MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling

Supplementary materials

MassiveFold v1.2.2 parameters that can be specified in the JSON parameter file:
"models_to_use": list of neural network models to use; by default all are used
"pkl_format": how to manage pickle files

- 'full' to keep the pickle files generated by the inference engine,
- 'light' to reduce its size by selecting main components, which are: number of recycles, PAE values, max PAE, plddt scores, ptm scores, iptm scores and ranking confidence values (stored in ./light_pkl directory)
- 'none' to remove them

Parameters added to AlphaFold in AFmassive v1.1.3 and used with *run_AFmassive.py*:
--*dropout_rates_filename*: provides dropout rates at inference from a JSON file.
--*early_stop_tolerance*: early stop threshold for recycling
--*bfd_max_hits*: max hits in BFD/uniref MSA
--*mgnify_max_hits*: max hits in mgnify MSA
--*uniprot_max_hits*: max hits in uniprot MSA
--*uniref_max_hits*: max hits in uniref MSA
--*start_prediction*: prediction to start with, can be used to parallelize jobs
--*end_prediction*: prediction to end with, can be used to parallelize jobs
--*stop_recycling_below*: after the first recycle step, only predictions with ranking confidence above this score will continue recycling; predictions below this threshold will still be present in *ranking_debug.json* and produced output.
--*min_score*: predictions with a score below this threshold will be excluded from the output
--*max_score*: terminates the computing process when a suitable prediction with a ranking confidence > max_score has been obtained
These parameters are accessible and can be used like the other flags of AlphaFold through the *run_AFmassive.py* script (instead of *run_alphafold.py*). All parameters will be listed with *run_AFmassive.py --help*. In the context of MassiveFold v1.2.2, these parameters can be set in the AFmassive JSON parameters file under the "AFM_run" section.

Parameters of ColabFold v1.5.5 (as described by the authors) that can be set in the ColabFold JSON parameters file for MassiveFold v1.2.2 and that refer to the *colabfold_batch* executable (details accessible through *colabfold_batch --help* except "model_preset" which was added):
"model_preset": multimer or monomer_ptm
"pair_strategy": How sequences are paired during MSA pairing for complex prediction. complete: MSA sequences should only be paired if the same species exists in all MSAs. greedy: MSA sequences should only be paired if the same species exists in at least two MSAs. Typically, greedy produces better predictions as it results in more paired sequences. However, in some cases complete pairing might help, especially if MSAs are already large and can be well paired.
"use_dropout": Activate dropouts during inference to sample from uncertainty of the models. This can result in different predictions and can be (carefully!) used for conformations sampling.
"num_recycle": Activate dropouts during inference to sample from uncertainty of the models. This can result in different predictions and can be (carefully!) used for conformations sampling.
"recycle_early_stop_tolerance": Specify convergence criteria. Run recycles until the distance between recycles is within the given tolerance value.
"stop_at_score": Compute models until pLDDT (single chain) or pTM-score (multimer) > threshold is reached. This speeds up prediction by running less models for easier queries.
"disable_cluster_profile": Experimental: For multimer models, disable cluster profiles.

Sets of parameters used for the massive sampling generation for H1140 with AFmassive:

Set 1 (Figure 2d):
- 3 NN versions, 5 NN models per version
- 5 predictions per NN model, totaling 75 predictions
- no dropout
- templates used
- recycling: 20 steps and early stop tolerance set to 0.5

Set 2 (Figure 2abe):
- 3 NN versions, 5 NN models per version
- 67 predictions per NN model, totaling 1005 predictions
- no dropout
- templates used
- recycling: 20 steps and early stop tolerance set to 0.5

Set 3 (Figure 2bf):
- 3 NN versions, 5 NN models per version
- 67 predictions per NN model, totaling 1005 predictions
- dropout activated (for Evoformer and structure module)
- no templates
- recycling: 20 steps and early stop tolerance set to 0.5

Set 4:
- NN version v1, one NN model, 10 predictions
- dropout activated (for Evoformer and structure module)
- no templates
- recycling: 1000 steps and early stop tolerance set to 0.5

Set 5 (Figure 2c and Table S1):
- NN version v1, one NN model, 10 predictions
- dropout activated (for Evoformer and structure module)
- templates not used
- recycling: 1000 steps and early stop tolerance set to 0.1

Two additional sets of parameters were used with ColabFold as a prediction engine:

Set 6 (Supplementary Figure 4a):
- 3 NN versions, 5 NN models per version
- 5 predictions per NN model, totaling 75 predictions
- no dropout
- templates not used
- recycling: 20 steps and early stop tolerance set to 0.5

