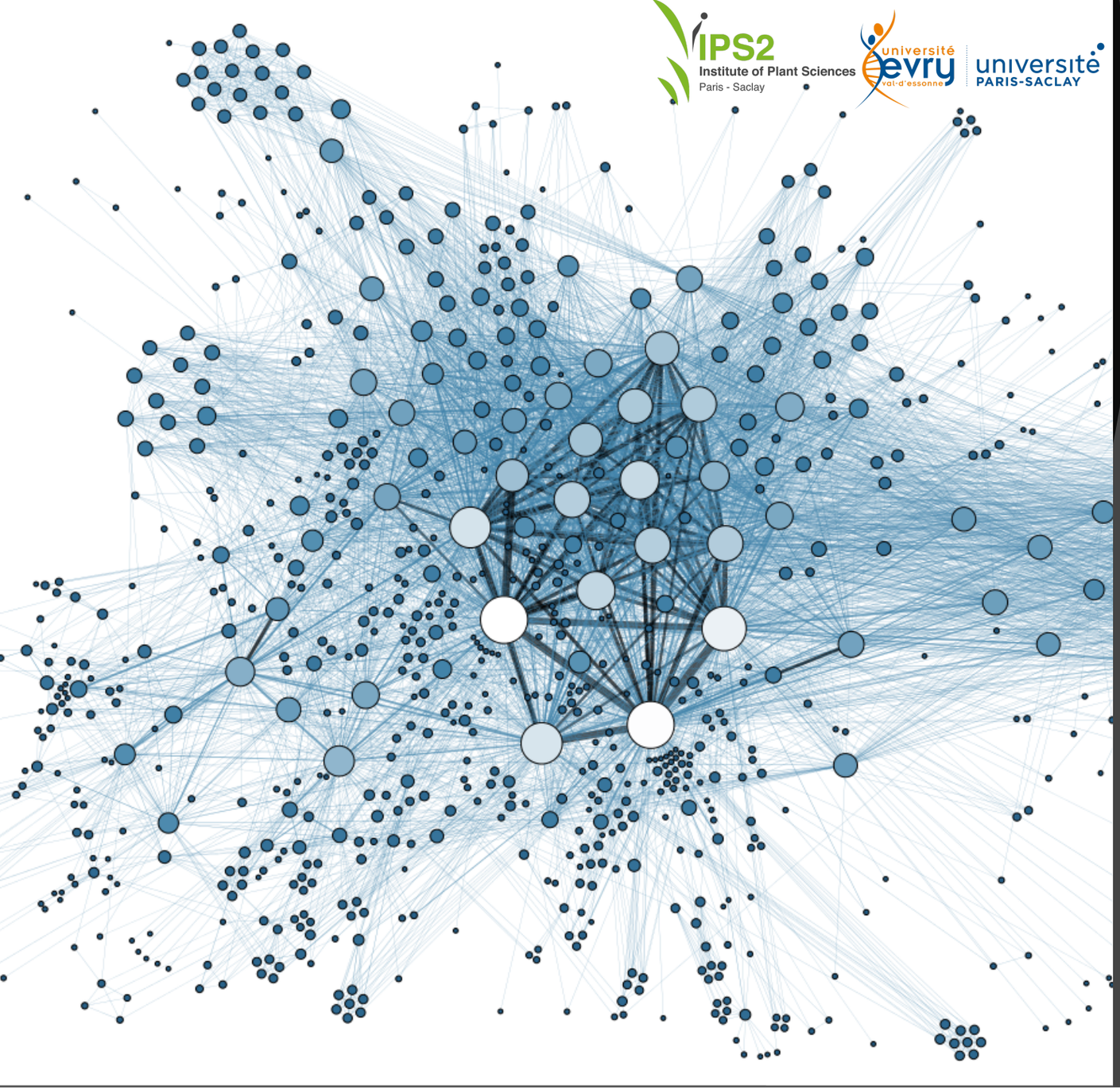
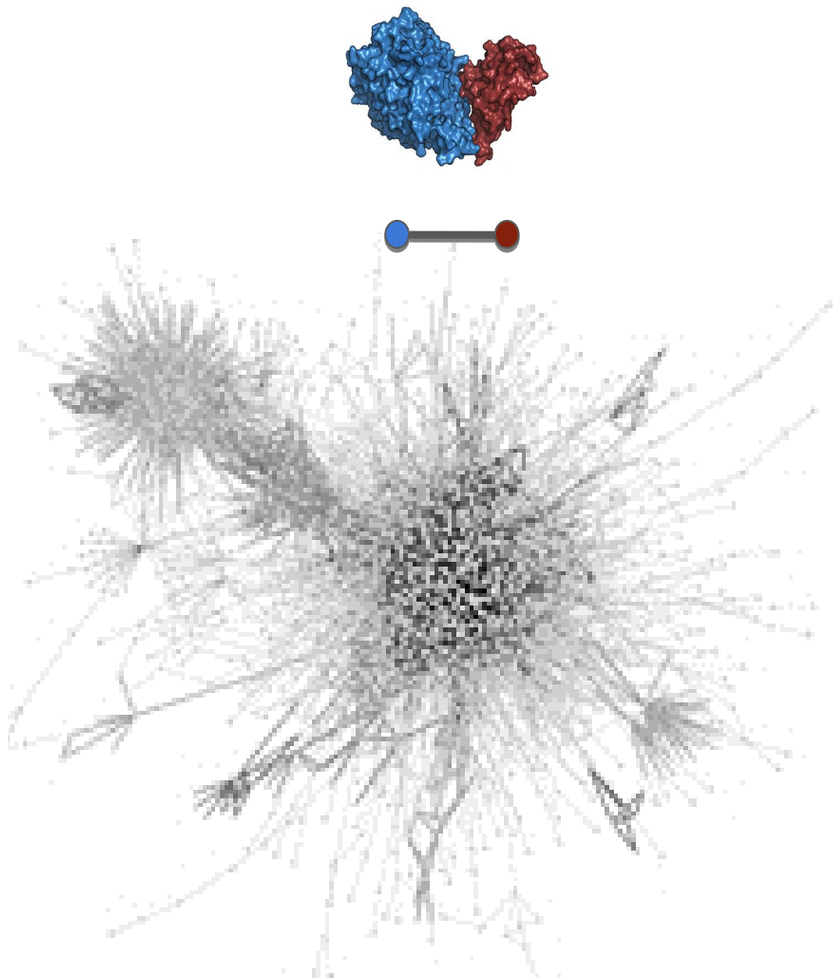


Prediction of interacting protein pairs and impact on the A. thaliana interactome

Marie-Hélène Mucchielli-Giorgi, Institute of Plant Sciences Paris-Saclay, Université d'Evry - Paris Saclay

Journées NetBio 2024, novembre 12, 2024

Search for clusters in a Protein-Protein Interaction (PPI) network: a way to annotate a proteome



A. thaliana interactome

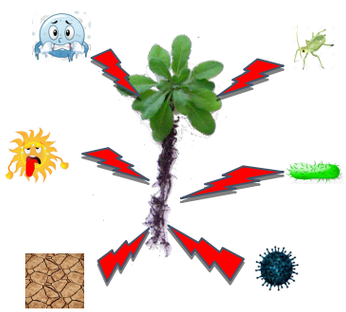
- A network that seems unreadable

but

- A network that is not randomly built

Search for clusters in a Protein-Protein Interaction (PPI) network: way to annotate a proteome

A. thaliana interactome



one color = one stress condition

- Proteins playing a role in the same process are highly interconnected
- They form sub-networks with a high density of interactions.

- Searching for clusters in a PPI network : a way to identify all proteins belonging to the same biological process

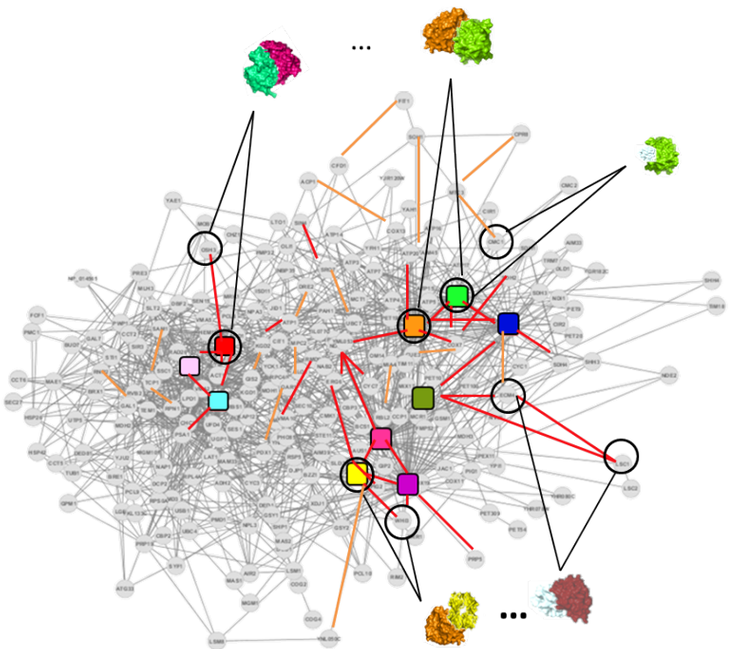
More annotated proteins

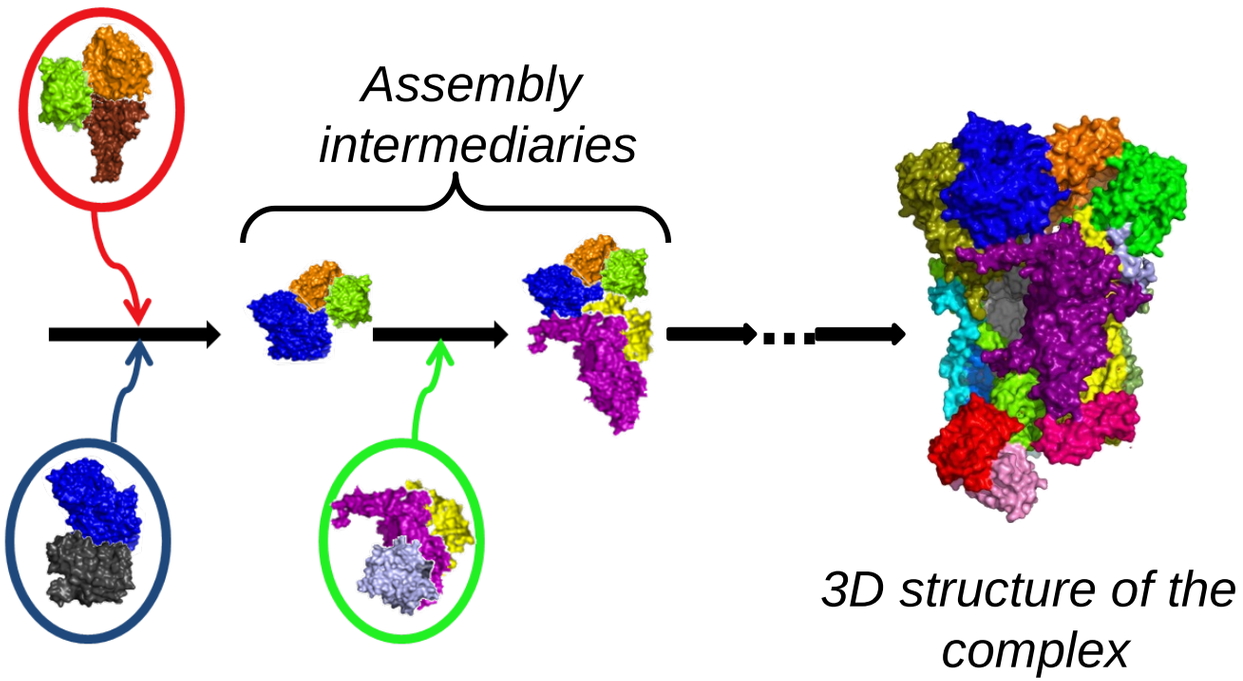
- Proof of concept on *S. cerevisiae*

Identification of a new protein involved in biogenisis of bc1 complex of the respiratory chain of S. Cerevisiae: the protein USB1 (Glatigny et al, BMC Sys. Biol. 2011).

