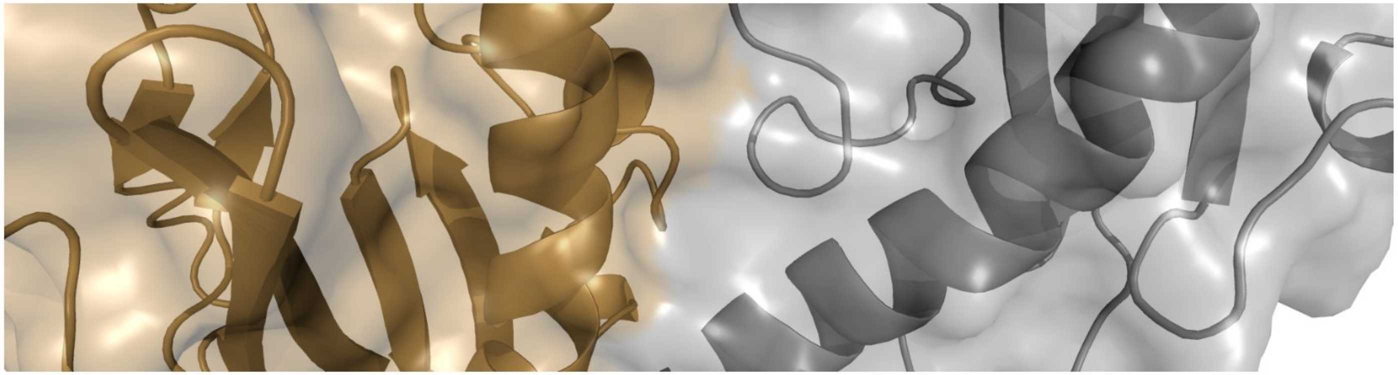


MassiveFold

Massive sampling data shared over CASP16-CAPRI

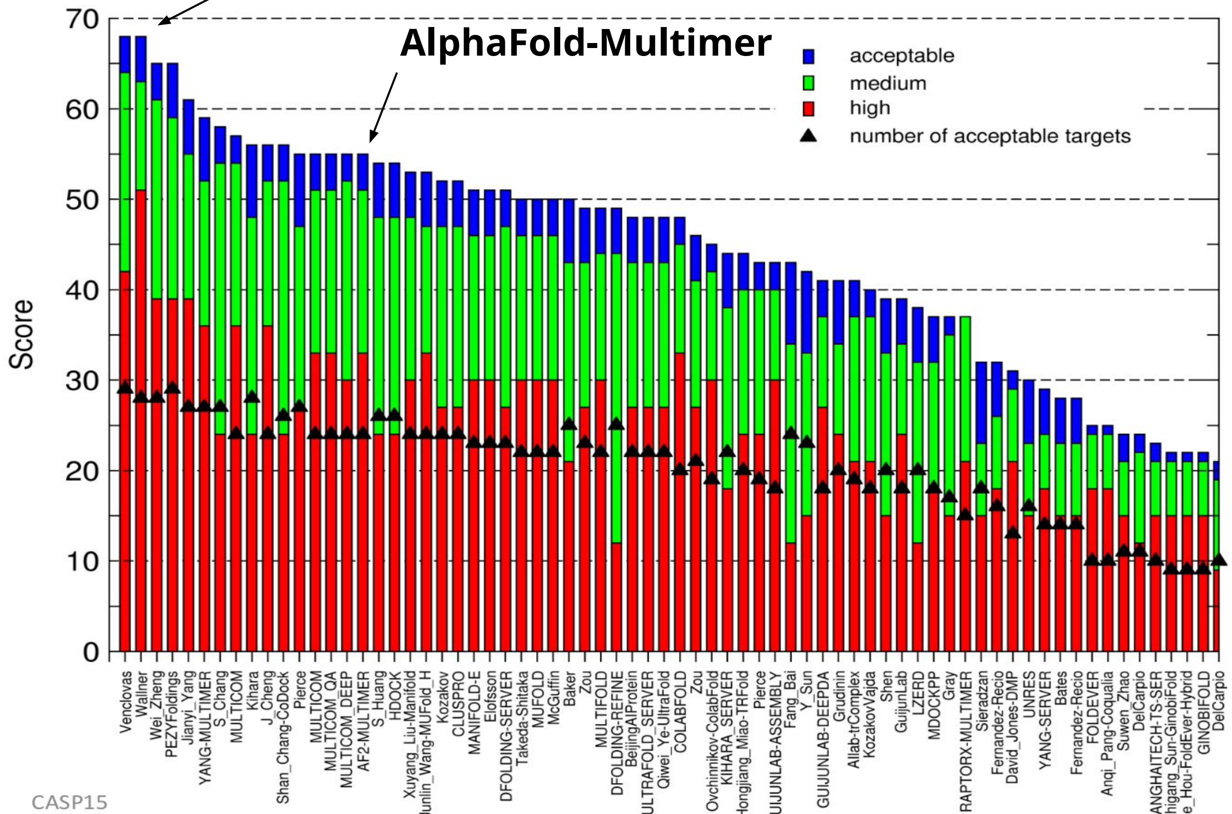
Nessim Raouraoua, Marc F. Lensink and Guillaume Brysbaert

