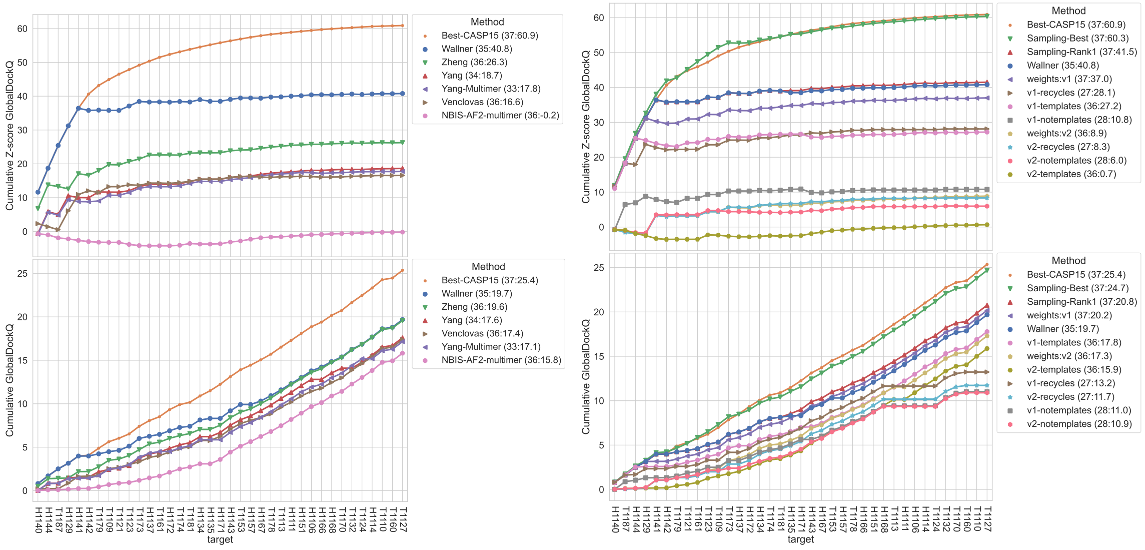
Performance to other CASP15 groups was measured by calculating Z-scores using DockQ for each group, i, and target, j:

$$Z_{i,j} = (DockQ_{i,j} - \langle DockQ_j \rangle)/\text{std}(DockQ_j)$$

where $\langle DockQ_j \rangle$ and std($DockQ_j$) are the average and standard deviation $DockQ$, respectively, for target j. The Z-score summed over each target is shown in Figure S2. However, to avoid a potential with Z-scores that poor models could obtain high Z-scores, in the sum, targets with no correct prediction (DockQ>0.2) by any group were excluded. For CASP15, it was only target T1176 for which no group obtained a correct prediction and that was filtered out. Thus, the total number of targets is 37. Opposite to CASP standard negative Z-scores were not set to zero, to better reflect the overall distribution of quality scores.

It is interesting to compare Wallner to the NBIS-AF2-multimer group since the input in terms of MSA and templates are the same for these two groups and the difference is in the amount of sampling and how the sampling is performed. NBIS-AF2-multimer is running AlphaFold multimer v2 with standard 25 models, while Wallner is using AlphaFold with the improved sampling protocol described in Methods. NBIS-AF2-multimer performs as the average group with a sum of Z-score close to zero (-0.2), this makes sense since almost every group is using AlphaFold in one way or another for their predictions. The Wallner group, on the other hand, has a sum of Z-score above 40 (40.8), and ends up at the very top of the table, clearly higher than the second-ranked group Zheng (26.3). This is great news since the Wallner method is completely automated and easily available as an update to the existing AlphaFold code.

To analyze the per-target contribution in more detail, the cumulative Z-score and DockQ scores were calculated by first ordering the targets by the maximum obtained Z-score, before calculating



(a) CASP15 groups

(b) Different settings

Figure 2: Cumulative Z-score (top) and DockQ (bottom) for rank 1 models (a), and for different settings, combinations, and other method (b). Targets are ordered by the highest Z-score. Group names are ordered by the corresponding sum, shown after the name together with the number of targets as (#targets:sum).

the cumulative sum in that target order. This ordering of targets can be seen as a measure of target difficulty, where the targets with larger Z-scores are difficult since not many groups performed well for these targets. The cumulative sums for the top groups, the baseline method NBIS-AF2-multimer, and the best possible prediction in CASP15 (Best-CASP15) are shown in Figure 2a. From the per-target analysis, it is clear that the main reason for the high sum of Z-score obtained by Wallner is the result of outstanding prediction for five targets: H1140, H1144, T1187, H1129, and H1141 (see section 3.3 below for a detailed discussion of these targets). These five targets contribute with 36.4 units to the total Z-score sum of 40.8.

