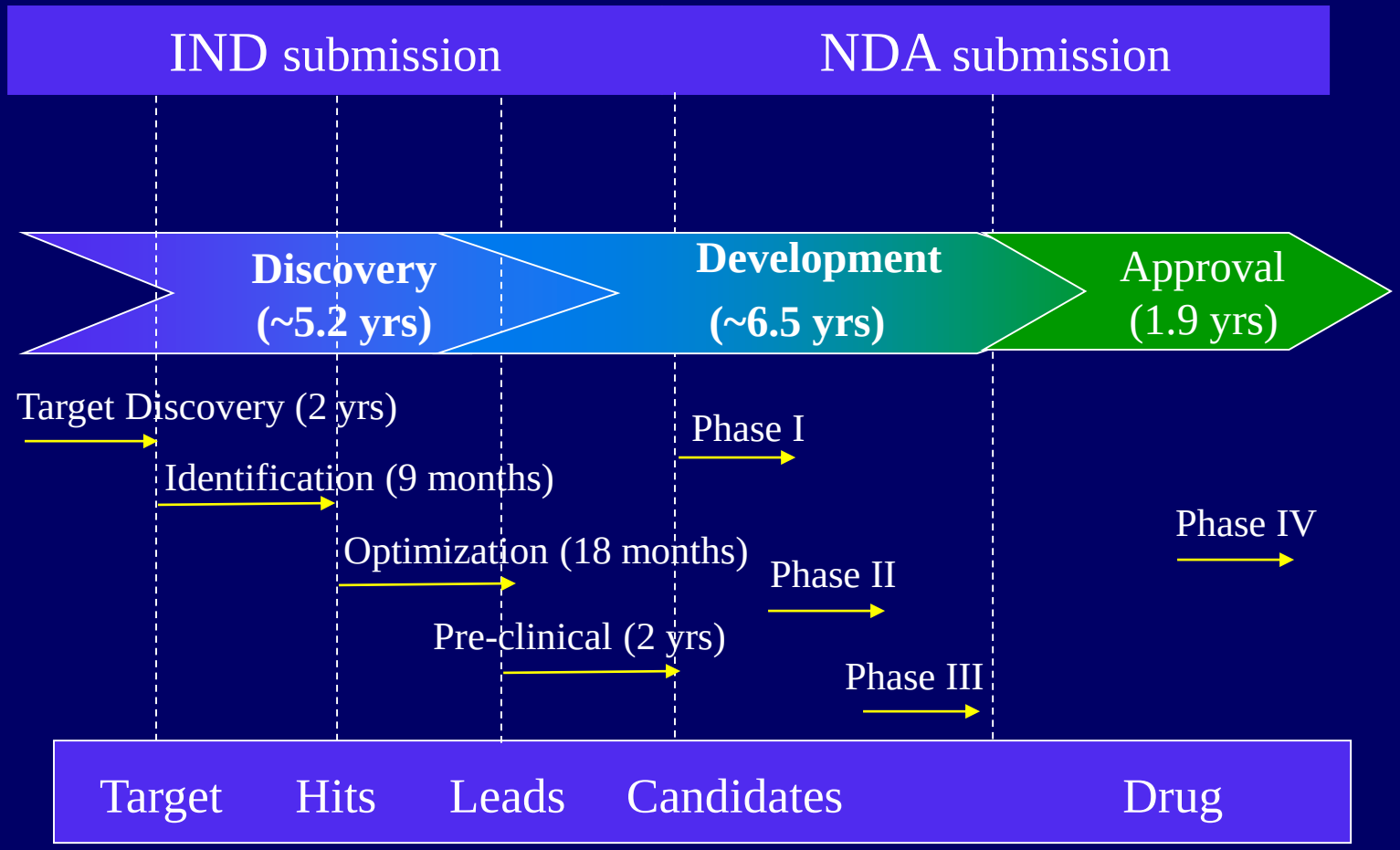
Lecture 2: Computer-Aided Drug Design

Junmei Wang
Department of Pharmacology, University of Texas Southwestern Medical Center at Dallas

Junmei.wang@utsouthwestern.edu

It is getting more difficult to bring a drug into market



average 13.6 years, $900M spending

The Chemical Space

1. Total chemical space: 10^{60} molecules
2. Total chemical substance in literature: 88 million
3. Total registered chemicals: 27 million
4. Number of small molecules within our bodies: a few thousands

   The biological relevant chemical space is only a minute fraction of the complete chemical space

1. 39,000 protein crystal structures
2. 367,000 small molecule X-ray structures

Nature, **2004**, *432*, 824

It is Extremely Challenging to Discover Small Molecules to Modulate the Function of Proteins

1. Quality of chemical libraries
2. Limit of valid drug targets
3. Quality of bio-assays

   HTS hits are likely to be different from assay to assay, and only about 30% of hits shown up in all three assays in one study.

Nature, **2004**, *432*, 824

Drug Discovery Approaches

- By chance

  penicillins, librium

- Random screening

  ‘war on cancer’ by NCI in 1970s

- Targeted screening

  HTS

- Drug metabolism studies

  sulindac, terfenadine HCl

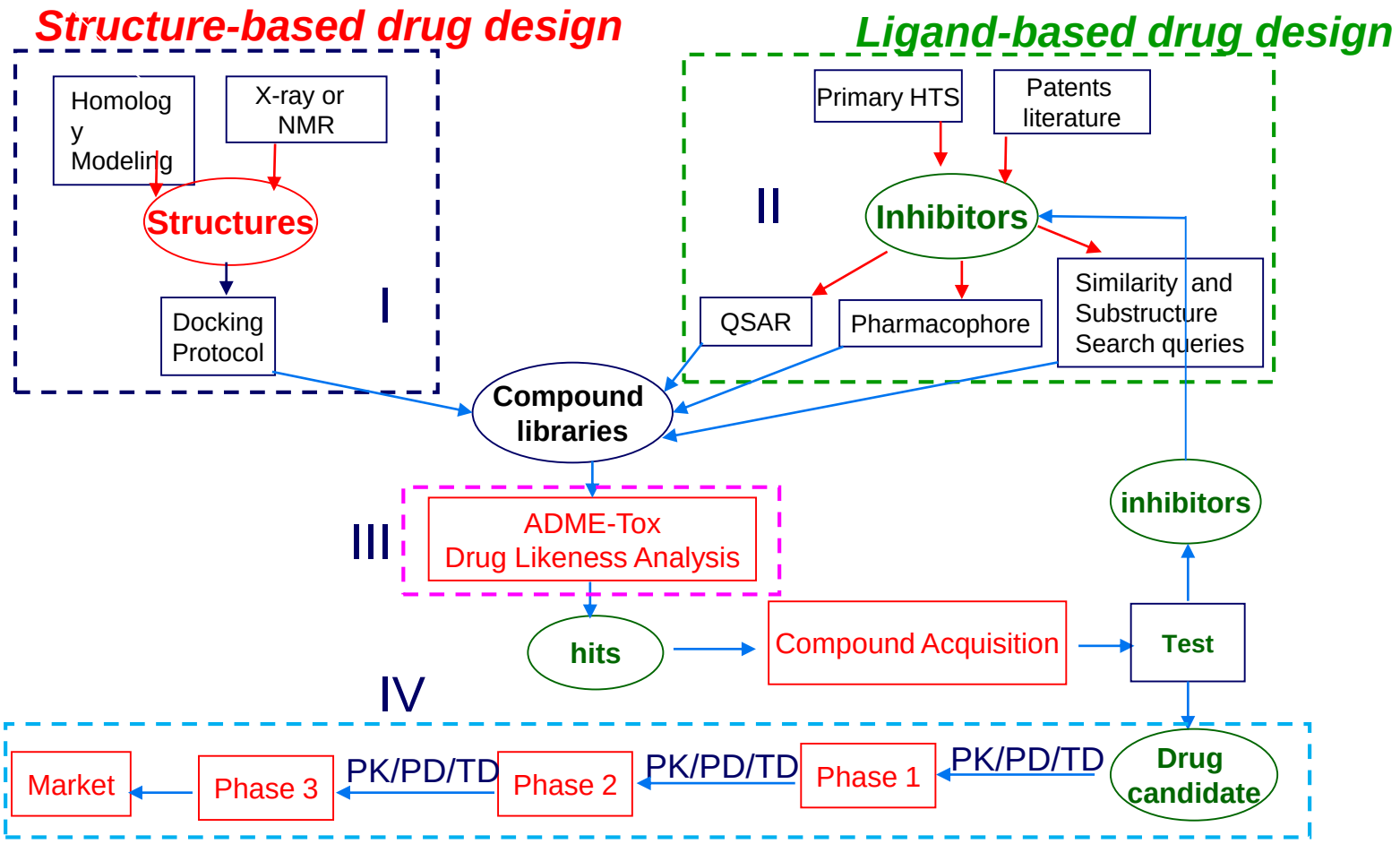
- Clinical observations

  dimenhydrinate tested at the allergy clinic, used for the treatment of seasickness and airsickness.

  bupropion HCl, sildenafil citrate

- Rational design

Computer-Aided Drug Discovery And Development



Some Famous Remarks on CADD

- **GIGO – Garbage in and garbage out**
- **85% Rule**

  If two compounds have 85% similarity, there is 85% chance the two compounds have similar activities

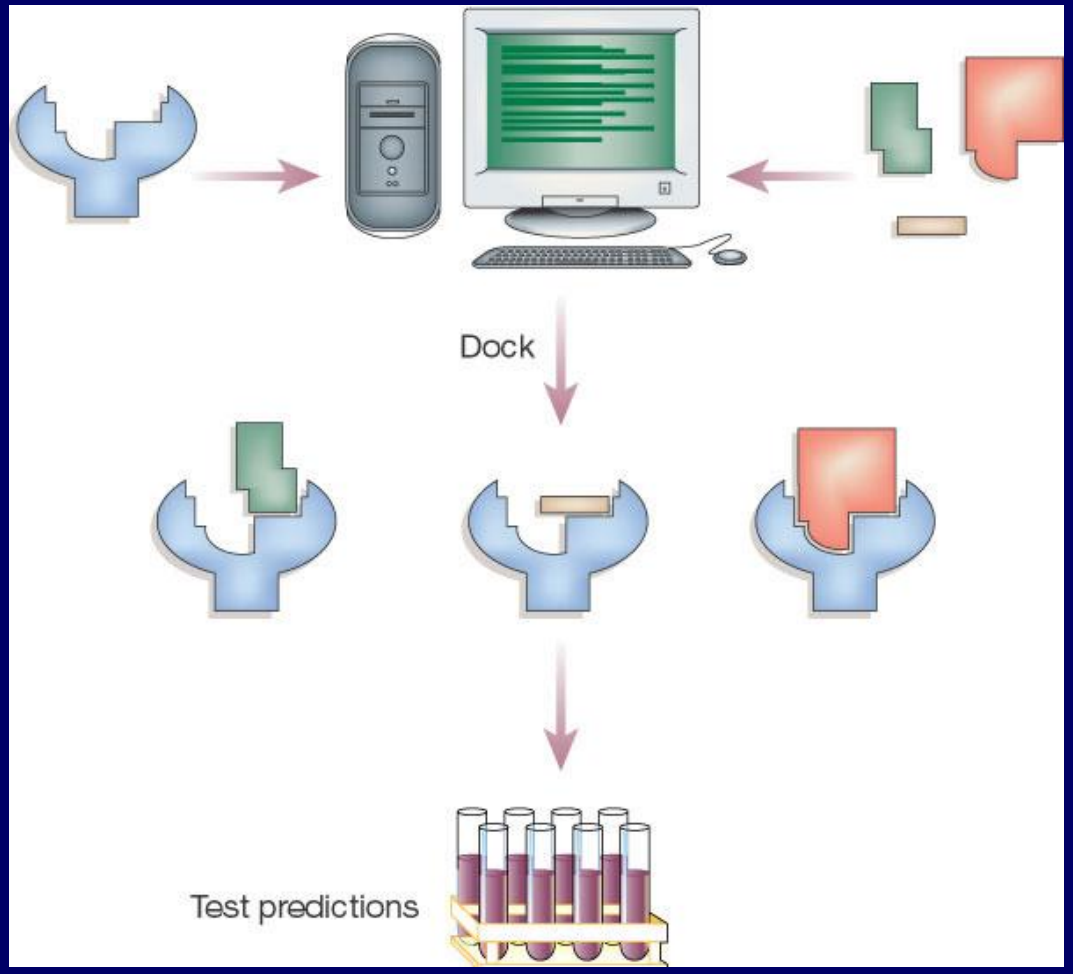
  Tanimoto similarity
- **Rule of 5**

  Lipinski

  Partition coefficient log *P* <= 5
  Molecular weight <= 500
  Number of hydrogen bond donors <= 5 (NH or OH)
  Number of hydrogen bond acceptors <= 10 (N and O)

  Polar surface area no greater than 140 $Å^2$
  Molar refractivity from 40 to 130

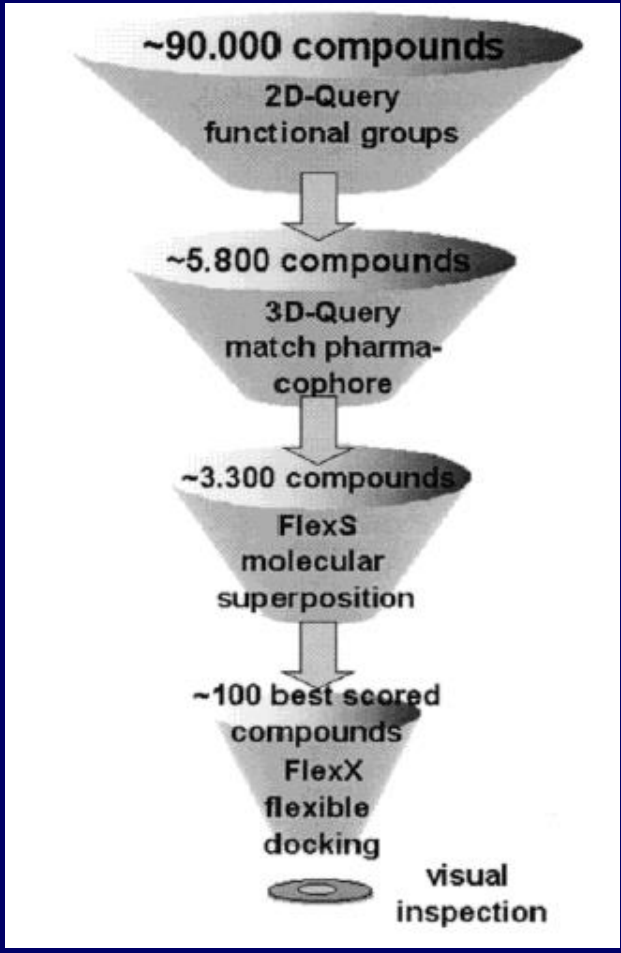
In silico screening (virtual screening)



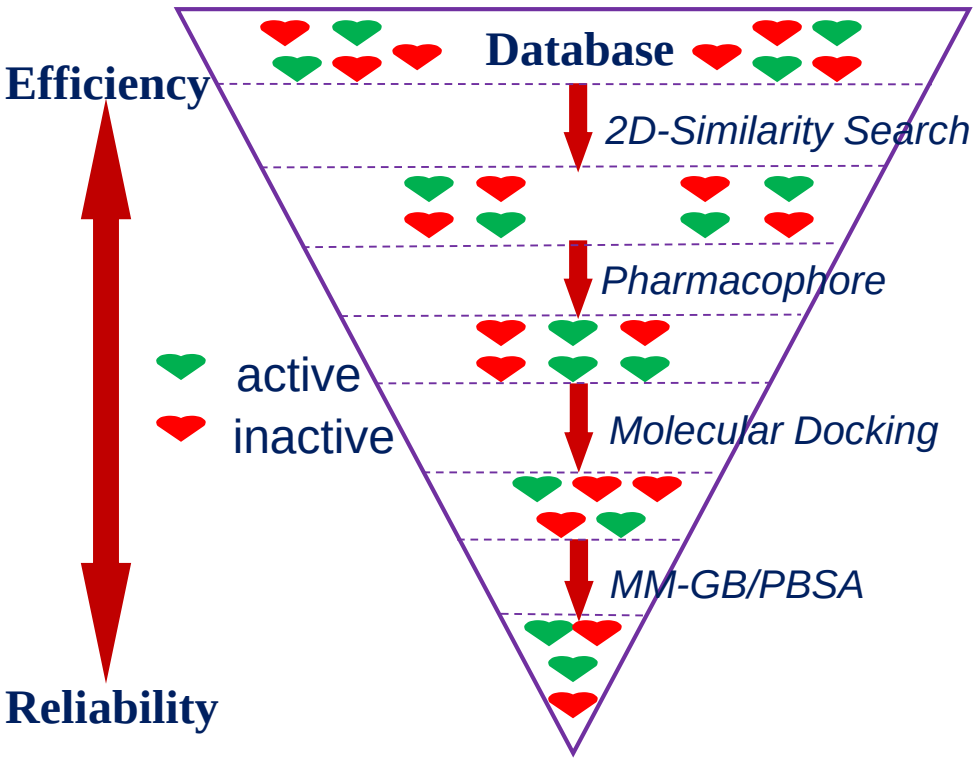
A Hierarchical Strategy for Virtual Screening

