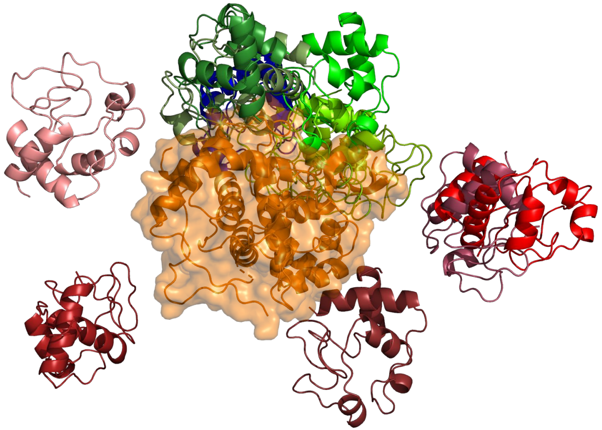
MeetDockOne

Team 1 Scoring: A machine learning strategy



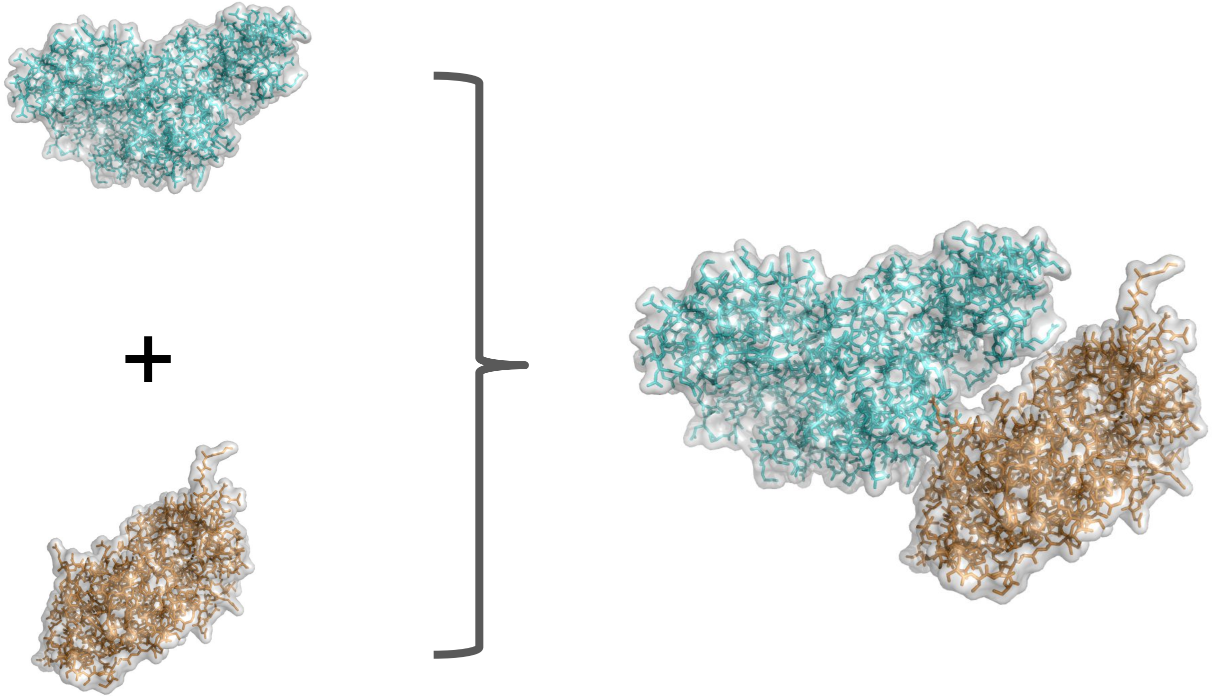
Paula Milán Rodríguez, François Gravey, Guillaume Delevoye, Ilyes Abdelhamid, Maxime Borry

