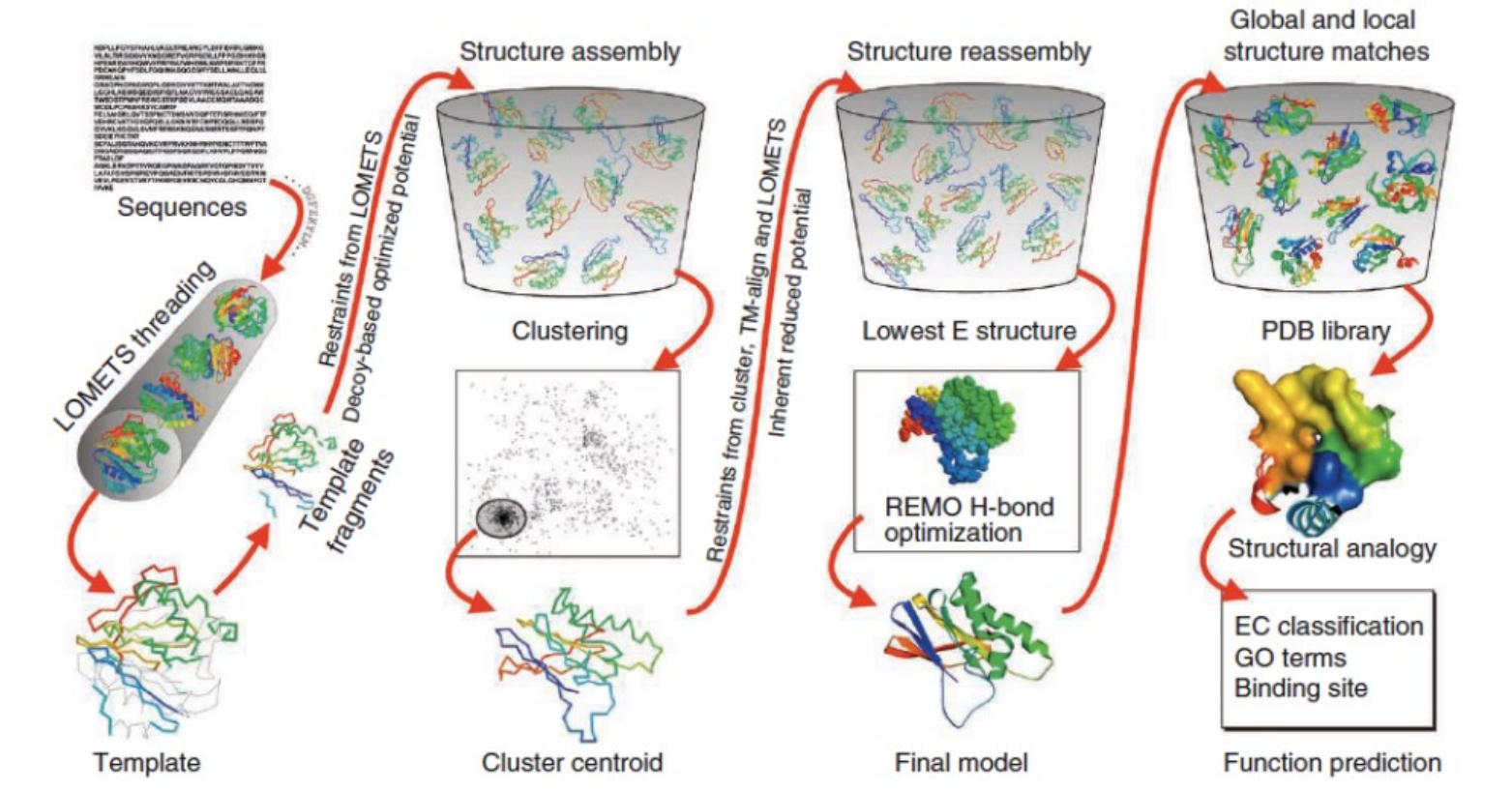
Fold prediction and statistical potentials



Methods for predicting the fold of a protein

Besides homology modeling, we have other methods to predict the structure of a protein

Threading

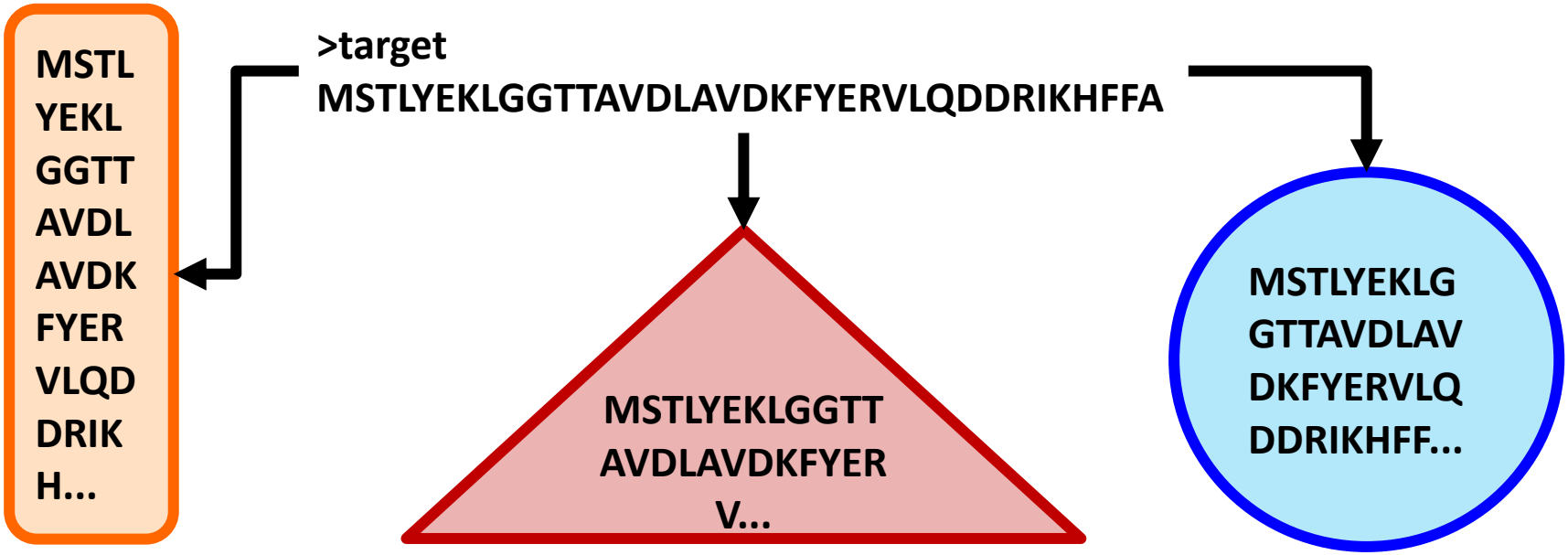
Ab initio

Molecular dynamics

Threading

Threading consists on forcing the fit of an amino acid sequence into a protein structure

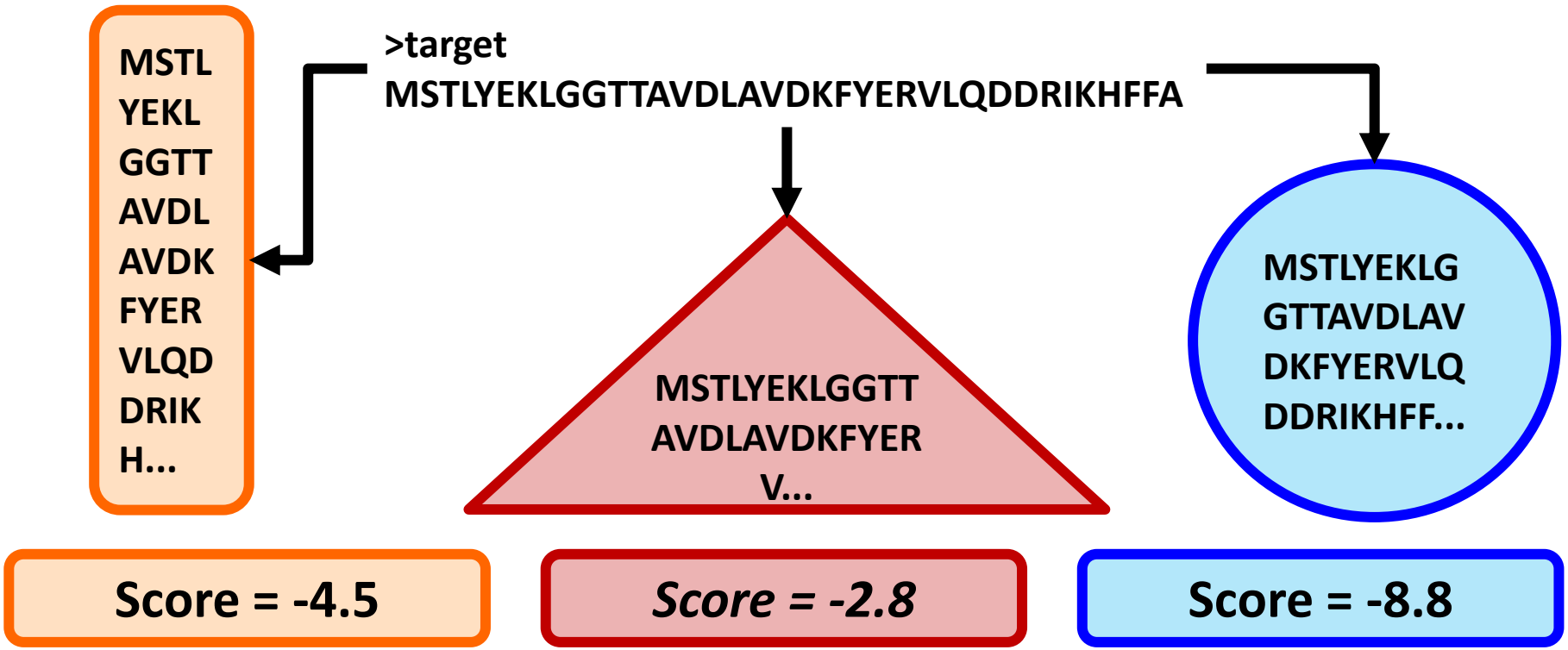
Threading programs usually have a database of protein structures, they fit the sequences in all of them and choose the best fit using scoring functions



Threading

Threading consists on forcing the fit of an amino acid sequence into a protein structure

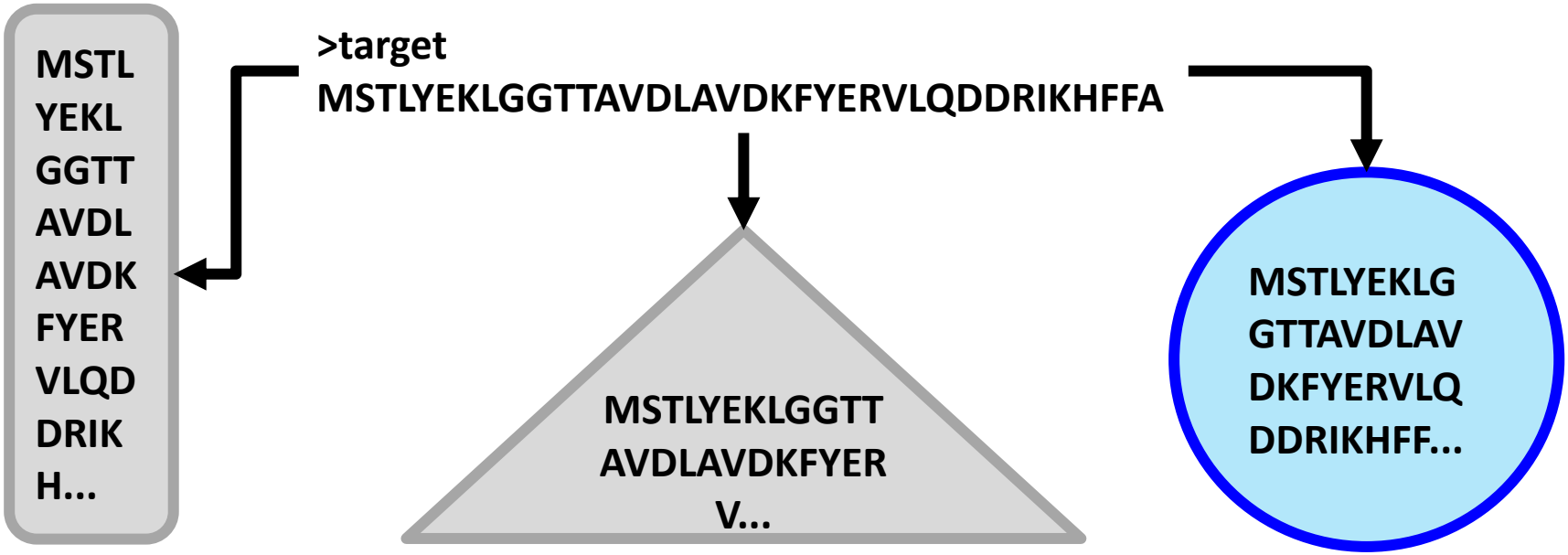
Threading programs usually have a database of protein structures, they fit the sequences in all of them and choose the best fit using scoring functions