1. Simple Filters - Lipinski ‘Rules of Five’
2. 2D-querys based on known inhibitors
3. 3D-querys
4. 3D-structural similarity search
5. Flexible docking
6. Visual inspection, 13 selected, 10 active

J. Med. Chem., **2002**, *45*, 3588



Lead Identification Through Virtual Screening Using A Set of Hierarchical Filters



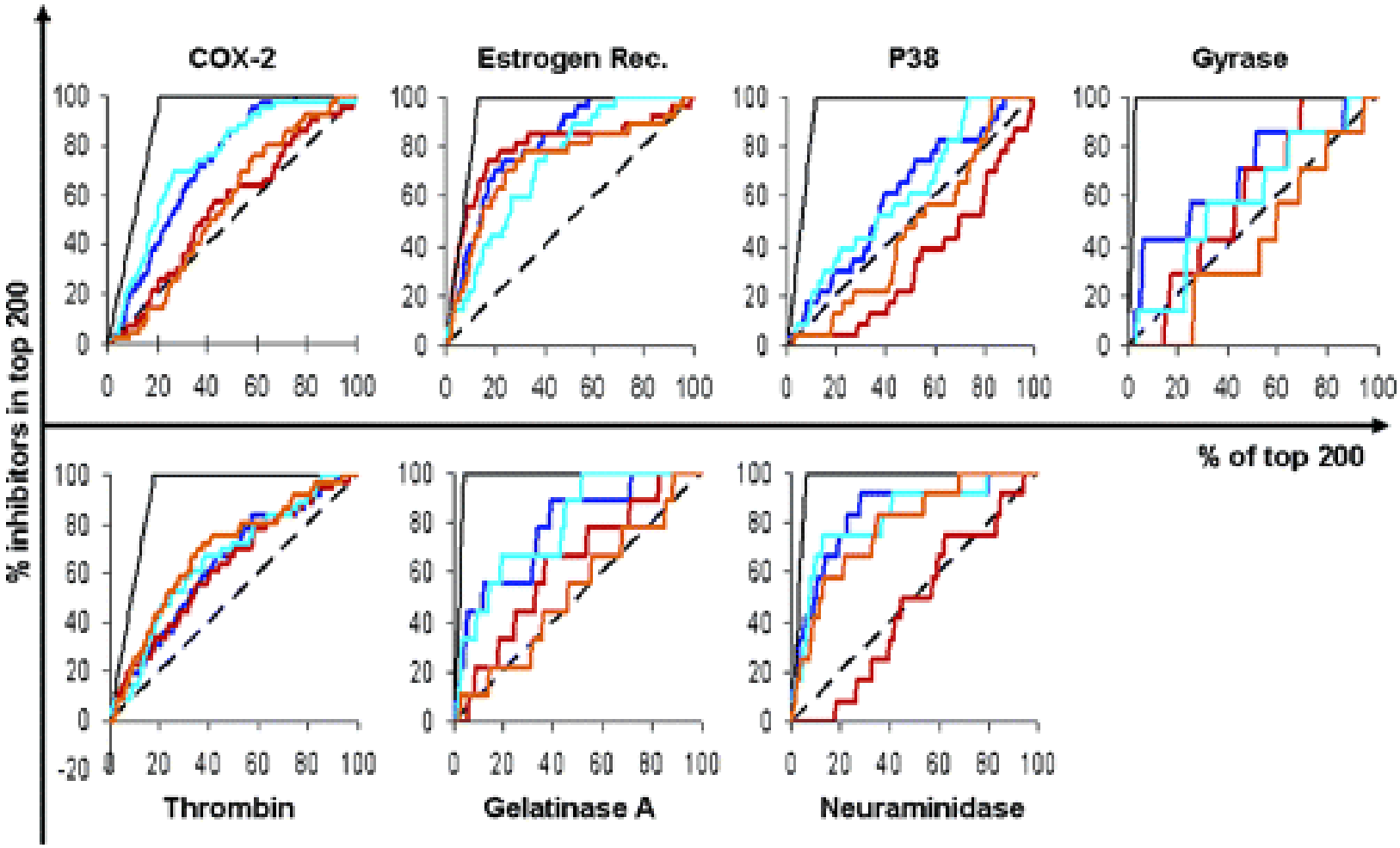
$$HR = \frac{m}{M}$$

$$EF = HR \times \frac{N}{n} = \frac{mN}{Mn}$$

HR : hit rate
EF : enrichment factor

N – total number of molecules in the library
n – number of hits
M – total number of known inhibitors
m – number of known inhibitors recognized as hits

Enrichment Curves



random selection (black, dashed)
ideal performance (black, solid)performance

Screening Databases

- SCD (Symyx Screening Compound Directory)
  http://www.symyx.com
  5.5 million compounds
- ZINC - a free database of commercially-available compounds for virtual screening
  http://zinc.docking.org/
  8 million compounds
- Pubchem
  http://pubchem.ncbi.nlm.nih.gov
  http://en.wikipedia.org/wiki/PubChem
  maintained by National Center for Biotechnology Information (NCBI)
  19 million compounds
- GDB-13: 970 million – J. AM. CHEM. SOC. 2009, *131*, 8732–8733

1D and 2D-Based Approaches – A Review

- **1D-based approach**

  Drug likeness analysis

  Lipinski's 'Rule of Five'

  MW < 500, clogp < 5.0, H-donor < 5, H-acceptor < 10

  PSA – polar surface area (<140 $Å^2$)

- **2D-based fingerprint**

  MDL, Daylight, Tripos

  Advantage of 2D approaches: fast, can essentially eliminate most unwanted compounds

  Tanimoto Coefficient = $N_{AB}/(N_A+N_B-N_{AB})$

  N_{AB} – number of features common to both A and B

  N_A – number of features in A, N_B – number of features in B

  T > 0.85 %

3D-Based Approaches – A Review

- **CoMFA** – Comparison of Molecular Field Analysis
- **HQSAR** – hologram QSAR
- **3D-Fingerprint**
- **Pharmacophore**

  Ligand-based:

  GASP, DISCO, DISCOtech, Galahad (Tripos), PHASE (Schrodinger), Catalyst (Accelrys), Discovery Studio (Accelrys) …

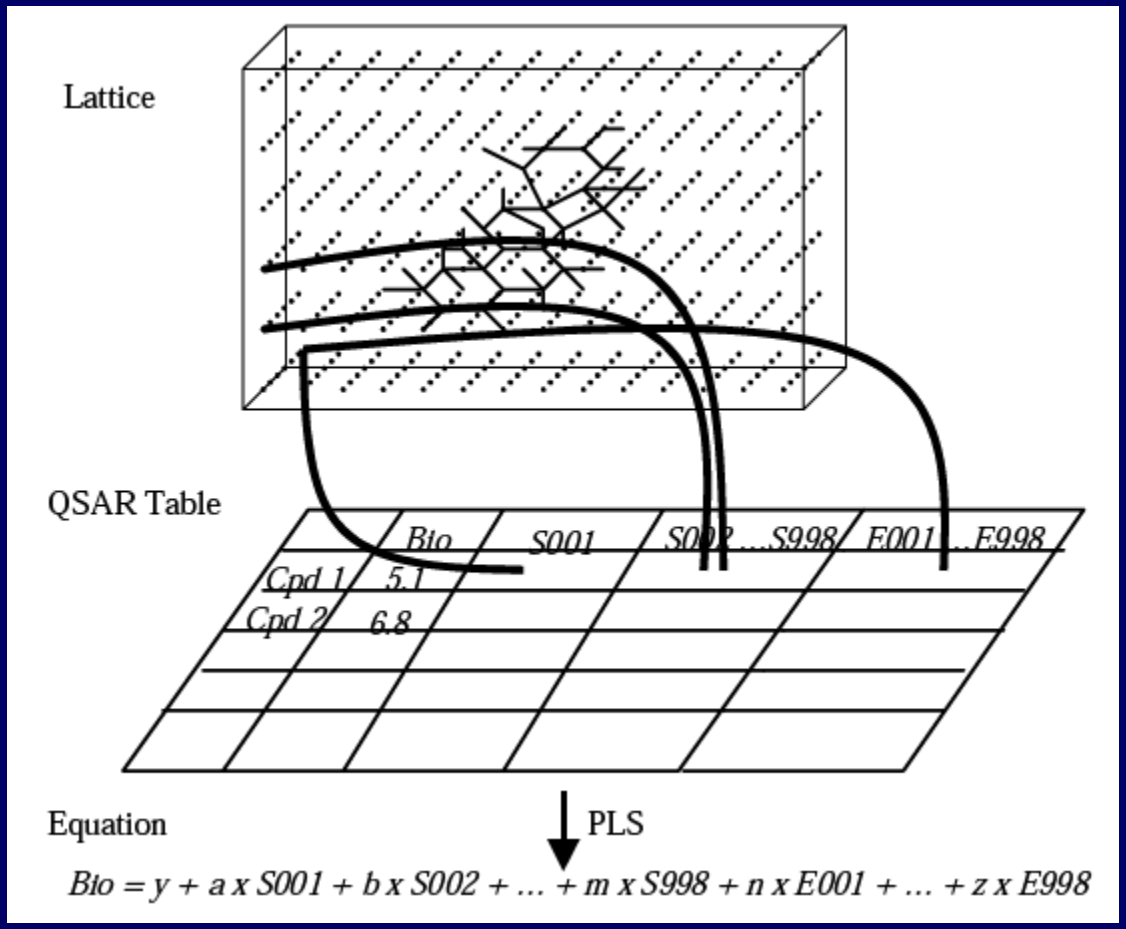
  Receptor-based:

  Unity (Tripos)
- **3D-property comparison**

  Shape – Rocs (OpenEye)

  Electrostatics – eon (OpenEye)
- **Molecular docking**

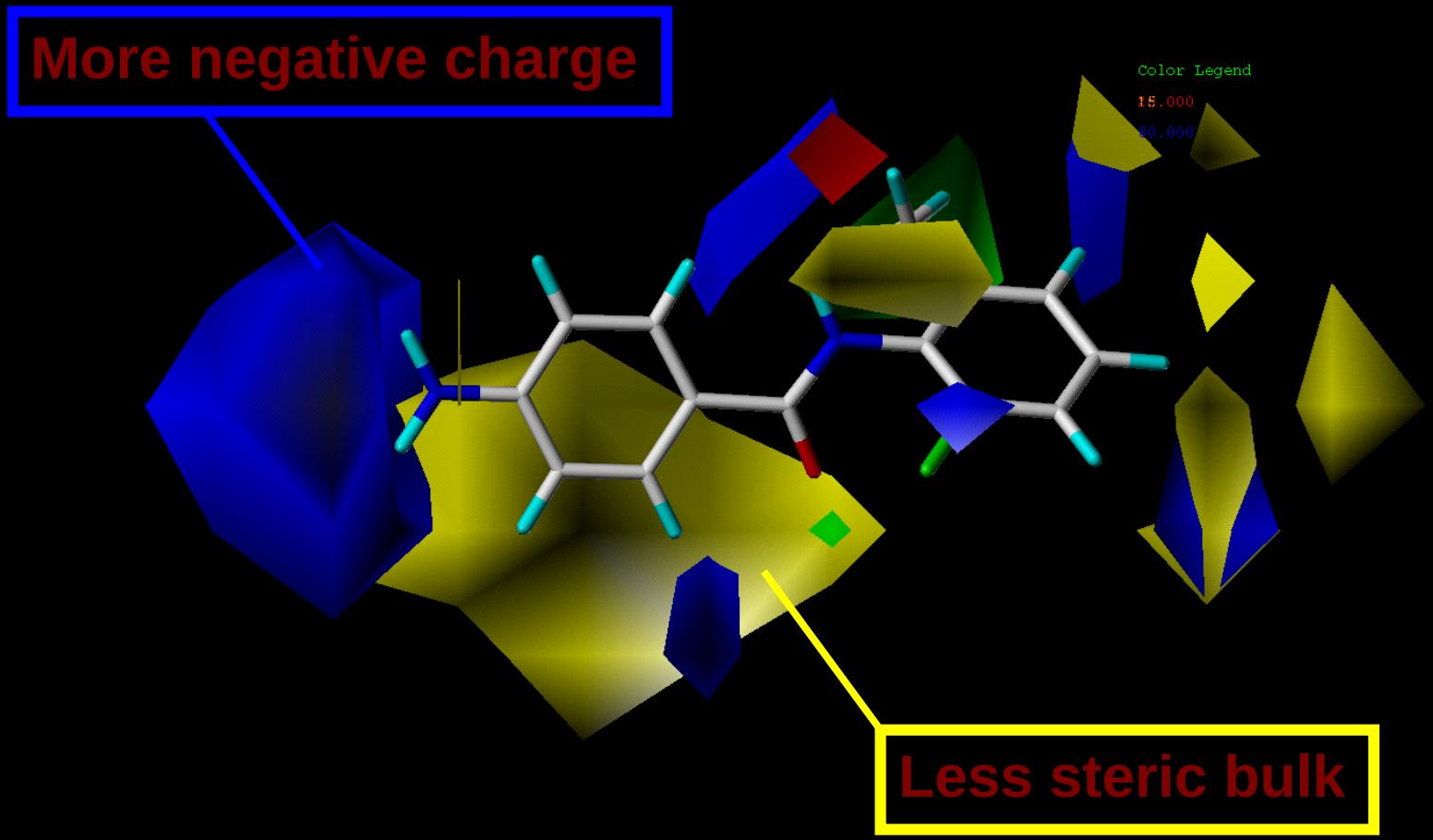
Principal of CoMFA



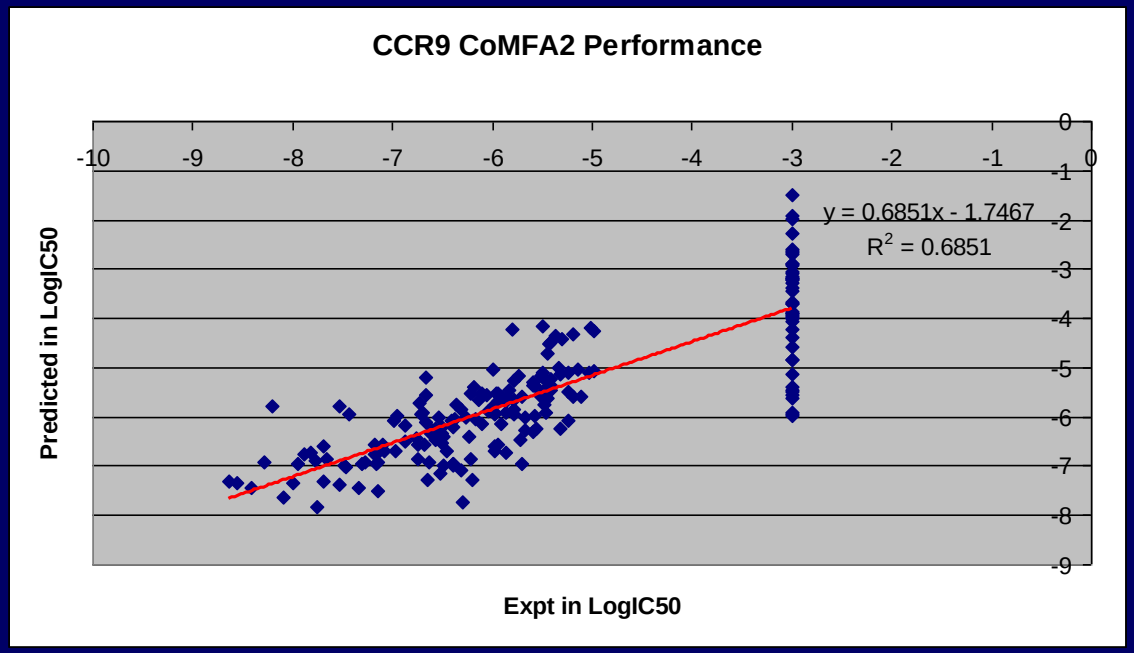
Adopted from Sybyl 7.3 Manual

Principal of CoMFA – continued

- High Coefficient (important) lattice points can be plotted around molecular structures



