Facultade de Química
Universidade de Vigo

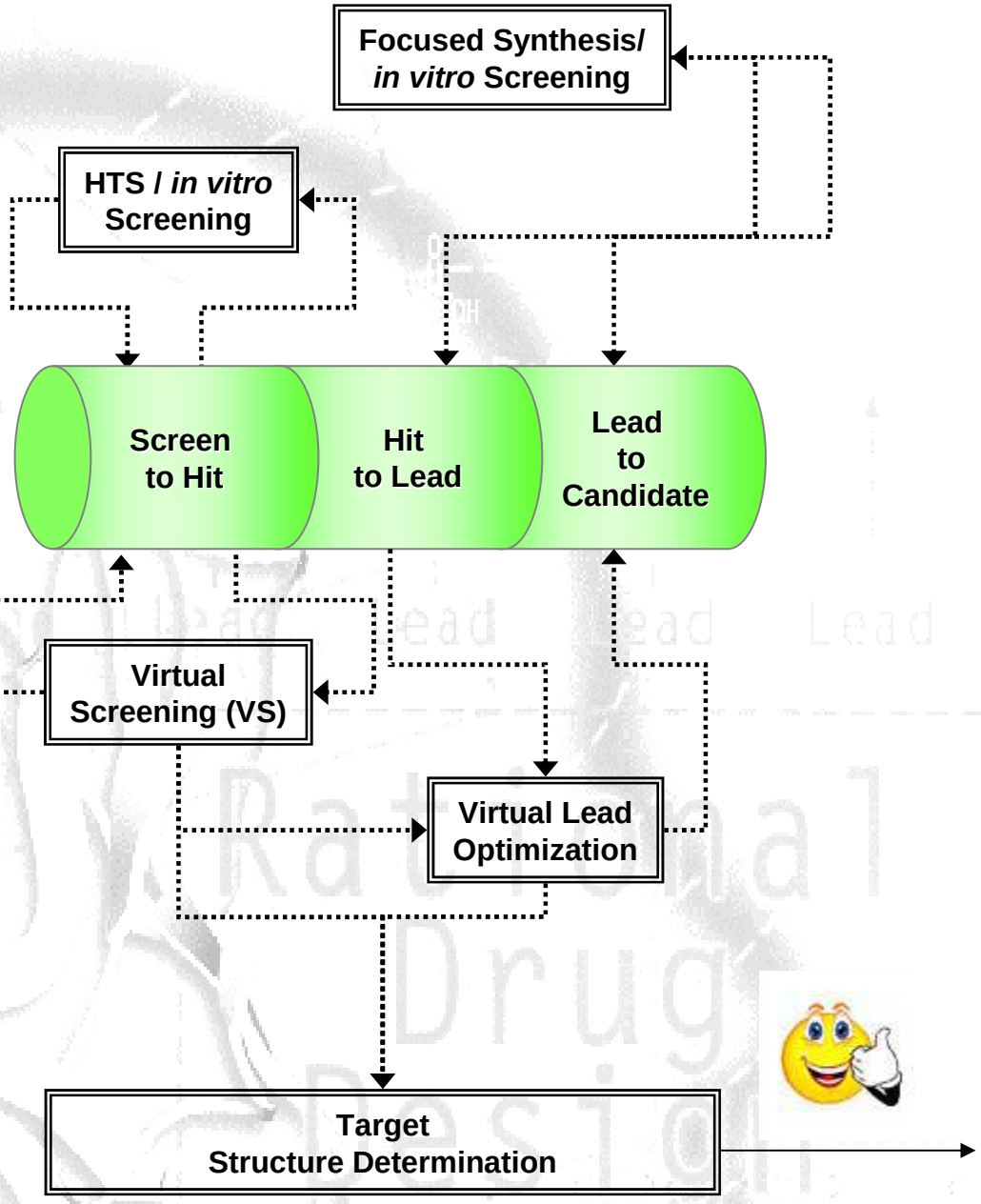
"Computational approaches in drug discovery: expectations and reality."

Stefano Moro

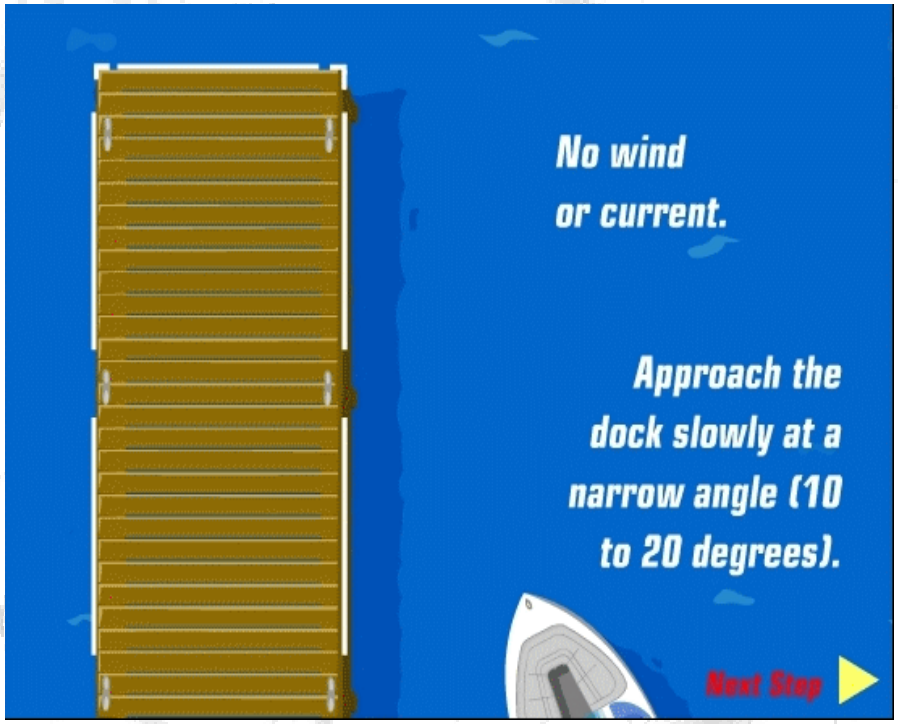
Molecular Modeling Section (MMS)

Department of Pharmaceutical and Pharmacological Sciences

University of Padova

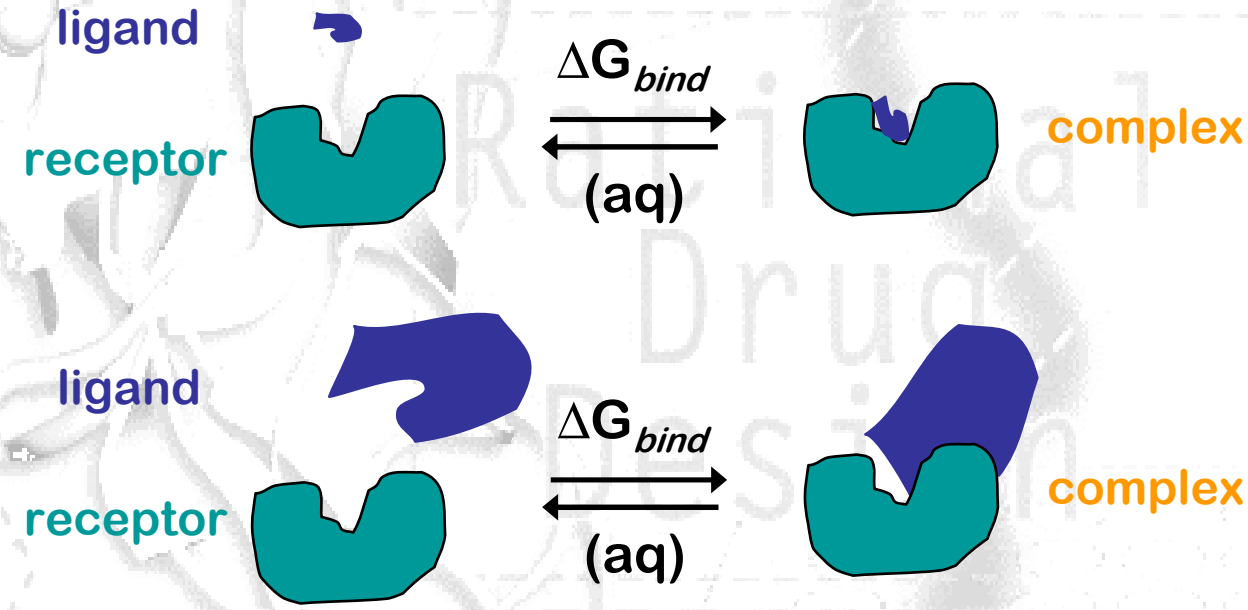

Focused Synthesis/
in vitro Screening
HTS / *in vitro*
Screening
Screen
to Hit
Hit
to Lead
Lead
to
Candidate
Virtual
Screening (VS)
Virtual Lead
Optimization
Target
Structure Determination

Docking & Scoring



What does it mean Docking?

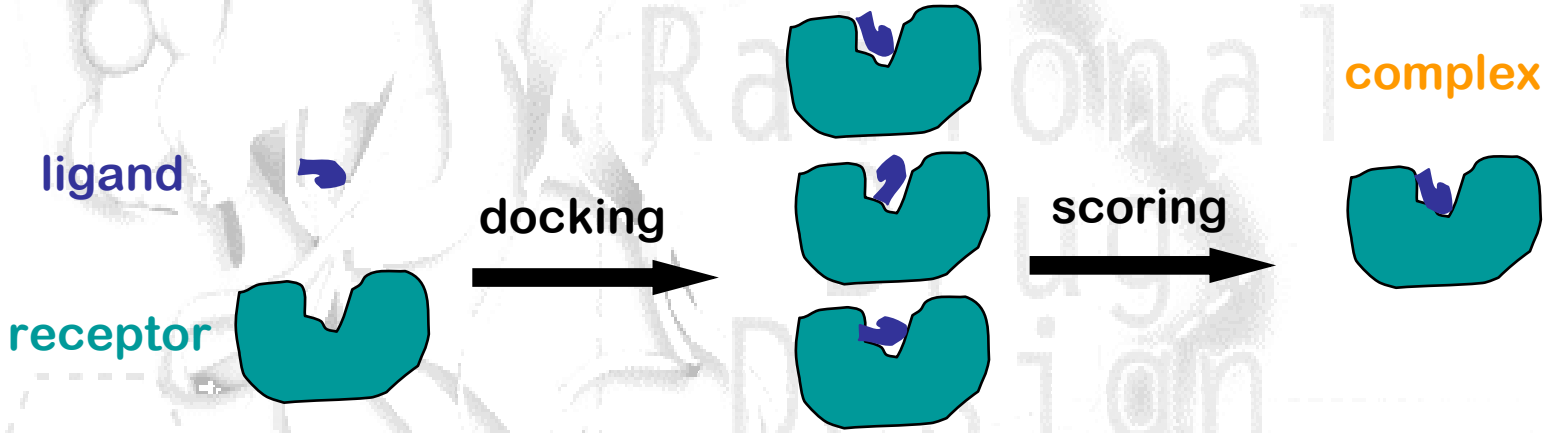
Generally speaking, any computational strategy that use 3D information about the "receptor" to predict *binding modes* and *affinities* for different ligands.