université PARIS DIDEROT PARIS 7

Meet-U
a meeting story

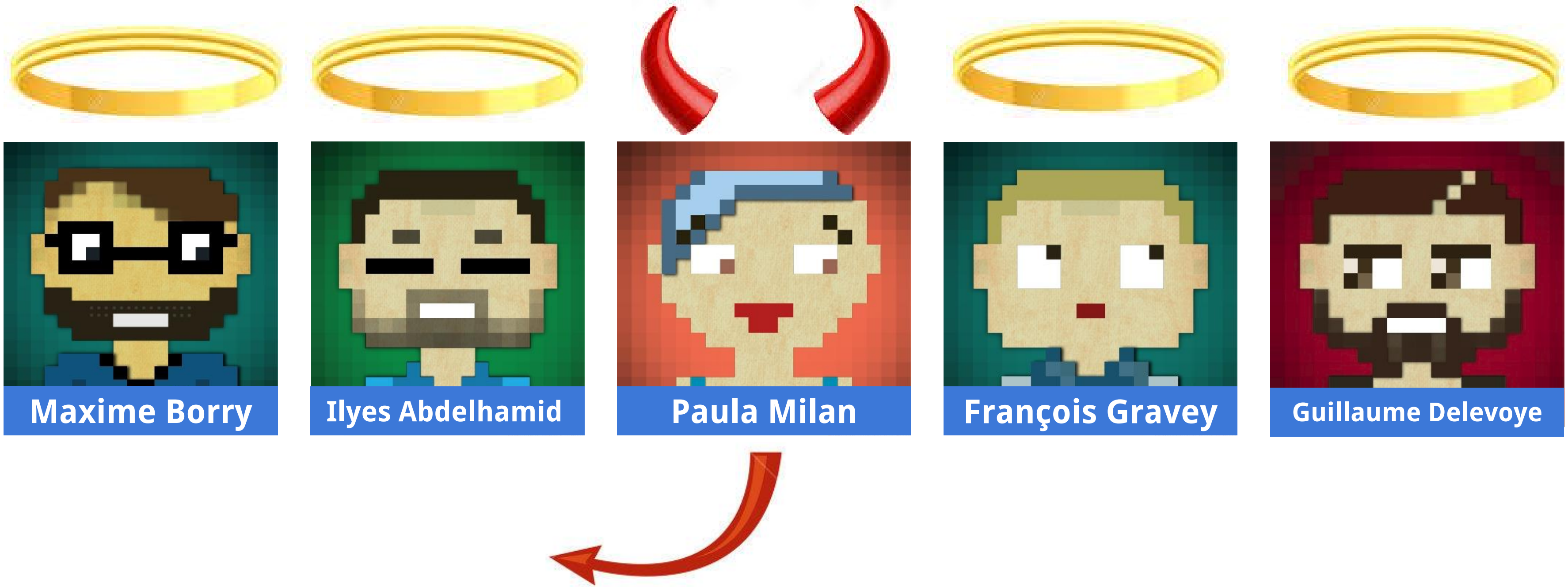
MASTER 2 Biologie Informatique

MEET THE PROBLEMATIC



BUT HOW ??