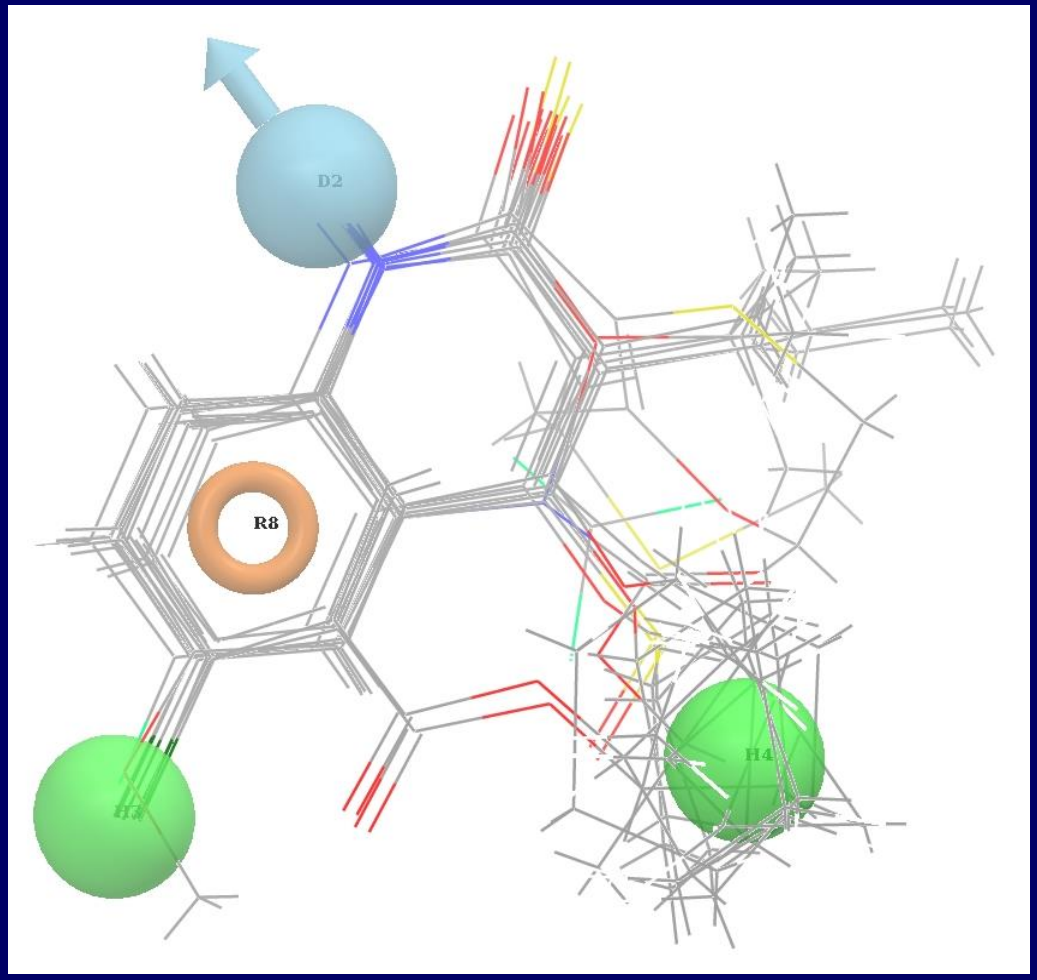
Case Study: 3D-QSAR



N = 198
Standard error = 1.20 log unit
$q^2 = 0.44$

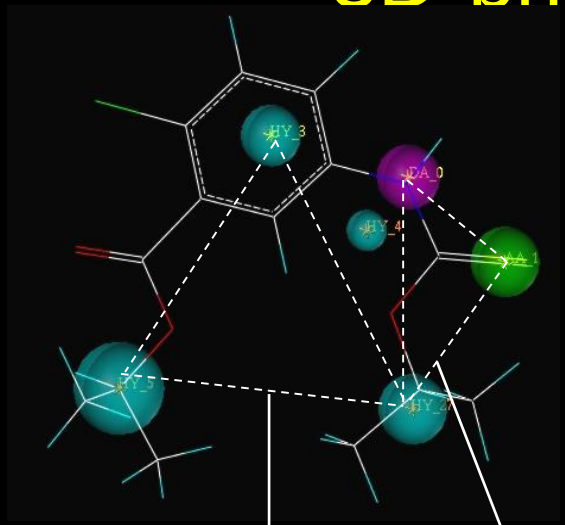
Screened 100,000 compounds, purchased 200 compounds, 42 have activity better than 10 μM

Pharmacophore and Auxophore



- **Pharmacophore** – the relevant groups on a molecule that interact with a receptor and are responsible for the activity
- **Auxophore** – other atoms are referred to as auxophore. Auxophore could be essential to maintain the integrity of the molecules and hold the pharmacophoric groups in appropriate positions.

3D-pharmcophore fingerprint



0	0	0	1	0	0	1	0	0	0

Five default features:

D*onor_atom*
A*cceptor_atom*
H*ydrophobic*
P*ositive_N*
N*egative_center*

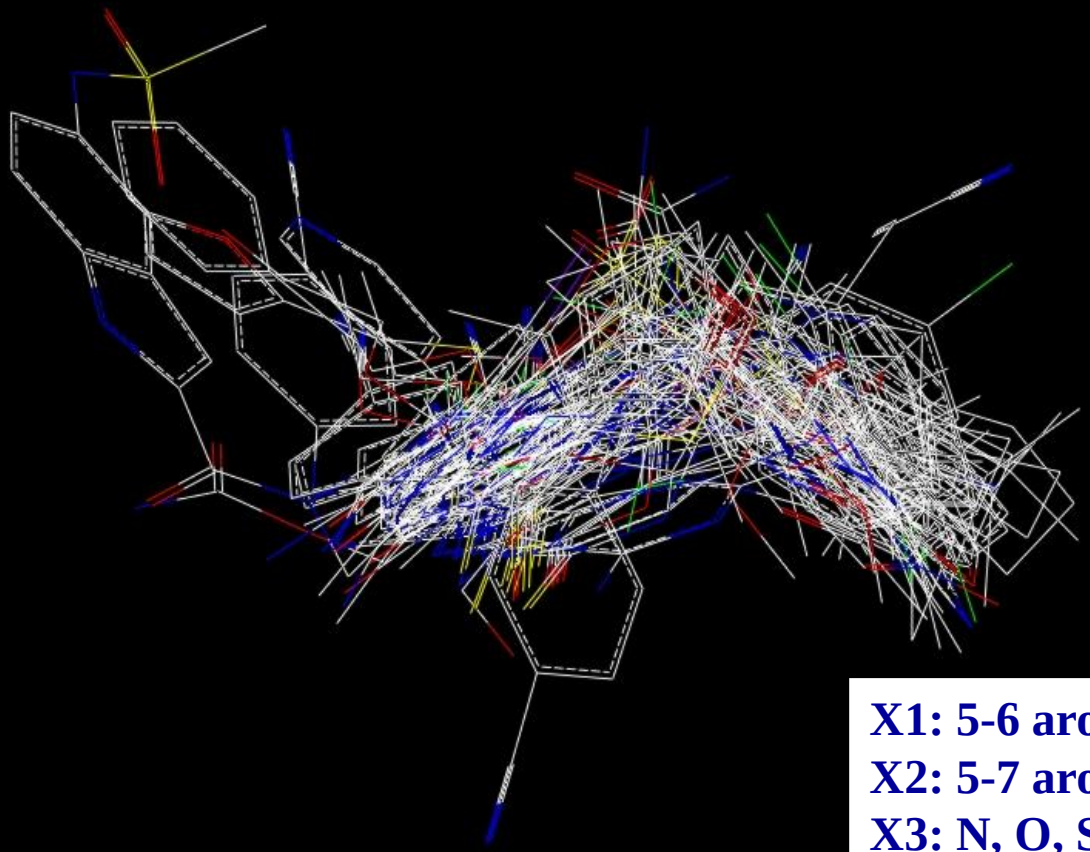
Triplet	DDD 111	DDD 211	DDA 311	DDH 321	DAH 442	DHH 444
Mol1	0	1	0	0	1	1
Mol2	0	1	0	0	0	1
Mol3	1	0	0	0	1	1
Mol4	0	1	1	0	0	1
Vector Sum	1	3	1	0	2	4
Feature Weight	3	3	3	4	4	5
Distance Weight	3	4	5	6	10	12
Bit Score	9	36	15	0	80	240

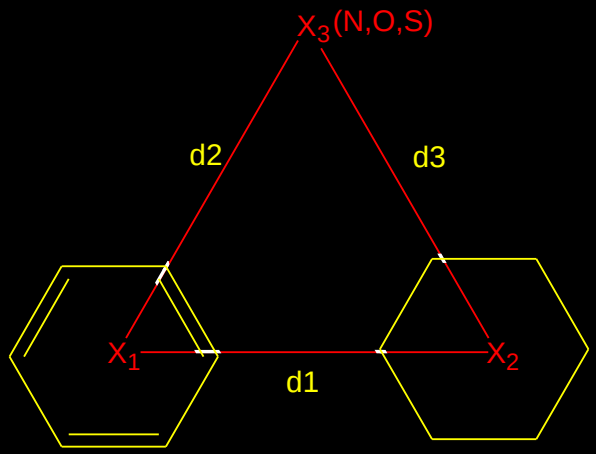
3D-pharmcophore fingerprint

Summary of tuplets virtual screenings for three typical

systems	Known inhibitors	Reference molecules	#hits of actives	#hits of reference molecules	HR	EF
HIV-1 RT	43	5327	31	72	0.72	74.0
thrombin	82	5230	75	457	0.91	11.4
HIV-1 PR	103	5357	53	210	0.51	25.5

A pharmacophore model conceived using a set of crystal structures





X1: 5-6 aromatic ring
X2: 5-7 aromatic or aliphatic ring
X3: N, O, S
Pattern 1: d1 4.5-6.0, d2 3.5-4.5, d3 4.5-6.5 Å
Pattern 2: d1 2.4-2.8, d2 3.5-4.5, d3 4.0-5.5 Å

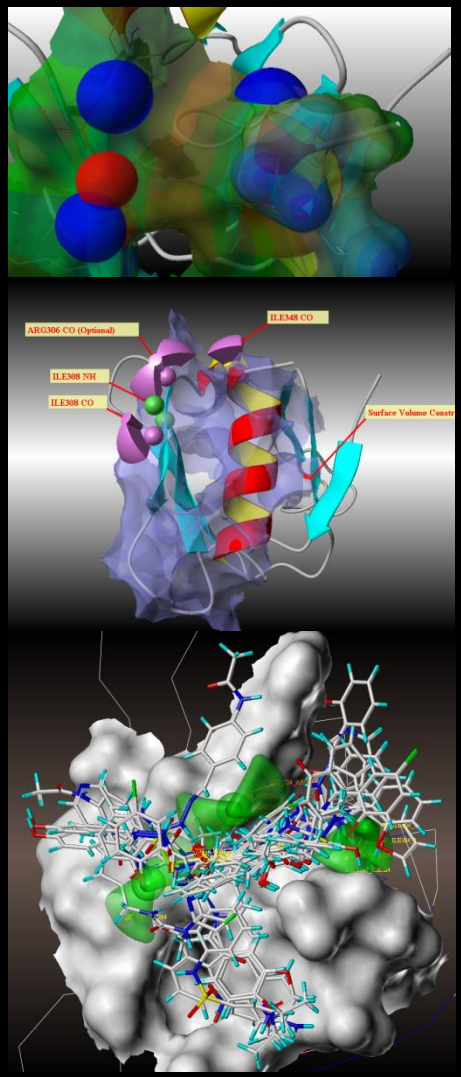
Identify Pharmacophore Based on A Protein Structure

Identify key interaction residues across the binding sites

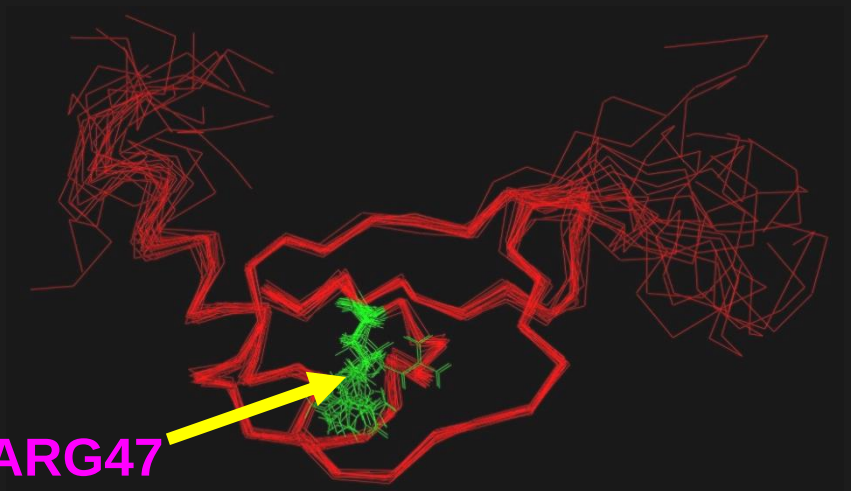
Generate Protein Pharmacophore Model

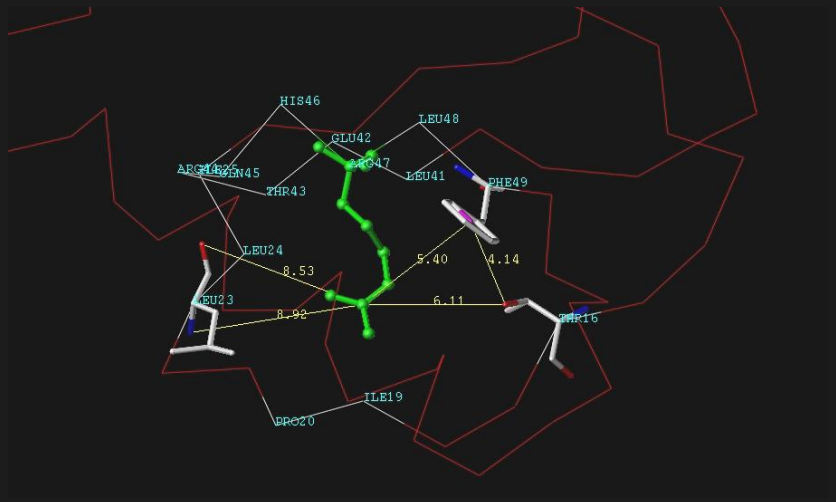
Refine the pharmacophore model by including surface volume constraints

8 out of 67 compounds proven to be active

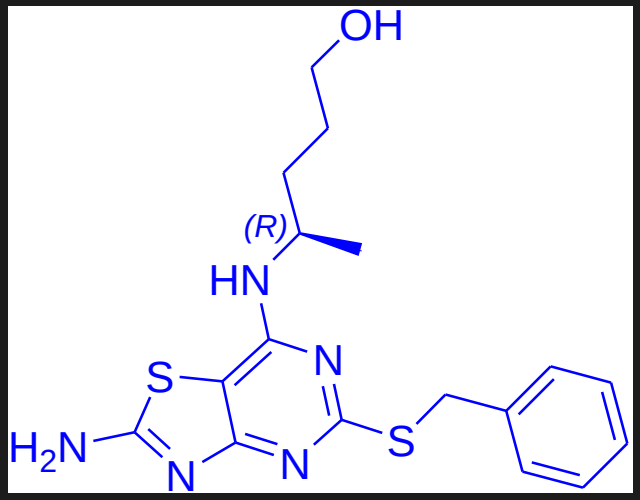


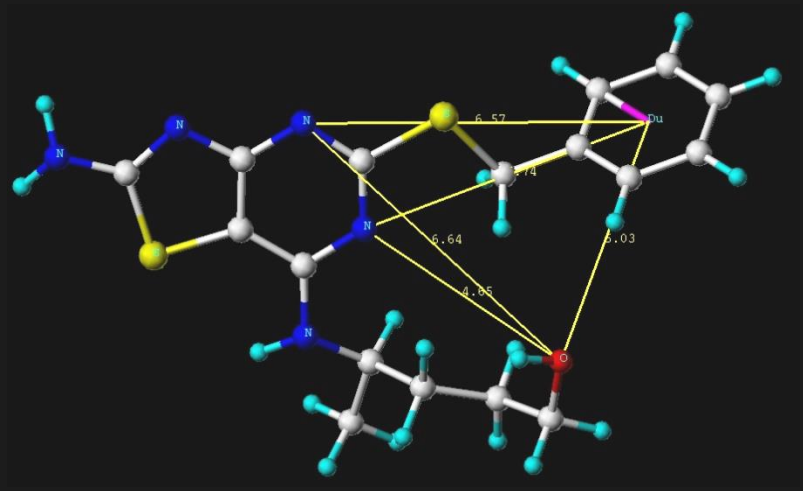
A Pharmacophore Model Based on A Peptide





