université PARIS DIDEROT PARIS 7

MASTER 2 Biologie Informatique

Team 4 - Downstream

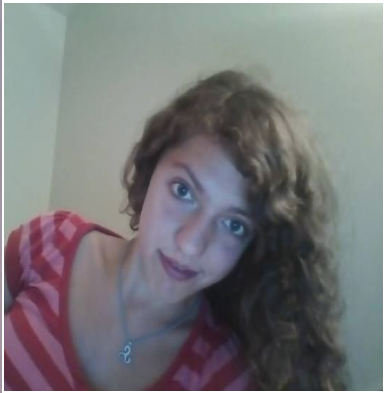
Gabriel Cretin - Hélène Kabbech - Tom Gutman - Flora Mikaeloff - Arnold-Franz Ake

Fold-U's team

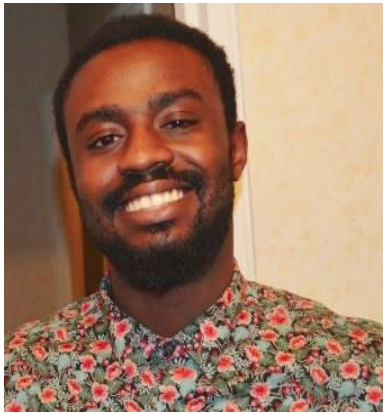
Gabriel Cretin
Git Master
Extreme coder

Hélène Kabbech
Manager
Extreme coder

Tom Gutman
Bibliography
Biologist



Flora Mikaeloff
Bibliography
Biologist



Arnold-Franz Ake
Code Review
Coder

Goal

Meet-U 2019 project

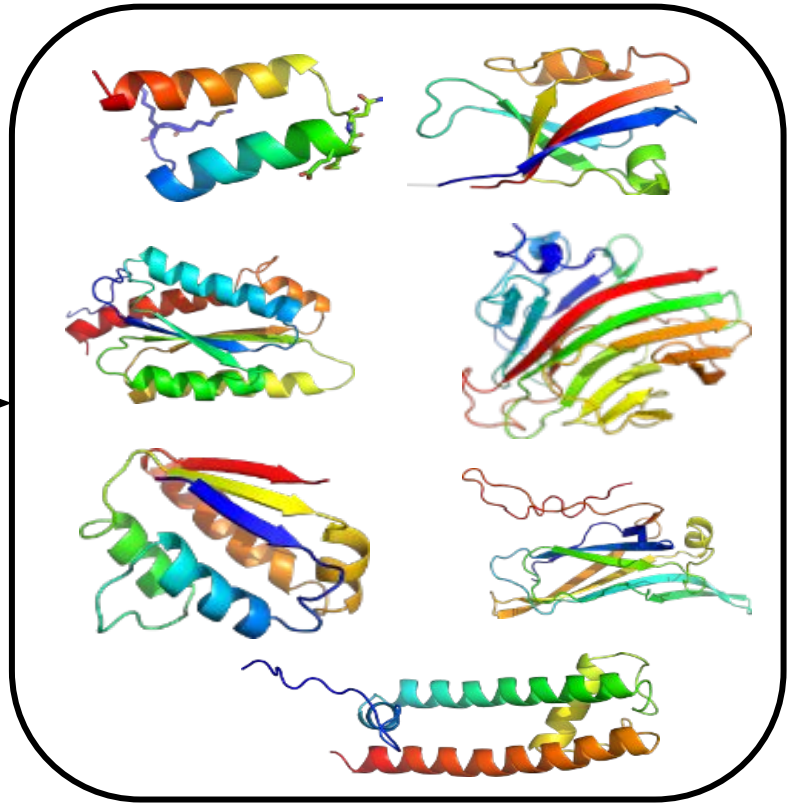
Structural prediction of a protein query from its amino acids sequence

Query sequence

ENVIREEA...RYCATEDGIT

3D structure ?

HOMSTRAD



Programming Tools and Strategy

Programming Tools

FOLDIU

Google Cloud Platform

Cluster processing

9.27/10

Pylint
Star your Python code!

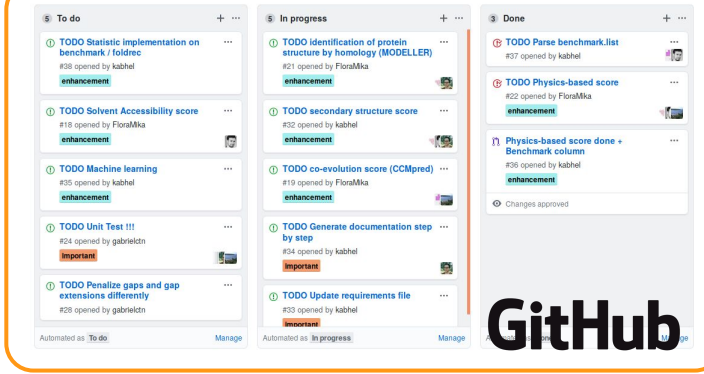
PEP8 Code Quality Checker

python™ pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
biopython NumPy

OOP Python 3 and Modules

SPHINX
Python Documentation Generator

Documentation



Pull request, Issues, Project management

Strategy Implemented scores

Alignment score from upstream team

+

Threading Score	Modeller Score	Co-Evolution Score	Secondary Structure Score	Solvent Accessibility Score	Blosum Score
DOPE energy of the query sequence threaded against templates structures	**DOPE energy** of the refined model based on alignment of query and template	**Adequation score of** evolutionary predictions between query and template structure	**Adequation score** between predicted and assigned secondary structures of query and template	**Adequation score** comparing buried/exposed residues between refined model and template	**Score** based on the comparison between query and template's likely mutations using BLOSUM62

