

MASSIVEFOLD

MassiveFold

Unveiling high quality predictions in MassiveFold data for CASP16-CAPRI

Nessim Raouraoua

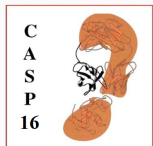
PhD student Univ. Lille, France

UGSF (Unité de Glycobiologie Structurale et Fonctionnelle) - Villeneuve d'Ascq

nessim.raouraoua@univ-lille.fr



What are CASP and CAPRI ?

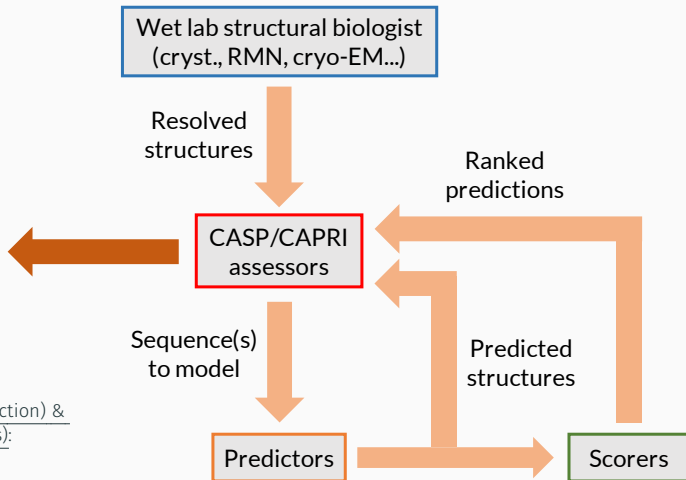


CASP & CAPRI conclusions:

- Groups ranking
- Method performance
- State of the art benchmark for future methods

CASP (Critical Assessment of protein Structure Prediction) & CAPRI (Critical Assessment of PRedicted Interactions):

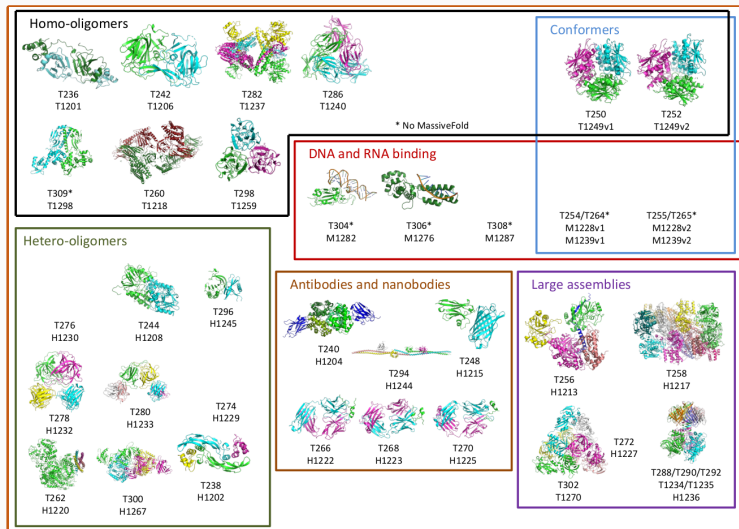
- Large joint experiment every 2 years
- Small CAPRI rounds in between



MassiveFold protocol

Parameters set's name	Sampling	Inference	Dropout		Templates	Recycles
			Evoformer	SM		
afm_basic	1005	AFmassive	✗	✗	✓	21
afm_woTemplates	1005	AFmassive	✗	✗	✗	21
afm_dropout_full	1005	AFmassive	✓	✓	✓	21
afm_dropout_full_woTemplates	1005	AFmassive	✓	✓	✗	21
afm_dropout_full_woTemplates_r3	1005	AFmassive	✓	✓	✗	3
afm_dropout_noSM_woTemplates	1005	AFmassive	✓	✗	✗	21
cf_woTemplates	1005	ColabFold	✗	✗	✗	21
cf_dropout_full_woTemplates	1005	ColabFold	✓	✓	✗	21
Total	8040					

