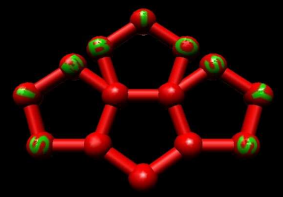
Algorithmic and mathematical challenges in protein-ligand docking and scoring



http://www.simbiosys.ca/

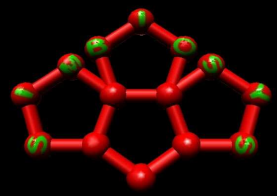
Zsolt Zsoldos

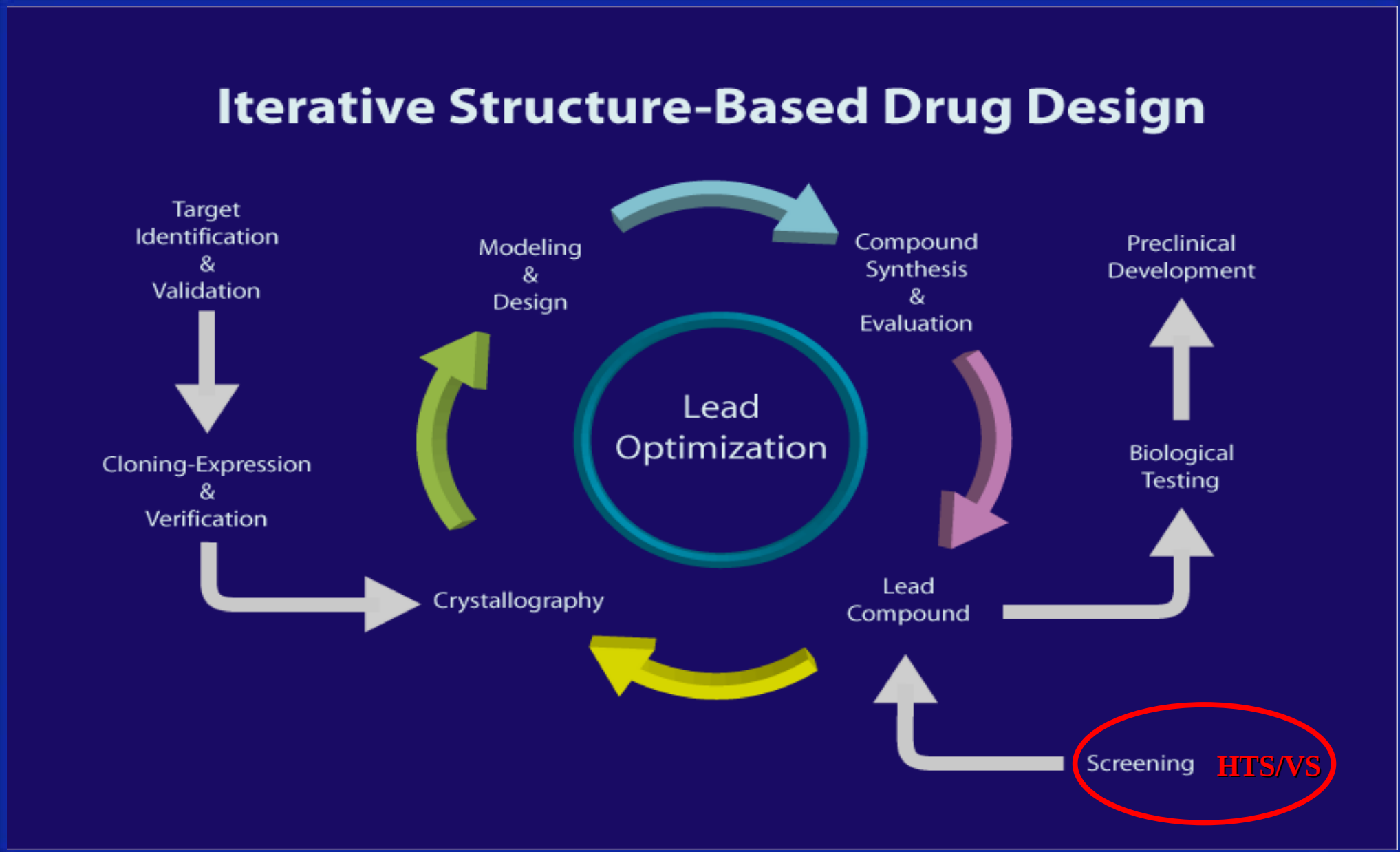
SimBioSys Inc., © 2010

Contents:

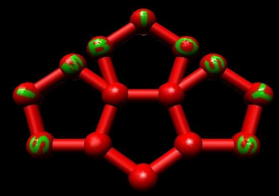
- Introduction: Virtual Screening methods in computational drug design
- The exhaustive docking algorithm of eHiTS
- A new statistically derived empirical scoring function
- LASSO: integrated VHTS filter
- Results

2. Structure Based Drug Design Cycle



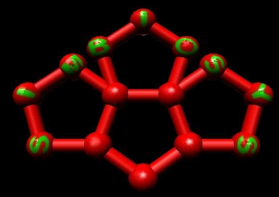


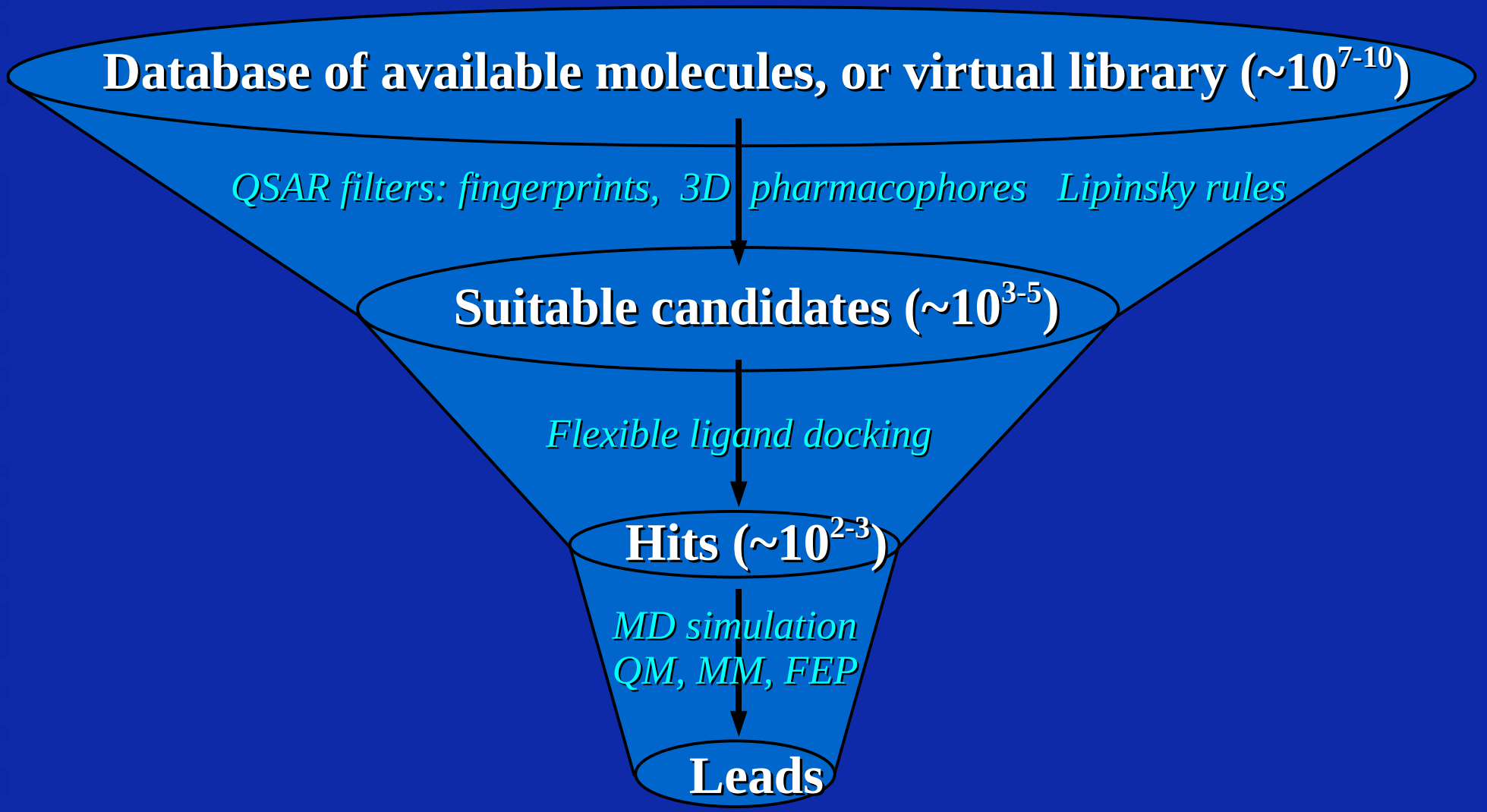
3. Comparison of 2 VS approaches



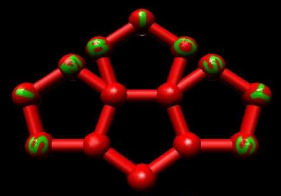
	Ligand based	Docking
Speed	fast (milliseconds)	slow (minutes)
Input	set of known actives	3D protein structure
Output Score	Similarity / arbitrary	Binding energy
Structure	None / overlay	Conformation, pose

4. Virtual Screening Funnel

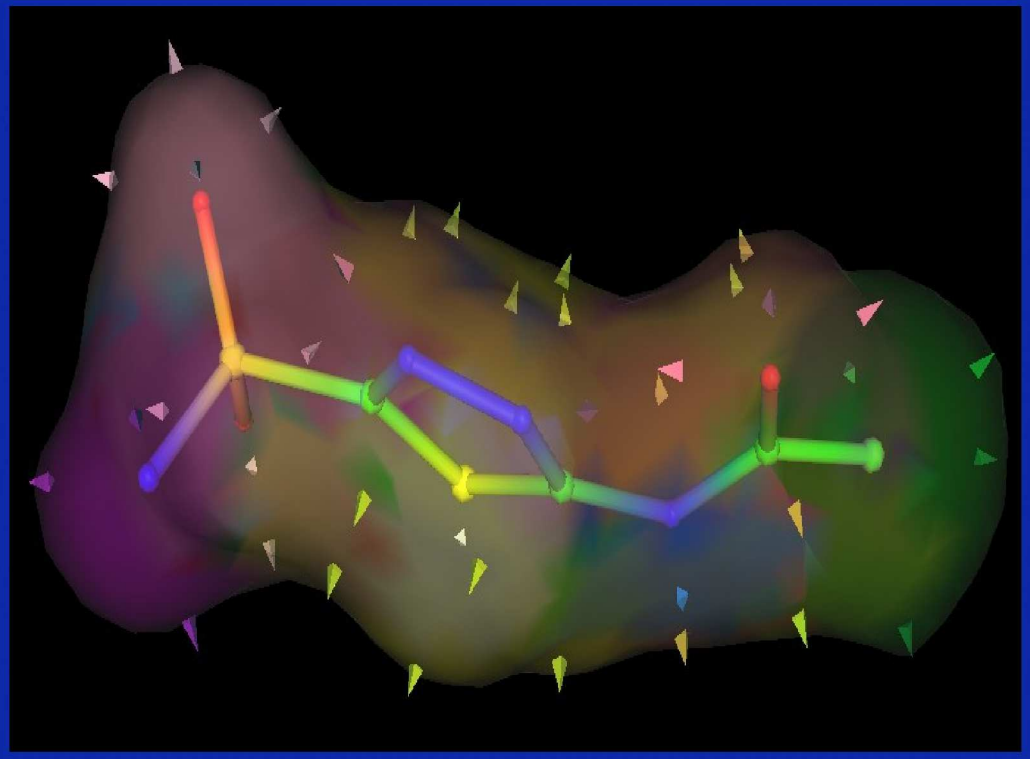




5. Ligand based screening methods

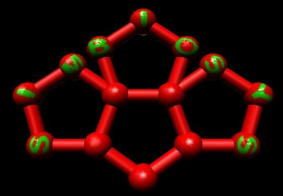


- Simple filters, e.g. Lipinsky rules of 5 (HB donors/acceptors, Log P, MW)
- 2D Fingerprints: existence of FGs, chain / substructure patterns
- 3D Pharmacophore match
- Feature tree matching
- 3D Shape matching, overlay
- Surface property matching

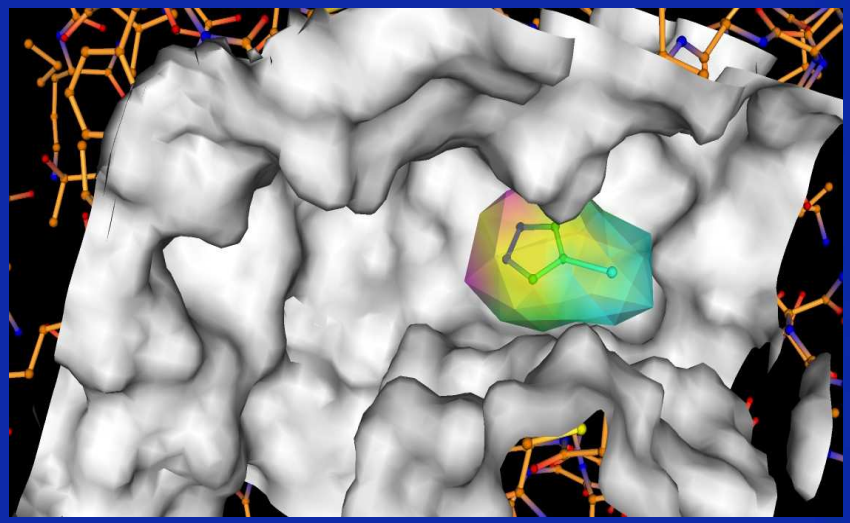


- CoMFA and other 3D grid mapping methods

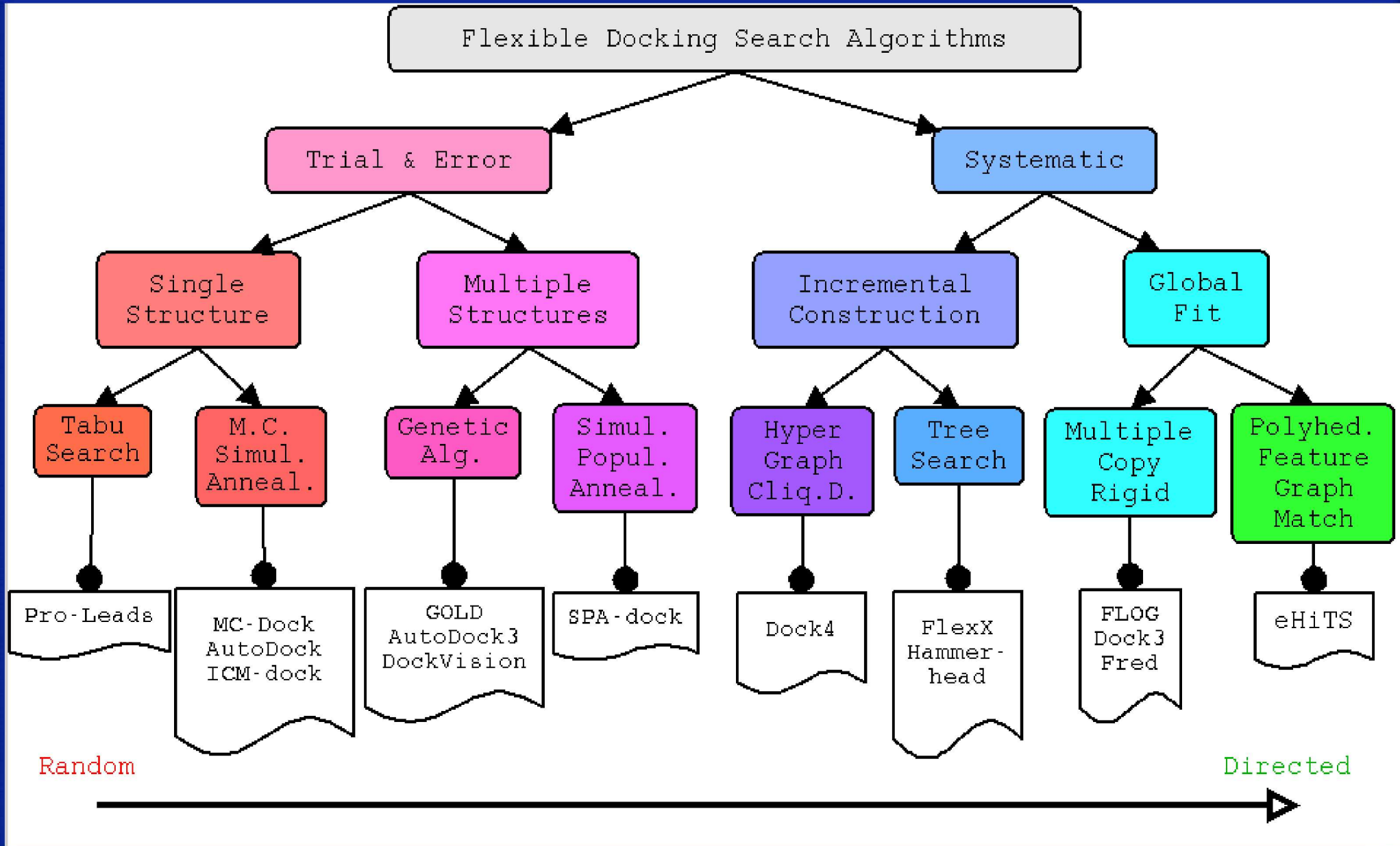
6. The docking approach



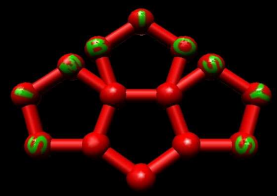
- 3D structure of target macro-molecule is known (X-ray, NMR)
  it has a binding site of certain shape (limited flexibility)
- A set of (flexible) small molecule ligands are given as candidates
- We need to fit the ligands to the cavity
- Interactions should be favourable
  (electrostatic, lipophil, H-bond, π-π)
- Estimate the binding free energy
- Output best conformation / pose



7. Yet Another Docking Algorithm, YADA, yada, yada...

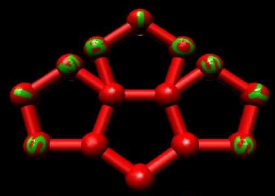


8. Contents: the eHiTS docking method

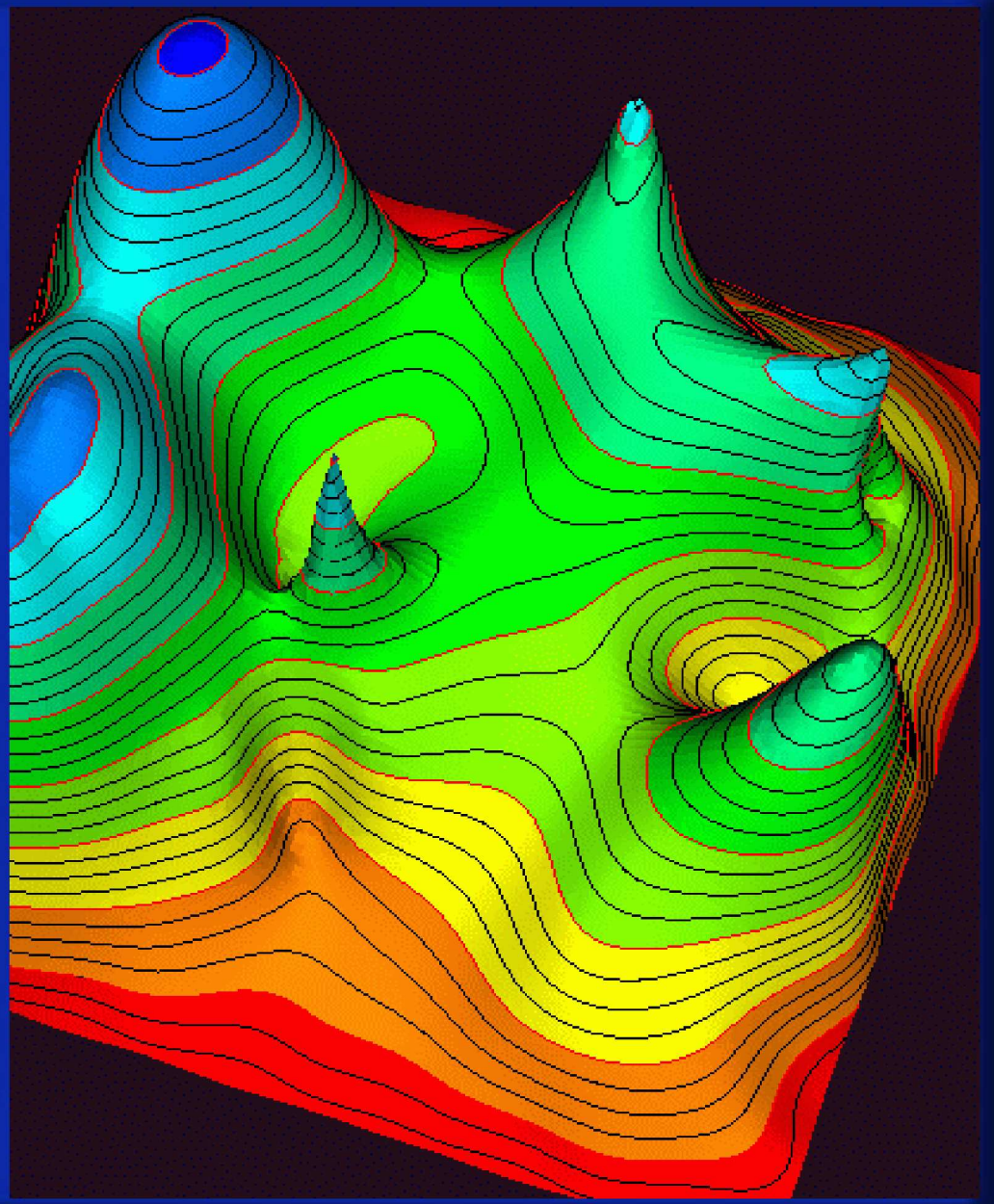


- **Problem size and sampling considerations**
  - What kind of exhaustive search do we want ?
  - Pose sampling resolution required
  - Statistics on bound ligand conformations
  - Search space size of exhaustive docking
- The docking algorithm of the eHiTS software
- A new statistically derived empirical scoring function
- Validation results
- LASSO: integrated VHTS filter