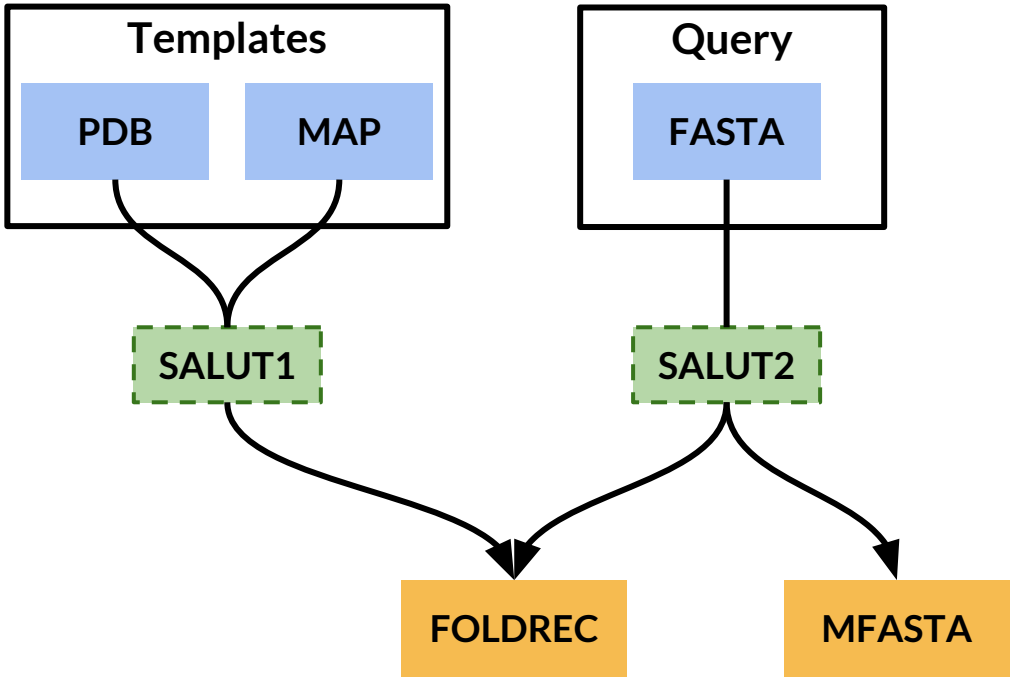
Normalization [0 - 1]

↓

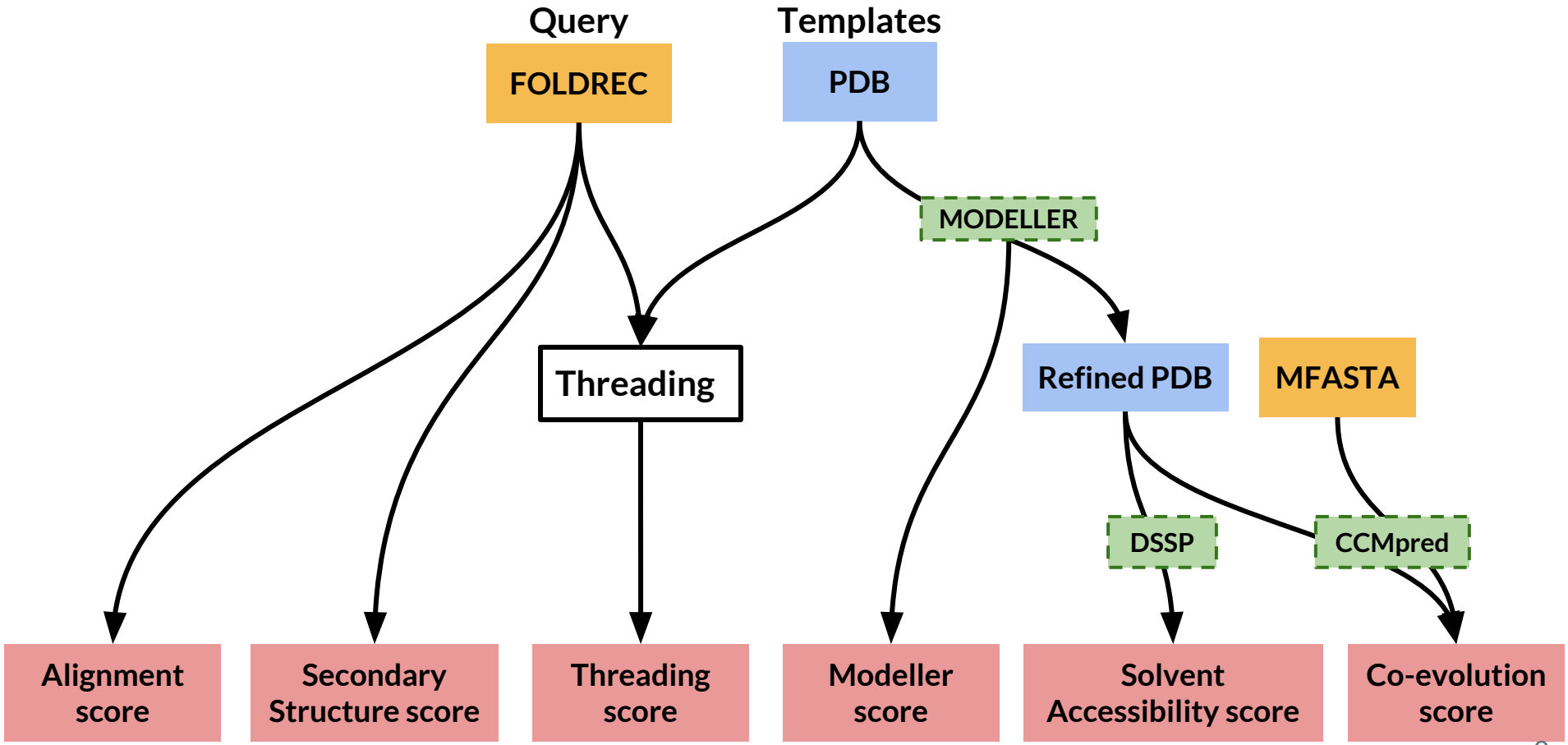
Machine Learning Score

Combination of scores, **weighted** by machine learning (Logistic Regression)