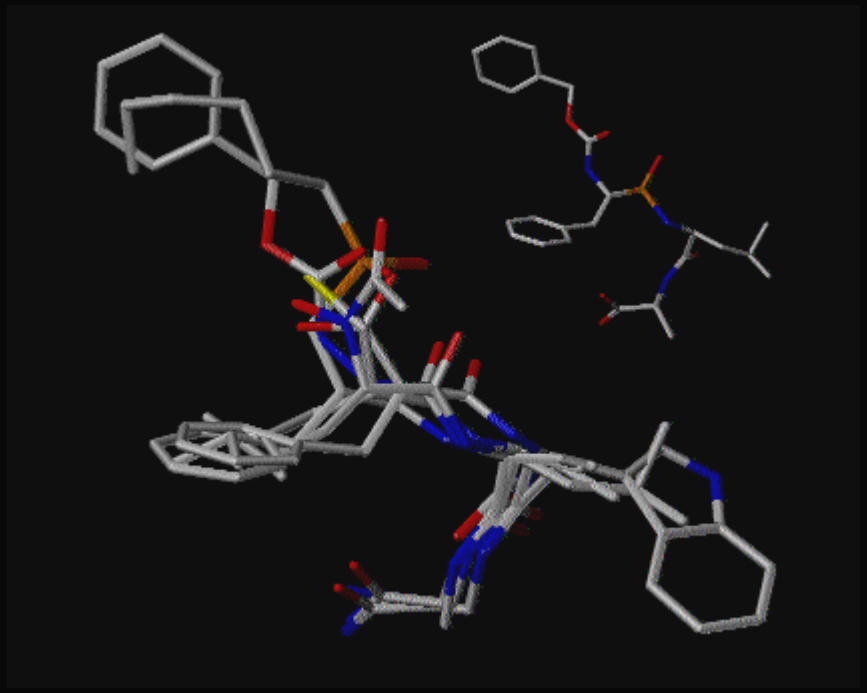
Chemokine domain of fractalkine





AstraZeneca

Pharmacophore Perception

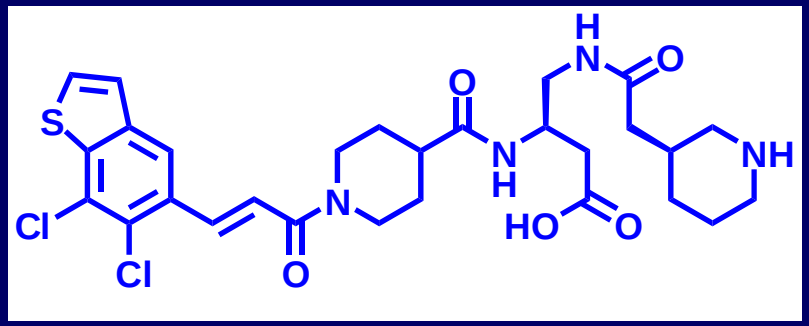


1. Structural alignment
2. Pharmacophore detection
3. Quantitative Structure-Activity relationships
4. *De novo* design

Combined fingerprint-based scores

S_{2d} - 2D similarity score (MDL, Openeye, Sybyl)
S_{shape} - Shape-based score (Rocs)
S_{elec} - Electrostatic similarity score (Eon)
S_{drug} - Drug-like score (Rule of 5, psa etc)

$$S = \sum_i w_i s_i$$

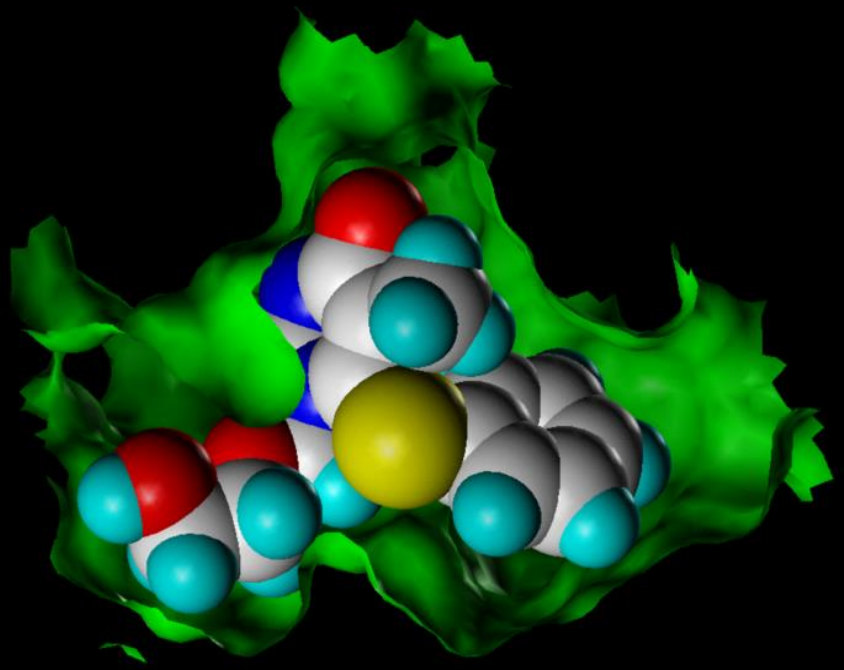


Screened 200,000 compounds, purchased 162 compounds,
12 have activity better than 1 μM

Summary of Virtual Screenings at A Pharmaceutical Company

Project	Methodology	# of Compounds purchased	Total expense ($)	# of Hits	HTS (100,000)
Project 1	**Pharmacophore**	**257**	**4138**	**1**	**No hits**
Project 2	**3D-QSAR**	**200**	**3088**	**42**	**N/A**
Project 3	**Docking**	**150**	**2506**	**0**	**About 80 hits, none of them is developable**
Project 3	**2D-fingerprint**	**96**	**2152**	**2**	
Project 3	**Combined fingerprint-based scores**	**162**	**3404**	**12**	

Molecular Docking



	2: FLEXX	13: G_	14: PMF	15: D_	16: CSCORE
REF1_001	−16.85	−209.21	−66.83	−125.35	4
REF1_002	−15.38	−215.58	−73.97	−127.04	4
REF1_004	−11.88	−197.37	−52.73	−139.73	3
_REF1_011	−10.23	−220.20	−45.63	−132.22	2
REF1_013	−9.77	−191.04	−75.08	−126.32	3
_REF1_018	−8.72	−58.30	−41.57	−42.56	0
REF1_019	−8.50	−190.67	−80.99	−123.83	3

Step 1: Docking a ligand into the binding site

Step 2: Evaluating the docking poses, i.e. calculating the docking scores

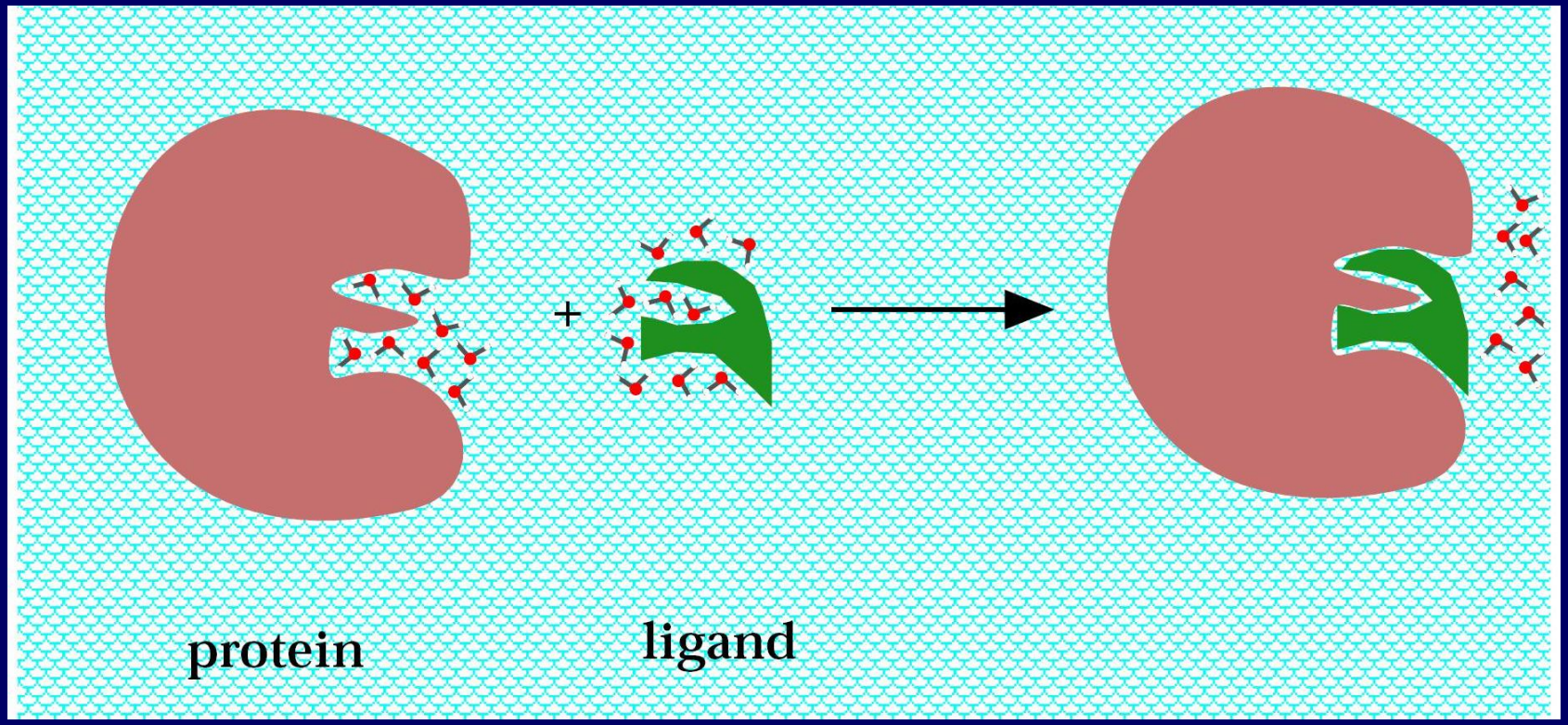
Docking Glossary

- **Receptor** or **host** – The "receiving" molecule, most commonly a protein or other biopolymer.
- **Ligand** or **guest** – The complementary partner molecule which binds to the receptor. Ligands are most often small molecules but could also be another biopolymer.
- **Docking** – Computational simulation of a candidate ligand binding to a receptor.
- **Binding mode** – The orientation of the ligand relative to the receptor as well as the conformation of the ligand and receptor when bound to each other.
- **Pose** – A candidate binding mode.
- **Scoring** – The process of evaluating a particular pose by counting the number of favorable intermolecular interactions such as hydrogen bonds and hydrophobic contacts.
- **Ranking** – The process of classifying which ligands are most likely to interact favorably to a particular receptor based on the predicted free-energy of binding.

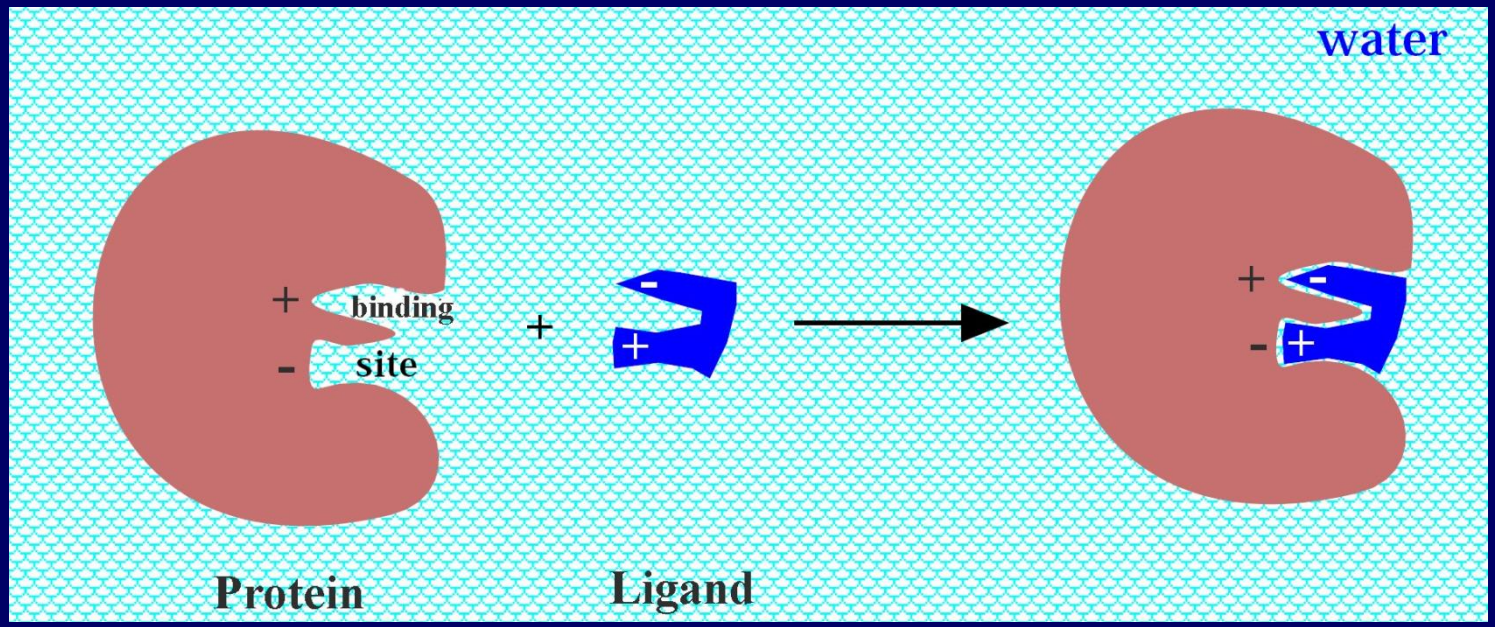
Docking Software Packages

1. **DOCK** – Developed in Tack Kuntz's group at UCSF (http://dock.compbio.ucsf.edu)
2. **GOLD** – Developed at Sheffield University, distributed by CCDC (http://www.ccdc.cam.ac.uk)
3. **FLEXX** – BioSolveIT (http://www.biosolveit.de/FlexX)
4. **FRED** – OpenEye Scientific (http://www.openeye.com)
5. **AUTODOCK** – Scripps Research Institute - http://autodock.scripps.edu/
6. **SURFLEX** – Developed by Ajay Jain at UCSF, distributed by Tripos (http://www.tripos.com)
7. **GLIDE** – Schrodinger LLC (http://www.schrodinger.com)

Ligand Binding is a Dehydration Process.



The Driving Forces for Protein-Ligand Complex Formation

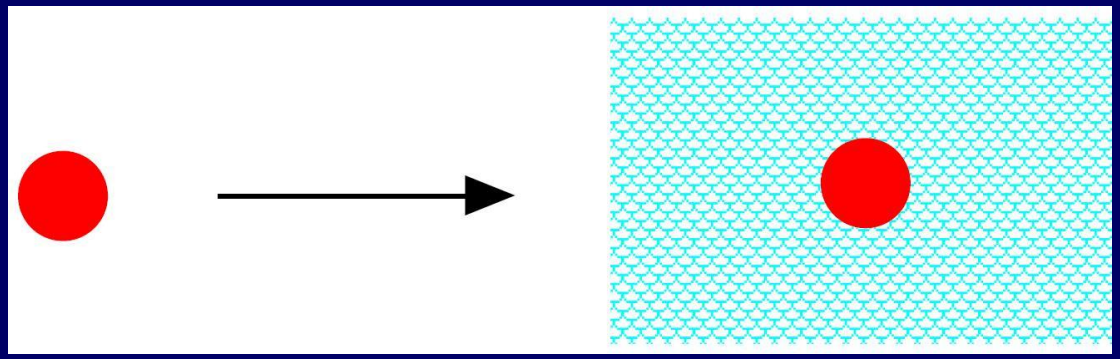


- Electrostatic interactions
- van der Waals interactions
- Entropic effects (ligand, protein, solvent)

Bottleneck: The system is not in vacuum!

Solvent Effect is the Bottleneck.

- Weaken the charge-charge interactions
- Self energies



$$\Delta G \equiv G_{solvent} - G_{vacuum} = -\frac{q^2}{2a}\left(1 - \frac{1}{\varepsilon_{water}}\right) \approx -\frac{q^2}{2a}$$

ε_{water}=78.3

Therefore, a charged group favors staying in aqueous environment.

Dehydration of a charge will cost energy.