In addition, the cumulative sum of DockQ, Figure 2a bottom panel, reveals that in terms of actual quality, the difference between Wallner and the second-ranked group Zheng is only very minor, 19.7 vs 19.6, respectively, or 0.56 vs 0.54 in the average. Indicating, that besides the five outstanding targets, the Zheng group is actually better for the other targets, which can be seen by the cumulative sum for Wallner converging to the Zheng group after an initial lead. Still, the difference between Wallner and NBIS-AF2-multimer is significant, 19.7 vs 15.8, respectively, for the sum, or 0.56 vs 0.42, respectively, for the average DockQ.

3.2 Comparing different settings

To analyze the possible reasons for the improved prediction. Cumulative Z-scores and sum of DockQ were calculated for the different settings (Table 1) used by the Wallner method, see Figure 2b. Additional methods corresponding to the best modeled generated in the sampling, *Sampling-Best*, and the rank 1 from the sampling, *Sampling-Rank1* are also included. The *Sampling-Rank1* is identical to Wallner rank 1, but with the two targets missing from the Wallner prediction added. Furthermore, the *Best-CASP15* and Wallner method are included as references. The Z-scores were calculated using the means and standard deviations from CASP15 predictions only, to make them comparable to the previously calculated Z-scores.

The *Sampling-Best* is on par with *Best-CASP15*, meaning that the pool of models generated by the Wallner method contains at least one model with similar quality as the best model submitted to CASP15, see Figure 2b. The fact that *Sampling-Rank1* is lower (0.56 vs 0.66 average DockQ) shows that there is room for improvement in selecting better models from the pool of generated models.

However, the most interesting result, is that *weights:v1*, using the initial version of the multimer neural network weights, performs almost as well as *Sampling-Rank1*, which includes all settings. Using *weights:v1* is much better than using *weights:v2*, with sum of Z-score 40.7 vs 8.8, and sum of DockQ, 20.2 vs 17.3, corresponding to average DockQ of 0.53 and 0.46, respectively. In fact, the sole reason for the success of the Wallner method can be attributed to sampling with v1 weights, while the v2 weights seem much less susceptible to improvement through sampling. This is actually something we noted in our previous study as well (Johansson-Åkhe and Wallner, 2022), but it is now also demonstrated in the blind testing provided by CASP.

While v2 seems to perform better than v1 in the absence of sampling, v1 seems to explore the conformational landscape in a more unbiased way. The major difference between v1 and v2 is the addition of a clash term penalty in the loss function when training v2. It is likely that this change has made the network more stringent and less explorative. Making an analogy to the case of structural refinement where it is often beneficial for sampling purposes to use a soft repulsive clash term, to avoid rejecting structures with minor clashes that are otherwise correct.

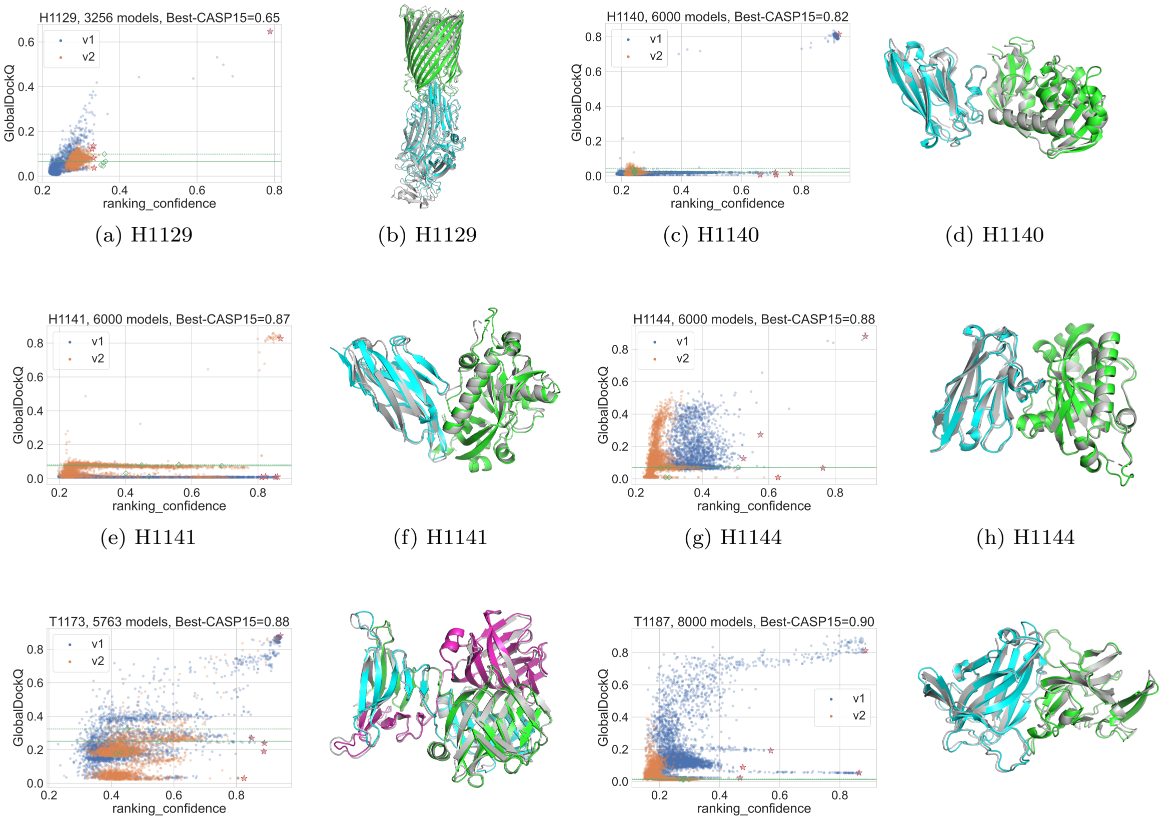
3.3 What went right?

