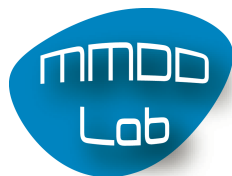




UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

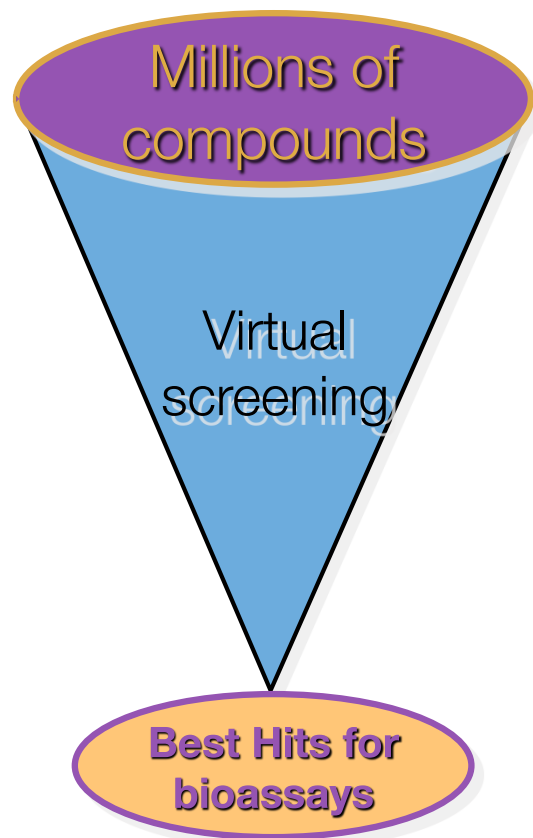


Structure-Based Virtual Screening

Giulio Rastelli

University of Modena and Reggio Emilia

Virtual Screening

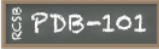


- **Rapid and inexpensive identification** of potential bioactive compounds from large collections of chemicals
- **Prioritize** compounds to be tested *in-vitro*
- **Molecular docking** is one of the most applied methods for virtual screening
- Docking is based on the analysis of ligand **complementarity** for the target active site in **geometric** and **energetic** terms
- By **ranking** compounds according to their predicted affinity score, the **prioritized** list of compounds can be used to rationally select a small subset of candidates for biological assays




Molecular Docking: Pros and Cons

- Fast and high-throughput method
- Less expensive compared to in-vitro screening
- Difficulties to simulate ligand and receptor flexibility
- Approximated scoring functions
- Poor agreement between estimated and experimental binding affinities
- False positives and negatives in the ranked lists

You need a crystal structure of the target

RCSB PDB PROTEIN DATA BANK  A MEMBER OF THE **PDB** | EMDatabank

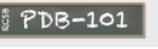


An Information Portal to **Biological Macromolecular Structures**
As of Tuesday Mar 12, 2013 at 5 PM PDT there are **88837** Structures | [PDB Statistics](#) | [Email](#) | [RSS](#) | [Facebook](#) | [Twitter](#)

Search  **Everything** Author Macromolecule Sequence Ligand 
Advanced 
Browse

[Search History](#) , [Previous Results](#)

Biological Macromolecular Resource

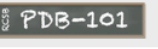
[Full Description](#)

Learn: Featured Molecules   **Biological Energy** 

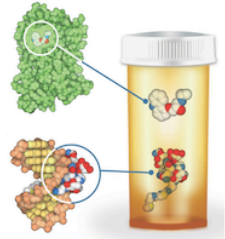
Structural View of Biology List View of Archive By: [Title](#) | [Date](#) | [Category](#)

Molecule of the Month
Erythrocrucorin
Hemoglobin comes in many shapes and sizes. In our blood, a **hemoglobin** with four chains carries oxygen from the lungs to cells throughout the body. Some plants build a single-chain hemoglobin to help protect sensitive nitrogen-fixing bacteria from oxygen, similar to the single chain **myoglobin** that stores oxygen in our muscle cells. Some bacteria also make simple forms of hemoglobin to help manage oxygen and other small molecules. Earthworms, however, are the champions when it comes to building huge hemoglobins. They, and a few other types of invertebrate animals, build enormous complexes of hemoglobin to carry their oxygen, termed erythrocrucorins.

[Full Article](#)

RCSB PDB News  [Weekly](#) | [Quarterly](#) | [Yearly](#)

2013-03-12
Access Drugs and Drug Targets in the PDB



Find drug stereoisomers, access DrugBank information, and browse

131k structures (June 2017)

www.rcsb.org

Filter: View: Reports: Sort:

✓ **1OHR**



VIRACEPT (R) (NELFINAVIR MESYLATE, AG1343): A POTENT ORALLY BIOAVAILABLE INHIBITOR OF HIV-1 PROTEASE

Authors: [Davies II, J.F.](#)

Release: 1998-12-09

Classification: [Aspartyl Protease](#)

Experiment: X-RAY DIFFRACTION with resolution of 2.10 Å

Residue Count: 198

Compound: 1 Polymer [[Display Full Polymer Details](#) | [Display for All Results](#)]
1 Ligand [[Display Full Ligand Details](#) | [Display for All Results](#)]

Citation: **Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease.** (1997) J.Med.Chem. **40**: 3979-3985 [[Display Full Abstract](#) | [Display for All Results](#)]

Molecule of the Month: [HIV-1 Protease](#)

Search Hit: Classification: ASPARTYL **PROTEASE** (Molecule of the Month:[HIV-1 Protease](#))

✓ **1HXW**



HIV-1 PROTEASE DIMER COMPLEXED WITH A-84538

Authors: [Park, C.H.](#), [Nienaber, V.](#), [Kong, X.P.](#)

Release: 1998-02-04

Classification: [Hydrolase/hydrolase Inhibitor](#)

Experiment: X-RAY DIFFRACTION with resolution of 1.80 Å

Residue Count: 198

Compound: 1 Polymer [[Display Full Polymer Details](#) | [Display for All Results](#)]
1 Ligand [[Display Full Ligand Details](#) | [Display for All Results](#)]

Citation: **ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans.** (1995) Proc.Natl.Acad.Sci.USA **92**: 2484-2488 [[Display Full Abstract](#) | [Display for All Results](#)]

Molecule of the Month: [HIV-1 Protease](#)

Search Hit: Classification: HYDROLASE/HYDROLASE INHIBITOR (Molecule of the Month:[HIV-1 Protease](#))

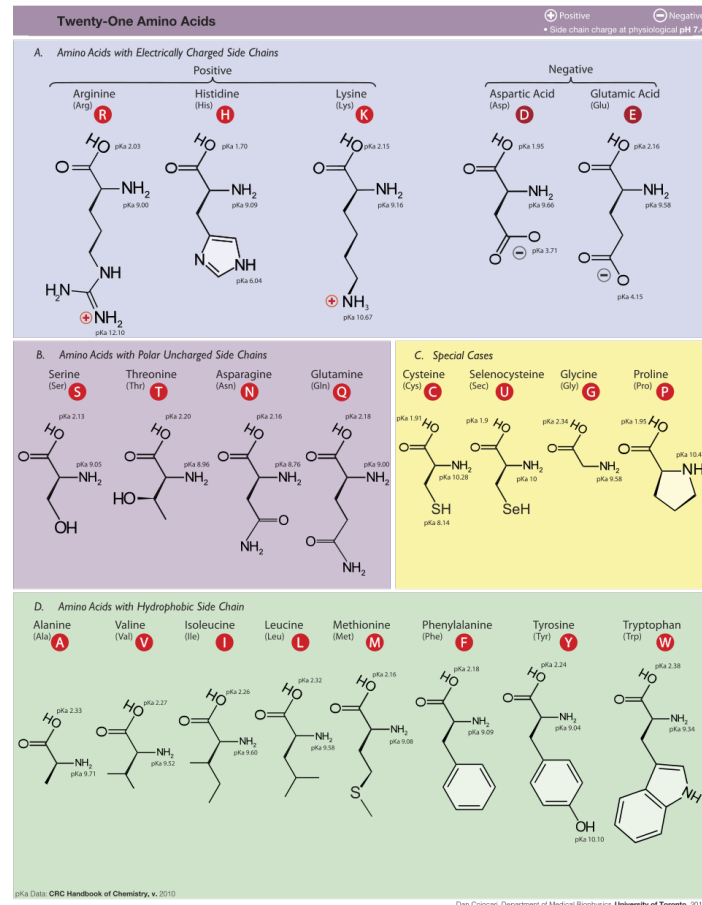
Then look for a (druggable) binding site

