MassiveFold: parallelized massive sampling with AlphaFold in CASP16

Nessim Raouraoua[1], Romuald Marin[2], Christophe Blanchet[2], Marc F. Lensink[1], and Guillaume Brysbaert*[1]

[1]UMR8576 (UGSF) – Centre National de la Recherche Scientifique - CNRS, Université Lille Nord (France) – France

[2]Institut Français de Bioinformatique – Commissariat à l'énergie atomique et aux énergies alternatives, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

AlphaFold2 is a deep learning tool that has revolutionised the field of structural bioinformatics. Since the publication of the inference code in 2021, the community has fully adopted AlphaFold2 and adapted it in many ways. Its creators demonstrated the superiority of the software's predictions for monomeric proteins at the biennial competition CASP14 (Critical Assessment of Structure Prediction) in 2020. An updated version was produced in 2021 and updated again in 2022. These versions can predict proteins in complexes and are named AlphaFold-Multimer. At CASP15 in 2022, AlphaFold-Multimer produced promising results in predicting protein complexes, though a researcher named Björn Wallner managed to substantially improve the models by using massive sampling with AlphaFold2 in a tool he developed called AFsample.

Massive sampling means increasing the number of predictions (25 with a default run of AlphaFold2) to several thousands and activating diversity parameters such as the usage of templates, the number of times the structure is recycled in the prediction process or activating the dropout at the inference. However, this approach requires a lot of GPU computing, and AlphaFold was not originally designed to parallelise computation on many GPUs.

Starting from a hackathon at the IDRIS (Institut du développement et des ressources en informatique scientifique) in 2023, we developed a tool named MassiveFold that allows to use AlphaFold2 with diversity parameters. MassiveFold is optimised to make several thousand predictions in parallel and can run on any infrastructure type, ranging from a computer with a single GPU to a GPU supercomputing cluster such as Jean Zay, the national CNRS supercomputer. MassiveFold is available to all users on Jean Zay.

In the 2024 CASP16 protein structure competition, we were granted 300,000 hours of computing time on the Jean Zay supercomputing cluster to conduct a systematic experiment involving massive sampling for all protein targets. These data were used in Phase 1 of the competition, in which predictors submitted their top five predictions, and in Phase 2, in which they were able to use our massive sampling data for the target to submit another top five. Our analysis of the competition results shows that very good predictions are present

*Intervenant

in our MassiveFold dataset but are not well ranked by the AlphaFold confidence score. This suggests that an efficient scoring function should be developed to identify them. The analysis also revealed that many 'easy' targets did not require massive sampling computation, whereas 'hard' ones benefited from this strategy and the choice of pushing the computation to massive sampling can be made a priori from a basic AlphaFold2 run.

All our MassiveFold data with AlphaFold2 ranking scores and assessments are available on the CNRS research data space of the Recherche Data Gouv repository.

Since the competition, DeepMind released AlphaFold3, which we have now integrated in the version 1.5 of MassiveFold. We are currently developing a new version of MassiveFold based on NextFlow, MassiveFold-NF.

MassiveFold was published here:

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.
Nature Computational Science 4, 824–828 (2024).
https://doi.org/10.1038/s43588-024-00714-4

And the preprint of CASP16-CAPRI is here:

Nessim Raouraoua, Marc F. Lensink, Guillaume Brysbaert
MassiveFold data for CASP16-CAPRI: a systematic massive sampling experiment
bioRxiv 2025.05.26.653955
https://doi.org/10.1101/2025.05.26.653955