AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling

Björn Wallner[1*]

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

December 21, 2022

Abstract

The AlphaFold neural network model has revolutionized structural molecular biology with unprecedented performance. We demonstrate that stochastically perturbing the neural network by enabling dropout at inference generates a more diverse set of models, from which better models can be selected using the self-assessment ranking score in AlphaFold. The method was benchmarked in CASP15 and ranked at the top in the protein assembly category. The method should be useful for anyone interested in modelling multimeric structures.

The unprecedented accuracy of AlphaFold (1) has transformed the field of computational and structural biology. It is now possible to achieve highly accurate predictions on par with experimentally determined structures. AlphaFold has rapidly become the go-to method for protein structure prediction for both monomers and multimer prediction pipelines (2; 3).

A key to the success of AlphaFold is its ability to assess the accuracy of its own predictions. AlphaFold estimates the per-residue accuracy using the predicted LDDT (4) (pLDDT). In addition, it also calculates a predicted TM-score (5) (pTM) from the predicted aligned error (PAE) matrix (1). The correlation for pLDDT and pTM to its actual values are 0.76 and 0.85, respectively (1), and this correlation is maintained even for high-quality predictions. For multimer prediction, an inter-chain predicted TMscore (ipTM) is calculated from the PAE of the inter-chain distances (1).

Given enough evolutionary-related sequences, AlphaFold predicts monomers with very high accuracy even without using structural templates (1). However, for multimers, this is not necessarily true (6), the evolutionary signal constraining multimers is much weaker, and for these cases, more sampling might improve the prediction. The need for more sampling was also the reason why the default number of sampled models in AlphaFold-multimer was increased from 5 in version 1 (v1) to 25 in version 2 (v2) (7). Furthermore, predicting transient interactions or interactions with flexible binding partners, such as short peptides or disordered regions, requires even more sampling to achieve optimal performance (8).

*To whom correspondence should be addressed (bjorn.wallner@liu.se

Version	Templates	Dropout	Recycles
multimer_v1	Yes	Yes	3
multimer_v1	No	Yes*	3
multimer_v1	No	Yes*	21
multimer_v2	Yes	Yes	3
multimer_v2	No	Yes*	3
multimer_v2	No	Yes*	9

Table 1: Different settings of AlphaFold used in AFsample, *Version* refers to the version of the multimer neural network weights, *Templates* refers to if structural templates were used or not, *Dropout* refers to if dropout was enabled, *Recycles* refers to how many recycles was used (default 3).
*No dropout in structural module

For complex cases, simply increasing the number of sampled models might not be enough if the evolutionary constraints have trapped the prediction in a local minimum in the conformational landscape (9) or the if the evolutionary constraints are weak. In such cases, increasing the number of times, the prediction is recycled in the network can improve performance (10). Another option is to randomly perturb or alter the input MSA, which has been shown to enable better sampling of the conformational landscape and prediction of multiple conformational states (11; 12). An alternative way to achieve more diversity among the generated models is to enable the dropout layers in the neural network (8; 10). The dropout layers in a neural network are commonly used only when training the networks to cause them to learn different redundant solutions to the same problem by stochastically dropping some of their weights (putting them to zero). The dropout rate in the AlphaFold network is 10%-25%, depending on the network module. By activating these layers at inference, the network will naturally sample the uncertainties (13) in the structure prediction, and the structural diversity of the sampled models will be increased.

Here, we present AFsample that significantly improves over AlphaFold2-multimer (AF2-multimer) baseline. The method was the most successful in CASP15 for multimer prediction (Figure 1a). It improved the average DockQ (14) from 0.41 for AF2-multimer (NBIS-AF2-multimer) to 0.55, with eight targets showing considerable improvements with more than 0.4 DockQ units, essentially going from incorrect to high-quality predictions (Figure 1b).

In short, AFsample generates a large pool of models using different settings (Table 1). The self-assessment ranking score (`ranking_confidence`), which for multimer is a linear combination of the predicted interface TMscore (ipTM) and the predicted TMscore (pTM), $0.8\text{ipTM} + 0.2\text{pTM}$, is used for selection. Six different settings were used, and they all involved using dropout to increase the model diversity. Both `v1` and `v2` of the multimer neural network weights and increased number of recycles were also utilized to increase the diversity further. The number of recycles for `v1` and `v2` were optimized in a previous study of peptide-protein predictions (8). Four settings involved selective dropout, where we only activated dropout in the Evoformer part of the network and not in the structural module. Despite the name, the bulk of structure prediction in AlphaFold is not performed in the structural module but in the much more extensive Evoformer network. In a previous study, we observed an improved corre-

lation between `ranking_confidence` and actual DockQ using selective dropout and no dropout in the structural module (8). For each setting, on the order of 1,000 models were generated per setting for a total of 6,000 models per target.