MEET US: PAULA’S ANGELS



Maxime Borry

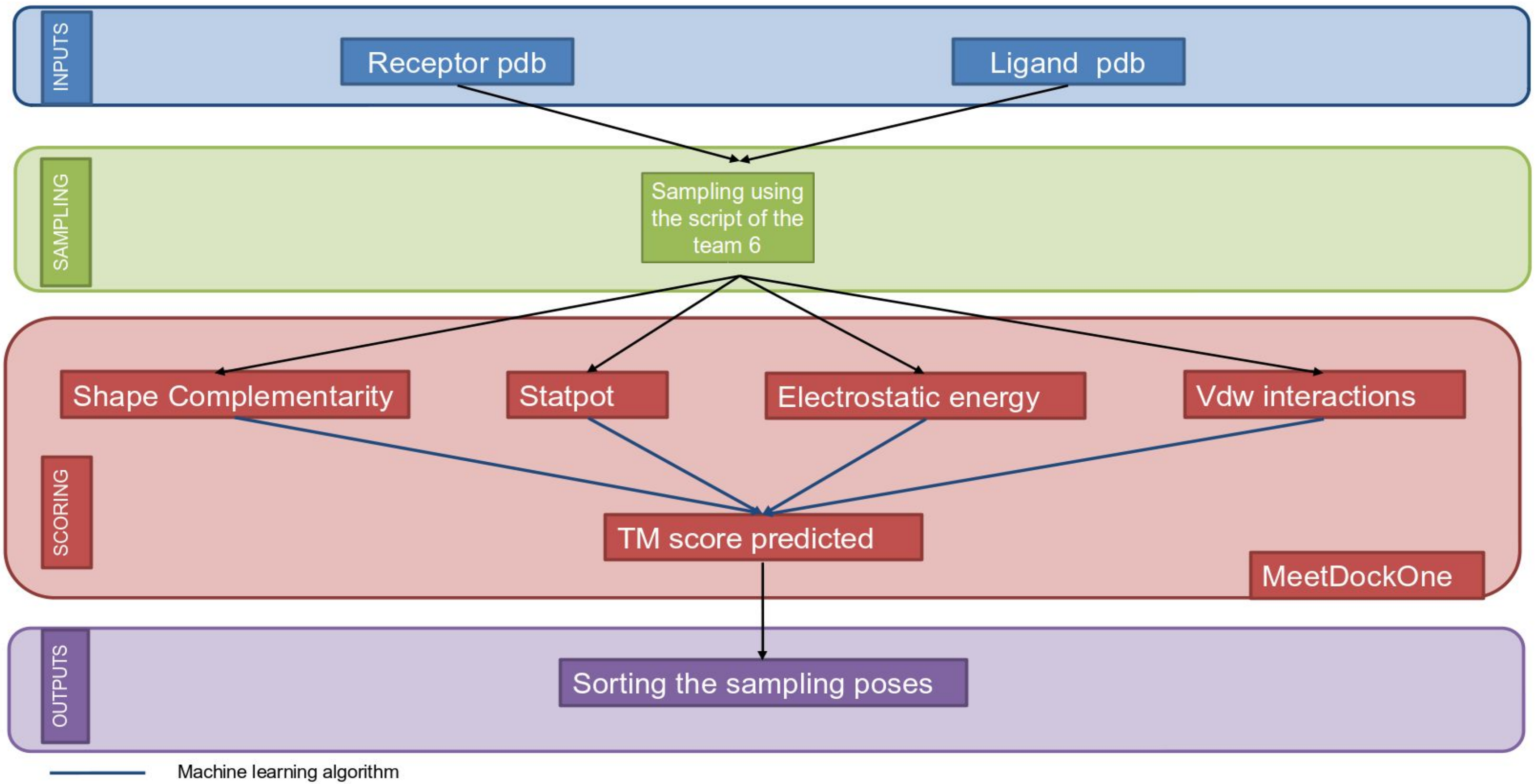
Ilyes Abdelhamid

Paula Milan

François Gravey

Guillaume Delevoye

MEET THE PIPELINE



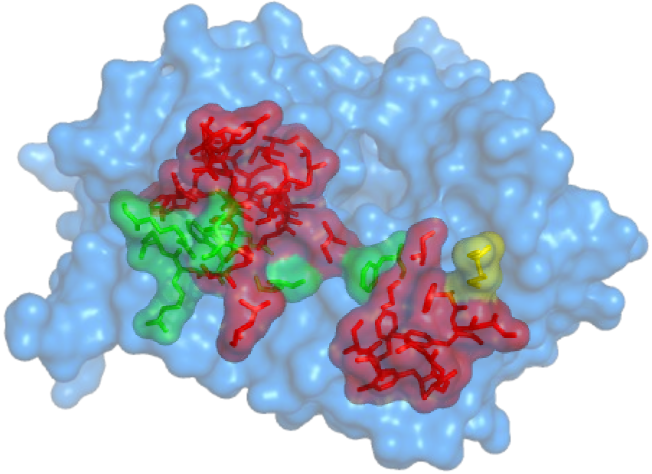
MEET THE SCORING

SCORING - RESIDUE DEPTH

Two methods :

MSMS -> adjustable cut-off (4Å by default)

NACCESS -> Accessible Surface Area (ASA)
Surface identified by the residue whose relative ASA is at least 25% of the total residue surface.



Surface representation of 2ZA4

Statistical Potential

Glaser et al.'s knowledge-based method

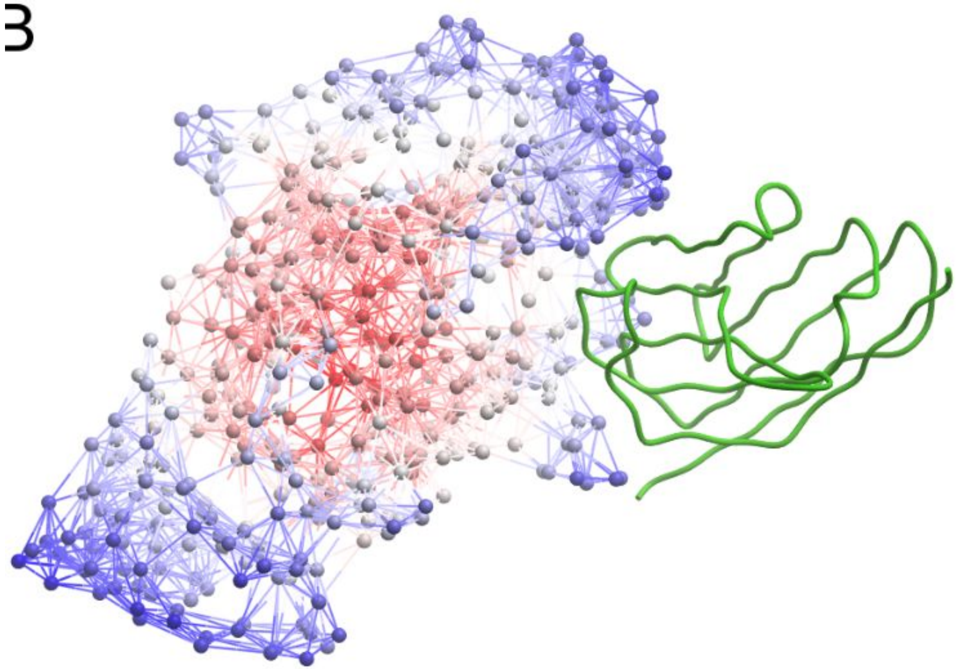
Reference : Pons, C., Glaser, F., and Fernandez-Recio, J. (2011). Prediction of protein-binding areas by small-world residue networks and application to docking. BMC Bioinformatics 12:378.

Methodology:

- Characterizing unbound proteins as networks
- Using different topology measures to predict protein protein binding sites
- Integrating these measures in pyDock (a docking scoring algorithm based on physico-chemical terms

Combining pyDock and network-based scoring :

$$pyDockCloseness_{p}^{d} = pyDock_{p} + wCloseness_{p}^{d}$$



Example of protein binding site predictions
Pons et al. BMC Bioinformatics 2011, 12:378

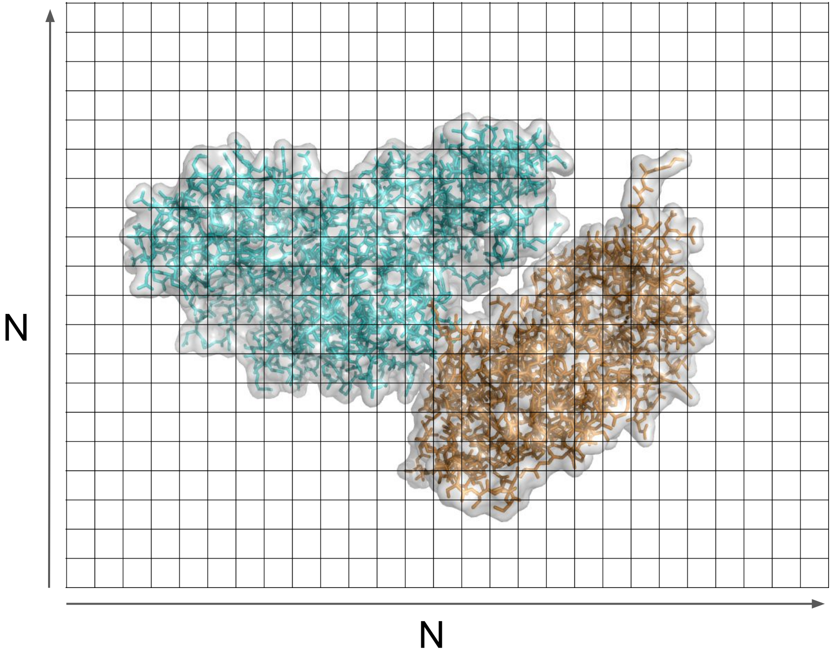
SHAPE COMPLEMENTARITY

- **N** x **N** x **N** grid
- grid point (**l**, **m**, **n** = 1, 2, ... **N**)
- **grid spacing** 2 Å
- **N** large enough for **R** and **L**

$$R_{SC}(l,m,n) = \begin{cases} 1 & \text{surface of } R \\ \rho i & \text{core} \\ 0 & \text{open space} \end{cases}$$

$$L_{SC}(l,m,n) = \begin{cases} 1 & \text{surface of } L \\ \rho i & \text{core} \\ 0 & \text{open space} \end{cases}$$