Guillaume Brysbaert
CNRS - France - Lille

CASP15/CAPRI - Dec 2022

Multimers

Björn Wallner = Massive sampling with AFsample



Marc F Lensink, CASP15, 2022

Massive sampling:

- thousands of predictions
- diversity parameters: neural network version, dropout, templates, recycles

Limitations:

- cost in GPU hours
- management of such a large computation

CNRS supercomputing cluster "Jean Zay" - France

cnrs

IDRIS - Paris

Partition CPU

- 28800 cœurs Intel Cascade Lake 6248 @ 2,5 GHz
- RAM 138 To
- 2,3 PFlop/s

Partitions GPU

4 GPU	8 GPU	4 GPU
1832 GPU V100	416 GPU A100	1456 GPU H100
OPA 100 Gb/s par GPU	OPA 100 Gb/s par GPU	IB NDR 400 Gb/s par GPU
50 To HBM2	33 To HBM2e	116 To HBM3
17,8 PFlop/s	8,2 PFlop/s	99,9 PFlop/s

MassiveFold

Started in March 2023 (GPU Hackathon at IDRIS with NVIDIA)

Nessim Raouraoua
Marc Lensink
Guillaume Brysbaert

Claudio Mirabello
Björn Wallner

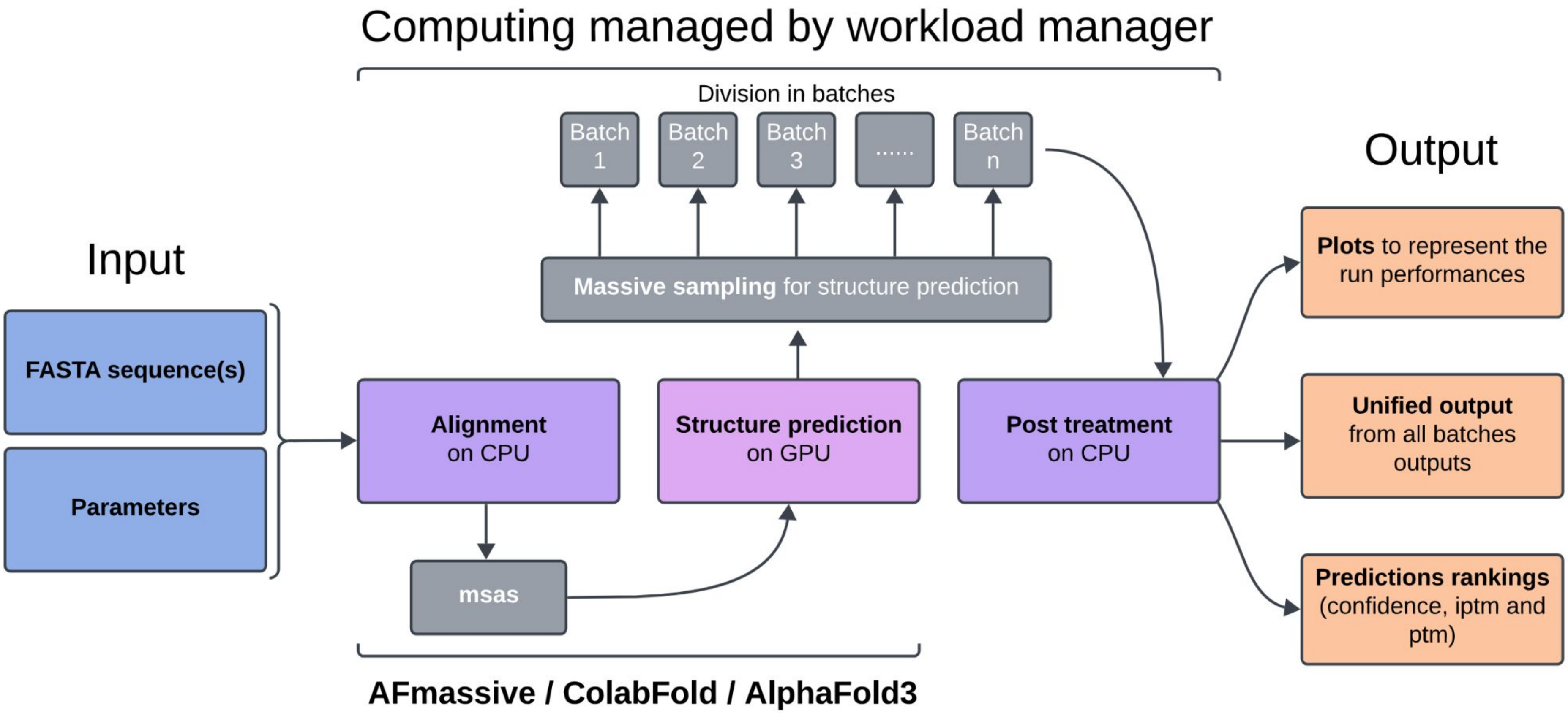
MUDIS4LS
Christophe Blanchet

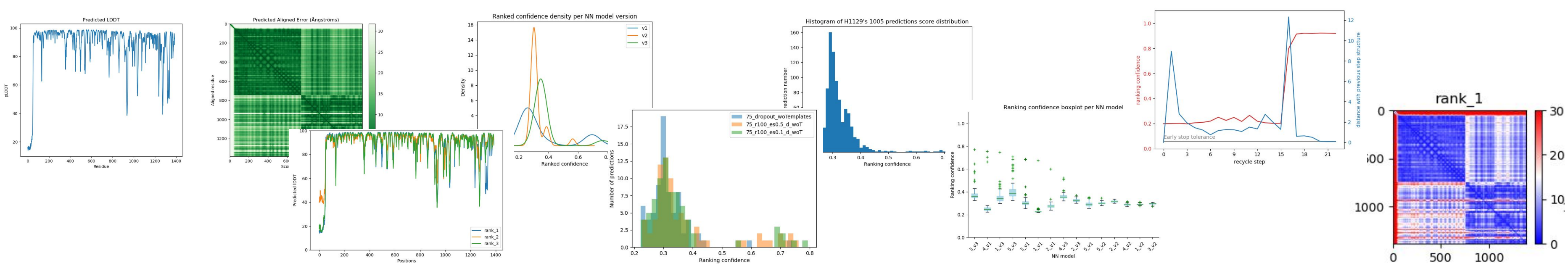
IDRIS
Supercomputing cluster Jean Zay
Thibaut Véry

Goals:

- Update **AFsample** => **AFmassive**, to use on the national cluster
- Optimization of the computing through **parallelization**

MassiveFold





Raouraoua Nessim et al. MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. 2024. *Nature Computational Science*, https://www.nature.com/articles/s43588-024-00714-4

CASP16/CAPRI - 2024

Statement:

- CAPRI 55 (February 2024): several groups ran massive sampling
- for CASP16-CAPRI, many groups would certainly do the same
- unfair for predictors who don't have access to many GPUs

Motivation for CASP16-CAPRI:

- provide massive sampling data to make the competition fairer
- avoid many groups burning GPU hours for the same type of computation
- boost scoring developments