Search for clusters in a Protein-Protein Interaction (PPI) network: a way to model protein complex assembly

Proteins in an assembly intermediate have more common partners with each other than with other proteins in the complex



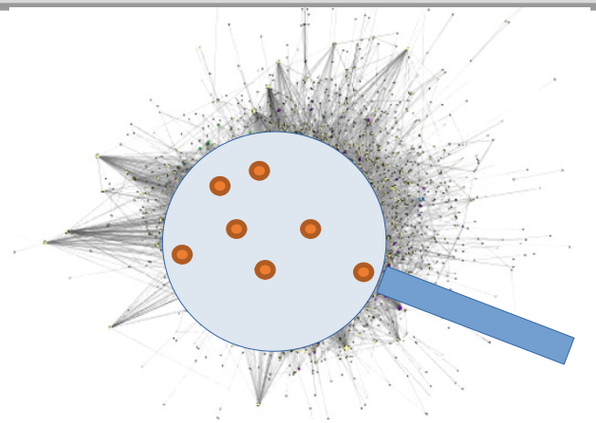


Glatigny et al. BMC Syst Biol. 2017 Jul 11;11(1):67.

The interactome quality: a limit for the clustering methods

137 690 PPIs for 27 469 *A. thaliana* proteins

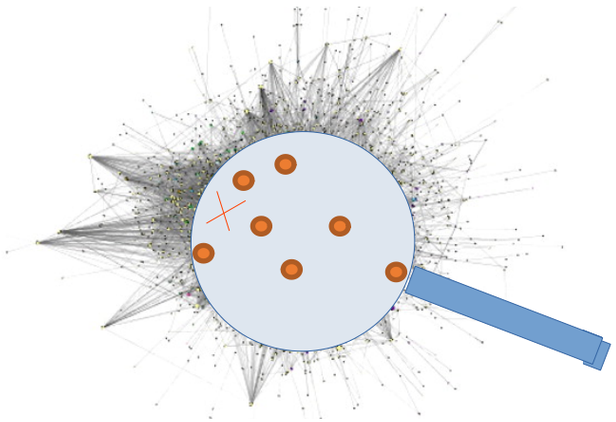
-> **Lots of missing PPIs** in the interactome of A. thaliana



Add new PPIs

PPIs from "high-throughput" experiences

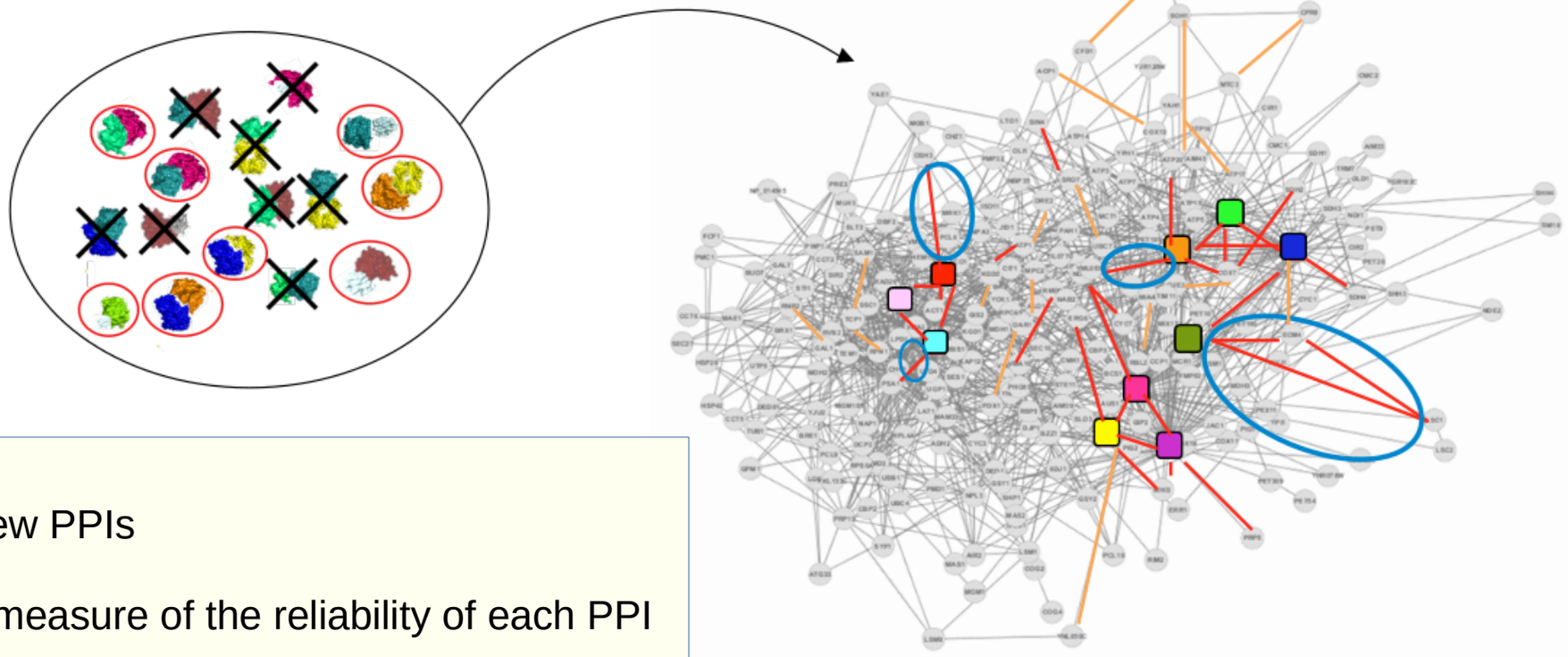
-> **Lots of false positives**



Remove false positives

The prediction of Interacting Protein Pairs (IPP) : a way to a more reliable network

prediction of the partners of network proteins

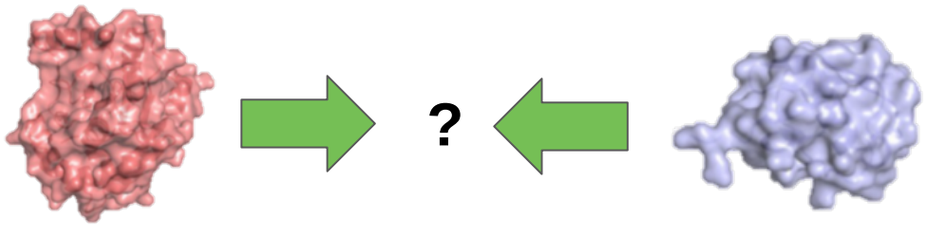


- New PPIs

- A measure of the reliability of each PPI

-Elimination of numerous false positives

How can we predict interacting protein pairs (IPP) ?

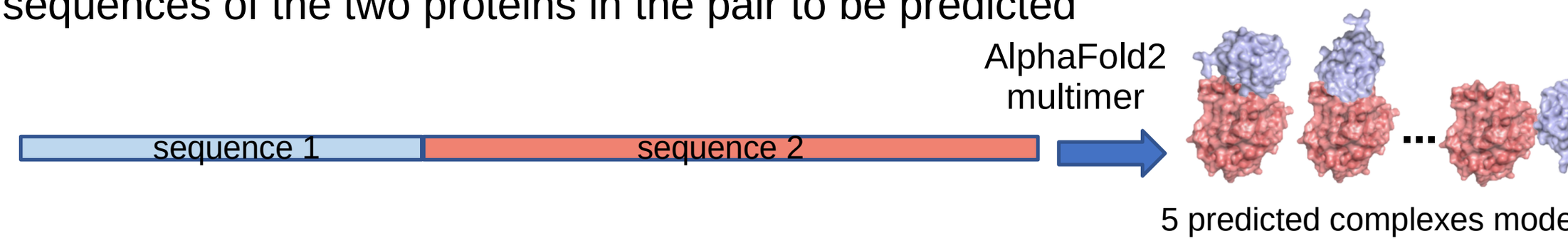


- **How ?**

- By using results provided by the deep learning approaches AlphaFold2 version Multimer,

- **What kinds of data are used ?**

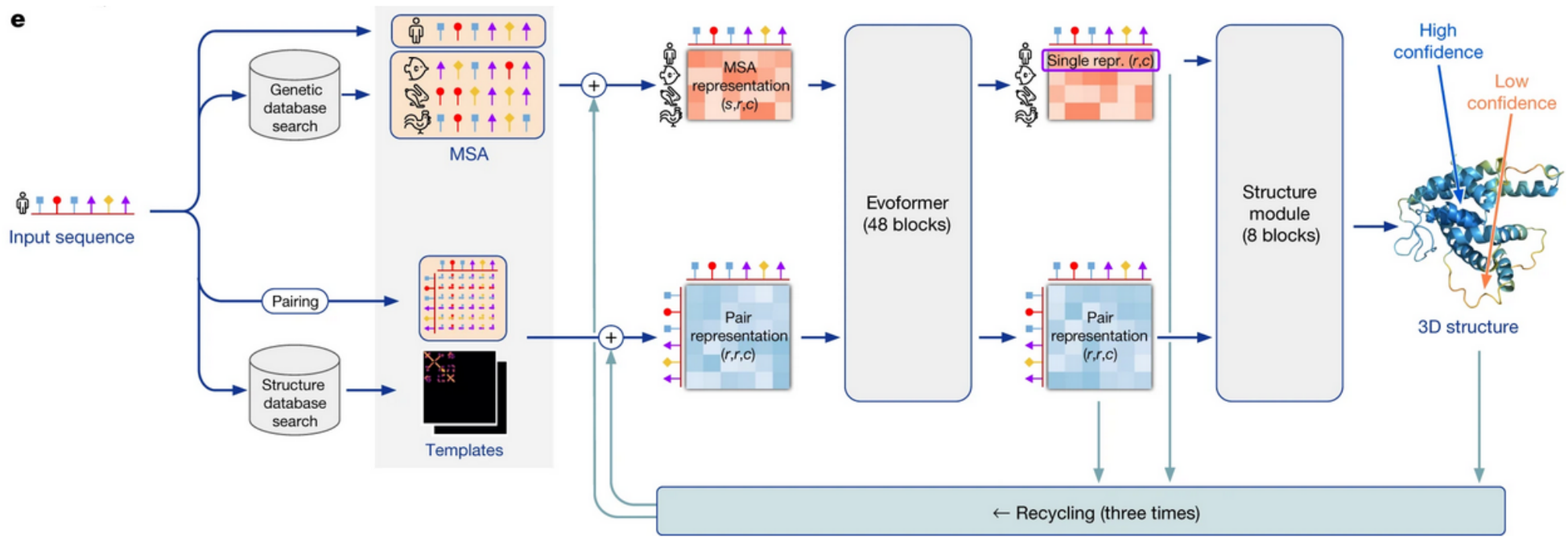
  The sequences of the two proteins in the pair to be predicted



-

Protein complex prediction with AlphaFold-Multimer (AlphaFold2 learned on protein complexes)

doi: https://doi.org/10.1101/2021.10.04.463034
BioRxiv, October 5, 2021



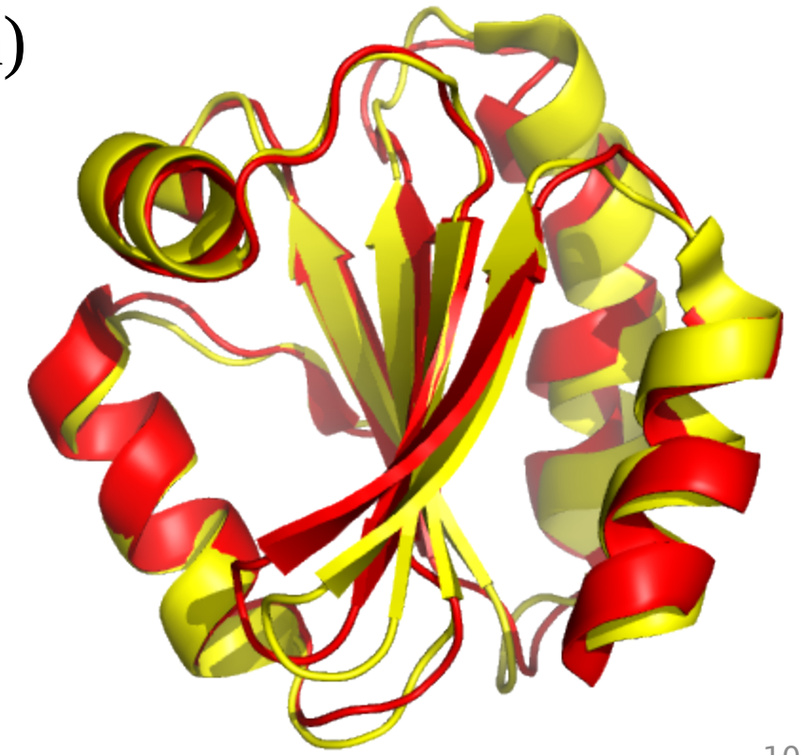
Quality scores used for the prediction of interacting protein pairs