Shape

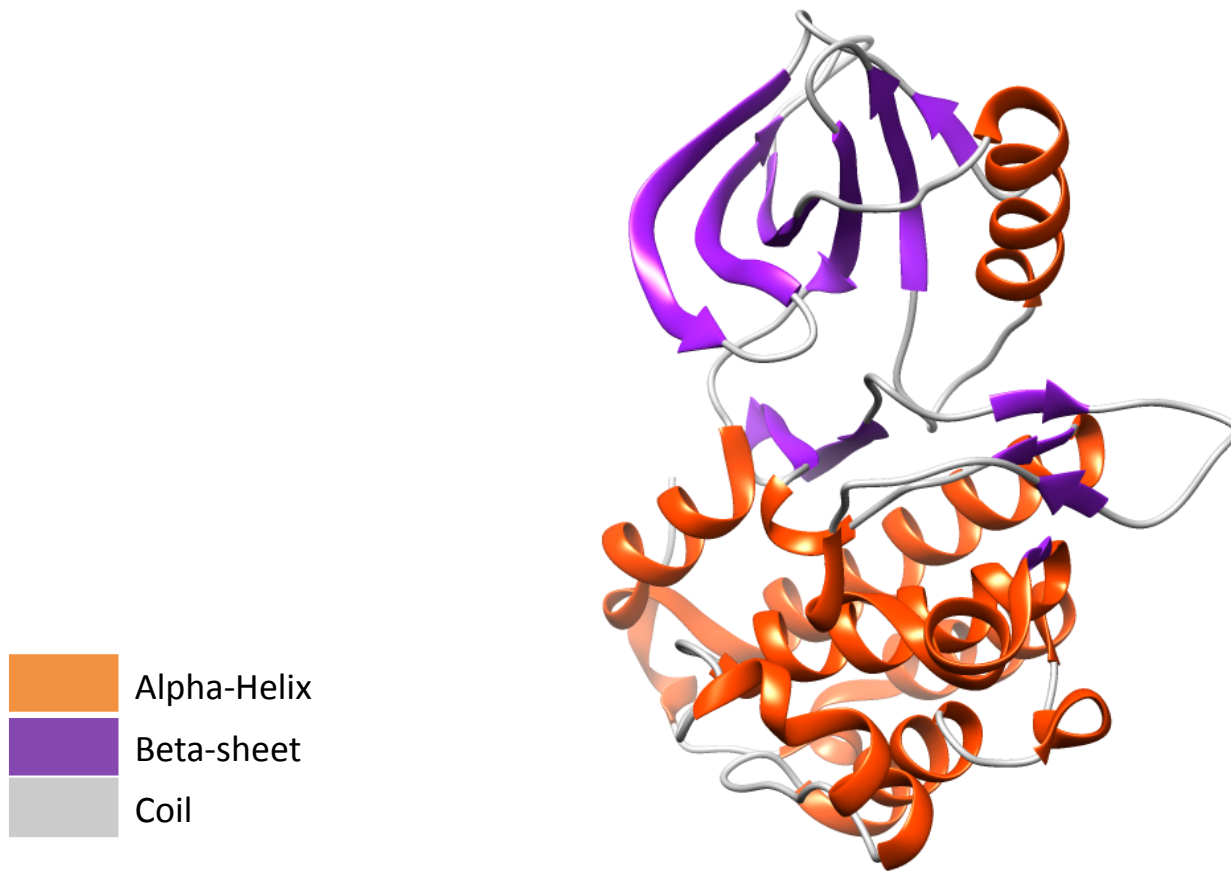
Composition

Solvent accessibility

Physchem properties



Visualization of a protein crystal structure



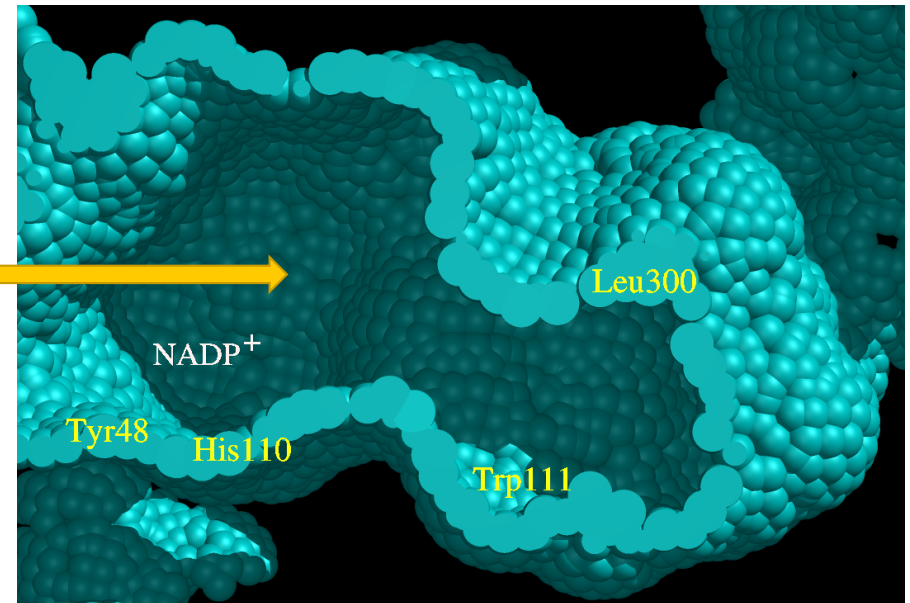
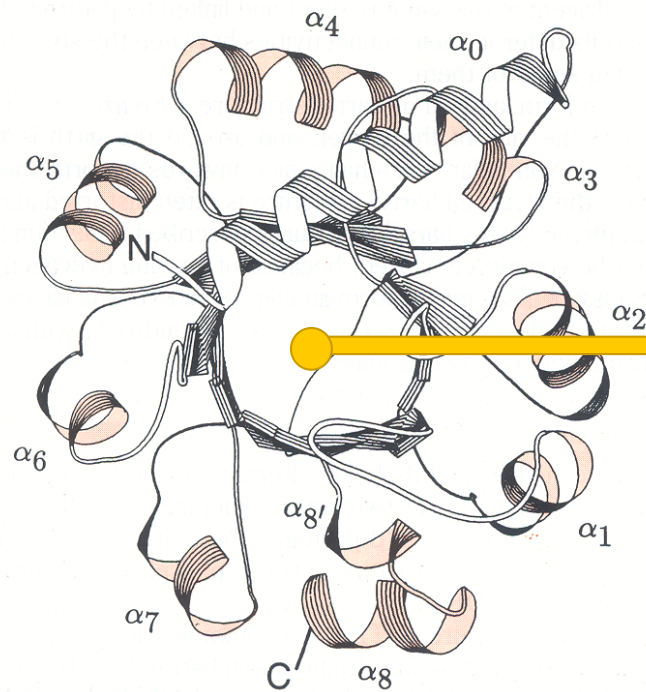
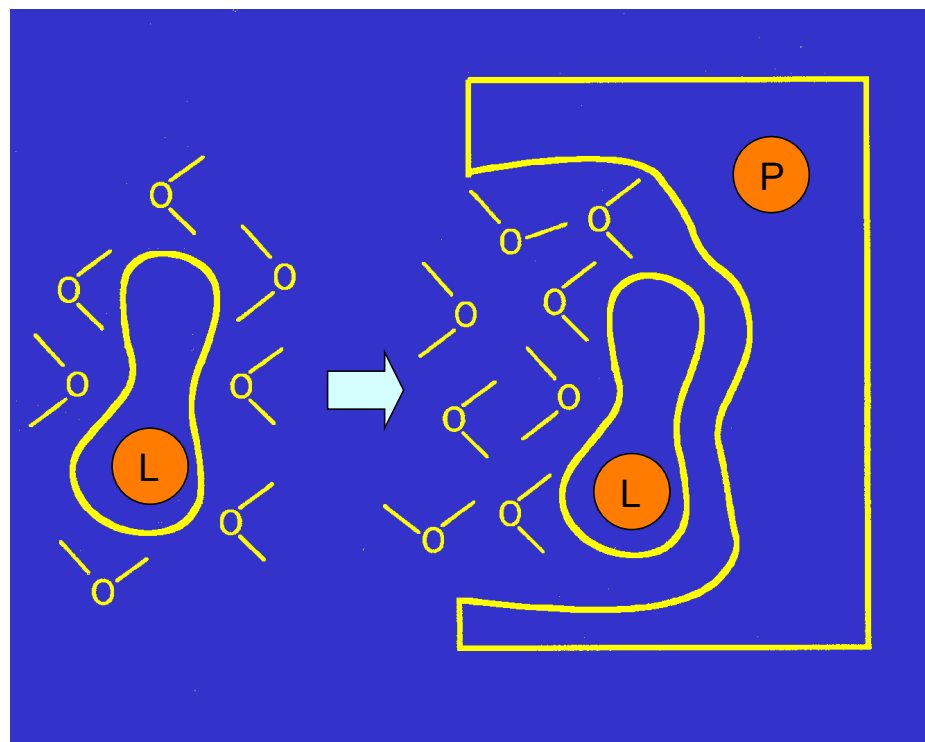
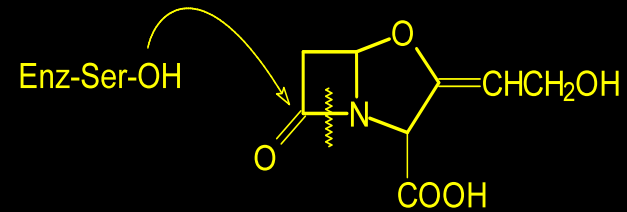


Figure 2.7. View of a typical $(\alpha/\beta)_8$ barrel protein down the axis of the barrel. The central β -strands all point out of the plane of the paper, whereas the α -helices project downwards. The particular structure shown is of indole glycerol phosphate synthase. It has an additional α -helix (α_0) at the N-terminus and a short helical segment ($\alpha_{8'}$) preceding helix 8. Reproduced with permission from Nermann, T. and Kirschner, K. (1990) *Protein Eng.*, 4, 137.

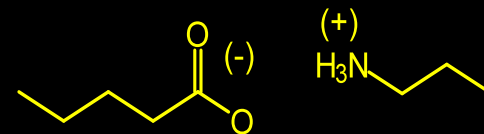
Prediction of a ligand (L) – protein (P) complex