9. What is an exhaustive search ?

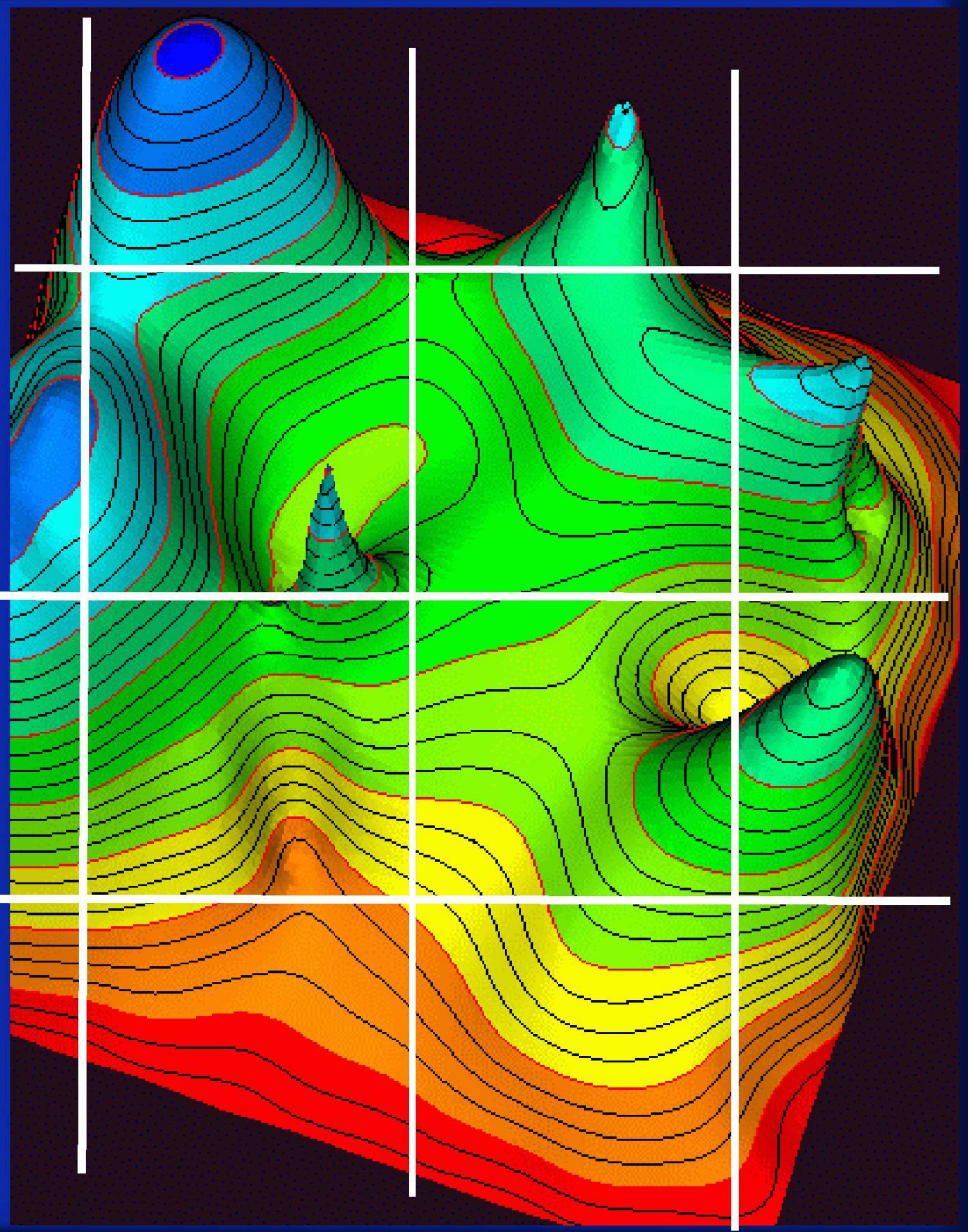


- Systematic
  - no random elements



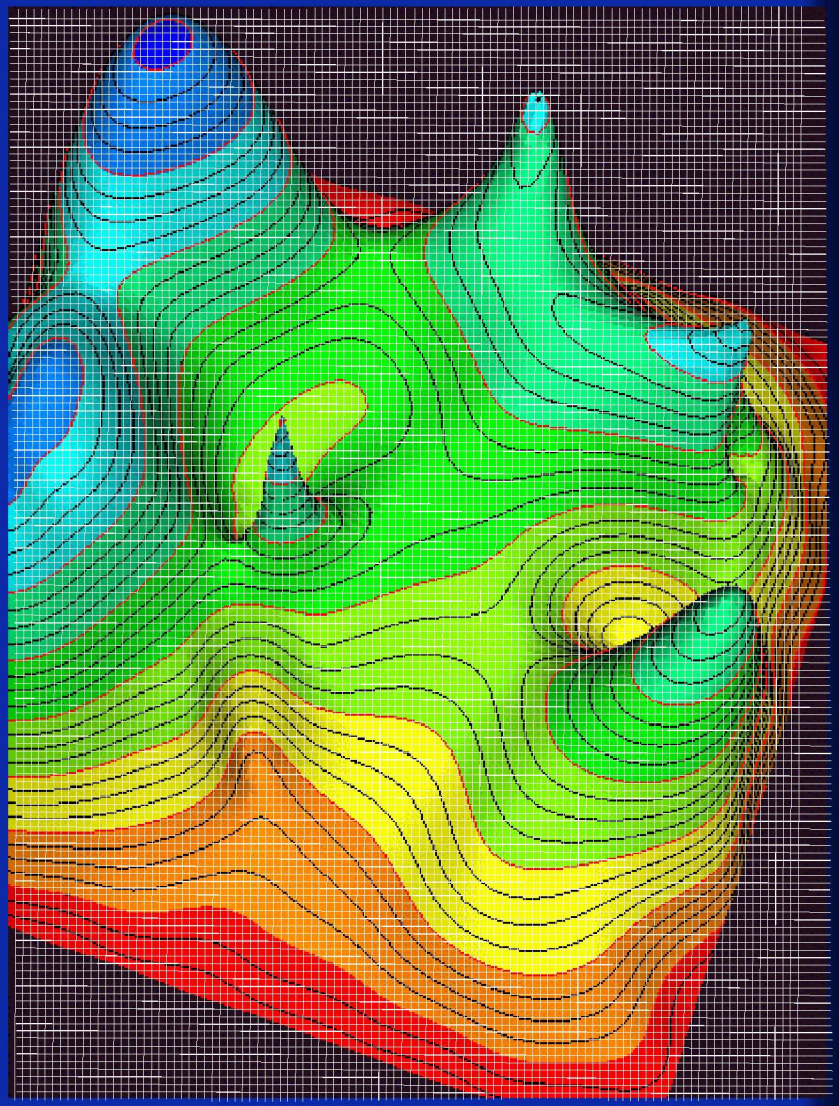
10. Systematic, but not exhaustive

- Systematic
  - no random elements
- The sampling resolution is not sufficient to find all the local maxima



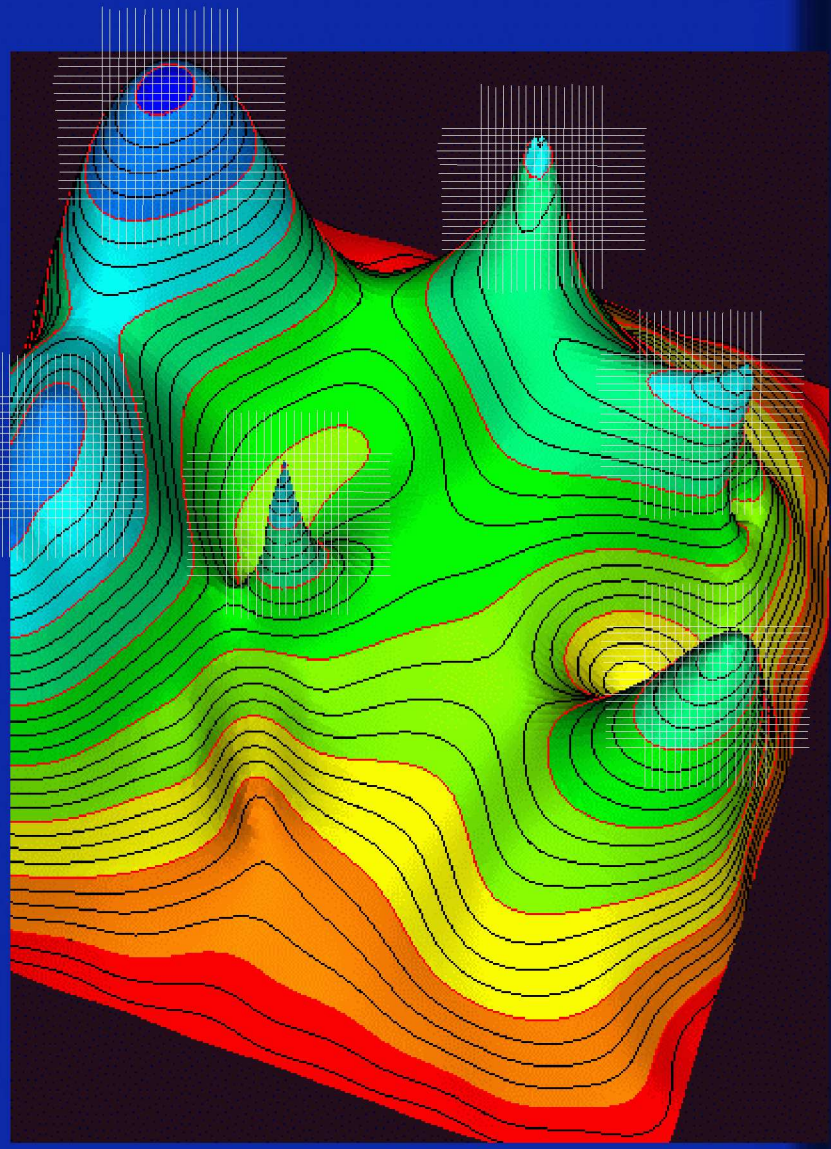
11. Brute-force exhaustive search

- Systematic
  - no random elements
- Sufficient sampling resolution
  - depends on goal function
- The whole search space is covered, not practical for difficult problems

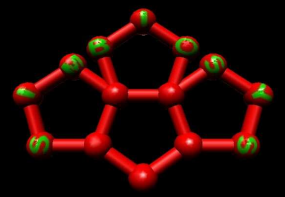


12. Intelligent exhaustive search

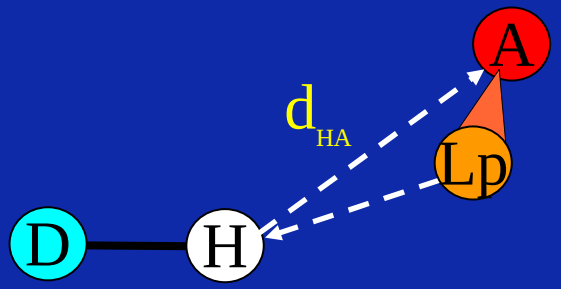
- Systematic
  - no random elements
- Sufficient sampling resolution
  - depends on goal function
- Solution space coverage
  - no need to sample the whole search space



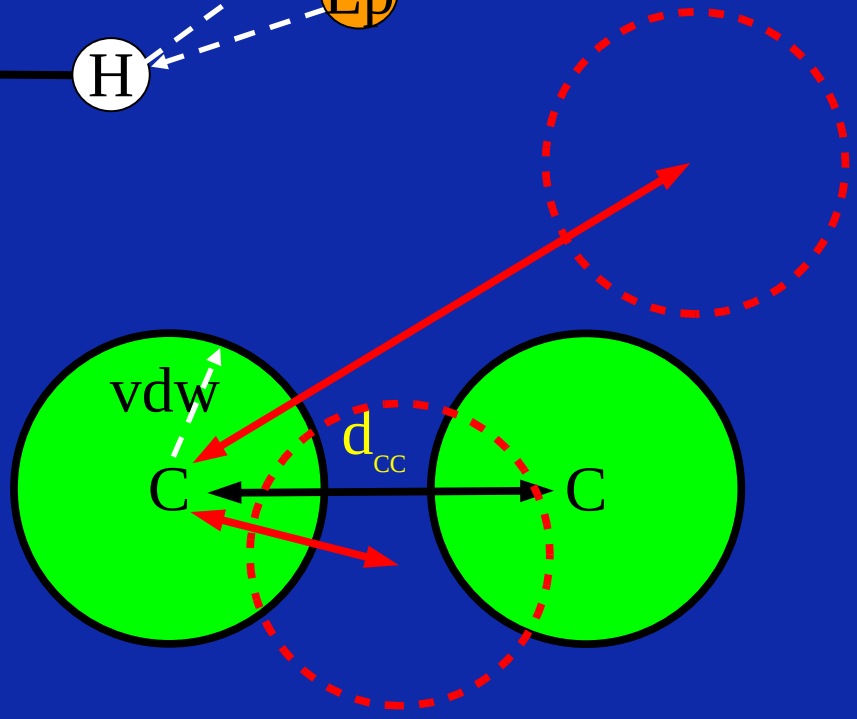
13. Pose sampling requirements of good interaction geometry for scoring



- H-bond geometry
  H-acceptor distance range
  1.6Å to 2.2Å, i.e. 1.9Å±0.3Å

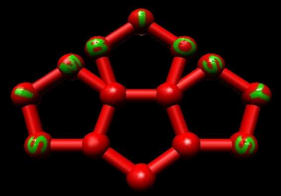


- Hydrophobic contact
  carbon-carbon distance range
  3.2Å to 4.2Å, i.e. 3.7Å±0.5Å

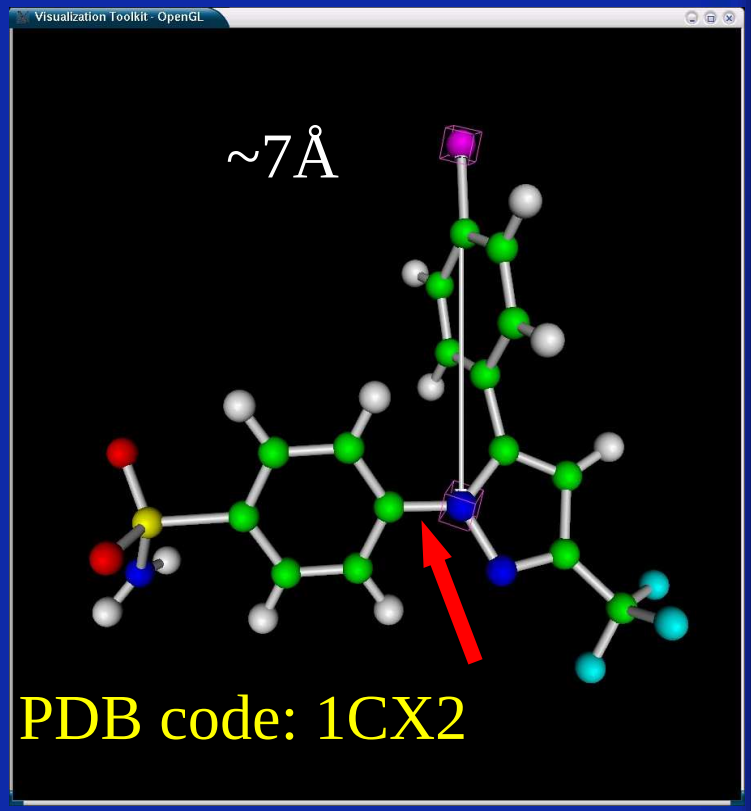


- discretization must be fine enough to sample *atom* placements every **0.5Å** or less

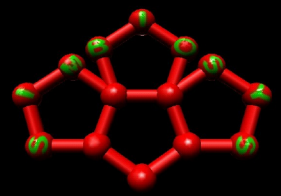
14. Orientation and dihedral angle precision required



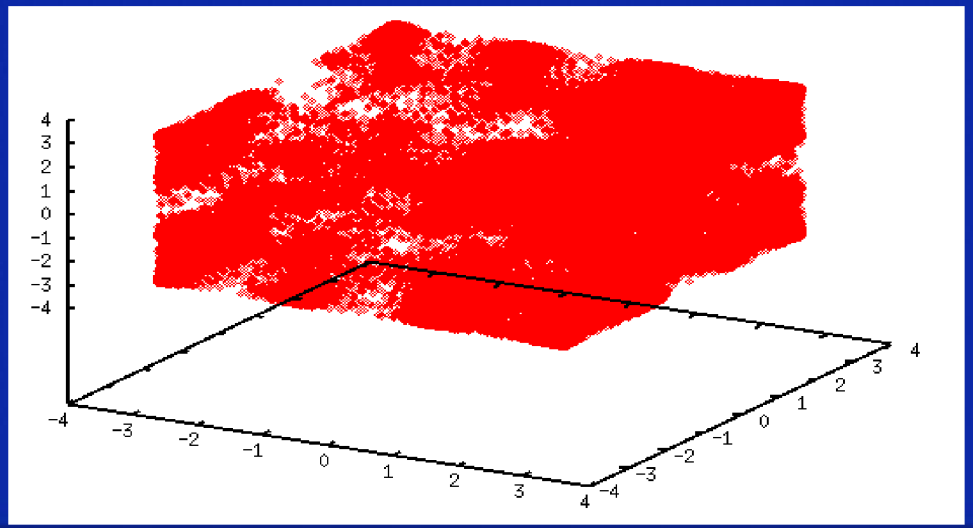
- Goal: sample atom placement every 0.5Å or finer
  - Drug-like molecules can have heavy atoms at 7Å distance from a rotation axis (see figure)
  - Simple trigonometric calculation: tangential movement of 0.5Å is caused by rotation of about 5° at a rotation radius of 7Å
  - Consequence: orientation and dihedral sampling *must be* finer than **5°**

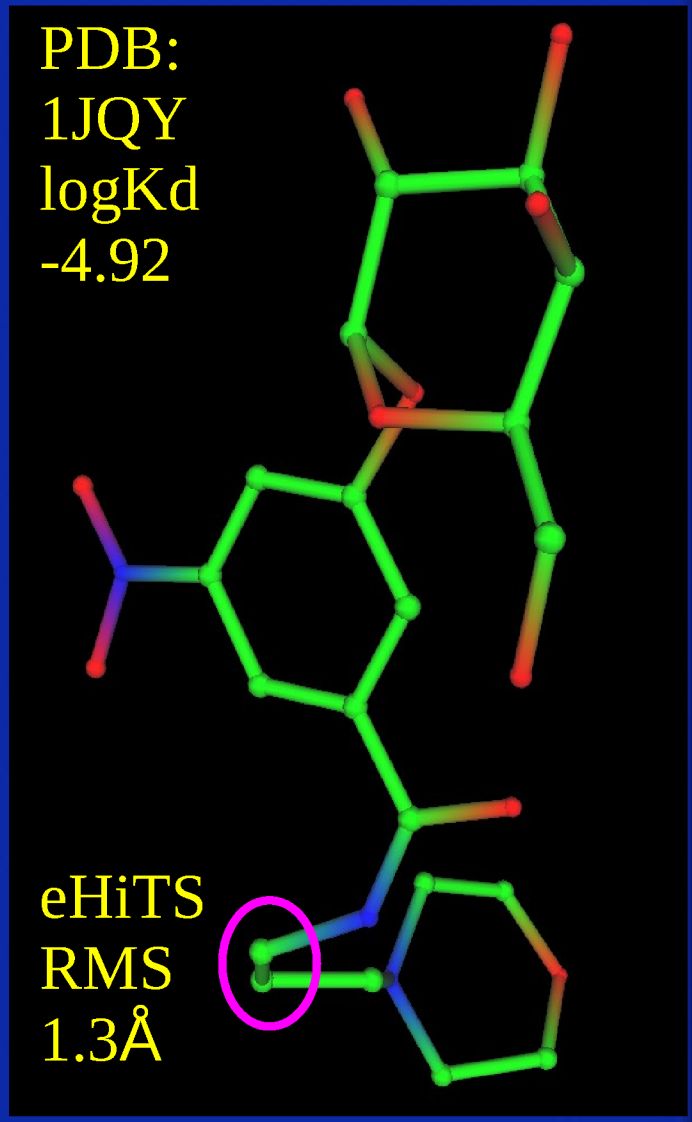


15. Statistics on bound ligand conformations (~5000 PDB <2.5Å)

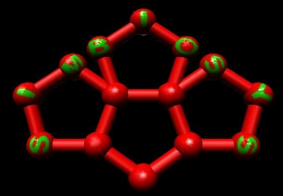


- High energy, strained torsions appear with high frequency (37%)
- Sampling every 60° for each rotatable bonds miss 97% of X-ray conformations by more than 5° error
- Dihedral angles are fully scattered and fill the whole range