The molecular docking problem:

To place a ligand (small molecule) into the binding site of a receptor in the manners appropriate for optimal interactions with a receptor (DOCKING).

To evaluate the ligand-receptor interactions in a way that may discriminate the experimentally observed mode from others and estimate the binding affinity (SCORING).



predict (x,y,z) etc

predict "ΔG_{bind}"

Molecular Docking: basic principles

- The association of molecules is based on interactions
  - *H-bonds, salt bridges, hydrophobic contacts*
  - *Electrostatic*
  - *Very strong repulsive (vdW) interactions on short distances.*
- The associate interactions are weak and short range.
  - *Strong binding implies surface complementarity.*
- Most molecules are flexible.
- The binding affinity is the energetic difference to the uncomplexed state.
  - *The surrounding medium (usually water) plays an important role.*
  - *Entropy has a significant impact on binding.*
- The binding affinity describes an ensemble of complex structures, not a single one.
  - *Tight binders often have a dominating binding mode.*
  - *What about weak binders?*

The three crucial components of virtualization process:

pre- and/or during docking:

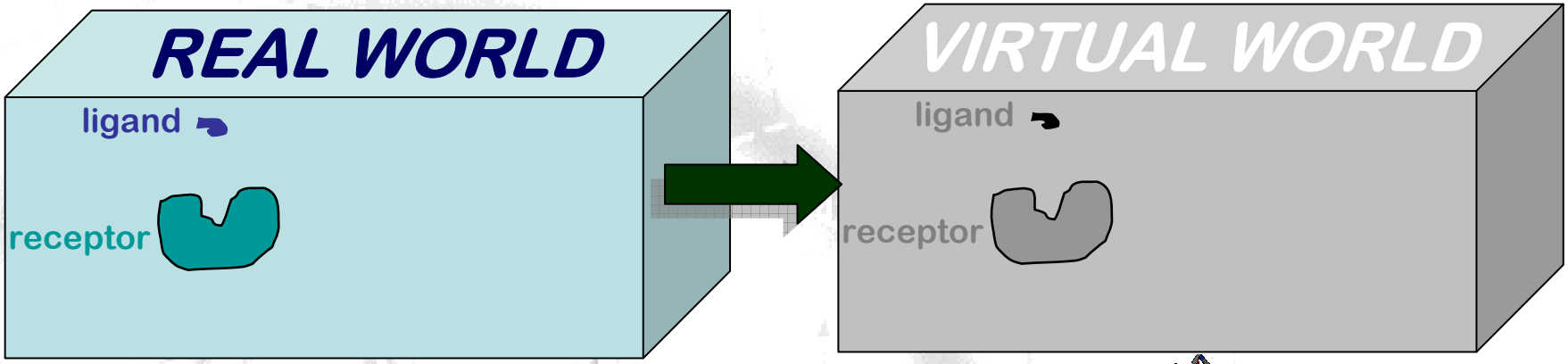
Representation of receptor binding site and ligand

during docking:

Sampling of configuration space of the ligand-receptor complex

during docking and scoring:

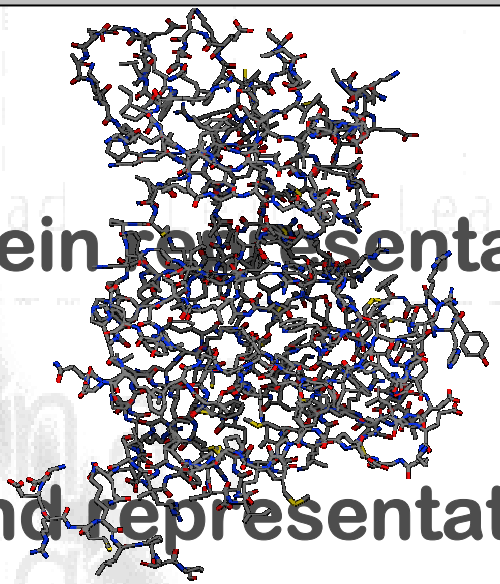
Evaluation of ligand-receptor interactions



Real protein

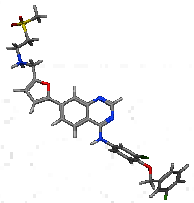
Epithelial Grow Factor Receptor (kinase domain)

3D protein representation



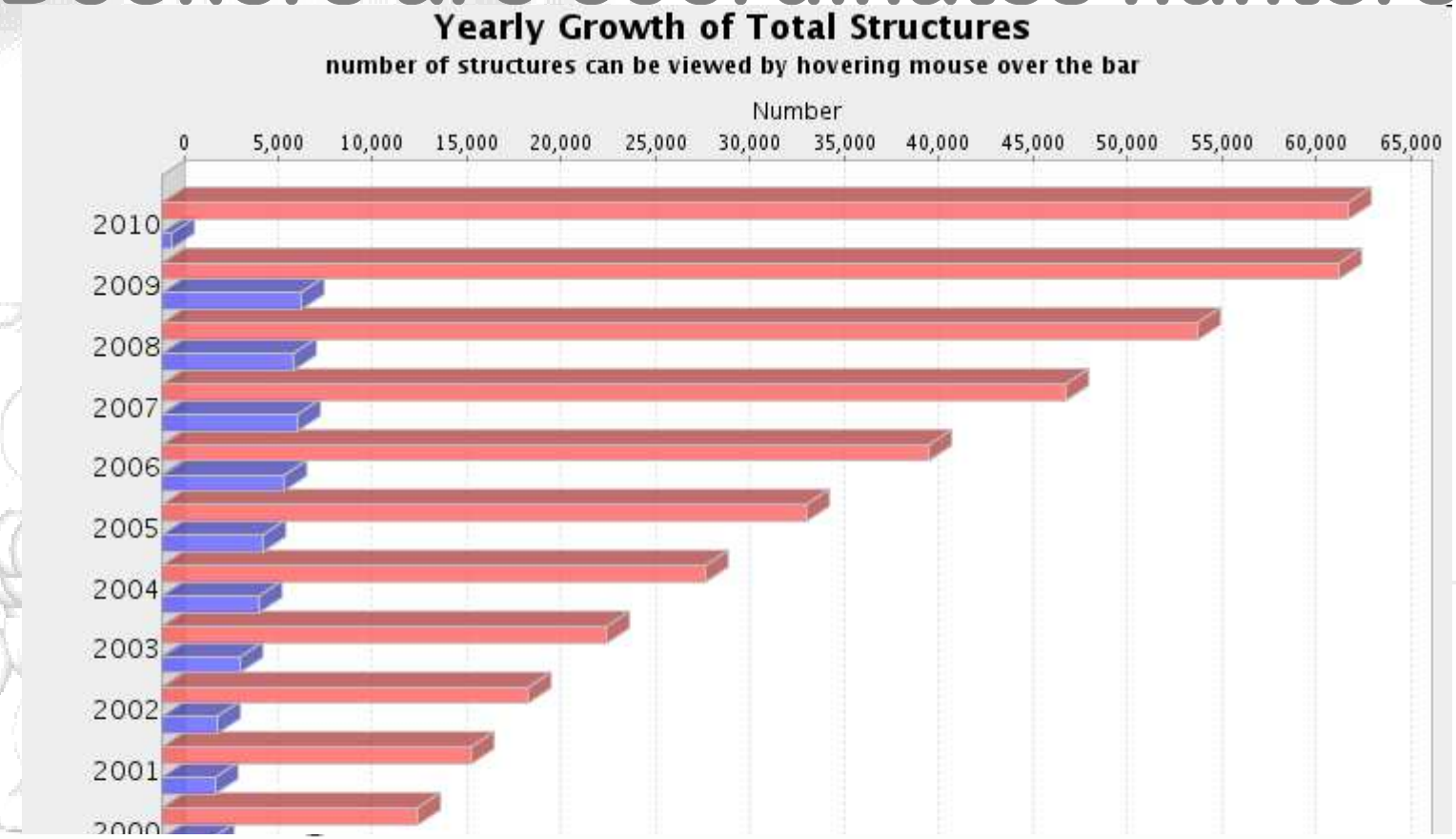
Real ligand

3D ligand representation



Lapatinib (Tykerb®, GSK)

Dockers are coordinates hunters!



Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	50741	1188	2328	17	54274
NMR	7175	887	152	7	8221
ELECTRON MICROSCOPY	182	17	70	0	269
HYBRID	18	1	1	1	21
other	120	4	4	13	141
Total	58236	2097	2555	38	62926

There are coordinates.... and coordinates!

PDB info must be carefully checked:

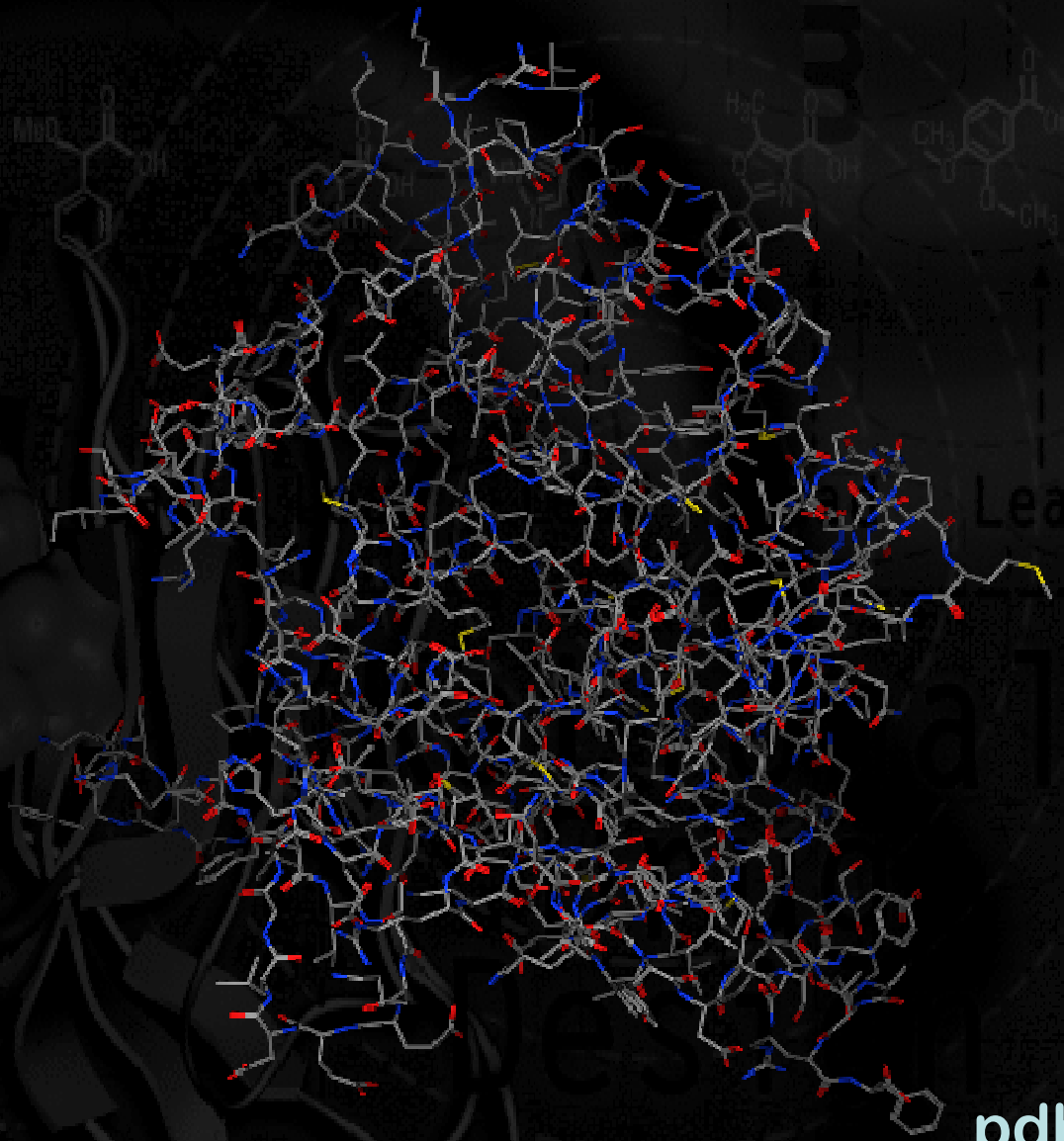
- **Resolution (preferable <2.5 Å);**
- **Missing residues;**
- **Mapping B-factors (flexible regions);**
- **Co-crystallizations (dimers or oligomers, co-factors, salts, crucial water molecules etc…)**
- **Inspection, comparison and selection of the most valuable structure if more than one is available;**

Original PDB must be "manipulated" before docking!

Several computational adaptations must be performed:

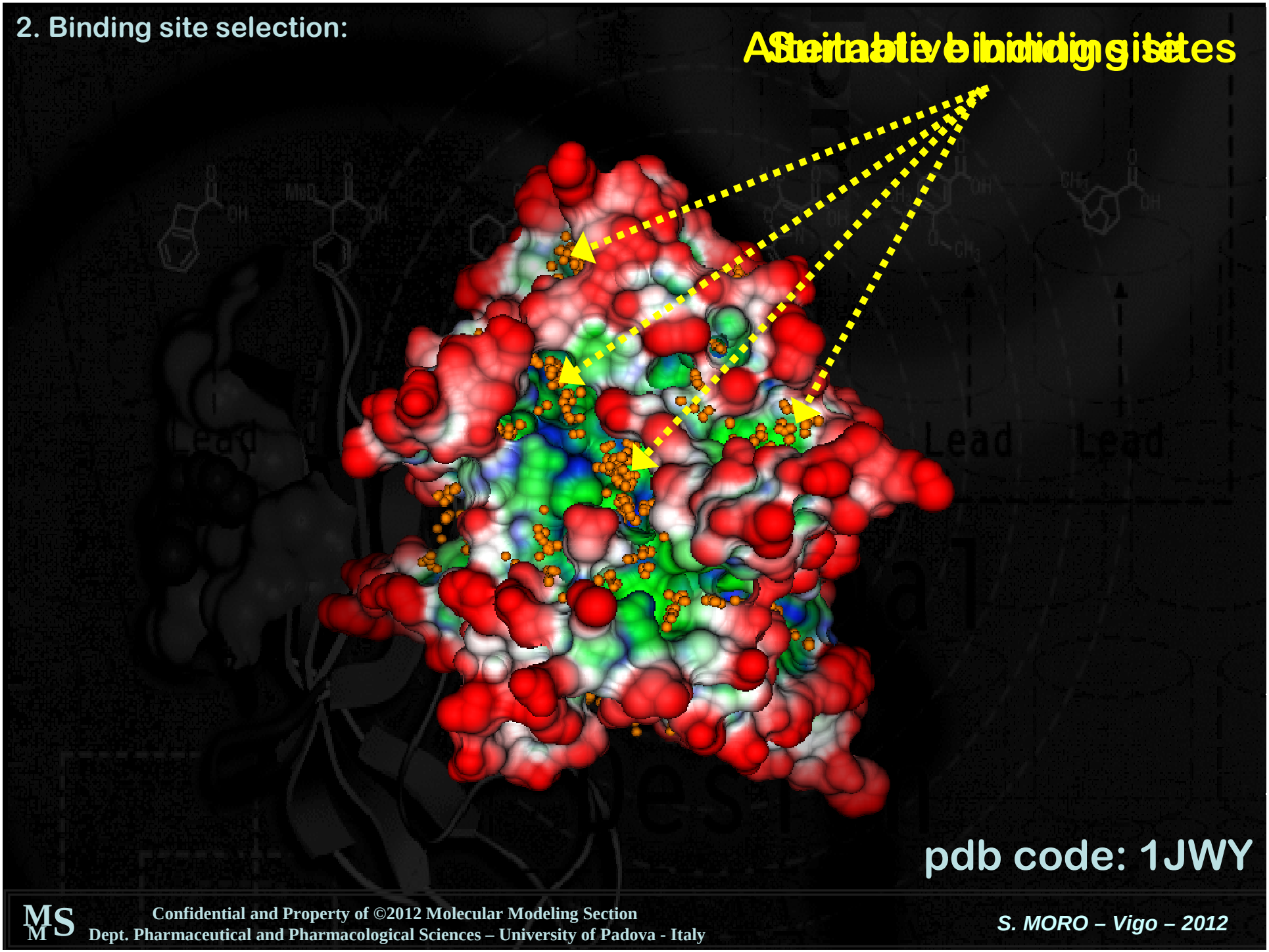
- H-adding and optimization (hydrogen atoms are naturally missing in any crystal structure);
- Check and fix the ionization form of all protonable side chains;
- Fix missing residues if necessary;
- Remove only the unnecessary co-crystallized partners (be careful!!!);
- Inspection of the possible binding sites and the eventually available binding motif;

Where a ligand can dock this target?



pdb code: 1JWY

2. Binding site selection:

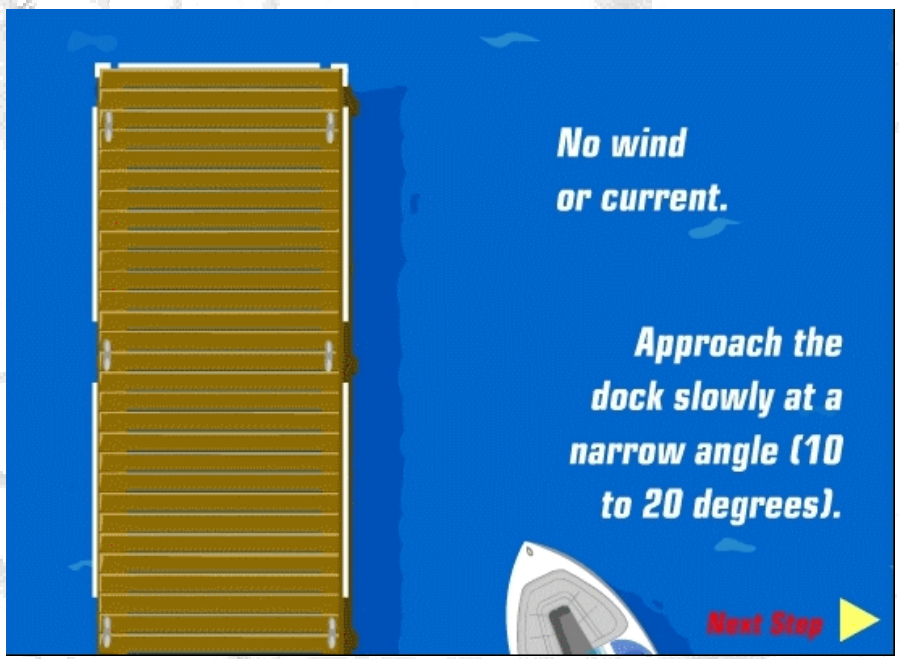


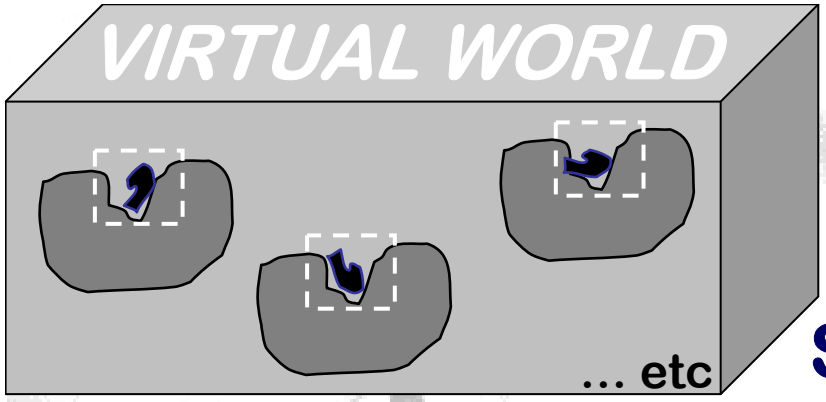
pre- and/or during docking:

Representation of receptor binding site and ligand

during docking:

Sampling of configuration space of the ligand-receptor complex

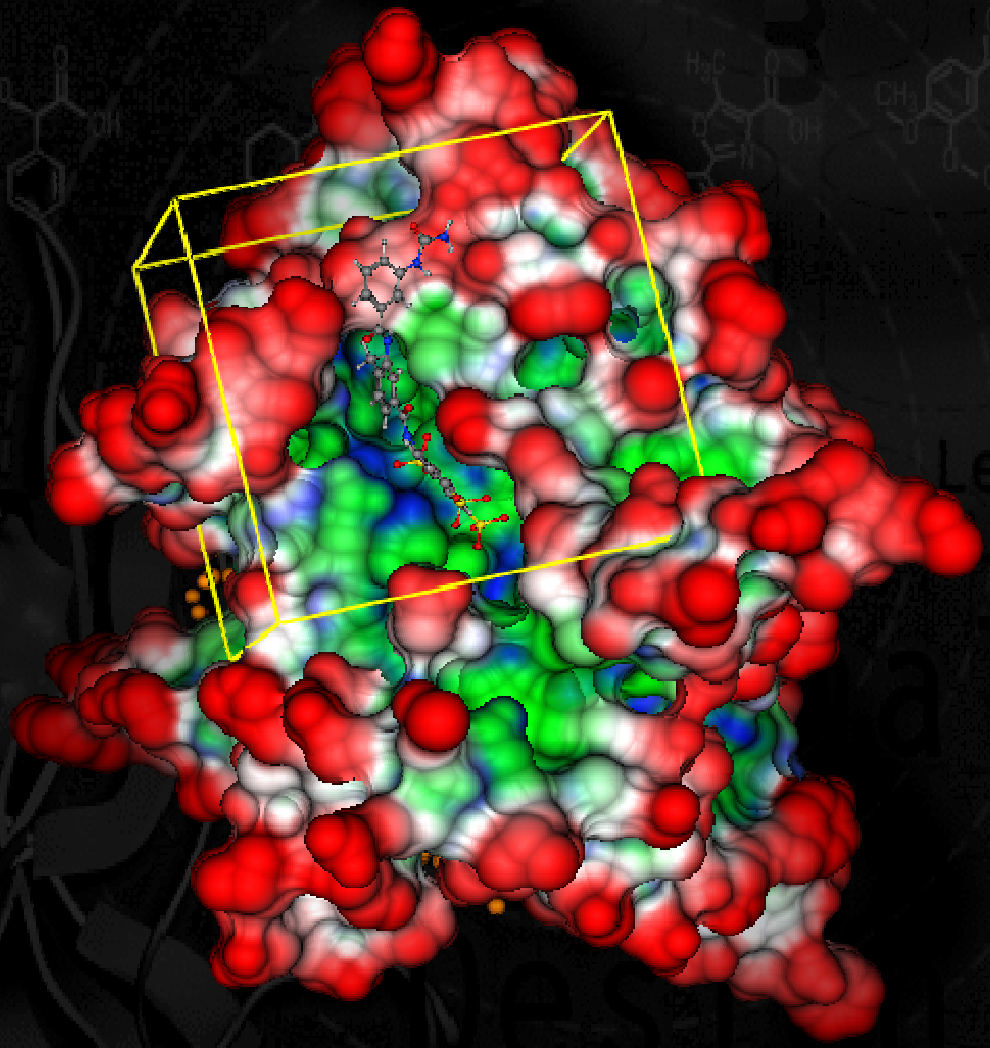




Some definitions:

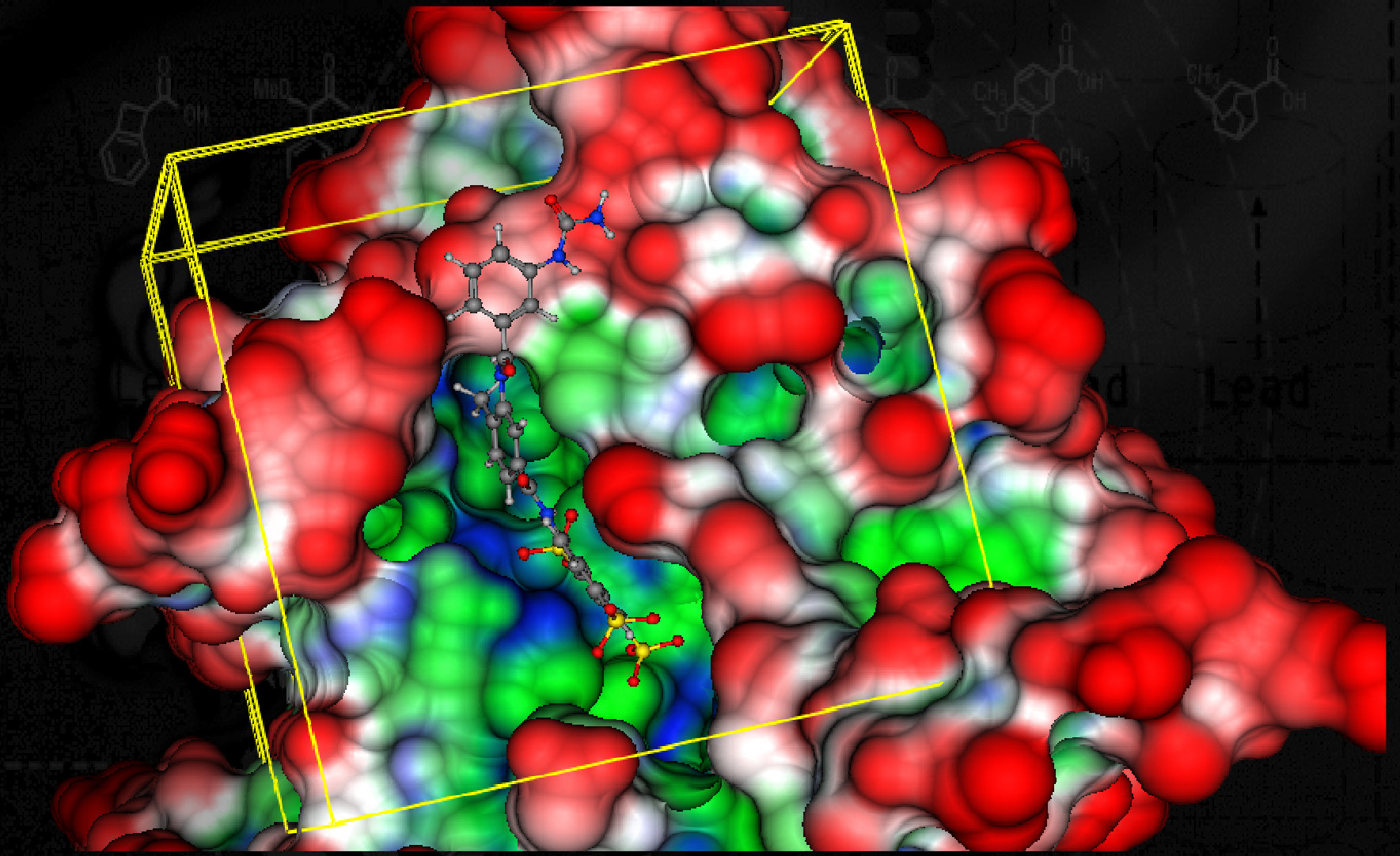
	Translations (3N) Rotations (3N)	Conf. LIG	Conf. REC
RIGID			
SEMI-FLEXIBLE			
FLEXIBLE			side-chains close to LIG

3. Docking box selection:



pdb code: 1JWY

4. Conformational (pose) sampling :

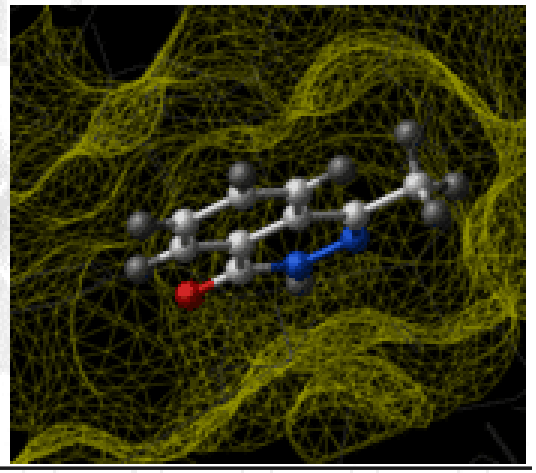


pdb code: 1JWY

Docking algorithms:

The best of these predict the experimental (X-ray!!) pose about 70% of the time, although selecting the program that will give the best results for any given target is not straightforward.

Clearly, good poses can be produced, but how does one pick the program that will do so reliably for the targets of interest?



pre- and/or during docking:

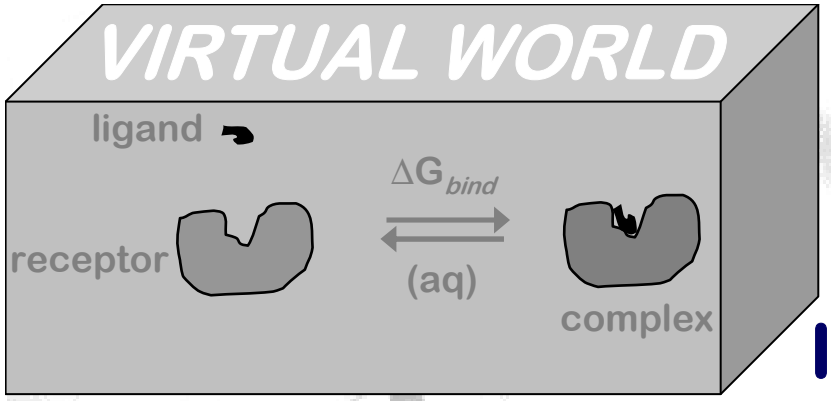
Representation of receptor binding site and ligand

during docking:

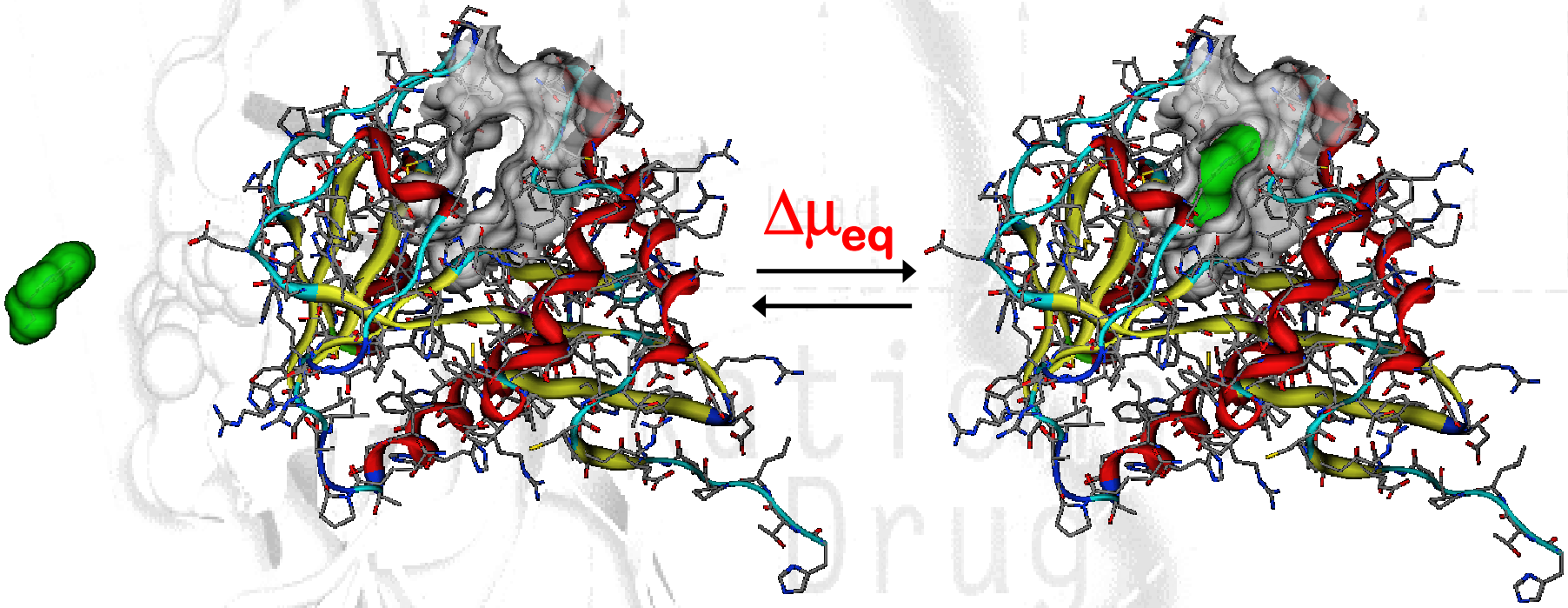
Sampling of configuration space of the ligand-receptor complex

during docking and scoring:

Evaluation of ligand-receptor interactions



In a time-independent contest

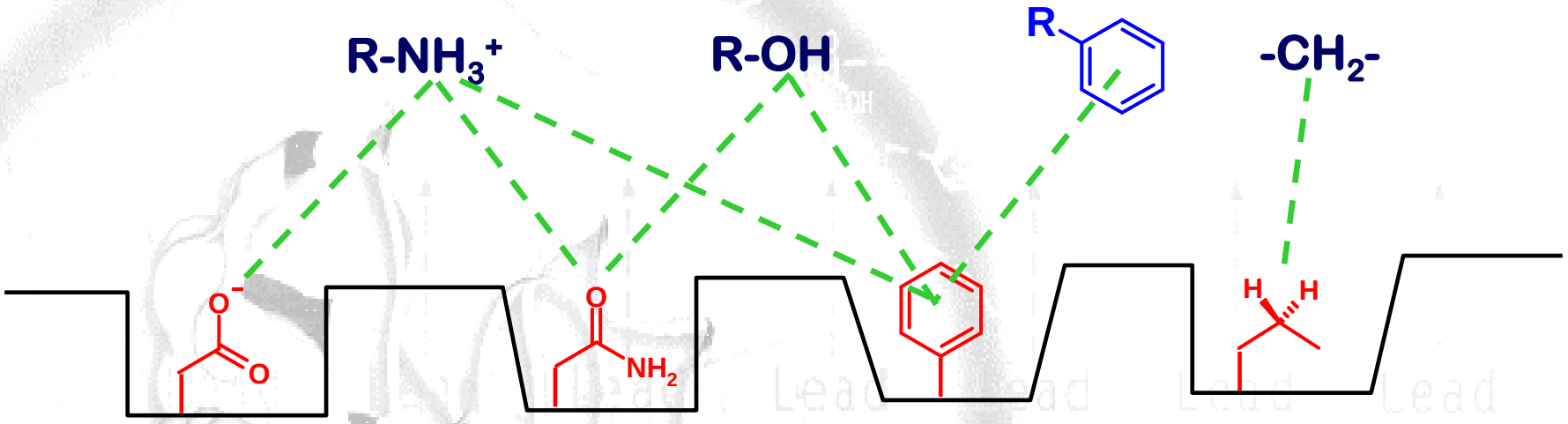


$$\Delta\mu_{eq} \cong \Delta Ep_{eq}$$

Types of scoring functions:

- **Force field based**: nonbonded interaction terms as the score, sometimes in combination with solvation terms
- **Empirical**: multivariate regression methods to fit coefficients of physically motivated structural functions by using a training set of ligand-receptor complexes with measured binding affinity
- **Knowledge-based**: statistical atom pair potentials derived from structural databases as the score
- *Other: scores and/or filters based on chemical properties, pharmacophore, contact, shape complementary*
- **Consensus** scoring functions approach

Force Field Based Scoring Functions



		(kcal/mol)
Interazione carica-carica (*legame ionico*):	$-\Delta G^0 \cong$	5 ÷ 10
Interazione carica-dipolo:	$-\Delta G^0 \cong$	1 ÷ 7
Interazione carica-π:	$-\Delta G^0 \cong$	8 ÷ 10
Interazione dipolo-dipolo forte (*legame idrogeno*):	$-\Delta G^0 \cong$	1 ÷ 7
Interazione mediata dal trasferimento di carica:	$-\Delta G^0 \cong$	1 ÷ 6
Interazione mediata dall'interazione π-π:	$-\Delta G^0 \cong$	1 ÷ 2
Interazione dipolo-dipolo debole (van der Waals):	$-\Delta G^0 \cong$	0.5 ÷ 1

Force Field Based Scoring Functions

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left(\frac{A_{ij}}{r_{ij}^{a}} - \frac{B_{ij}}{r_{ij}^{b}} + 332 \frac{q_i q_j}{Dr_{ij}} \right)$$

e.g. AMBER FF in DOCK

- **Advantages**
  - FF terms are well studied and have some physical basis
  - Transferable, and fast when used on a pre-computed grid
- **Disadvantages**
  - Only parts of the relevant energies, i.e., potential energies & sometimes enhanced by solvation or entropy terms
  - Electrostatics often overestimated, leading to systematic problems in ranking complexes

Empirical Scoring Funcion:

- **Use MLR to fit coefficients to a set of physically motivated terms in order to reproduce the experimental binding affinity of a training set of known protein-ligand complexes.**

- **Data: A set of protein-ligand complexes with known 3D structures and binding affinities (ΔG).**

$$\Delta G = \Delta G_0 + \Delta G_{rot} N_{rot} + \Delta G_{hb} \sum_{neutral\_Hbonds} f(\Delta R, \Delta\alpha)$$
$$+ \Delta G_{io} \sum_{ionic\_\mathrm{int}} f(\Delta R, \Delta\alpha) + \Delta G_{aro} \sum_{aro\_\mathrm{int}} f(\Delta R, \Delta\alpha)$$
$$+ \Delta G_{lipo} \mid A_{lipo} \mid$$

- **ΔG_0: Lost of transformation entropy.**
- **Δg_{rot}:Lost of conformational DOF (ligand entropy).**
- **N_{rot}: Number of rotatable bonds immobilized during complex formation.**
- **ΔG_{hb}, Δg_{io}: Hydrogen bonds (neutral, charged).**
- **ΔG_{aro}: Interaction between aromatic groups.**
- **ΔG_{lipo}: Accounts for lipophilic interactions.**
- **A_{lipo}: Receptor-ligand lipophilic contact surface area.**

	GSF (Gaussian Scoring Function)	PLP	ChemGauss	ChemScore	ZapBind
Shape	Favors heavy atoms contacts	Piecewise Linear Potential Interactions	Uses GSF	Favors Lipophilic interactions	Surface Area Contact Term
Metals		Favorable interaction with HB acceptors	Favorable interaction with HB acceptors	Favorable interaction with HB acceptors	Poisson-Boltzman electrostatics
H-Bonds / Electrostatics		HB based on heavy atom distance.	HB based on Lone-Pair and Polar hydrogen positions.	HB interactions with angle and distance constraints	Poisson-Boltzman electrostatics
PI-Stacking			Favorable interactions between PI ring of aromatics		
Entropic				Penalty for frozen rotatable bonds	
Desolvation		Penalty for HB atoms near non-polar atoms			Poisson-Boltzman electrostatics

Consensus Scoring

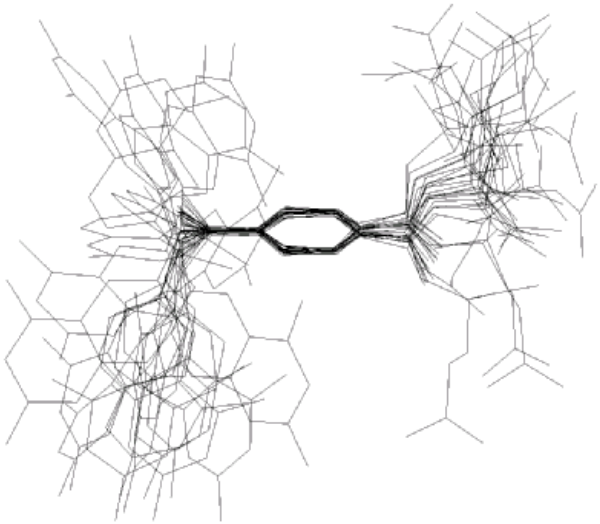
- Too many scoring functions, none prevails in terms of predictivity;
- ***Combined approach***: using in parallel several scoring functions and prioritize only the most converging selected candidates from the top x% (*e.g.* 5-10%)

Reduce false positives!

	SF_1	SF_2	SF_3	SF_4	SF_5
1	☺				
2		☺		☺	
3	☺			☺	☺
n			☺		

Scoring functions: drama…

"Windows of activity" is very small: the ΔG_{bind} difference between the best ligand that one might reasonably expect to identify using virtual screening (potency, ~ 50 nM) and the experimental detection limit (potency, ~ 100 μM) is only 4.5 kcal/mol.



The ΔG contributions due to conformational factors alone for typical druglike ligands (which are usually neglected in most scoring functions) can be as large as this!

Scoring functions:

we can consider any scoring function a mathematical tool to evaluate the "quality" of ligand-receptor coordinates… nothing more, unfortunately!