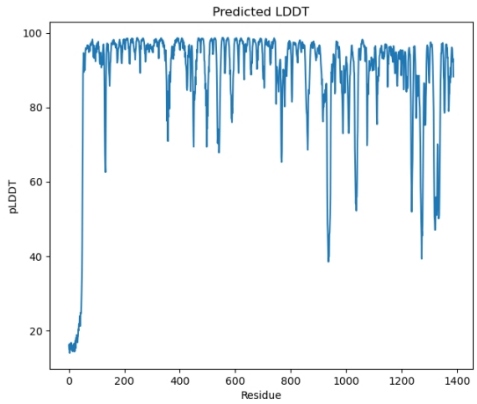
Set 7 (Supplementary Figure 4b):
- 3 NN versions, 5 NN models per version
- 67 predictions per NN model, totaling 1005 predictions
- dropout activated
- templates not used
- recycling: 20 steps and early stop tolerance set to 0.5

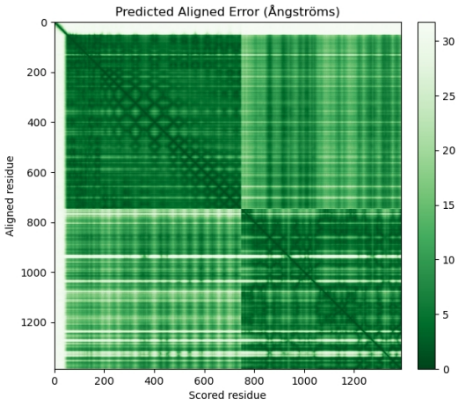
Supplementary Figures and Table

```
{
    "dropout_rate_msa_row_attention_with_pair_bias": 0.15,
    "dropout_rate_msa_column_attention": 0.0,
    "dropout_rate_msa_transition": 0.0,
    "dropout_rate_outer_product_mean": 0.0,
    "dropout_rate_triangle_attention_starting_node": 0.25,
    "dropout_rate_triangle_attention_ending_node": 0.25,
    "dropout_rate_triangle_multiplication_outgoing": 0.25,
    "dropout_rate_triangle_multiplication_incoming": 0.25,
    "dropout_rate_pair_transition": 0.0,
    "dropout_rate_structure_module": 0.1
}
```

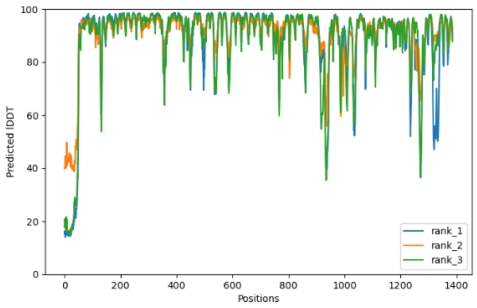
Supplementary Figure 1: List of dropout rates of the Evoformer and of the structure module (last entry)