Quality scores of the predicted structure provided by AlphaFold2,used for IPP prediction

- pTM (predicted Template Modeling score)
- pLDDT (predicted Local Distance Difference Test)
- PAE (Predicted Alignment Error)
- ipTM (interface predicted Template Modeling score)
- contact probabilities

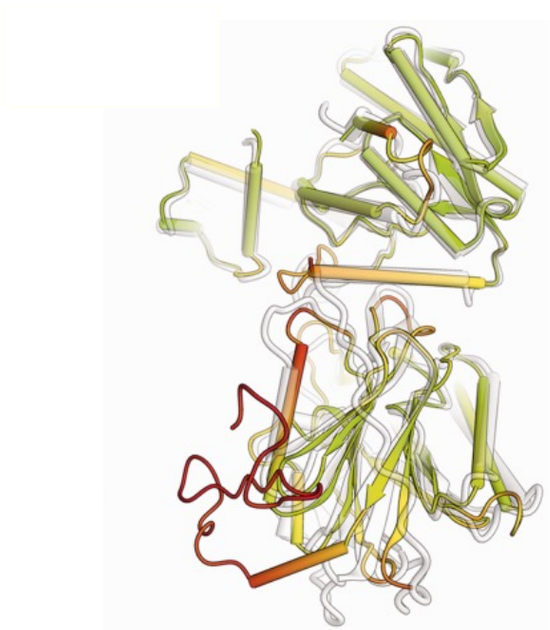
pTM (predicted Template Modeling score)

- The TM score measures the difference between the experimental structure and the predicted structure, normalized by protein length.
- Varies from 0 to 1 (1 being a perfect match)
- pTM is a predicted TM score



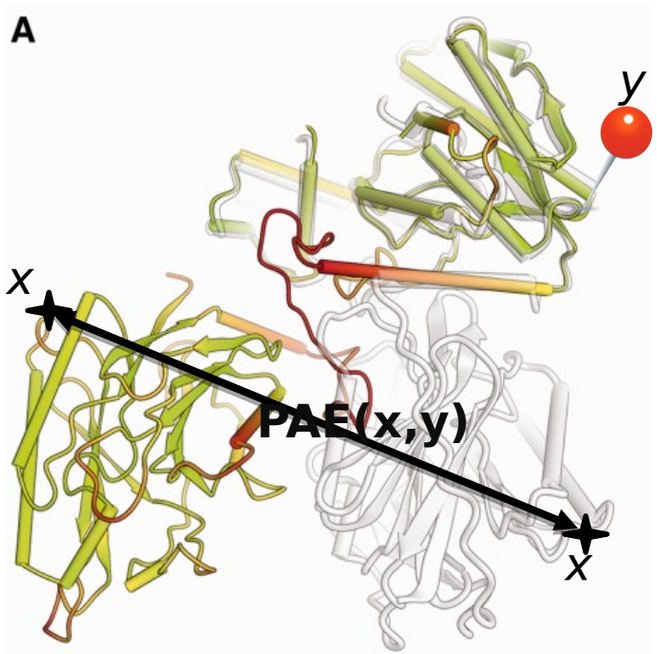
pLDDT (predicted Local Distance Difference Test)

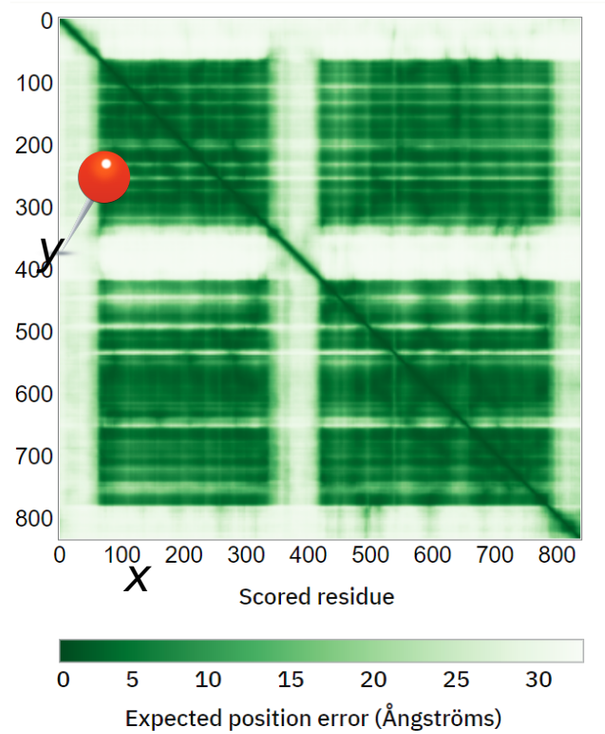
- LDDT locally compares experimental structure and prediction
- Gives a measure of the quality of the prediction of each amino acid's environment
- The pLDDT is a predicted LDDT.



PAE (Predicted Alignment Error)

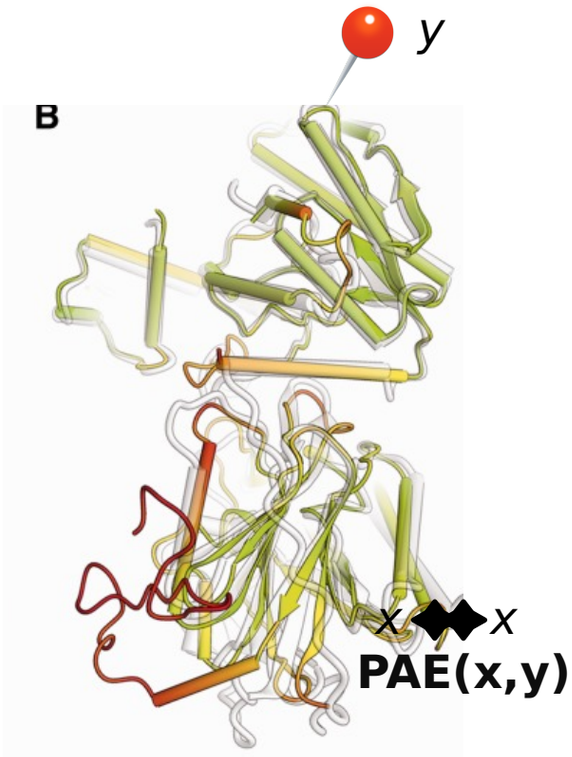
Indicates, **for each *x* position**, the difference between the experimental structure and the predicted structure **when the two structures are aligned at the *y* position**.

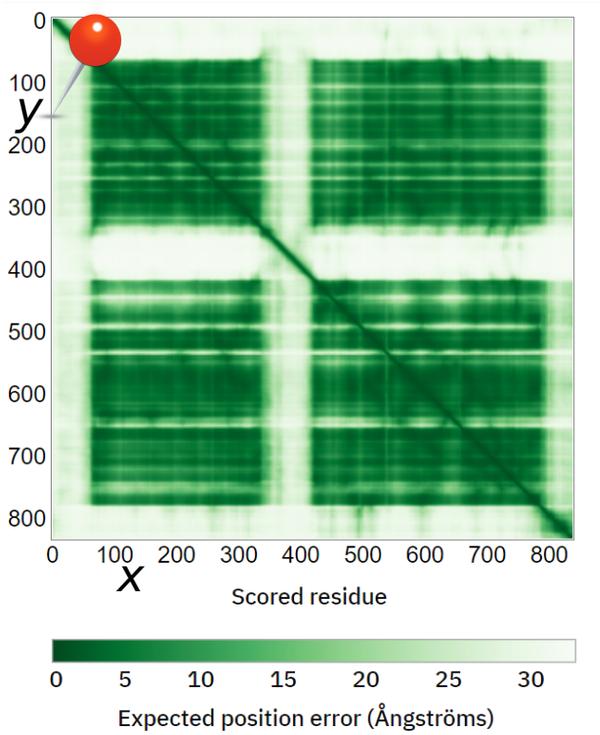




PAE (Predicted Alignment Error)

Indicates, **for each *x* position**, the difference between the experimental structure and the predicted structure **when the two structures are aligned at the *y* position**.



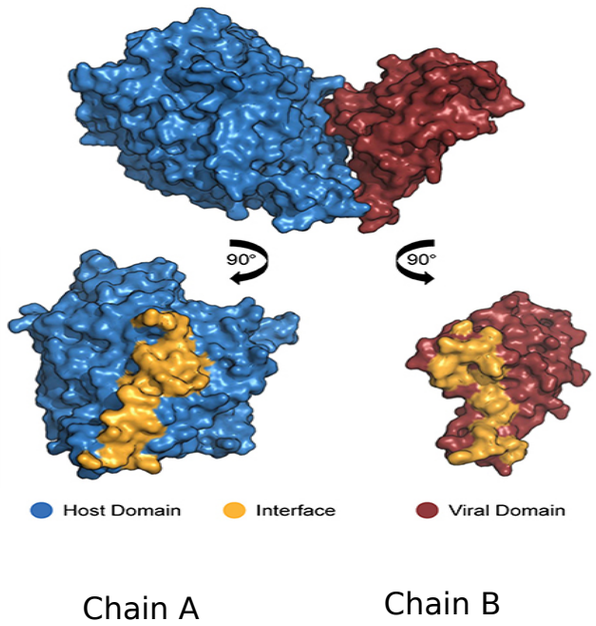


ipTM (interface predicted Template Modeling score

- ipTM (interface predicted Template Modeling score)

ITM : TM score of residues of the interface of the chain not used for the structural alignement of experimental and predicted structures on residues of th interface.

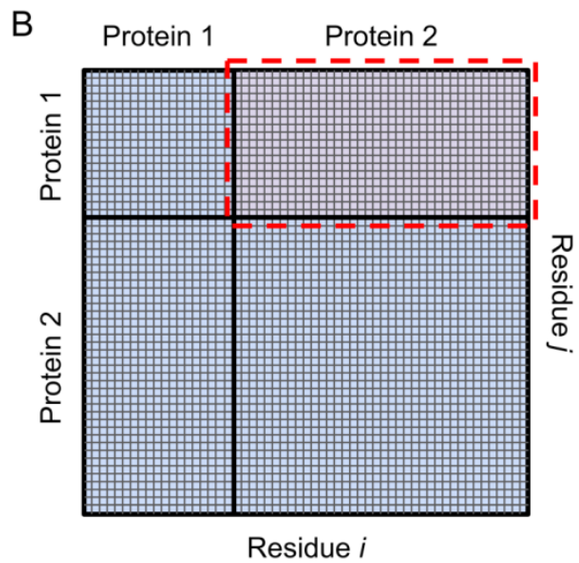
ipTM is a predicted ITM score



Model confidence for ranking the predicted complexe

$$\text{model confidence} = 0.8 \cdot \text{ipTM} + 0.2 \cdot \text{pTM}$$

Contact probabilities

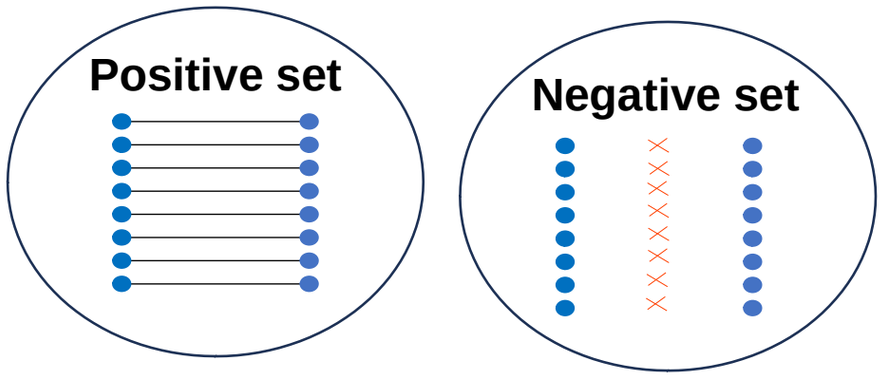


Matrix of contact probabilities : probabilities of each amino acid pair being within 12 Å

R. Humphreys *et al.*, « Computed structures of core eukaryotic protein complexes », *Science*, vol. 374, nº 6573, p. eabm4805, dec. 2021, doi: 10.1126/science.abm4805.