CASP16-CAPRI



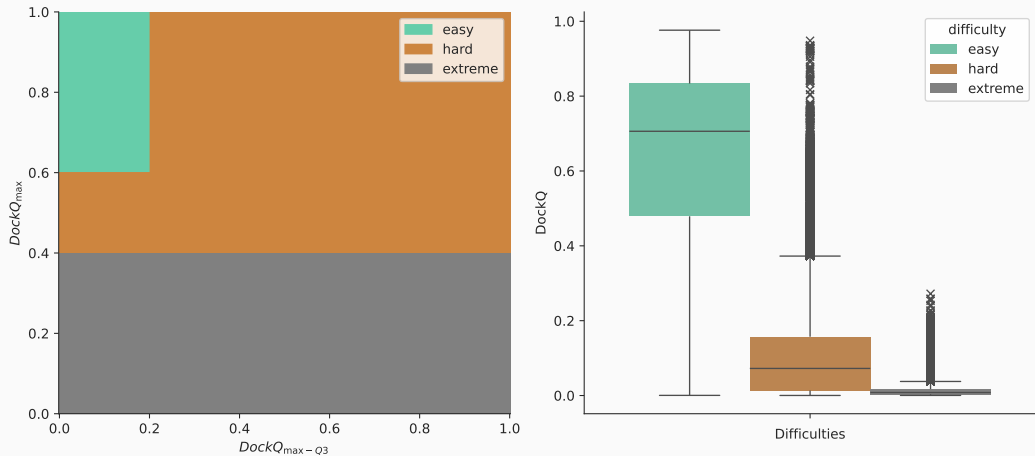
Monomer	31
Assemblies	39 (65*)
Total	70 (65*)

Some statistics

- **Accessibility:** Access to vast computing resources.
- **Infrastructure:** Estimated cost of \$100,000 (covering power, maintenance, etc..)
- **Storage:** 2.2 terabytes of optimized output data.
- **Environmental Impact:** 7.3 tons of CO2 emissions.

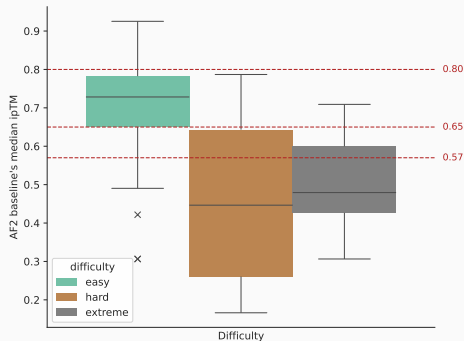
	Predictions	Computing time (kh)
Monomer	262 640	43
Assemblies	288 605	222
Total	551 245	265

CASP16-CAPRI: division by difficulty



→ A posteriori classification: DockQ distribution dispersion

CASP16-CAPRI: targets selection strategies (S1, S2 and S3)



ipTM median for AF2 basic run

- 0.8: challenging interfaces are below
- 0.65: separate 'easy' and challenging
- 0.57: best F1-score separator

→ Strategies S1, S2 and S3

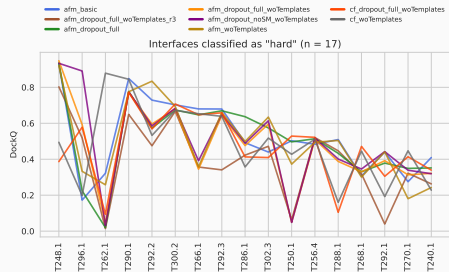
Strategy rollout with basic AF2 run:

- median ipTM > threshold: interface is 'easy', stop sampling
- median ipTM < threshold, interface is 'hard', producing massive sampling

CASP16-CAPRI: reduced scale strategies (S4, S5 and S6)