Covalent bonds



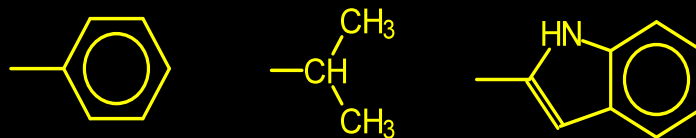
Electrostatic interactions

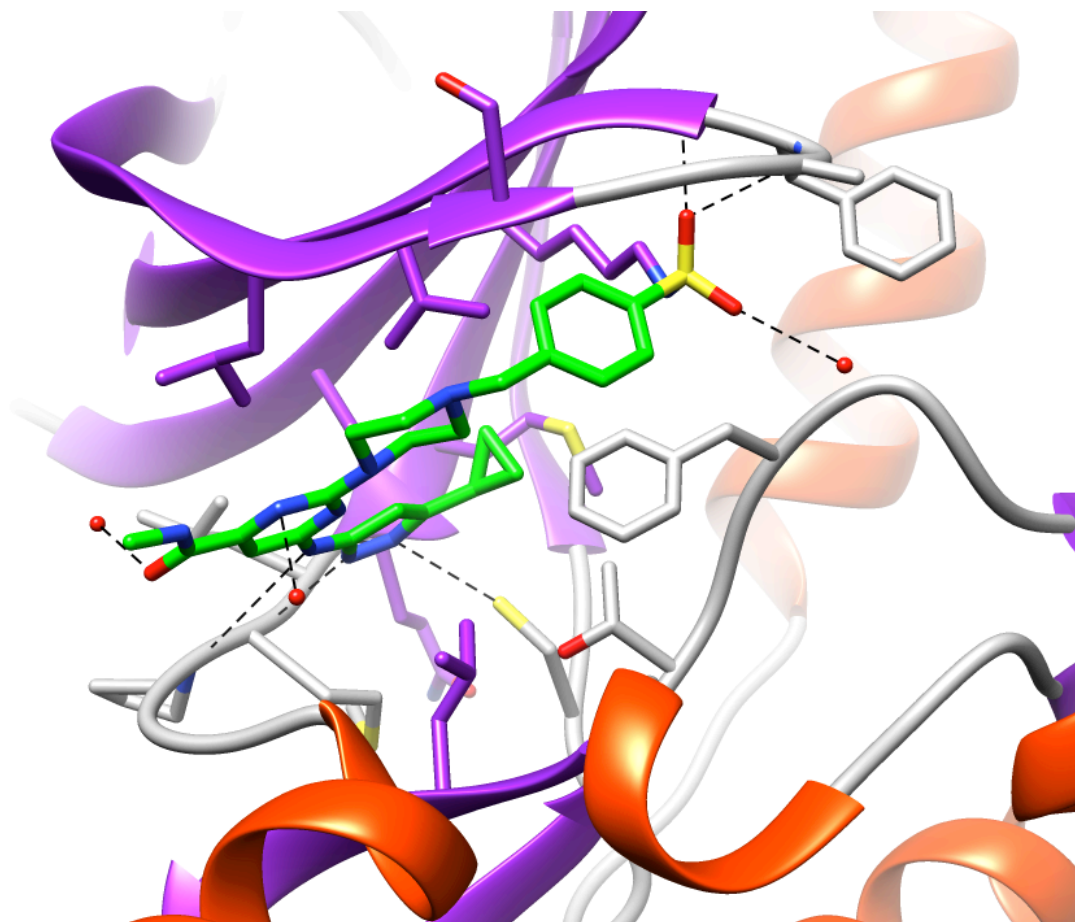


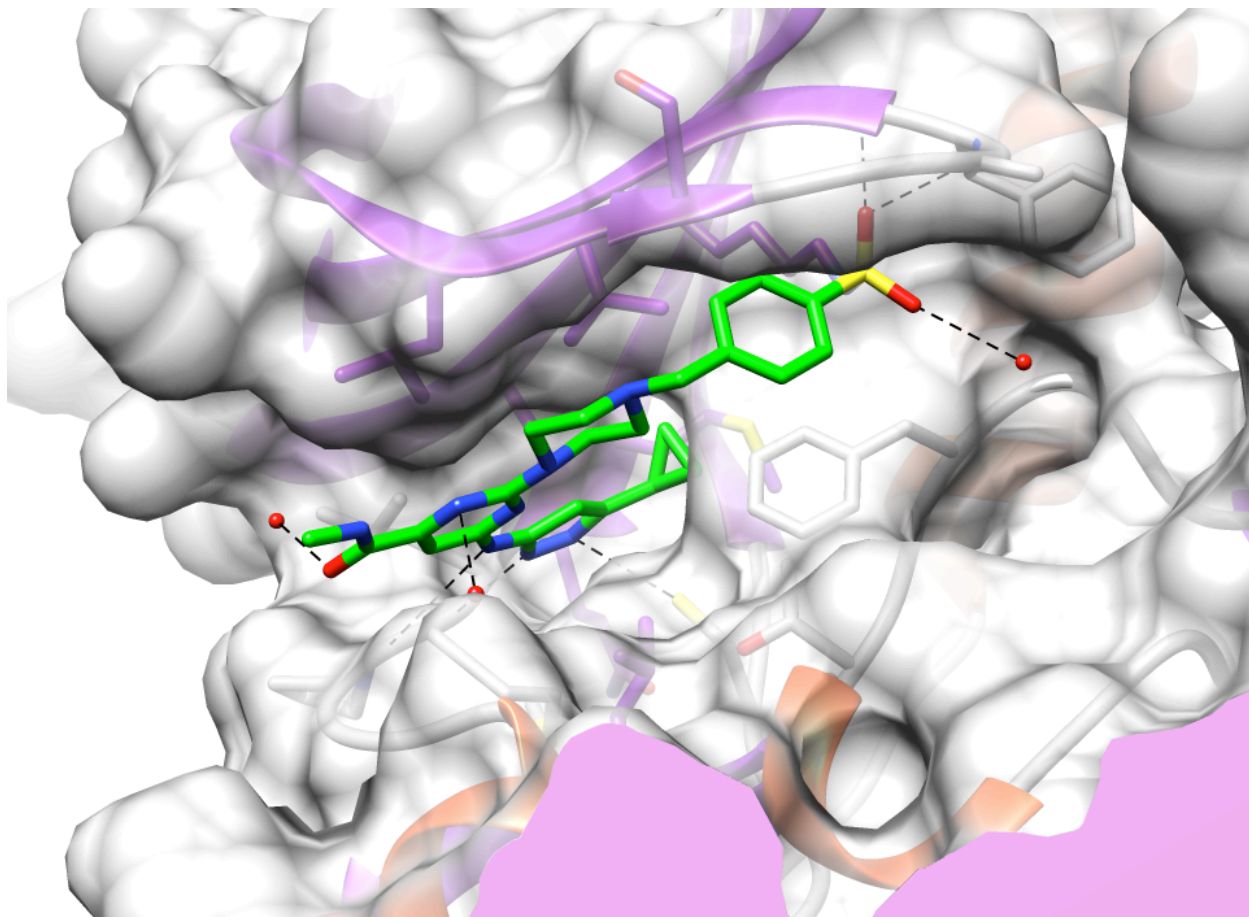
Hydrogen bonds

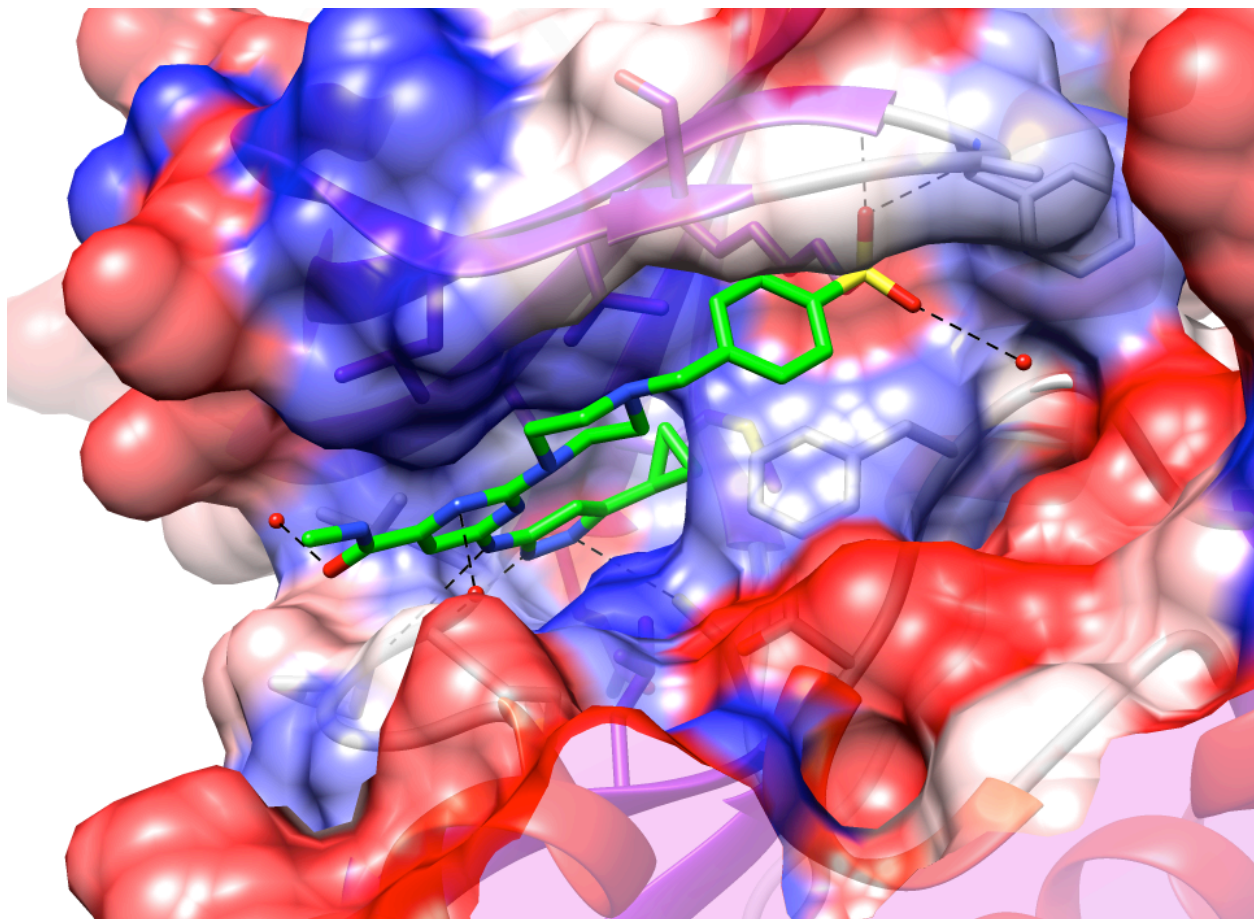


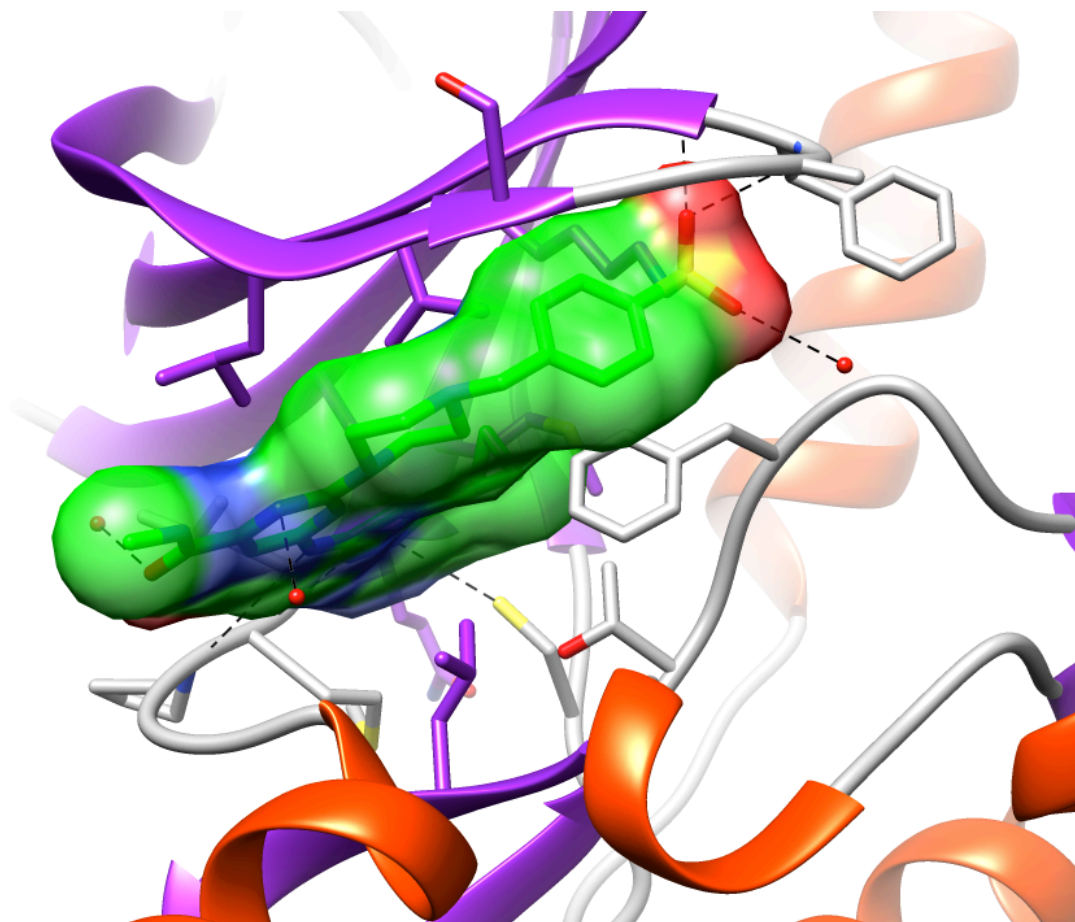
Hydrophobic interactions

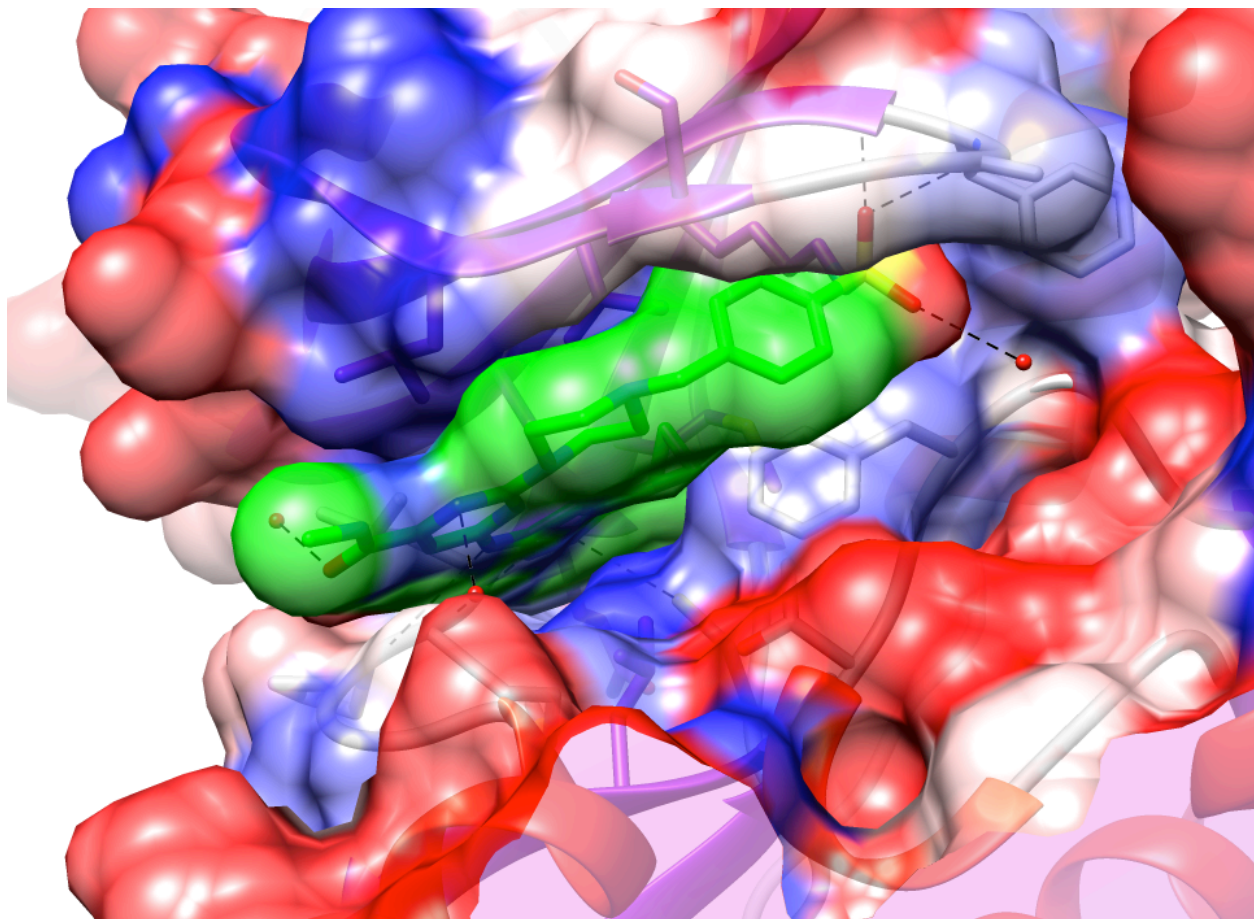




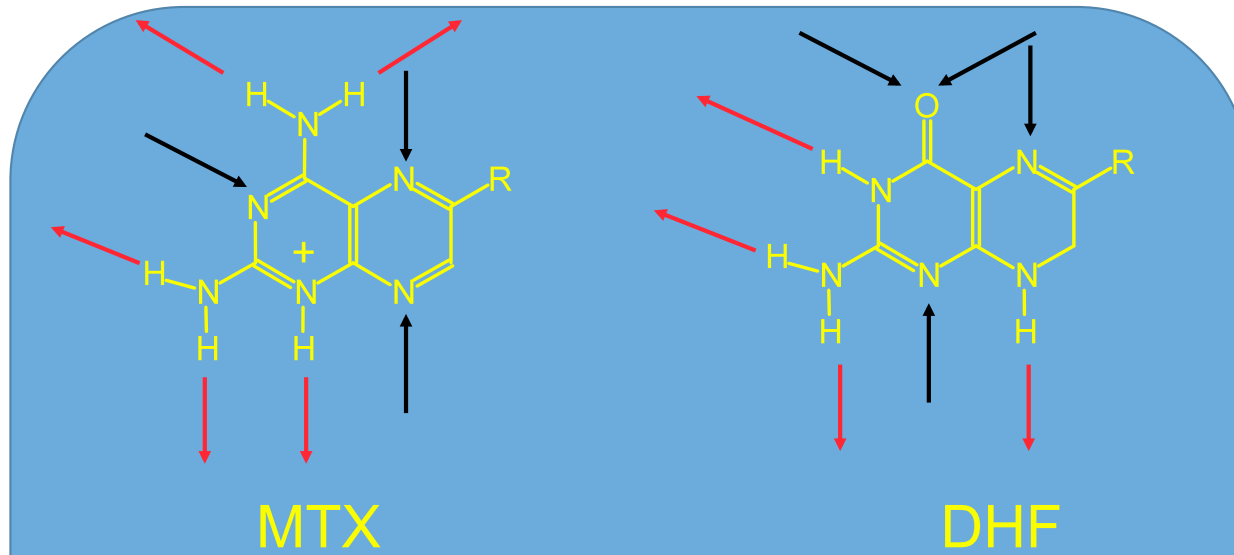






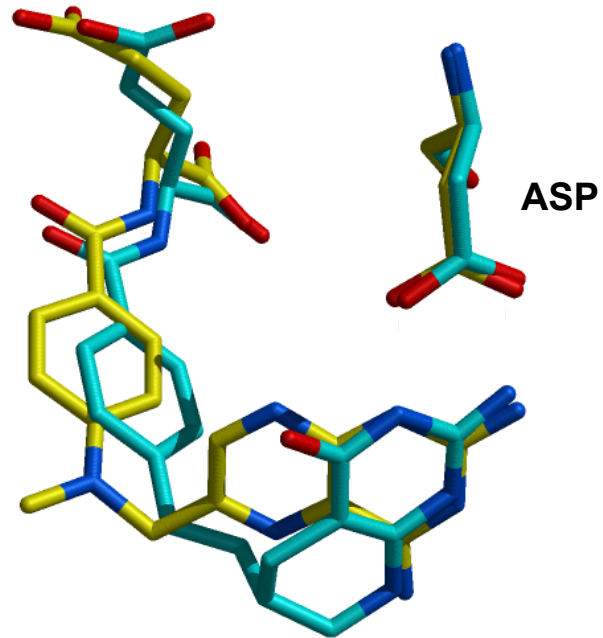


The importance of hydrogen bonds

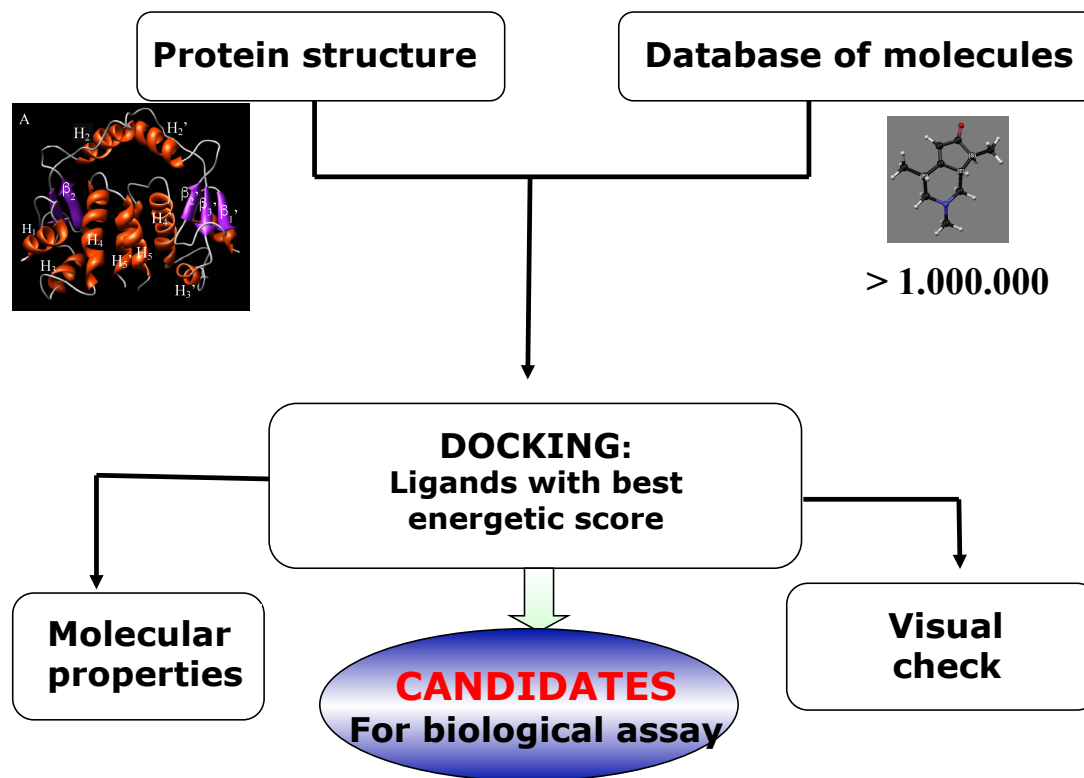


Testing the reliability of cristallographic structures

Overlapping of Methotrexate (**MTX, yellow**) in complex with DHFR and one substrate analog (**DHF, cyan**)



Structure-based virtual screening



What is molecular docking?

The aim of different docking methods is to **predict the interactions** of potential ligands in the binding site of a biological target.

Once the **affinity** (ΔG) of potential ligand-target binding is estimated, it is possible to **rank** the compounds, meaning that compounds are sorted with respect to the score (from more negative and favorable ΔG to less negative scores)

THE «DOCK» METHOD

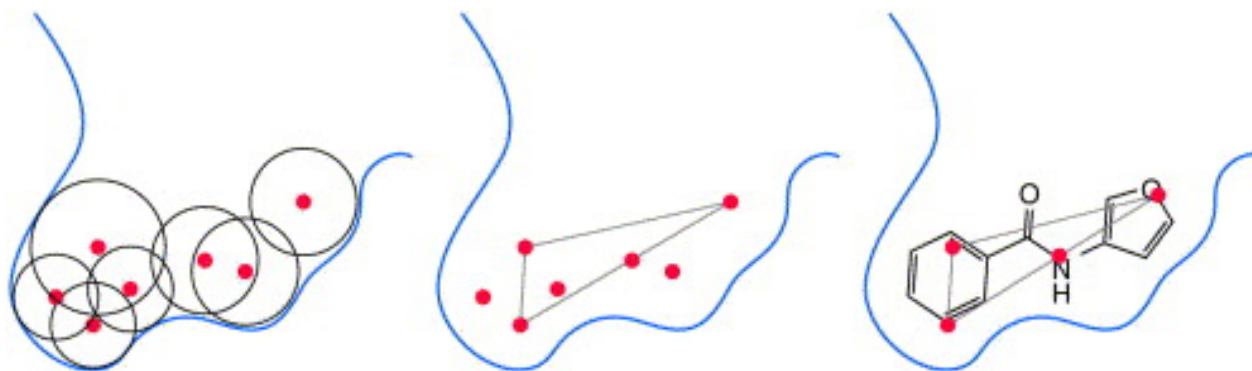
1) Generation of spheres describing the surface of the active site

Spheres have variable dimensions, in order to perfectly describe the concave and convex parts of the molecular surface of the active site of the macromolecules. Spheres can overlap.

2) Matching

To *steer* the ligand inside the active site part of the spheres' centers describing the active site will be *overlapped* with atoms of the ligands. The center of the spheres will be paired with the atoms of the ligand.

This matching step generates as many orientations of the ligand inside the site as the possible spheres/atoms overlapping.



3) Scoring

Any orientation is scored (positively or negatively) through **two criteria** :

a) Steric conflict with the macromolecule: if the orientation generates steric conflict between the ligand and the macromolecule, the orientation will be discarded

b) Highest ligand-macromolecule interaction energy : a score is assigned to each orientation that satisfies the steric criteria (point a)) through calculations of the ligand-macromolecule interaction energy. This is calculated adding the Van der Waals interactions (hydrophobic) and electrostatic energies.

$$E_{\text{int}} = \sum E_{\text{vdW}} + E_{\text{elect}}$$

Methodological aspects involved in molecular docking

What can we expect from docking?

Given the tridimensional structure of the target and one database of possible ligands we can expect to find molecules, different from known ones, able to bind the target. We can also expect to be able to predict their affinity.