A first prediction method

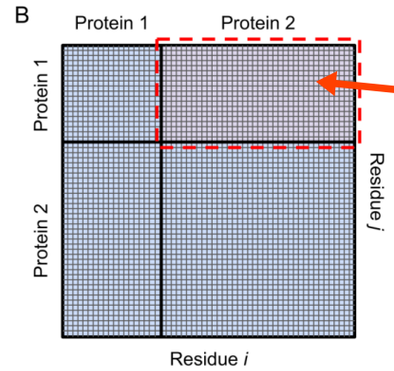
For each proteins pair of a data set



Positive set : pairs of proteins whose interaction has been experimentally proven by at least 3 different methods and published in at least two different articles.

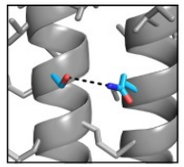
Negative set : Pairs of random proteins never identified as interacting in protein-protein interaction databases

AlphaFold2

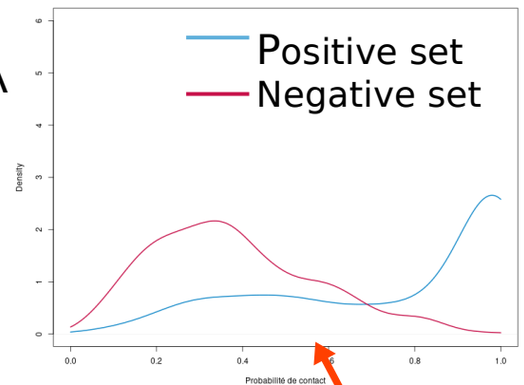


Matrix of probabilities of each amino acid pair being within 12 Å

Pmax=Maximum contact probability



Distribution of the Pmax values of all protein pairs



T

For a given threshold value T, each protein pai is predicted as an IPP if its Pmax > T and as a random pair if its Pmax < T.

Issues associated with IPP prediction

- The proportion of IPP *versus* random pairs is very low :

  for *A. thaliana* around 300 000 IPP estimated versus 771 million random pairs

- Many pairs of proteins have thus to be tested: 771,518,121 for *A. thaliana*

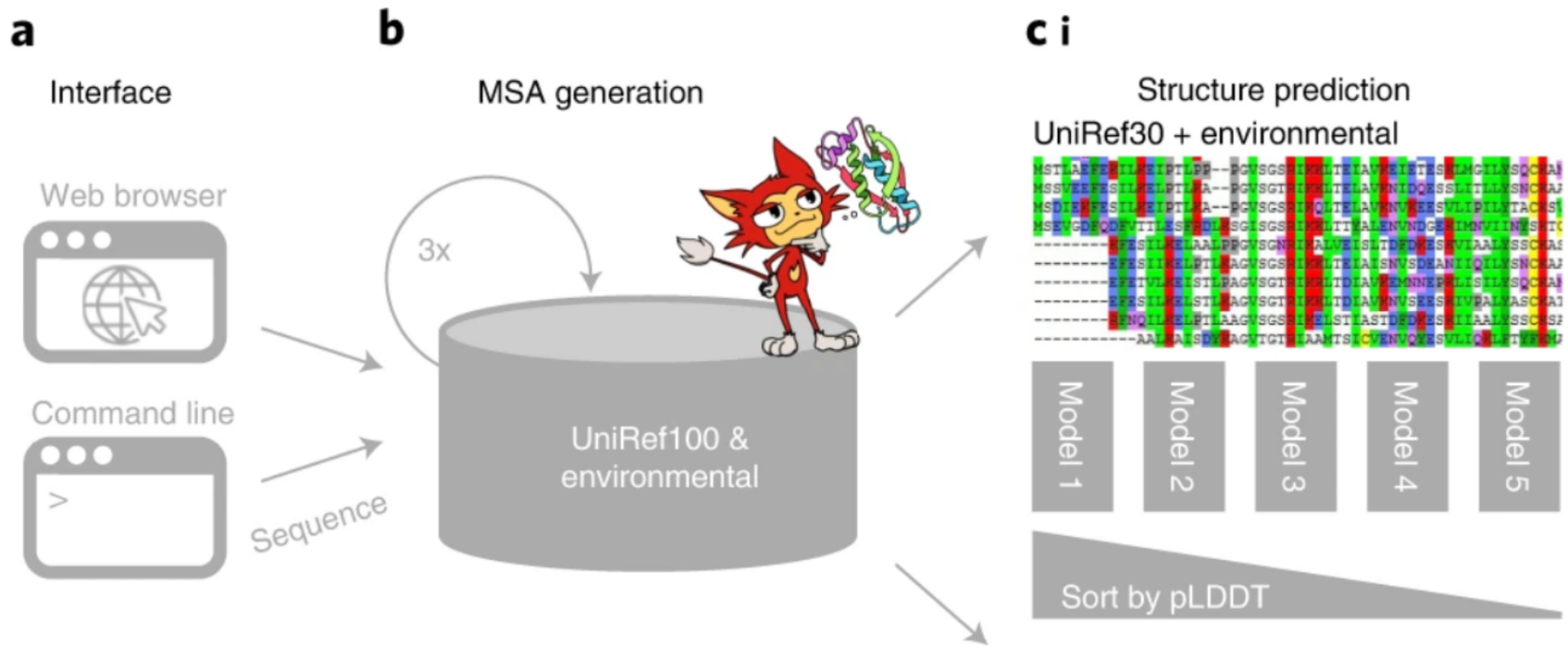
- **AlphaFold2-multimer** is **high computing time consuming** :

  impossible to compute on an interactome

  → We had to **find solution to reduce the computing time**

Reduce the computation time for interactome exploration by using ColabFold

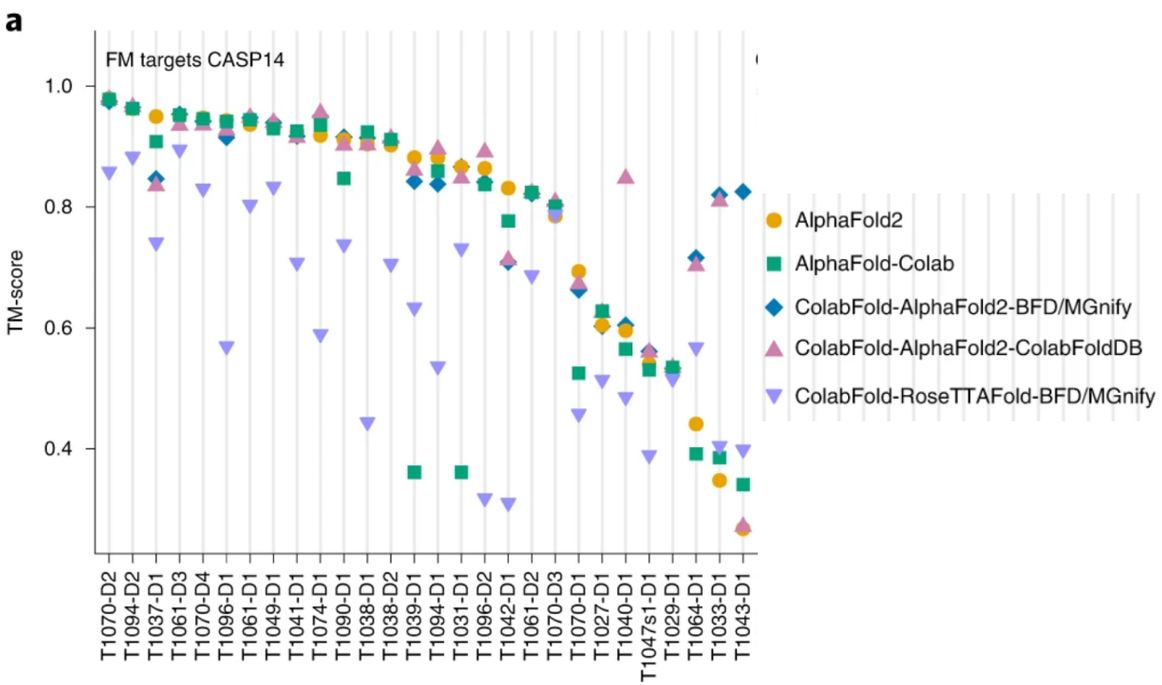
- **Solution 1** : **Use ColabFold** , an accelerated Alphafold2 : accelerated MSA generation using the MMseqs2 algorithm on databases where redundancy has been reduced to a minimum

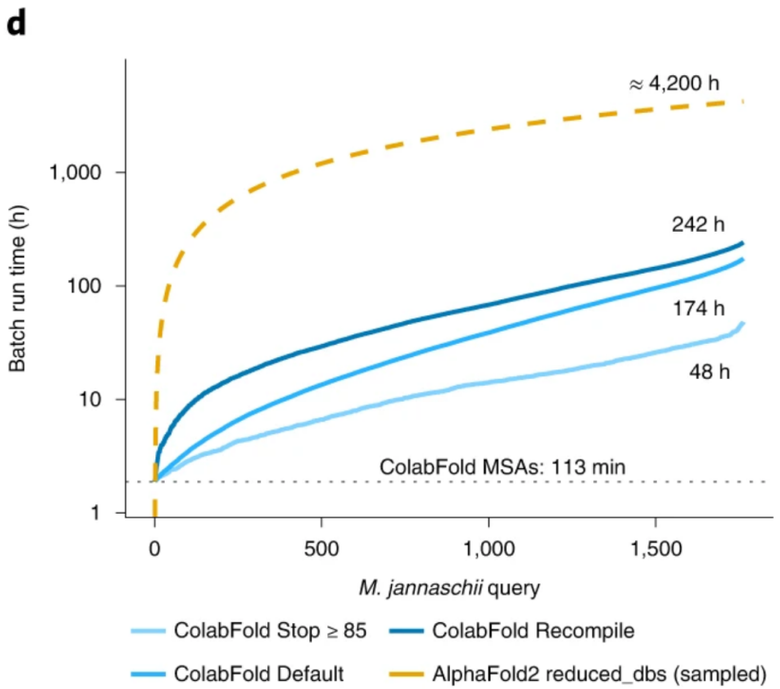


Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679-682. doi: 10.1038/s41592-022-01488-1

Reduce the computation time for interactome exploration by using ColabFold

Pipeline 40 to 60 times faster with very little loss of quality





Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679-682. doi: 10.1038/s41592-022-01488-1

Find the best AF parameters for a compromise between calculation time and prediction performance

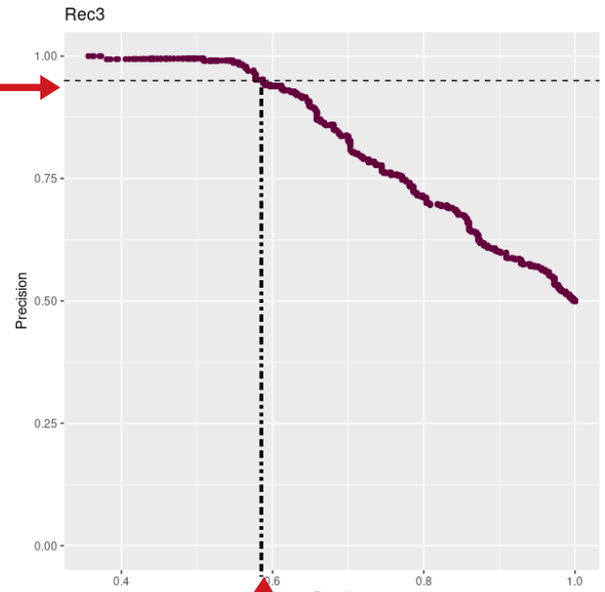
Dataset : 548 IPPs and 1612 proteins random pairs of ***S. cerevisiae***

