

MASSIVEFOLD

Practical work

Massive sampling

IFB training - AlphaFold and beyond - 10th to 12th December 2025

T1206

The T1206 CASP16 target is a homodimer of the Porcine Astrovirus 4 Capsid Spike. We are going to model this homodimer and see if it is worth it enhancing the sampling for this target.

- Copy the 'AFmassive_params.json' file to 'AFmassive_params_v3_default.json' and edit "model_to_use" using only neural network v3 of AlphaFold2 (AlphaFold2 baseline)

```
"massivefold": {  
  "run_massivefold": "run_AFmassive.py",  
  "run_massivefold_plots": "../massivefold/massivefold_plots.py",  
  "data_dir": "$DSDIR/AlphaFold-2024-04",  
  "uniref_database": "",  
  "jobfile_headers_dir": "./headers",  
  "jobfile_templates_dir": "../massivefold/parallelization/templates",  
  "output_dir": "./output",  
  "logs_dir": "./log",  
  "input_dir": "./input",  
  "scripts_dir": "../massivefold/parallelization",  
  "models_to_use": "model_1_multimer_v3,model_2_multimer_v3,model_3_multimer_v3,model_4_multimer_v3,model_5_multimer_v3",  
  "pkl_format": "light"  
},
```

- Run AFmassive:

```
./run_massivefold.sh -s input/T1206.fasta -r afm_v3_default -f AFmassive_params_v3_default.json
```

T1206 - Results

- Look at the “ranking_iptm.json” scores
- Precomputed results are also available here

<https://nextcloud.univ-lille.fr/index.php/s/dnQLmBywbcQdx56>

=> Do you think it is worth it enhancing the sampling ?

H1267

The H1267 CASP16 target is a tetramer involved in SPARDA system. Here again, we want to model this homodimer and see if it is worth it to enhance the sampling for this target.

- Perform the same computation as T1206, but setting 'jeanzay_jobarray_time' to 00:45:00 and the parameter -b to 1
- Look at the "ranking_iptm.json" scores
- Precomputed results are also available here

<https://nextcloud.univ-lille.fr/index.php/s/dnQLmBywbcQdx56>

=> Is it worth it enhancing the computation ?

H1267

Let's try massive sampling with AFmassive, ColabFold and AlphaFold3.

We defined 5 sets of parameters for AFmassive and ColabFold and showed in Raouraoua et al., Proteins, 2025 that computing 33 predictions per neural network model is enough for the majority of cases

=> 33 predictions x 15 NN models x 5 conditions = 2475 predictions

- Get the 5 JSON files from the 'Parameters_JSON' folder of the shared WORK storage (also available on Nextcloud: <https://nextcloud.univ-lille.fr/index.php/s/dnQLmBywbcQdx56>)

```
rsync -avx $ALL_CCFRWORK/training_Alphafold_202512/Parameters_JSON/* $CCFRSCRATCH/MassiveFold/massivefold_runs/
```

- Use them to launch these 5 runs:

```
./run_massivefold.sh -s input/H1267.fasta -r afm_basic_33 -p 33 -b 5 -f AFmassive_params_basic.json  
./run_massivefold.sh -s input/H1267.fasta -r afm_dropout_full_33 -p 33 -b 5 -f AFmassive_params_dropout_full.json  
./run_massivefold.sh -s input/H1267.fasta -r afm_dropout_noSM_woTemplates_33 -p 33 -b 5 -f AFmassive_params_dropout_noSM_woTemplates.json  
./run_massivefold.sh -s input/H1267.fasta -r afm_woTemplates_33 -p 33 -b 5 -f AFmassive_params_woTemplates.json  
./run_massivefold.sh -s input/H1267.fasta -r cf_woTemplates_33 -p 33 -b 5 -f ColabFold_params_woTemplates.json
```

N.B.: if the alignments are not run yet, use the '-j' parameter of 'run_massivefold.sh' to wait for the alignment job to finish for the 4 jobs after the AFmassive basic one.

and an extra run with AlphaFold3:

```
./run_massivefold.sh -s input/H1267.fasta -r af3_basic_33 -p 33 -b 5 -f AlphaFold3_basic.json
```

H1267

- Display your jobs in the queue:

```
squeue --me -o "%.18i %.9P %.50j %.8u %.2t %.10M %.6D %R %.8Q %.10l %.10P" --sort=-p
```

- Because the computation for these 5 sets would require 400 to 500 H100 hours, cancel them all

```
scancel --me
```

N.B.:

- Once completed, you should be able to use the 'gather_runs.py' script in the 'massivefold' folder to gather all the predictions of these 5 sets in one folder, with a full reranking.

```
../massivefold/gather_runs.py --help
```

- AlphaFold3_params.json contains parameters to use or not the paired, unpaired and templates. These parameters could be used in massive sampling (not evaluated yet)