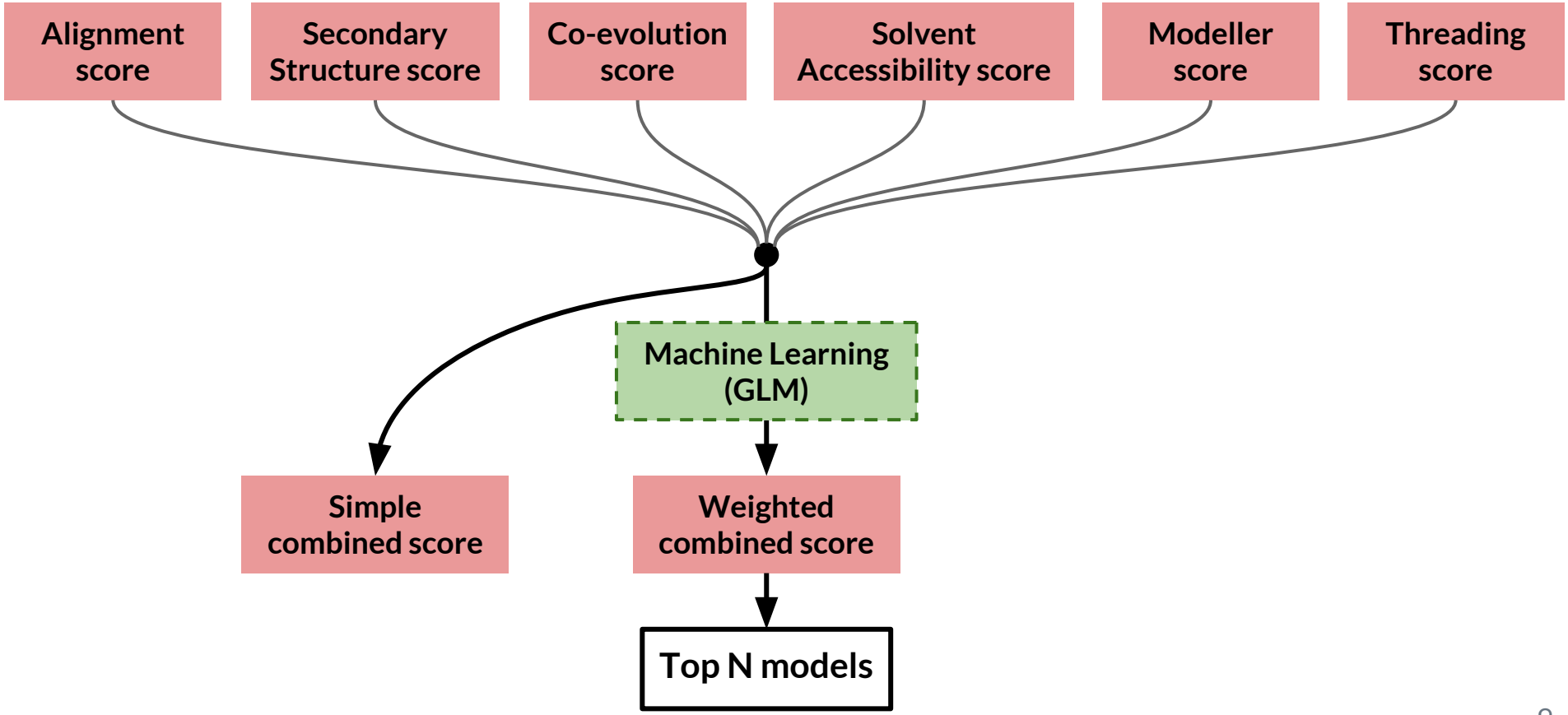
Pipeline Upstream



Pipeline Downstream



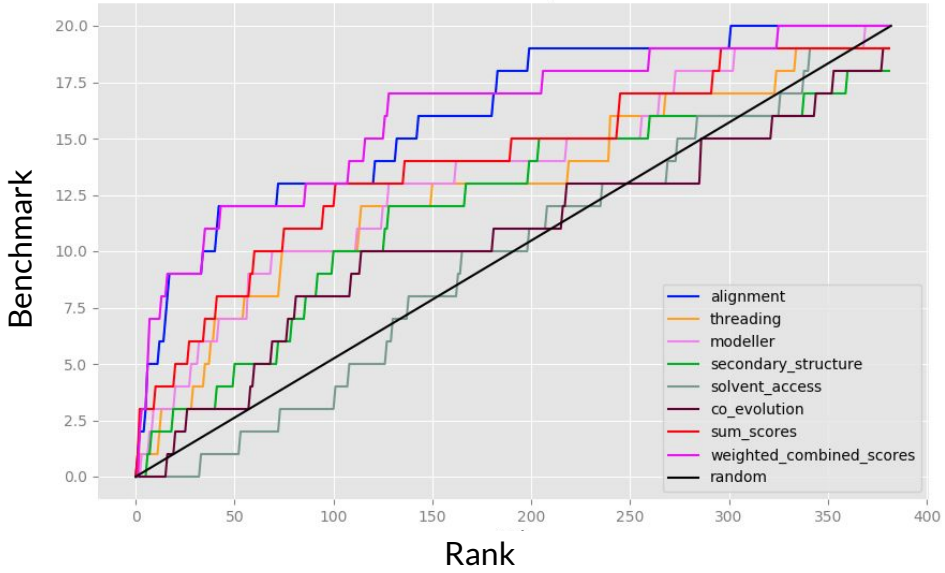
Pipeline Downstream



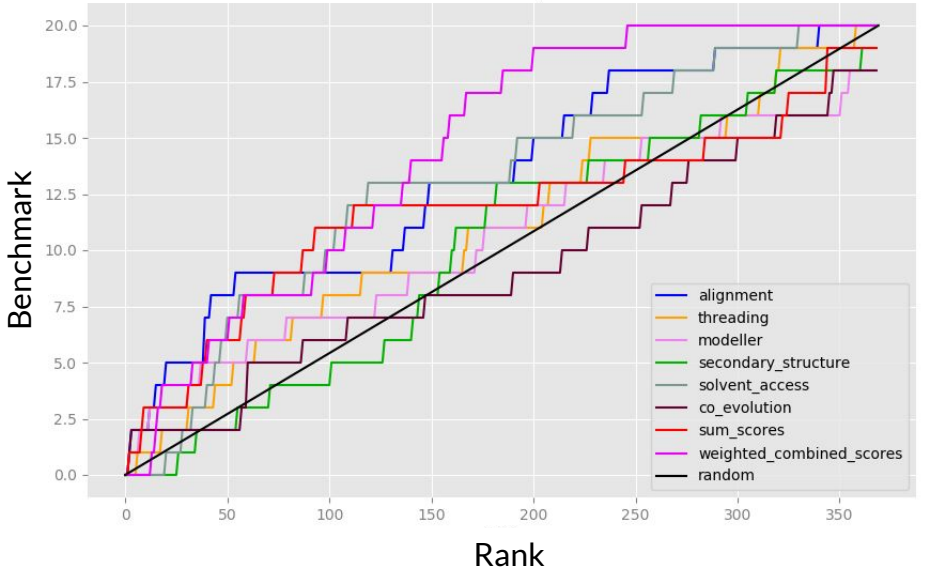
Benchmarking Results

Enrichment plots

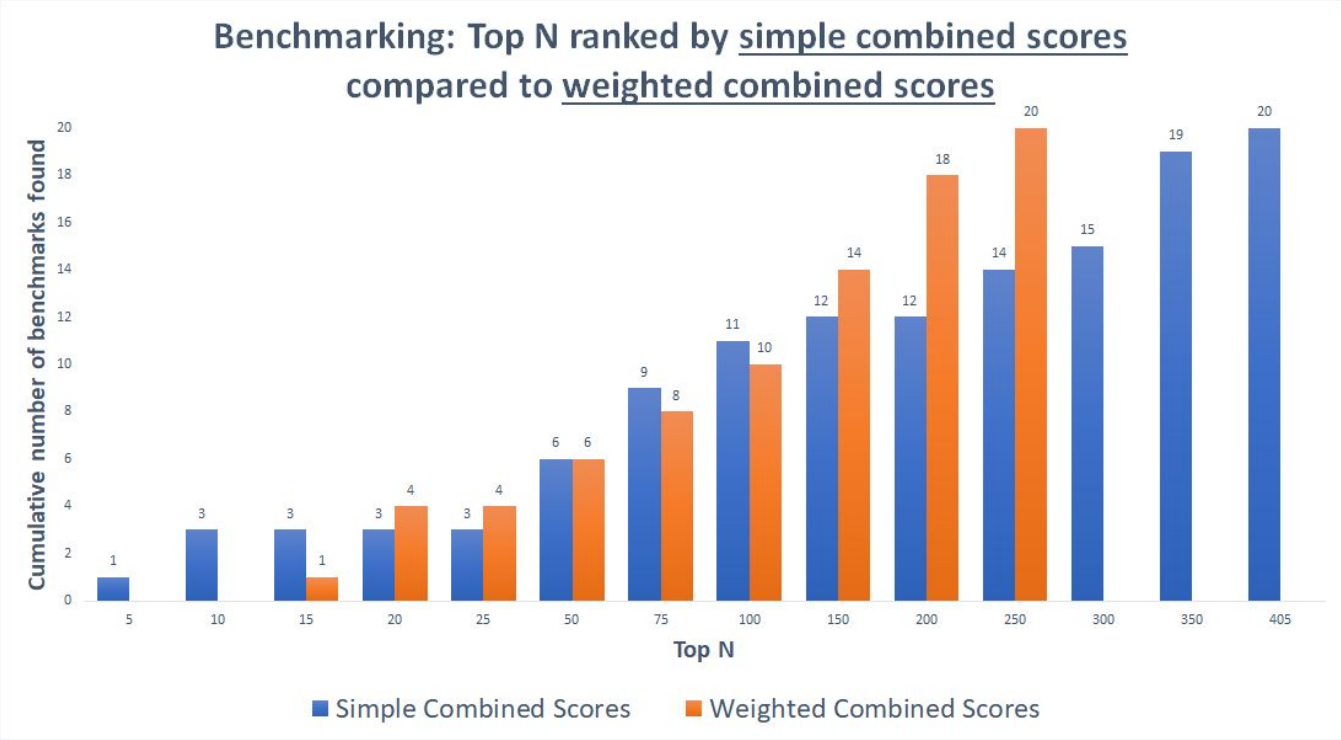
ORION Foldrecs



Upstream team Foldrecs



With upstream team Foldrecs



Machine learning: less specific, more sensitive detection !

Mysterious Sequence Results

Mysterious Sequences Downstream

Mysterious sequences	Top 1 Template found	SCOP id	Teams which found similar class SCOP ids
trQ6HKV8 Damien REMY	hormone	a.26.1.1	3/9
trQ6HLB4 Damien REMY	fkbp	d.26.1.1	1/9
feuA Hugues Herrmann	serpin	e.1.1.1	0/9
G1G14-4353 Hugues Herrmann	cah	b.74.1.1	1/9
recN Nicolas Methais **(gaps removed)**	citrate_synt	a.103.1.1	3/5
G1G14-3311 Mathieu Legras	DNA_photolyase	a.99.1.1	2/9
scdA Marie Merle	tms	d.117.1.1	0/9
hmp SELVAM Marie-Jeanne Lucie	COX1	f.24.1.1	3/9
mfd SELVAM Marie-Jeanne Lucie **(too long)**			
celB Dylan MOUTACHI	tubulin	c.32.1.1	1/9
arcD Dylan MOUTACHI	aldedh	c.82.1.1	1/9

Discussion & Conclusion

Discussion & Conclusion

Fold-U

Program able to generate N structural models based on :

- **Protein threading**
- **Additional scoring :**
  Sequence -and structure- based features
- **Machine learning :**
  Less specific but more sensitive Fold detection