3 recycles seems to be a good compromise between calculation time and prediction performance

3 recycles

False Positive Rate = 0.05

$$Precision = \frac{TP}{TP + FP}$$



$$Recall = \frac{TP}{TP + FN}$$

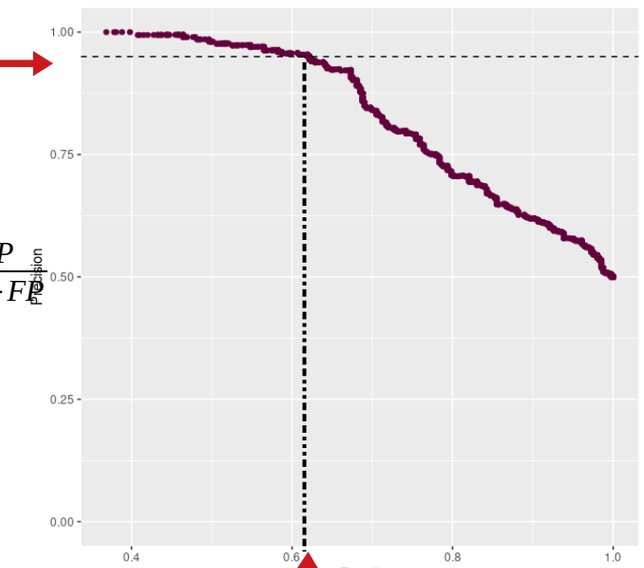
59 % IPPs predicted
14 min

versus
versus

20 recycles

False Positive Rate = 0.05

$$Precision = \frac{TP}{TP + FP}$$



$$Recall = \frac{TP}{TP + FN}$$

62 % IPPs predicted
1h18

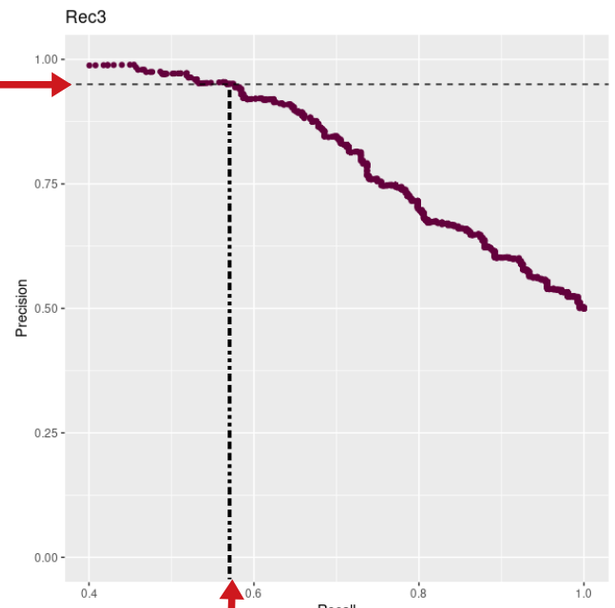
Effect of the number of models on prediction performances

With 3 recycles

One model=the best model

False Positive Rate = 0.05

$$Precision = \frac{TP}{TP+FP}$$



$$Recall = \frac{TP}{TP+FN}$$

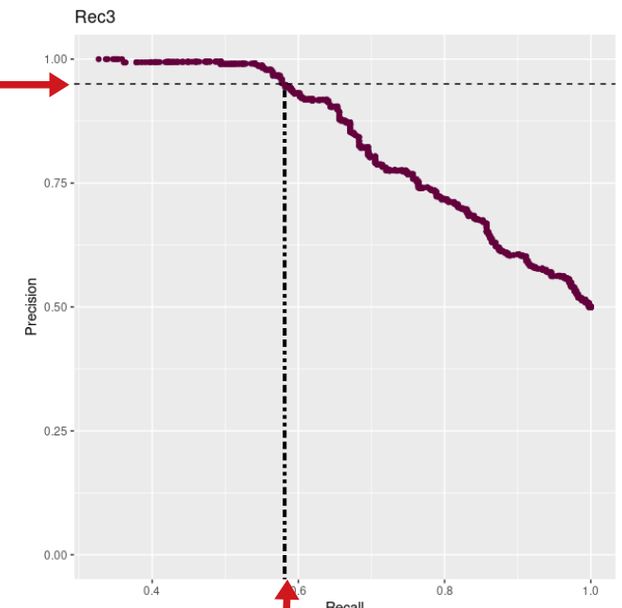
Slight drop in performance : 59 % predicted IPP

versus

Five models

False Positive Rate = 0.05

$$Precision = \frac{TP}{TP+FP}$$



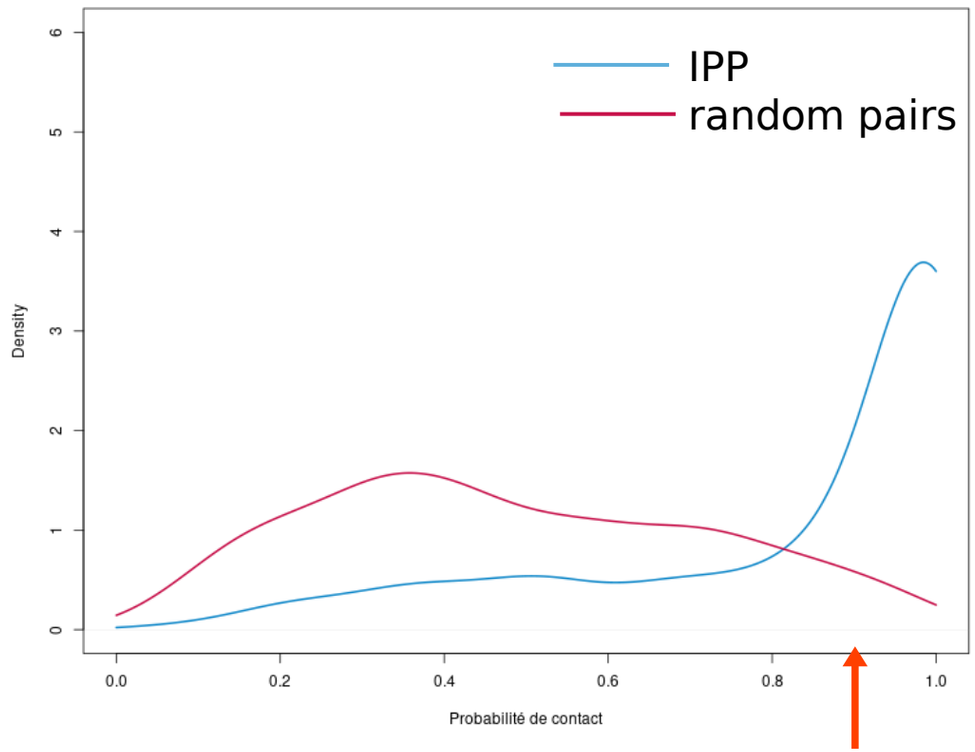
$$Recall = \frac{TP}{TP+FN}$$

62 % predicted IPP

Why PPI prediction is better with 5 models ?

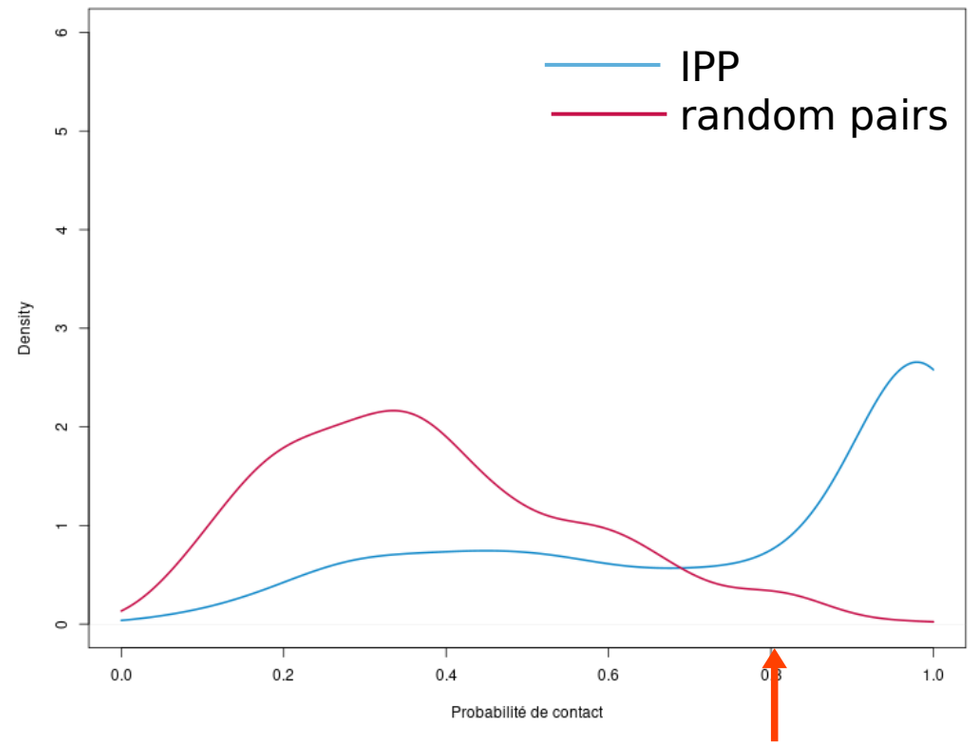
Distribution of contact probabilities obtained with 3 recycles

One model=the best model



Probability threshold=0.951
for a False Positive Rate=0.05

Five models

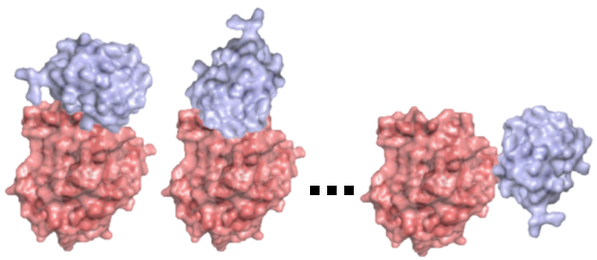


Probability threshold=0.806
for a False Positive Rate=0.05

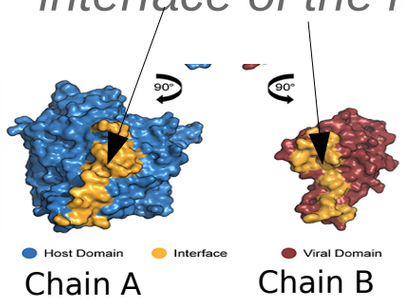
Better separation of contact probabilities distributions with 5 models

Why PPI prediction is better with 5 models ?