GRAZIE
PER LA PAZIENZA

Outline:

Molecular Docking & Scoring;

Docking Applications:

- *Virtual Libraries;*
- *High-Throughput Virtual Screening (HTVS);*

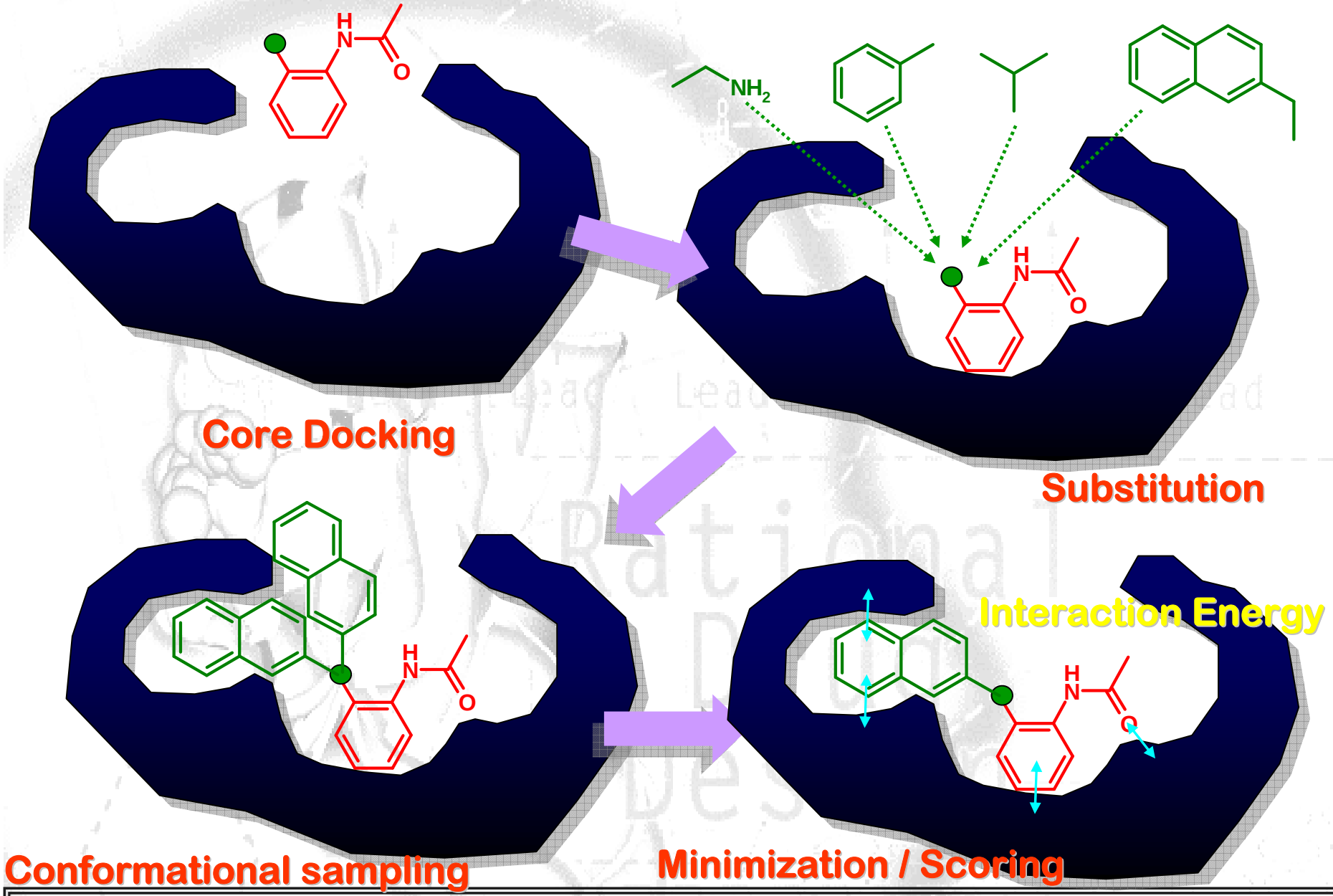
Docking Applications

- Determine the lowest free energy structures for the receptor-ligand complex
- Search database and rank hits for lead generation
- Calculate the differential binding of a ligand to two different macromolecular receptors
- Study the geometry of a particular complex
- Propose modification of a lead molecules to optimize potency or other properties
- de novo design for lead generation
- Library design

Docking of Combinatorial Libraries

- **Combinatorial docking problem:** given a library of ligands, calculate the docking score (and the geometry of the complex) for each molecules of the library
- **R-group selection problem:** given a library, select molecules for the individual R-groups in order to form a smaller sublibrary with an enriched number of hits
- ***de novo* library design:** given a catalog of available reagents, design a library (incl. The rules of synthesis) that will optimize the number of hits
- **The incremental construction method:** PRO_SELECT, CombiDOCK (Sun, Ewing et al. 1998), FlexXc
- **Docking of the fully enumerated library followed by plate optimization or cherry-picking**

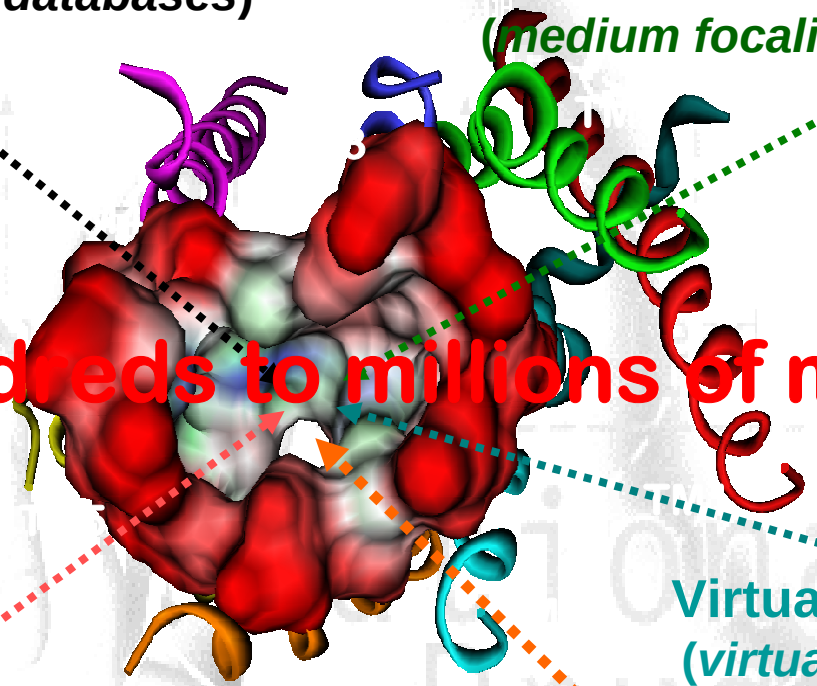
Protein based design of combinatorial libraries



Scouting chemical space:

Random Virtual Screening
(*large known chemical databases*)

Focalized Virtual Screening
(*medium focalized chemical databases*)



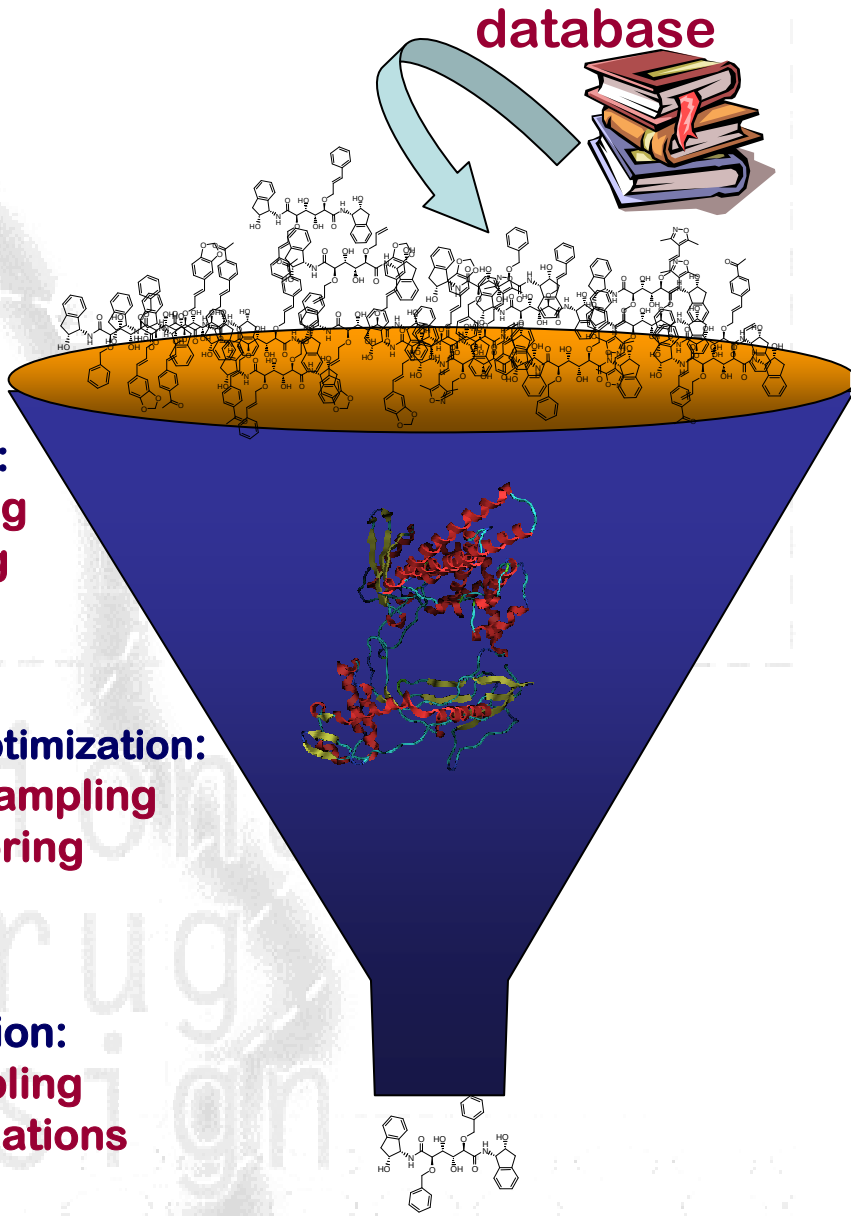
From hundreds to millions of molecules.

Virtual Libraries Screening
(*virtual chemical databases*)

Target-based
De Novo Design

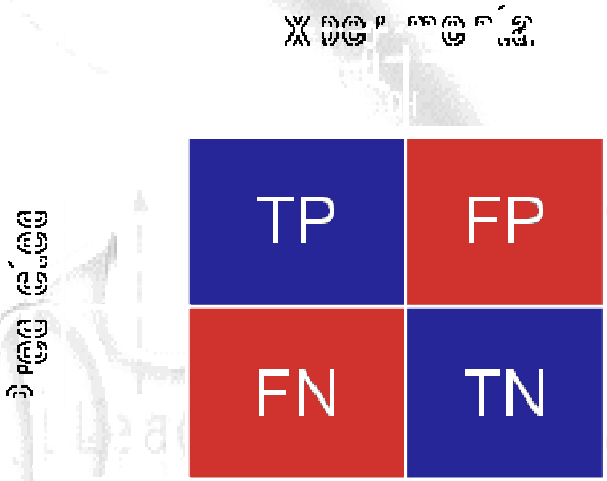
Classical MedChem Screening

Speed *versus* accuracy:



# compounds	speed	accuracy	methods
~10^6			virtual screening: automated docking empirical scoring
~10^3			lead identification/optimization: conformational sampling forcefield scoring
~10^1			final optimization: extensive sampling free energy simulations

False Positive and False negative Rates



- A database contains 100000 compounds with 20 good binders.
- With a false positive rate of 1% we will get 1000 false positives.
- In order to achieve a 90% chance of finding at least one true hit the false positive rate must be 0.2%

Virtual screening requires intensive computing, of the order of few TFlops during one day to compute 1 million docking poses or 1000 ligands on one target protein. So...