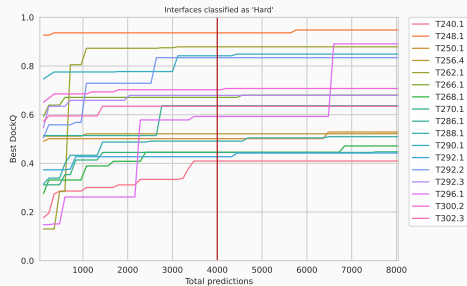
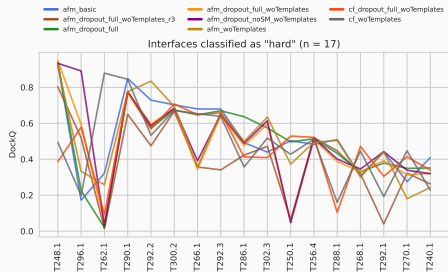
Parameters set's name	Sampling	Inference	Dropout		Templates	Recycles
			Evoformer	SM		
afm_basic	1005	AFmassive	✗	✗	✓	21
afm_woTemplates	1005	AFmassive	✗	✗	✗	21
afm_dropout_full	1005	AFmassive	✓	✓	✓	21
afm_dropout_full_woTemplates	1005	AFmassive	✓	✓	✗	21
afm_dropout_full_woTemplates_r3	1005	AFmassive	✓	✓	✗	3
afm_dropout_noSM_woTemplates	1005	AFmassive	✓	✗	✗	21
cf_woTemplates	1005	ColabFold	✗	✗	✗	21
cf_dropout_full_woTemplates	1005	ColabFold	✓	✓	✗	21
Total	8040					

CASP16-CAPRI: reduced scale strategies (S4, S5 and S6)



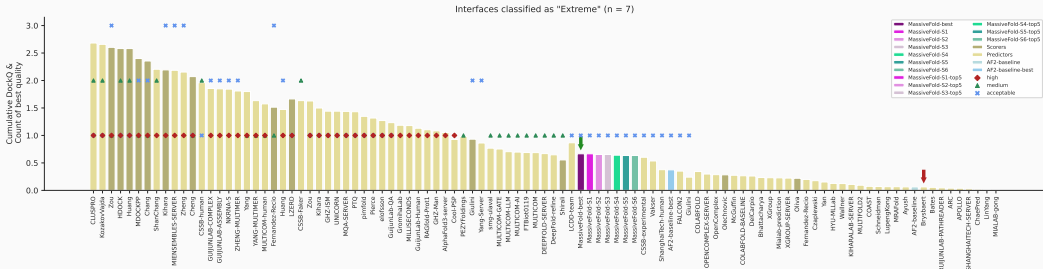
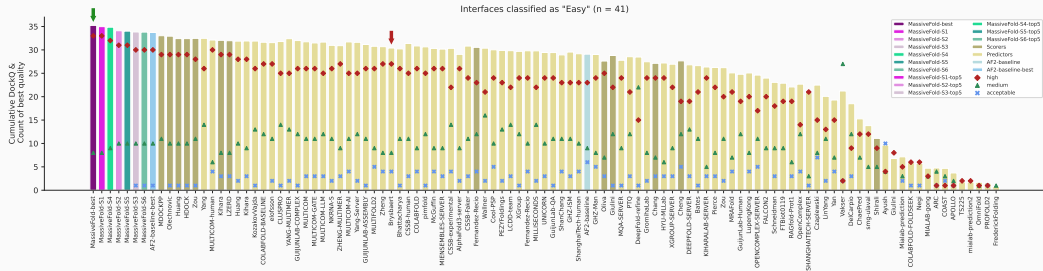
Parameters set's name	Sampling	Inference	Dropout		Templates	Recycles
			Evoformer	SM		
afm_basic	1005	AFmassive	✗	✗	✓	21
afm_woTemplates	1005	AFmassive	✗	✗	✗	21
afm_dropout_full	1005	AFmassive	✓	✓	✓	21
afm_dropout_full_woTemplates	1005	AFmassive	✓	✓	✗	21
afm_dropout_full_woTemplates_r3	1005	AFmassive	✓	✓	✗	3
afm_dropout_noSM_woTemplates	1005	AFmassive	✓	✗	✗	21
cf_woTemplates	1005	ColabFold	✗	✗	✗	21
cf_dropout_full_woTemplates	1005	ColabFold	✓	✓	✗	21
Total	8040 5025					