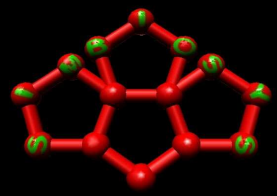


16. Search space size for exhaustive flexible ligand docking



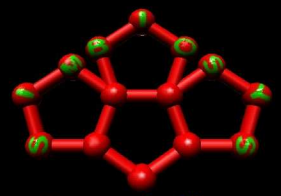
- Number of poses to examine with sampling defined:
  Translations(0.5Å) * Rotations(5°) * Dihedrals(5°) =
  $20^3 * 72^3 * 72^n$
  for n=6 rot.bonds => **$2*10^{20}$ poses *per ligand***
- Brute force evaluation 2000/s => 3 billion years
- Stochastic methods can only explore a tiny fraction of this space with no driving force towards coverage
- Comparative evaluation of 11 scoring functions for molecular docking
  Renxiao Wang, Yipin Lu, and Shaomeng Wang,
  J.Med.Chem. 2003, **46**, 2287-2303
  - concludes: greatest problem is pose sampling

17. Contents: the docking algorithm

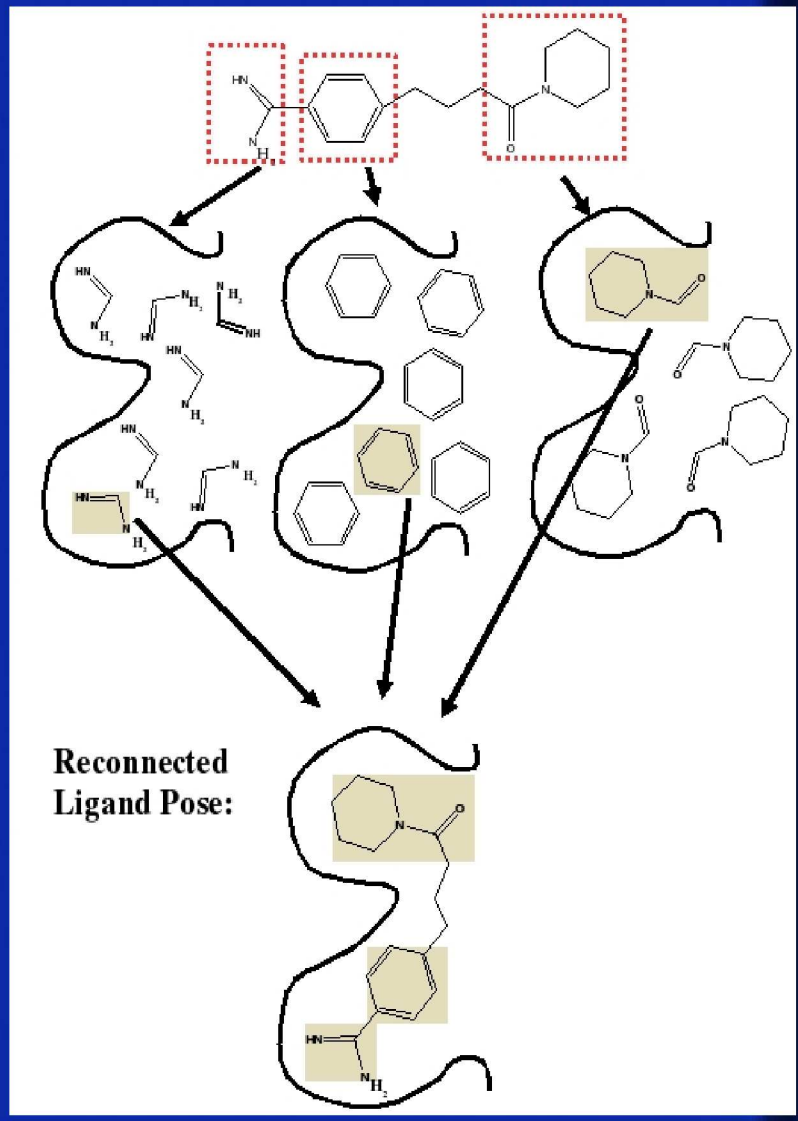


- Docking algorithm of the eHiTS software
  - Overview of the algorithm
  - Fragmentation of the ligand
  - Rigid fragment docking
  - Exhaustive pose matching algorithm – clique detection
  - Best-Match algorithm
  - Flexible chain fitting
  - Reconstruction and energy minimization
  - Protonation state handling
  - Fragment pose database to speed-up screening
- A new statistically derived empirical scoring function
- LASSO: integrated VHTS filter
- Results

18. eHiTS algorithm overview: intelligent exhaustive search

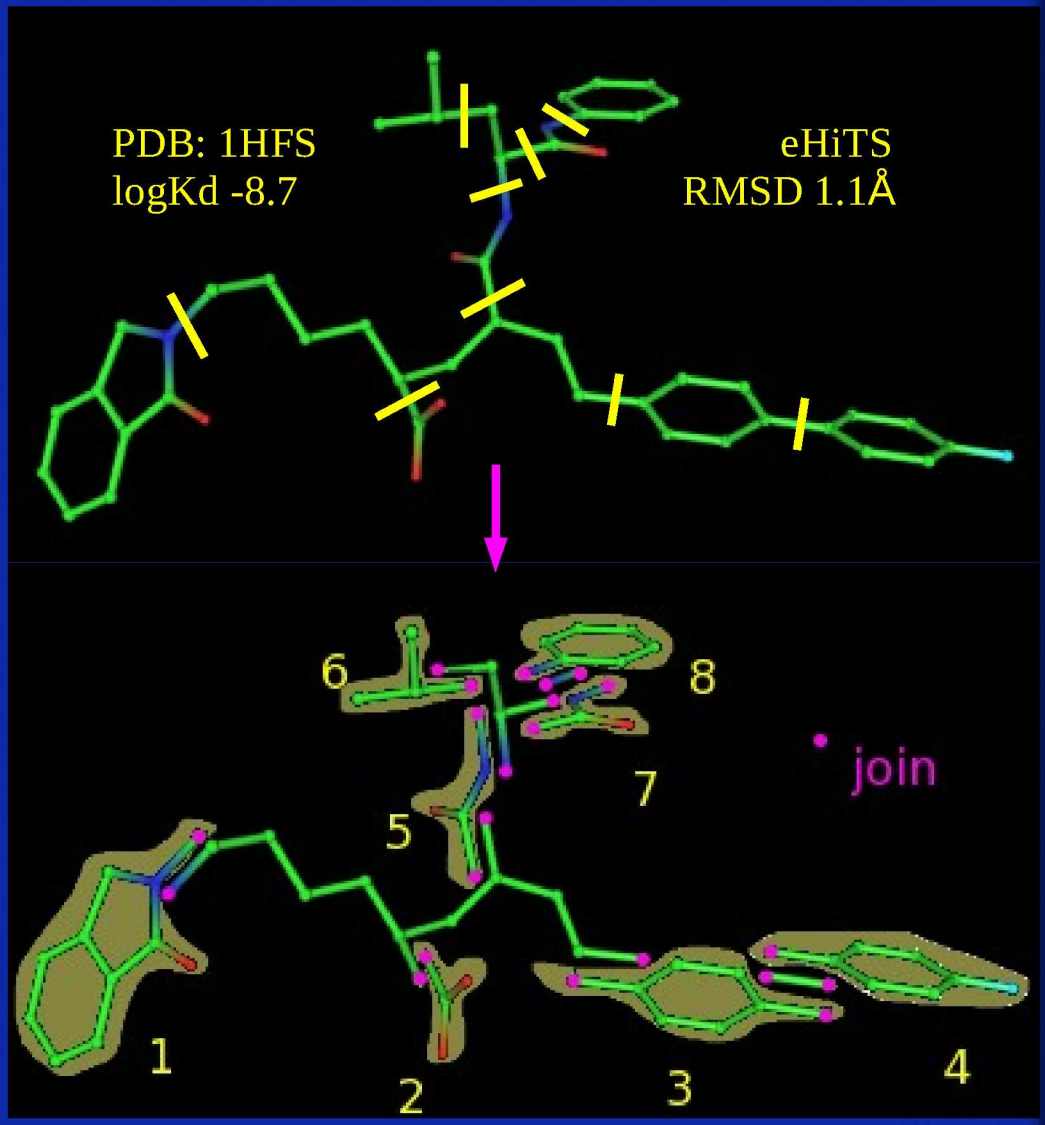


- Ligand is divided into rigid fragments, flexible chains
- All rigid fragments are docked independently (many poses)
- Pose matching (clique detection)
- Flexible chain fitting (continuous)
- Local energy minimization

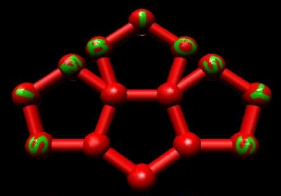


19. Fragmentation of the ligand

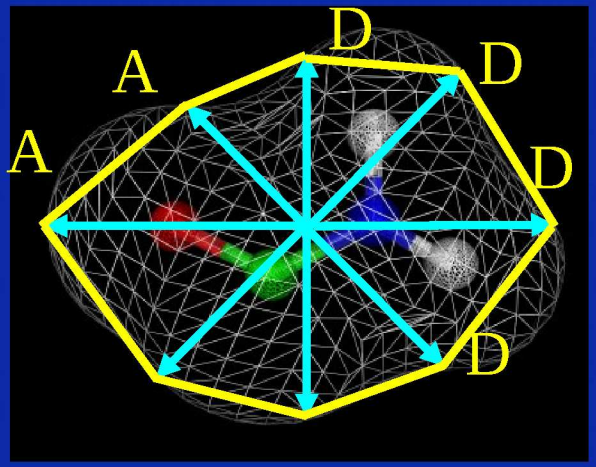
- Rigid fragments
  - Ring systems
  - Double & resonance bonds
  - Terminal fragments
- Flexible chains
  - Single bonds, sp^3 atoms
  - May contain junctions

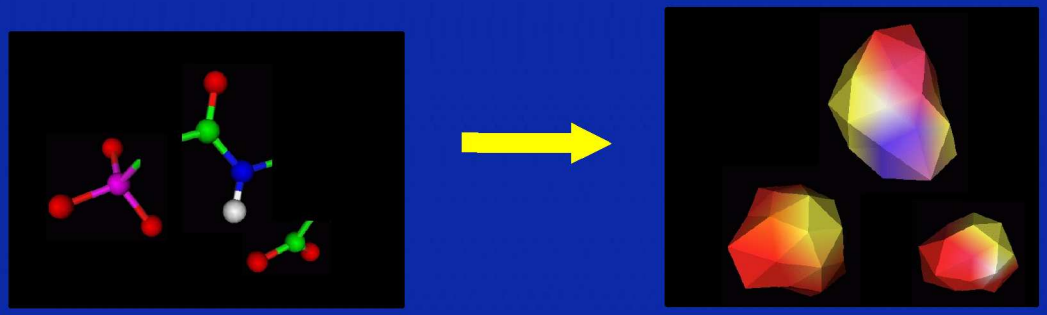


20. Rigid fragment docking based on Chemical feature mapped polyhedra

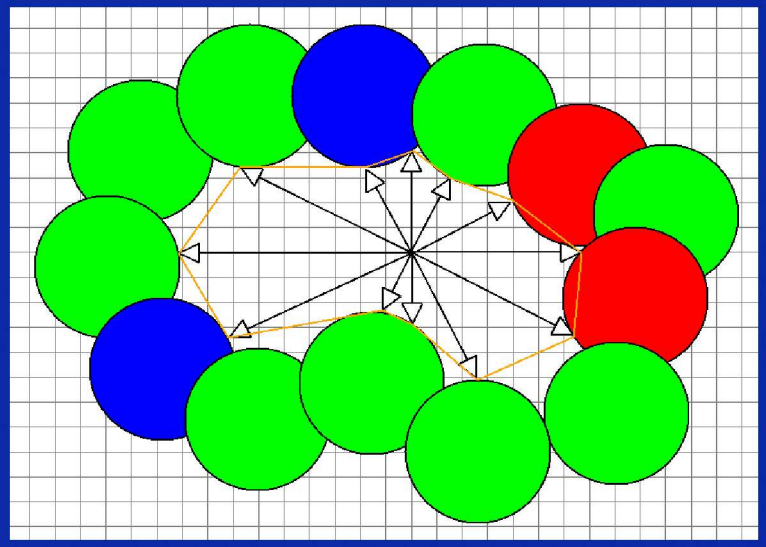


- Polyhedron shrink-wrapped onto molecular surface (Connolly)





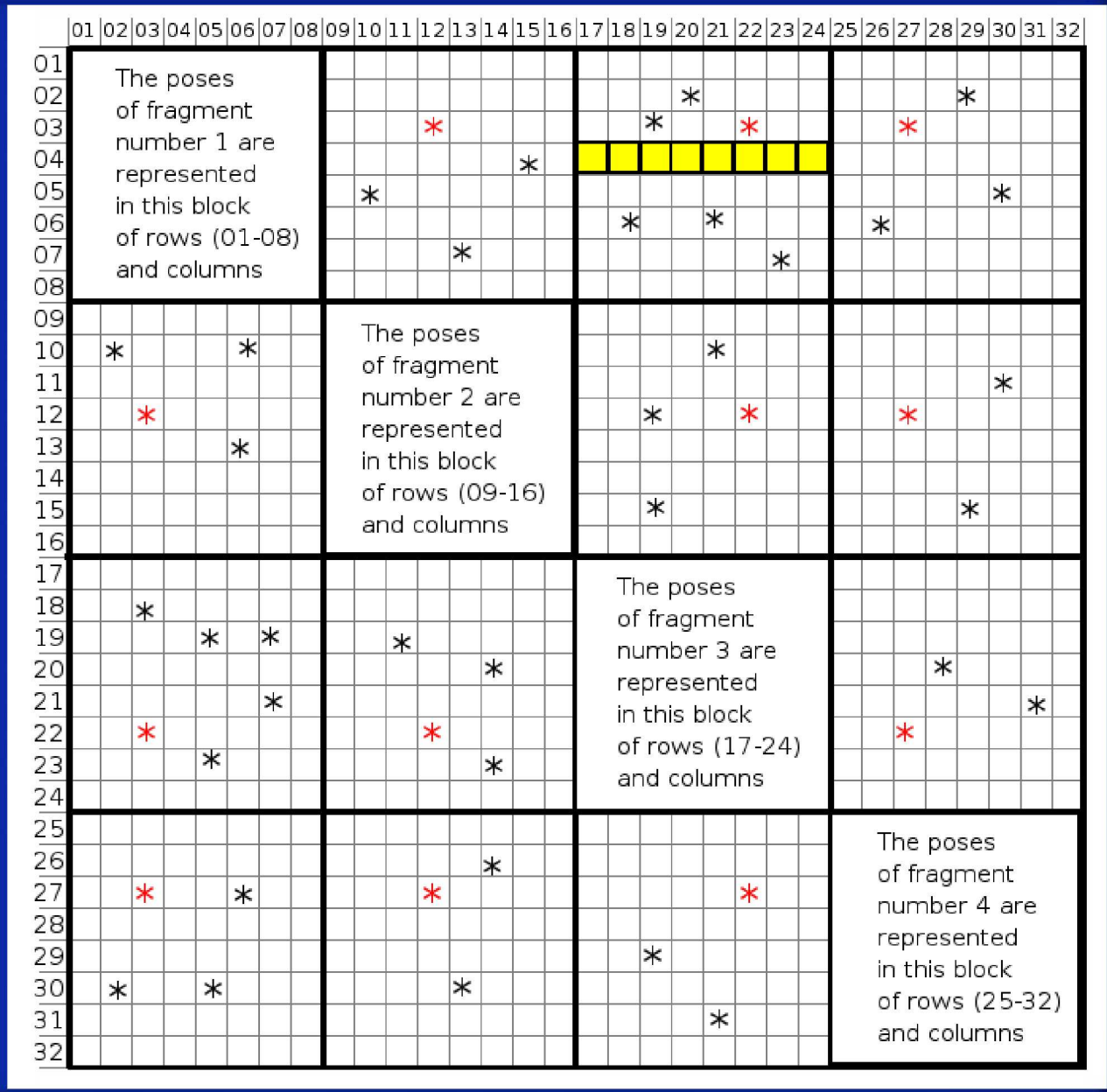
- Chemical feature flags on vertices



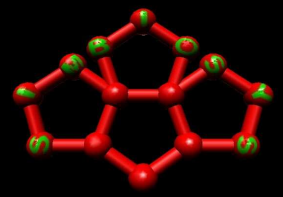
- Analogue cavity representation
- Rapid mapping of ligand and cavity polyhedra

21. Exhaustive pose matching by clique detection

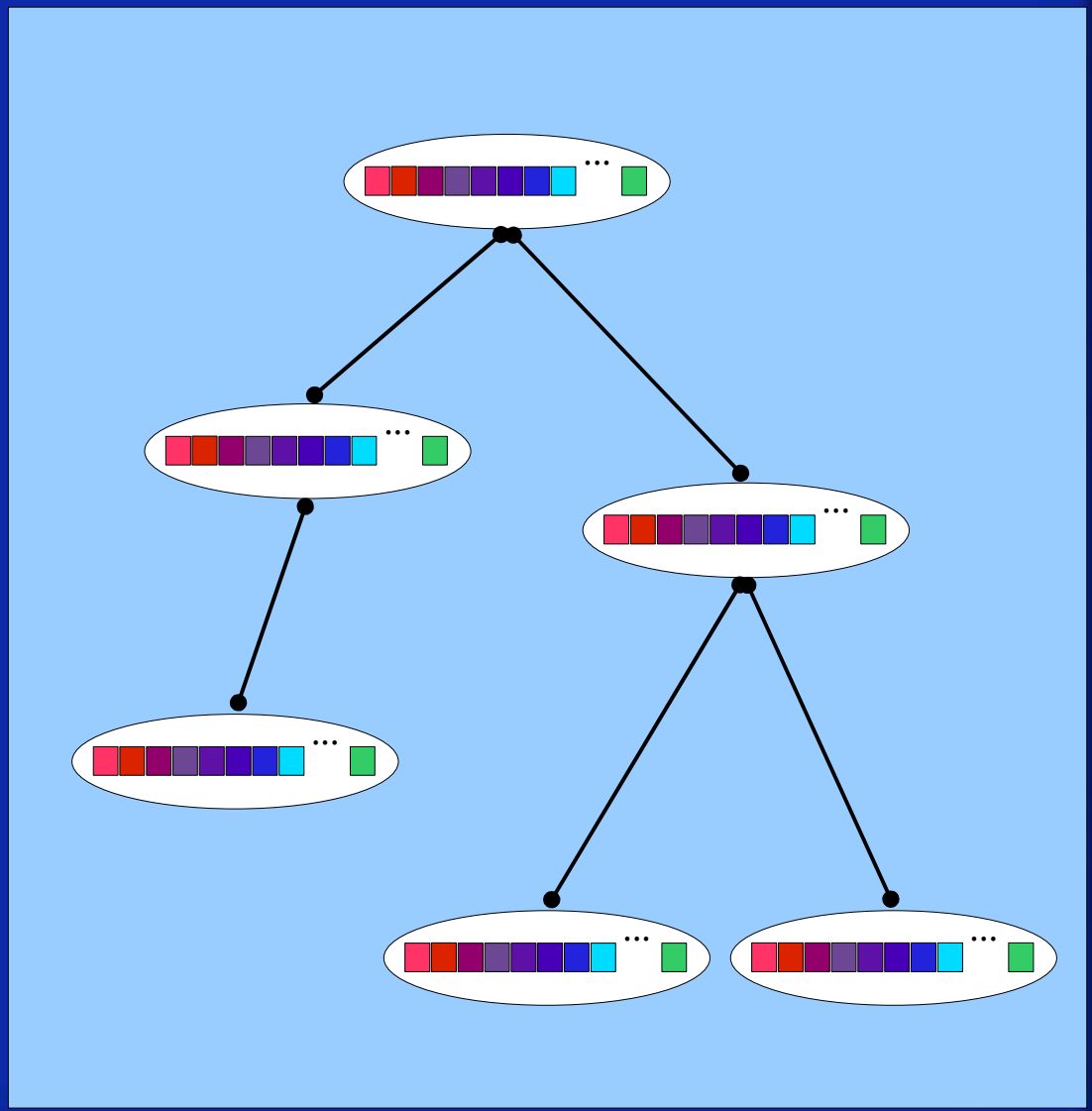
- Nodes:
  - Rigid fragment poses
- Edges:
  - Different fragments
  - No bump
  - Compatible distance