CASP16/CAPRI - 2024

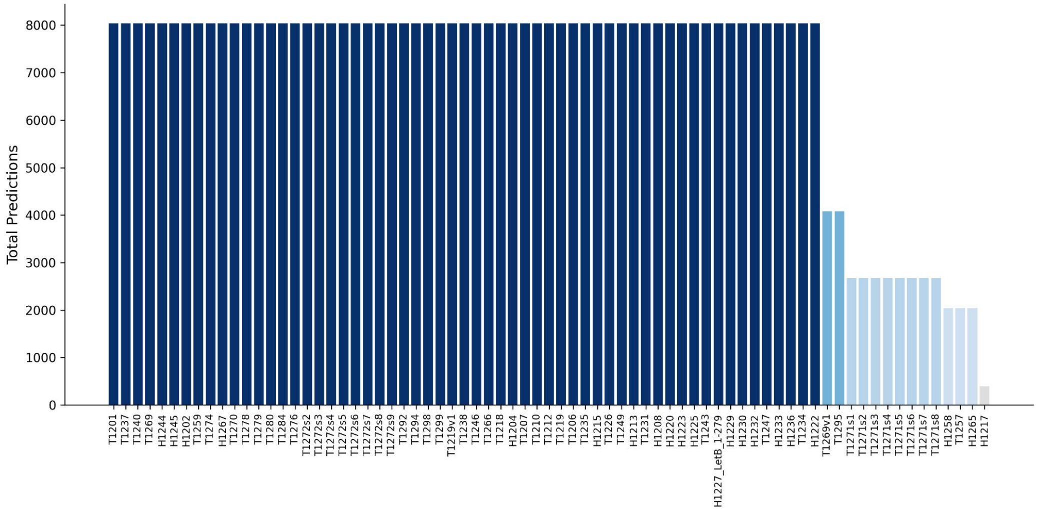
Statement:

- CAPRI 55 (February 2024): several groups ran massive sampling
- for CASP16-CAPRI, many groups would certainly do the same
- unfair for predictors who don't have access to many GPUs

Motivation for CASP16-CAPRI:

- provide massive sampling data to make the competition fairer
- avoid many groups burning GPU hours for the same type of computation
- boost scoring developments

Up to 8040 predictions per target



CASP16/CAPRI - 2024

Statement:

- CAPRI 55 (February 2024): several groups ran massive sampling
- for CASP16-CAPRI, many groups would certainly do the same
- unfair for predictors who don’t have access to many GPUs

Motivation for CASP16-CAPRI:

- provide massive sampling data to make the competition fairer
- avoid many groups burning GPU hours for the same type of computation
- boost scoring developments

Stage 0: stoichiometry

Stage 1: predictions

Participation as a baseline
Top-5 following the AF confidence score

Stage 2: MassiveFold

Predictions provided to predictors
(including “light” pickle files)

CASP16/CAPRI - 2024

(Up to) 8040 MassiveFold predictions = 8 x 15 NN x 67 predictions

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**

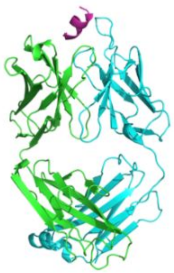
Early stop tolerance set to 0.5

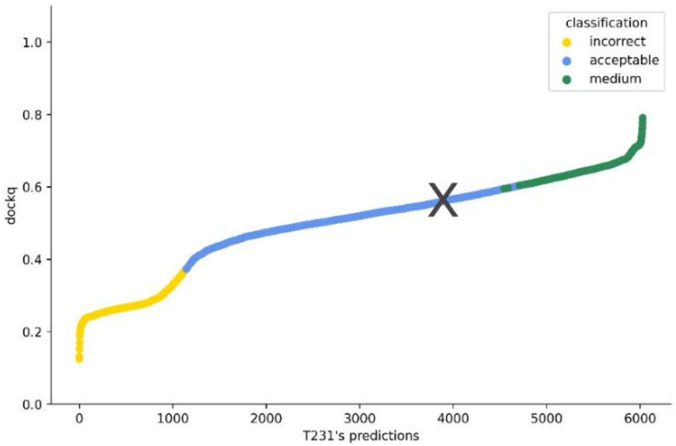
CASP16/CAPRI - 2024 - Computation on Jean Zay