Boxplot of the percentages of amino acids in common between the interfaces of the models for IPP and random pairs

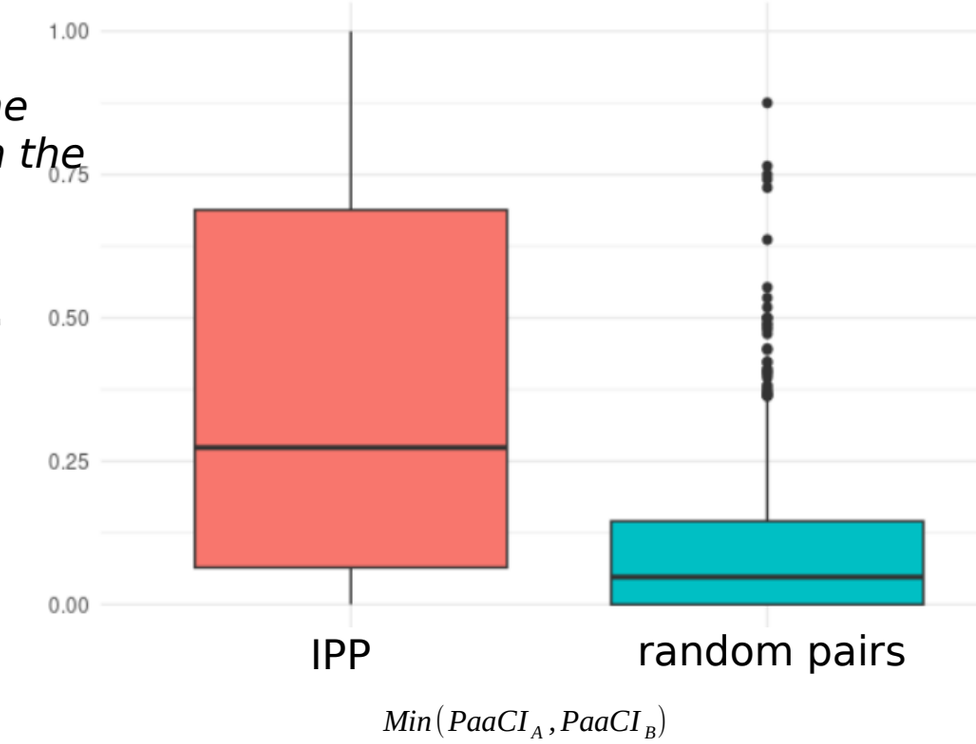


IM_i = *amino acids at the interface of the model i*



$PaaCI_j$ =*Percentages of amino acids in Common between the Interfaces of the 5 models on the cain j*

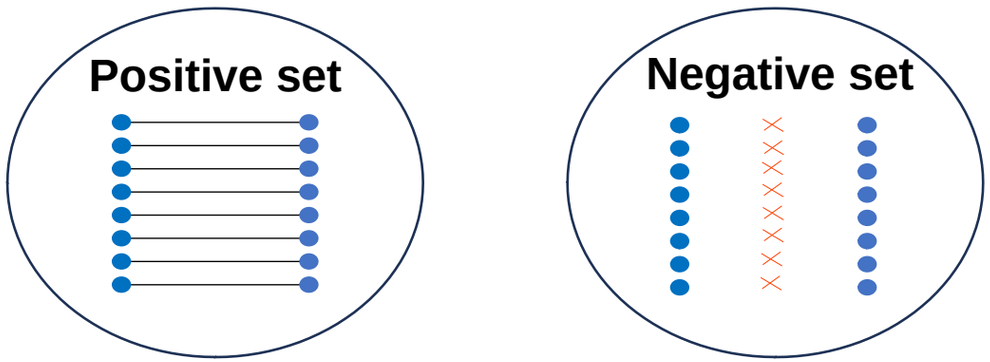
$$\frac{|IM1 \cap IM2 \cap IM3 \cap IM4 \cap IM5|}{|IM1 \cup IM2 \cup IM3 \cup IM4 \cup IM5|}$$



The 5 models have much different interfaces for random pairs than for IPPs

Prediction of *A. thaliana* Interacting Protein Pairs

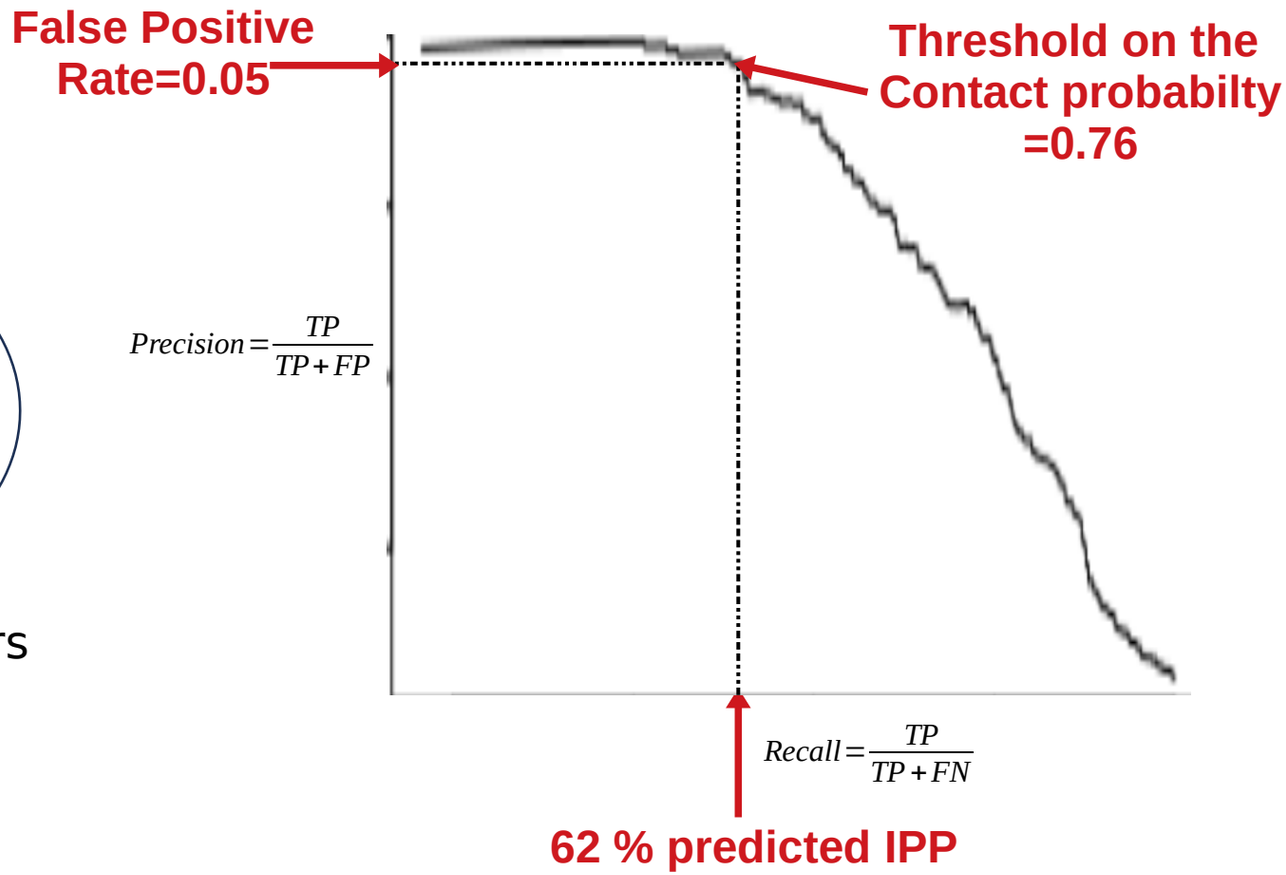
A. thaliana datasets



398 IPP (114 + 286)

1204 random pairs

Prediction results with 3 recycles and 5 models



To increase the percentage of predicted IPP

Use a **random forest** to **compute an interaction probability** for each protein pair **from the different quality scores**, probability of contacts, percentage of common amino acids at the interface of the 5 models **and the status of the protein pair** (IPP = 1 or random pair = 0).

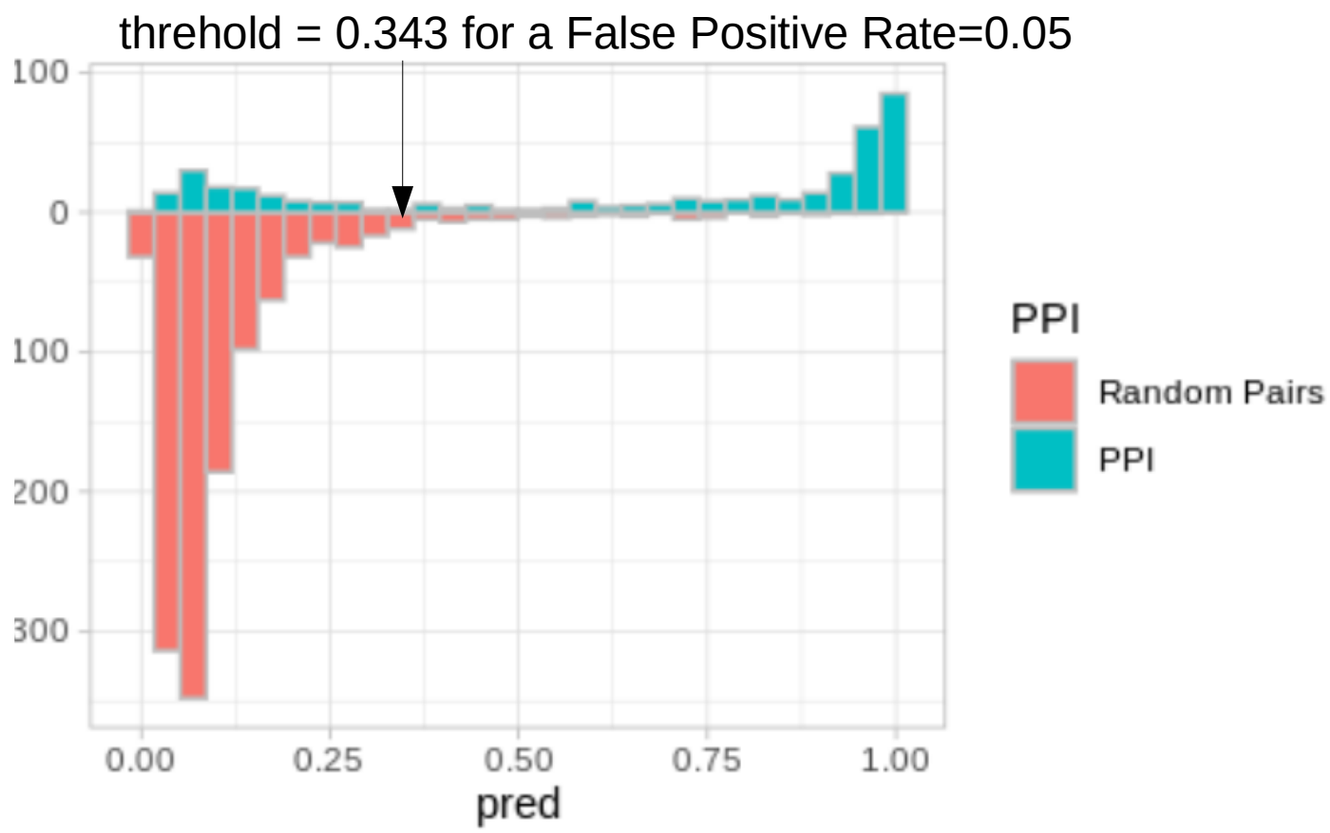
max_iptm	avg_iptm	max_contact_prob	avg_contact_prob	intersect_rat	PPI.bin
0.40	0.376	1.0010	0.9820	0.189	0
0.20	0.172	0.4924	0.4383	0.000	0
0.20	0.162	0.7030	0.4836	0.000	0
0.37	0.314	0.7617	0.6744	0.000	0
0.41	0.404	0.9990	0.9952	0.252	0
0.60	0.564	0.9946	0.9610	0.015	0
0.19	0.134	0.7592053	0.4442	0.0000	1
0.23	0.198	0.4434502	0.3287	0.1081	1
0.11	0.096	0.3482123	0.2093	0.0000	1
0.25	0.236	0.7603451	0.6604	0.1371	1
0.15	0.140	0.4650292	0.2925	0.0000	1
0.31	0.238	0.8095289	0.7064	0.3864	1

Random Forest

pred
0.019908694
0.081602907
0.033929093
0.327494963
0.022264888
0.248700949
0.88418454
0.83666050
0.62356131
0.97516634
0.81933785
0.99171244

Results: a higher percentage of predicted IPP

Distribution of the interaction probabilities



70 % IPP predicted

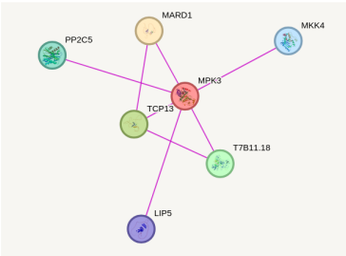
Limits

- The number of percentage of predicted IPP remains low.
- Poor partner prediction for non-globular proteins, particularly proteins with large, unstructured loops.
- Impossible to make predictions on the whole of the *A. thaliana* interactome because there are too many pairs.

Perspectives

Applying our IPP prediction method on a reduced scale.