Threading

Threading consists on forcing the fit of an amino acid sequence into a protein structure

Threading programs usually have a database of protein structures, they fit the sequences in all of them and choose the best fit using scoring functions



Score = -4.5

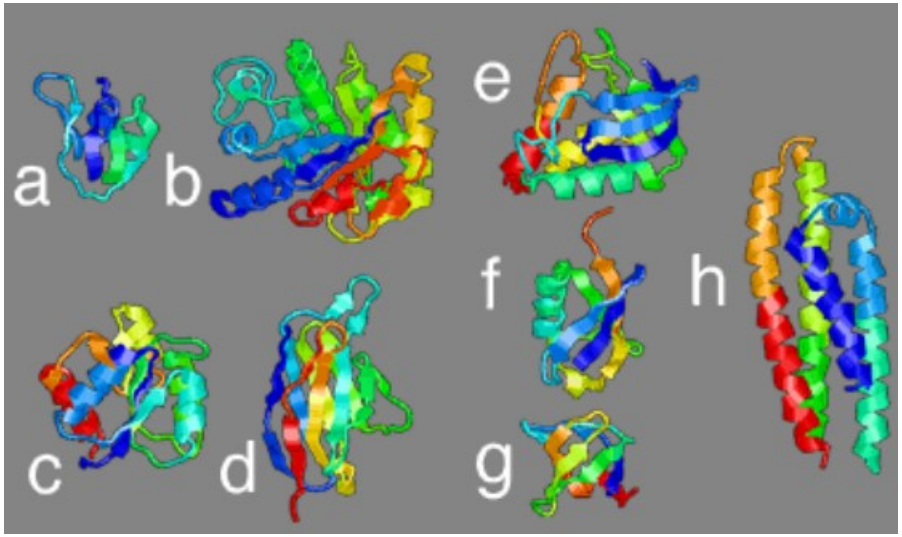
Score = -2.8

Score = -8.8

Threading

The reasoning behind threading is that as new protein structures are solved, barelly new folds are discovered

This raises the idea that the number of folds in nature is limited, and therefore any target protein should have the fold of an already known structure



Examples of threading programs are **threader**, **RAPTOR** or **phyre**

Ab initio

***Ab initio* methods make a prediction of protein structure without using homologs or any other information about protein structures**

The ab initio term is more defining of the input information used by the program than the algorythm itself. That is why ab initio programs have many diverse algorithms:

Neural networks
Alphafold

Threading of protein framents
Rosetta, I-Tasser

Mutual information

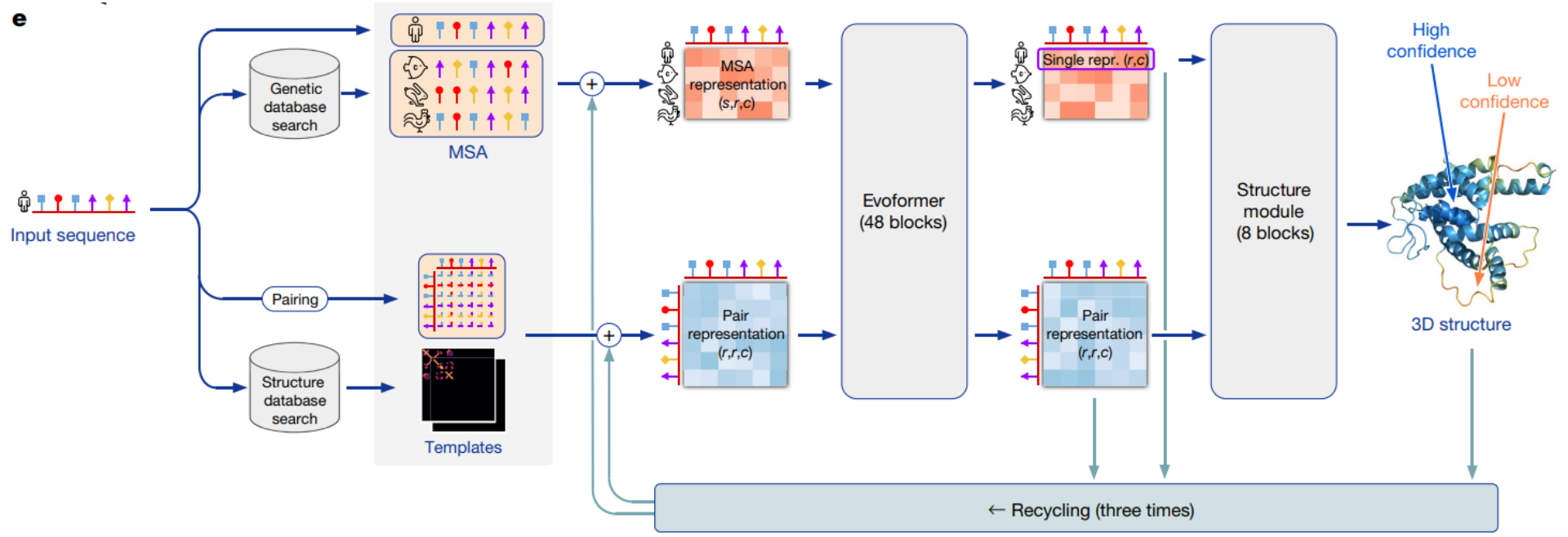
Ab initio

***Ab initio* methods make a prediction of protein structure without using homologs or any other information about protein structures**

Neural networks
Alphafold

Threading of protein framents
Rosetta, I-Tasser

Mutual information



Ab initio

***Ab initio* methods make a prediction of protein structure without using homologs or any other information about protein structures**

Neural networks
Alphafold

Threading of protein framents
Rosetta, I-Tasser

Mutual information

Alphafold deserves its own class, we will cover that next week!

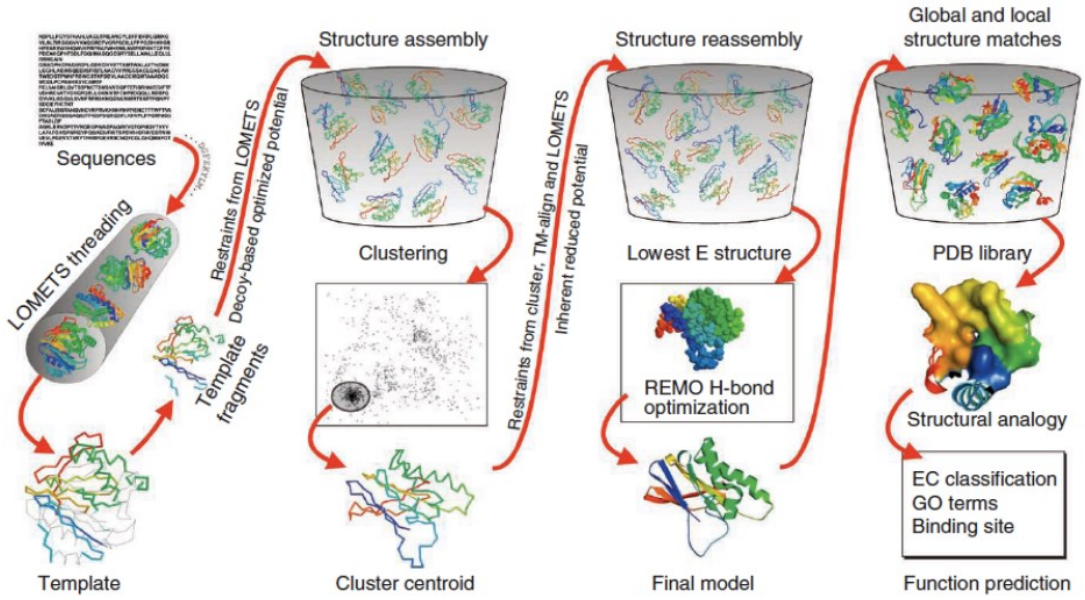
Ab initio

***Ab initio* methods make a prediction of protein structure without using homologs or any other information about protein structures**

Neural networks
Alphafold

Threading of protein framents
Rosetta, I-Tasser

Mutual information



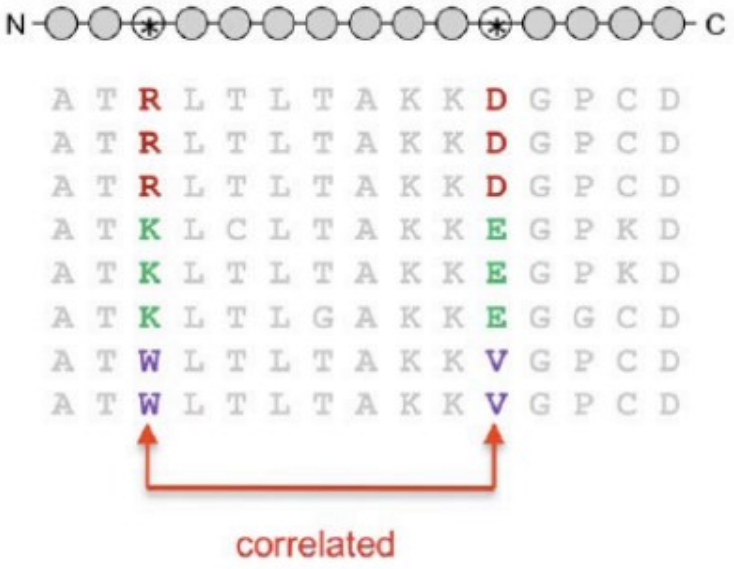
Ab initio

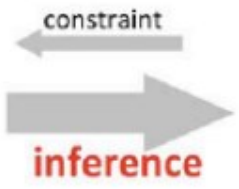
***Ab initio* methods make a prediction of protein structure without using homologs or any other information about protein structures**

Neural networks
Alphafold

Threading of protein framents
Rosetta, I-Tasser

Mutual information

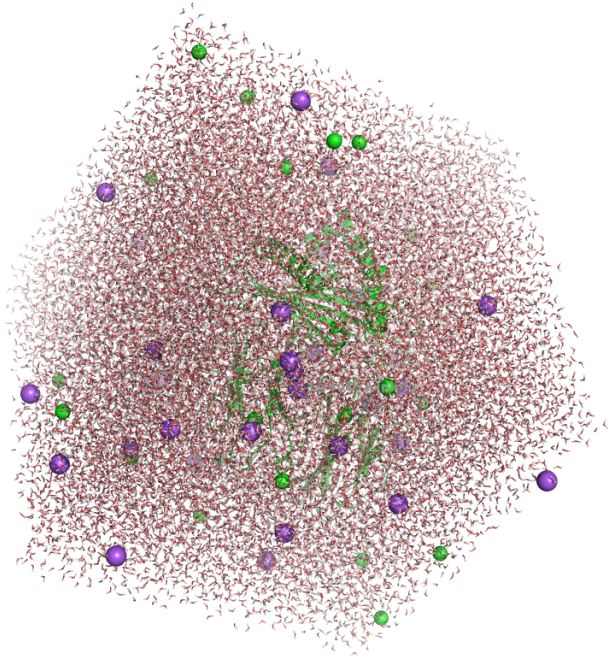




Molecular dynamics

Molecular dynamics methods simulate the protein during its folding process, as well as the solvent molecules arround the protein

This is the method you would use if you had infinite computational power and infinite time to make the computation. It's not just simulating the protein, it's simulating the whole folding process!