It is clear the sole reason for the improved performance of AFsample is improved sampling as AFsample was using identical MSAs as the AF2-multimer baseline. The need for sampling is illustrated further by target H1144 (Figure 1c and d). Target H1144 is a nanobody interaction, nanobodies and regular antibody interactions are particularly challenging to model since the interaction with the antigen is highly specific, and the exact shape and location of the epitopes can vary significantly depending on the antigen (15). Here, only three out of 6,000 models obtained a `ranking_confidence>0.8`, out of which all were of high quality with a `DockQ>0.8`. In fact, only five high-quality models could be found in the whole set of 6,000 models sampled. Thus, without substantial sampling, most likely, no high-quality model would have been generated at all.

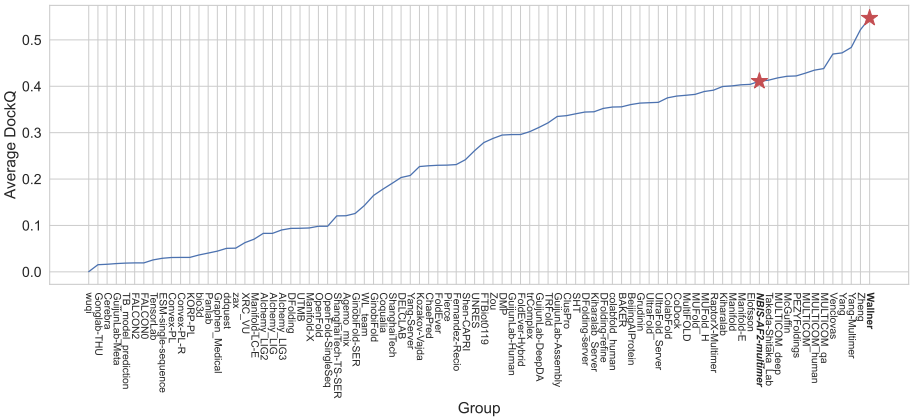
The other example is T1187o (Figure 1e and f). Target T1187o is a dimer of the Uniprot ID `Q94EW1`. The AlphaFold prediction of the monomer is almost perfect (`https://alphafold.ebi.ac.uk/entry/Q94EW1`). However, the model of the dimer is completely wrong with a DockQ very close to zero (Figure 1e). However, using sampling, it is possible to generate several high-quality models (38/6000). Still, selecting the best possible model was not straightforward as one-third of the models with a `ranking_confidence>0.8` actually had a wrong domain orientation.

AFsample requires more computational time than AF2-multimer baseline, as it generates 240x models, and including the extra recycles, the overall timing is around 1,000x more costly than the baseline. Thus, a good strategy is to first run the AF2-multimer baseline prediction and only use more sampling if needed. A good criterion for this is to use a `ranking_confidence>0.8` and require at least a couple of structures at this level of confidence. However, given that AlphaFold inference in many cases is relatively fast, a minute or less per model, a couple of days one a single GPU to get high-quality results for a difficult target is not an enormous amount of computational time.