Our strategy in CASP15 was using AlphaFold with the improved sampling strategy we developed. The tremendous success demonstrated above clearly shows that sampling is the way forward. By comparing the per-target performance with NBIS-AF2-multimer, which was using identical input, we can see which targets improved over the baseline, see Figure S3. The targets, H1129, H1140, H1141, H1144, T1173, and T1187 showed improvements with +0.6 in DockQ, while T1123 and H1134 improved 0.4. Three targets, H1167, H1168, and T1124 got worse and those will be discussed below.

The sampled model ensembles visualized as the ranking_confidence score against the DockQ score and the predicted models superimposed on the reference are shown for the successful cases in Figure 3. Of the six success cases, four are a direct consequence of using v1 weights (H1129, H1140, H1144 , T1187), for T1173, the first ranked models were generated by v1, but there are models of similar quality generated by v2, and for H1141 the first ranked model is generated by v2. Even though the choice of network weight clearly influences the results it is impossible a priori to know which network weights to use, thus all sets of network weights have to be sampled. As demonstrated by the successful cases, it is possible to improve both the sampling and selection of high-quality models. Importantly, the fact that the ranking_confidence score improves from relatively low scores (<0.4) for the baseline method to scores >0.8 after sampling indicates that the method is not only able to sample high-quality models but also to identify them as such.

3.4 What went wrong?

To pinpoint the targets where our performance was sub-optimal we compared our per-target performance with the performance of the best overall and best rank 1 models not submitted by us to CASP15, we also added the performance of the best possible model generated through the sampling, see Figure S4. In principle, there are two types of mistakes, either the scoring function is not able to select the best model, or the sampling is not able to generate good models. In addition, it is also possible that both these mistakes occur at the same time. We classify the target as having a scoring problem if the ΔDockQ>0.2 between the selected and best-sampled model, in a similar manner, we classify targets as having a sampling problem if the ΔDockQ>0.2 between best-sampled model and the best model in CASP15. By using these definitions, six targets were classified as having potential scoring problems and six targets as having problems with sampling, see Table 2.



(a) H1129 (b) H1129 (c) H1140 (d) H1140

(e) H1141 (f) H1141 (g) H1144 (h) H1144

(i) T1173 (j) T1173 (k) T1187 (l) T1187

Figure 3: Successfully modelled targets from CASP15 illustrated by ranking_confidence vs DockQ and structural superposition on reference structure (grey). The ranking_confidence vs DockQ are separated based no weight, the red stars show the submitted predictions by Wallner, green diamonds show the submission by NBIS-AF2-multimer, where the solid green line is the first ranked, and the dashed line the best submitted.