Born, *Z. Phys.*, 1920

Hydrophobic Effect of Solvent

Docking Scoring Function

- **Knowledge-based** – atom pairs in contact

$$Score = \sum_{r<cutoff} A_{ij}(r)$$

PMF, Drug Score (*J. Med. Chem.*, **2005**, *48*, 6296)

- **Energy-based**
  1. No solvation term (Dock, Gold, LigandFit)
  2. Parameterized solvation term (Glide)
  3. Free energy based

Docking Scoring Function – to be continued

- **Simple scoring function**

$$E = E_{vdw} + E_{elec} + E_{tor}$$

$$E_{vdw} = \sum_{ij} \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}$$

$$E_{elec} = \sum_{ij} \frac{q_i q_j}{r_{ij}}$$

$$E_{tor} = \sum_{i} v_i (1 + \cos n\phi - \theta)$$

- **Empirical scoring function**

$$G = G_0 + n_{hbond} \times G_{hbond} + n_{metal} \times G_{metal} + n_{lipo} \times G_{lipo} + n_{rot} \times G_{rot} + SAS \times G_{sas} + \ldots$$

- **Free energy**

$$\begin{array}{ccccc} A_{aqueous} & + & B_{aqueous} & \xrightarrow{\Delta G_{binding}} & AB_{aqueous} \\ \downarrow -\Delta G_{solv}^{A} & & \downarrow -\Delta G_{solv}^{B} & & \uparrow -\Delta G_{solv}^{AB} \\ A_{gas} & + & B_{gas} & \xrightarrow{\Delta G_{gas}} & AB_{gas} \end{array}$$

$$\Delta G_{binding} = G^{AB} - (G^{A} + G^{B})$$

How to Calculate Free Energy of a Molecule?

$$G = G_{gas} + G_{solv} = H_{gas} - TS_{MM} + G_{solv}$$
$$\approx E_{gas} - TS_{MM} + G_{solv} \quad (1)$$

$$G_{PBSA/GBSA} = G_{PB/GB}^{elec} + G_{SA}^{nonpolar} \quad (2)$$

$$G_{GB}^{elec} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{\varepsilon r_{ij}} - \frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right) \sum_{i=1}^{N} \frac{q_i^2}{\alpha_i} \quad (3)$$

$$\nabla . \varepsilon(r) \nabla \phi(r) = -4\pi\rho(r) \quad (4)$$

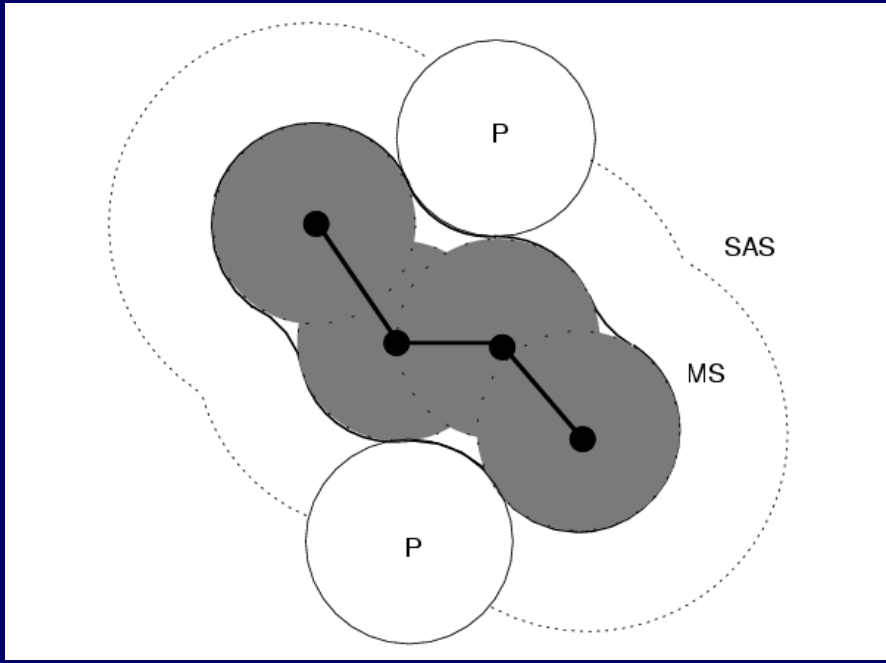
$$G_{SA}^{nonpolar} = G_{cav} + G_{vdw} = \gamma SAS + b \quad (5)$$

Surface Area Definitions

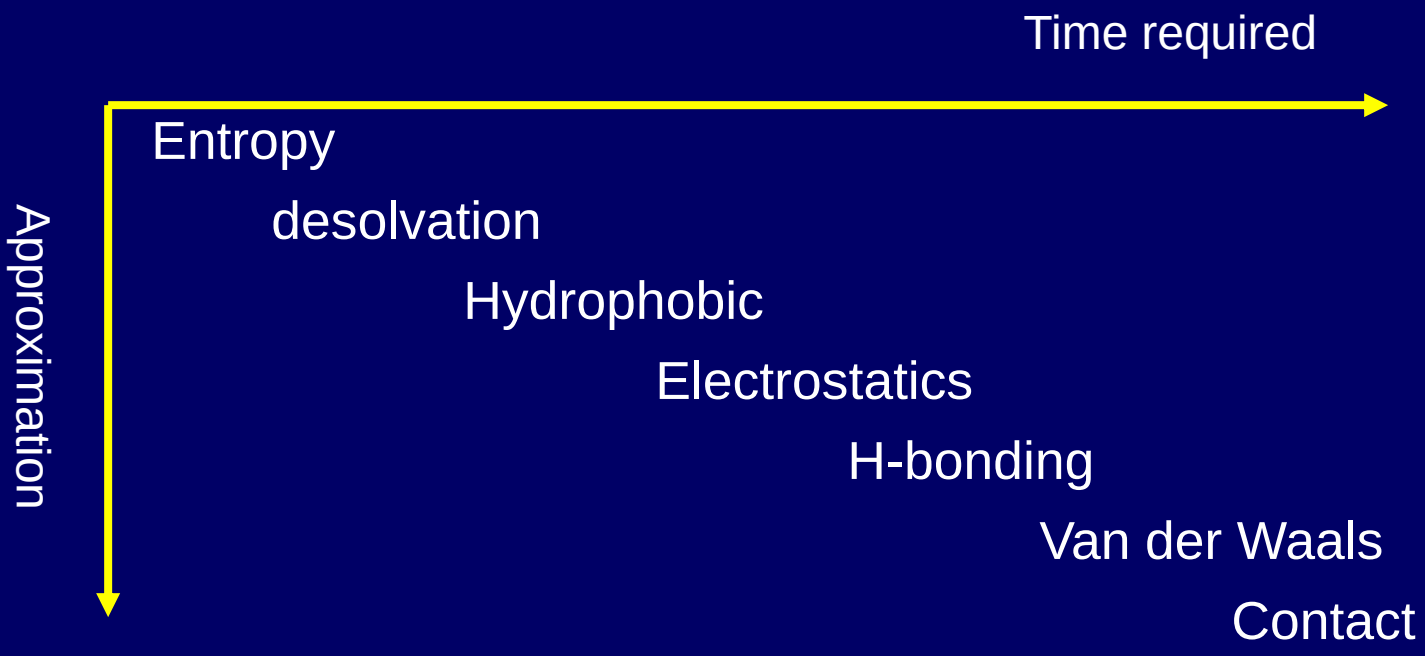
Van der Waals

SES

SAS



Docking Scoring Function



Example: Autodock

- Autodock uses pre-calculated affinity maps for each atom type in the substrate molecule, usually C, N, O and H, plus an electrostatic map
- These grids include energetic contributions from all the usual sources

$$\Delta G = G_1 \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + G_2 \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right)$$
$$+ G_3 \sum_{i,j} \frac{q_i q_j}{\epsilon(r) r_{ij}} + G_4 \Delta G_{tor} + G_5 \sum_{i,j} S_i V_j e^{-r_{ij}^2 / 2\sigma^2}$$

Stouten Pairwise Atomic Solvation Parameters

Favorable for C, A ; Unfavorable for O, N

Proportional to the absolute value of partial charges

Docking Algorithms

- Conformational search

  Omega - OpenEye

- Docking pose searching

  Searching space:

  Rigid docking

  Flexible docking (flexible ligands or flexible ligands and receptor)

  Algorithms:

  1. Genetic algorithms (Gold and AutoDock)
  2. Complementarity methods (DOCK, Fred, Glide, Surflex)

     Shape complementrity

     SAS, overall shape, geometric constraint)

     Binding complementrity

     hydrogen binding, hydrophobic contacts, van der Waals interactions

     Distance geometry

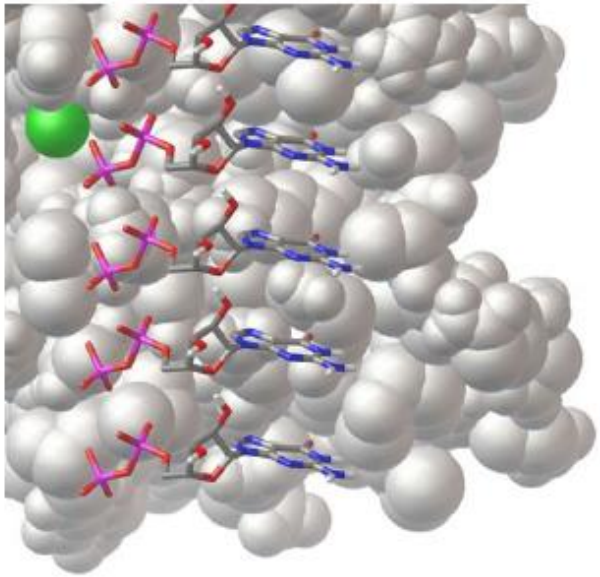
Approaches to Flexibility

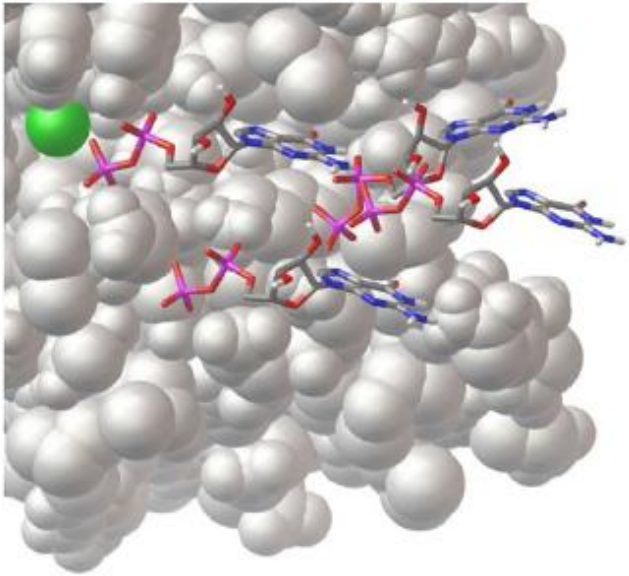
- A relatively simple molecule with 10 rotatable bonds has more than 10^9 possible conformation if we only consider 6 possible positions for each bond
- Monte Carlo, Simulated Annealing and Genetic Algorithm can help navigate this vast space
- Other methods have been developed to again circumvent this problem

Flexibility

- Some algorithms (call Place & Join algorithms) break the ligand up into pieces, dock the individual pieces, and try and reconnect the bound conformations
- FlexX uses a library of precomputed, minimized geometries from the Cambridge database with up to 12 minima per bond. Sets of alternative fragments are selected by choosing single or multiple pieces in combination
- Flexible docking via molecular dynamics with minimization can handle arbitrary flexibility, however it is extremely slow

Two Kinds of Search





Systematic

- Exhaustive, deterministic
- Outcome is dependent on granularity of sampling
- Feasible only for low-dimensional problems

Stochastic

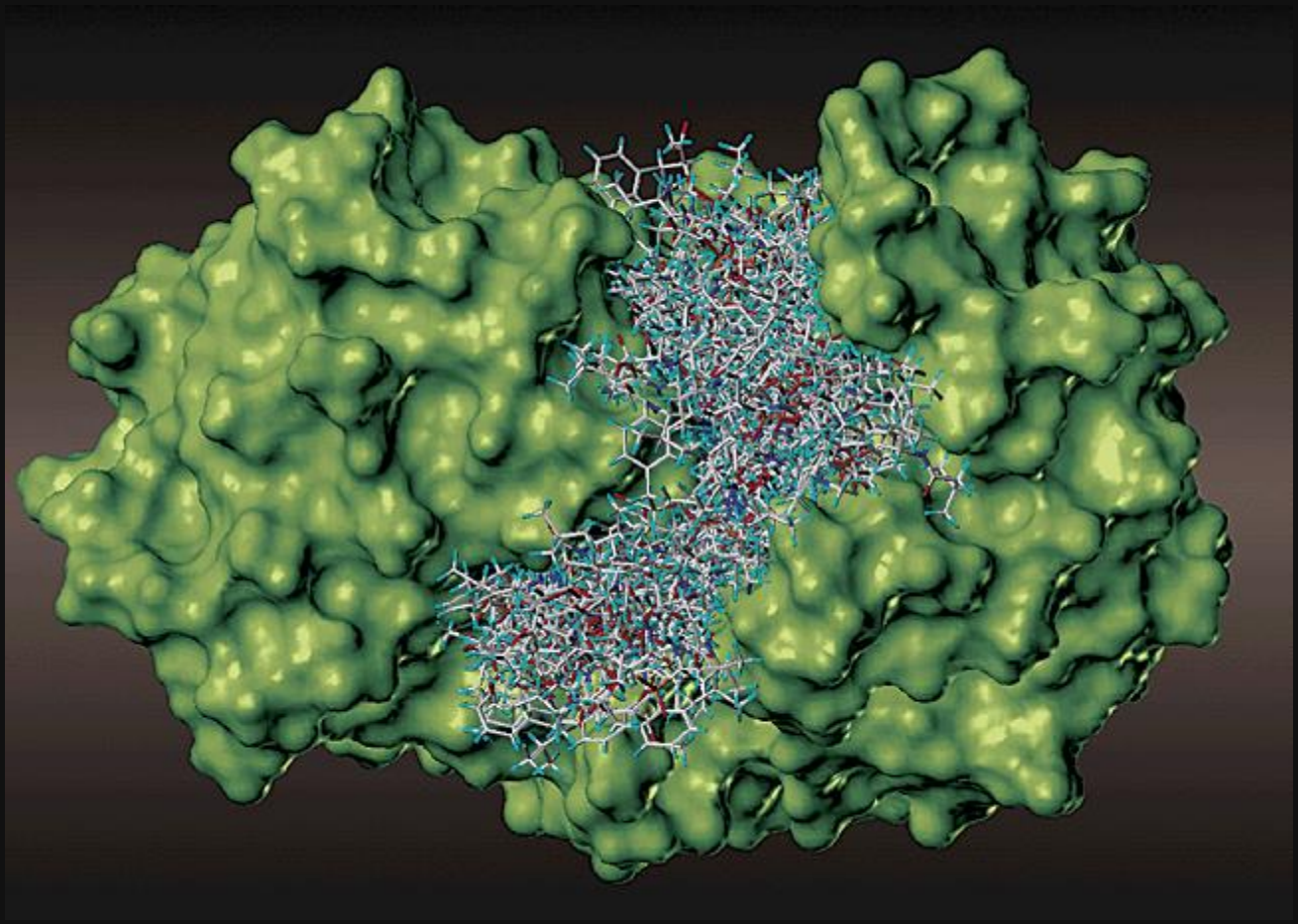
- Random, outcome varies
- Must repeat the search or perform more steps to improve chances of success
- Feasible for larger problems

Stochastic Search Methods

- Simulated Annealing (SA)*
- Evolutionary Algorithms (EA)
  - Genetic Algorithm (GA)*
- Others
  - Tabu Search (TS)
  - Particle Swarm Optimisation (PSO)
- Hybrid Global-Local Search Methods
  - Lamarckian GA (LGA)*