1. Prediction of the *A. thaliana* MKP3 interactors (collab with J. Bigeard, IPS2) :



More than 100 additional partner proteins (targets or regulators of MPK3) predicted
→ Biological validations in progress

2. Prediction of the interactome of *A. thaliana* chlorosplast in progress (collab with E. Delanoy & D. Monacello,IPS2) :

- Prediction of the 5625 PPIs between the 75 proteins encoded by the chloroplast genome
- Prediction of the partners of 6 chloroplast proteins of interest with the 1500 proteins localised in the chloroplast

→ In progress

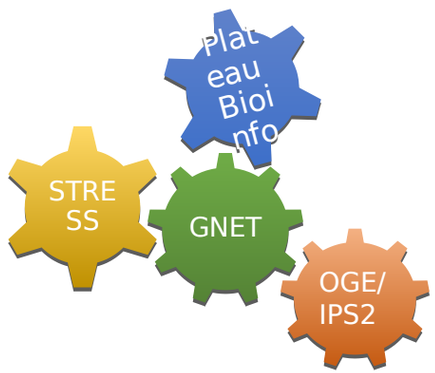
Thanks to

GNET team : Simon Gosset, Jean-Philippe Tamby

OGE team : Dario Monacello, Etienne Delanoy

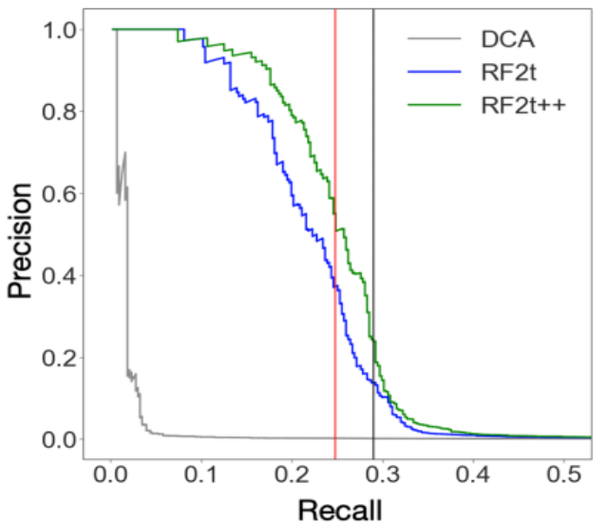
STRESS team : Jean Bigeard

Plateau Bioinfo : Frederic Desprez

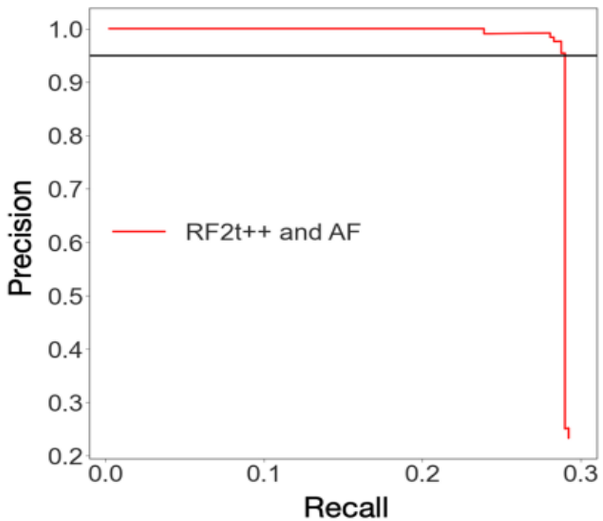


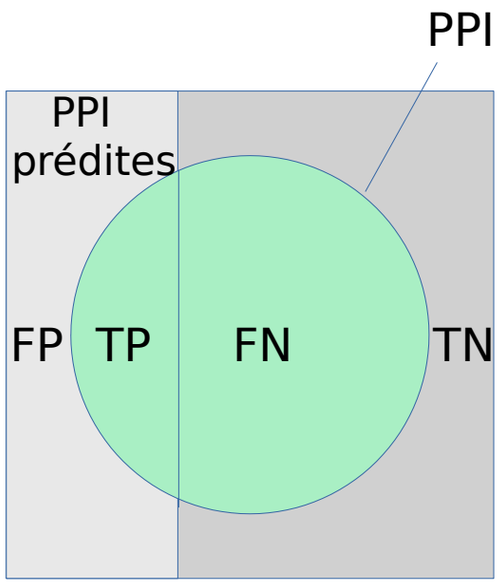
RoseTTAFold + AlphaFold2: an excellent tool for predicting protein complex structure but a poor tool for predicting PPIs

RoseTTA results on a set of 768 PPI +768000 non PPI of *S. cerevisiae*



AlphaFold2 results on PPIs predicted by by Rosetta Fold



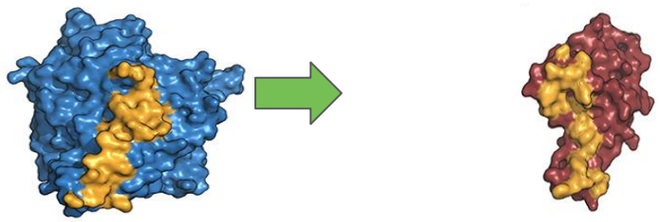


Precision=TP / (TP+FP)=0.95

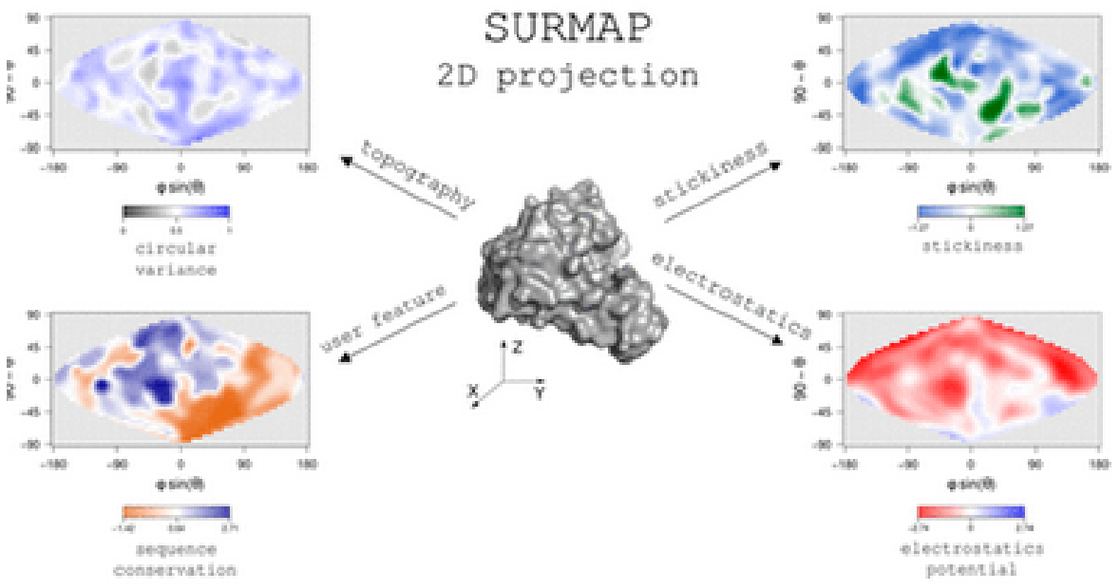
Recall=TP / (TP+FN)=0.29

Predicted PPIs are very safe but only 29% of PPIs are predicted as such

Our project: Understanding what distinguishes protein partners correctly predicted from those missed by Rosetta+AlphaFoldMultimer

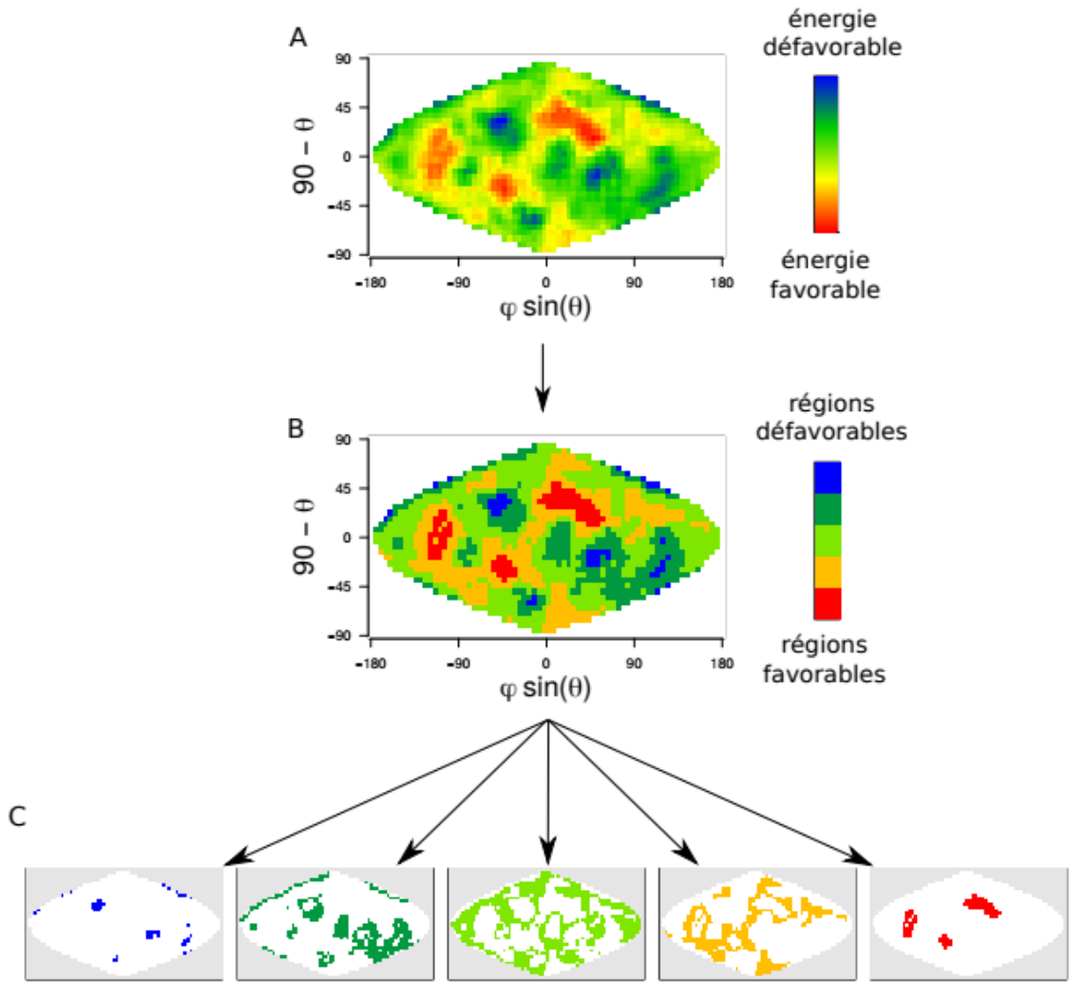


- To characterize surface properties of two proteins in interaction
- We developped A Software for Mapping in Two Dimensions Protein Surface Features (SURFMAP)