Methods for predicting the fold of a protein

Besides homology modeling, we have other methods to predict the structure of a protein

Threading

Ab initio

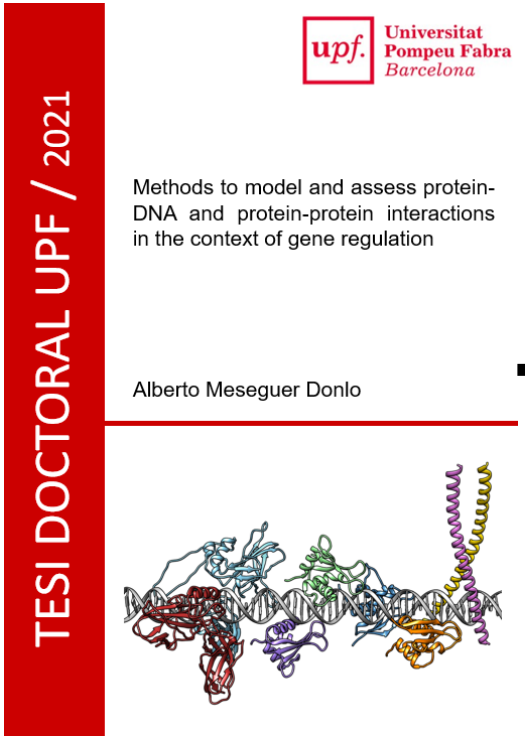
Molecular dynamics

All these methods generate a huge amount of models. How do we know what models are correct?

Statistical potentials

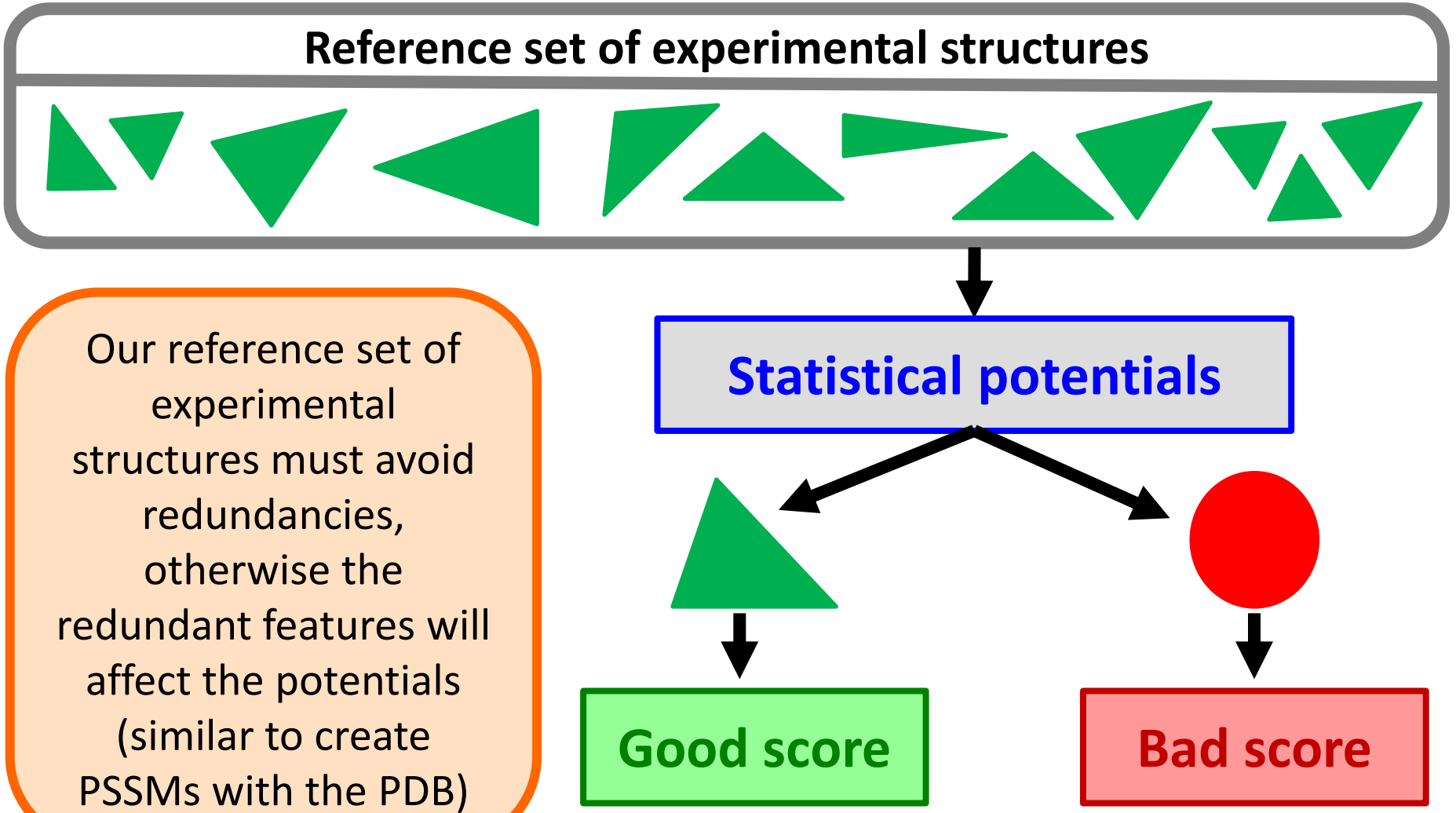
All the methods seen before (plus homology modeling) can generate lots of models (many of which can be wrong)

We can use statistical potentials to identify the good models and discard the wrong ones



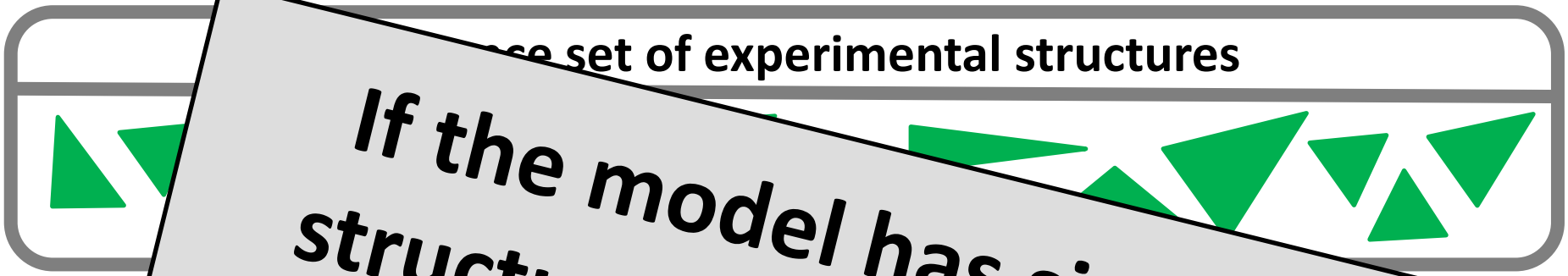
My thesis was about the development of statistical potentials to score protein-DNA interactions

Statistical potentials



Our reference set of experimental structures must avoid redundancies, otherwise the redundant features will affect the potentials (similar to create PSSMs with the PDB)

Statistical potentials



If the model has similar structural features to the proteins in the reference set, statistical potentials will provide good scores

Good score

Statistical potentials

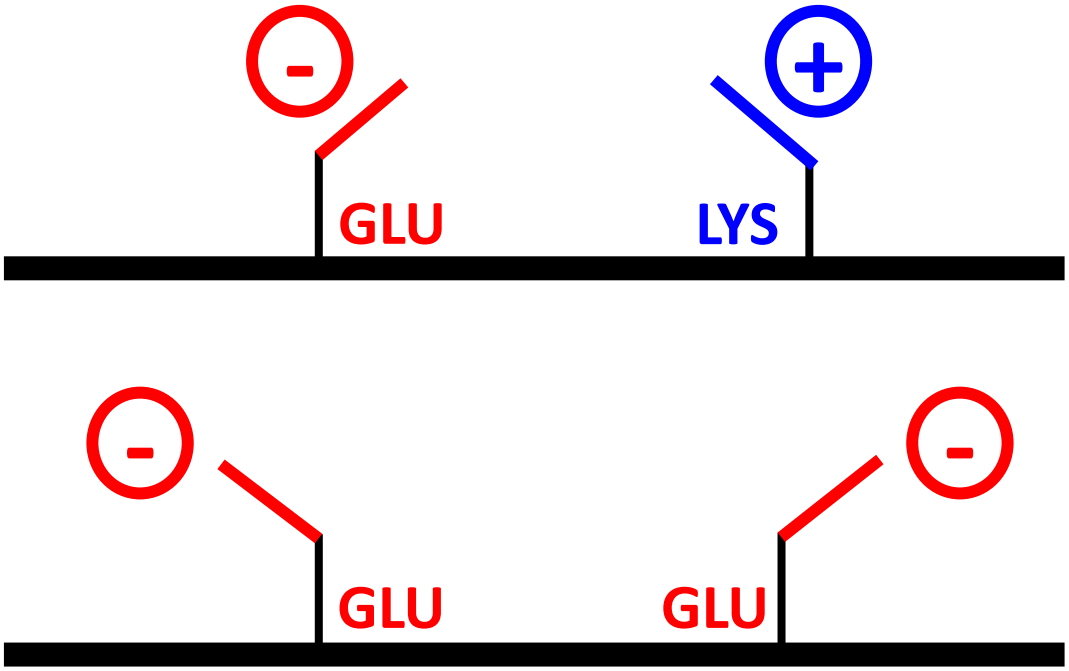
What are the structural features that statistical potentials use?

Amino acid contacts

Amino acid exposure

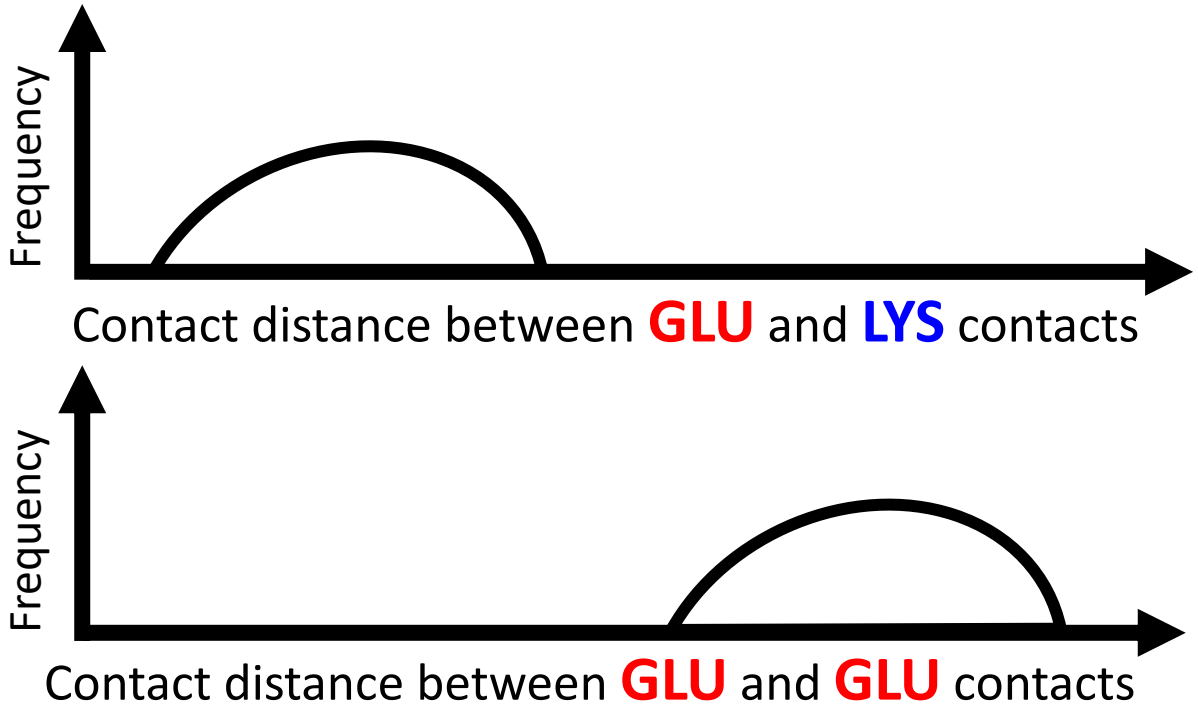
Statistical potentials: contacts between pairs of amino acids