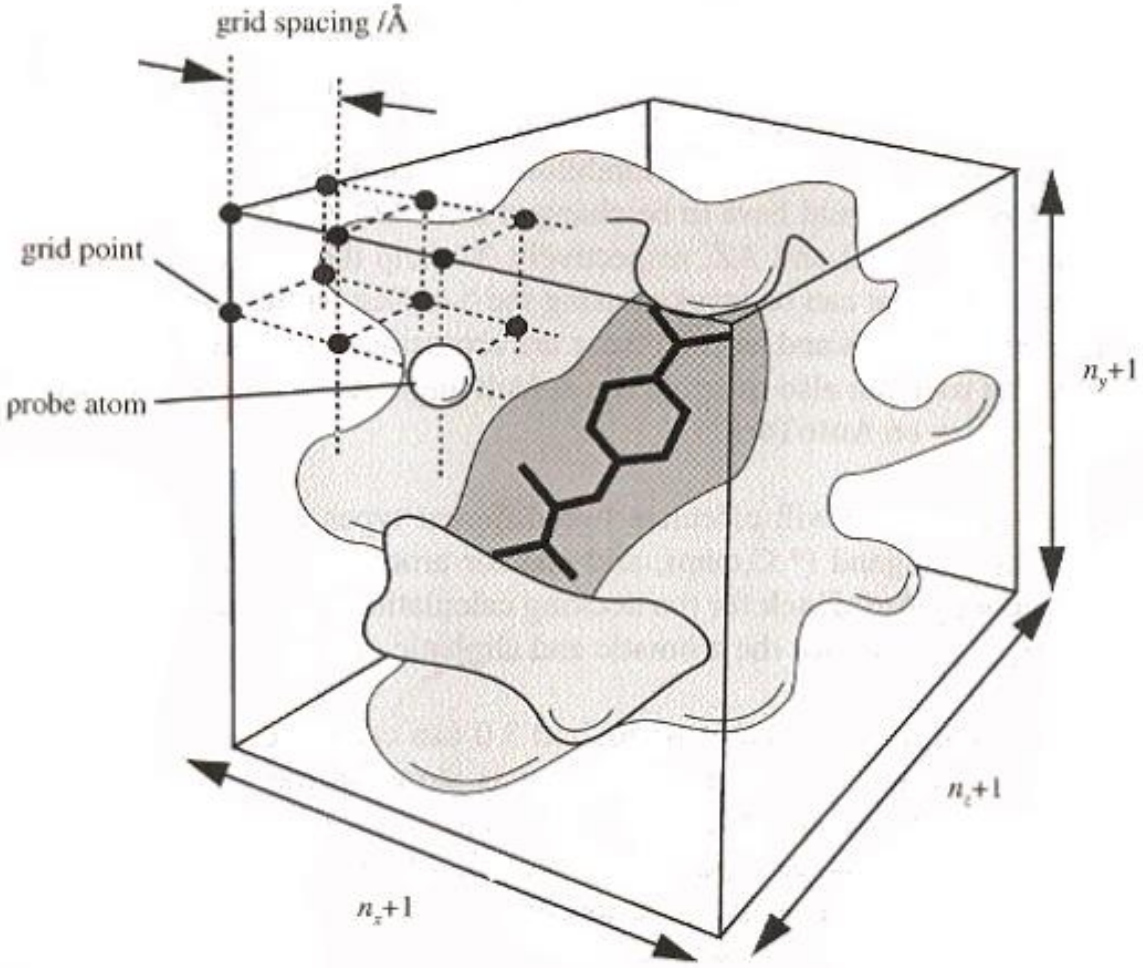
*Supported in AutoDock

Ligand Conformational Sampling By Autodock



Grid Maps

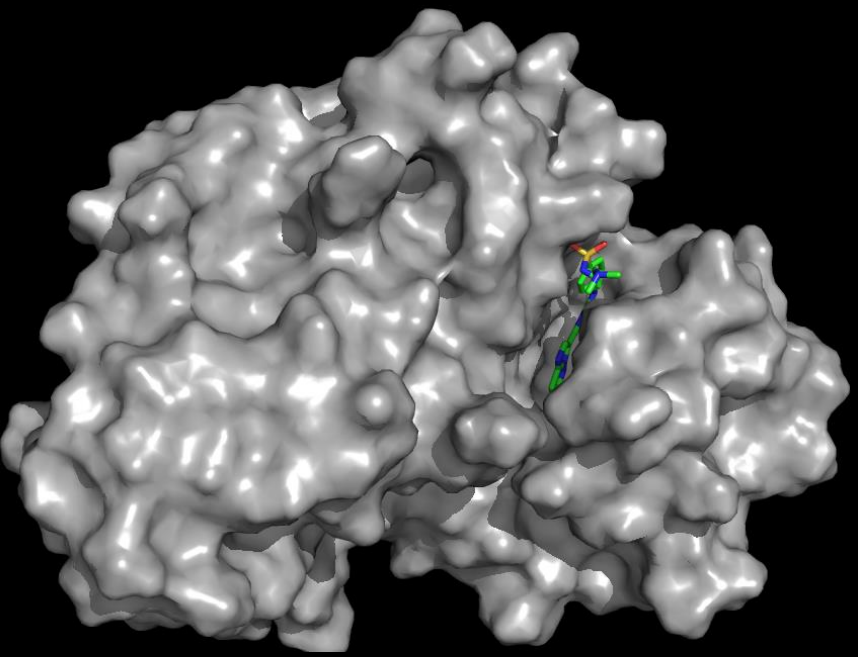
Each type of atom is placed at each individual grid point and the change in free energy is calculated

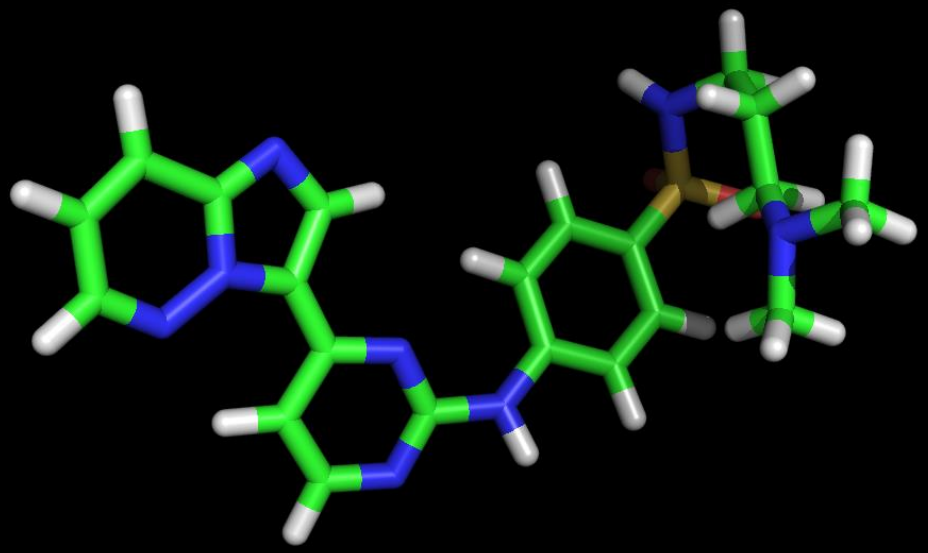


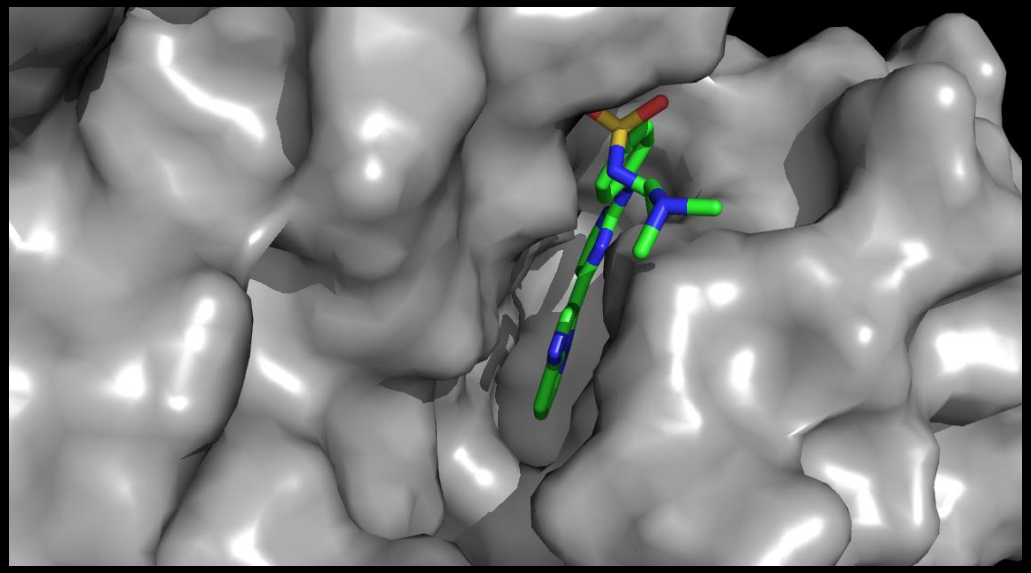
Using AutoDock: Step-by-Step

- Set up ligand PDBQT—using ADT's "Ligand" menu
- OPTIONAL: Set up flexible receptor PDBQT—using ADT's "Flexible Residues" menu
- Set up macromolecule & grid maps—using ADT's "Grid" menu
- Pre-compute AutoGrid maps for all atom types in your set of ligands—using "`autogrid4`"
- Perform dockings of ligand to target—using "`autodock4`", and in parallel if possible.
- Visualize AutoDock results—using ADT's "Analyze" menu
- Cluster dockings—using "analysis" DPF command in "autodock4" or ADT's "Analyze" menu for parallel docking results.

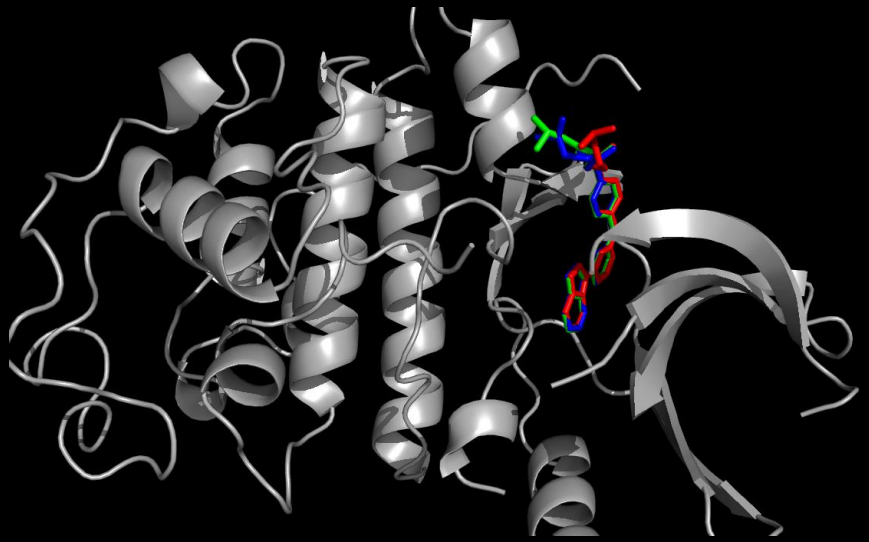
Example: Cyclin-Dependent Kinase (CDK2)

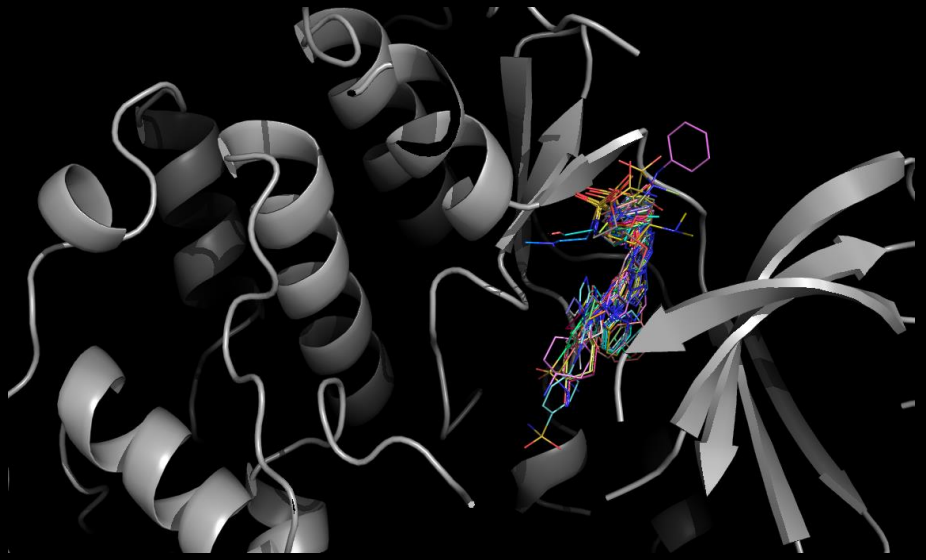






Example: Cyclin-Dependent Kinase (CDK2)



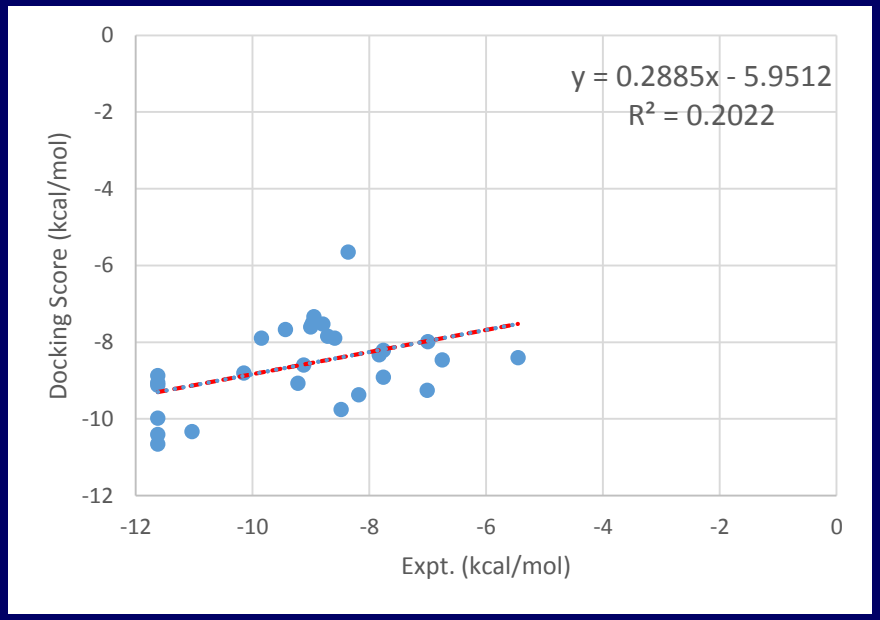


Red – crystal structure
Blue – Glide XP
Green – Glide SP

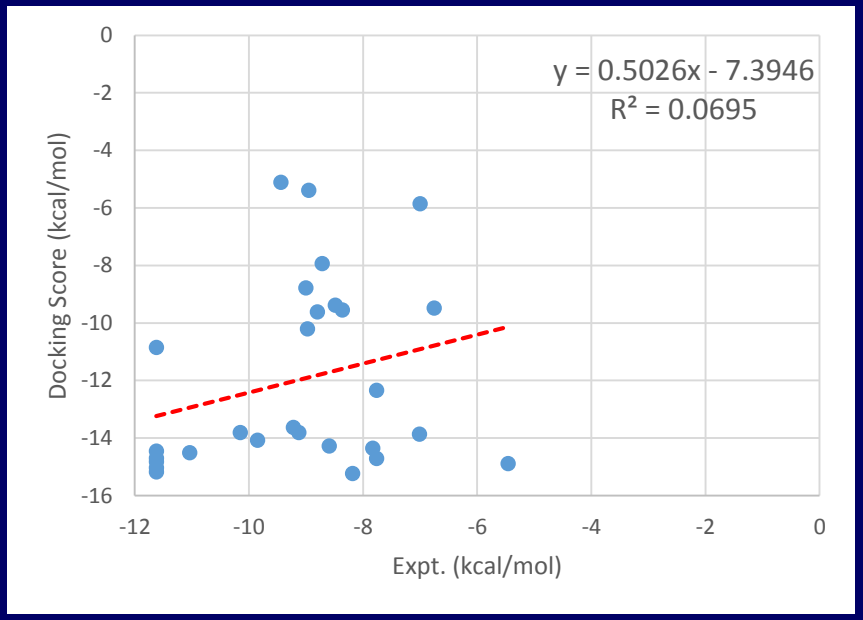
SP – 1.517 Å
XP – 1.310 Å

28 Inhibitors

Glide Docking Performance



Glide SP



Glide XP

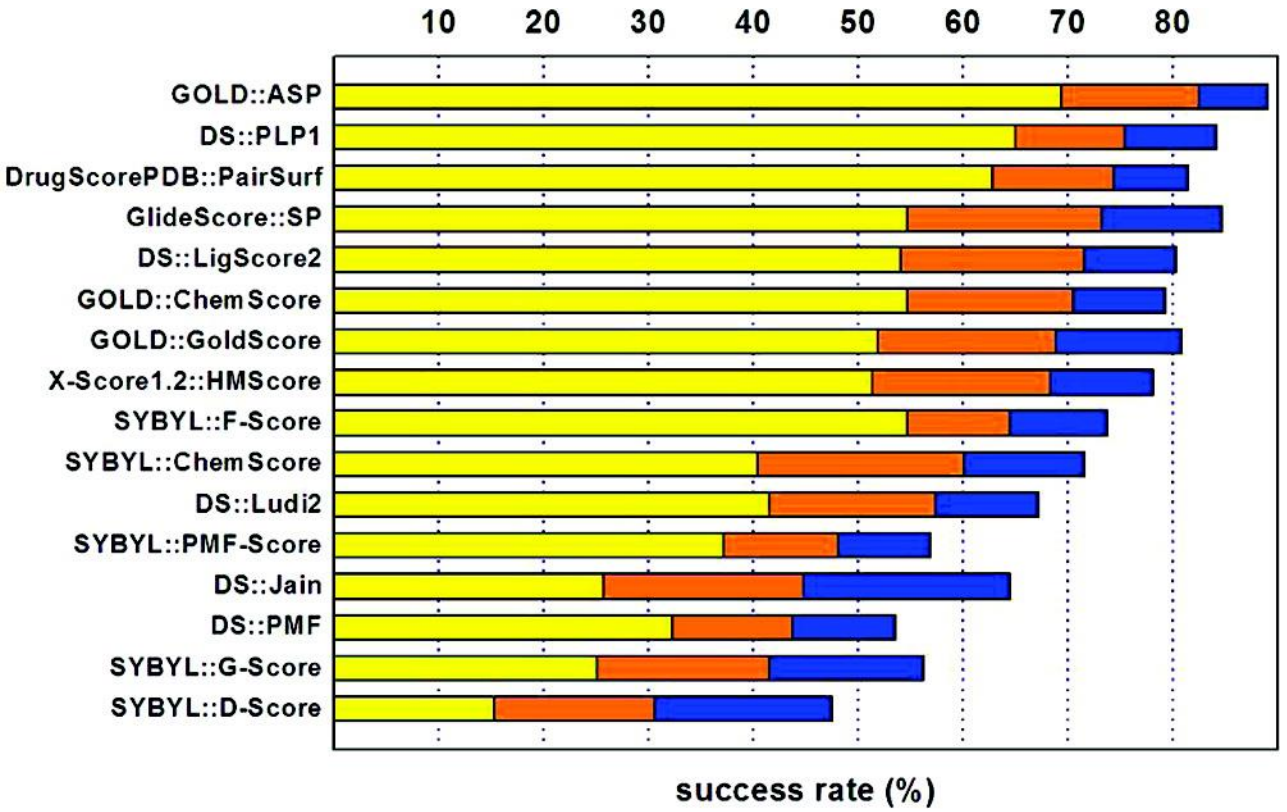
Critical Assessment of Docking Scoring Functions

1. **DOCK Power** –the ability to identify the true ligand binding pose among computer-generated decoys

2. **Ranking Power** - the ability to correctly rank different ligands bound to the same protein according to their binding affinities when the correct binding poses of these ligands are known.