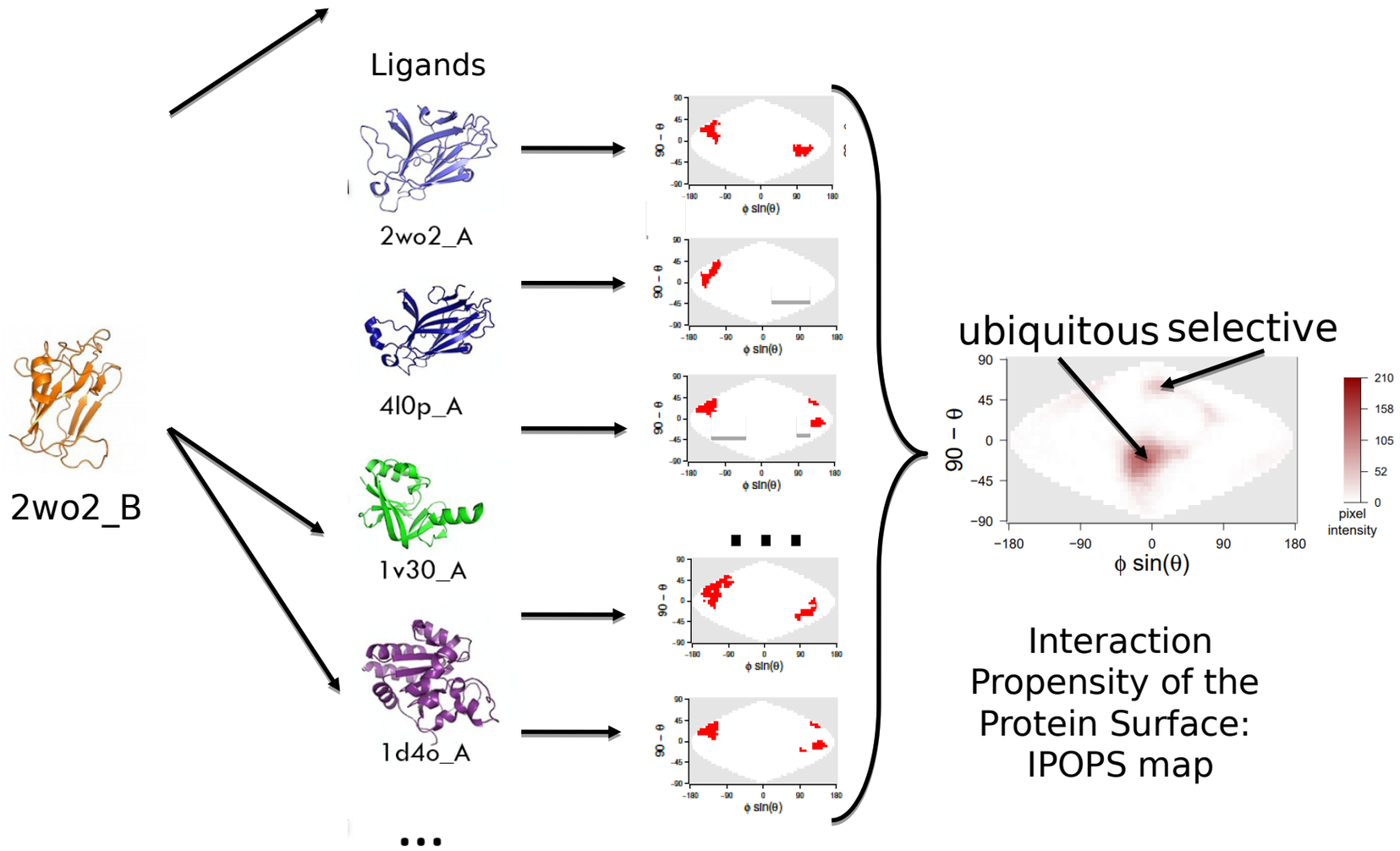


Schweke H, Mucchielli MH, Chevrollier N, Gosset S, Lopes A. SURFMAP: A Software for Mapping in Two Dimensions Protein Surface Features. J Chem Inf Model. 2022 Apr 11;62(7):1595-1601.

Discretization and class separation

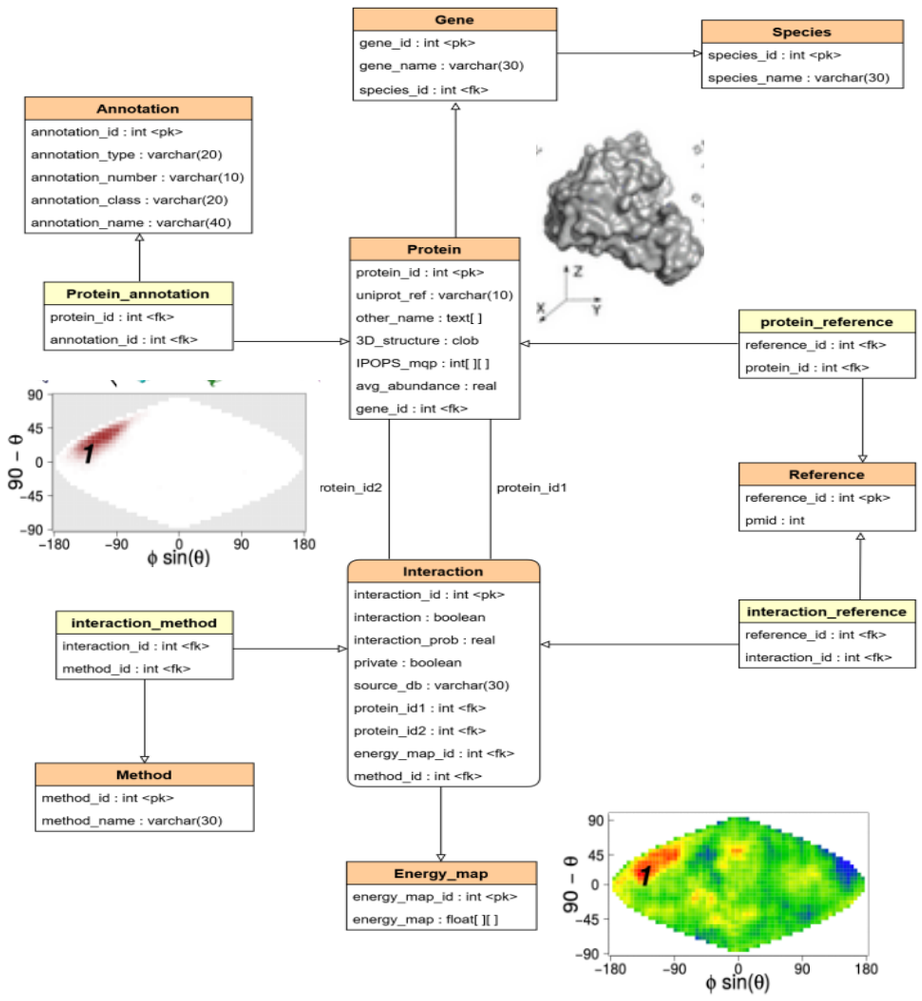


Interaction propensity of the protein surface map



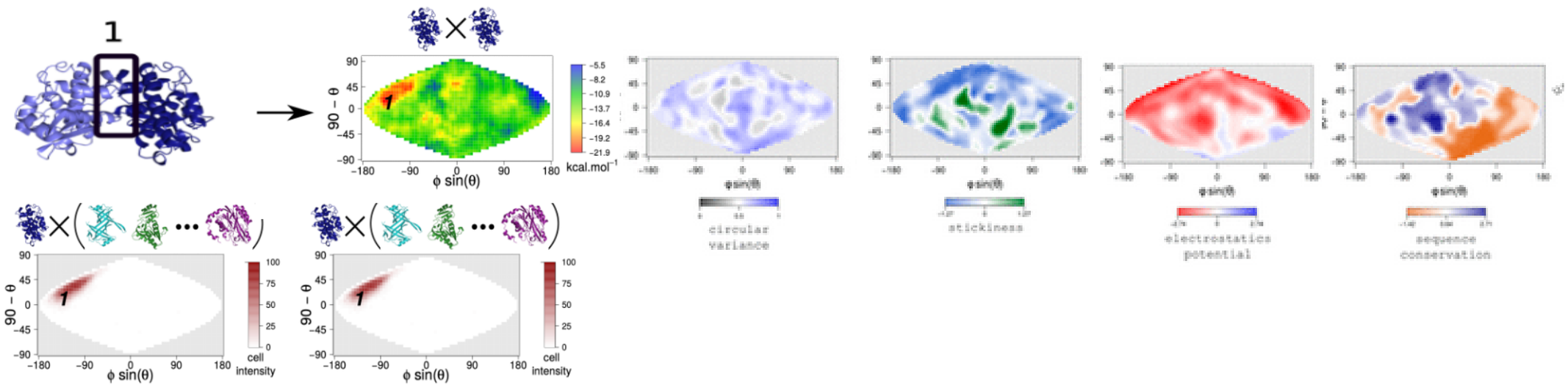
Schweke, Mucchielli, Sacquin-Mora, Bei & Lopes, JMB, 2020

PPIDB: A Protein-Protein Interaction Database



Dock&Co4PPIP: creating the first predicted interactome of chloroplastic proteins of *A. thaliana*

- Interactions between 1519 proteins located in the chloroplast are being identified by double hybride
- On the experimentally identified PPIs, RosettaFold +AlphaFoldMultimer will be applied
- Statistical analysis of the surface properties of the two proteins in interaction and of the properties of the interface in order to discriminate the set of PPIs predicted as PPI of the set of PPIs predicted as non PPI



Objective: refine the RosettaFold+AlphaFoldMultimer PPI prediction

Thanks to

GNET team : Simon Gosset, Cécile Guichard, Marie-Laure Martin, Jean-Philippe Tamby

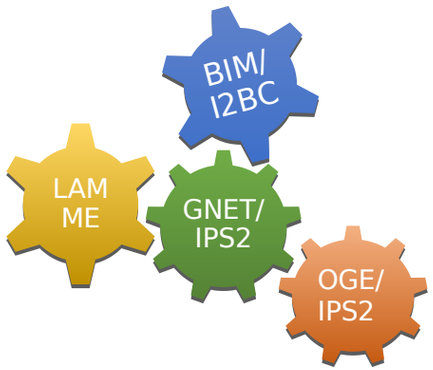
OGE team : Dario Monacello

Plateau Bioinfo : Frederic Desprez

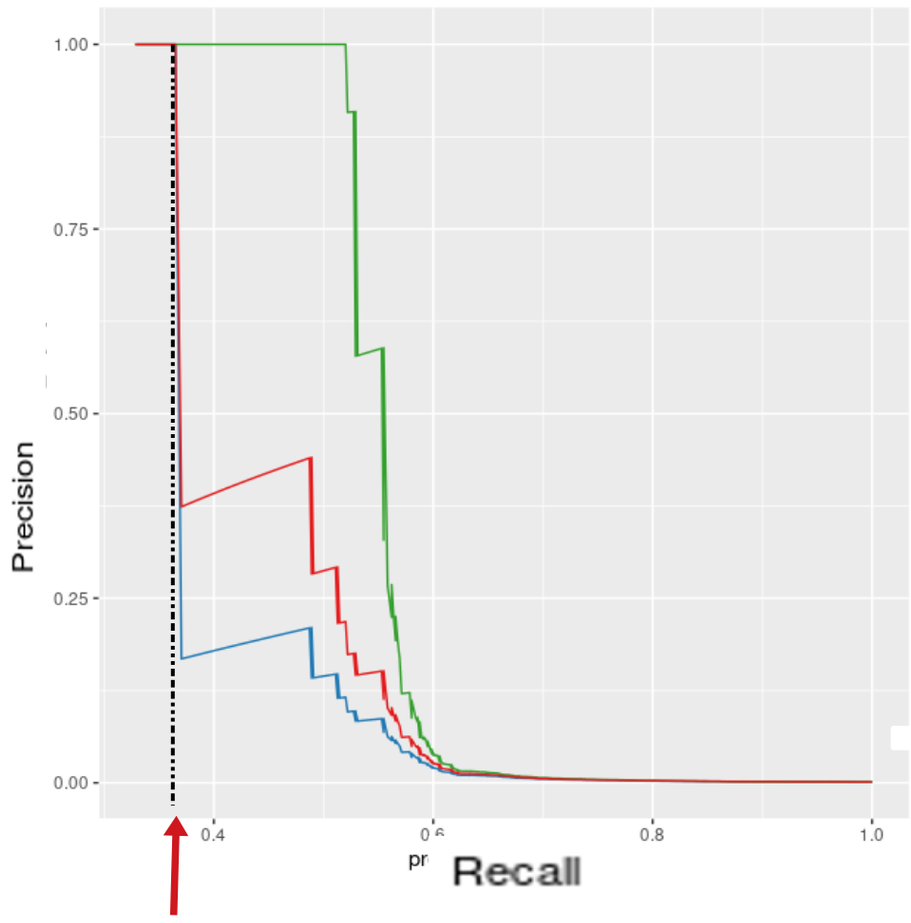
Franck Samson

Anne Lopes
Hugo Schweke

Sjoerd de Vries



Estimated prediction performance when the imbalance between IPPs and radom set is 1 per 1000



In the worst case 37 % IPP predicted