target	**Res**	**Stoichiometry**	**DockQ sampled**	**DockQ rank1**	**DockQ best5**
scoring problem					
T1109	227	A2	0.81	0.25	0.76
T1121	381	A2	0.53	0.15	0.32
T1124	384	A2	0.82	0.59	0.81
T1161	48	A2	0.68	0.22	0.68
H1167	560	A1B1C1	0.66	0.39	0.63
H1168	567	A1B1C1	0.83	0.62	0.62
sampling problem				**DockQ CASP**$_{\text{best}}$	
H1137	3939,	A1B1C1D1E1F1G2H1I1	0.28	0.66	0.26
H1142	347	A1B1	0.08	0.31	0.03
T1160	48	A2	0.20	0.71	0.18
H1171	366	A6B1	0.52	0.75	0.48
H1172	366	A6B2	0.48	0.83	0.41
T1179	261	A2	0.43	0.81	0.27

Table 2: Target classified as having scoring or sampling problems. **Res** is the number of residues, **DockQ sampled** is the best DockQ generated. **DockQ rank1** is the DockQ for rank 1, **DockQ best5** is the best DockQ of the five submitted models, **DockQ CASP**$_{\text{best}}$ is the best DockQ achieved by any group in CASP.

3.4.1 Scoring problem

Targets with scoring problems fail to rank the best model at rank 1. However, it turned out that in all cases, there is at least an acceptable model (DockQ>0.2), and often even better, considering the best out of the five submitted predictions, see Table 2. It is often small differences in score, but a large difference in model quality. For T1124, the five submitted models have ranking_confidence between 0.91 and 0.92, while the DockQ is in the range 0.59-0.81, see Figure S5a-c, and for H1167, the top three predictions have ranking_confidence between 0.75-0.80, while the DockQ is between 0.39-0.63, see Figure S5d-f. One should also bare in mind that the ranking_confidences are predicted by 10 different neural networks, network model 1-5 for v1, and v2, respectively, and it is possible that the scores are not perfectly calibrated even though they try to predict the same quantity.

Below we discuss a couple of targets that, at first glance, seem to suffer from a scoring problem, but in reality, seem to sample different conformations.