22. Best-Match algorithm

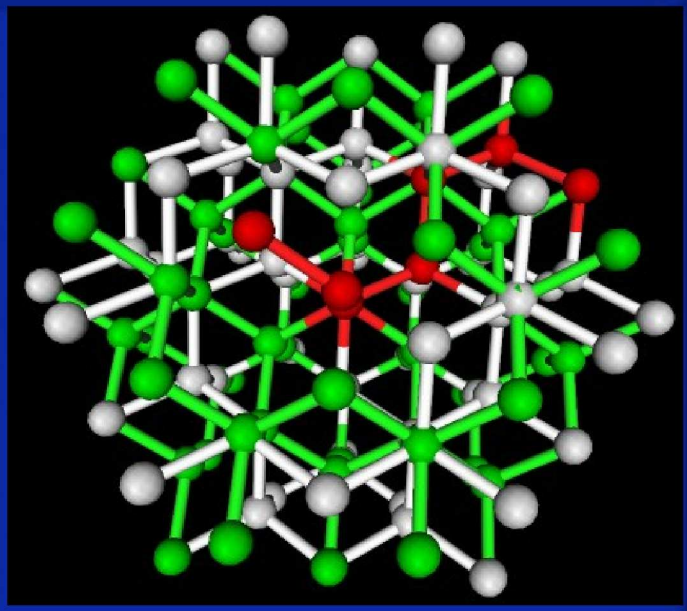


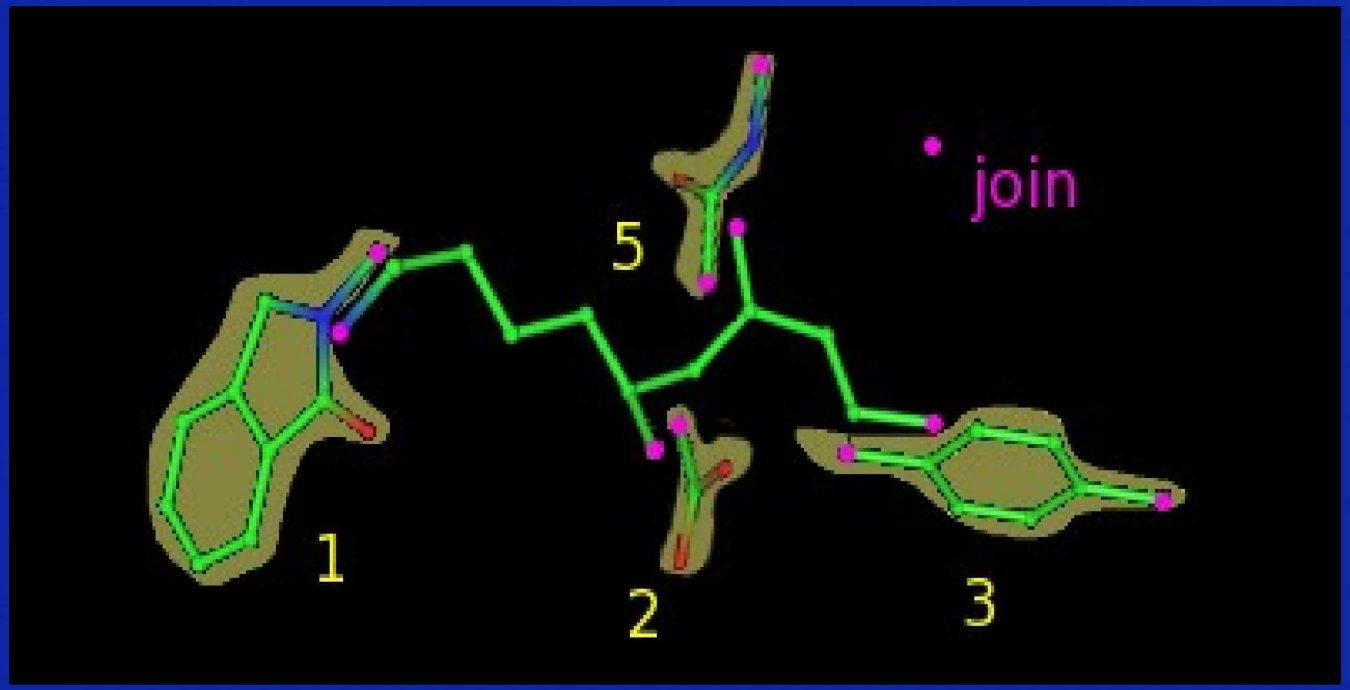
- Rigid fragment docking: 5K-100K poses, <0.5Å RMSD
- Tree representation
- Distance requirement for connected fragments only
- Find best-k scoring pose sets that are compatible recursively
- Grid based proximity data structure for compatible pairs
- Linear time complexity

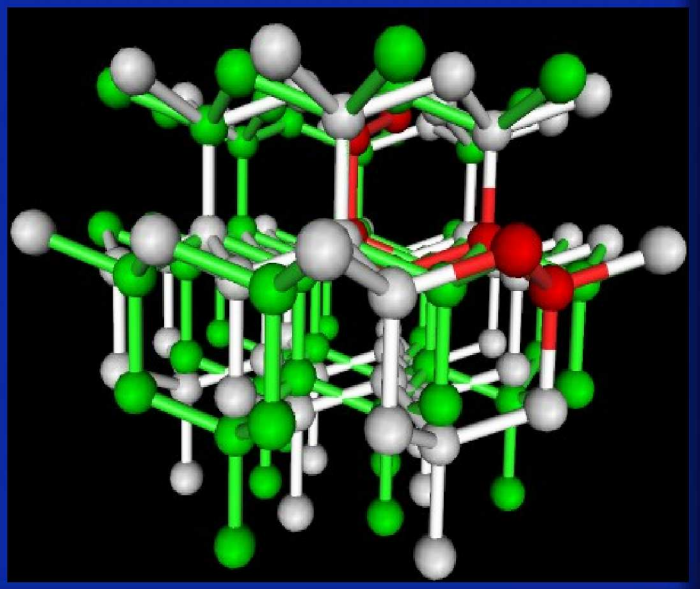


23. Flexible chain fitting

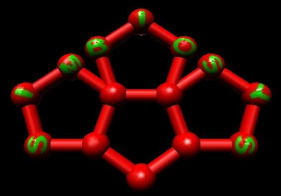
- Find chain conformation to link given rigid fragment poses:
  - Double diamond latice to start from
  - Tweak the chain to fit end point pairs and avoid boundary (PLSF)



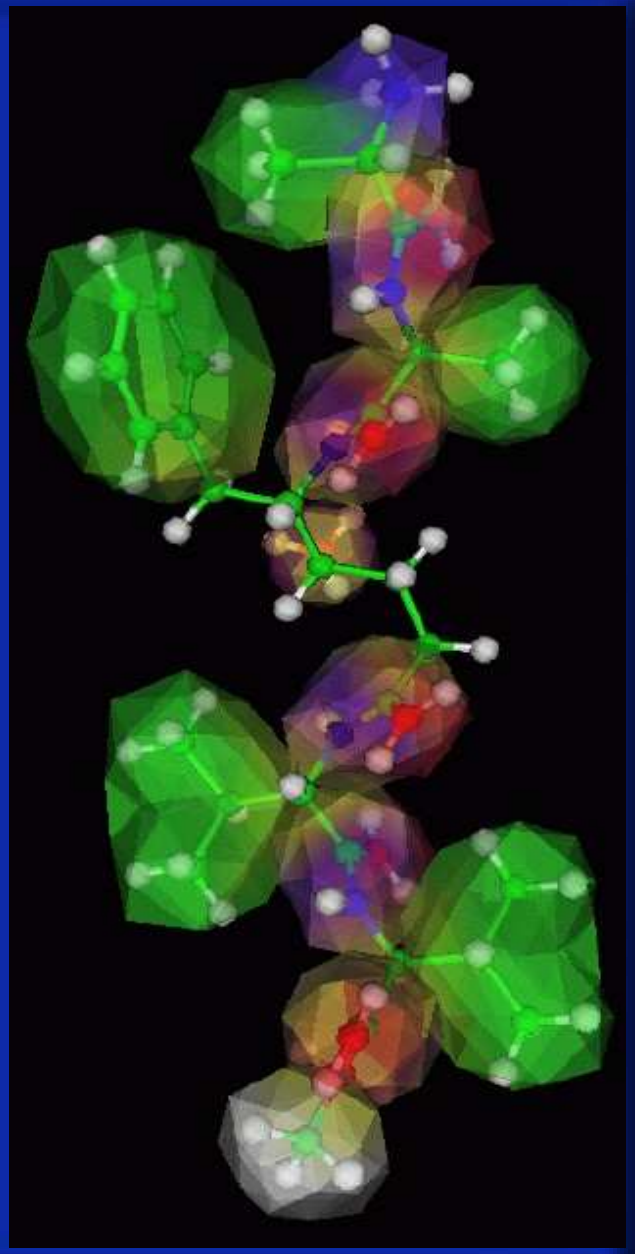


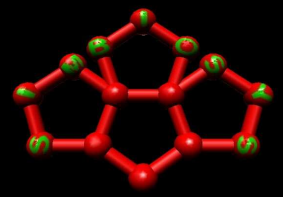


24. Reconstruction and local energy minimization



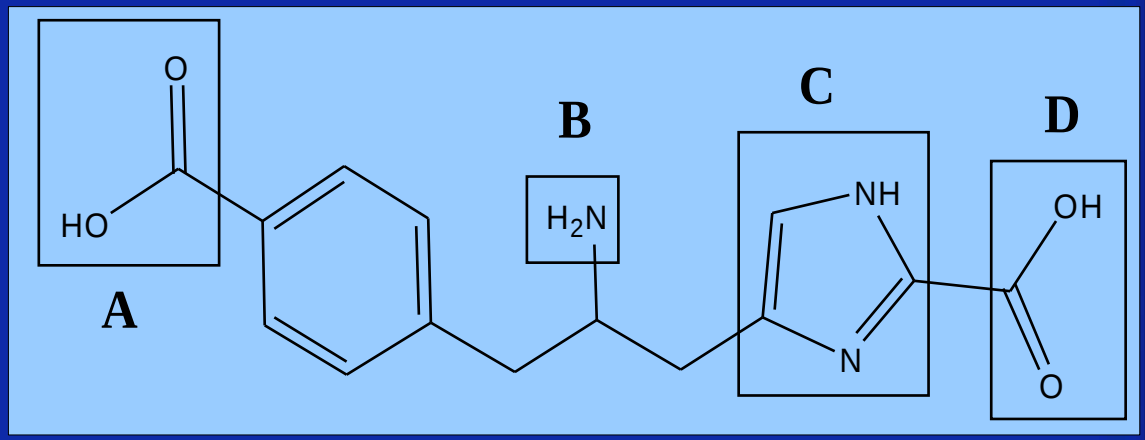
- Fragments are joined (transformed to maintain bond lengths and angles)
- Local minimization:
  - Modified Powell algorithm
  - Rigid body rotations and translations: 3 axes
  - Dihedral angles of rotatable single bonds
  - Goal function includes: receptor-ligand interactions, internal strain, solvation

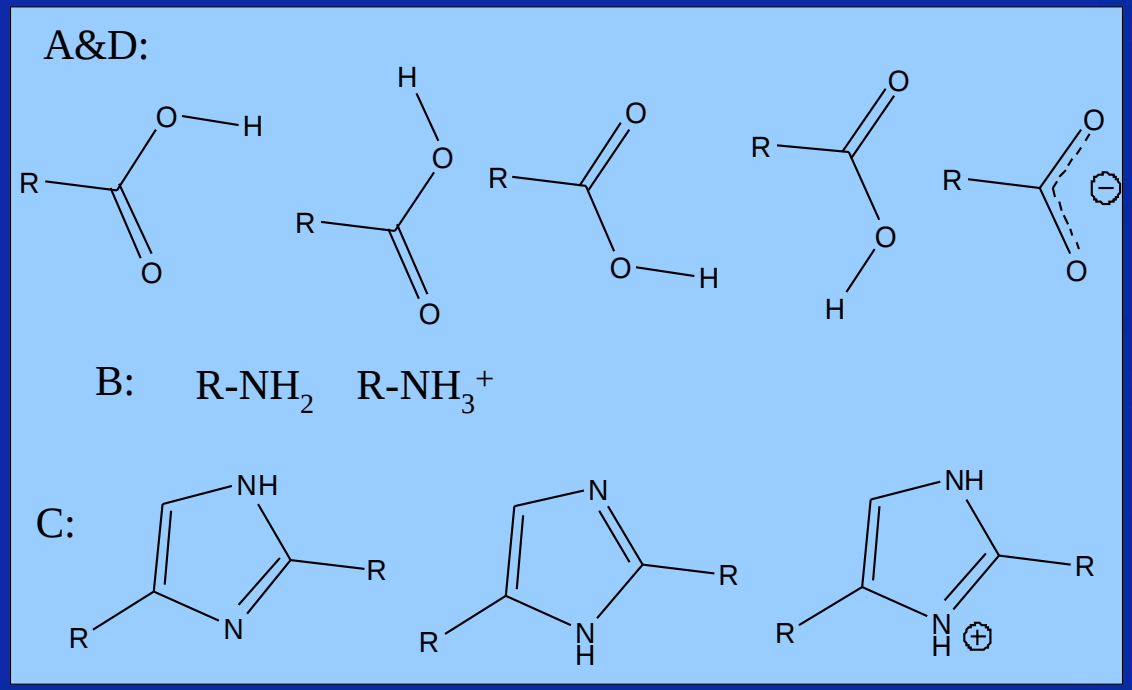




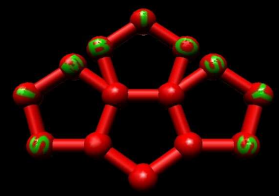
25. Universal protonation handling

- Generic form using alternative flags (H/Lp)
- Scoring picks better one for each atom
- Example:
  - 150 states enumerated
  - 11 independent H/Lp



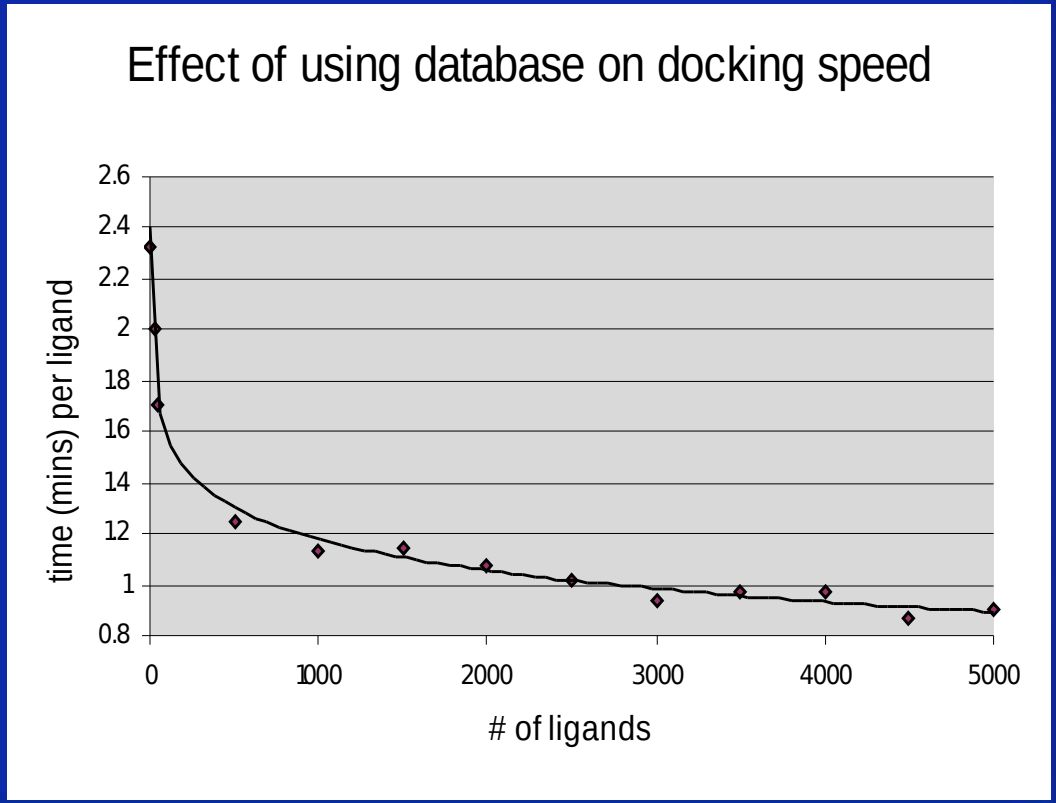


26. Dock-Table SQL-DB to speed-up screening

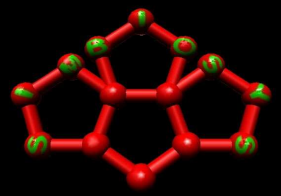


- Significant portion of CPU time spent in RigiDock
- Rigid fragment poses & scores are independent
- Stored in SQL DB and re-used for other ligands

- Disk space required:
  - Separate table per receptor
  - ~1MB per rigid fragment
  - Limited to 10K => 10GB

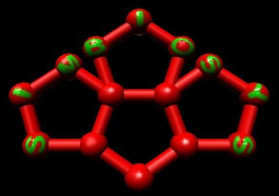


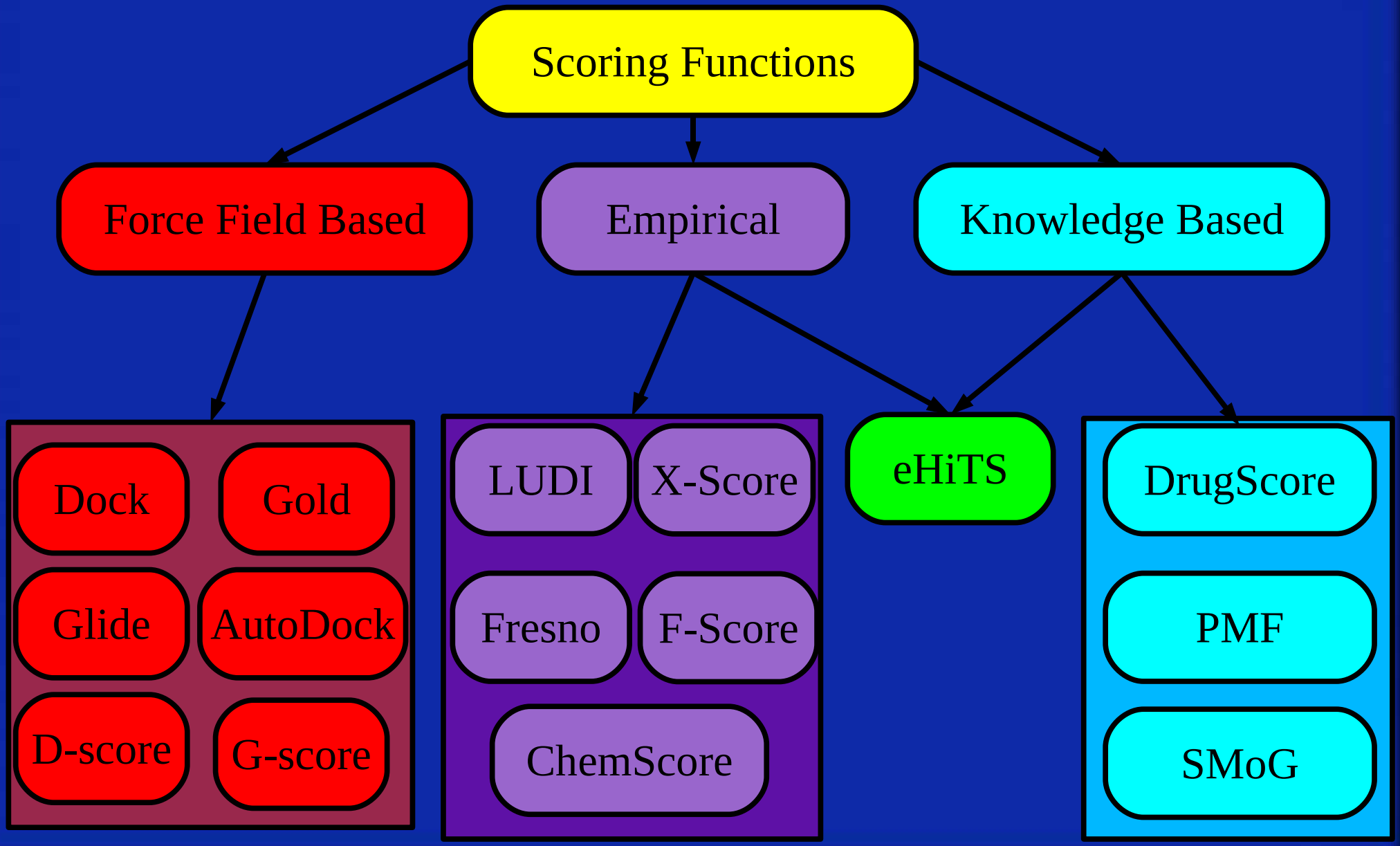
27. Contents: New statistically derived empirical scoring function



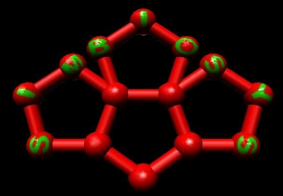
- The docking algorithm of the eHiTS software
- **New statistically derived empirical scoring function:**
  - Overview of various scoring methods
  - Problem with atom-center based scoring functions
  - Interaction Surface Points (ISP)
  - Interaction geometry description using ISPs
  - PDB curation and statistical data collection
  - Fitting empirical functions to the statistical data
  - Additional terms of the scoring function
  - Protein family recognition and clustering
  - Family based weight tuning
- LASSO: integrated VHTS filter
- Results

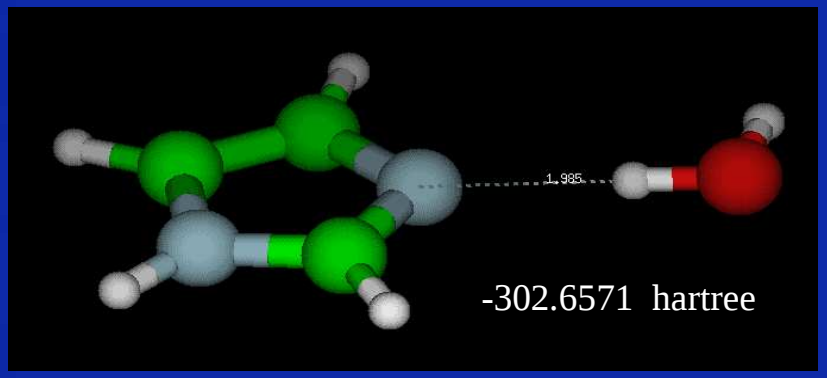
28. Various scoring methods



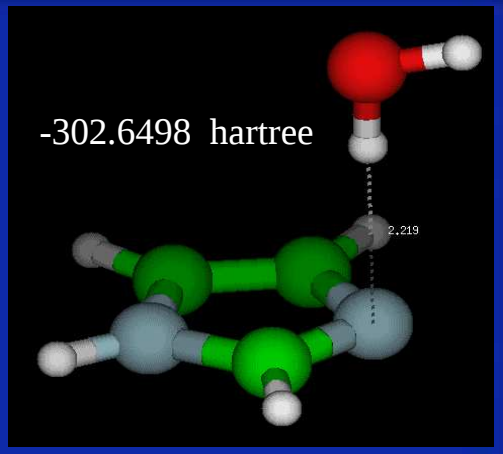


29. Problem with atom-centre based scoring functions

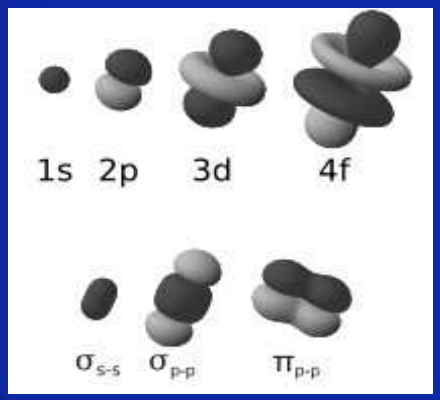




ΔE 4.5 kcal/mol

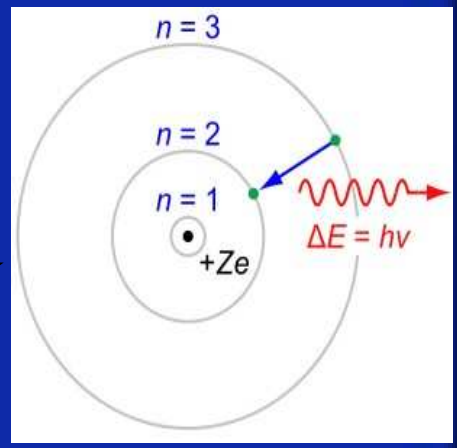


- Imidazole: 4.5 kcal/mol difference between lone-pair direction and above plane direction based on QM calculaton
- Atom-center based QM-fitted point charge FF model => no difference!
- Fundamental contradiction between QM and FF models:

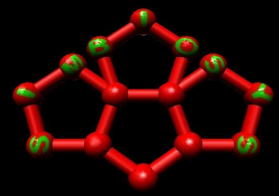


QM: all about electron density (location probability)

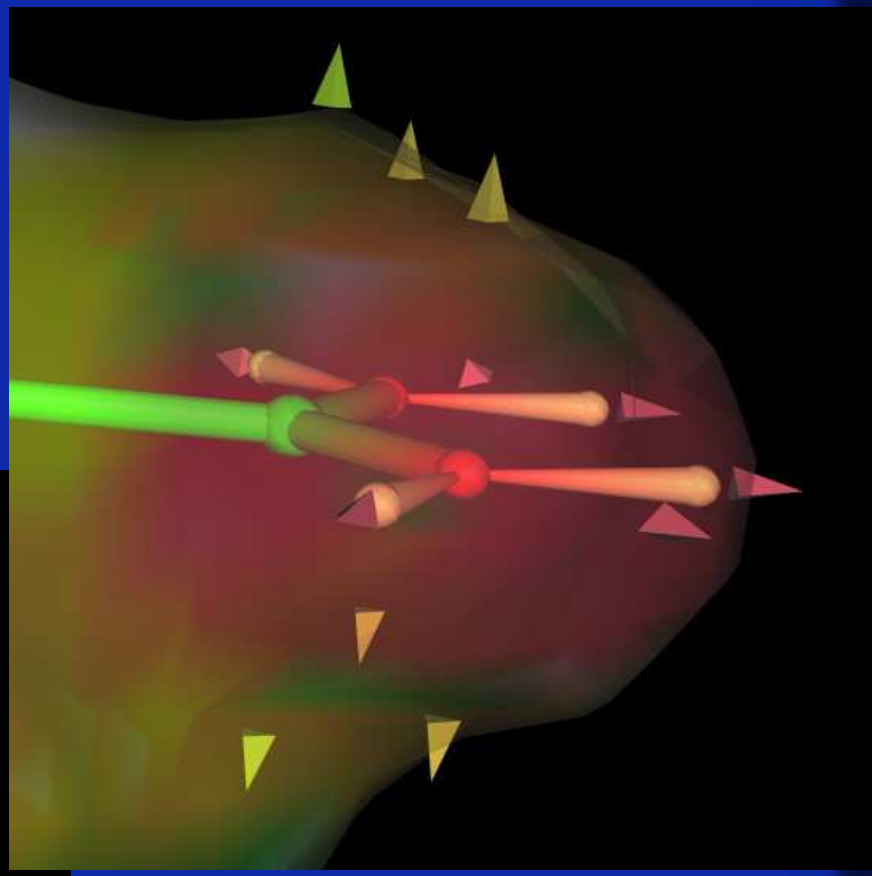
FF: ignores electron density ~ century old Bohr model

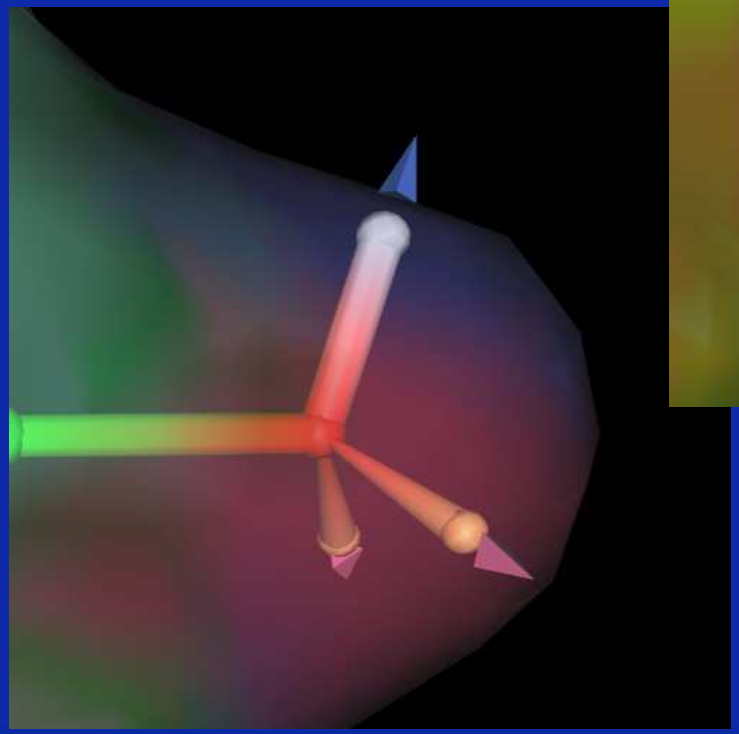


30. Interaction Surface Points (ISP)

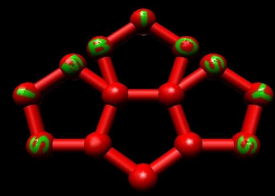


- eHiTS places directional surface points in specific locations on the surface of molecules to represent various interaction capabilities:
  - H atoms,
  - lone electron pairs,
  - π electrons

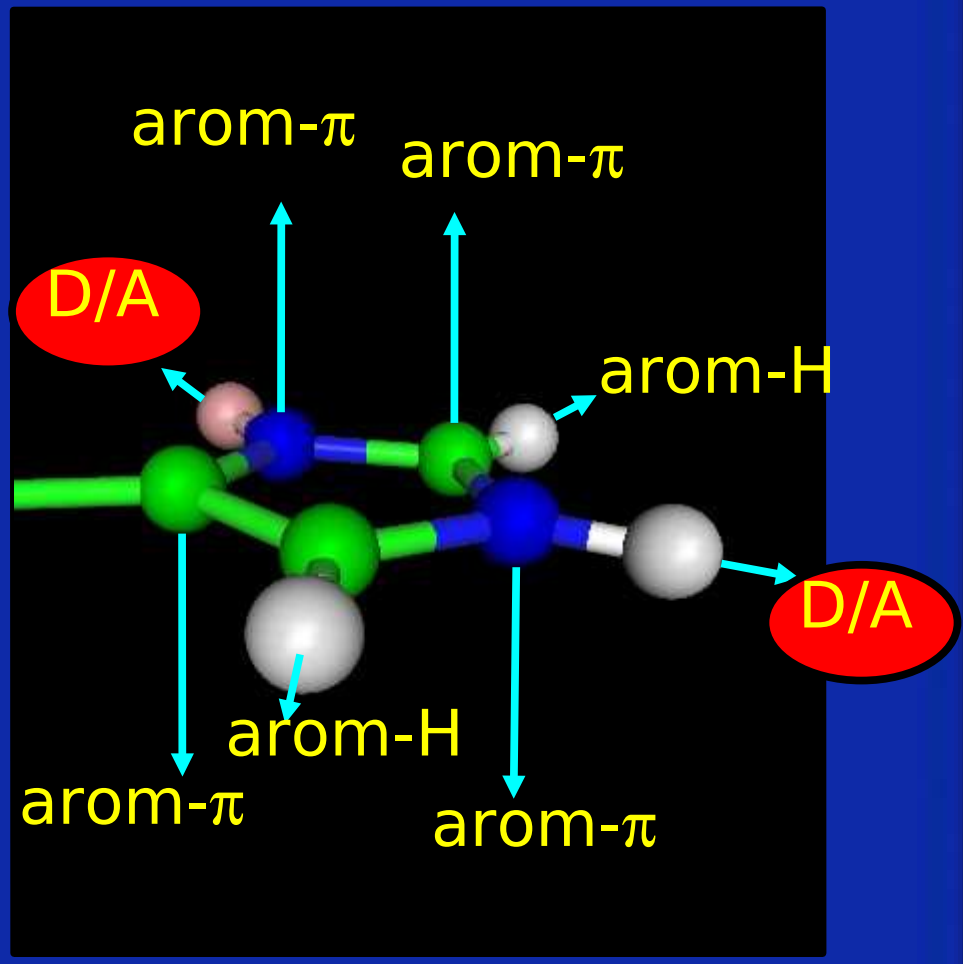




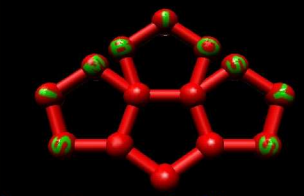
31. Interaction surface point (ISP) types

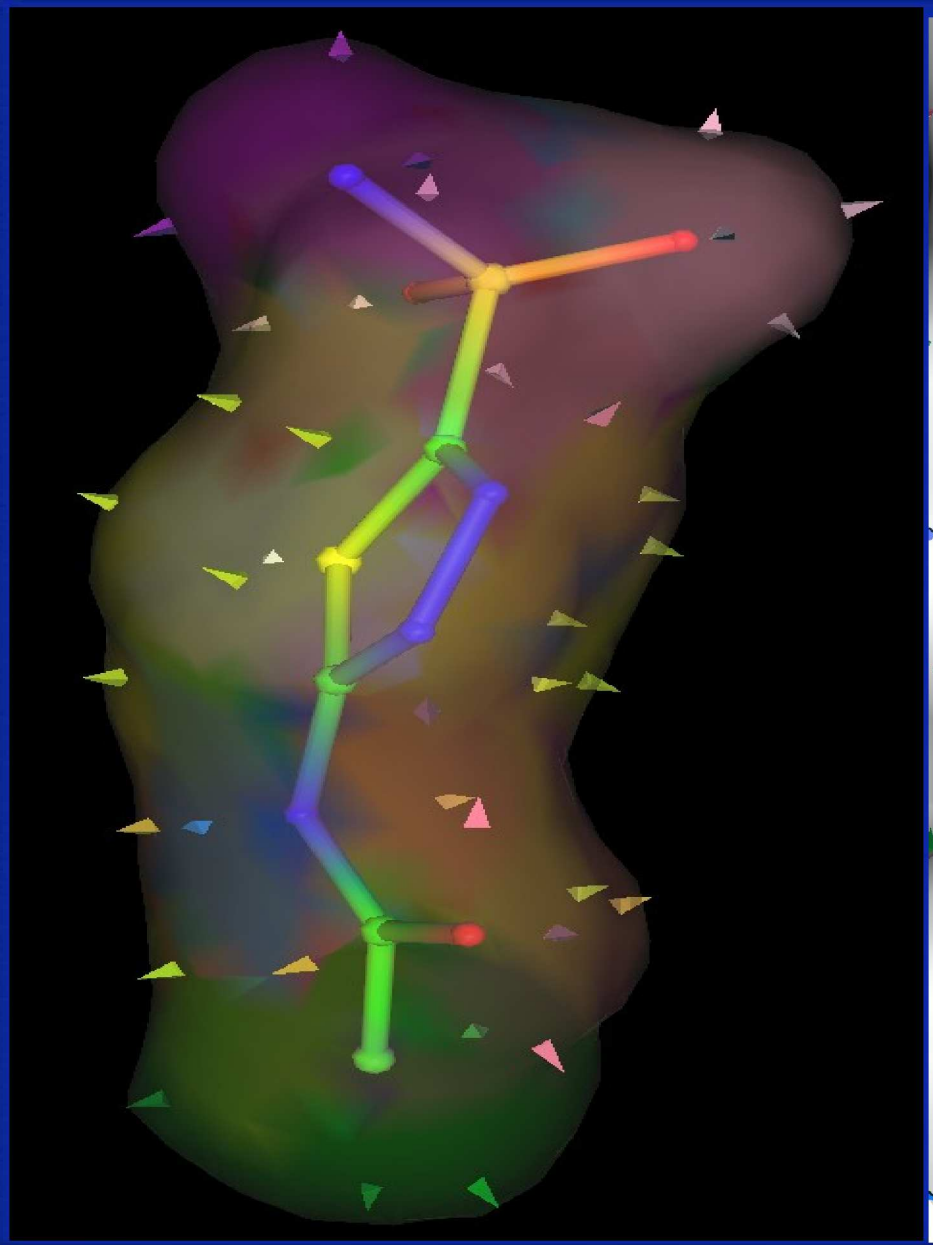


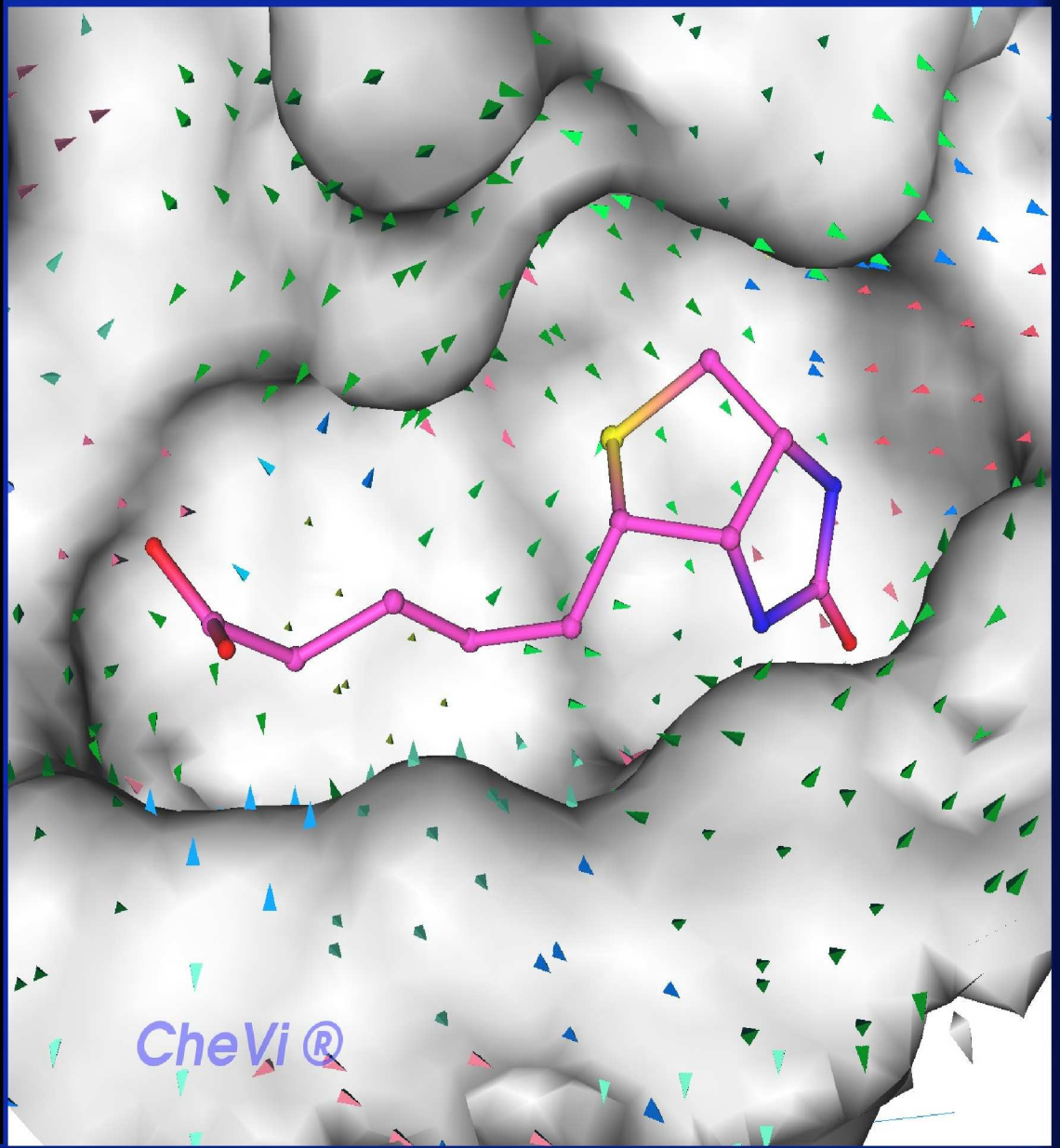
- H-bond Donors (5 kinds)
- H-bond Acceptors (5 kinds)
- Ambivalent H donor/acceptor
- Aromatic Pi-stacking (5 kinds)
- Hydrophobic (3 kinds)
- Metal ion
- Misc (Sulfur, Halogens)



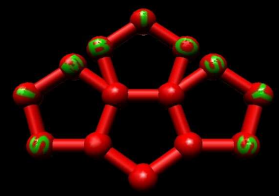
32. ISP set examples



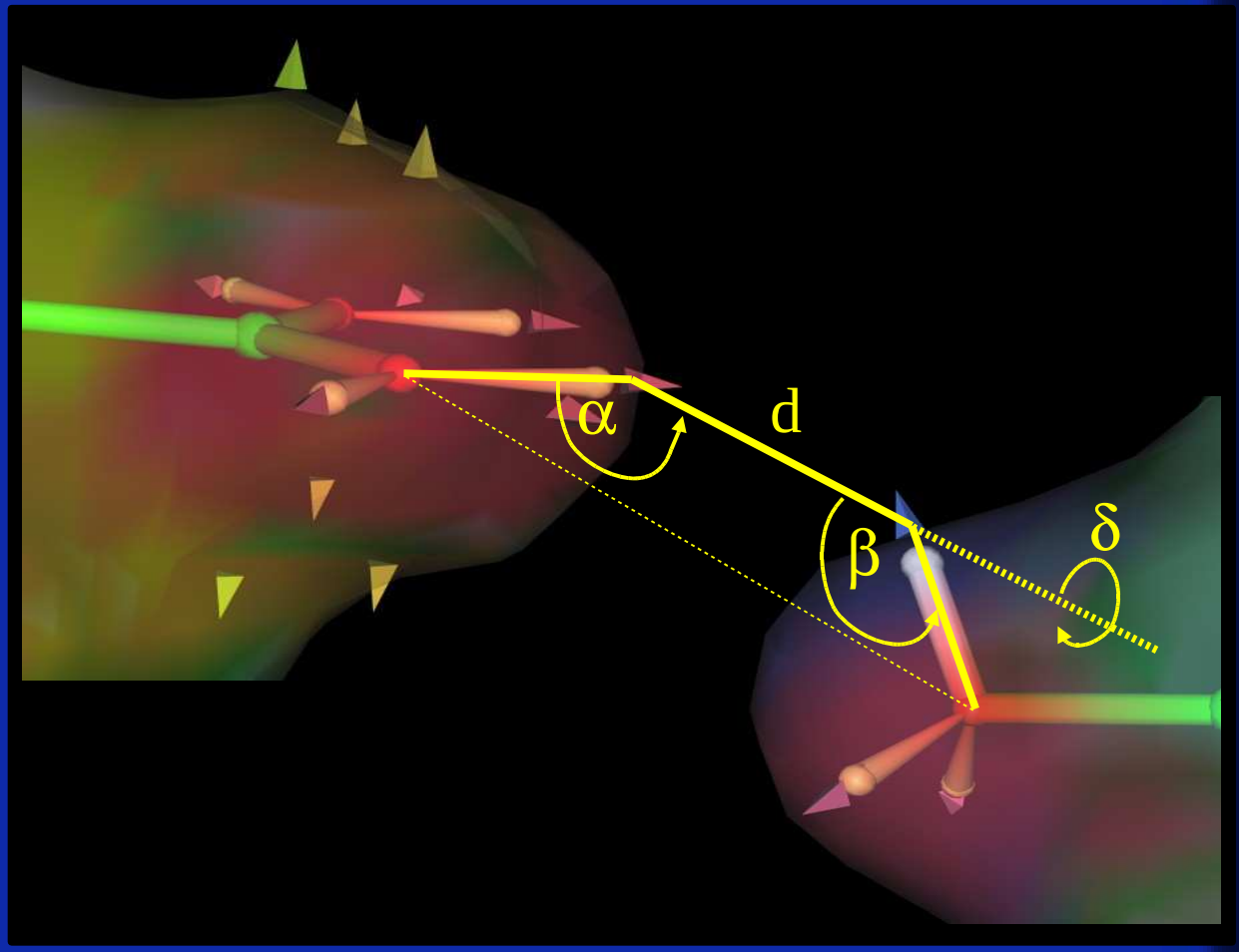




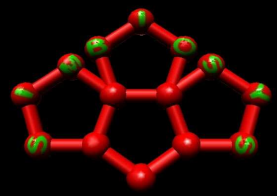
33. Interaction Geometry description



- Interaction distance (d)
- Angles between ISP vectors and interaction direction (α,β)
- The torsion angle (δ) of the two ISP vectors



34. PDB file filtering and curation



1. Protein-ligand complexes from the PDB, xray resolution 2.5Å or better: ~21,000
2. The PDB-report created by the WHAT_CHECK software was used for filtering:

   Errors in protein structures. R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Nature (1996) 381, 272-272

- Major modeling errors
- High bond length or bond angle deviations
- Ramachandran Z-score very low
- chi-1/chi-2 angle correlation Z-score very low
- Abnormal packing environment or Z-score
- Backbone conformation Z-score very low
- Side chain planarity problems
- C/N-terminal problems
- Unusual residues or torsional angles
- Connections to aromatic rings out of plane
- Abnormal packing for sequential residues
- Low packing Z-score for some residues

5. HIS, ASN, GLN side chain flips are detected (H-bonding) and corrected
6. Duplicate, unexpected atoms and water clusters without H-bonding are omitted
7. The Uppsala Electron-Density Server was used to detect and filter local errors

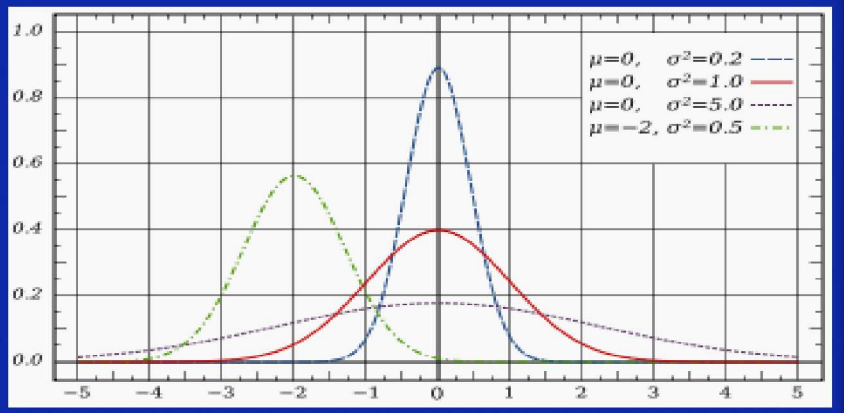
   GJ Kleywegt, MR Harris, JY Zou, TC Taylor, A Wählby & TA Jones (2004), Acta Cryst. D60, 2240-2249

3. Structures with major errors or too many residue errors are omitted: ~12,000 left
4. Residues with significant errors (RSCC<0.85, RSR>0.2, OWAB>40) are omitted

35. Statistical data collection

- ~12000 high resolution (<2.5Å) crystal structures – millions of inter.
- Probability of atom being at distance *d* (Gaussian distribution):

$$p(d) = \left(\frac{B}{4\pi}\right)^{-3/2} \int_0^{\pi}\int_0^{2\pi} \exp\left(\frac{-4\pi^2 r_{\alpha\beta}^2}{B}\right) d^2 \sin(\alpha)\, d\alpha\, d\beta$$

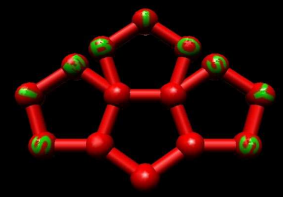


- Probability of distance *d* to occur between two heavy atoms:

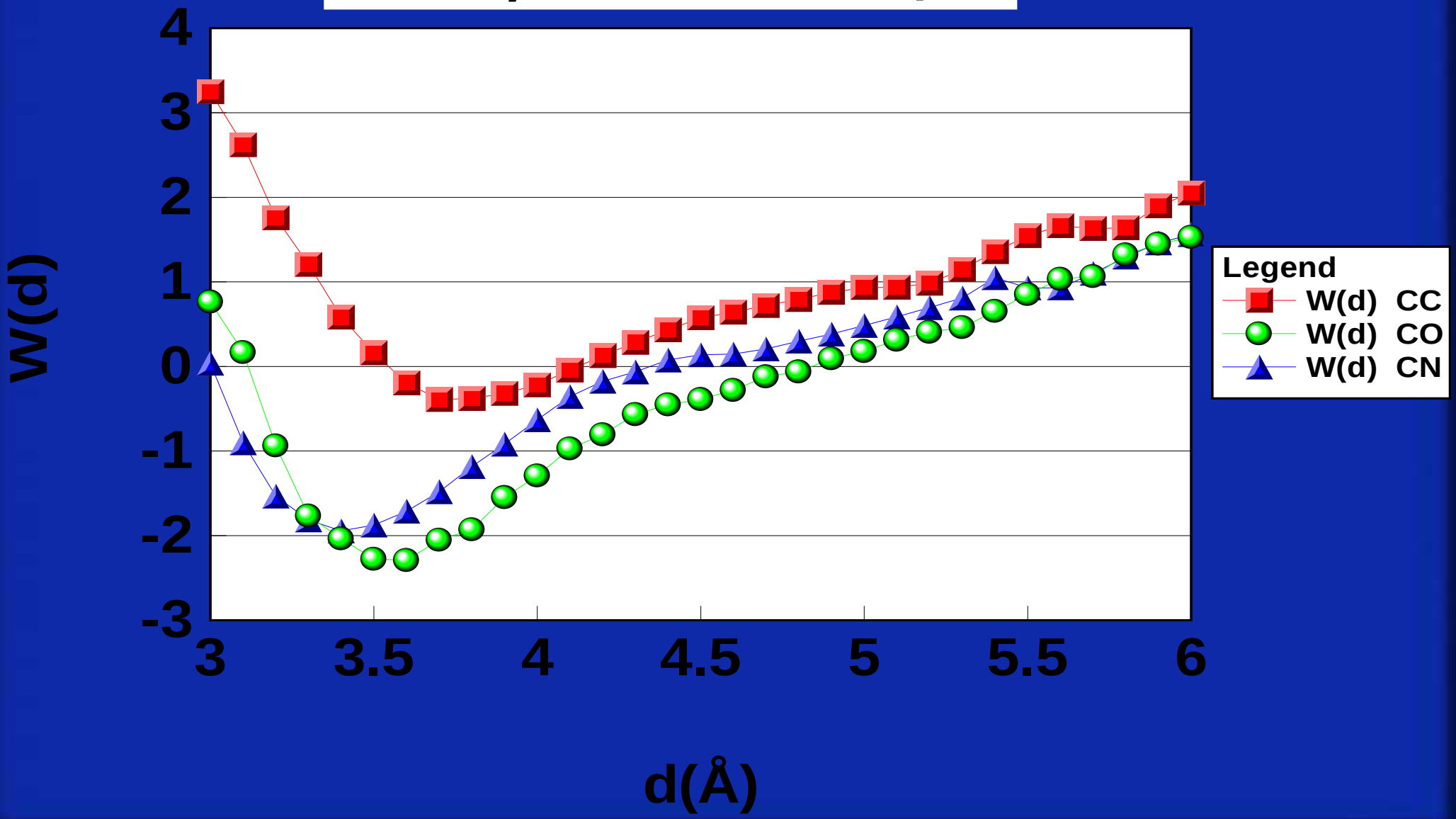
$$P(d) = \left(\frac{4\pi}{B_0 + B1}\right)^{\frac{3}{2}} d^2 \int_0^{\pi}\int_0^{2\pi} \exp\left(\frac{-4\pi^2}{B_0 + B_1}\|P_0 - P_1 + P_s\|^2\right) \sin\alpha\, d\alpha\, d\beta$$

- Similar formulae for angle and torsional components
- 4D data collection using fine numerical integral sampling

36. Energy by Boltzmann distribution



$$W(d,\alpha,\beta,\delta)=-kT\ln g(d,\alpha,\beta,\delta)$$

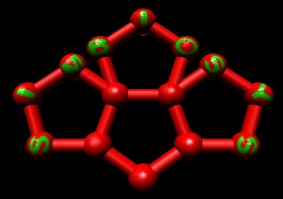


37. Fitting empirical functions to the statistical data

- Direct usage of 4D data array is impractical (Mem. + CPU)
- Analytical function parameters are fitted to reproduce the data:

$$
\begin{aligned}
g(d,\alpha,\beta,\delta) &= e_0\, s(d)\, l(\alpha)\, r(\beta)\, t(\delta) \\
s(d) &= p_{10}d + p_{11}d^2 + p_{12}d^3 + p_{13}a(d) + p_{14}c(d) + p_{15} \\
a(d) &= p_8(d - p_9)^2 \\
c(d) &= \begin{cases} \cos(a(d)) & \text{if } a(d) > -\pi \wedge a(d) < \pi \\ -1 & \text{otherwise} \end{cases} \\
l(\alpha) &= p_0 \cos\alpha + p_1 \cos^2\alpha + p_2\sqrt{\cos\alpha} + p_3 \\
r(\beta) &= p_4 \cos\beta + p_5 \cos^2\beta + p_6\sqrt{\cos\beta} + p_7 \\
t(\delta) &= p_{16} \cos\delta + p_{17} \cos^2\delta + p_{18}\sqrt{\cos\delta} + p_{19}
\end{aligned}
$$

38. The Scoring matrix

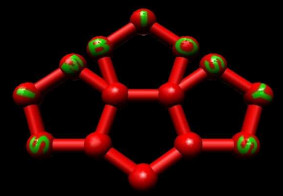


polar/H-bond **π –π interactions** **π /aromatic--cation/H-donor**

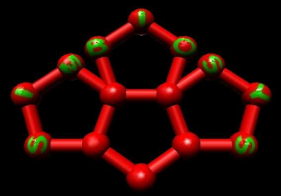
Re\Li	DonH+	Amine	Don-H	PO3--	AcidL	AccLp	WS-Lp	Ambiv	Rot-H	RotLp	CLipo	AromH	WSlip	Neutr	AromP	Res+-	Res_C	Sp2+-	Sp2_C	Halog	Join.	
METAL	-9.99	1.78	0.02	2.18	2.12	0.57	0.43	5.54	0.32	0.12	-4.8	-4.18	-3.19	0.27	-4.11	0.15	-1.49	0.98	-1.71	-6.46	4.39	METAL 0
DonH+	-5.15	-6.13	-5.38	2.57	3.8	1.14	-0.49	-0.79	-2.52	0.1	-2.95	-1.86	-1.23	-3.04	-5.31	-0.51	-1.69	-6.4	-6.6	-1.71	-1.46	DonH+ 1
Amine	-7.94	0.24	-5.21	2.13	1.37	1.47	-0.4	-0.75	0.3	1.43	-3.64	-1.72	-0.85	-5.5	-3.87	-0.75	-3.32	-1.94	-3.78	-1.89	0.91	Amine 2
Don-H	-9.99	-0.26	-1.78	2.86	2.18	3.27	2.36	0.49	-0.59	1.78	-1.7	-1.9	-0.72	-0.15	-3.95	0.01	-2.32	-2.68	-3.27	-0.65	-1.21	Don-H 3
WSdon	-9.99	-2.93	-5.56	-0.21	-0.16	-0.09	2.83	0.64	0.12	0.62	-0.64	-1.29	-0.97	-0.12	-0.74	-0.79	-2.48	-3.36	-3.21	-0.37	-0.53	WSdon 4
PO3--	-0.92	1.26	2.86	-0.72	-1.31	-1.7	-1.08	-0.58	3.77	-0.52	-1.28	-1.64	1.34	0.2	-4.65	-0.72	-0.65	-1.53	-3.57	-3.52	-0.89	PO3-- 5
AcidL	3.58	3.11	2.34	-0.66	-1.64	-0.72	-3.87	0.52	3.94	0.37	-0.99	-1.65	2.24	0.23	-5.04	-0.92	-2.45	0.12	-0.13	-0.86	-0.76	AcidL 6
AccLp	3.05	1.67	3.09	-3.8	-2.39	-1.7	-2.98	-0.01	0.51	-2.26	-0.29	0.45	0.6	0.8	-3.51	-2.42	0.14	-1.07	-0.97	-1.41	0.12	AccLp 7
Ambiv	-3.31	-0.98	2.1	3.02	1.41	0.9	0.65	4.8	1.08	1.63	-1.94	0.09	0.03	0.74	-2.79	-0.46	-2.73	1.23	-4.65	-1.02	0.68	Ambiv 9
Rot-H	-0.45	-9.99	0.24	3.15	3.78	0.89	-0.24	2.47	-4.02	-0.2	-0.61	0.05	0.19	0.54	-4.43	0.53	0.24	-3.71	-9.99	0.12	-1.09	Rot-H 10
RotLp	3.75	-1.25	2.63	0.01	0.06	-2.09	-4.05	3.75	-0.01	-2.46	-0.65	0.18	-0.98	1.34	-6.19	-1.37	-0.83	-0.79	-2.15	-1.46	1.35	RotLp 11
CLipo	-5.5	-3.79	-2.97	-2.53	-2.48	-0.87	-0.46	-1.75	-1.66	-1.25	0.83	0.78	-0.12	-0.02	1.91	0.34	0.59	0.06	1.27	1	-1.27	CLipo 12
AromH	-9.22	-3.12	-3.76	-1.88	-1.33	-1.11	0.11	-2.76	-1.94	-1.47	-0.01	0.61	-0.14	-0.11	1.28	1.08	1.87	0.03	0.4	0.39	-2.58	AromH 13
WSlip	0.01	-0.26	-1.25	-0.09	-1.07	0.12	0.16	-0.14	-1.05	-0.7	-0.09	0.27	-0.11	-0.1	1.58	2	1.62	0.35	0.62	0.76	-0.84	WSlip 14
Neutr	-9.99	-2	-0.44	0.03	0.2	-0.64	0.67	-0.63	-0.06	0.57	-0.38	-0.14	-0.68	-0.13	-0.27	0.42	0.47	0.81	-1.08	-0.26	-0.33	Neutr 15
AromP	-9.99	0.23	-3.21	-6.67	-4.18	-2.75	-1.83	0.14	-2.14	-1.67	3.61	3.12	3.75	3.29	4.88	3.14	4.56	4.6	3.66	2.61	0.79	AromP 16
Res+-	-1.56	0.16	0.46	-1.96	-2.88	-2.12	-1.42	-0.52	0.02	0.21	2.86	3.09	2.54	2.97	4.05	3.35	5.08	4.69	3.84	4.22	-4.38	Res+- 17
Res_C	-4.02	-2.1	-1.04	-2.78	-4.97	-1.94	-3.65	-2.12	-1.71	0.38	2.49	3.37	2.05	2.5	4.86	2.78	3.85	1.4	3.1	4.14	-1.28	Res_C 18
Sp2+-	-9.99	-9.04	-3.58	-2.43	-1.62	-2.64	-0.48	-2.43	-0.78	0.11	2.38	2.79	1.2	1.03	5.25	2.36	4.17	2.7	3.58	4.32	-2.88	Sp2+- 19
Sulfu	-0.03	-0.43	-1.03	-3.09	-3.24	-3.65	-0.98	-9.99	-1.24	-3.38	0.03	1.83	0.09	0.36	1.21	1.29	0.94	-0.16	1.98	0.07	1.04	Sulfu 22
Re/Li	DonH+	Amine	Don-H	PO3--	AcidL	AccLp	WS-Lp	Ambiv	Rot-H	RotLp	CLipo	AromH	WSlip	Neutr	AromP	Res+-	Res_C	Sp2+-	Sp2_C	Halog	Join.	

metal-ion **hydrophobic**

39. Additional scoring terms



- De-solvation: continuous model, ISP type dependent
- Steric clash penalty: distance-square from Connolly surface
- Pocket depth: signed distance of atoms from convex hull
- Protein family data based coverage (ISP type pairs)
- Ligand strain energy (torsional probability + vdw LJ 6-12)
- Ligand intra-molecular interaction score (ISP pair ~ receptor)
- Ligand entropy loss (frozen rotatable bonds)



40. Protein “family” clustering

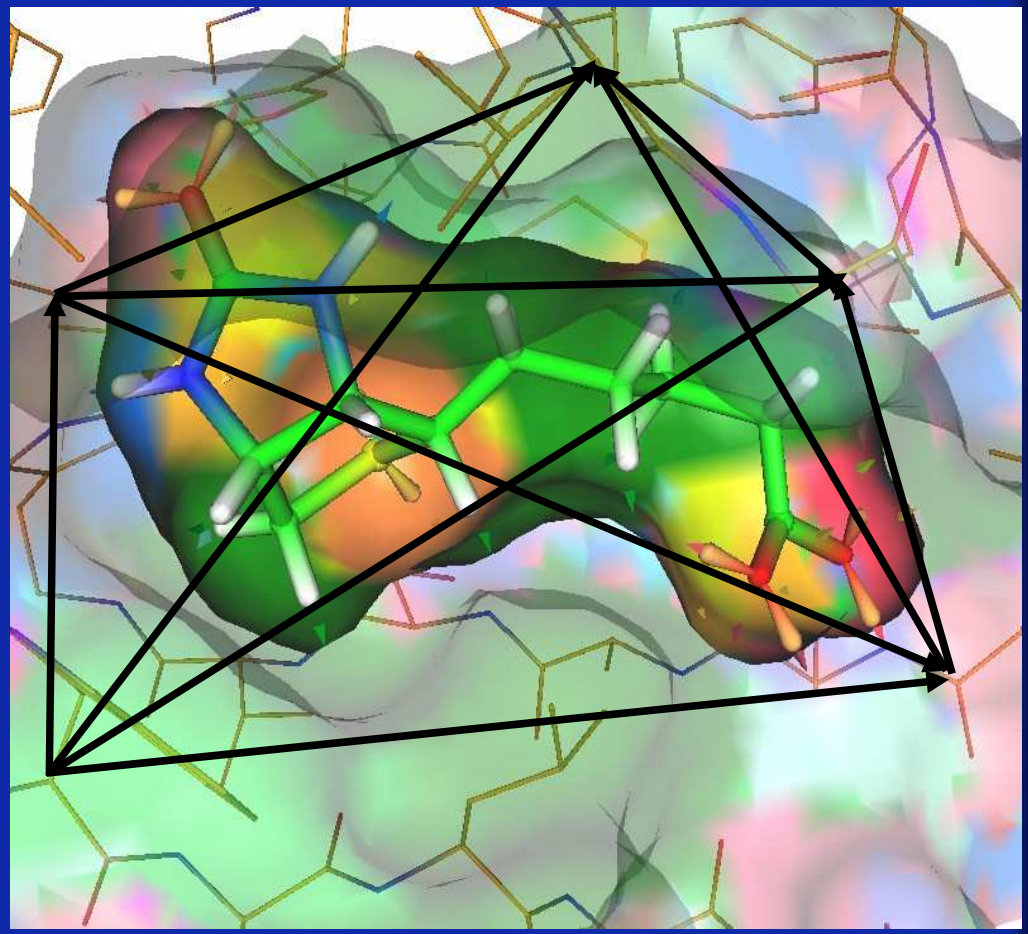
~12,000 PDB Complexes are clustered automatically into ~500 protein sets.

Geometric clustering is based on binding site residue Cα distance matrix.
- distance tolerance (default 3Å)
- matching subset size minimum (5)
- minimum set-size (5 entries)

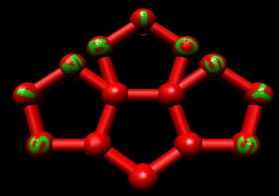
Correspondence to biological activity family is not exact, e.g. Kinase DFG-in DFG-out is separate, but thrombin and trypsin in same set.