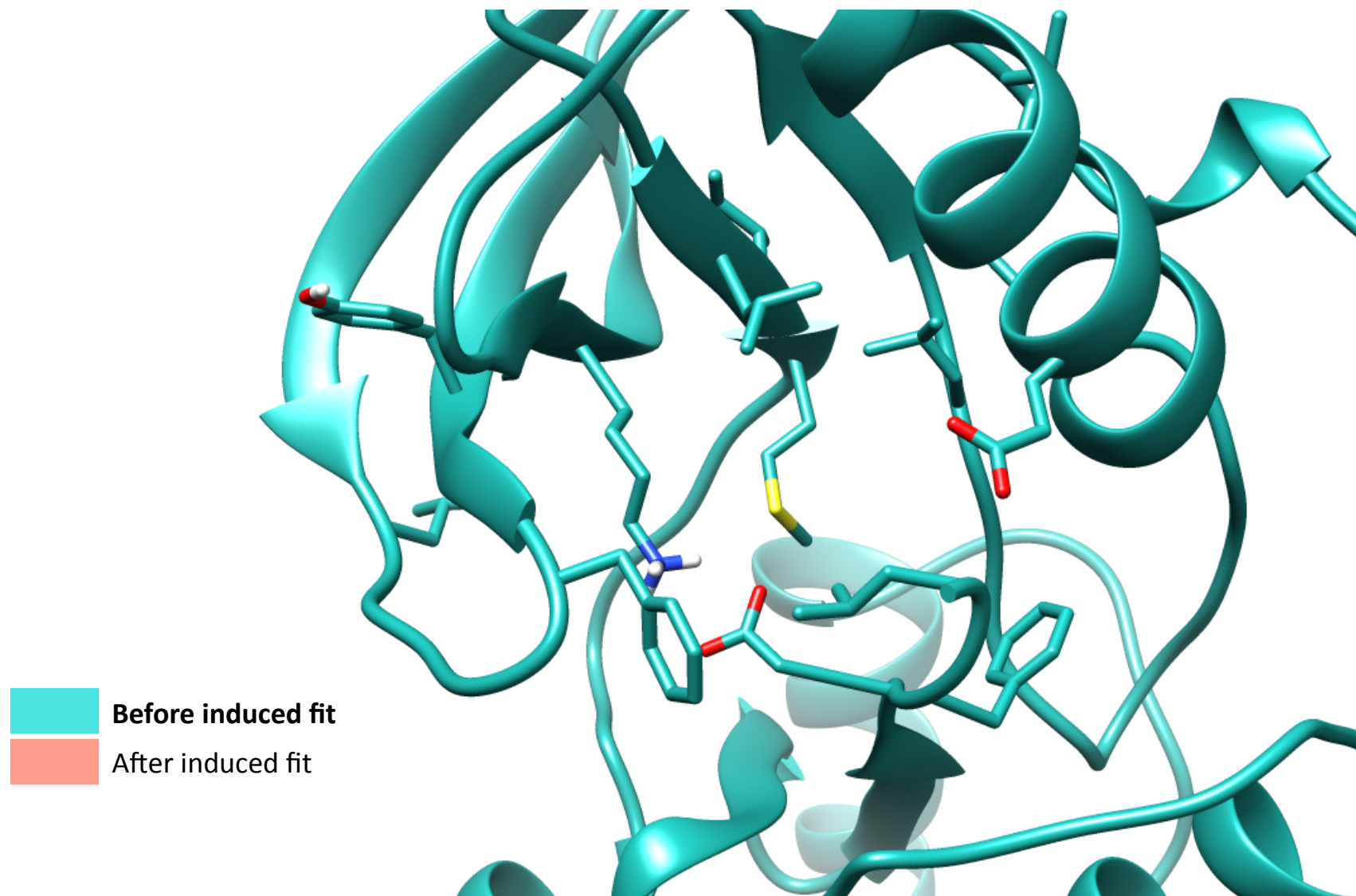
Which factors are mostly affecting the quality of docking results?

1. Ligand conformational freedom
2. Target conformational freedom
3. Solvent contribution
4. Effect of partial atomic charges
6. Calculation speed
7. Calculation of binding ΔG : scoring functions

1. Ligand flexibility : almost every docking algorithm used today relies on the possibility to evaluate the conformations accessible to the ligand, and to obtain optimal conformations to bind the target site (induced fit).
The programs usually differ on the methods used for the conformers generation.

2. Receptor flexibility : the structure of the target to be used in docking calculation represents one of the possible stable conformations of the target. It is rigid, especially when it comes from a crystal structure. Most of docking methods are not able to consider target flexibility, or can only take into account a limited flexibility, related to a small selection of aminoacids. Usually, the accuracy is also limited.

Which is the best structure to be used for molecular docking?



3. Solvation: The effect of the solvent in assembling the ligand-target complex is pivotal. The contribution of solvation may be calculated with different methods and is added to the other components of the function:

$$\Delta G_{\text{bind}} = \Delta G_{\text{interaction}} - \Delta G_{\text{solv (L)}} - \Delta G_{\text{solv (R)}}$$

4. Effect of partial atomic charges: partial atomic charges of the ligand and receptor's atoms can be assigned with different methods. A.e. Gasteiger-Marsili, semi-empirical methods. The «quality» of the used charges plays a big role on the accuracy of docking results.

6. Calculation speed: A compromise has to be made between protocol accuracy and required time for each molecule, especially in virtual screening procedures with many molecules.

7. Scoring function:

SCORING FUNCTIONS

Scoring methods must predict the orientation (or pose) of the potential ligand and predict its affinity for the target. An overall score is thus assigned to the ligand-target complex. This score has to be representative of the interaction energy (ΔG) between ligand and target.

$$\text{score} \propto \Delta G$$

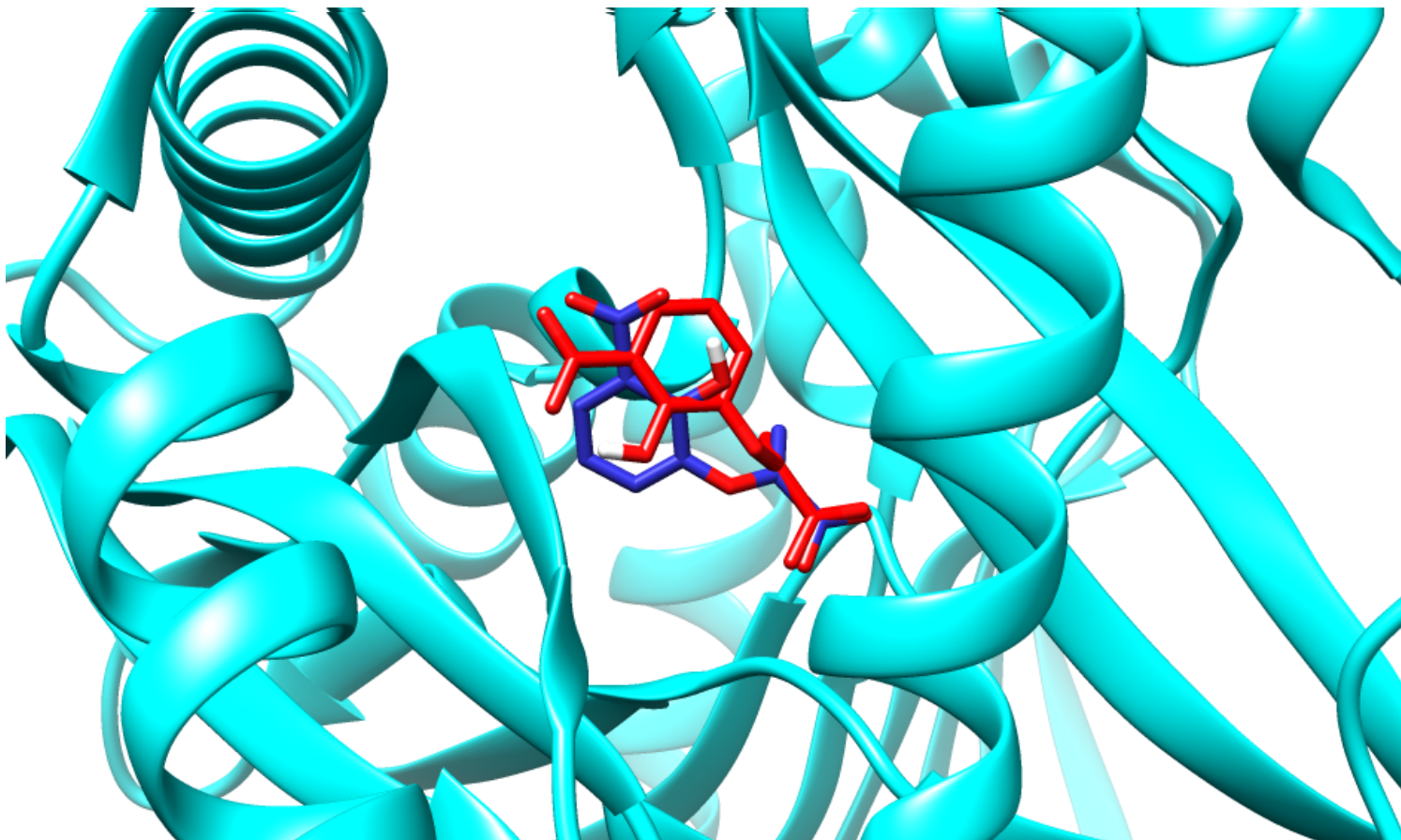
Energy interactions are correlated to the affinity of a given ligand for that target by this formula:

$$\Delta G_{\text{binding}} = -RT \ln K_{\text{affinity}}$$

This means that by computing the binding ΔG it is possible to adequately estimate the activity of a ligand toward a target. Therefore this value makes it possible to screen big databases and select the best ligands

In order to have an efficient scoring function, some requirements should be met:

- a. The generated ligand orientations have to be correctly ordered, meaning that the pose with a higher similarity to the experimental one should have a better score.**
“REDOCKING”
- b. If more than one ligands are docked in the same active site, the relative binding energy has to be correctly sorted. This means that ligands with higher affinity should have better score with regards to ligands with lower affinity, and clearly distinct from inactive molecules.**
- c. The scoring function has to be fast enough to be included in a scoring program. This is particularly true when these methods must be applied to screen a high number of chemical compounds.**



SCORING FUNCTIONS

- **FORCE FIELD based**
- **KNOWLEDGE-BASED**
- **EMPIRICAL**
- **CONSENSUS**

FORCE FIELDS based SCORING FUNCTIONS

These methods use a classic energetic function from molecular mechanics (Amber, Charmm,... force fields) to compute the score. Binding energies of the ligand-target complex are approximated as a sum of the Van der Waals and electrostatic interactions between pairs of atoms of the ligand-target complex. A correction is applied to account for solvation effects.

The free energy $\Delta G_{\text{binding}}$ is:

$$\Delta G_{\text{binding}} = \Delta H_{\text{binding}} - T\Delta S_{\text{binding}} = \Delta G_{\text{interaction}} + \Delta G_{\text{solvation}} - T\Delta S_{\text{binding}}$$

The contribute of the solvation energy ($\Delta G_{\text{solvation}}$) is actually decomposed in two components:

$$\Delta G_{\text{solv}} = \Delta G_{\text{elettrostatic solv}} + \Delta G_{\text{non polar solv}}$$

where:

$\Delta G_{\text{elettrostatic solv}}$ electrostatic component usually computed with the Poisson-Boltzmann equation or with the Generalized Born

$\Delta G_{\text{non polar}}$ non polar component generally proportional to the area of the surface accessible to the solvent

The entropic contribution (ΔS) is difficult to predict and very often is overlooked.

Advantages: Accuracy

Disadvantages: methods based on these scoring function tend to require longer computational times, due to the number and the complexity of the energetic terms.

KNOWLEDGE-BASED SCORING FUNCTIONS

1. More frequent ligand-target interactions are favored from an energetic point of view and thus they have a positive contribution on the binding affinity.
2. Using Boltzmann distribution equation it is possible to convert the probability of finding an **atom A** of the ligand at a **distance r** from **atom B** of the protein in terms of energy interaction between A and B as functions of r.