Amino acid contacts



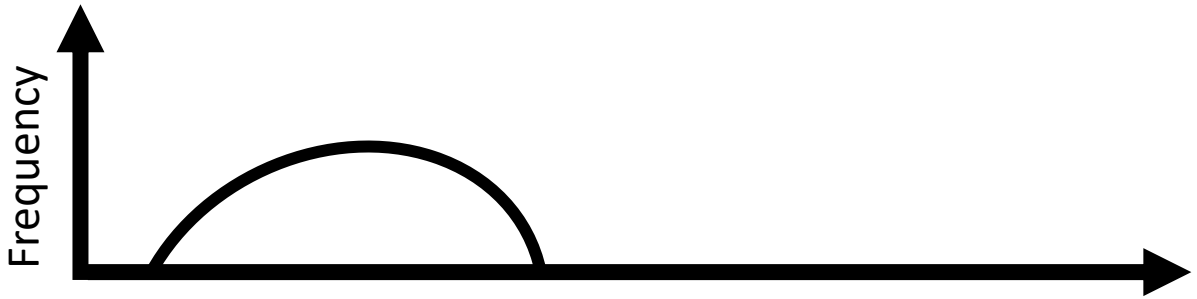
Statistical potentials: contacts between pairs of amino acids

Amino acid contacts

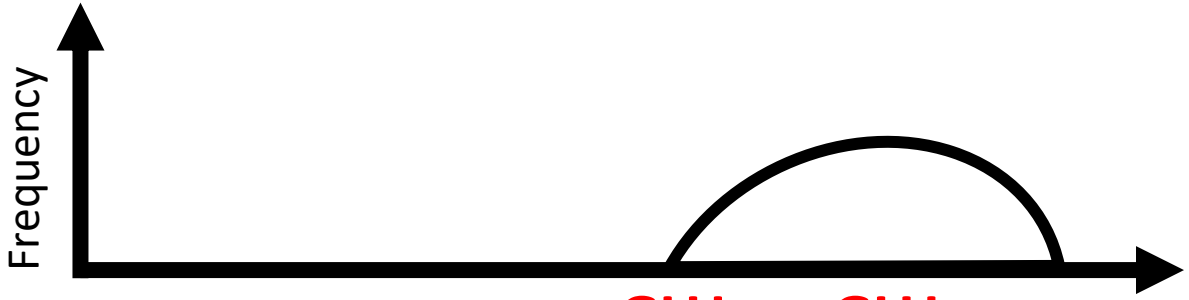


Statistical potentials: contacts between pairs of amino acids

The frequency of contacts can be affected by the fact that there is higher density of Aa in the core of the protein than in the surface



Contact distance between **GLU** and **GLU** contacts in the **core**



Contact distance between **GLU** and **GLU** contacts in the **surface**

Statistical potentials: contacts between pairs of amino acids

The frequency of contacts can be affected by the fact that there is higher density of Aa in the core of the protein than in the surface

This is the **reference state** problem, and different statistical potentials programs handle it in different ways. The two main ways are:

Assuming no difference between surface and core (1 reference state)

Pros: The data is not fragmented, so you have more data to calculate probabilities

Cons: You loose accuracy

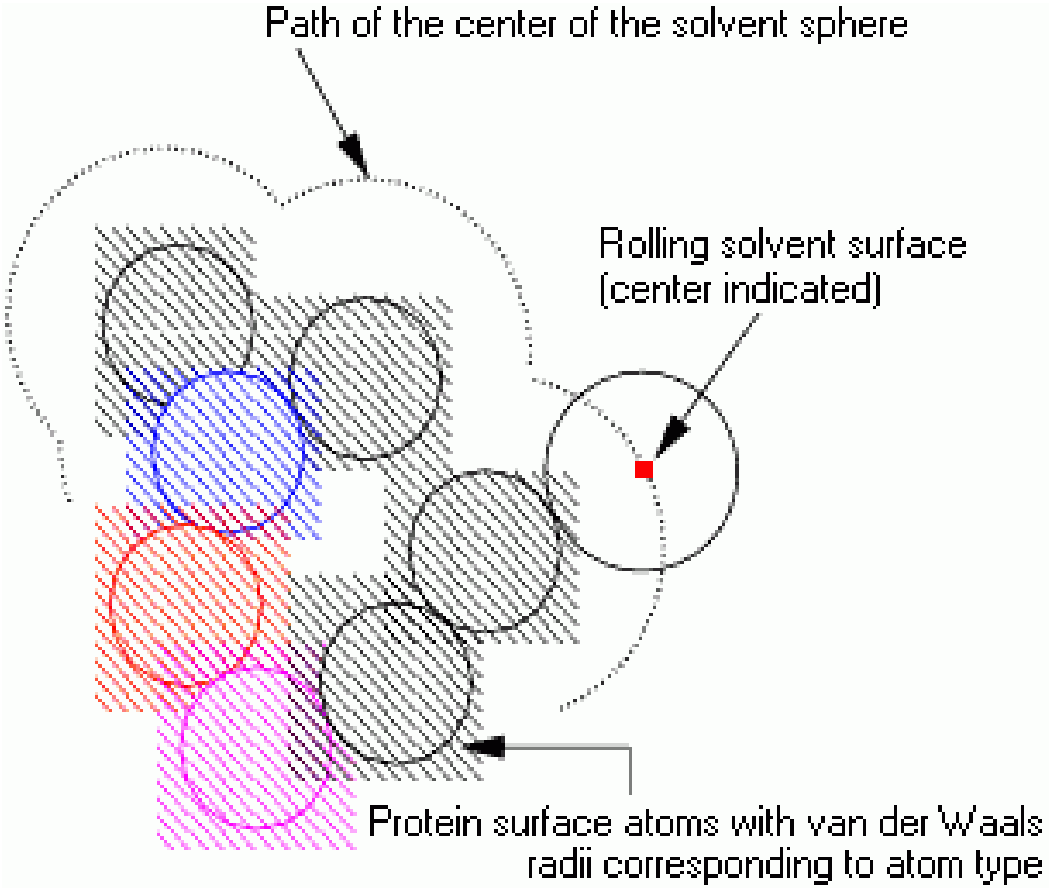
Splitting the data according to amino acid density (several reference states)

Pros: You win accuracy

Cons: The data is more fragmented, so you may miss data to calculate reliable probabilities

Statistical potentials: amino acid exposure

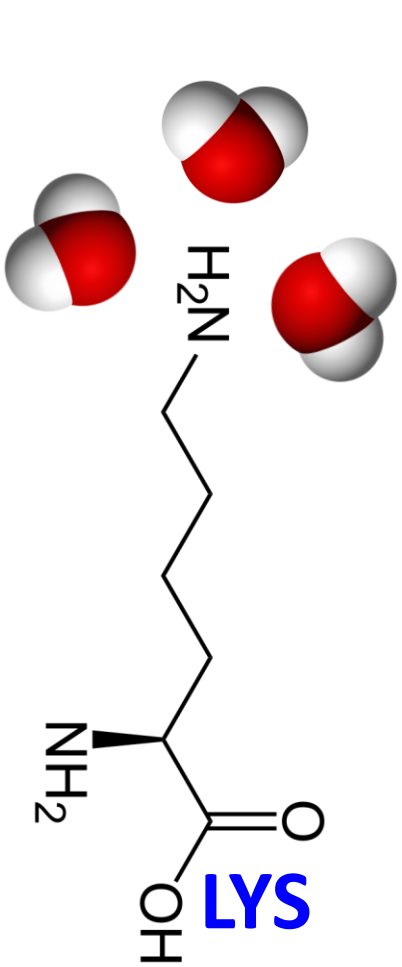
We determine the degree of exposure by measuring the accessible surface area (ASA) of each amino acid



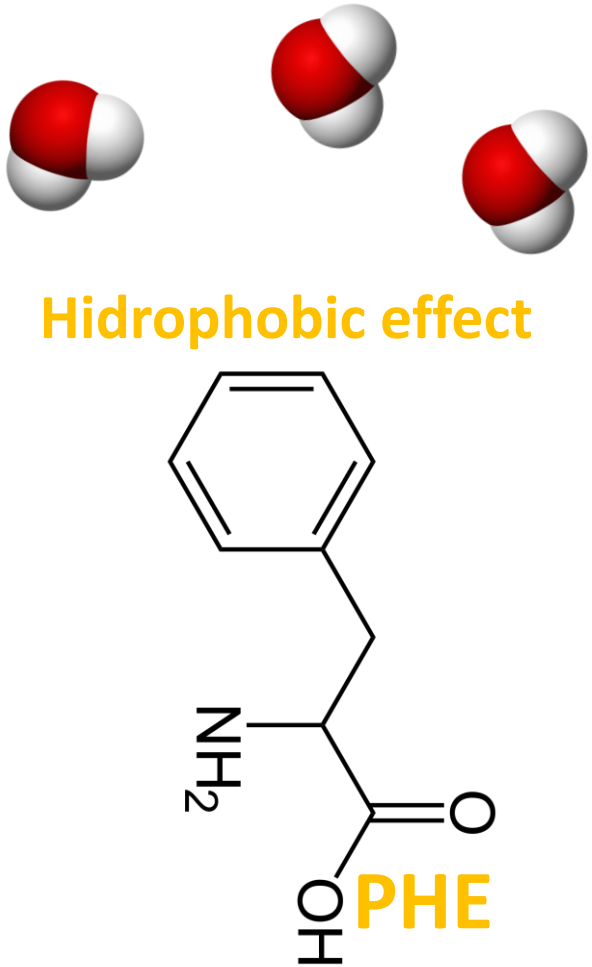
B. Lee and F. M. Richards (1971).