$i = \sqrt{-1}$ $\quad \rho = 9.$



ELECTROSTATIC AND VDW

Electrostatic energy

$$V\,elec = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{qi\,qj}{4\,\pi\,\varepsilon 0\,rij}$$

Computed for residues:
TYR, HIS, CYS, ASP, GLN, LYS, ARG

Lennard-Jones potential

$$V_{ij} = 4\varepsilon \left[\left(\frac{\theta}{r_{ij}} \right)^{12} - \boxed{\left(\frac{\theta}{r_{ij}} \right)^{6}} \right]$$

Default values:

ε = 10
θ = 3.9

Van der Waals force

MACHINE LEARNING

Goal : Predict the Tm score from the other MeetDockOne function

scikits learn
machine learning in Python

Learning data : 17 native complexes; 5936 decoys

Data preparation : Scikits learn pipeline

- Missing values : Sklearn Imputer, Median strategy
- Normalisation : StandardScaler

Three machine learning algorithms :

- Linear regression
- Decision tree regressor
- Random forest regressor

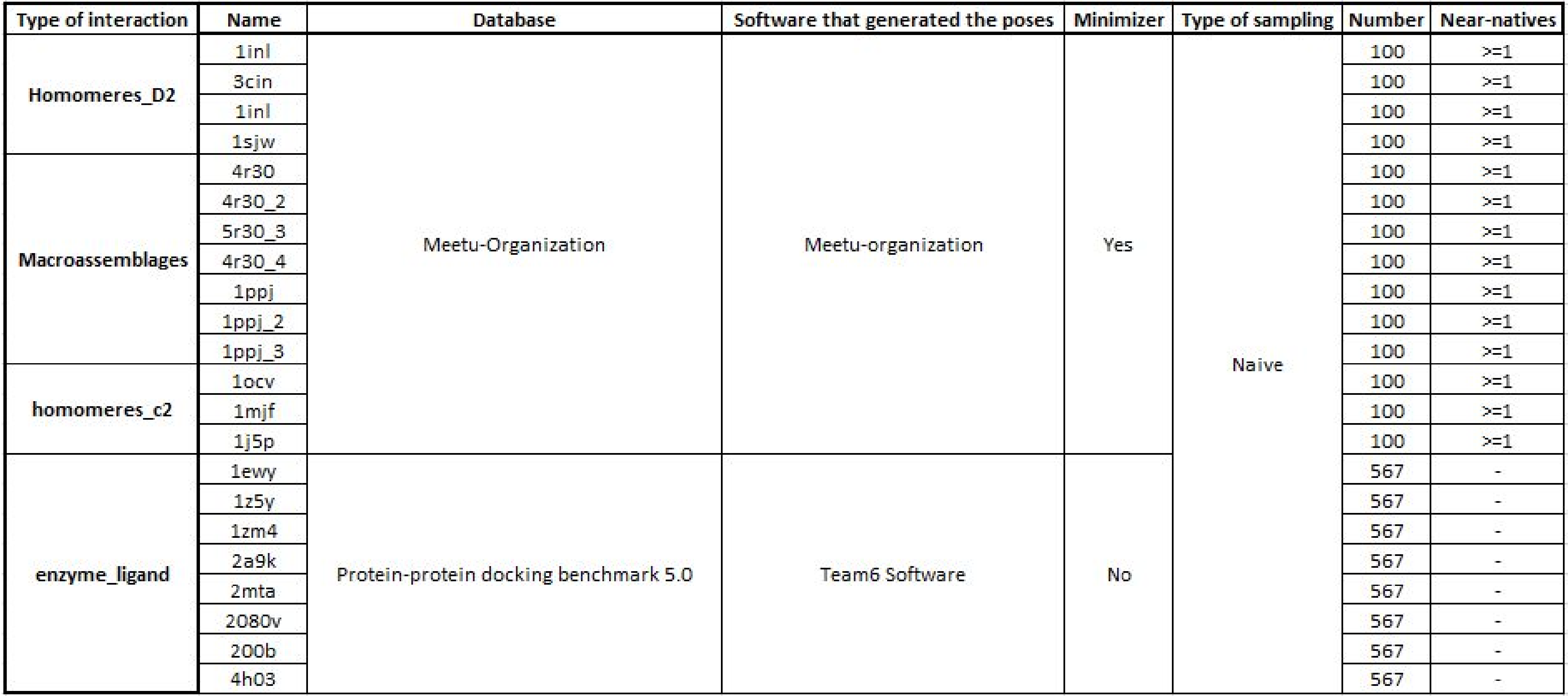
Algorithms evaluation : Cross Validation

- Mean squared error
- ‘K fold’ cross validation (n = 10)