Knowledge-based functions derive from the observation of statistical analysis on interatomic contacts between ligands and proteins of a wide sample of crystallographic structures of complexes in the PDB.

They are based on the probability that a given interaction might happen between a determined pair of atoms (or better said, “atom types”)

The score is proportional to the sum of the interactions between all atom pairs and it is «weighted» on the probability that a given interaction might actually happen.

EMPIRICAL SCORING FUNCTIONS

The energetic score is represented as a binding ΔG .

The scoring function is calibrated on a set of protein-ligand complexes with known affinity binding data.

These functions are based on a series of empirical rules that take into account all atom types and their geometries for the different kinds of interaction.

The binding free energy is estimated as the sum of terms that resemble force field based scoring functions. However, in this case, contributions are **empirically calculated**.

In the sum, **the «weight» of every term in empirically parametrized**, so that the total scores ($\Delta G_{\text{binding}}$) for known ligand are the closest possible to the $\Delta G_{\text{binding}}$ values related to experimental binding constants of a given series of target-ligand complexes.

The first function of this kind is Bohm's function. It has five terms representing hydrogen bonds, ionic interactions, lipophilic interactions, number of rotatable bonds.

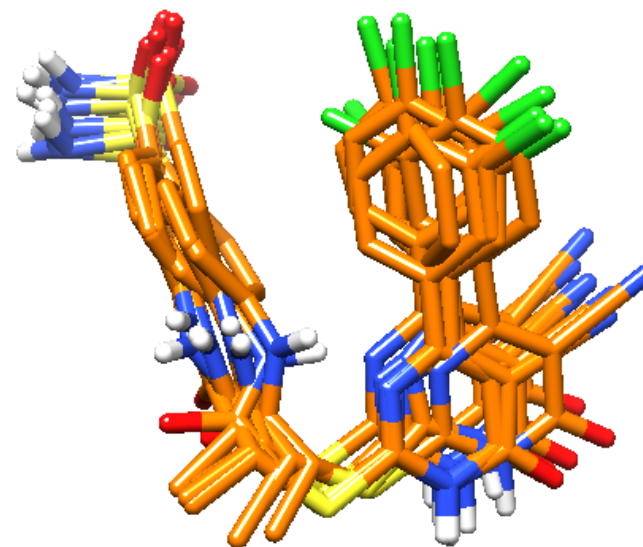
$$\begin{aligned}\Delta G_{\text{binding}} = & \Delta G_0 + \Delta G_{\text{hb}} \sum_{\text{hbonds}} f(\Delta R, \Delta \alpha) \\ & + \Delta G_{\text{ionic}} \sum_{\text{ionic int}} f(\Delta R, \Delta \alpha) \\ & + \Delta G_{\text{lipo}} \mid A_{\text{lipo}} \mid \\ & + \Delta G_{\text{rot}} N_{\text{rot}}\end{aligned}$$

CONSENSUS SCORING FUNCTIONS

In this approach several scoring function are used and then combined.

Highly scored ligand-target complexes in **two or more** scoring function are considered strong indication for binding.

This method drastically reduces the presence of **false positives**, either in choosing the most promising molecules for biological tests, but also in choosing the most “correct” orientation for the selected compounds



Enrichment factor

- **How to evaluate the «performance» of a scoring method?**
- The performance of a docking method can be evaluated taking into consideration:
 - 1) The ability to reproduce the correct orientation of ligands for which there are crystallographic complexes. (**redocking**)
 - 2) The ability to highly score known active ligands with respect to known inactive or untested molecules for a given target. (**enrichment factor**).
 - 3) The ability to identify new biologically active molecules inside database. These molecules will be selected and biologically tested (**hit rate**).

To determinate enrichment factors, a mixed database composed of known ligands and “**decoys**” is built.

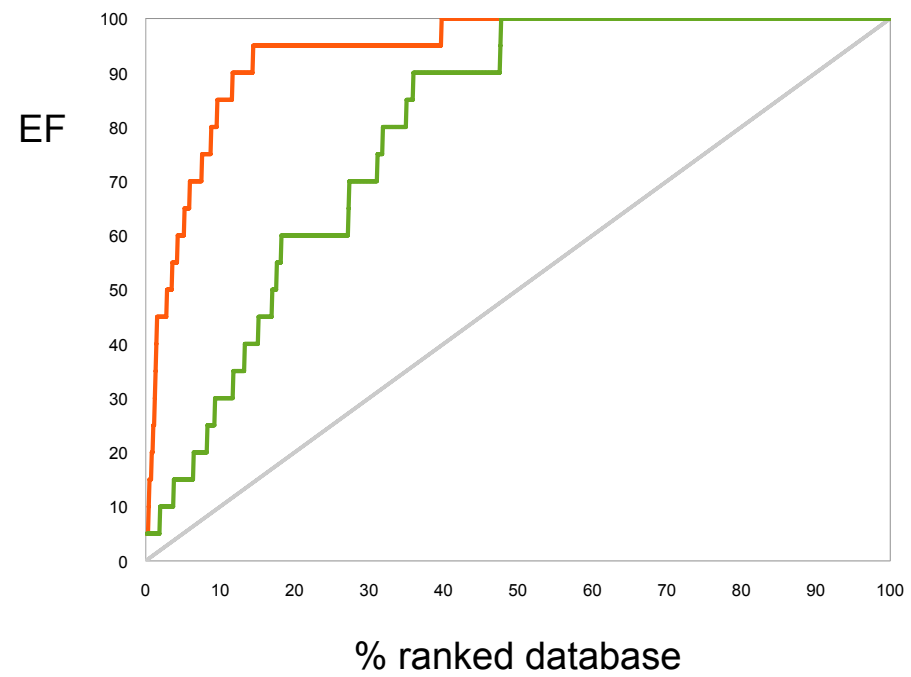
Decoys are molecules with **different chemical structure** from the ligands (so that they can not, theoretically, be considered as proper «ligands» for the target), but with **similar physico-chemical properties** such as Molecular Weight, LogP, number of groups able to make H bonds, ...

For each known ligand a fixed number of decoys is included (to keep, for example, a 1:50 ratio)

A docking-based virtual screening is performed, and compounds are then sorted based on their $\Delta G_{\text{binding}}$

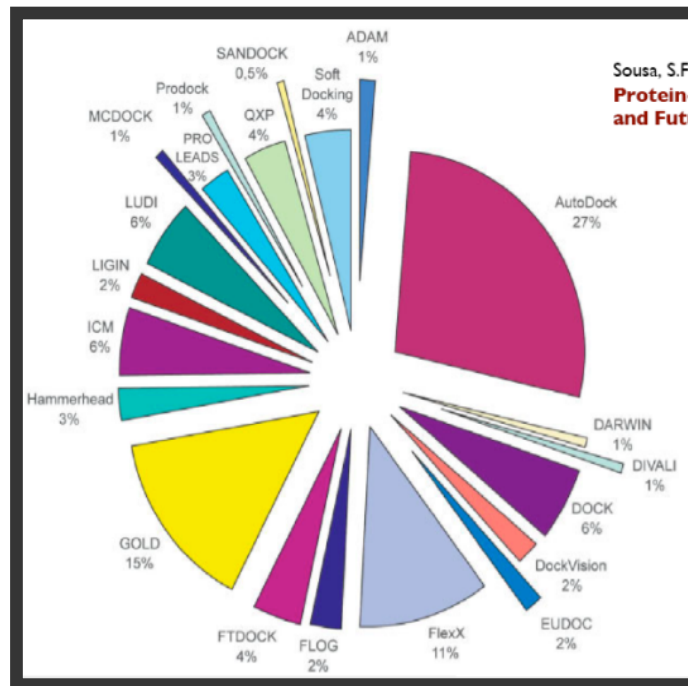
The enrichment factor is calculated evaluating how many known ligands are ranked with a high score, compared to the rest of the database.

$$\mathbf{EF} = \frac{\text{n}^\circ \text{ known ligands/percentage of database}}{\text{n}^\circ \text{ total ligands/entire database}}$$



MOLECULAR DOCKING METHODS

There are several docking algorithms, whose differences are both in the method used to look for the **orientation** of the potential ligand, but also on the **scoring** method.



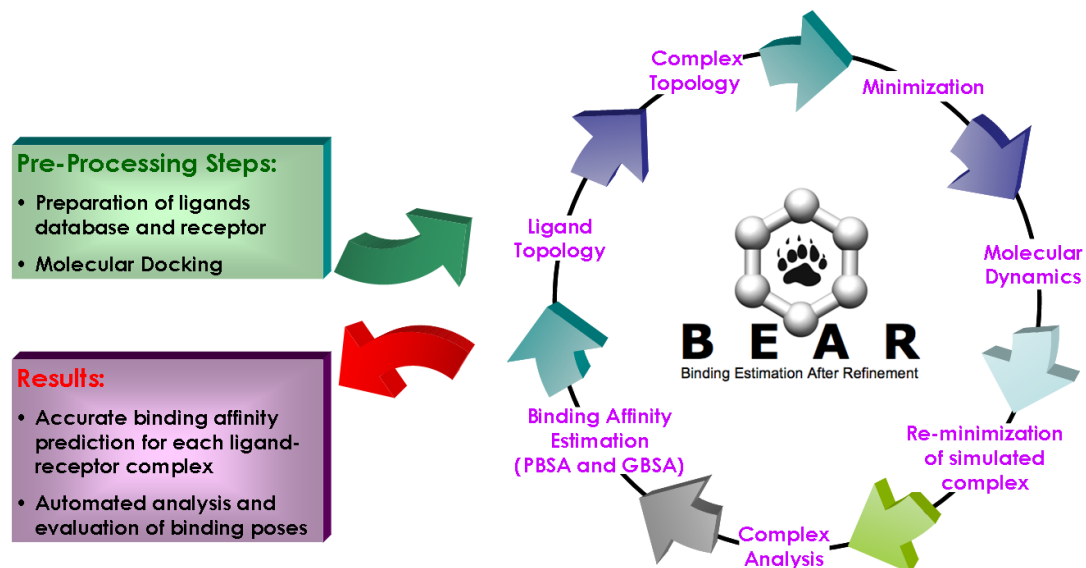
Post-docking approaches

- Develop an automated **post-docking method** specifically designed to improve docking results
 - Improved simulation of flexibility
 - MD subtask specifically devised to help overcome potentially high energy barriers between different conformations of the ligand in the target-binding site
 - Improved evaluation of binding affinity
 - MM-PBSA and MM-GBSA scoring functions taking into account the solvation contribution to the binding energy

BEAR (Binding Estimation After Refinement)