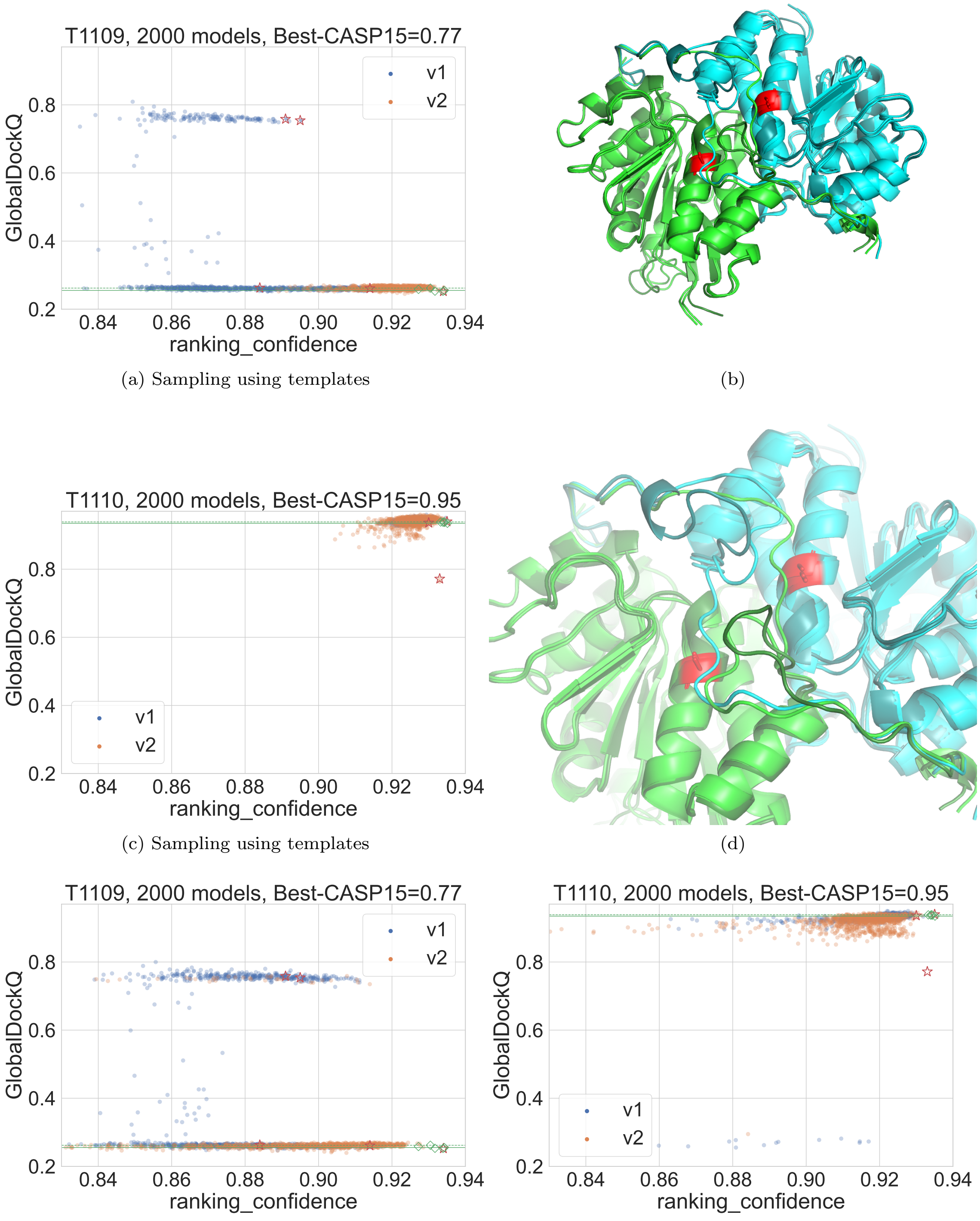
3.4.2 T1109 and T1110

Target T1109 and T1110 is a 227-residue homo dimer of isocyanide hydratase from *Ralstonia solanacearum.* T1110 is the wild-type (WT) , and T1109 is a D183A mutation. The mutation causes the C-termini to make a 360-degree turn and alters the C-termin interaction from *intra-chain* in the WT to *inter-chain* in the mutant by swapping the interaction with the C-terminal tails. For the WT, target T1110, the whole sampled population is correct and of high-quality DockQ>0.9, Figure 4c. For the mutant, T1109, there are two populations, the larger is actually the WT conformation, while the smaller generated by v1 weights contains the correct structure for the mutant. The ranking_confidence scores are higher for the WT population ≈0.93 vs ≈0.89 for the mutant population. Representative structure of the WT (rank 1) and mutant cluster (rank 3) are superimposed on the mutant reference structure shown in Figure 4b. Rank 3 is the one that follows the reference structure (in darker colors).

It is interesting to compare the ranking_confidence score for T1109 and T1110, see Figure 4a,c. As the sequences only differ by a single point mutation, the input data is virtually the same for both targets. In addition, since the scores were very high for these targets, only the settings using templates were used to generate these models, i.e. *v1-templates*, and *v2-templates* from Table 1. The structural templates are very similar to the WT, including the conformation of the C-termini, explaining why the prediction for T1110 is almost perfect and why the largest cluster for T1109 is also close to the WT. The the ranking_confidence score distribution for v2 is very tight for both T1109 and T1110, while the same distribution is wider for v1 in the case of the mutant T1109, see

Figure 4a,c. This indicates that v2 relies more on the template compared to v1, as it is only v1 that is able to sample outside the template distribution for the mutant.

To verify this hypothesis, after CASP15, we rerun targets T1109 and T1110 using the no templates settings. Indeed, without templates, the population for the mutant conformation is larger and also contains models generated by v2, see Figure 4e,f. Again, this underlines the importance of running with different initial settings to maximize the diversity in the sampling.



(a) Sampling using templates

(b)

(c) Sampling using templates

(d)

(e) Sampling without using templates

(f) Sampling without using templates

Figure 4: (a) T1109 ranking_confidence vs DockQ, there is some overplotting the highest v1 score is 0.92, (b) superposition of model 1 and model 3 on the reference structure colored by chain. The reference structure is in darker green and cyan. The mutation D183A is highlighted in red. (c) T1110 wild-type ranking_confidence vs DockQ, (d) Zoom in on the key difference in loop conformation of the C-termini. Colors as in (b). (e) T1109, without templates, ran after CASP15. (f) T1110, without templates, ran after CASP15.