Discussion & Conclusion

- ⊕ Optimized and clear code
- ⊕ Machine learning
- ⊕ Interesting and relevant results
- ⊖ Troubles with co-evolution and solvent accessibility scores
- ⊖ Machine learning : Can be improved using more data

Perspectives

Tweak scores parameters

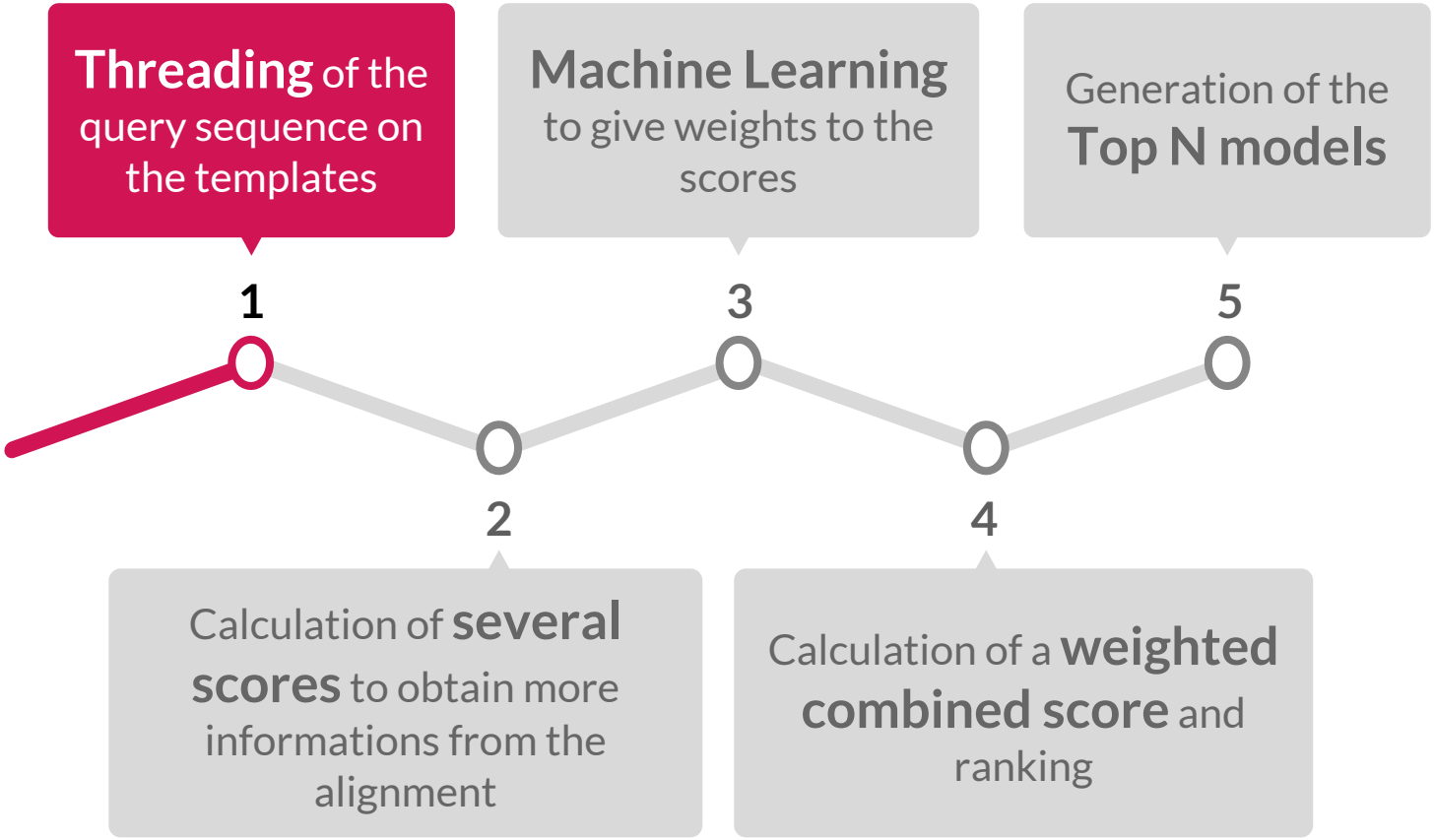
Changing the algorithm of solvent accessibility