Statistical potentials: amino acid exposure

Polar and charged are more likely to be exposed (higher ASA) because of their tendency to interact with water molecules

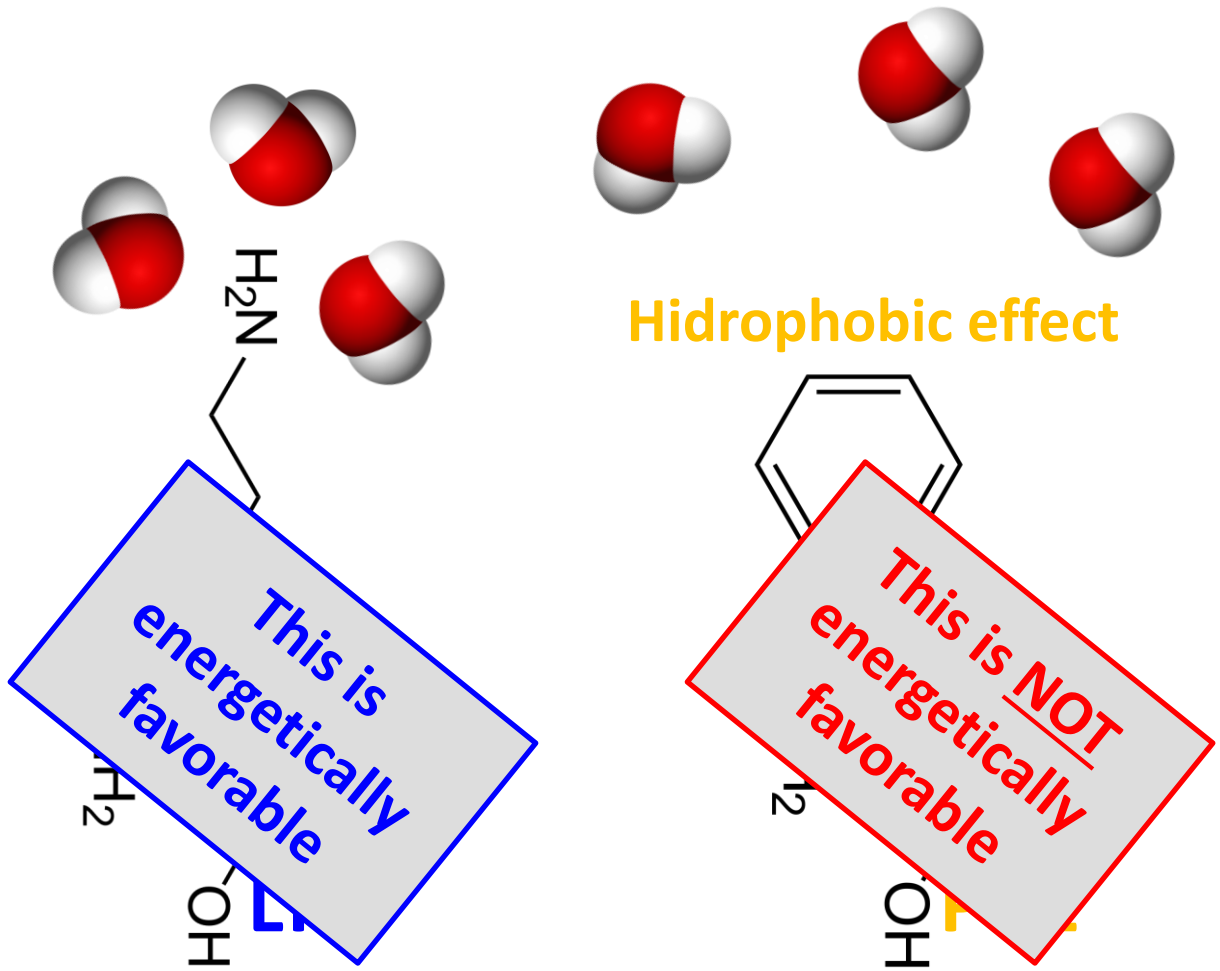


Hidrophobic effect



Statistical potentials: amino acid exposure

Polar and charged are more likely to be exposed (higher ASA) because of their tendency to interact with water molecules



Deriving statistical potentials from Boltzmann's Law

Statistical potentials are computed using Boltzmann Law

This is the formula I showed you in the seminars:

$$P=(1/z)(e^{(-E/kT)})$$

And this is the formula I showed you in biophysics:

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

Are they the same formula?

Deriving statistical potentials from Boltzmann's Law

Yes! Z equals the partition function
(physicists use q to refer to the partition function)

$$P=(1/z)(e^{(-E/kT)})$$

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

$$Z = \sum e^{-E_n/kT}$$

Can you imagine what is Z?

Deriving statistical potentials from Boltzmann's Law

Let's assume that we apply Boltzmann's Law to particles in different energy levels (as we did in biophysics)

Probability of having 1 particle in a specific energy level (e.g. 2 KJ/mol)

Probability of having 1 particle in all the energy levels

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

Deriving statistical potentials from Boltzmann's Law

Let's assume that we apply Boltzmann's Law to particles in different energy levels (as we did in biophysics)

Probability of having 1 particle in a specific energy level (e.g. 2 KJ/mol)

Probability of having 1 particle in all the energy levels

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

See that these calculations are only possible with discrete energy levels, which is an oversimplification of reality

Deriving statistical potentials from Boltzmann's Law

Let's assume that we apply Boltzmann's Law to contacts between pairs of amino acids (as statistical potentials programs do)

Probability of having a GLU and a GLY at a contact distance of 5Å

Probability of having all Aa making contacts with all Aa at all distances

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

Deriving statistical potentials from Boltzmann's Law

Let's assume that we apply Boltzmann's Law to contacts between pairs of amino acids (as statistical potentials programs do)

Probability of having a GLU and a GLY at a contact distance of 5Å

Probability of having all Aa making contacts with all Aa at all distances

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

This number is a constant and is impossible to calculate!!!

Deriving statistical potentials from Boltzmann's Law

Let's assume that we apply Boltzmann's Law to contacts between pairs of amino acids (as statistical potentials programs do)

Probability of having a GLU and a GLY at a contact distance of 5Å

Probability of having all Aa making contacts with all Aa at all distances

$$\frac{N_i}{N} = \frac{e^{-E_i/kT}}{\sum e^{-E_n/kT}}$$

→ **This number (Z) is a constant and is impossible to calculate!!!**

↓

If you cannot calculate Z, just find the way to avoid it!

Deriving statistical potentials from Boltzmann's Law

We will operate Boltzmann Law to avoid the calculation of Z

Variables

$$P=(1/z)(e^{(-E/kT)})$$

Constants

Deriving statistical potentials from Boltzmann's Law

We will operate Boltzmann Law to avoid the calculation of Z

Variables

$$P=(1/z)(e^{(-E/kT)})$$

Constants

We put logarythms in both sides and isolate the energy

$$E= -KTlnP + KTlnZ$$

Deriving statistical potentials from Boltzmann's Law

We will operate Boltzmann Law to avoid the calculation of Z

Variables

$$P=(1/z)(e^{(-E/kT)})$$

Constants

We put logarythms in both sides and isolate the energy

$$\Delta E = -KT\ln P + 1000$$

I can give a fix value to Z and instead of calculating absolute energies, calculating changes in energy. These changes in energy (ΔE) will be our statistical potentials scores.

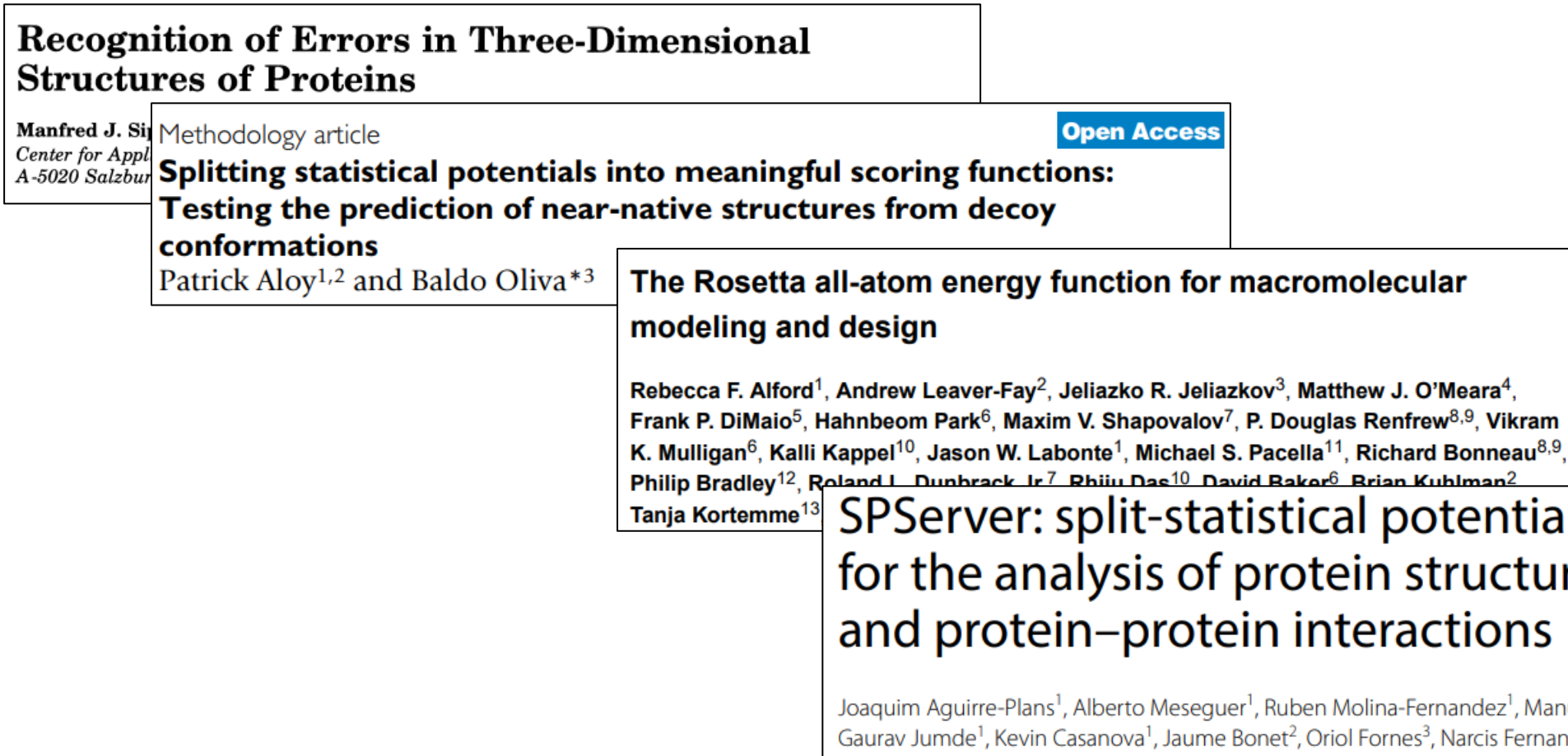
Statistical potentials are human made tools

Statistical potentials are human made tools

Statistical potentials are human made tools

Statistical potentials are human made tools

Many research groups have developed their own statistical potentials scoring functions



Statistical potentials are human made tools

Some statistical potentials scores also include scoring elements based on the laws of physics

One example of this is to use Coulomb's law to measure the electrostatic forces taking place between the amino acids of the protein

```
def elec_int(at1, at2, r):
    '''Electrostatic interaction energy between two atoms at r distance'''
    return 332.16 * at1.xtra['charge'] * at2.xtra['charge'] / MH_diel(r) / r
```

Pros

Including physics based scores can improve the performance of your potentials

Cons

Including physics based scores increases the computational cost and the execution times of your potentials

Statistical potentials are human made tools

You can obtain different statistical potentials depending on what terms you use and the relevance that you give to each term

This can be related with the scenario at which the potentials are supposed to work

When working with **transmembrane proteins** we don't have the membrane in the structure. Then, exposure analysis would result on hydrophobic residues being in the surface (which is not true).

Final score = Contacts · 0.3 + VanDerWaals · 0.4 + Electrostatics · 0.3

Statistical potentials are human made tools

You can obtain different statistical potentials depending on what terms you use and the relevance that you give to each term

This can be related with the scenario at which the potentials are supposed to work

When working with **transmembrane proteins** we don't have the membrane in the structure. Then, exposure analysis would result on hydrophobic residues being in the surface (which is not true).

Final score = Contacts · 0.3 + VanDerWaals · 0.4 + Electrostatics · 0.3

Electrostatics are very important in **protein-DNA interactions** because DNA is a negatively charged molecule.

Final score = Contacts · 0.3 + Exposure · 0.3 + Electrostatics · 0.4

Statistical potentials are human made tools

You can obtain different

How can you know what statistical potentials are the best???