Under represented sets and singletons are treated as a fall-back general set

The same matching criteria is used to find the “family” of the target protein in the preprocessing step of a docking run



41. Protein “family” based weight tuning



Docking is performed for all members of a “family” (training PDB set) to generate 300+ poses using default scoring weight parameters

All scoring term values are recorded for each pose along with the RMSD from the x-ray pose

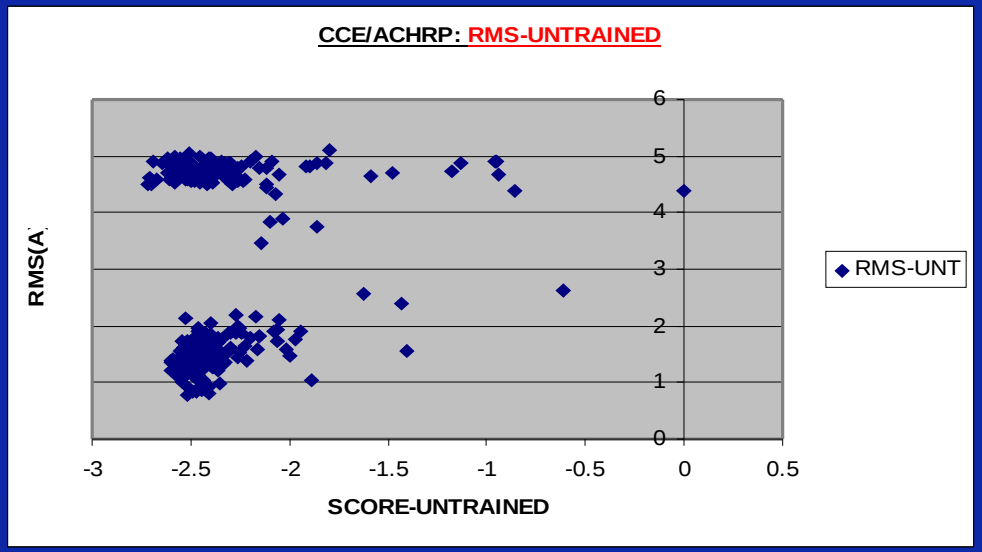
The interaction score-matrix is divided into 5 categories (metal, H-bond, hydrophobic, pi-pi and other): we use 1 weight parameter per category

Along with the additional terms, we have 20 weight parameters to tune

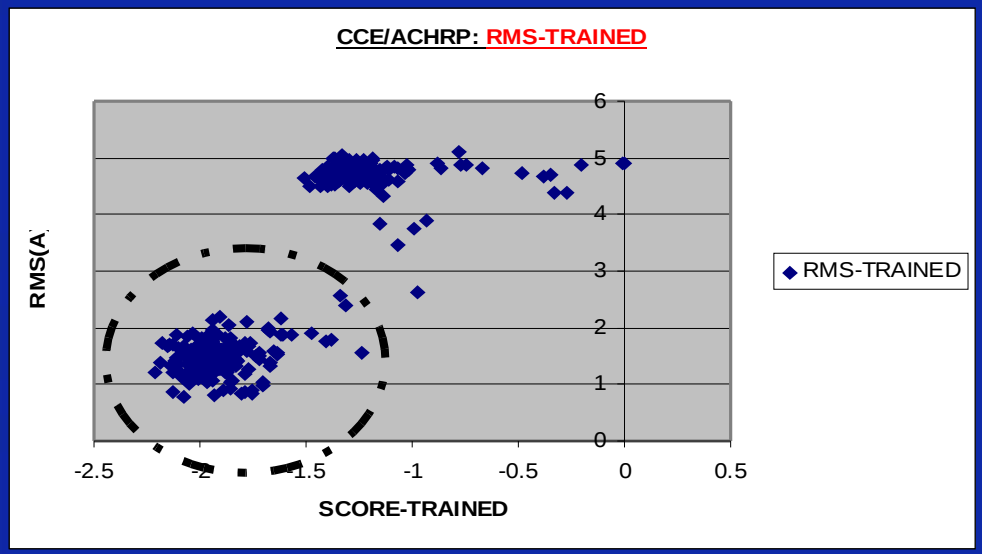
The goal function of the weight tuning optimization process includes:
- RMSD of the top-rank pose from each of the complexes in the set
- rank position of the closest pose to the x-ray among all poses
- score difference between the closest pose and the top-rank pose

Tuning does not influence the generated set of poses (rank-order only)

42. Effect of rank tuning

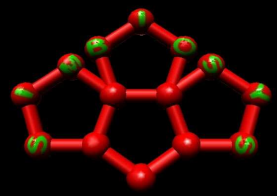


Untrained Scoring: Note that while there are many low RMS solutions in the good score regime there are also high RMS solutions with same score range



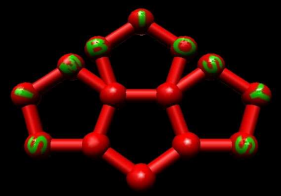
Trained Scoring: There is dramatic score-separation of the ‘correct’ pose RMS-re-gime (circled) at low scores from poor scoring –high RMS results

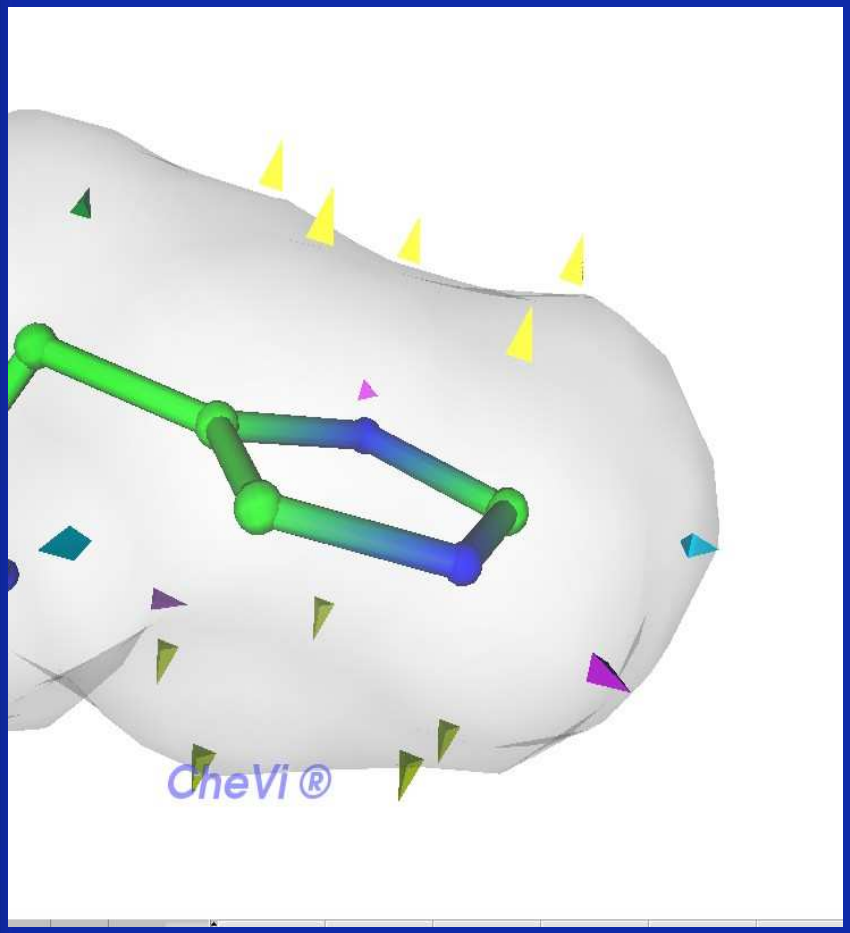
43. LASSO: integrated VHTS filter



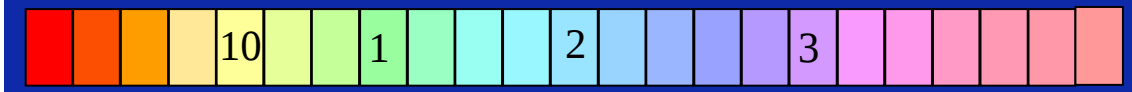
- Docking algorithm of the eHiTS software
- A new statistically derived empirical scoring function
- **LASSO: integrated VHTS filter**
  - **Conformation independent QSAR descriptor**
  - **Machine learning using neural network**
  - **2D structural diversity vs ISP vector diversity**
  - **Scaffold hopping: 2D diversity vs enrichment**
- **Results**

44. Conformation independent QSAR descriptor





Interacting Surface Point Count Vector:



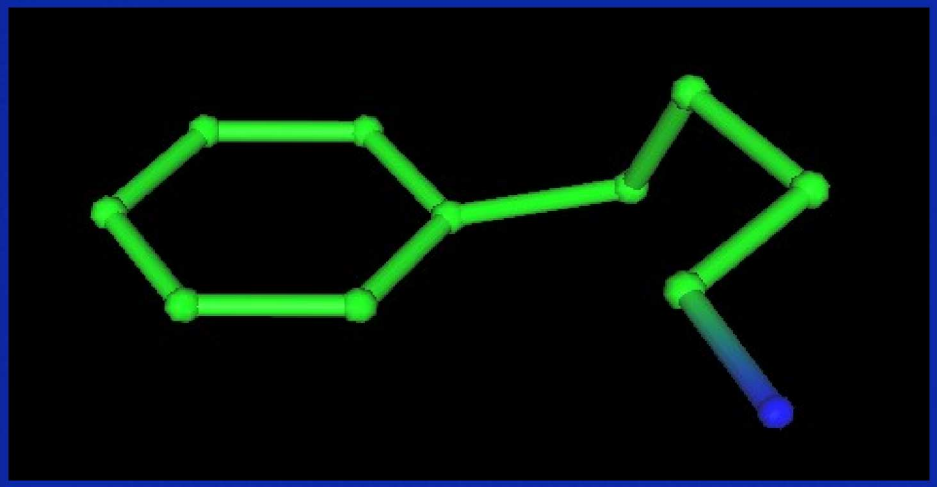
The eHiTS -LASSO VHTS Filter is based on the count vector of the interacting surface point types (ISPT) of the ligand molecules.

Each ISPT has associated chemical properties (indicated by their color), such as H-bond donor, H-bond acceptor, hydrophobic, π-stacking, etc.

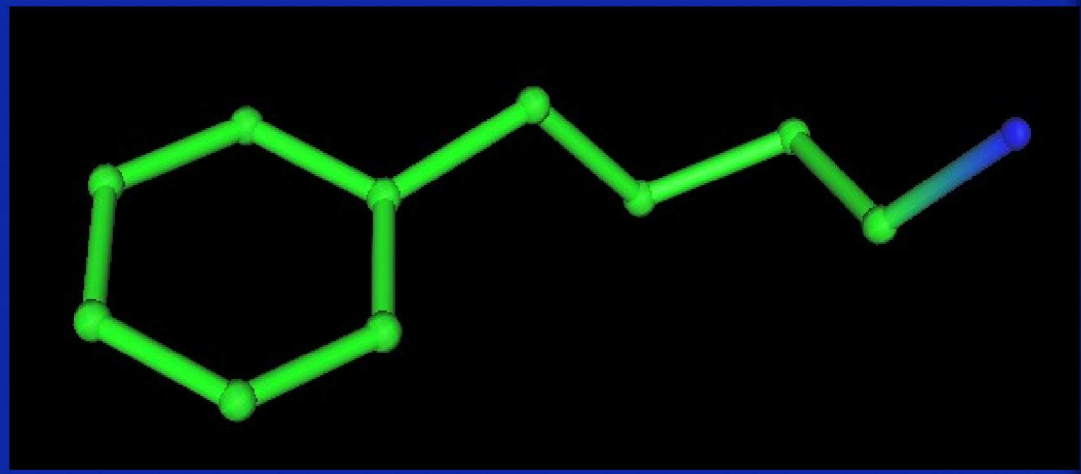
The Filter is based on the assumption that ligands with similar feature vectors have similar activity.

45. Conformation Independence

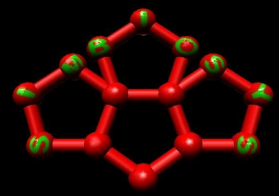
- Each heavy atom gets a specific number (up to 5) interaction surface points fully determined by its connectivity and hybridization, i.e. does not depend on the conformation
- The type of the ISP is determined by the local environment (e.g. perceived functional group) of the atom, i.e. depends only on 2D connectivity, not conformation.
- Only the 3D orientation of the ISP depends on the conformation of the molecule, the count of each ISP type is independent!



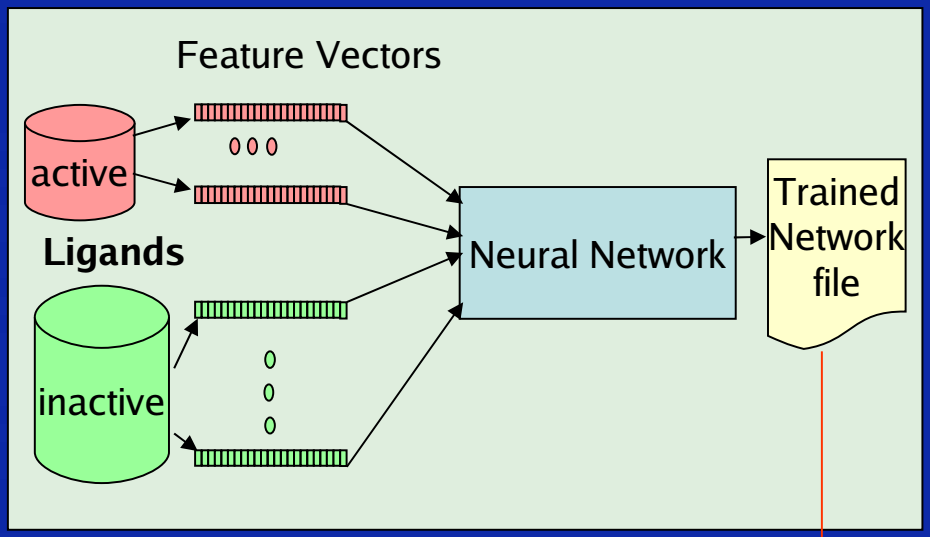
=



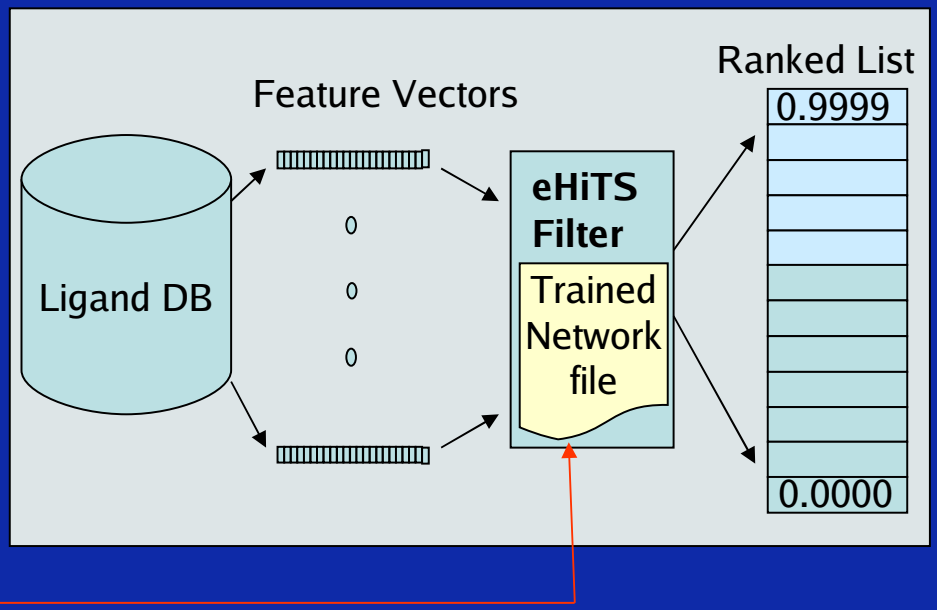
46. Machine learning using Neural Network



Training eHiTS LASSO

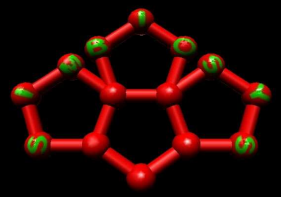


Screening with eHiTS LASSO

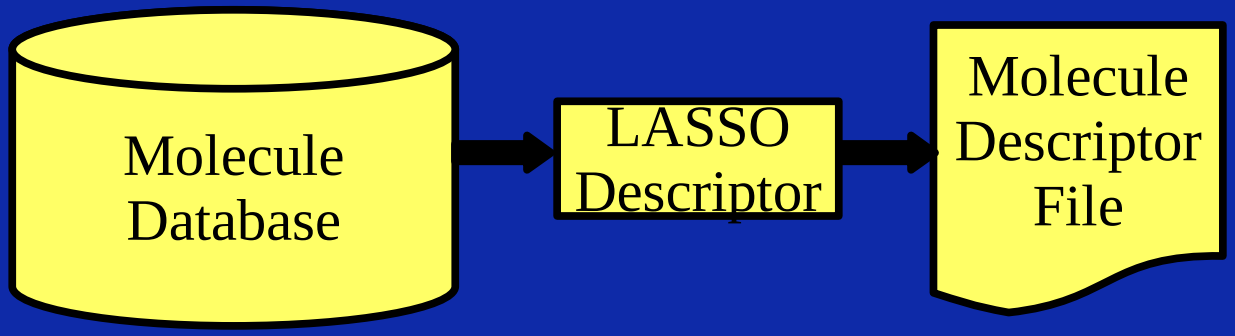


Feed-forward neural network with a single hidden layer:
Input layer 23 nodes, hidden layer 5 nodes, output layer 1 node

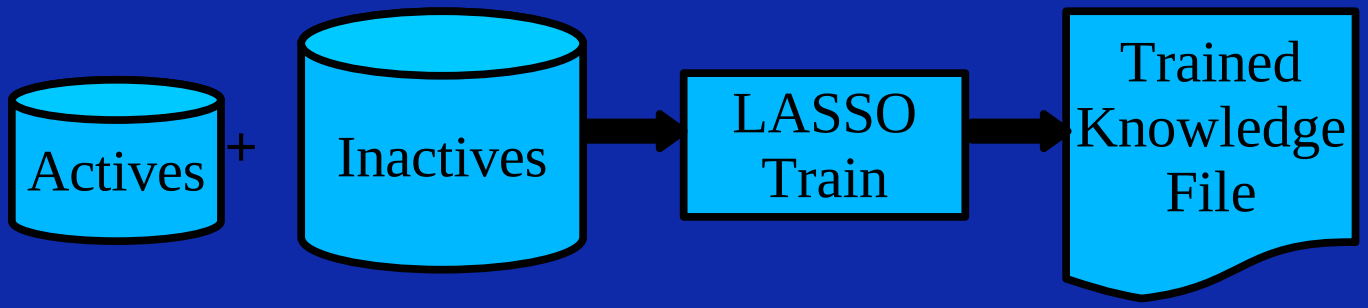
47. Filtering by Neural Network



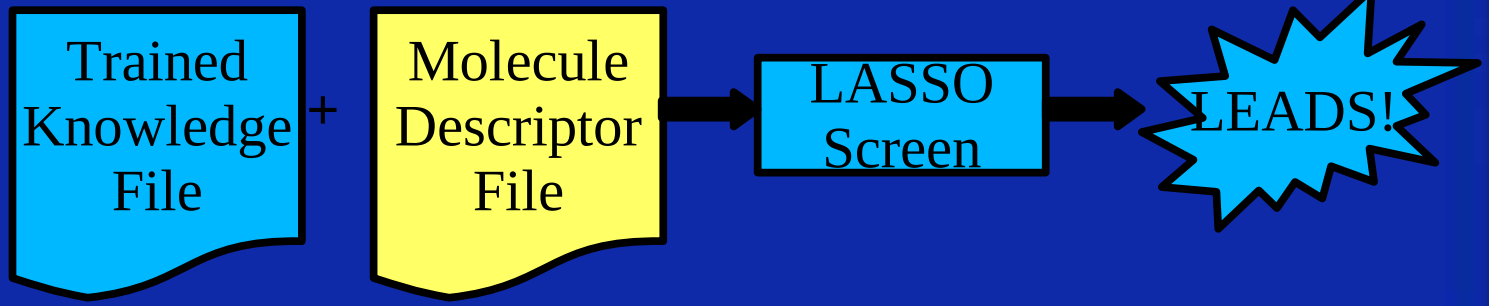
Step 1:
Create a LASSO descriptor file of the database you wish to screen
speed: thousands / min.



Step 2:
Create a knowledge base file using known active ligands and inactive molecules

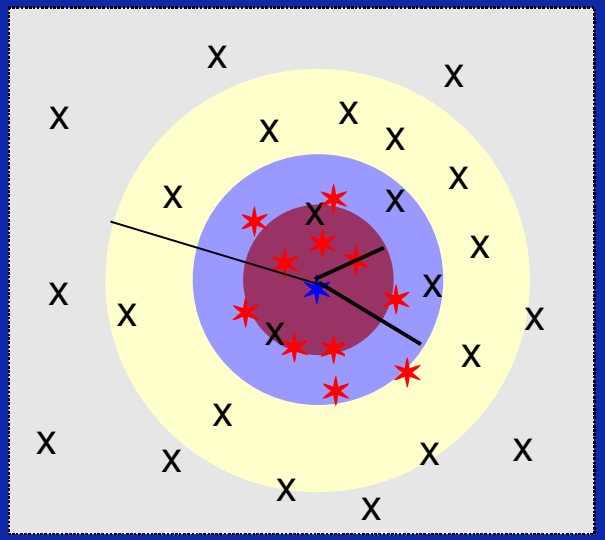


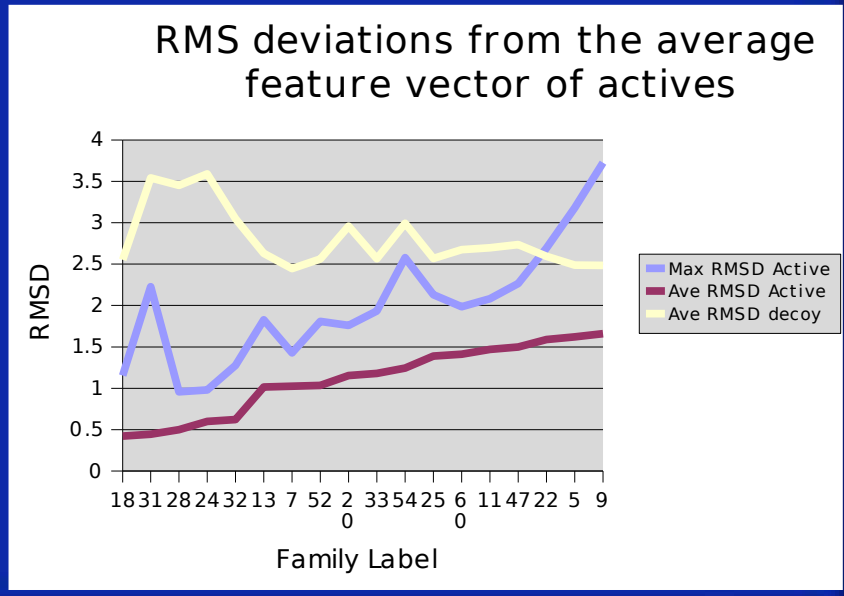
Step 3:
Screen your database using your knowledge base to find new leads
speed: millions / min.