Improvement of machine learning

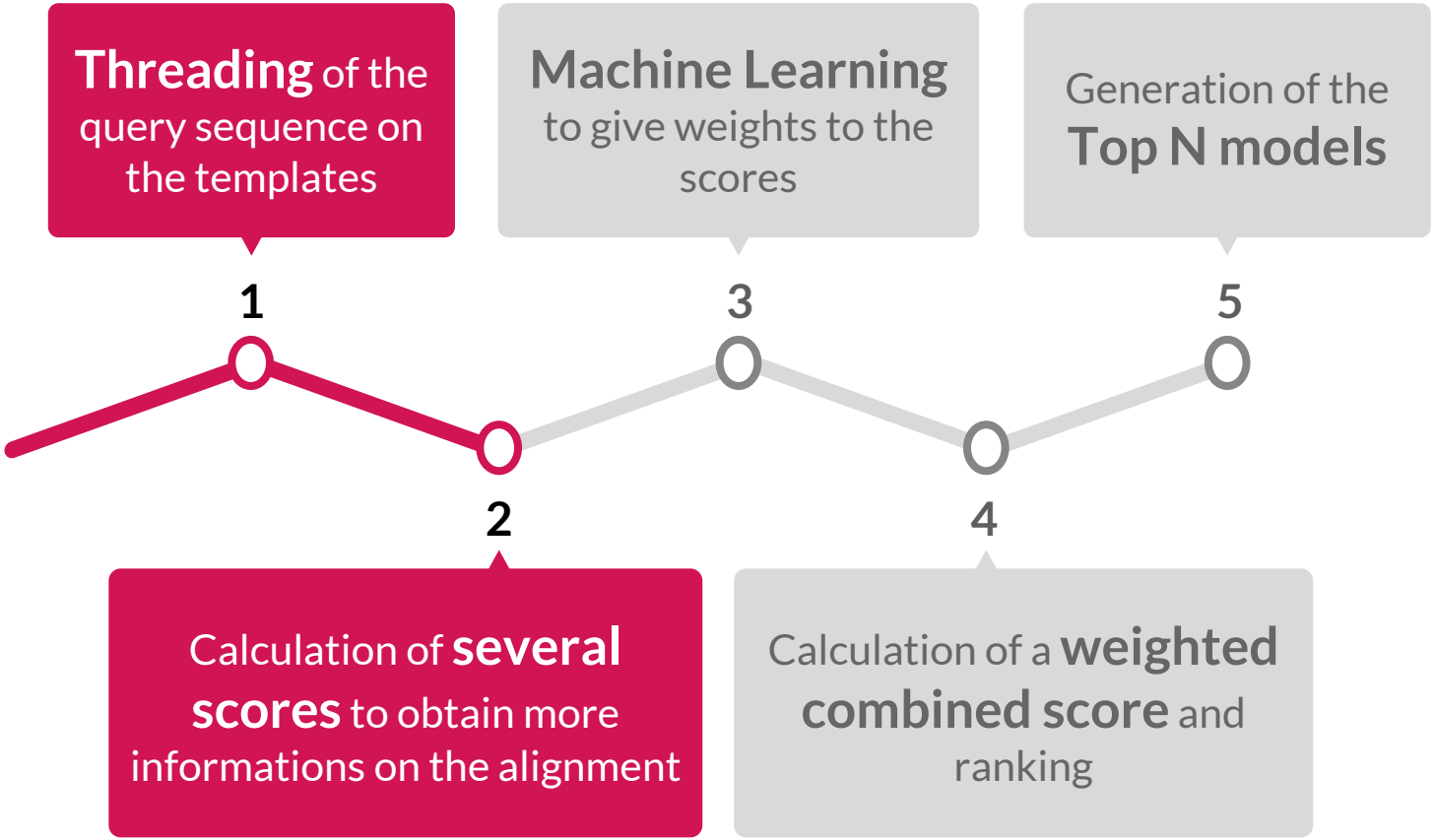
Add other scoring functions

Annexes

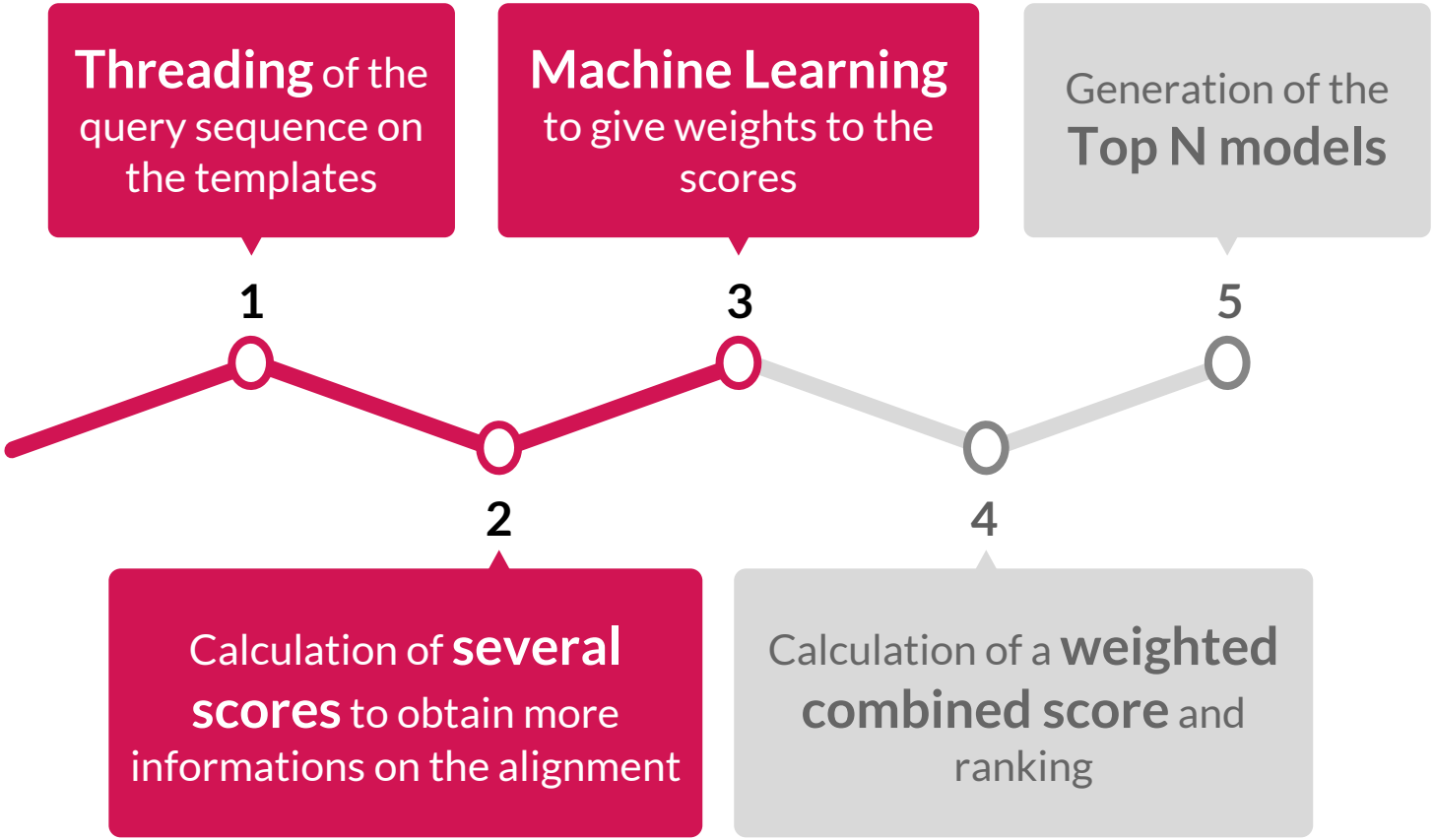
Fold-U's strategy