...ntacts · 0.3 + Exposure · 0.3 + Electrostatics · 0.4

Competitions in the field of structural bioinformatics

Every few years there are competitions on the different predictive methods in the field os structural bioinformatics

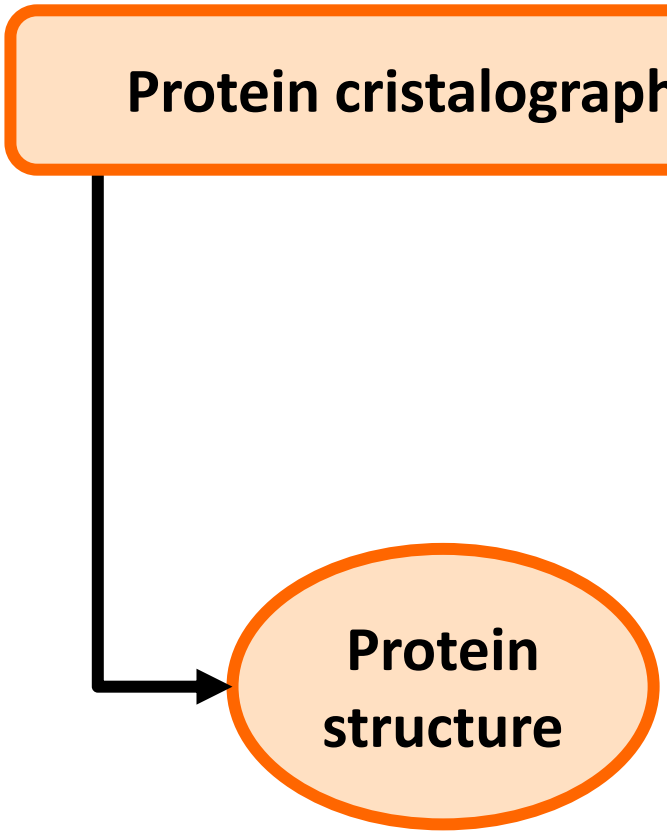
CASP
Predicting protein folds

CAPRI
Predicting protein-protein interactions

Scoring Functions
For both CASP and CAPRI

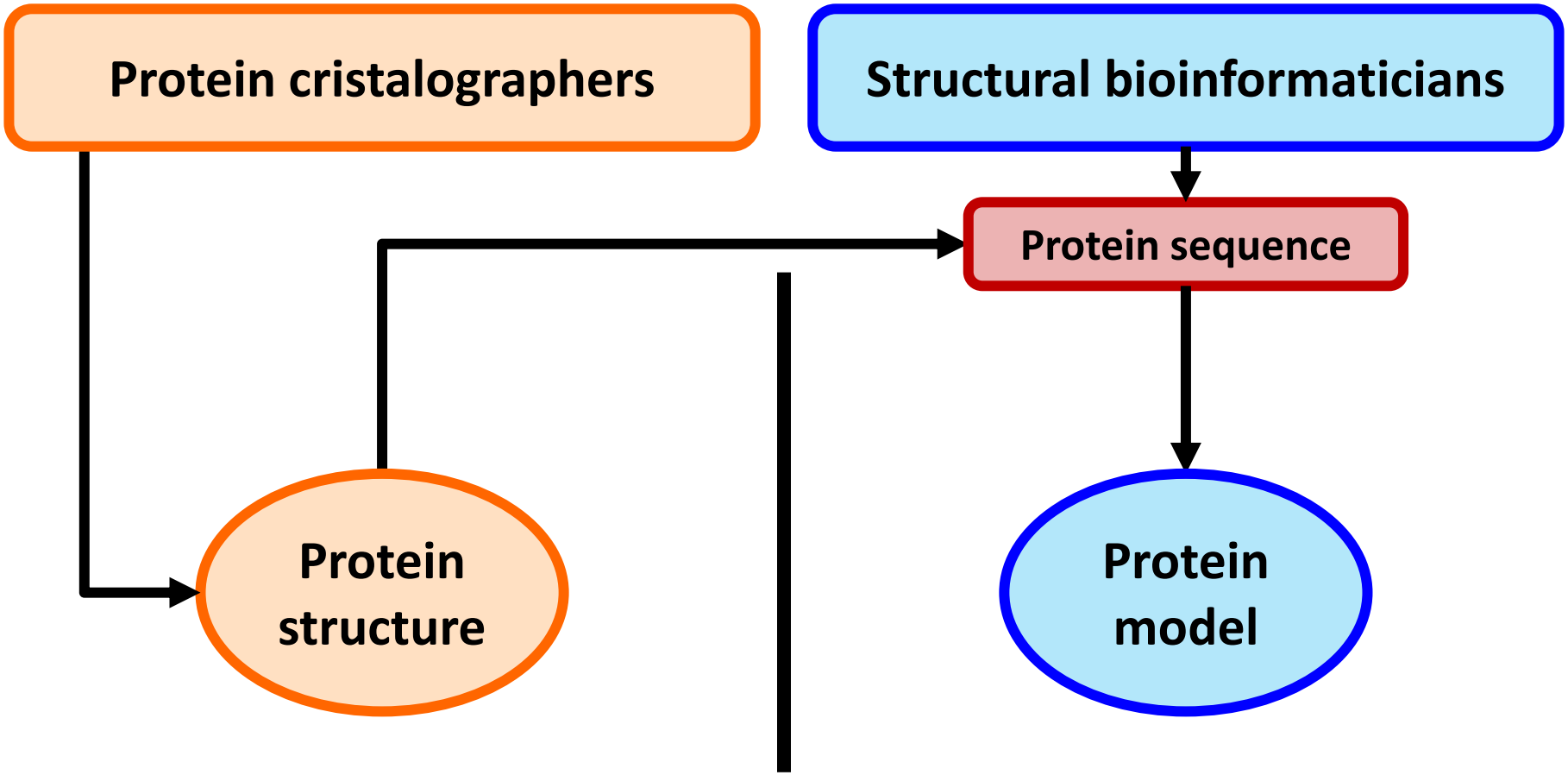
CASP: Critical Assessment of Structure Prediction

Worldwide competition for protein structure prediction, it takes place every two years since 1994



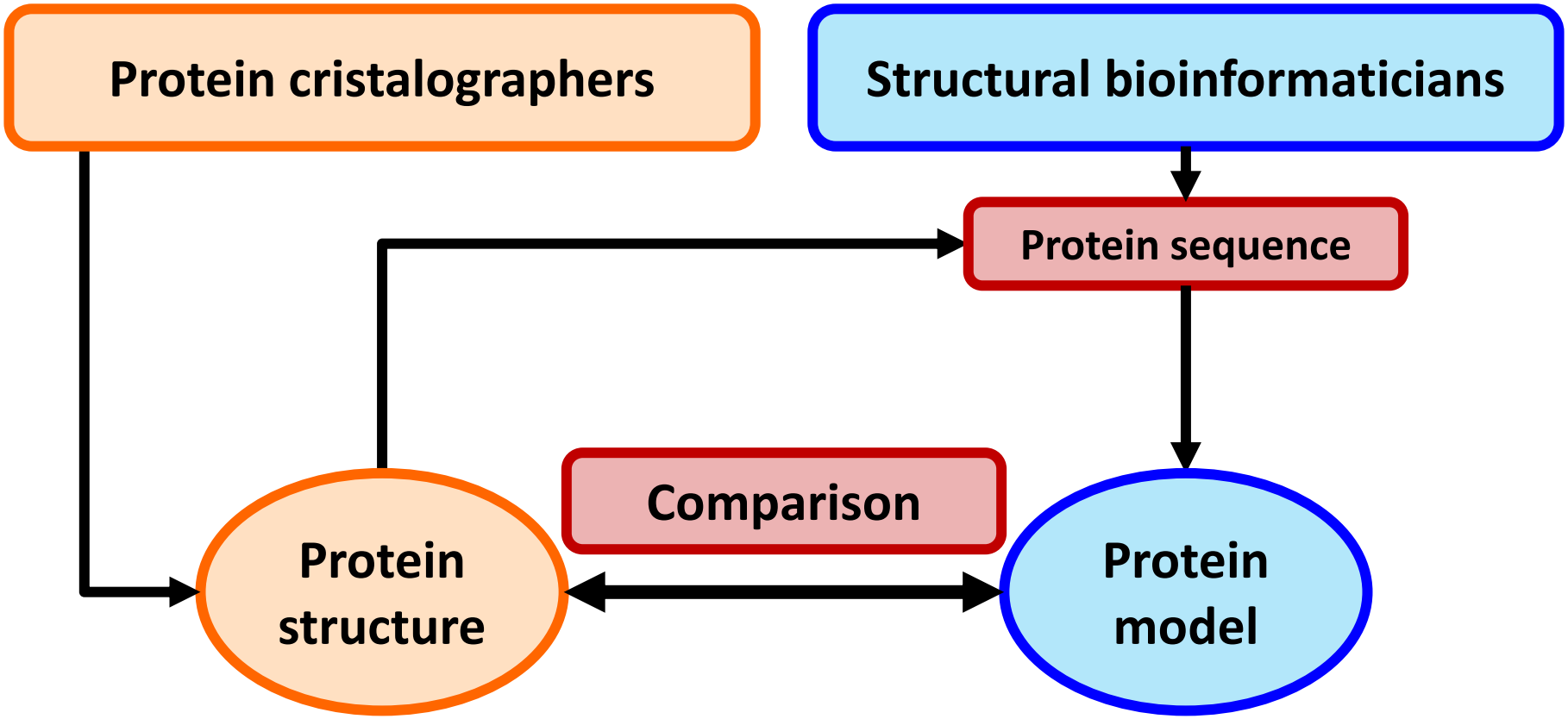
CASP: Critical Assessment of Structure Prediction

Worldwide competition for protein structure prediction, it takes place every two years since 1994



CASP: Critical Assessment of Structure Prediction

Worldwide competition for protein structure prediction, it takes place every two years since 1994



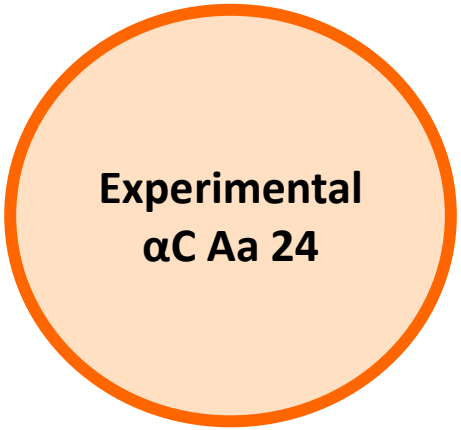
CASP: Critical Assessment of Structure Prediction

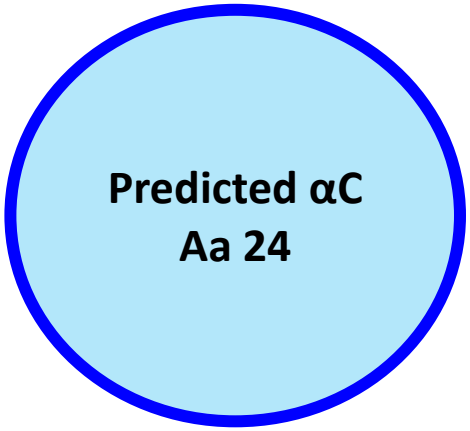
The performance in the CASP performance is evaluated with the GDT_TS score

Distance (Å)	Inside threshold
0.5	
1	
1.5	
2	
2.5	
3	
3.5	
4	

Until 10 Å...