Best hyper-parameters values : GridSearchCV

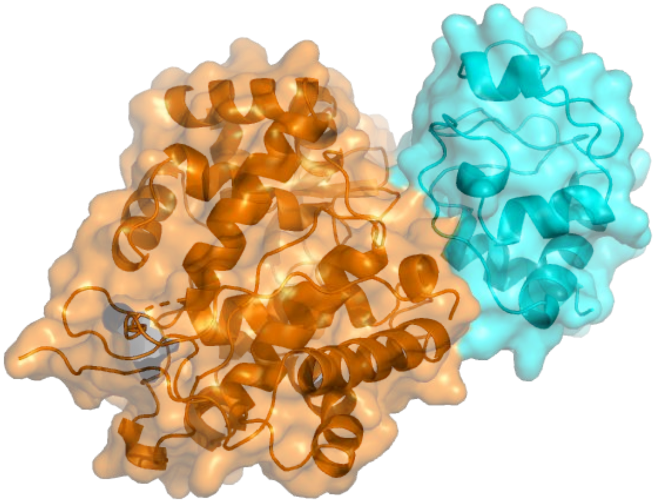
- Number of trees (n=10, 50 or 100)
- Number of features (n=1, 2, 3, or 4)
- Bootstrap (True or False)

MACHINE LEARNING - Learning Dataset

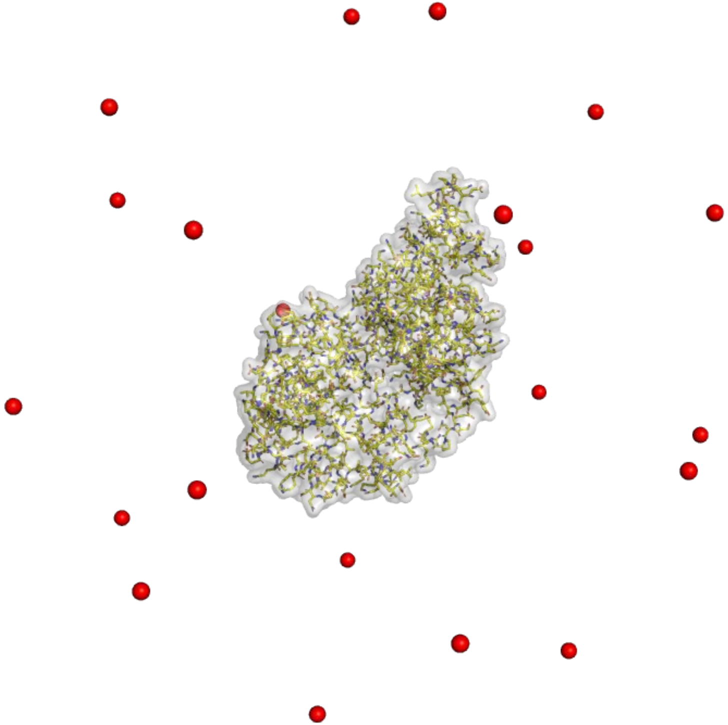


Type of interaction	Name	Database	Software that generated the poses	Minimizer	Type of sampling	Number	Near-natives
Homomeres_D2	1inl	Meetu-Organization	Meetu-organization	Yes	Naive	100	>=1
	3cin					100	>=1
	1inl					100	>=1
	1sjw					100	>=1
Macroassemblages	4r30					100	>=1
	4r30_2					100	>=1
	5r30_3					100	>=1
	4r30_4					100	>=1
	1ppj					100	>=1
	1ppj_2					100	>=1
	1ppj_3					100	>=1
homomeres_c2	1ocv					100	>=1
	1mjf					100	>=1
	1j5p					100	>=1
enzyme_ligand	1ewy	Protein-protein docking benchmark 5.0	Team6 Software	No		567	-
	1z5y					567	-
	1zm4					567	-
	2a9k					567	-
	2mta					567	-
	2080v					567	-
	200b					567	-
	4h03					567	-