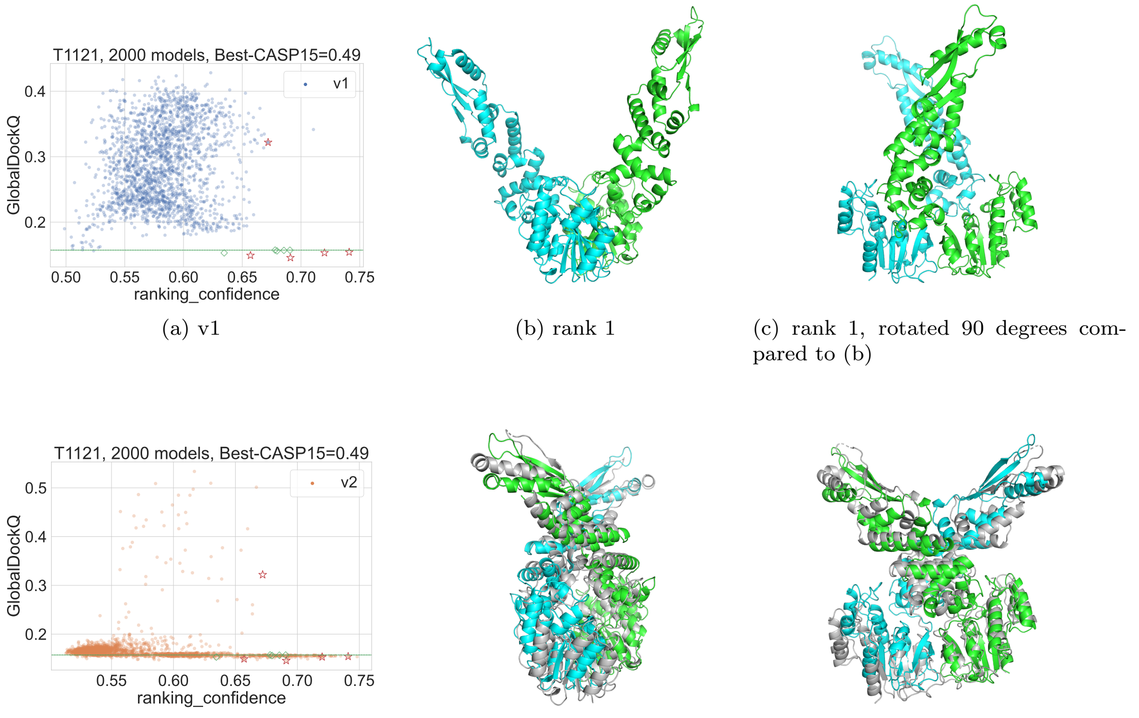
3.4.3 T1121

Target T1121 is a DNA nuclease JetD from *Pseudomonas aeruginosa* 381-residue dimer, since it binds and cleaves DNA, it most likely has several conformations. The structure used as the reference is a closed autoinhibited conformation (Deep *et al.*, 2022). The best-sampled conformations have ranking_confidence > 0.7, the scores from v2 are slightly higher than the scores from v1, see Figure 5a,d. If the closed autoinhibited structure is used as a reference (pdb:7til), the best models have a DockQ of 0.33, acceptable quality, were generated by v1, and were the four highest scoring cluster overall.

On the other hand, the rank 1 model by our method generated by v2 has a DockQ of 0.0. The overall shape of the monomers is modeled correctly but the relative orientation of the subunits is different compared to the reference model, forming a relatively open conformation, see Figure 5b, compared to the closed conformation of the reference structure. Interestingly, Deep *et al.* (2022) proposed a model of the open active state, which is actually very similar to the rank 1 model. Thus, one could speculate that the rank 1 model is actually not incorrect, but simply represent the open active conformation. The fact that our sampling method seems to be able to generate and select both these conformations indicates that the method, indeed, could be used to generate conformational ensembles for proteins with several states.

3.4.4 Clustering problem: H1168

H1168 is a three-chain protein, where the main problem is to predict the interface between B:C. This target illustrates a problem with our filtering scheme to avoid submitting too similar targets using MMalign. According to MMalign, there are only two clusters, and we only submitted two models for this target, see Figure S6. The TMscore for rank 1 against the model with the best DockQ is 0.88 using the default setting for the length-dependent normalization factor (d0=8.37Å). However, forcing d0 to be 3.5Å the TMscore drops to 0.68. This shows that it is important to control for the d0 when using MMalign on larger complexes. In the future, we will use the updated version of DockQ (Mirabello and Wallner, 2023) to compare complexes.



(a) v1

(b) rank 1

(c) rank 1, rotated 90 degrees compared to (b)

(d) v2

(e) rank 4 from v1

(f) rank 4, rotated 90 degrees to (e)

Figure 5: ranking_confidence vs DockQ for version v1 (a) and v2 (d). Structural models of rank 1 (b,c), a potential open active conformation, and rank 4 (e,f), a closed conformation that is similar to the reference structure.