We superimpose the experimental and the predicted structures. Then we evaluate if equivalent amino acids are within certain threshold distances:





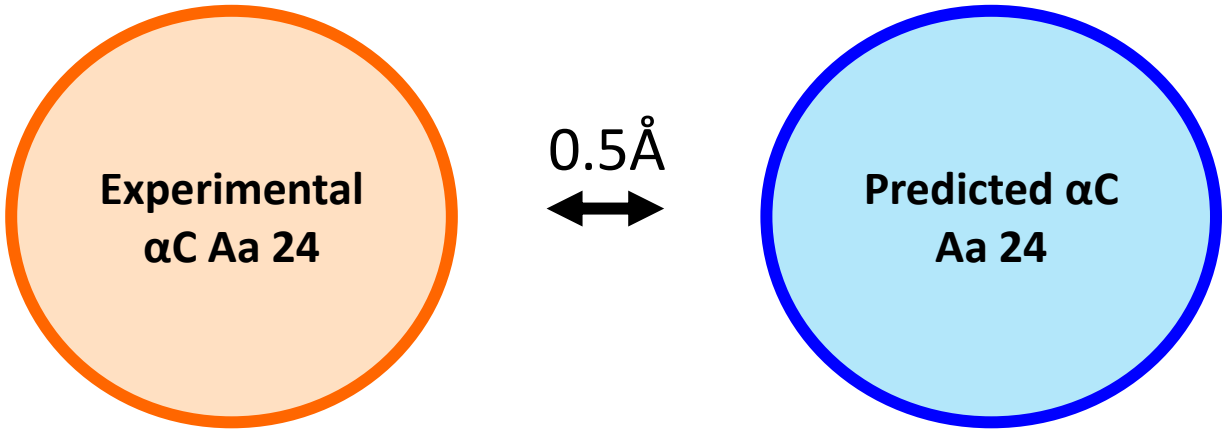
CASP: Critical Assessment of Structure Prediction

The performance in the CASP performance is evaluated with the GDT_TS score

Distance (Å)	Inside threshold
0.5	No
1	
1.5	
2	
2.5	
3	
3.5	
4	

Until 10 Å...

We superimpose the experimental and the predicted structures. Then we evaluate if equivalent amino acids are within certain threshold distances:



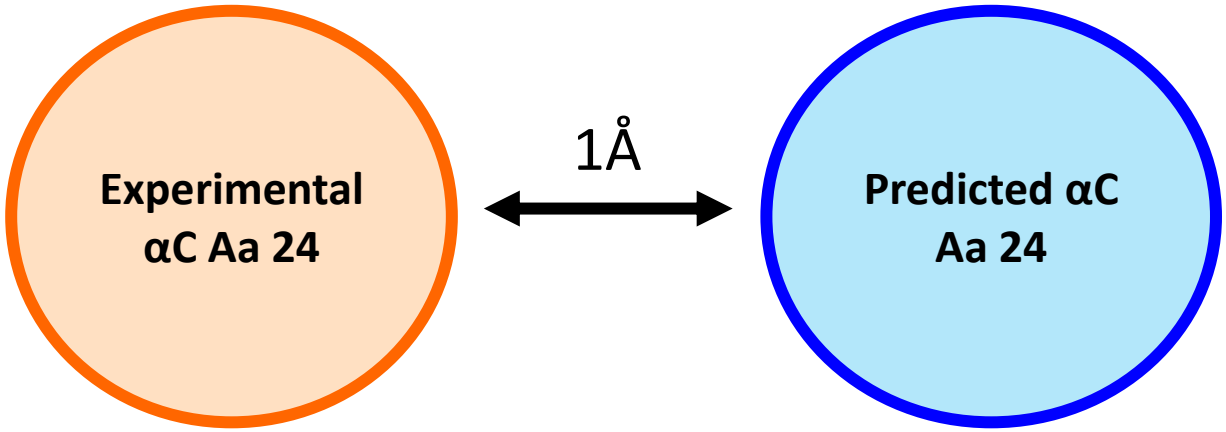
CASP: Critical Assessment of Structure Prediction

The performance in the CASP performance is evaluated with the GDT_TS score

Distance (Å)	Inside threshold
0.5	No
1	No
1.5	
2	
2.5	
3	
3.5	
4	

Until 10 Å...

We superimpose the experimental and the predicted structures. Then we evaluate if equivalent amino acids are within certain threshold distances:



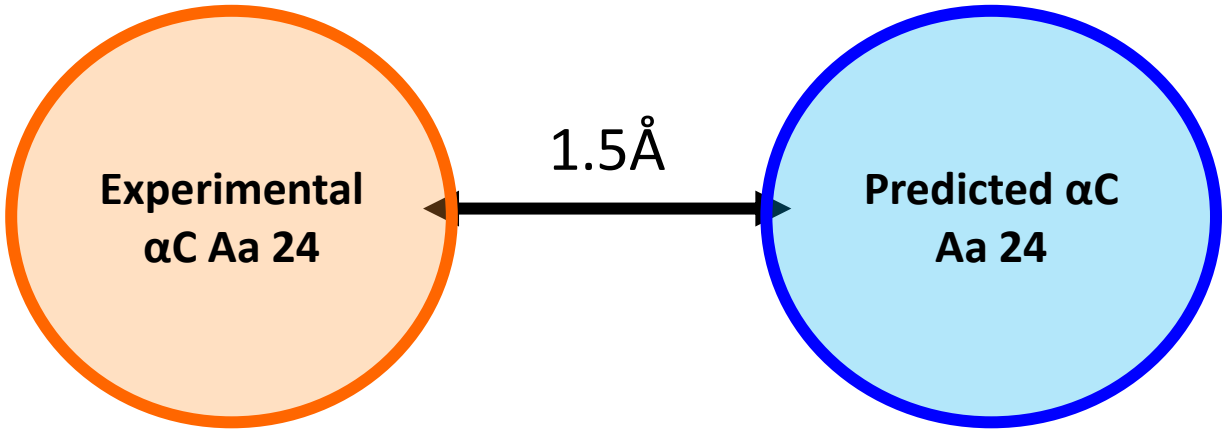
CASP: Critical Assessment of Structure Prediction

The performance in the CASP performance is evaluated with the GDT_TS score

Distance (Å)	Inside threshold
0.5	No
1	No
1.5	YES
2	YES
2.5	YES
3	YES
3.5	YES
4	YES

Until 10 Å...

We superimpose the experimental and the predicted structures. Then we evaluate if equivalent amino acids are within certain threshold distances:



CASP: Critical Assessment of Structure Prediction

The performance in the CASP performance is evaluated with the GDT_TS score

Distance (Å)	Inside threshold
0.5	10%
1	25%
1.5	40%
2	50%
2.5	60%
3	65%
3.5	70%
4	75%

Until 10 Å...

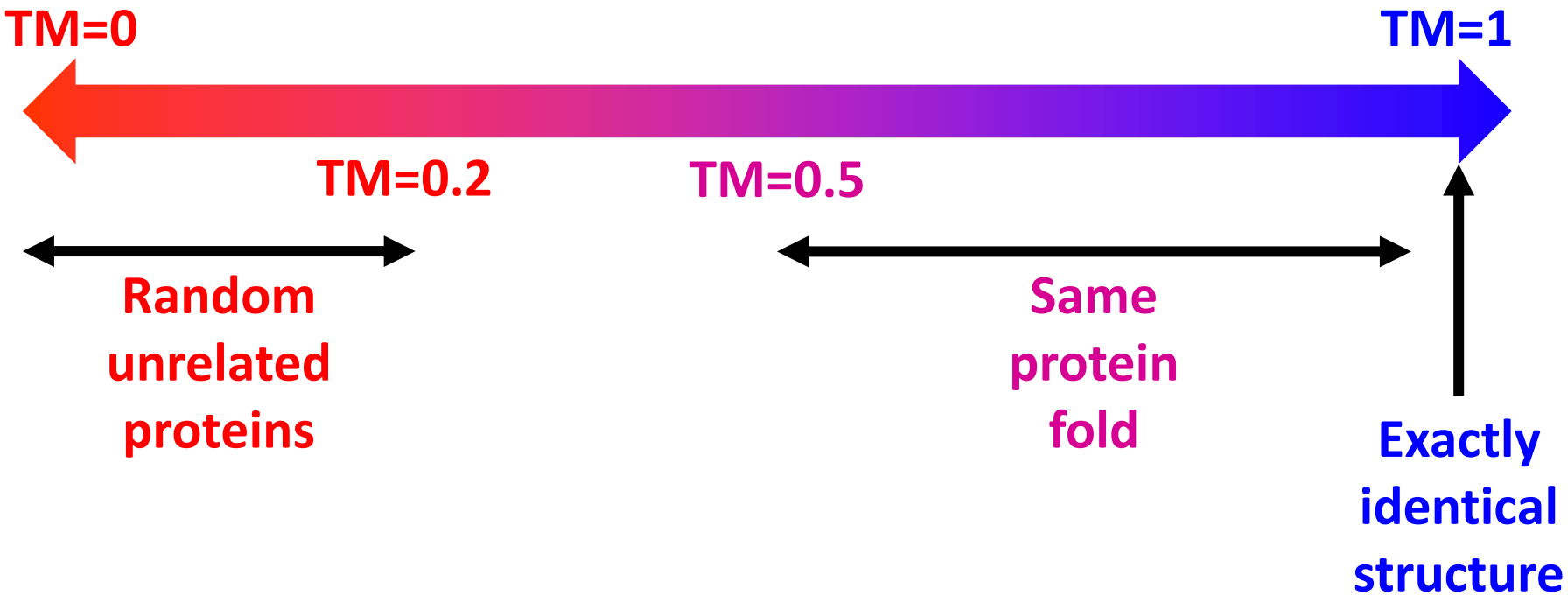
Then we calculate the % of amino acids that fit inside the threshold at different distances

The GDT_TS score is obtained ad the average of the percentage of amino acids within the threshold for distances of 1, 2, 4 and 8 Å

CASP: Critical Assessment of Structure Prediction

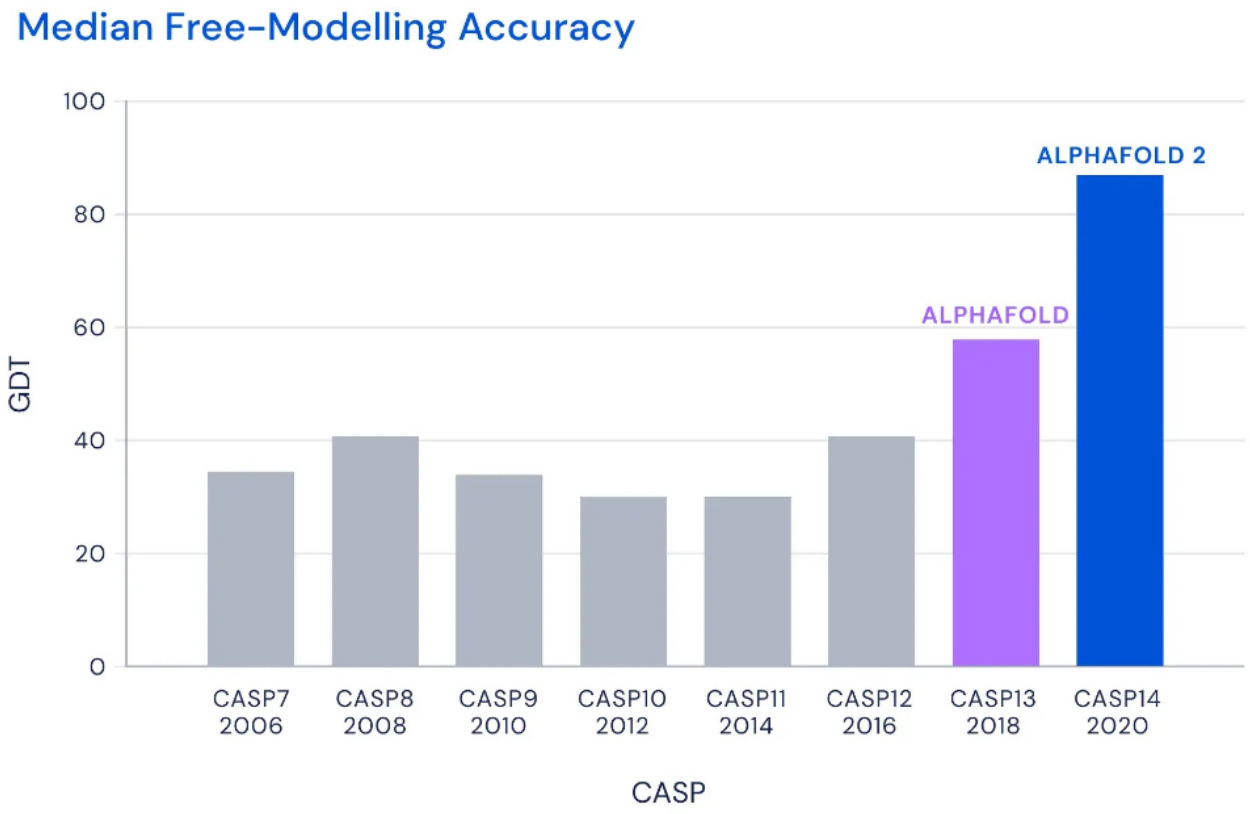
The performance in the CASP performance is also evaluated with the TM-score

