MassiveFold

Massive sampling with MassiveFold in CASP16-CAPRI

Nessim Raouraoua[1], Marc F. Lensink[1] and Guillaume Brysbaert[1]

[1] *Univ. Lille, CNRS, UMR 8576 – UGSF - Unité de Glycobiologie Structurale et Fonctionnelle, F-59000 Lille, France*

guillaume.brysbaert@univ-lille.fr

***Key:** Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; DeepL:Y; AF:AF2; EMA:Y; MD:N*

The results of CASP15 have shown that increasing the number of predictions while including diversity in the inference process led to a significant improvement for multimer predictions[1]. However, this massive sampling strategy requires access to a large GPU infrastructure to be able to generate the predictions in a short period of time, and is therefore not accessible to all predictor groups.

For CASP16, we used MassiveFold[2] (https://github.com/GBLille/MassiveFold), which allows massively expanding the sampling of structure predictions by optimizing the computing of AlphaFold[3,4] based predictions. It improves the parallelization of the structure inference by splitting the computing on CPU for alignments, running automated batches of structure prediction on GPU, and gathering the results in a single output directory, with a consolidated ranking and a variety of plots. MassiveFold uses AFmassive (https://github.com/GBLille/AFmassive) inference engine, an updated version of AFsample[6] that offers additional diversity parameters for massive sampling. MassiveFold can also use ColabFold[5].

We used a large GPU cluster to generate 8040 predictions for the majority of the targets, submitted our top 5 and made the ensemble of predictions available to all predictors for a CASP16 phase 2 prediction round where they could use our predictions in any way they wanted to submit an updated top 5.

Methods

In collaboration with the IDRIS national french GPU supercomputing center of GENCI for the structure inference and the PLBS/SINBIOS platform for data storage, we generated massive sampling data based on AlphaFold and ColabFold. The inference integrated diversity following mainly Björn Wallner's parameterization used in CASP15[1] but exploiting the three neural network models currently available for AlphaFold for multimers. In addition, the last CAPRI conference in Feb 2024 showed that generating a high amount of structures with default AlphaFold parameters in many instances also leads to an increase in good quality models[7], thus we added this condition.

Therefore, to integrate the maximum number of parameters in a massive sampling strategy but limiting the structures to a reasonable amount, we predicted, **for multimers**, 67 structures per AlphaFold neural network model (5 NN models x 3 NN versions = 15 for multimers) with AFmassive, resulting in 1005 structures per condition for the 6 first sets of parameters listed in **Table 1**, for a total of 6030 structures. To complement them with more diversity in particular with the ColabFold_DB and MMseqs2 approach for alignments, we computed an extra 1005 predictions with ColabFold for an additional two sets of parameters, adding 2010 ColabFold predictions to the total (Table 1).

This results in a total of **8040** structure models for each target, except for the largest ones which would require too many GPU hours. For these, we reduced the number of predictions according to the size of the target.

For monomers, because only one neural network version was published by DeepMind and not three like multimers, we compensated by computing 67x3=201 predictions per neural network model, therefore obtaining 8040 structure models as well for each target, again except for the largest targets for which we reduced the number of predictions according to their size.

Predicted structures were not relaxed.

All predictions were ranked based on the AlphaFold confidence score and the top 5 structures were submitted to CASP16-CAPRI phase 1 as a MassiveFold baseline. All our predictions were then made available to all the predictors for phase 2. For the specific case of targets presenting alternative conformations, human input was involved in structure selection for the top 5.

Setup	Dropout Evoformer	Dropout structure module	Templates	Recycles	Structure inference engine
afm_basic			X	**21**	**AFmassive**
afm_woTemplates				**21**	**AFmassive**
afm_dropout_full	X	X	X	**21**	**AFmassive**
afm_dropout_full_woTemplates	X	X		**21**	**AFmassive**
afm_dropout_full_woTemplates_r3	X	X		**3**	**AFmassive**
afm_dropout_noSM_woTemplates	X			**21**	**AFmassive**
cf_woTemplates				**21**	**ColabFold**
cf_dropout_full_woTemplates	X	X		**21**	**ColabFold**

Table 1: Sets of parameters used for massive sampling in CASP16-CAPRI

Results

We computed massive sampling with MassiveFold for 73 targets (34 monomers and 39 multimers). 8040 predictions were generated for each target, except for the following ones because of their large size:

- H1217 (5878 residues): 395 predictions
- H1227 (5689 residues): 45 structures were generated and the top 5 were submitted for phase 1, but for phase 2, the structure was trimmed to 2101 residues and 8040 predictions were generated
- H1258 (3092 residues); T1257 (3789 residues) and H1265 (3924 residues): 2040 predictions
- T1271 subunits: 2680 predictions each
- T1295 (3752 residues) and T1269 (2820 residues): 4080 predictions

In total, 510475 structures were produced.

Availability

MassiveFold is available here: https://github.com/GBLille/MassiveFold

Acknowledgements

This work was granted access to the HPC resources of IDRIS (under the allocations 2024-AD010715407, 2024-AD010715332 and 2024-AD010715333) made by GENCI (Grand Equipement National de Calcul Intensif).

1. Wallner B. 2023. Improved multimer prediction using massive sampling with AlphaFold in CASP15. Proteins.
2. Raouraoua N, Mirabello C, Véry T, Blanchet C, Wallner B, Lensink MF, et al. 2024. MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. Research Square
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596: 583–589.
4. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. 2021. Protein complex prediction with AlphaFold-Multimer.
5. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. Nat. Methods 19: 679–682.
6. Wallner B. 2023. AFsample: improving multimer prediction with AlphaFold using massive sampling. Bioinforma. Oxf. Engl. 39: btad573.
7. Raouraoua, Nessim, Lensink, Marc F, Brysbaert, Guillaume. 2024. Massive sampling strategy for antibody-antigen targets in CAPRI Round 55 with MassiveFold. Available from https://doi.org/10.22541/au.172592104.47153431/v2.