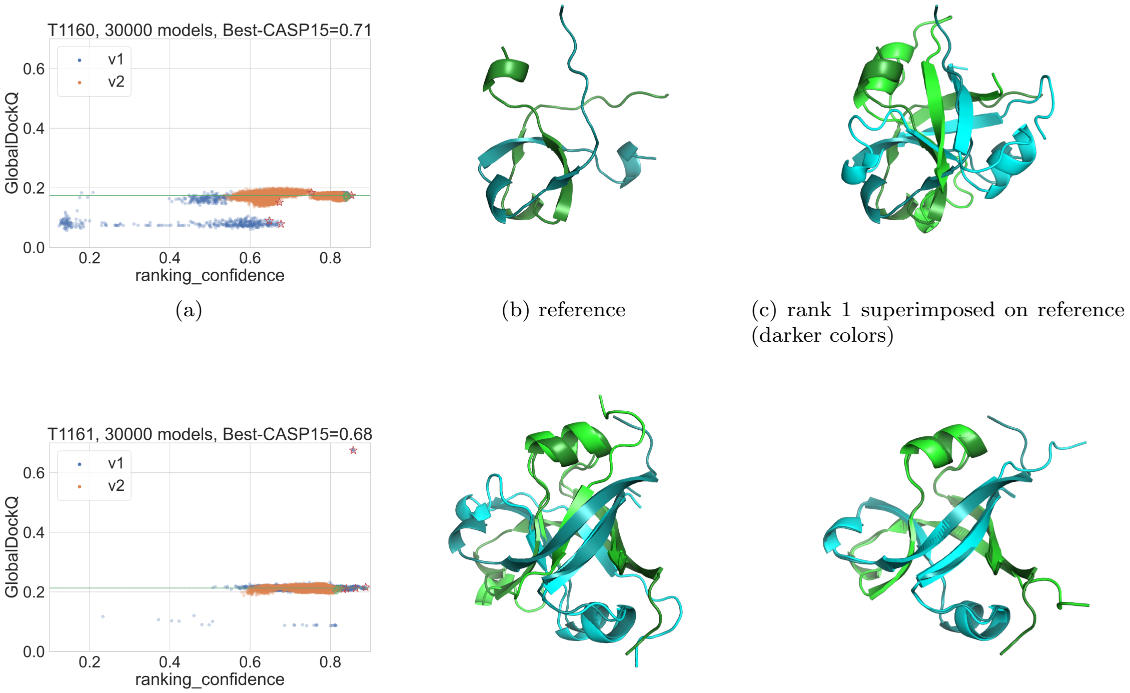
3.5 Sampling problem

The criteria for classifying a target as having a sampling problem was that any group in CASP submitted a better model than was generated by the massive sampling in the Wallner method. In general, AlphaFold does not work as well for large assemblies, which is understandable as it folds everything from scratch. In addition, the massive sampling is hampered by the computational time to generate even a single model, for some targets, e.g. H1137, as long as 3 days on an Nvidia V100.

Here, there is clear room for improvement by folding and assembling in a stepwise manner as well as using templates and symmetry for multimer interactions.

3.5.1 T1160 and T1161

T1160 and T1161 are small, 48 amino acids, dimeric, ancient protein reconstructions, the sequences are similar (differ by three amino acids) but the crystallization conditions are different, which leads to different structures. Of course, it is difficult to take crystallization conditions into account, but one could at least hope that the correct topology could be generated through the sampling. Since the sequences are reconstructions, there are no homologous sequences in the multiple sequence



(a)

(b) reference

(c) rank 1 superimposed on reference (darker colors)

(d) T1161

(e) rank 1, DockQ=0.20

(f) rank 3, DockQ=0.68

Figure 6:

alignment, but there are several templates with the wrong topology. 30,000 models were sampled for each target, the most sampling performed for any of the CASP15 targets. Despite the massive sampling, the best models for T1160 only have DockQ≈0.20, Figure 6a. However, for T1161, there are actually three models with DockQ≈0.68, Figure 6d, and one of these were submitted as prediction rank 3, Figure 6f. The fact that a correct prediction for T1161 was generated in only 3 out of 30,000 attempts (0.01% success rate) indicates that even more sampling could be needed for T1160. Indeed, the reference structure for T1160 is certainly a lot less folded than the reference structure for T1161, Figure 6b,f, which could be difficult to sample. Comparing the ranking_confidence score distribution for T1160 and T1161, they are actually quite similar Figure 6a,d, with a tight cluster of low-quality models around ranking_confidence 0.6-0.8. The v1 models are slightly more explorative, even more so for T1160, but it is for T1161 that three of the models from v1 turned out to be correct. There are no structures between DockQ 0.2 and 0.6, which has to do with the fact that the protein is small and intertwined, and that any structure is either wrong or right. From a sampling perspective, this is also problematic since there is also no guidance toward the correct state.