Grid computing opens new perspectives for virtual screening as it gives access to very large computing resources. But...

Lessons from the WISDOM:

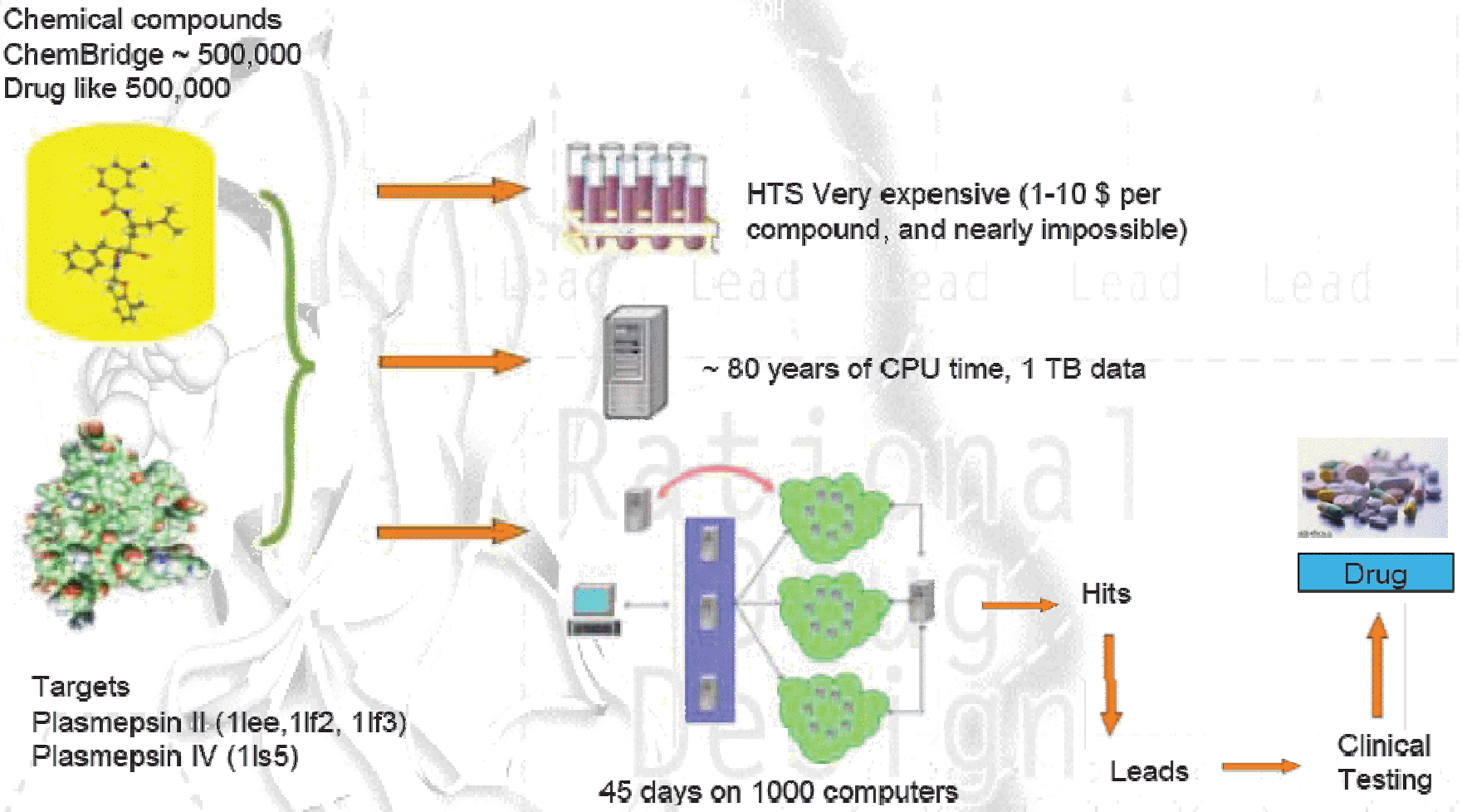
(Word-wide *In Silico* Docking On Malaria)

This first large scale docking experiment ran on the EGEE grid production service from the 11 July 2005 until 19 August 2005.

The biological goal was to propose new inhibitors for a family of aspartic protease (*plasmepsins*) produce by *Plasmudium falciparum*. Plasmepsins are responsible for the initial cleavage of human hemoglobin and later follow by other proteases.

Lessons from the WISDOM:

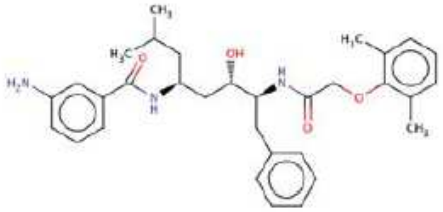
Drug Discovery Work Flow

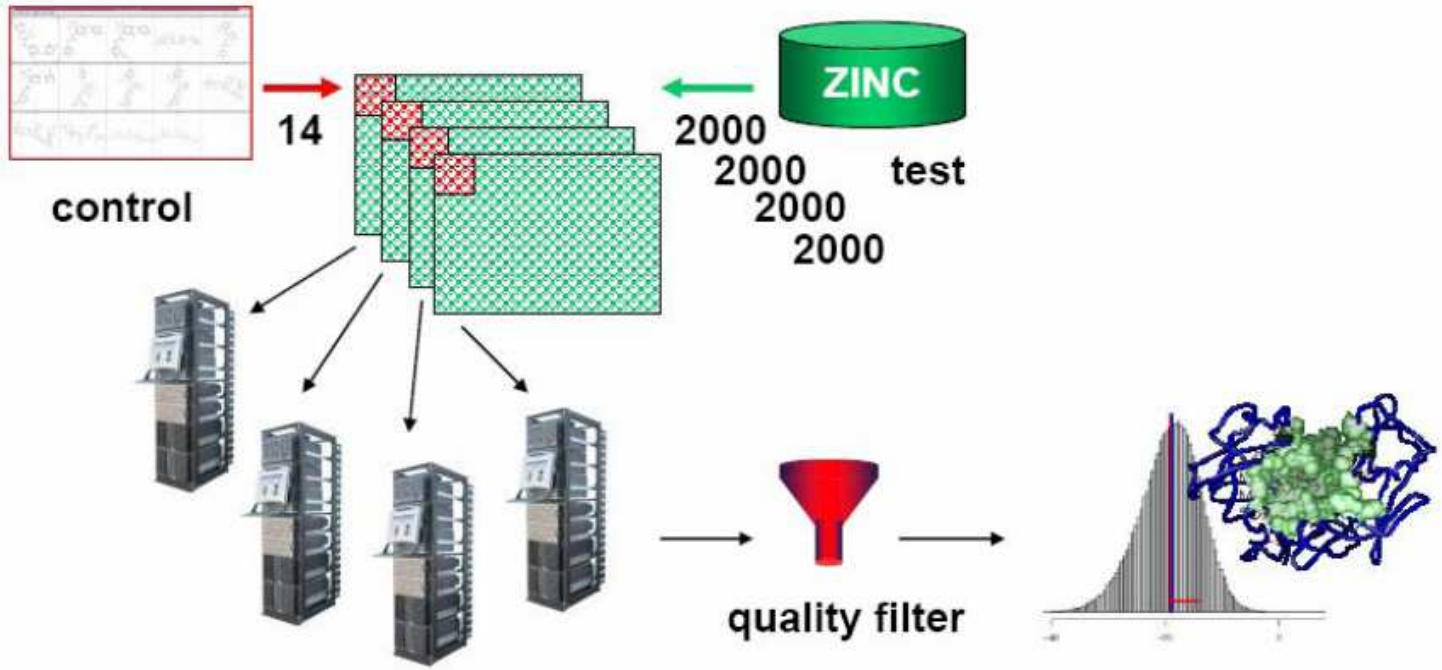


Lessons from the WISDOM:

Testing under control condition

- Known ligands
  - 5 co-crystalized ligands
  - 9 ligands from literature

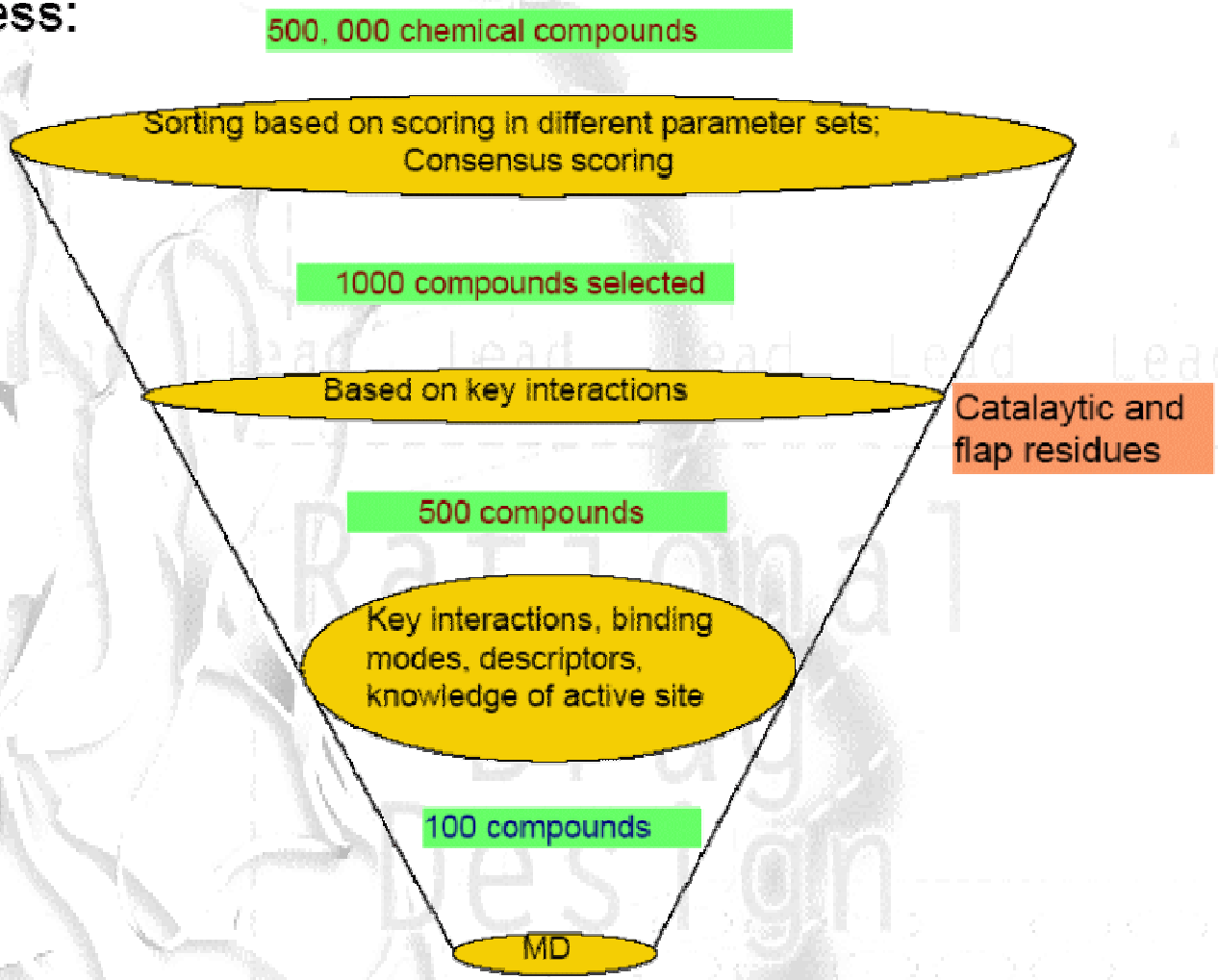




Lessons from the WISDOM:

Identification of novel compounds

Filtering process:



Summary of the WISDOM activity:

Total number of completed dockings	2×10^6
Estimated duration on 1 CPU[a]	88.3 years
Duration of the experience	6 weeks
Cumulative number of Grid jobs	54,000
Maximum number of concurrent CPUs	2000
Size of data	~ 1TB
Overall speedup	767.37
Distribution efficiency[b]	38.4%
Number of countries	17

[a] The measurement was done on a PC with one Xeon 2.8 GHz CPU and 2GB RAM;
[b] Distribution efficiency here is approximated as the ration between the overall speedup and the maximum number of concurrently CPUs.

Critical issues:

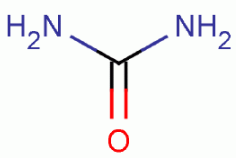
The overall grid efficiency was as average lower than 50%. This means that a large fraction of the job had to be resubmitted. This generated a significant extra workload on the user.

About 1TB of data were produced by the 60000 jobs submitted. Collection, registration and backup of these output data turned out to be a heavy task.

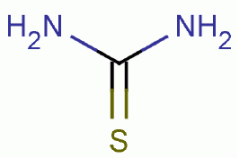
Post-processing of the huge amount of data generated was incredibly demanding task as millions of docking scores had to be compared.

Lessons from the WISDOM:

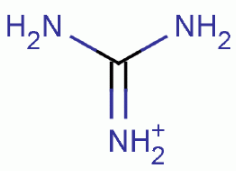
Identification of novel compounds



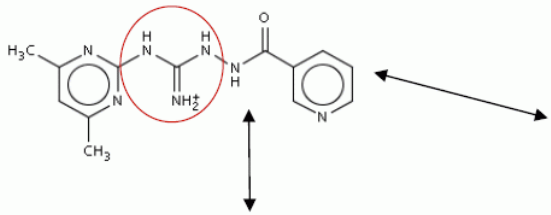
Urea compounds



Thiourea compounds



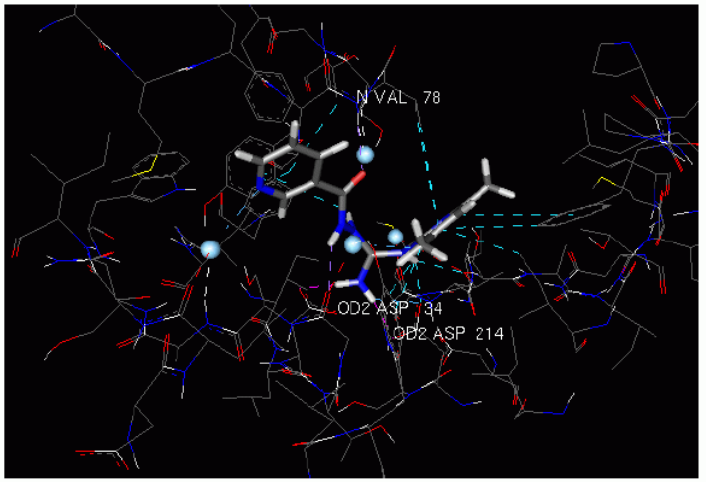
Guanidino compounds



Terminal

File Edit View Terminal Tabs Help

No.	Lig. Atom	Lig. ANo.	Ligand IA-Type	Rec. Aton	Rec. AA	Rec. Chain	Rec. AANo	Receptor IA-Type
1	N1	5	h_acc	water			58	h_don
1	N7	19	h_acc	water			39	h_don
1	N7	19	phenyl_center	C	TYR	A	77	amide
1	C7	13	amide	CG	TYR	A	77	phenyl_center
1	C13	21	ch3_phe	CG	TYR	A	192	phenyl_center
1	N1	5	phenyl_center	CE1	PHE	A	294	phenyl_ring
1	N1	5	phenyl_center	CG2	VAL	A	78	ch3_phe
1	N1	5	phenyl_center	CD1	ILE	A	300	ch3_phe
1	N1	5	phenyl_center	CE2	TYR	A	192	phenyl_ring
1	N1	5	phenyl_center	CG1	VAL	A	78	ch3_phe
1	C3	3	phenyl_ring	CG	PHE	A	294	phenyl_center
1	N3	8	h_don	OG1	THR	A	217	h_acc
1	N3	8	h_don	OD1	ASP	A	214	h_acc
1	N4	10	h_don	OD1	ASP	A	34	h_acc
1	N4	10	h_don	OD2	ASP	A	214	h_acc
1	N4	10	h_don	OD1	ASP	A	214	h_acc
1	C12	20	phenyl_ring	CG	TYR	A	77	phenyl_center
1	N7	19	phenyl_center	CD2	TYR	A	77	phenyl_ring
1	O1	14	h_acc	N	VAL	A	78	h_don
1	N6	12	h_don	O	GLY	A	36	h_acc



Note: Guanidino compounds are likely to be novel, so far, not identified as inhibitors for Plasmepsins

Concluding:

Nowadays, docking suffers from 3 serious limitations:

1. ***It requires a crystal structure,***
2. ***It is still pretty slow,***
3. ***It cannot predict affinity.***

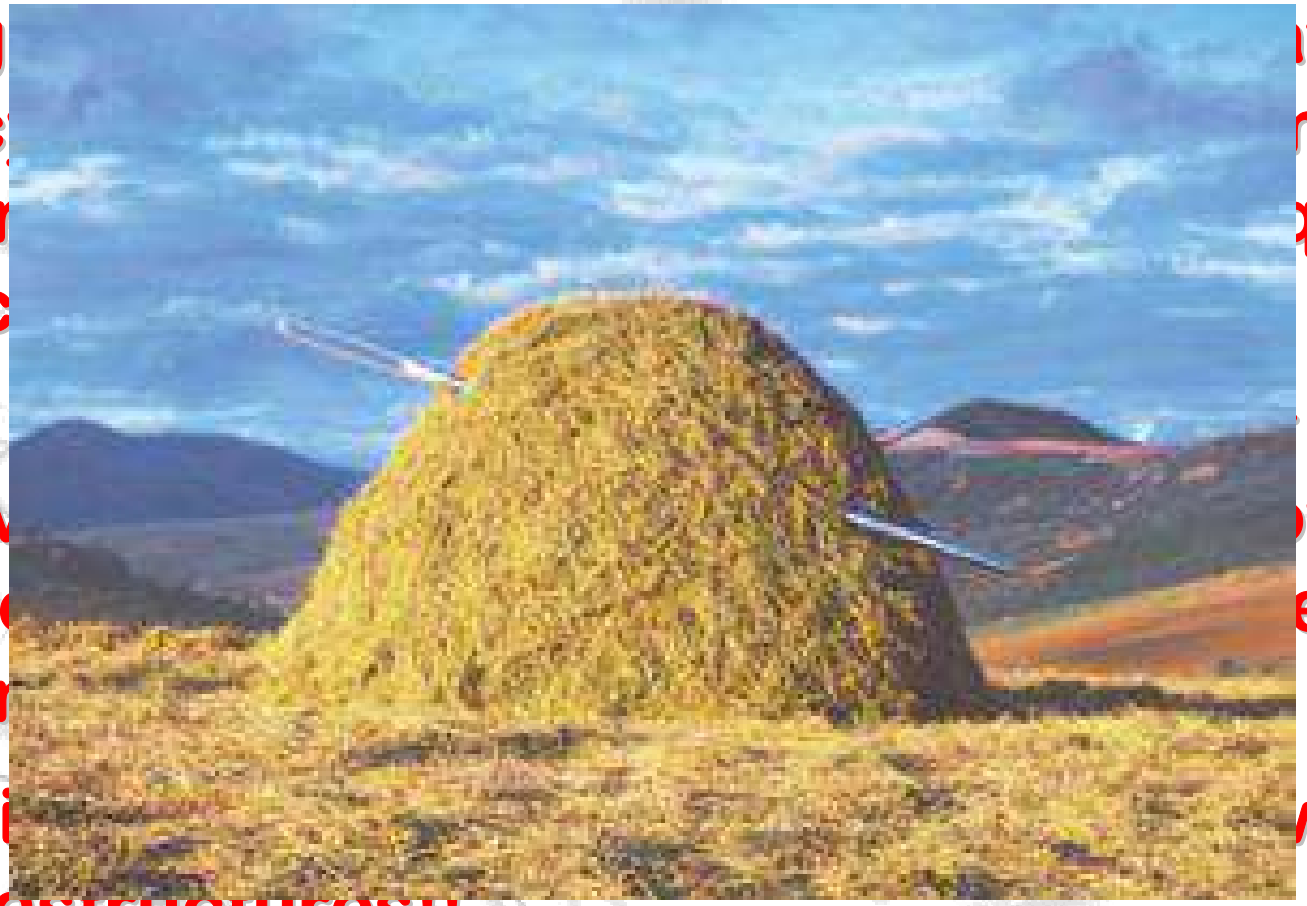
Bottom line: current docking is almost always better than random, but still way too inaccurate to be a sole or dominant approach for lead generation.

Perspectives:

Docking ateau in
accurac nportant
breakthr quantum
dynamic

Cluster highly
predictiv w many
structur elatively
little mar

Better wledge's
and infrastructures!!



Makes the needle in the haystack considerably larger

Backstage!