3. **Scoring Power** - the ability of producing binding scores that are correlated, preferably in a linear manner, with experimentally measured binding affinities when protein−ligand complex structures are known

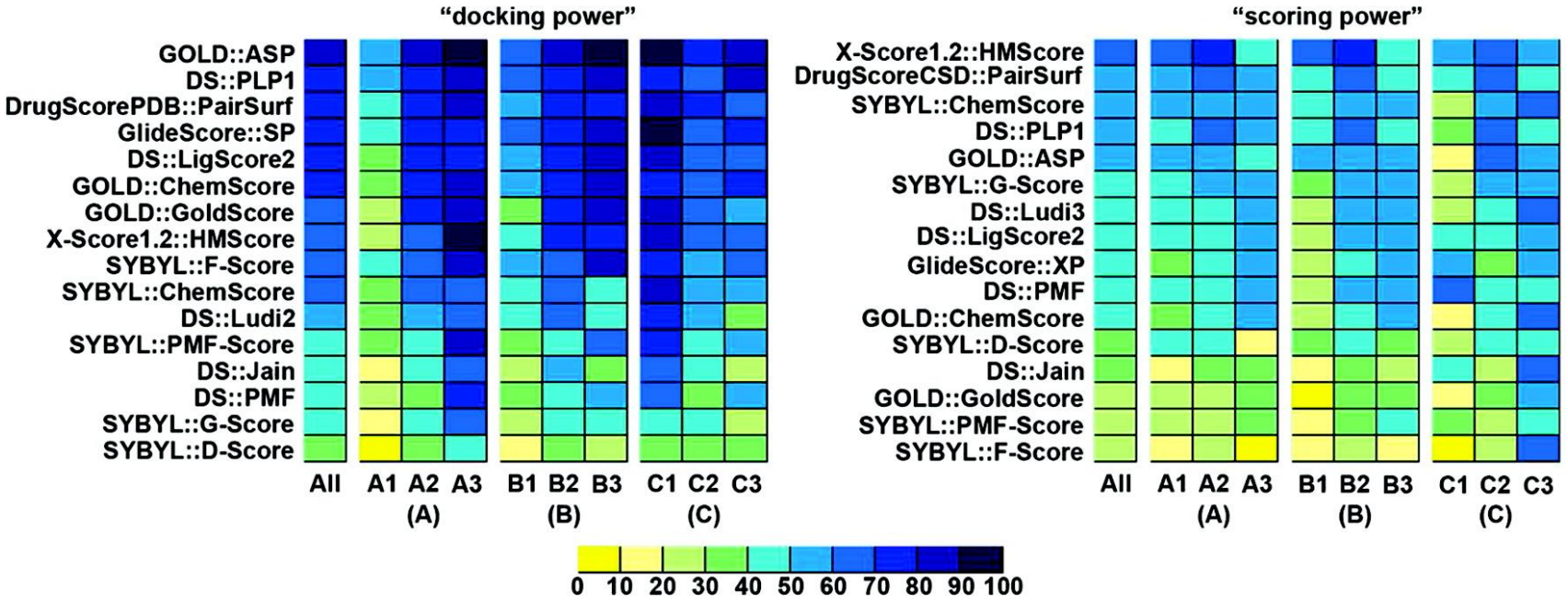
Performance of Reproducing Binding Poses



Comparison of the success rates of 16 scoring functions on the primary test set when the cutoff is rmsd < 1.0 Å (yellow bars), < 2.0 Å (orange bars), or < 3.0 Å (blue bars), respectively. The true ligand binding poses were included in the decoy sets in this test. Scoring functions are ranked by the success rates when the acceptance cutoff is rmsd < 2.0 Å.

Published in: Tiejun Cheng; Xun Li; Yan Li; Zhihai Liu; Renxiao Wang; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093.
DOI: 10.1021/ci9000053

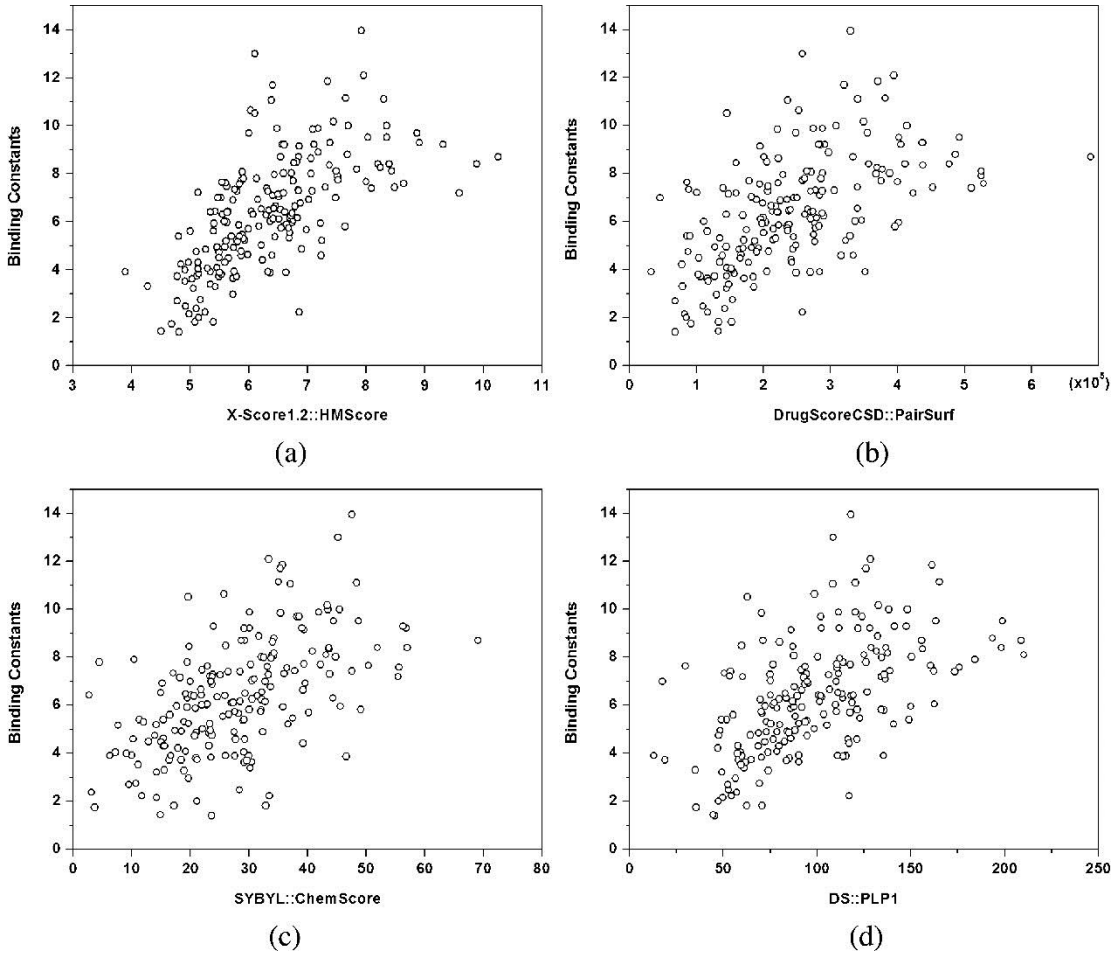
Docking Power And Scoring Power of 16 Scoring Functions



"Docking power" and "scoring power" of all 16 scoring functions on the subsets in the primary test set. Three sets of subsets were classified by (A) buried percentage of the solvent-accessible surface area of the ligand, (B) buried percentage of the molecular volume of the ligand, and (C) the hydrophobic index of the binding pocket. Here, scoring functions are ranked by their performance on the entire primary test set.

Published in: Tiejun Cheng; Xun Li; Yan Li; Zhihai Liu; Renxiao Wang; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093.
DOI: 10.1021/ci9000053

How Well Do Docking Scores Correlate With Measured Binding Constants



Correlations between the experimentally measured binding constants (in −log *K*d units) of the 195 protein−ligand complexes in the primary test set and the binding scores computed by (a) X-Score::HMScore (*R* = 0.644), (b) DrugScoreCSD::PairSurf (*R* = 0.569), (c) SYBYL::ChemScore (*R* = 0.555), and (d) DS::PLP1 (*R* = 0.545).

Published in: Tiejun Cheng; Xun Li; Yan Li; Zhihai Liu; Renxiao Wang; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093.
DOI: 10.1021/ci9000053

CAPRI

- Just like the CASP competition in the protein folding field, there is a bi-annual competition capped CAPRI: the Critical Assessment of PRedicted Interactions

- J. Janin et al.
  "CAPRI: a Critical Assessment of Predicted Interactions"
  Proteins (2003) 52:2-9

- Mendez et al.
  "Assessment of blind predictions of protein-protein interactions: Current status of docking methods"
  Proteins (2003) 52:51-67

Consensus Score

Enrichments	Approach	Single methods	Consensus method
Hit rates (%)	Intersection using three scoring functions	3	18
Hit rates (%)	Intersection using three scoring functions	10	65–70
Top compounds containing all actives (%)	Voting using three scoring functions	20	8.4

Drug Discovery Today, **2006**, *11*, 1359-6446

Consensus Score - to be continued

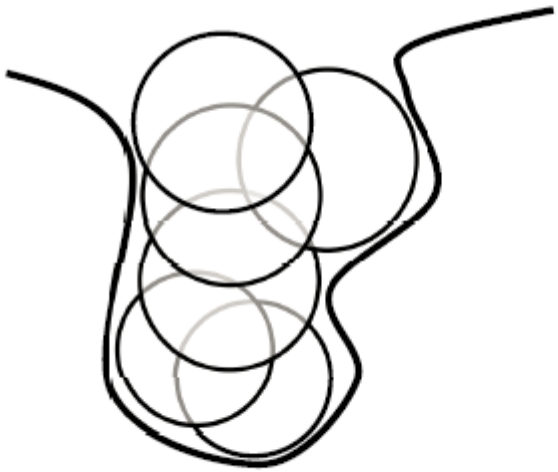
Docking Pose	Approach	Single methods	Consensus method
Ligands with top docked pose within 2Å of the crystal structure (%)	ConsDock	39–56	60
Ligands with top docked pose within 3Å of the crystal structure (%)	Average rank using three functions	66–76	80–84

Consensus Score - to be continued

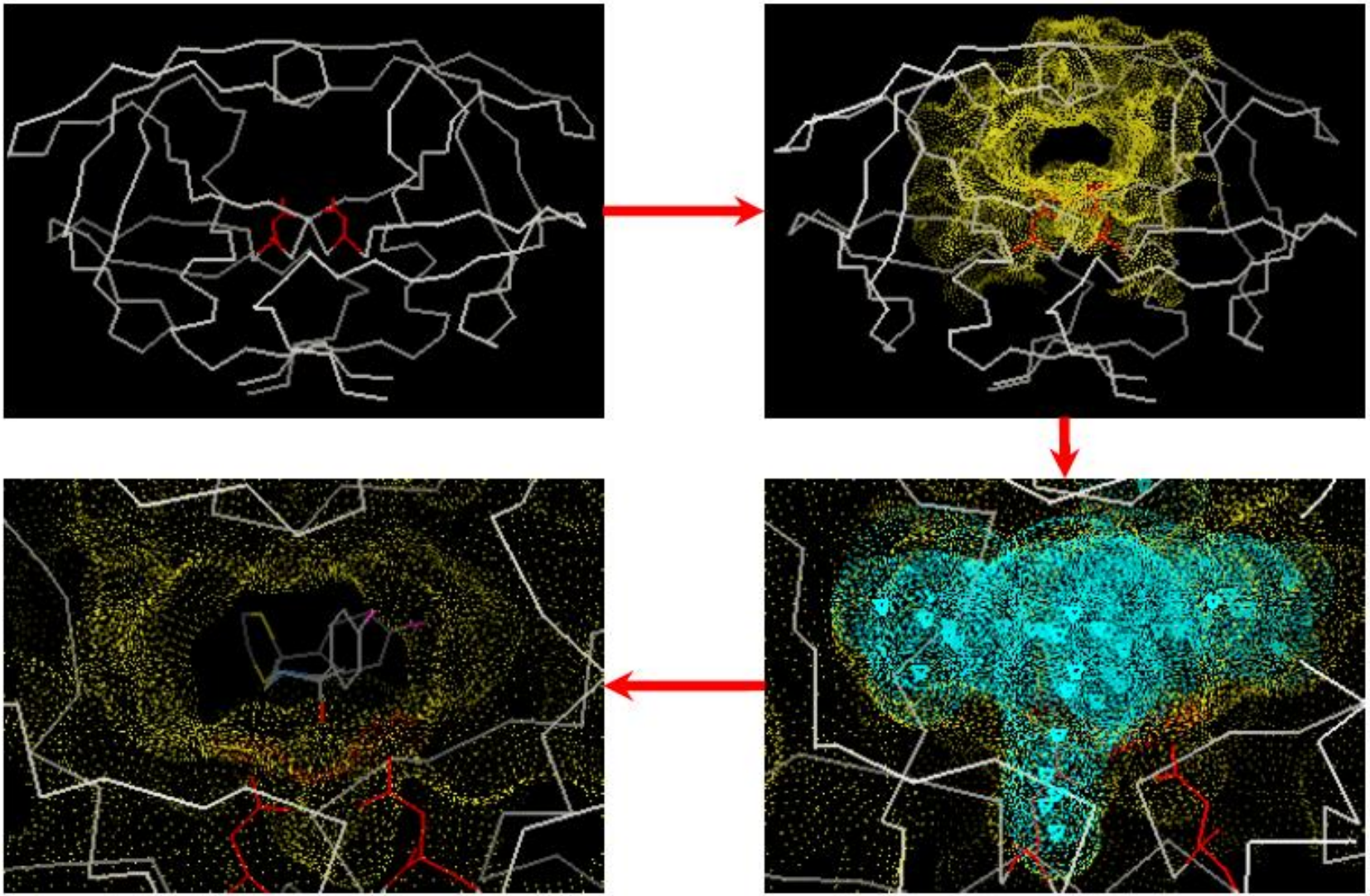
Docking Scores	Approach	Single methods	Consensus method
Rank correlation of predicted and experimental binding energies	Sum-rank	0.13–0.92	0.54–0.85
Rank correlation of predicted and experimental binding energies	CScore	0.13–0.92	0.60–0.86
Correlation (r^2) between predicted and experimental binding energies	Average rank	0.16–0.32	0.34
Correlation (r^2) between predicted and experimental binding energies	PLS	0.10–0.56	0.68
RMS error (kJ/mol) between predicted and experimental binding energies	Average rank	3.00–4.93	2.49

DOCK

- The DOCK program is from the Kuntz group at UCSF
- It was the first docking program developed in 1982
- It represents the (negative image of the) binding site as a collection of overlapping spheres