The TM-score is a score derived from the RMSD of the superimposition of two proteins



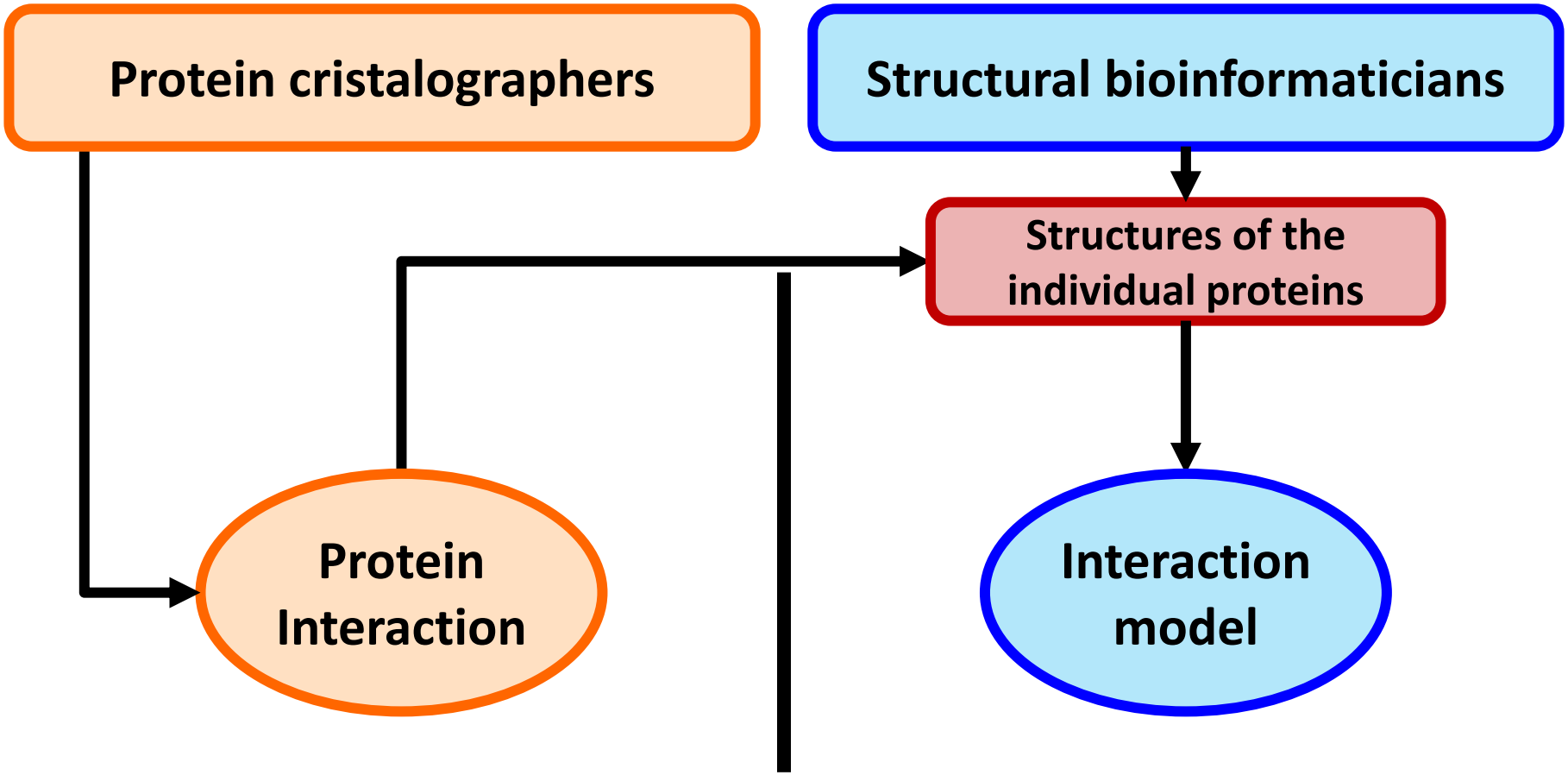
CASP: Critical Assessment of Structure Prediction

Alphafold became famous in the CASP competition, in which it showed an outstanding accuracy predicting protein structures



CAPRI: Critical Assessment of PRediction of Interactions

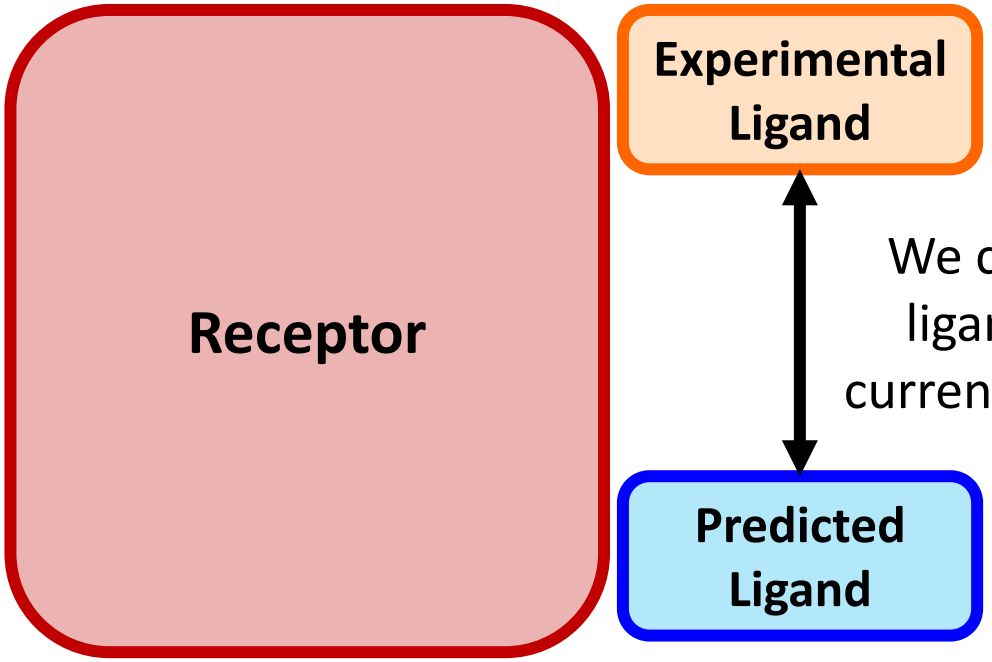
Similar competition to CASP, but instead of predicting protein folds they predict protein-protein interactions



CAPRI: Critical Assessment of PRediction of Interactions

Performance in CAPRI is evaluated by the ligand RMSD.

When predicting protein-protein interactions, the receptor is the bigger protein and the ligand is the smaller protein

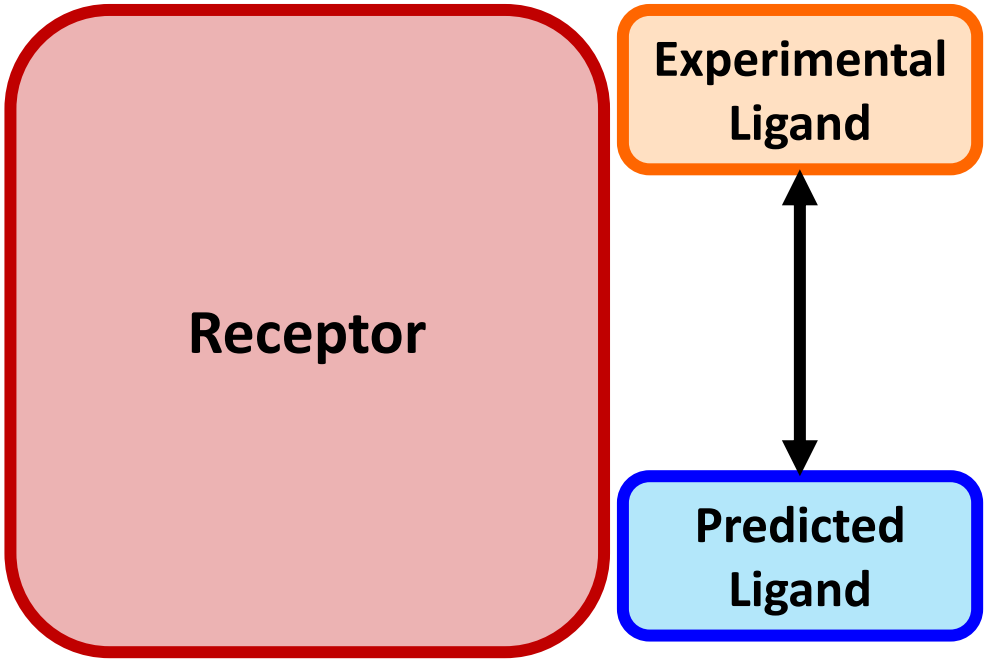


We compare the coordinates of the two ligands and calculate an RMSD at their current locations (without superimposition)

CAPRI: Critical Assessment of PRediction of Interactions

Performance in CAPRI is evaluated by the ligand RMSD.

When predicting protein-protein interactions, the receptor is the bigger protein and the ligand is the smaller protein



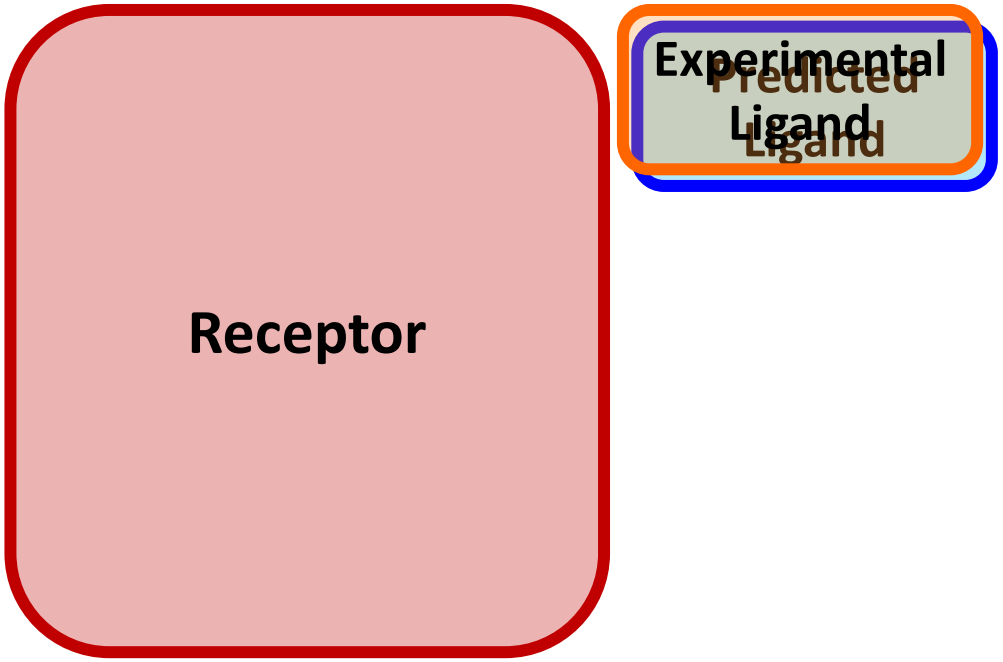
Ligand RMSD >> 0

Bad prediction

CAPRI: Critical Assessment of PRediction of Interactions

Performance in CAPRI is evaluated by the ligand RMSD.

When predicting protein-protein interactions, the receptor is the bigger protein and the ligand is the smaller protein



Ligand RMSD ≈ 0

Good prediction