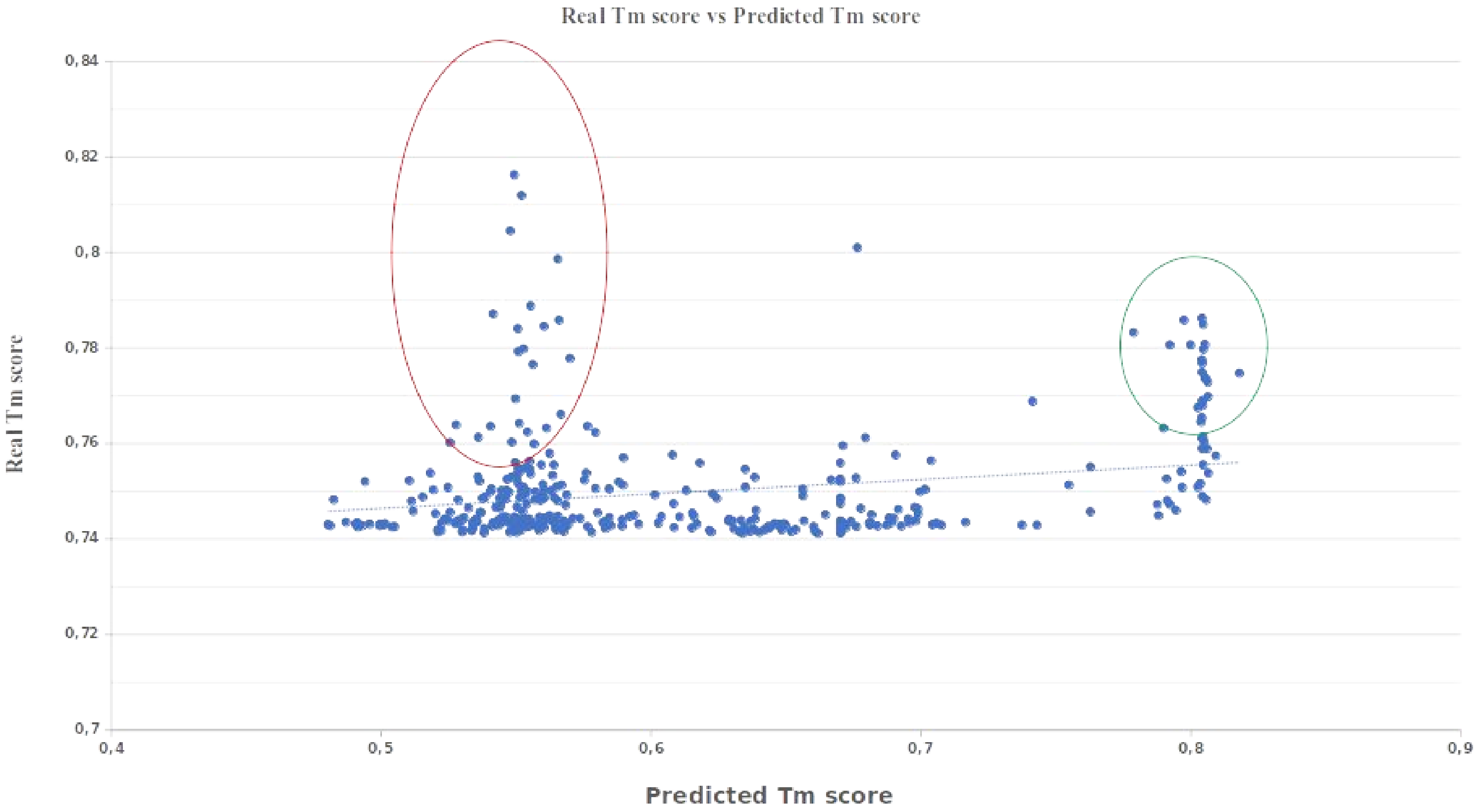
A-Z WITH 2PCC



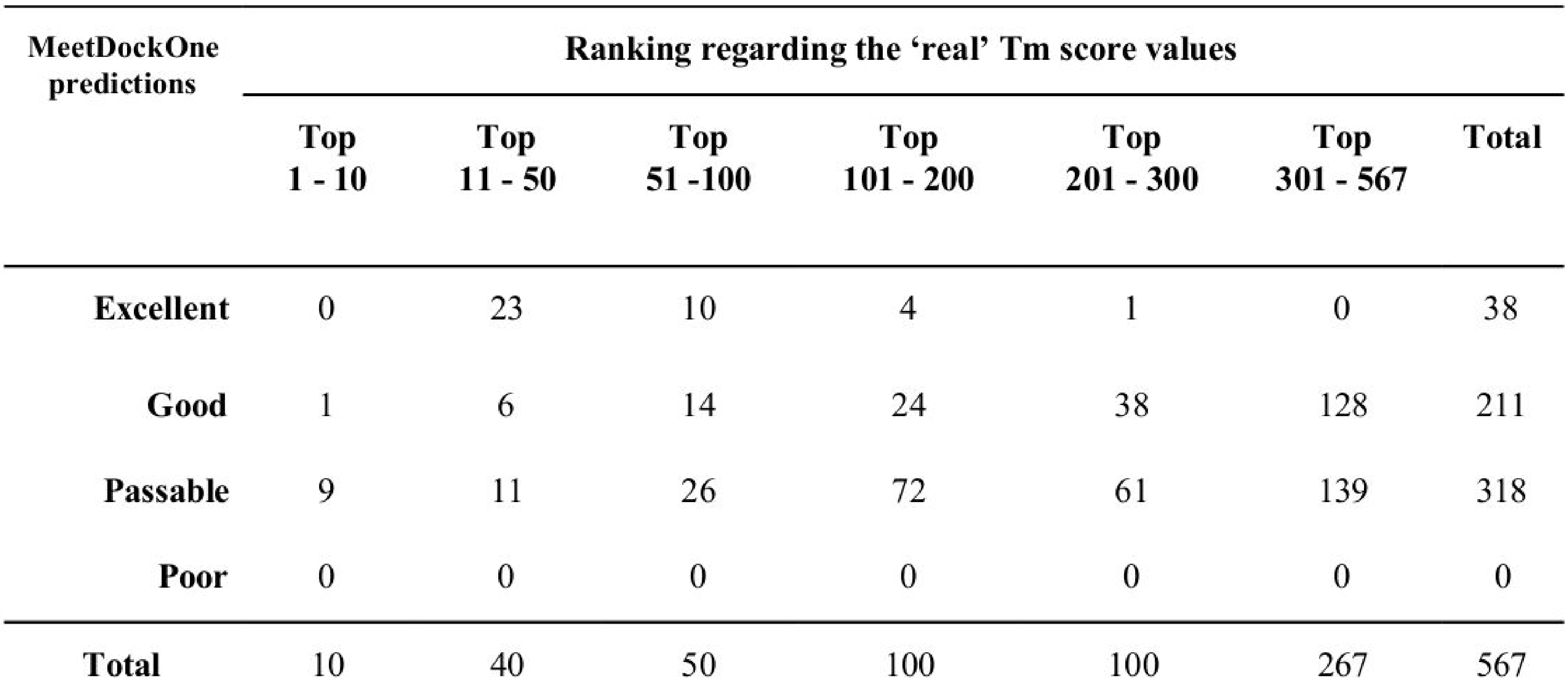
CYTOCHROME C PEROXIDASE AND CYTOCHROME C