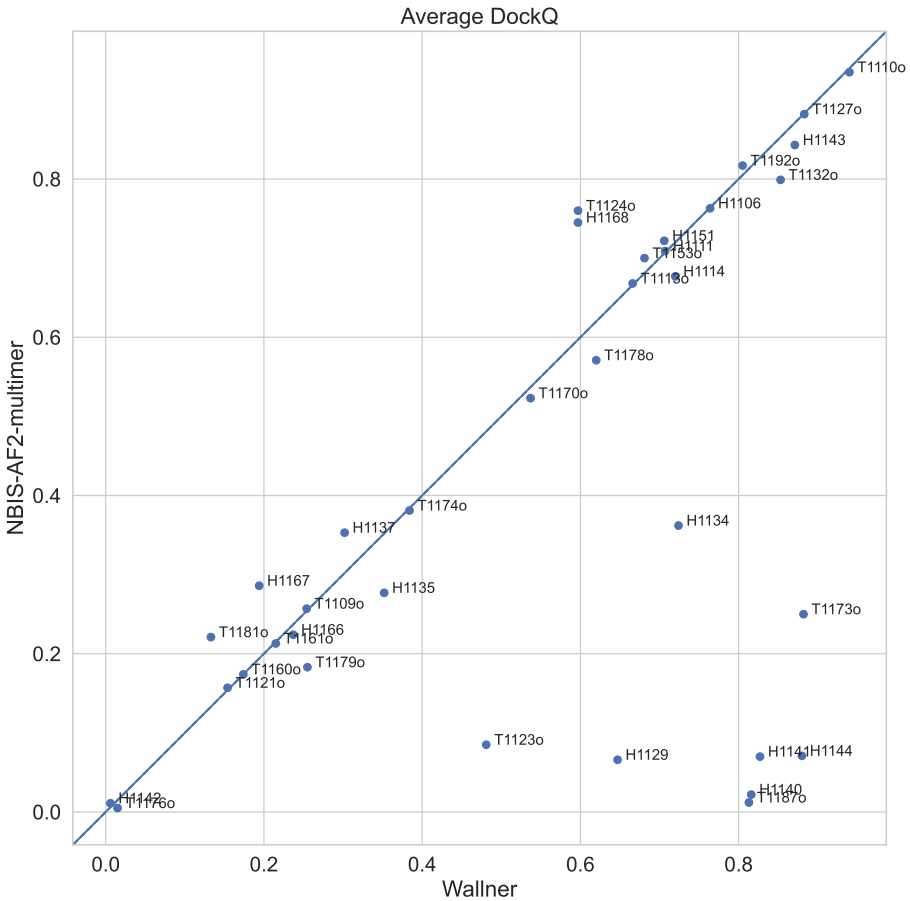
The results from CASP15 demonstrate that the best way to model multimeric protein assemblies today is to use AFsample. However, it is very likley that sampling will improve the performance of AlphaFold in any setting. Dynamics, flexibility, or simply the sheer complexity of large molecular assemblies will all require more sampling. The success of sampling relies heavily on the excellent internal scoring function in AlphaFold, which so far has stand up to the test.

Acknowledgments

This work was supported by a WASP-DDLS grant from Knut and Alice Wallenberg (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 the Swedish National Infrastructure for Computing (SNIC), Knut and Alice Wallenberg (KAW) (Berzelius), and Linköping University (LiU) at the National Supercomputer Center (NSC) in Linköping.

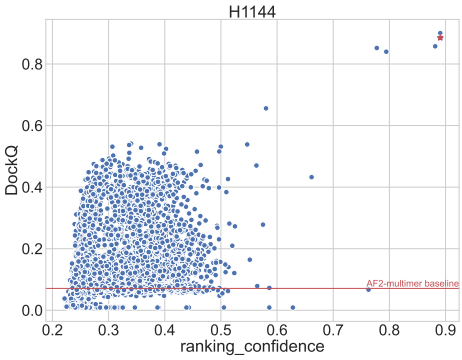


(a) Average DockQ over multimeric targets for all groups in CASP15



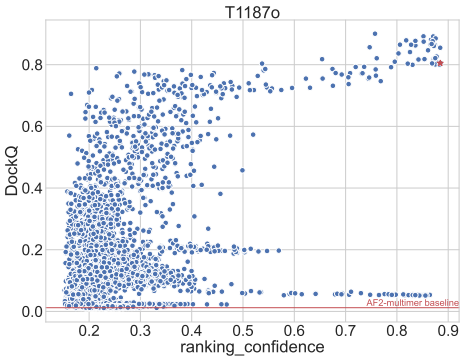
(b) DockQ comparison to AF2-multimer baseline per CASP15 target

Figure 1: Performance in CASP15, continues on the next page...

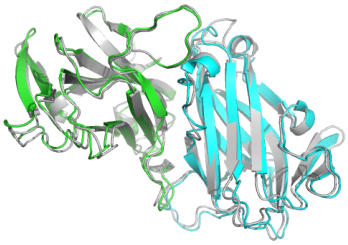


(c) Ranking_confidence score vs DockQ for models sampled for CASP target T1187

(d) CASP target H1144, native chain A(green) and B (cyan), rank 1 prediction in gray, DockQ=0.88



(e) Ranking_confidence score vs DockQ for models sampled for CASP target T1187



(f) CASP target T1187o, native chain A(green) and B (cyan), rank 1 prediction in gray, DockQ=0.81

Figure 1: Example targets from CASP15

References

[1] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11 (2021).

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

[3] Terwilliger, T. C. *et al.* Improved AlphaFold modeling with implicit experimental information. *Nature Methods* **19**, 1376–1382 (2022).

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

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

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

[7] Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2021).

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

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

[10] Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022).

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

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

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

[14] Basu, S. & Wallner, B. Dockq: a quality measure for protein-protein docking models. *PloS one* **11**, e0161879 (2016).

[15] Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies* **8**, 55 (2019).