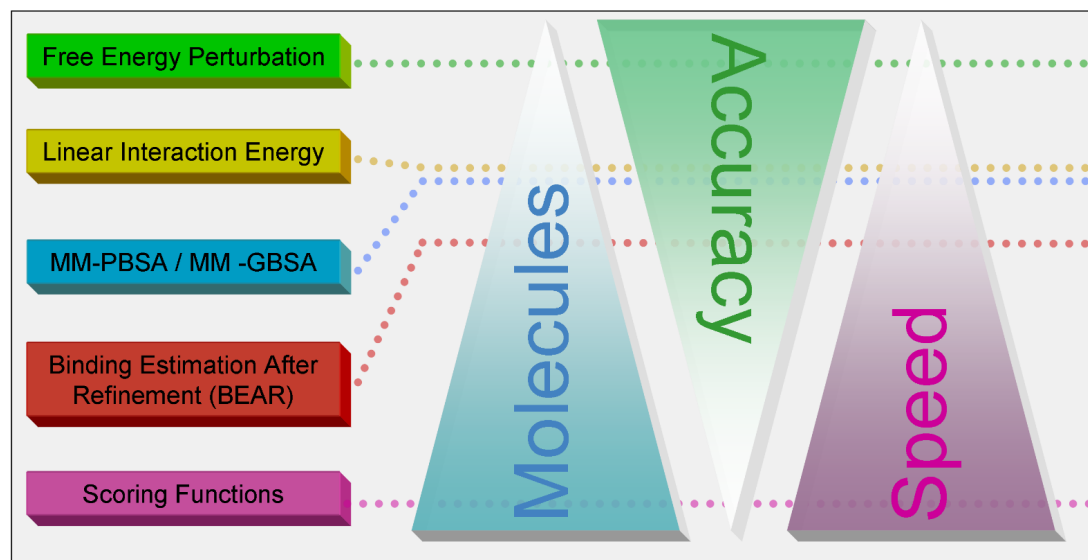
Simulation of flexibility using MD

Prediction of binding affinity using free energy-based scoring functions (MM-PBSA and MM-GBSA)



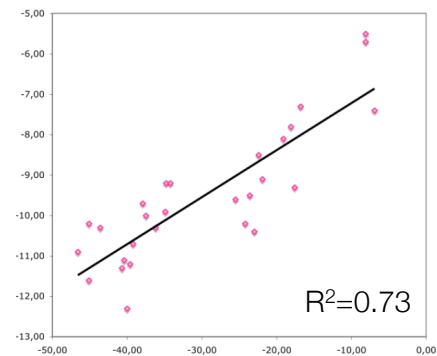
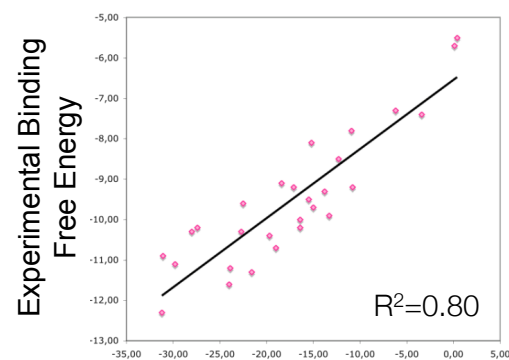
Force-field based scoring functions

More accurate but generally computationally intensive methods are applicable to a small number of compounds, while more approximate methods are usually faster but less accurate in predicting binding affinities

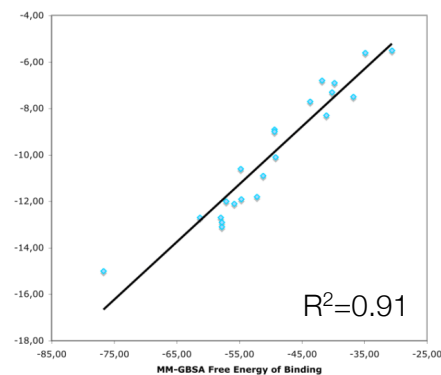
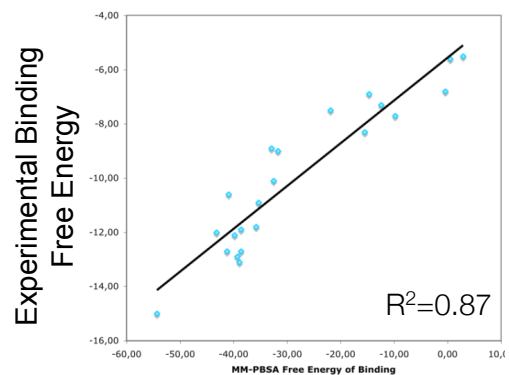


Scoring function reliability

ALR2



DHFR



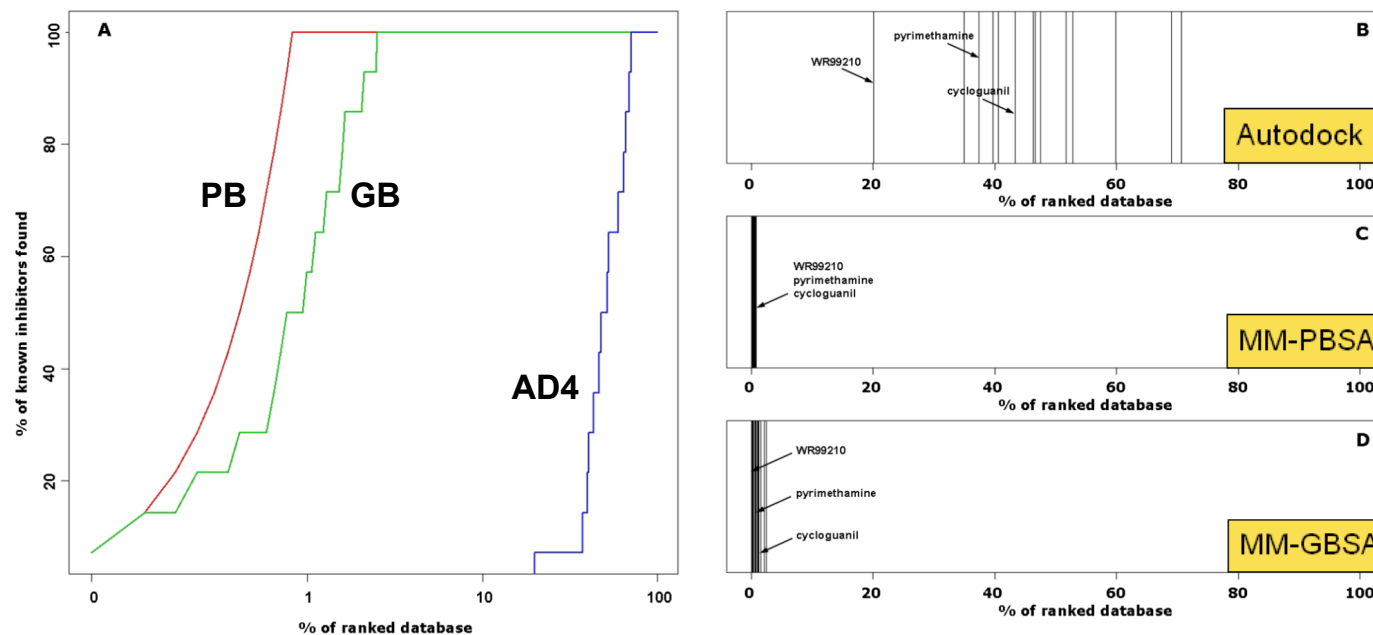
MM-PBSA Binding Free Energy

MM-GBSA Binding Free Energy

Enrichment Factors: DHFR / NCI-div

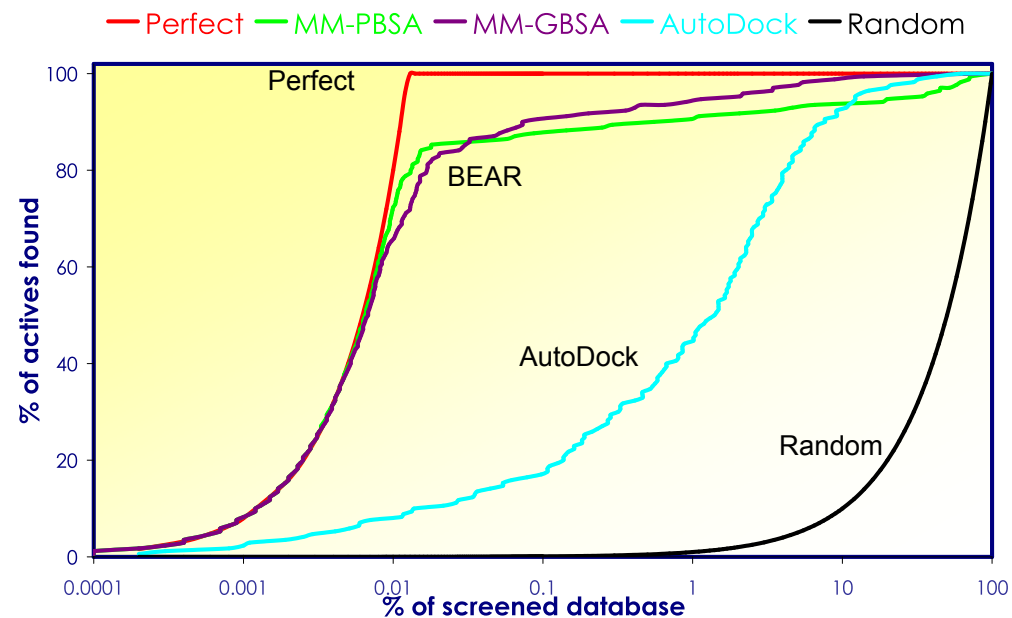
- Target: DHFR (PDB code 1J3I)
- Compounds:
 - NCI diversity set (**1720 compounds**)
 - 14 known inhibitors (**1 known inhibitor/~120 cpds**)

Docking with Autodock 4
BEAR refinement and rescore



Enrichment Factors: DHFR / ZINC

- Target: DHFR
- Compounds:
 - ZINC Database Lead-Like subset (~1,5 million compounds)
 - ~170 known inhibitors (1 known inhibitor/~9000 cpds)



Applications in drug discovery



- Initiative for drug discovery against neglected and emergent diseases
- International collaboration with partners from Europe, Asia and Africa
- Based on virtual screening on computing GRID

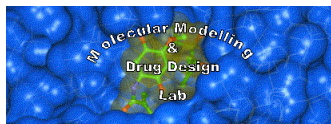
Docking of
~4 million
compounds on
several targets



Molecular docking (FlexX)
~413 CPU years, 1.738 TB data
~100,000 dockings per minute



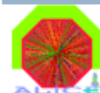
Data challenge on EGEE
~90 days on ~5000 computers



EGEE computing grid



Grid Projects Collaborating in LHC Computing Grid



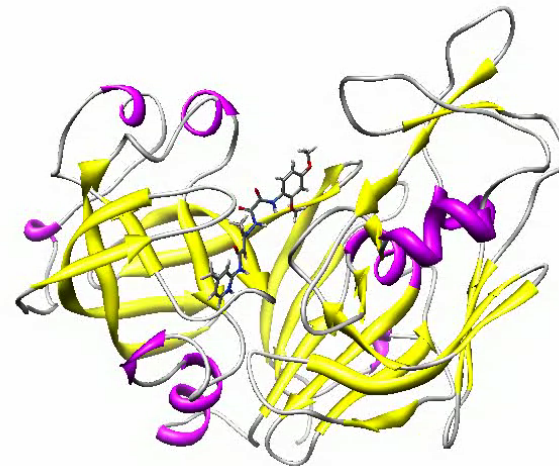
EGEE Operations Information

Active Sites	177
Available CPU	33723
Available Storage (TB)	14170

LastBuild: Sat Mar 17 10:16:01 GMT 2007 GstatQuery: 2006-12-15

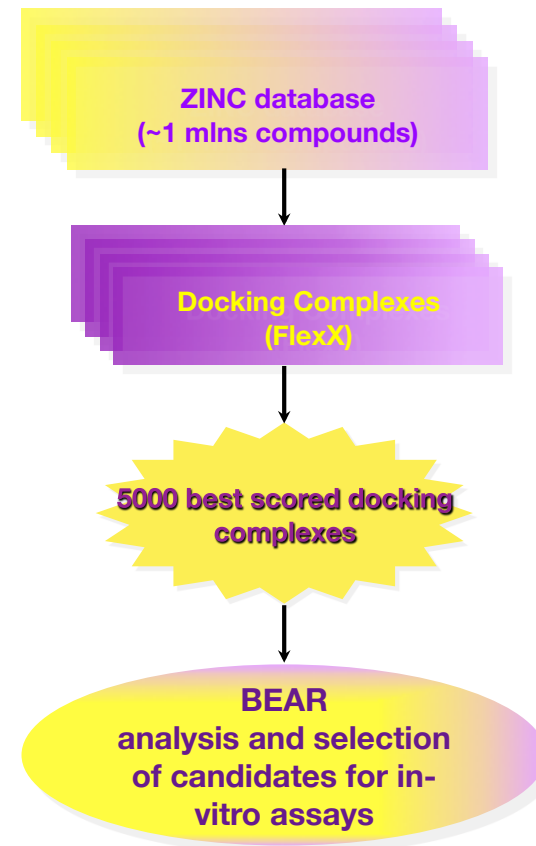
Data challenge on *P. falc* Plasmepsin II