DOCK



http://dock.compbio.ucsf.edu

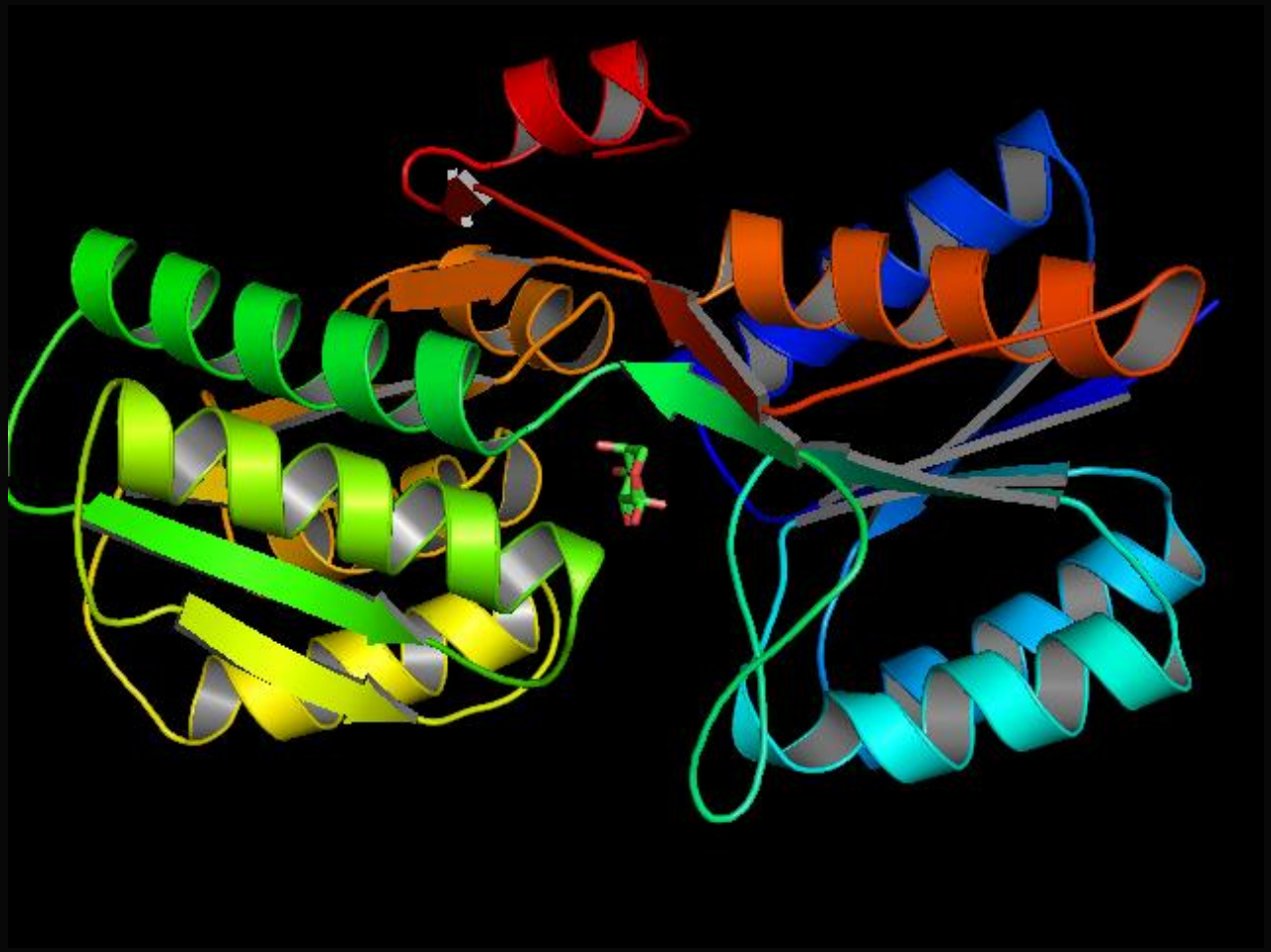
Case Study Using DOCK

- Structural preparation
  download pdb file – 1ABE from www.pdb.org

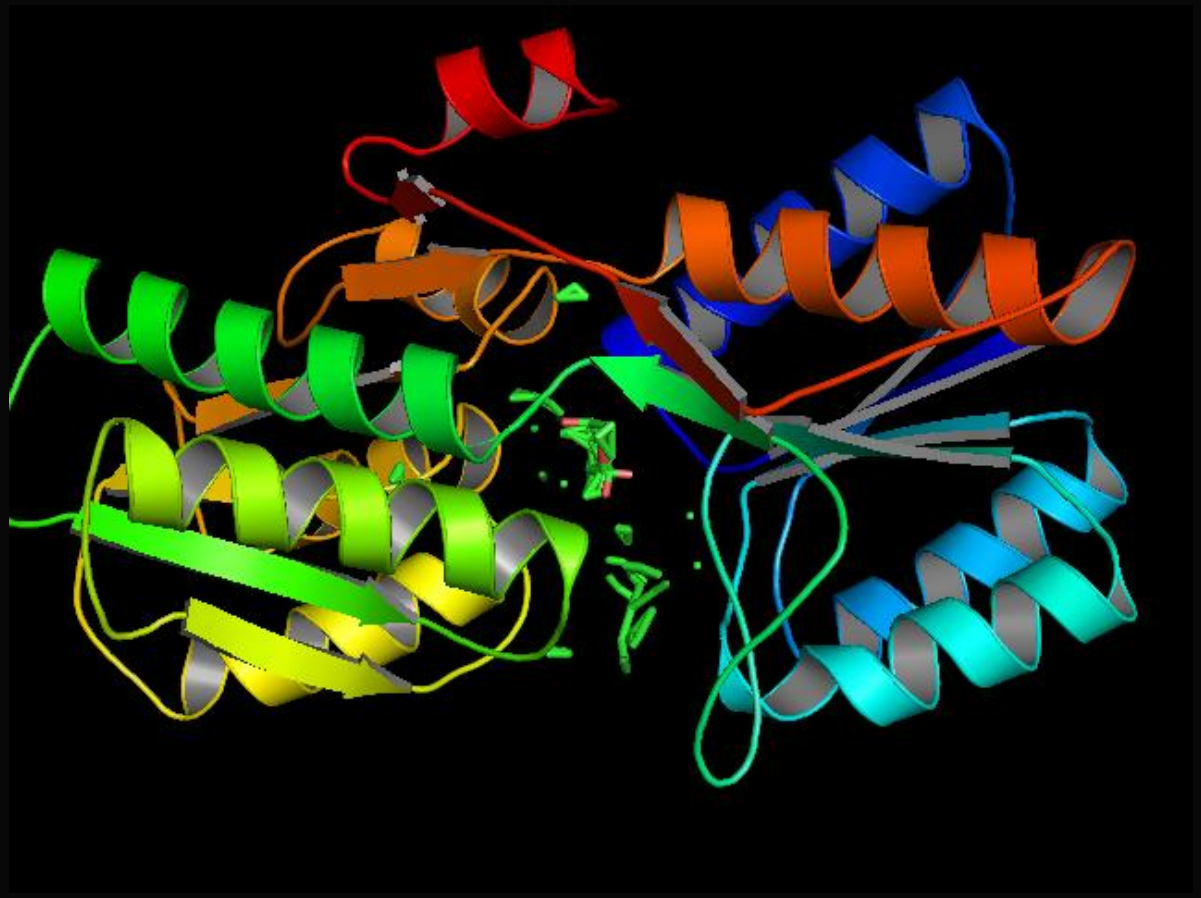
  rec.pdb – add hydrogen, load AMBER charges
  rec_noH.pdb - generate SAS using the *dms* program
  lig.pdb – use *antechamber* or *sybyl* to generate Gasteiger charges

- Identify binding site
  run *sphgen* to generate “negative spheres” that complementarily match protein surface
  Manually select a sphere cluster that best describes the binding site
  select_spheres.sph

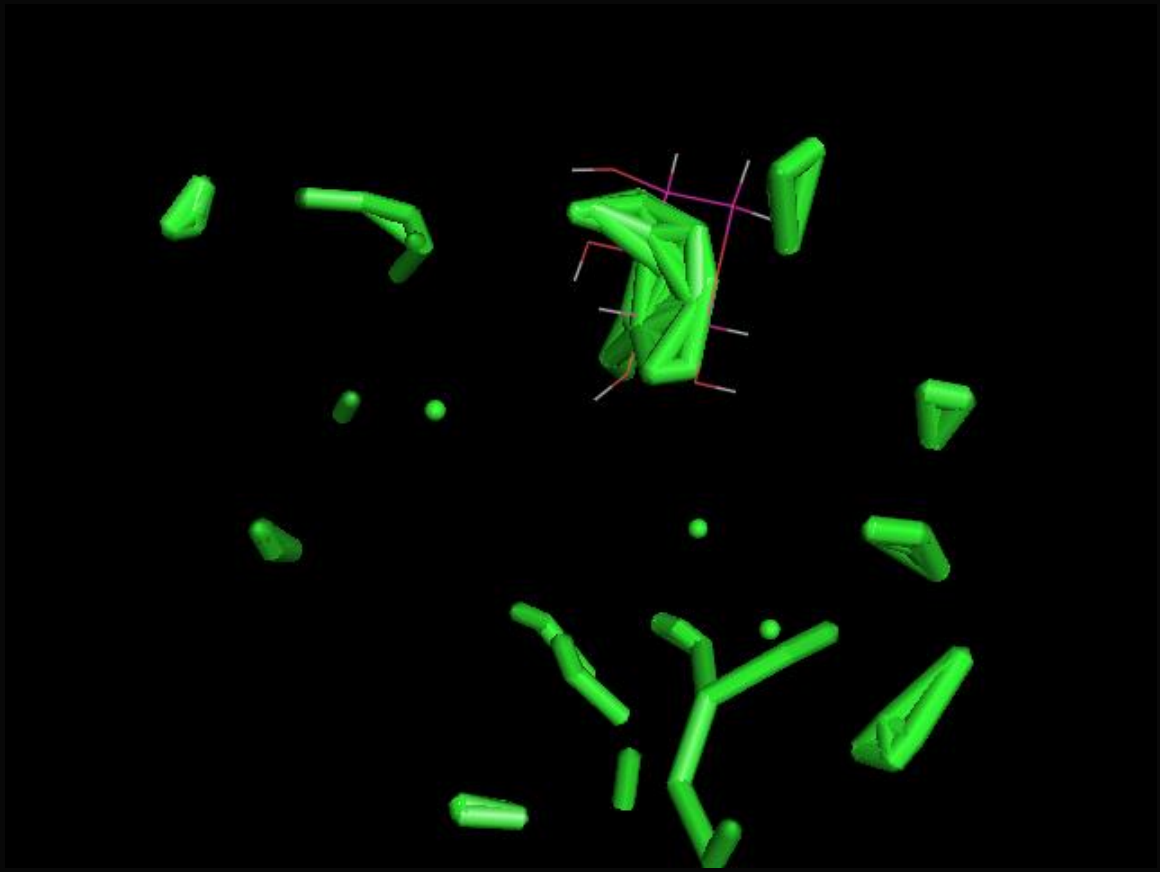
Case Study Using DOCK6



Case Study Using DOCK6

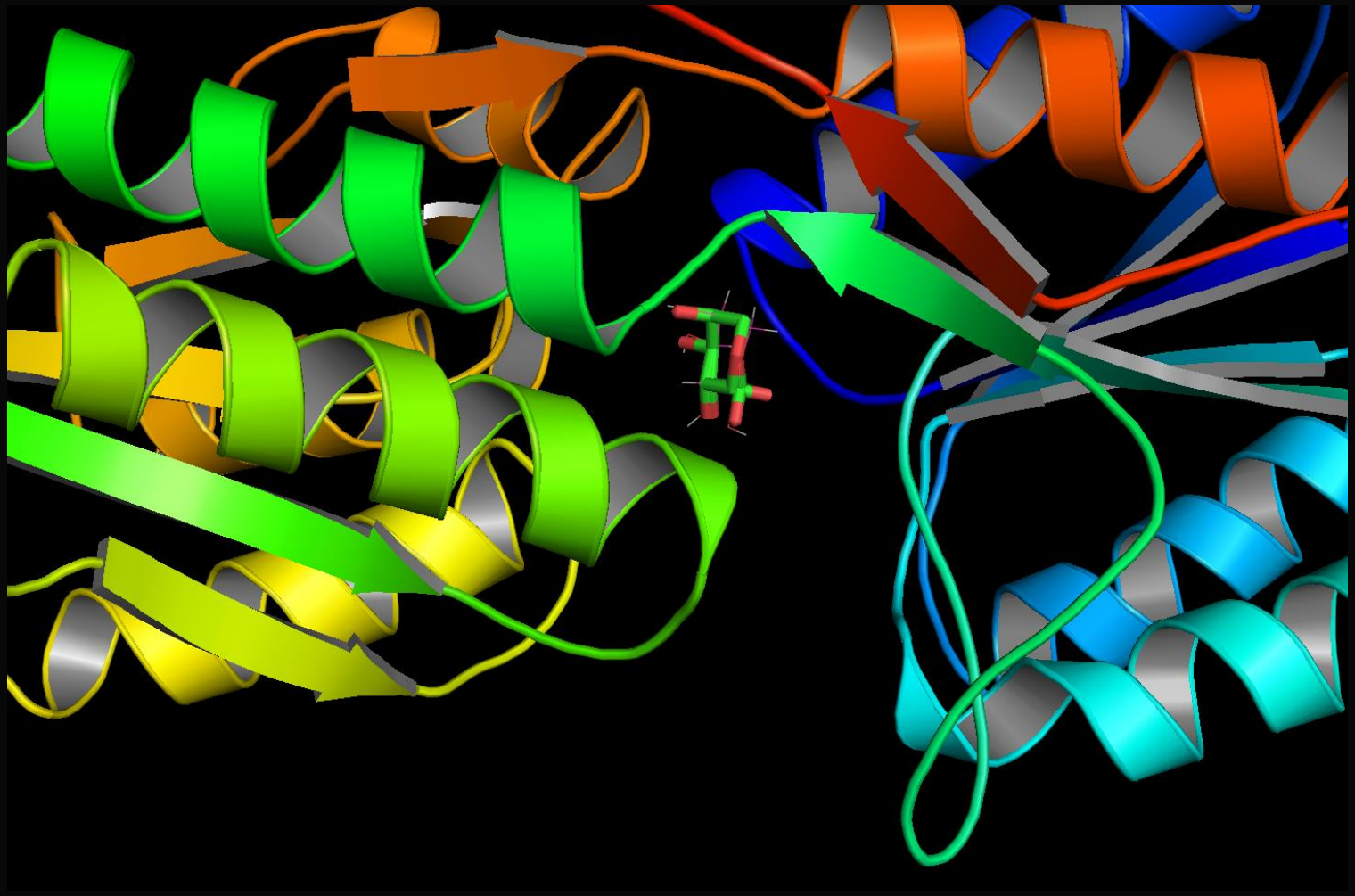


Case Study Using DOCK6

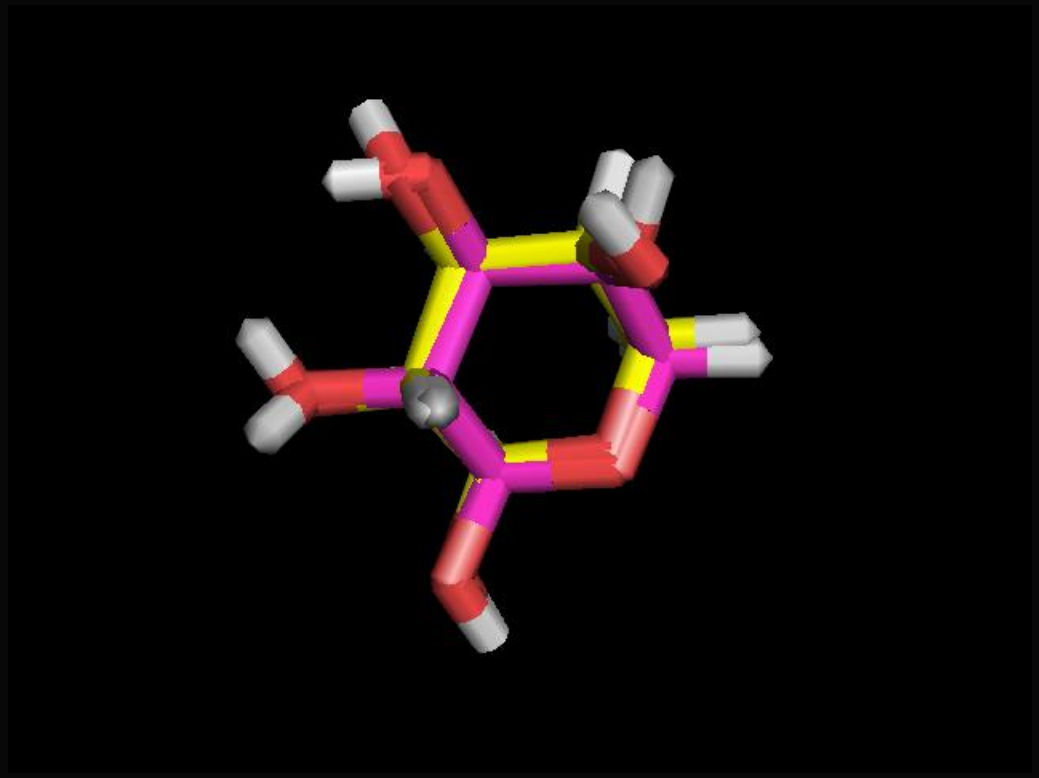


Green: Negative Spheres where ligand atoms may occupy

Case Study Using DOCK6



Case Study Using DOCK6



Case Study Using DOCK6 – to be continued

- **Calculate grid potentials**
  calculate the grid potentials around the selected spheres

- **Perform docking**
  Rigid docking
  Grid Score = -28.34
  vdw: -22.26
  es: -6.07

  Flexible docking
  Grid Score = -33.21
  vdw: -22.11
  es: -11.09

Lab Section

- ZINC (zinc.docking.org)
- OpenBabel (openbabel.org)
- AutoDock (autodock.scripps.edu, mgltools.scripps.edu)

Assignment and project are posted on course webpage (https://Mulan.swmed.edu)