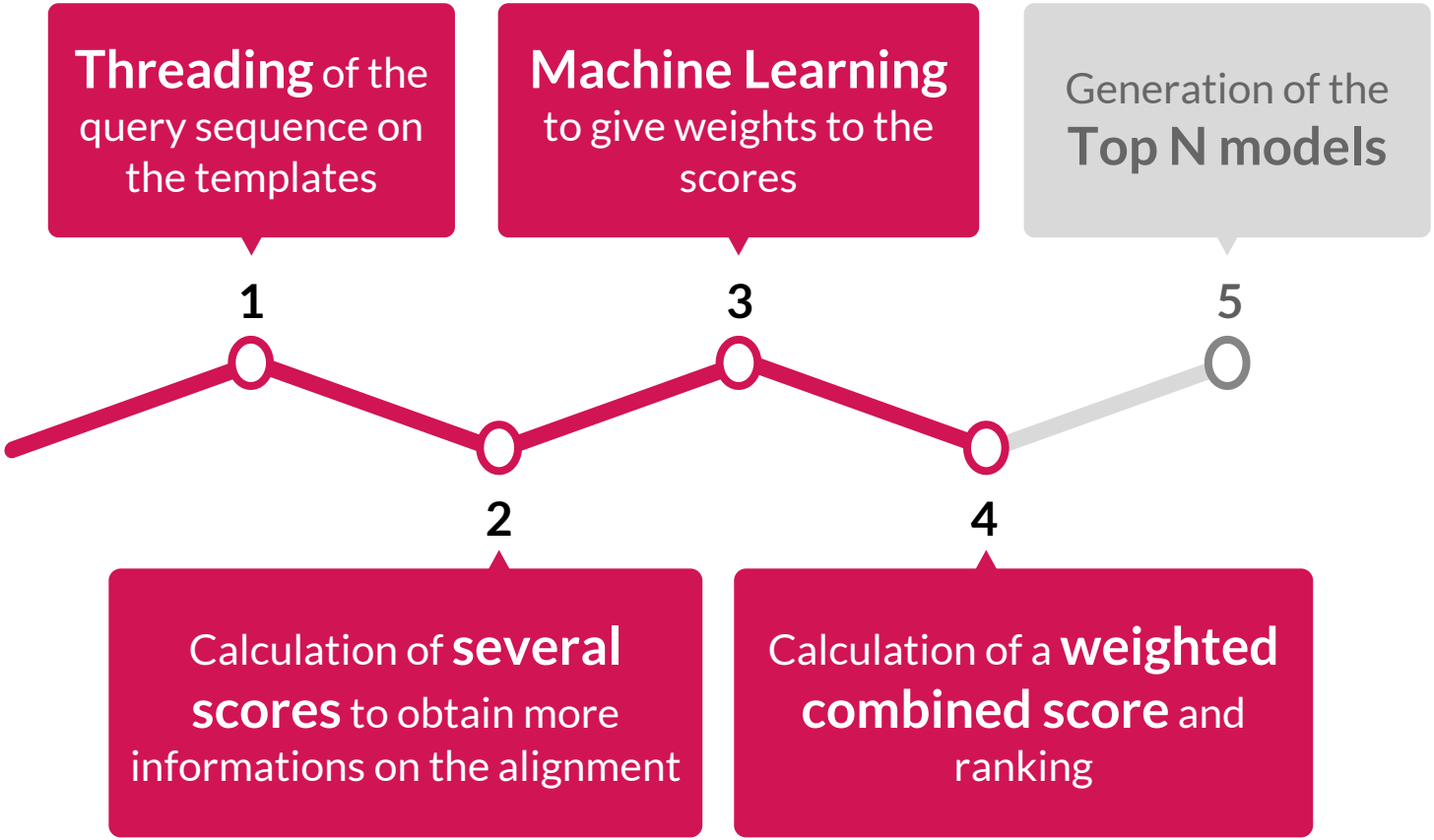
Fold-U's strategy



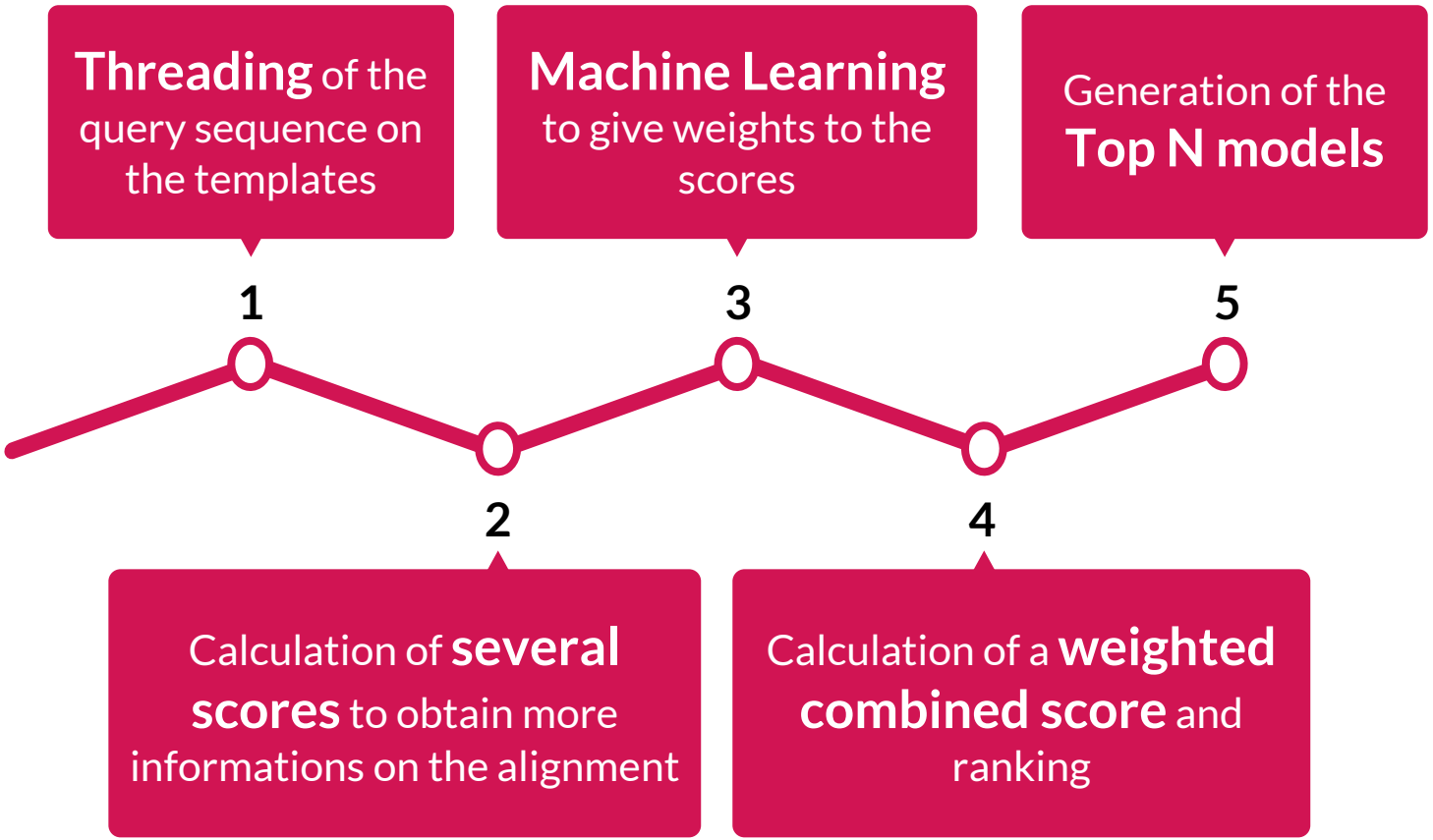
Fold-U's strategy

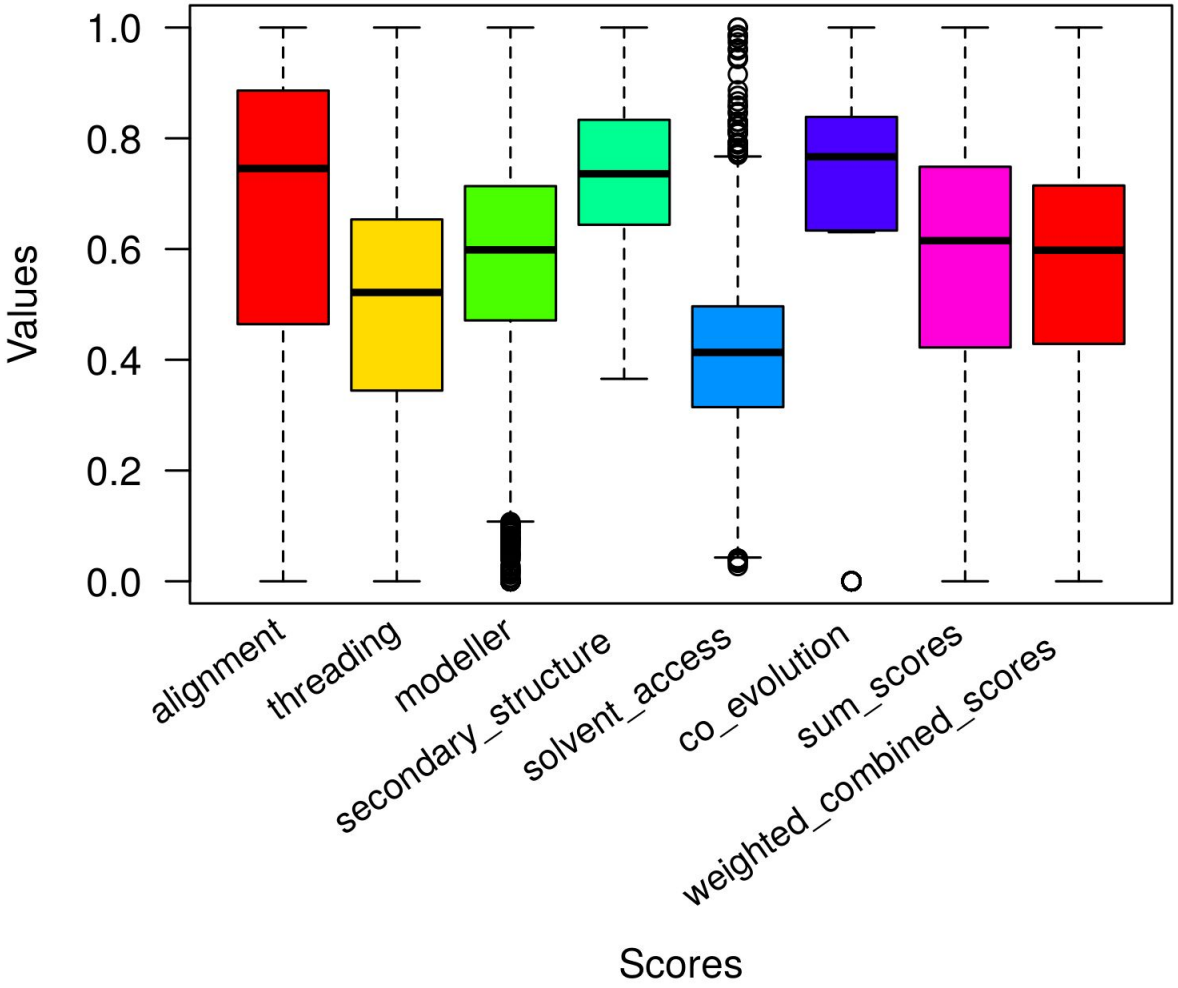


Fold-U's strategy



Fold-U's strategy




1.0
0.8
0.6
0.4
0.2
0.0
Values
alignment
threading
modeller
secondary_structure
solvent_access
co_evolution
sum_scores
weighted_combined_scores
Scores