2PCC RESULTS

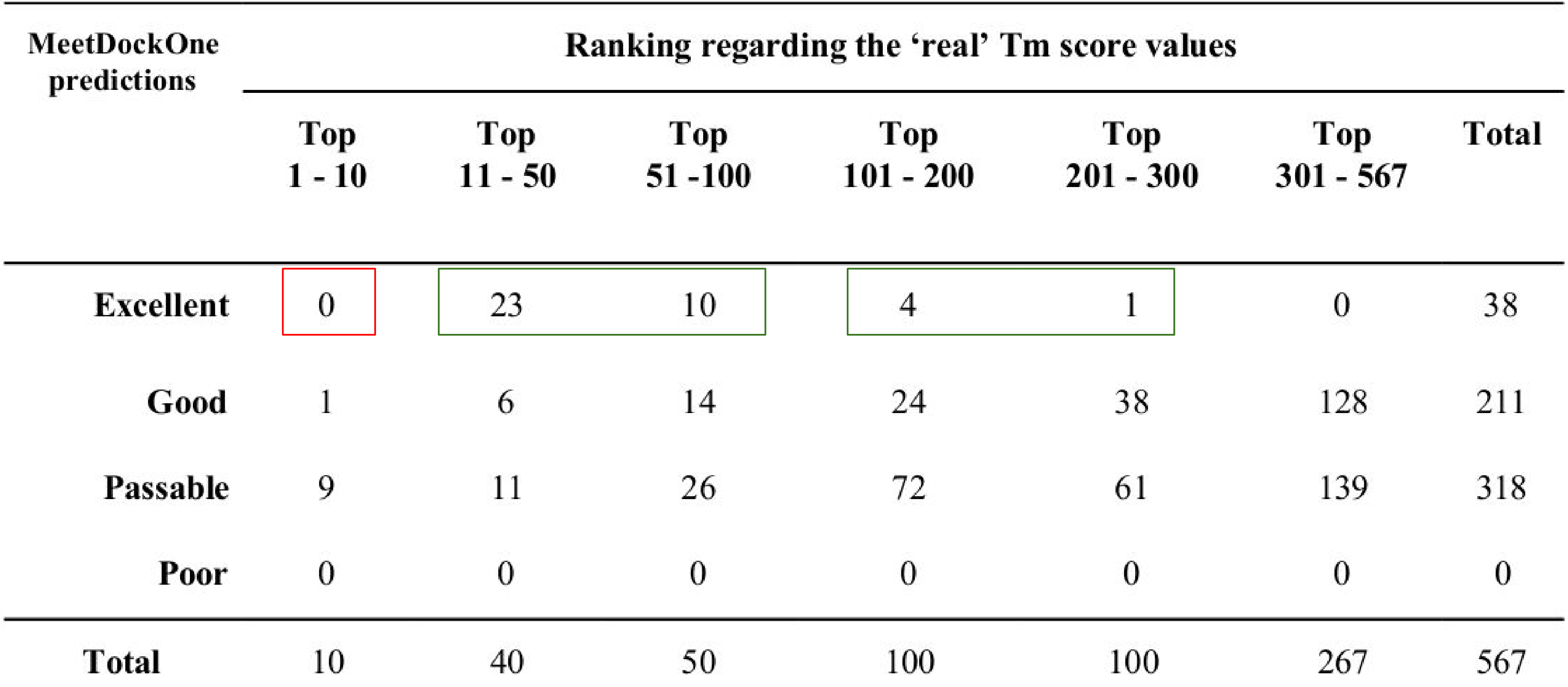


2PCC RESULTS



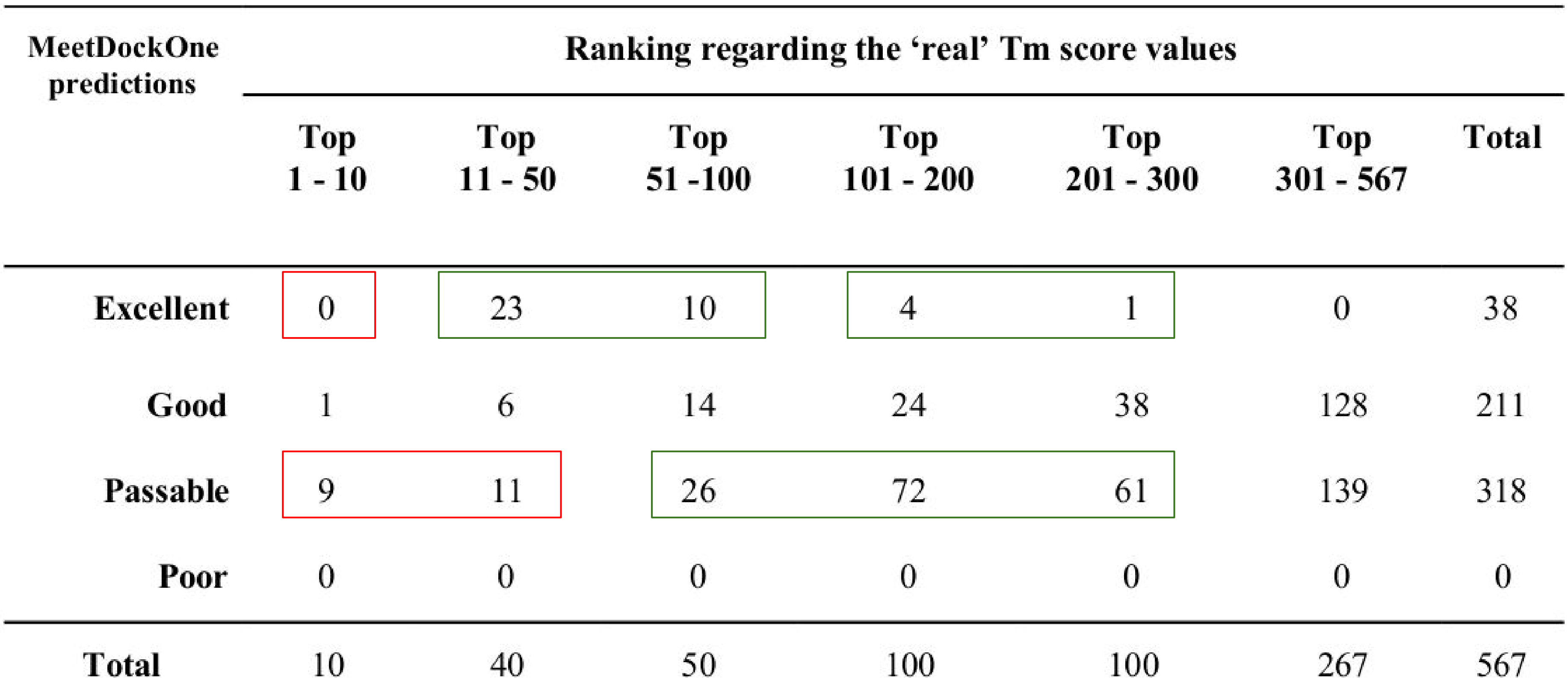
MeetDockOne predictions	Ranking regarding the ‘real’ Tm score values						
	Top 1 - 10	Top 11 - 50	Top 51 -100	Top 101 - 200	Top 201 - 300	Top 301 - 567	Total
Excellent	0	23	10	4	1	0	38
Good	1	6	14	24	38	128	211
Passable	9	11	26	72	61	139	318
Poor	0	0	0	0	0	0	0
Total	10	40	50	100	100	267	567