- Plasmodium *falciparum* aspartic protease
- Key enzyme for the parasite metabolism, responsible for the initial cleavage of haemoglobin during the intra-erythrocyte stage of the parasite infection
- WISDOM (Wide in Silico Docking on Malaria) targets



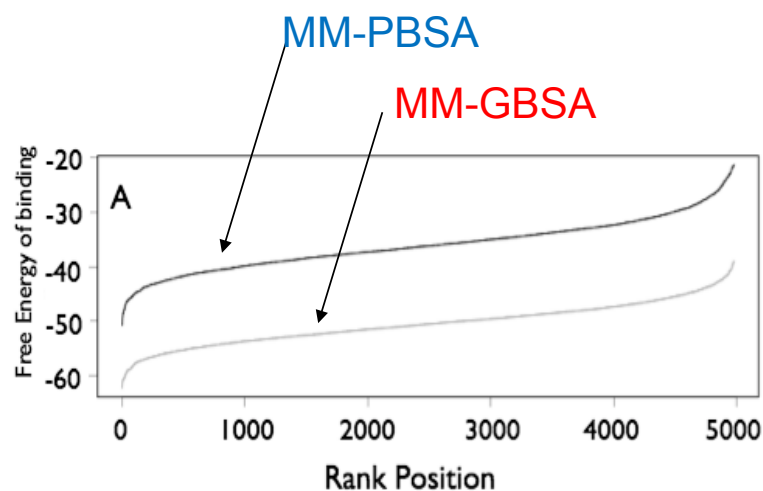
Virtual screening protocol

- Protein structure: Plasmepsin II from PDB
- Ligands: ~1 million cpds from ZINC database
- Docking software: FlexX
- Docking results analysis
- BEAR post-processing and results analysis
- Visual inspection of the complexes
- Compound selection for *in-vitro* assays



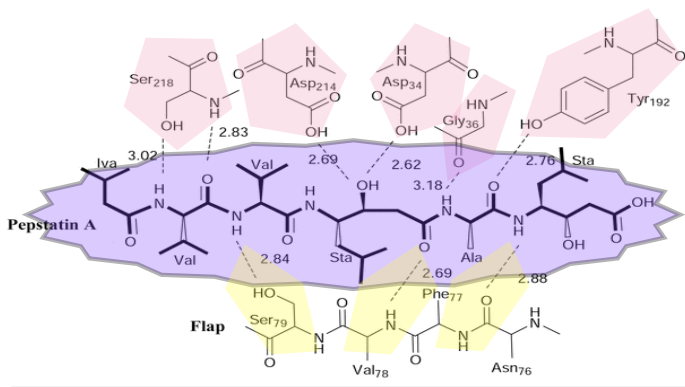
BEAR rankings

Results from BEAR analysis were ranked according to the two scoring functions used by BEAR (MM-PBSA/MM-GBSA)



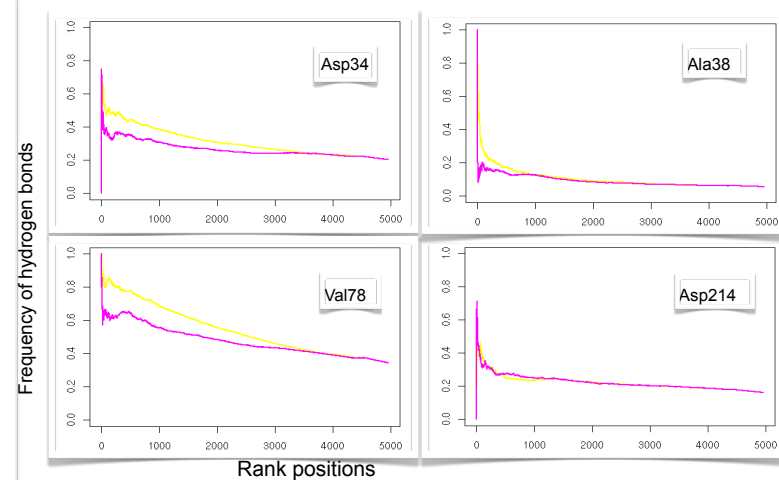
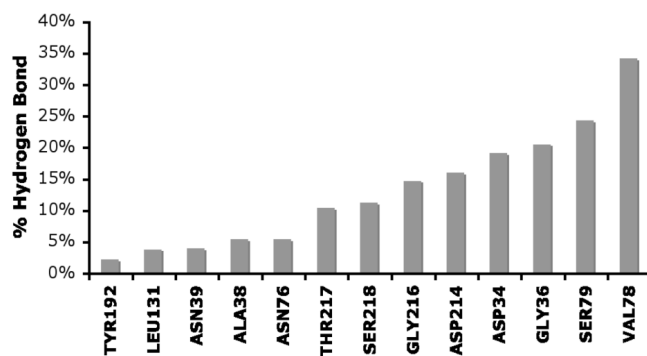
Potent (nM) Plasmeprin inhibitors **Pepstatin A**, **RS367**, **RS370** were retrieved at the first positions of the ranked lists, whereas these compounds ranked several thousand positions downstream in the original docking list

Analysis of ligand-plasmepsin interactions



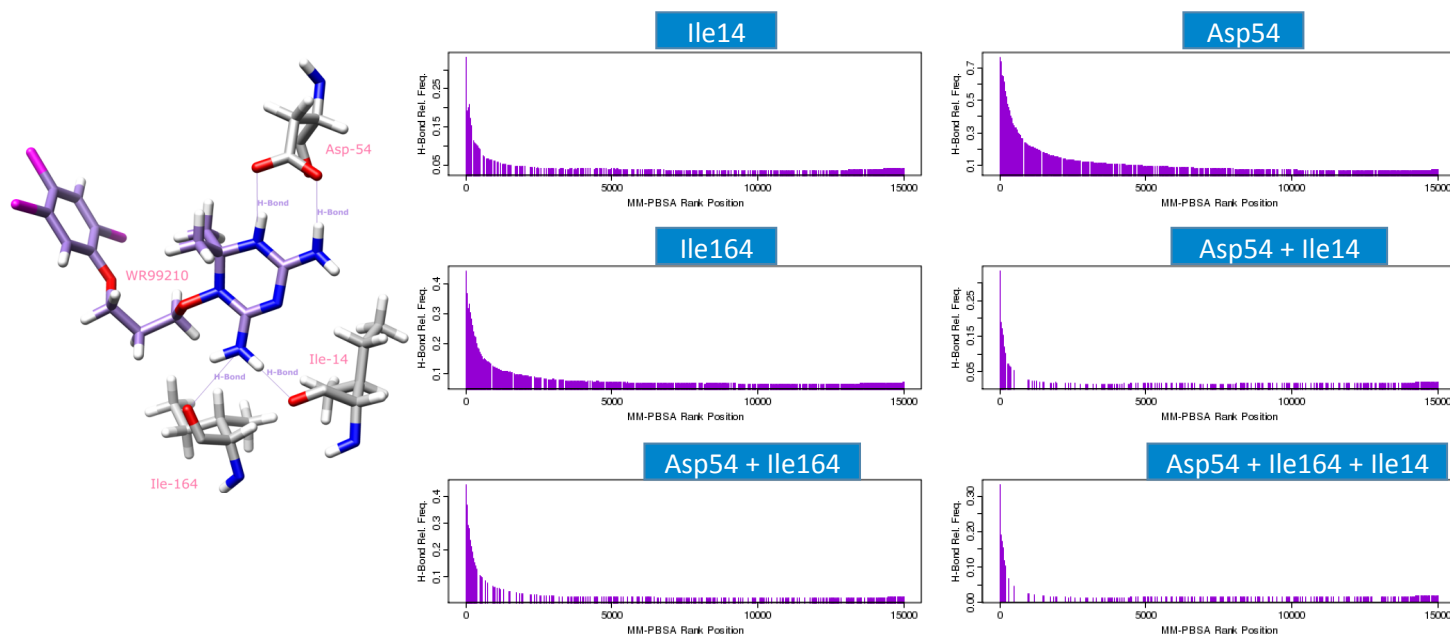
Analysis of the interactions with the PLM active site residues involved in binding of known inhibitors such as Pepstatin A

Best scoring compounds establish key interactions with the protein



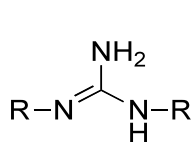
Analysis of ligand-target interactions

- Pf DHFR crystal structure
- 4.3 million cpds (ZINC database)
- Docking with FlexX
- Post-docking with BEAR of **15.000 best compounds**.

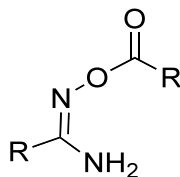


Selection of compounds for testing

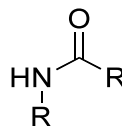
- Selection was made from the **200 best ranking** compounds in both MM-PBSA and MM-GBSA ranked lists
- **Interactions** with active site residues (visualization)
- **Chemical diversity**: selection of compounds that interact with Asp214 and Asp34 with different scaffolds:



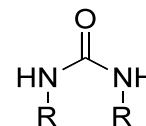
Guanidine



N-alkoxyamidine



Amide



Urea/Thiourea

30 compounds selected for biological assays

Biological assays

- Assay method: FRET analysis
- **Compounds tested: 30**
- **Active compounds identified: 26**
- Inactive compounds: 4
- Range of activity:
4.3 nM - 1.8 μ M
- **HIT RATE ~85%**
- **Two entirely new classes of inhibitors**

Table 3. Measured IC₅₀ values of the thirty tested compounds and of three reference inhibitors used for comparison.

Mol.	IC ₅₀ (nM)	Mol.	IC ₅₀ (nM)	Mol.	IC ₅₀ (nM)
1	305.1 \pm 1.5	12	237.4 \pm 1.5	23	87.5 \pm 0.1
2	5.5 \pm 2.0	13	1087.6 \pm 0.7	24	4.4 \pm 0.8
3	6.4 \pm 0.7	14	9.5 \pm 1.1	25	122.9 \pm 1.1
4	42.6 \pm 1.5	15	96.1 \pm 0.2	26	146.4 \pm 1.0
5	236.4 \pm 0.7	16	30.0 \pm 1.8	27	201.1 \pm 1.3
6	145.2 \pm 2.4	17	n.i.	28	7.6 \pm 1.1
7	4.3 \pm 0.6	18	187.1 \pm 3.1	29	1831.3 \pm 1.9
8	62.1 \pm 0.6	19	n.i.	30	38.9 \pm 2.4
9	118.1 \pm 1.9	20	189.0 \pm 1.4	RS367	18 ^[a]
10	8.8 \pm 0.8	21	57.3 \pm 0.4	RS370	30 ^[a]
11	n.i.	22	n.i.	Pep.A	4.3 \pm 0.9

[a] IC₅₀ values taken from Silva et al.^[18]

n.i.: no inhibition