3.5.2 H1171 and H1172

Targets H1171 and H1172 contain two proteins from the Recombination UV complex, RuvA, and RuvB. RuvB is an ATPase that forms hexamers, and RuvA is a 48-residue DNA-binding domain. H1171 has one RuvA bound to RuvB (A6B1 stoichiometry), while H1172 has two RuvA bound to RuvB (A6B2 stoichiometry). RuvB is a symmetric hexamer, but since there is no way to enforce symmetry in AlphaFold, the overall predicted structures are slightly asymmetric, resulting in sub-optimal model quality scores Figure S7a,d and superpositions, see Figure S7b,e. However, the 1:1 interaction between RuvA and RuvB is almost perfectly predicted, see Figure S7c,f. In the A6B2 case, the binding is predicted between wrong subunits, but it is not surprising since the reference structure has the two RuvA subunits binding two neighboring subunits asymmetrically, see gray RuvA subunit next to the blue in Figure S7e, while the prediction is binding symmetrically, the orange subunit at the bottom of Figure S7e.

The ranking_confidence from v1 and v2 is clearly showing different behaviors, see Figure S7a,d. While the DockQ scores for the generated models are similar, the ranking_confidence from v2 is consistently +0.3 higher than from v1. This is also a case where the sampling does not help at all since all sampled models are worse than the baseline. Again, this demonstrates that there is room for improvement in sampling large oligomeric structures.

4 Conclusions

The results by the Wallner method in CASP15 demonstrate that sampling by running AlphaFold with dropout activated at inference and using different settings is a relatively simple approach to obtain improved performance. Compared to running the AlphaFold multimer baseline (NBIS-AF2-multimer), there is virtually no performance loss, instead, there is a massive gain in performance for several targets (+0.6 in DockQ), with the mean DockQ increasing from 0.43 to 0.56. Of course, the sampling is time-consuming and should only be performed if the ranking_confidence is low for the baseline method (<0.7)

We observed that multimer version 1, v1, of the neural networks benefit much more from sampling compared to v2. This is interesting since the v1 weights have been accused of producing highly clashing models in the past. This might still be true, but since these clashing models do not receive a high ranking_confidence score, they are filtered out in the sampling.

The sampled models seem to contain different conformational states, as exemplified by T1121, where v2 predicts (and the baseline method) an open conformation, but v1 samples the closed conformation, which happened to be the reference structure in CASP15. Another example is the sampled models of single point mutation T1109, which contains both the WT and mutant conformational states.

Large assemblies are challenging for AlphaFold as templates are only used for monomers, and there are no symmetry constraints to limit the search space, thus the relative orientations of all subunits in a multimer structure have to be assembled from scratch. It should be relatively straightforward to include multimer templates, which would have

Funding

This work was supported by the Wallenberg AI, Autonomous System and Software Program (WASP) from Knut and Alice Wallenberg Foundation (KAW), Swedish Research Council grant, 2020-03352, The Swedish e-Science Research Center, and Carl Tryggers stiftelse för Vetenskaplig Forskning, 20:453. The computations were performed on resources provided by KAW and NSC (Berzelius).

References

Alamo, D. d., Sala, D., Mchaourab, H. S., and Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, **11**, e75751.

Basu, S. and Wallner, B. (2016). DockQ: a quality measure for protein-protein docking models. *PloS one*, **11**(8), e0161879.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports*, **7**(1), 10480.

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, **13**(1), 1265.

Cramer, P. (2021). AlphaFold2 and the future of structural biology. *Nature Structural & Molecular Biology*, **28**(9), 704–705.

Deep, A., Gu, Y., Gao, Y.-Q., Ego, K. M., Herzik, M. A., Zhou, H., and Corbett, K. D. (2022). The SMC-family Wadjet complex protects bacteria from plasmid transformation by recognition and cleavage of closed-circular DNA. *Molecular Cell*, **82**(21), 4145–4159.e7.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Johansson-Åkhe, I. and Wallner, B. (2022). Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Frontiers in Bioinformatics*, **2**, 959160.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, pages 1–11.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21), 2722–2728.

Mirabello, C. and Wallner, B. (2023). DockQ2: improved quality measure for protein-protein docking models. *biorxiv*, **11**(8), e0161879.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, **45**(D1), D170–D176.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, **19**(6), 679–682.

Mukherjee, S. and Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Research*, **37**(11), e83–e83.

Roney, J. P. and Ovchinnikov, S. (2022). State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Physical Review Letters*, **129**(23), 238101.

Steinegger, M. and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, **9**(1), 2542.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6), 926–932.

Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L., and Kern, D. (2022). Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv*, page 2022.10.17.512570.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, **57**(4), 702–710.