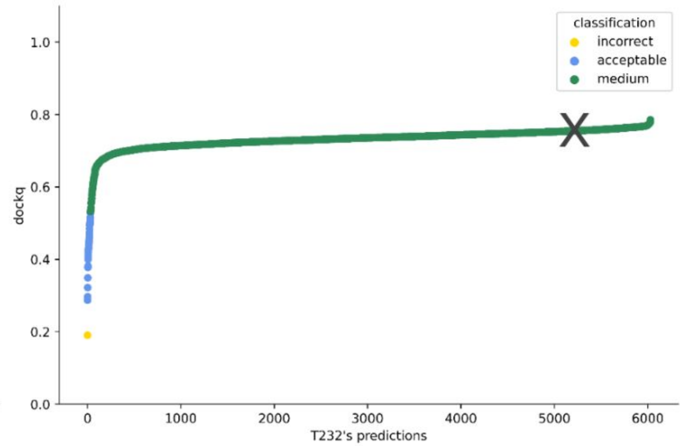
- **265 000 GPU hours** used (eq V100)
- **95 000 € ≃ $100 000**
- **7.3 CO_2 tons ≃ 9** round-trip flights Paris/Punta Cana
- **2.2 To** data shared for **73** targets in total (with “light” pickles)

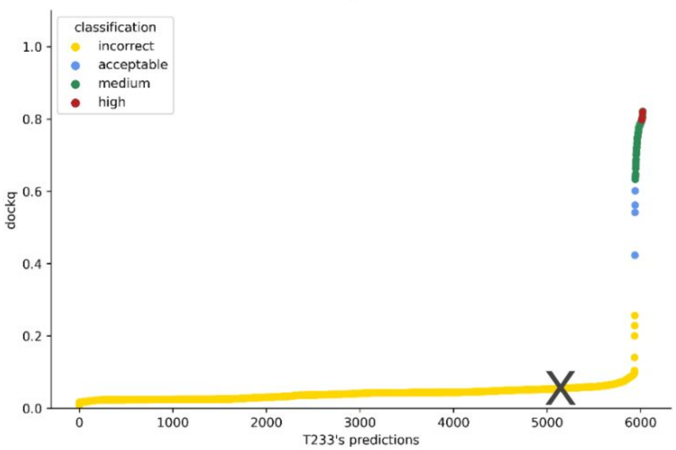
Target type	Number of predictions generated	Number of GPU hours used
Monomers	**262 640**	**43 000**
Assemblies	**288 605**	**222 000**
Total	**551 245**	**265 000**

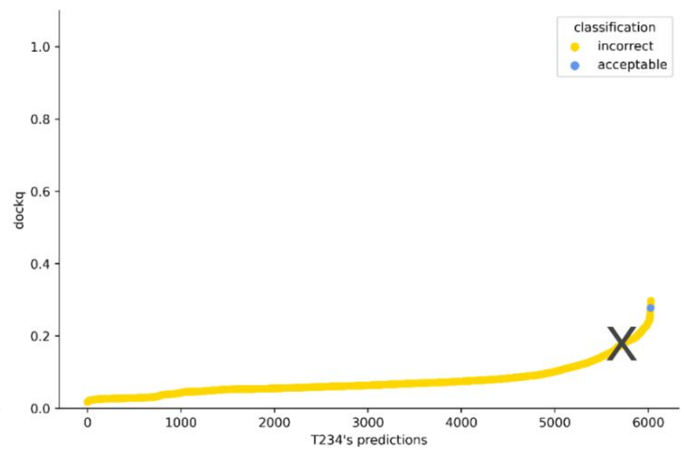
Expectations like CAPRI round 55











=> scoring

Conclusion

MassiveFold

- handles computing with AFmassive and ColabFold on CPU and many GPUs
- now also includes AlphaFold3

CASP16-CAPRI

- stage 1: baseline using AF2 confident score
- stage 2: up to 8040 predictions per target shared / > 500 000 predictions

An accurate **scoring** function is required => let's see CASP16-CAPRI's results!

https://github.com/GBLille/MassiveFold
https://github.com/GBLille/AFmassive

CAPRI

+ Nessim's POSTER

generated with ChatGPT 4