2PCC RESULTS



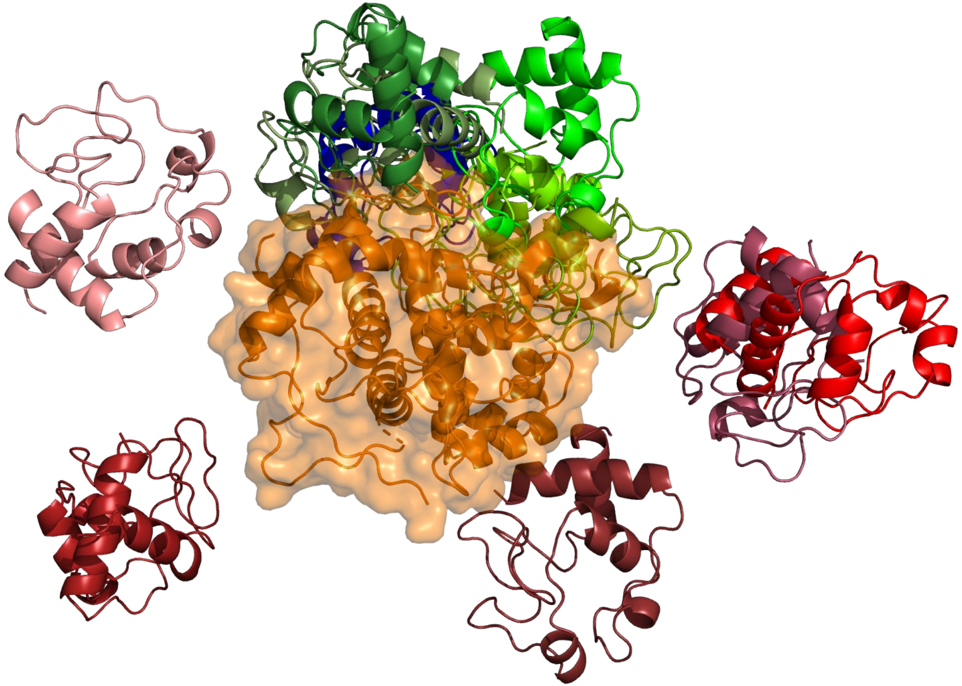
MeetDockOne predictions	Ranking regarding the ‘real’ Tm score values						
	Top 1 - 10	Top 11 - 50	Top 51 -100	Top 101 - 200	Top 201 - 300	Top 301 - 567	Total
Excellent	0	23	10	4	1	0	38
Good	1	6	14	24	38	128	211
Passable	9	11	26	72	61	139	318
Poor	0	0	0	0	0	0	0
Total	10	40	50	100	100	267	567

2PCC RESULTS



MeetDockOne predictions	Ranking regarding the 'real' Tm score values						
	Top 1 - 10	Top 11 - 50	Top 51 -100	Top 101 - 200	Top 201 - 300	Top 301 - 567	Total
Excellent	0	23	10	4	1	0	38
Good	1	6	14	24	38	128	211
Passable	9	11	26	72	61	139	318
Poor	0	0	0	0	0	0	0
Total	10	40	50	100	100	267	567

2PCC RESULTS



Best predicted pose

Predicted TM-score	Real TM-score
0.81	0.77

Best pose

Predicted TM-score	Real TM-score
0.55	0.81

CONCLUSION : limitations / opportunities of improvement

Limitations:

Issues with the minimizer : many distant poses which decreased the performance of our algorithm (important distances resulted in having many “0” values)

Few protein-protein interactions nature included into our “machine learning database”

“Low” throughput technique ~ 10 seconds per pose

Opportunities of improvement:

Add more features into our code :

- Desolvatation energy
- Sequence alignment docking

Docking using flexible binding domain for receptor

CONCLUSION : Added value

Enriched input data : original scoring function based on several approaches (complementarity, knowledge based, energetic)

The use of a machine learning program increases the reliability of our results

Multithreading coding allowed us to minimize running time