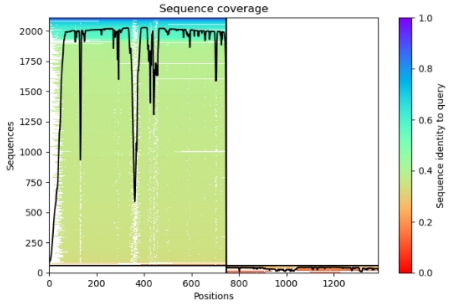
a - DeepMind

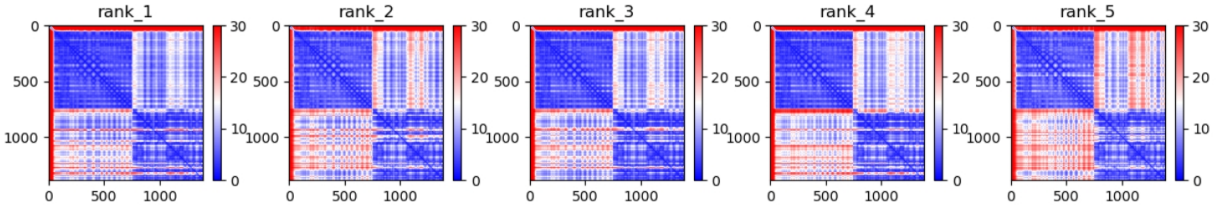




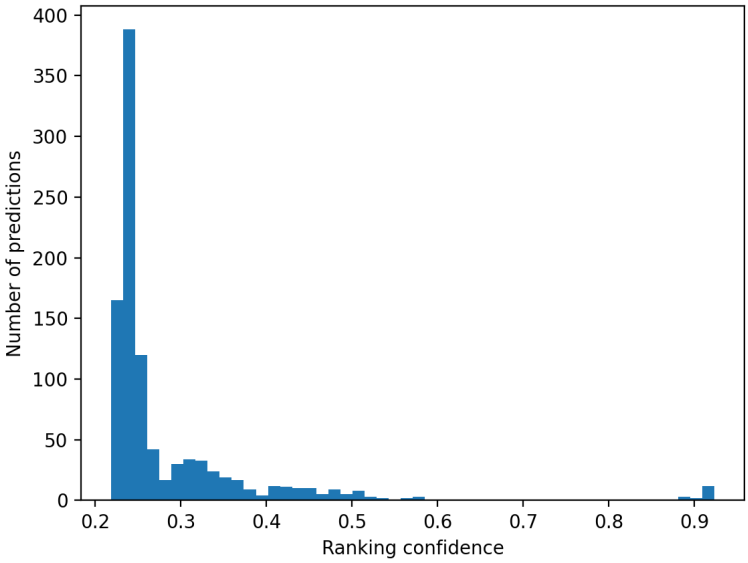
b - ColabFold



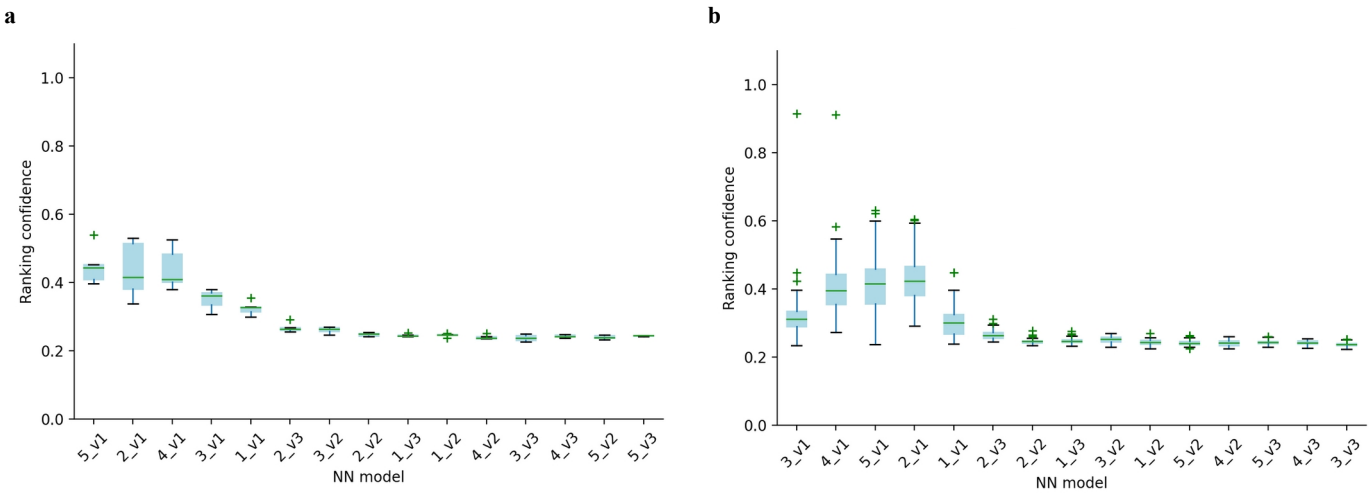




Supplementary Figure 2: pLDDT and Predicted Aligned Error plots following (**a**) the DeepMind style (one of each per predicted structure) and (**b**) the ColabFold style (multiple graphs in the same plot, left for pLDDT and bottom for Predicted Aligned Error); the ColabFold plots also show the sequence coverage (right-hand plot).



93 **Supplementary Figure 3:** Distribution of AlphaFold confidence scores for a prediction run of 1005 structures for CASP15 target H1140.



94 **Supplementary Figure 4:** Boxplots of the ranking confidence for each NN model generated by MassiveFold using ColabFold for structure
95 prediction for CASP15 target H1140, without templates, 20 recycles and early stop tolerance set to 0.5: (**a**) computing 75 predictions without
96 dropout activated, (**b**) computing 1005 predictions with dropout activated

Early stop tolerance	0.5	0.1
Confidence Scores	0.922	0.923
	0.920	0.923
	0.919	0.921
	0.917	0.921
	0.358	0.921
	0.245	0.921
	0.240	0.920
	0.234	0.920
	0.233	0.919
	0.189	0.918

Supplementary Table 1: Comparison of scores between 10 predictions for CASP15 target H1140, using the first neural network v1, dropout activated in the Evoformer and structure modules, without templates, with up to 1000 recycles and two different early stop tolerance thresholds.