CASP16-CAPRI: reduced scale strategies (S4, S5 and S6)

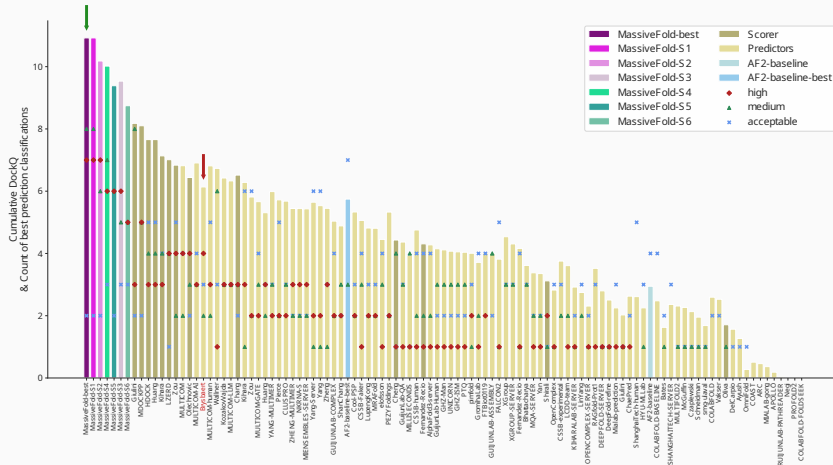


Parameters set's name	Sampling	Inference	Dropout		Templates	Recycles
			Evoformer	SM		
afm_basic	1005 495	AFmassive	✗	✗	✓	21
afm_woTemplates	1005 495	AFmassive	✗	✗	✗	21
afm_dropout_full	1005 495	AFmassive	✓	✓	✓	21
afm_dropout_full_woTemplates	1005	AFmassive	✓	✓	✗	21
afm_dropout_full_woTemplates_r3	1005	AFmassive	✓	✓	✗	3
afm_dropout_noSM_woTemplates	1005 495	AFmassive	✓	✗	✗	21
cf_woTemplates	1005 495	ColabFold	✗	✗	✗	21
cf_dropout_full_woTemplates	1005	ColabFold	✓	✓	✗	21
Total	8040 5025 2475					

CASP16-CAPRI: easy & extreme interfaces



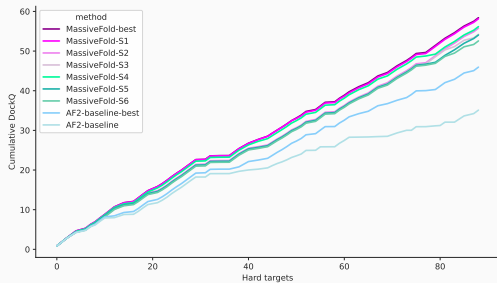
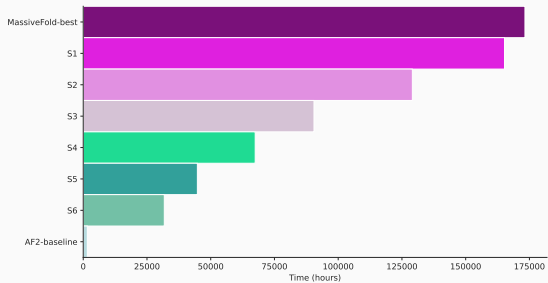
CASP16-CAPRI: on hard interfaces



Points

- Scoring
- Strategies

CASP16-CAPRI: strategies simulation



Efficiency:

- High computing time deduction
- Conserves high performance

Take-Home Messages

1 MassiveFold development

Protein structure prediction tool to efficiently parallelize massive sampling with many tools and diversity parameters.

2 CASP16-CAPRI participation

Production of MassiveFold data for baseline in stage 1 and shared in stage 2 in order to boost effort on scoring methods development.

3 Sampling strategies

With only the basic AlphaFold2 run ipTM runs, we manage to identify when to or not to produce massive sampling.