[16] Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170–D176 (2017).

[17] Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

[18] Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542 (2018).

[19] Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).

[20] Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology* **13**, e1005659 (2017).

Methods

AFsample is available as a command line interface using a modified version of the official AlphaFold release. The modified version is streamlined to produce many models, contains functionality to parallelize the computations to independent jobs, and exposes several internal parameters of AlphaFold to allow using it at its full potential. A script `run_afsample.sh` is provided that reproduce the method that participated in CASP15 under the name `Wallner`.

Filter the output. To enable the generation of thousands of models, the amount of data saved per model was limited to the predicted aligned error matrix (`PAE`), per residues predicted LDDT (`pLDDT`), overall predicted TMscore(`pTM`), predicted interchain TMscore (`ipTM`), and `ranking_confidence`. This reduces the size of the data saved per model by approximately a factor 100. The flag `--output_all_results` will restore the default behavior and output all data structures.

Checkpointing. Checkpointing is made default, it will not recalculate MSAs if they exists and if a model exists it will continue to the next model.

Added functionality. The most important modifications and exposed functionalities are described below:
`--model_preset` Modified to allow using both v1 and v2 models, the following presets are allowed: `multimer_v1`, `multimer_v2`, `multimer` (defaults to v2), `multimer_all` (both v1 and v2)
`--nstruct` the number of structures to output, works both for monomer and multimer protocols and replaces the `--num_multimer_predictions_per_model` which was exclusive to the multimer protocol.
`--max_recycles` the number of times a prediction is recycled in the neural network, default is 3.
`--dropout` enables the dropout layers at inference.
`--dropout_structure_module` if `False` it will not have the dropout layers enabled in the structure module.
`--suffix` option to add a descriptive suffix to each model name to enable output to the same output folder.
`--no_templates` do not use any templates, faster than filter by date which requires parsing all hits.
`--seq_only` will only run the sequence searches to create the MSAs and template hits, useful to prepare input files for larger runs.
`--input_msa` option to input a multiple sequence alignment in STOCKHOLM format.
`--nstruct_start` which structure to start with, useful to split a large job in many smaller by using `--nstruct_start 20` and `--nstruct 21` it will create models `20` and `21`.
`--models_to_use` option to specify which neural network models from `model_preset` to use. i.e `model_1_multimer_v2, model_3_multimer_v2` will only run model 1 and model 3 from `multimer_v2`.

Databases. Sequence databases were downloaded on April 22, 2022, and the PDB was updated May 2, 2022, using the download scripts provided by Deep-

Mind (`https://github.com/deepmind/alphafold/scripts/`). The following version were used:

- Uniclust30 (1) version: `UniRef30_2021_03`
- Uniref90 (2) from April 22, 2022.
- Uniprot, TrEMBL+SwissProt, from April 22, 2022.
- BFD database (3)
  `bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_cs219.ffindex`
  `MD5 hash: 26d48869efdb50d036e2fb9056a0ae9d`
- Mgnify version: `2018_12`
- PDB from May 2, 2022.

All-atom relaxation. In regular AlphaFold each model is constrained relaxed in the Amber99sb force field (4) using openMM (5). To save computational time the all-atom relaxation step is skipped for each model. Instead a script `run_relax_from_results_pkl.py` is provided, that performs the relaxation step for a given result `pickle`. Since none of the scores depend on this step, the relaxation can be performed only for a smaller subset of the models that are high scoring or are selected by some other criteria.

Benchmark.In the CASP15 benchmark both AFsample (`Wallner`) and AF2-multimer baseline was run with exactly the same multiple sequence alignments (MSAs) and templates. The alignments were created with the large database setting: `--db_preset=full_dbs` using the AF2-multimer baseline server. They were made available by the CASP organisers, and these were the MSAs used by the `Wallner` group in CASP15. The DockQ (6) scores for all methods that participated in CASP15 were downloaded from the CASP15 website. In the case of multiple interfaces, DockQ is calculated for each interface and then averaged. The rank 1 models from each method were used to calculate the average DockQ for the multimer targets for each method.

Data Availability

The MSAs and template information used in the CASP15 benchmark is available here: `http://bioinfo.ifm.liu.se/casp15/`

Code Availability

AFsample is free, open-source software (Apache) and is available from here: `http://wallnerlab.org/AFsample`

References

[1] Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170–D176 (2017).

[2] Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

[3] Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542 (2018).

[4] Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).

[5] Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology* **13**, e1005659 (2017).

[6] Basu, S. & Wallner, B. Dockq: a quality measure for protein-protein docking models. *PloS one* **11**, e0161879 (2016).