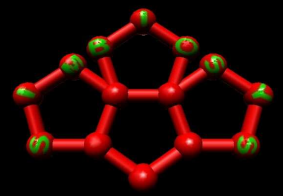
48. Diversity of Actives and decoys

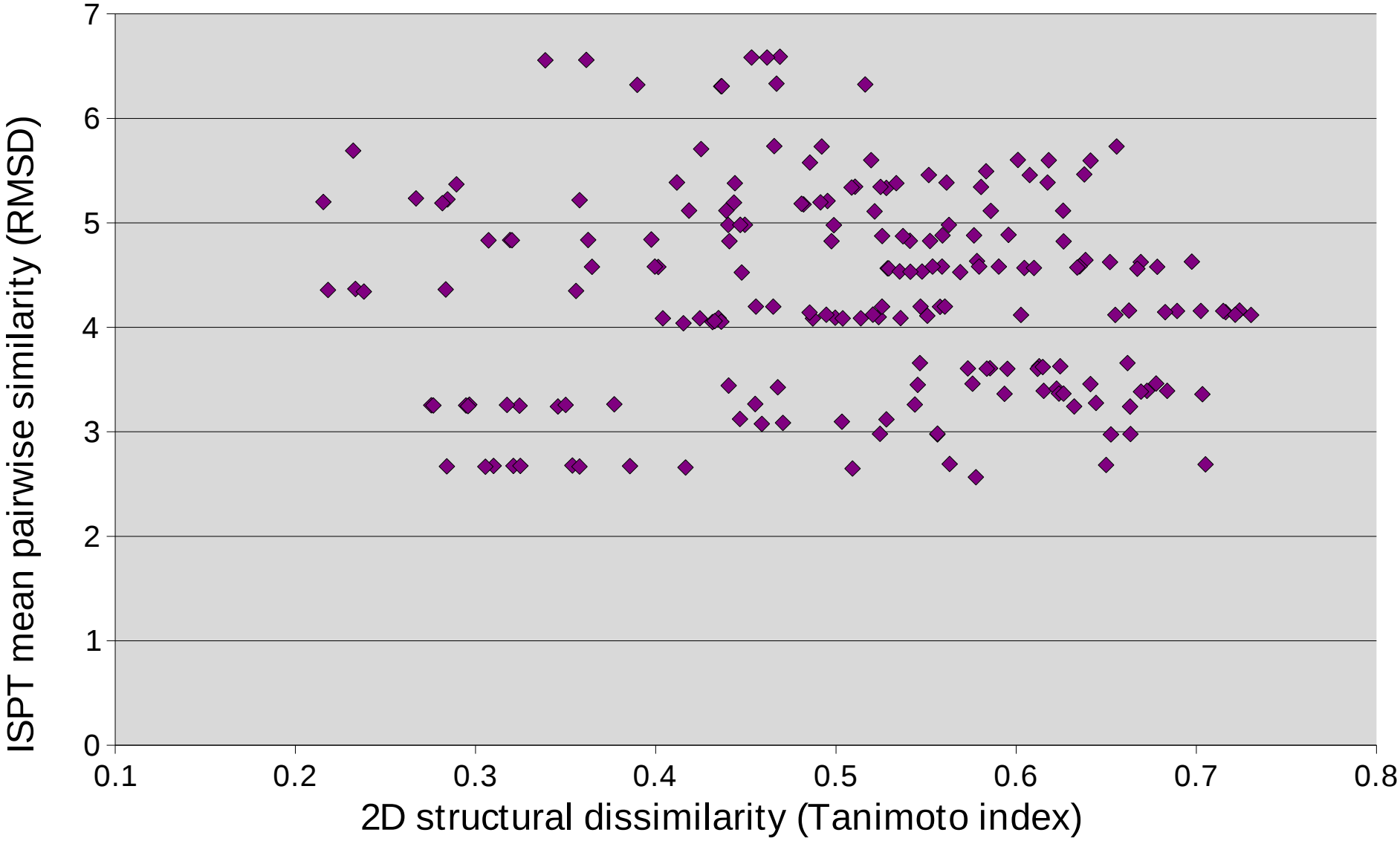
- For each set of actives, the average feature vectors was calculated (represented by the blue star)
- The RMSD from this feature vector was calculated for each active and decoy. The plot below shows the average RMSD for the actives and the decoys, as well as the MAX RMSD for the actives
- For 15 of the 18 codes even the max RMSD of the actives is less than the average RMSD of the decoys



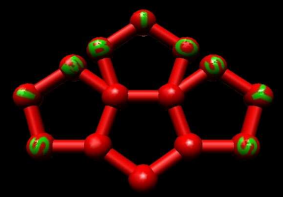


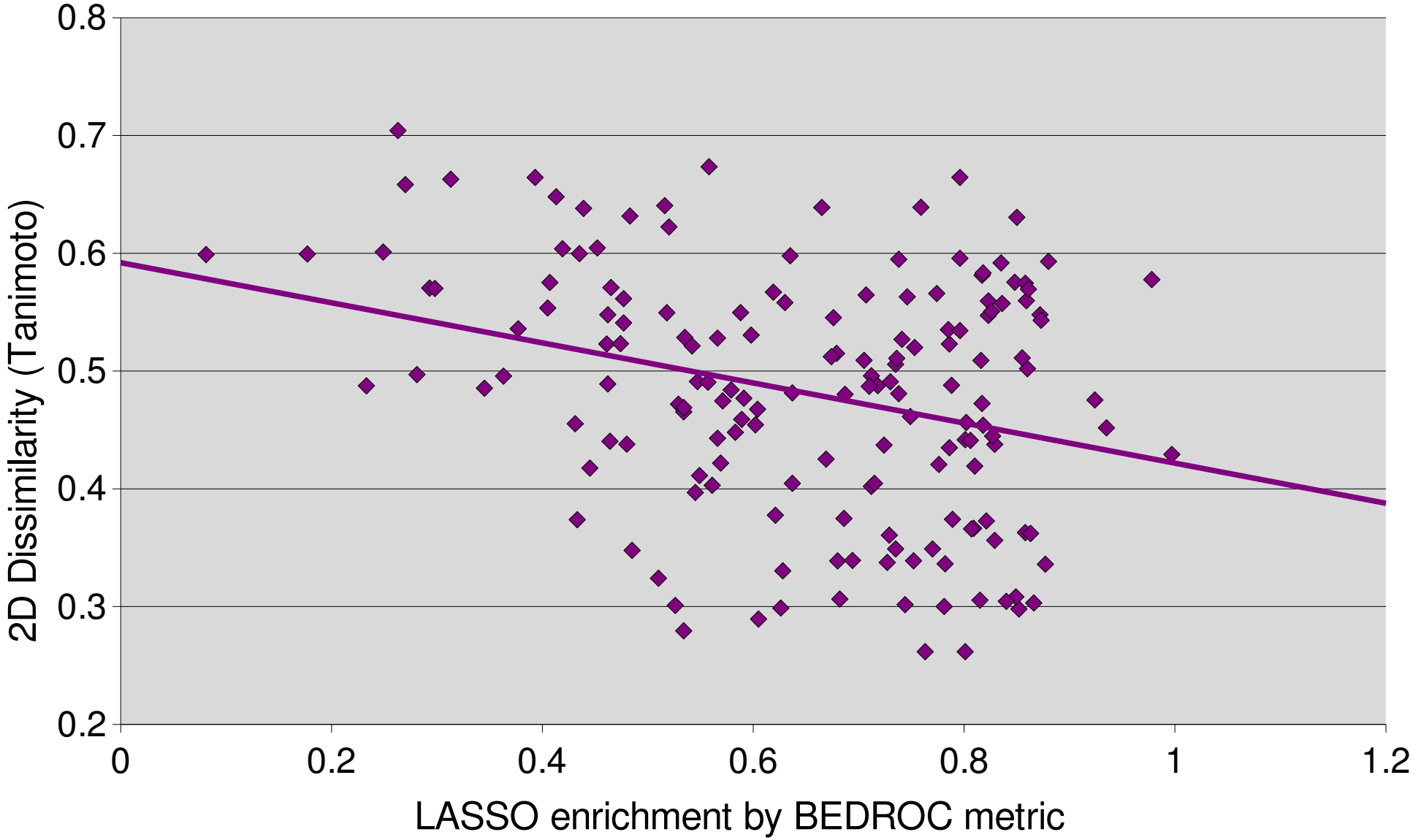
49. Structural diversity (Tanimoto) vs. diversity of LASSO descriptor



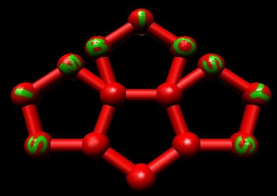


50. Scaffold hopping: 2D structure diversity versus LASSO enrichment results





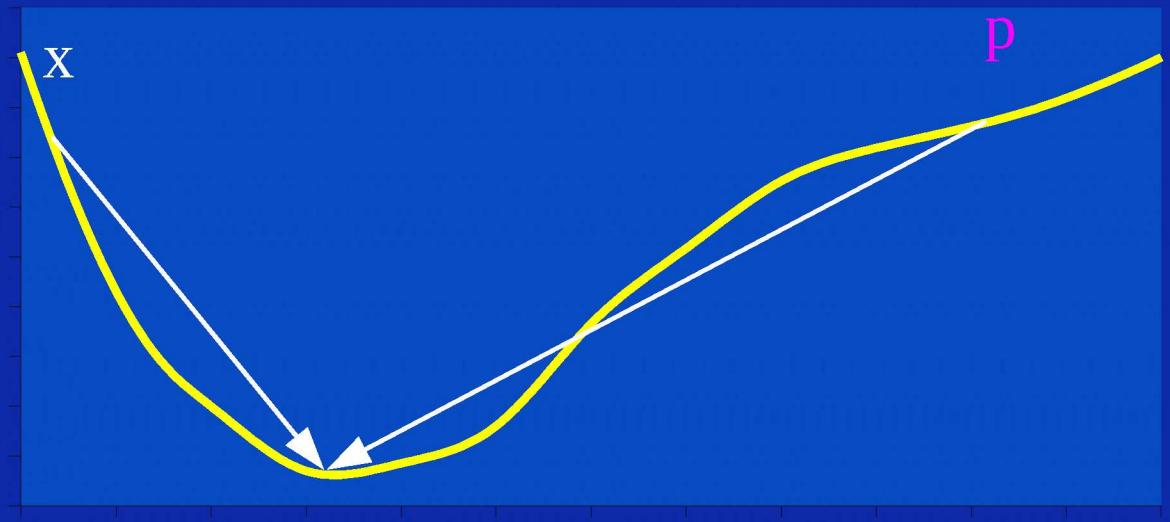
51. Contents: Results

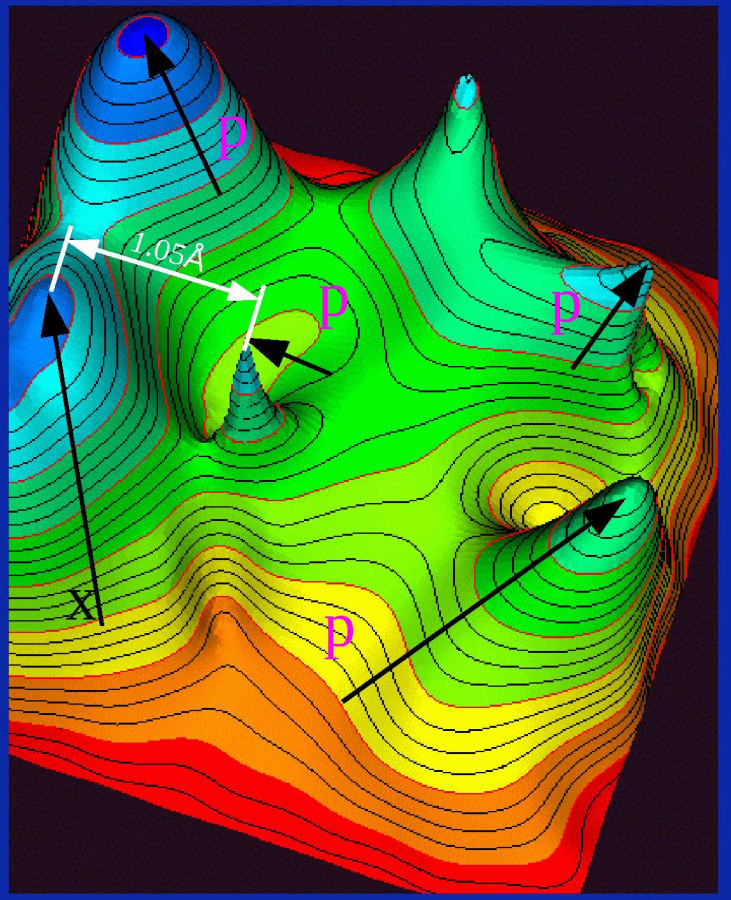


- The docking algorithm of the eHiTS software
- New statistically derived empirical scoring function
- LASSO: integrated VHTS filter
- **Results**
  - Warning about RMSD result comparisons
  - Validation results on 1568 diverse drug-like PDB complexes
  - Cross-docking validation results
  - Correlation of eHiTS-score with experimental binding data
  - Comparative study of various scoring functions
  - Screening, enrichment performance results
  - Speed performance in different modes

52. Warning about RMSD result comparisons

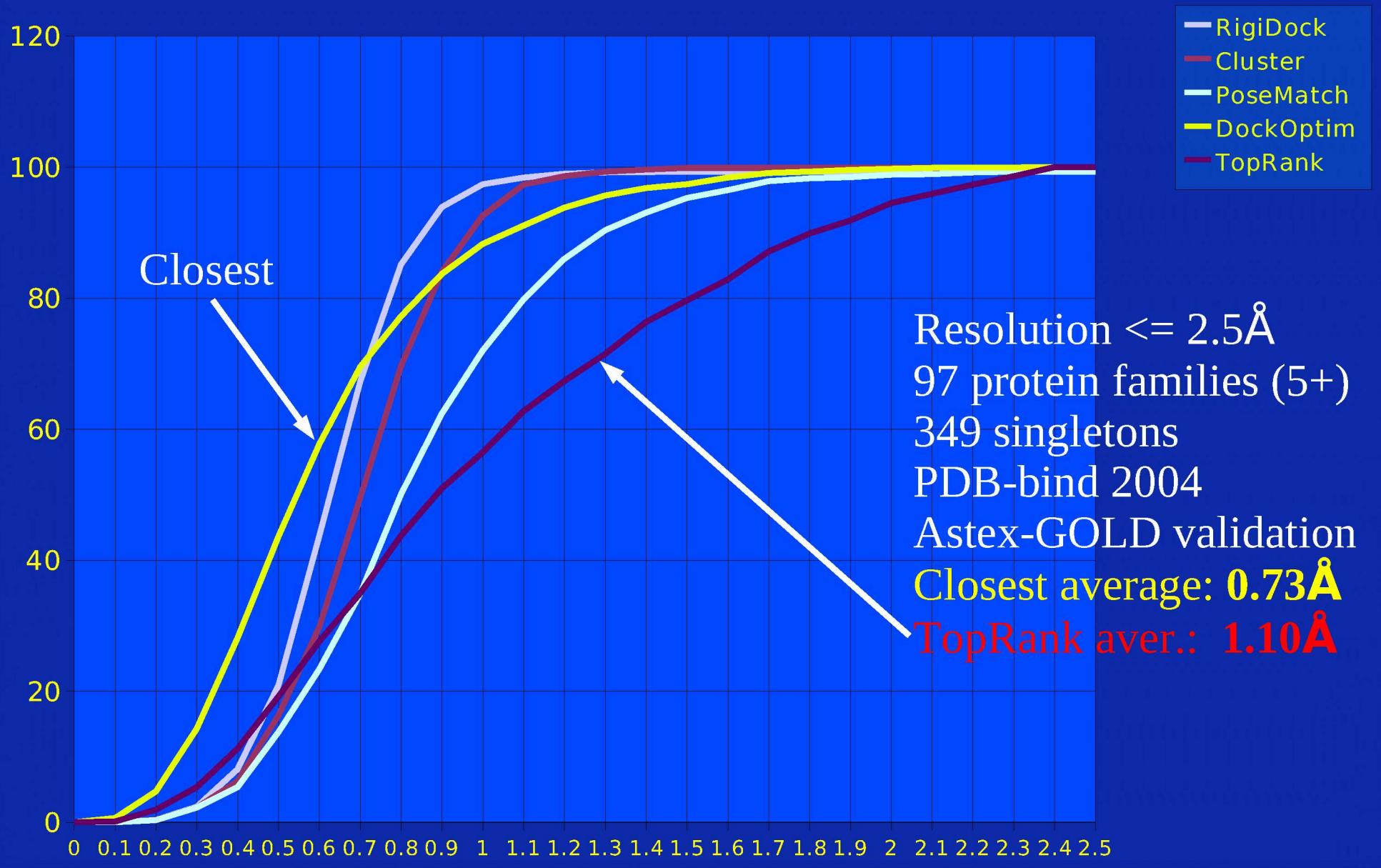
Some vendors publish RMSD results measured against pre-optimised ligand instead of the original X-ray structure





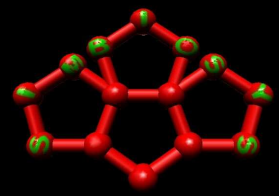
Such RMSD is a property of the scoring function shape (distance of two local minima) rather than a measure of docking accuracy. This number should be **ZERO**!

53. RMSD results on 1568 PDB complex with drug-like ligands



Resolution <= 2.5Å
97 protein families (5+)
349 singletons
PDB-bind 2004
Astex-GOLD validation
Closest average: **0.73Å**
TopRank aver.: **1.10Å**

23. Cross-docking results



- 2-3 PDB codes per family as listed in the publication referenced below

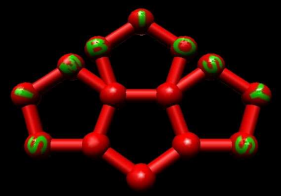
Cross-docking results for Glide are also taken from the publication (without induced fit)

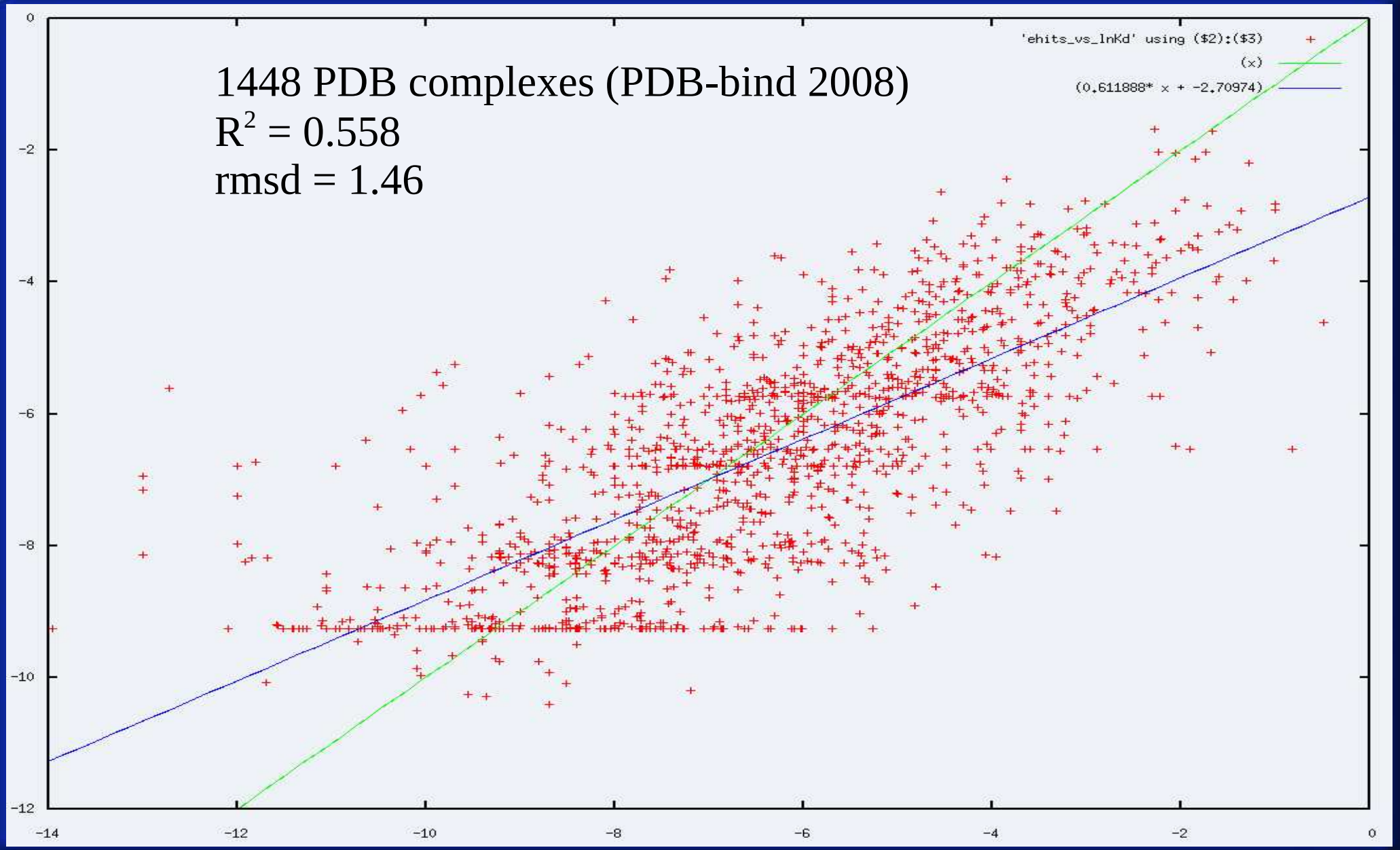
Last coulmn (Self) contains validation results of eHiTS

Family	Glide	EhiTS	Self
Aldose reductase	6.5	4.2	0.3
CDK-2	4.4	2.8	0.5
COX-2	8.8	4.5	1.2
Estrogen	3.8	1.5	0.7
Factor Xa	7.3	2.2	2
HIV-RT	7.2	1.4	0.6
Neuraminidase	2.8	2.3	1.8
PPAR	9.4	6.4	0.9
Thymidine kinase	2.6	2.7	0.6
Average:	5.87	3.11	0.96

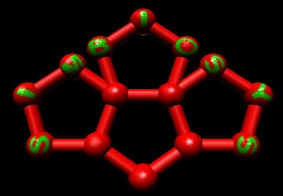
Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects
Woody Sherman, Tyler Day, Matthew P. Jacobson, Richard A. Friesner, and Ramy Farid
J. Med. Chem.; 2006; **49**(2) pp 534 - 553;

55. Correlation of eHiTS-score with experimental binding data





56. Results from a recently published scoring comparison study



J. Chem. Inf. Model., **2009**, *49* (6), pp 1568–1580

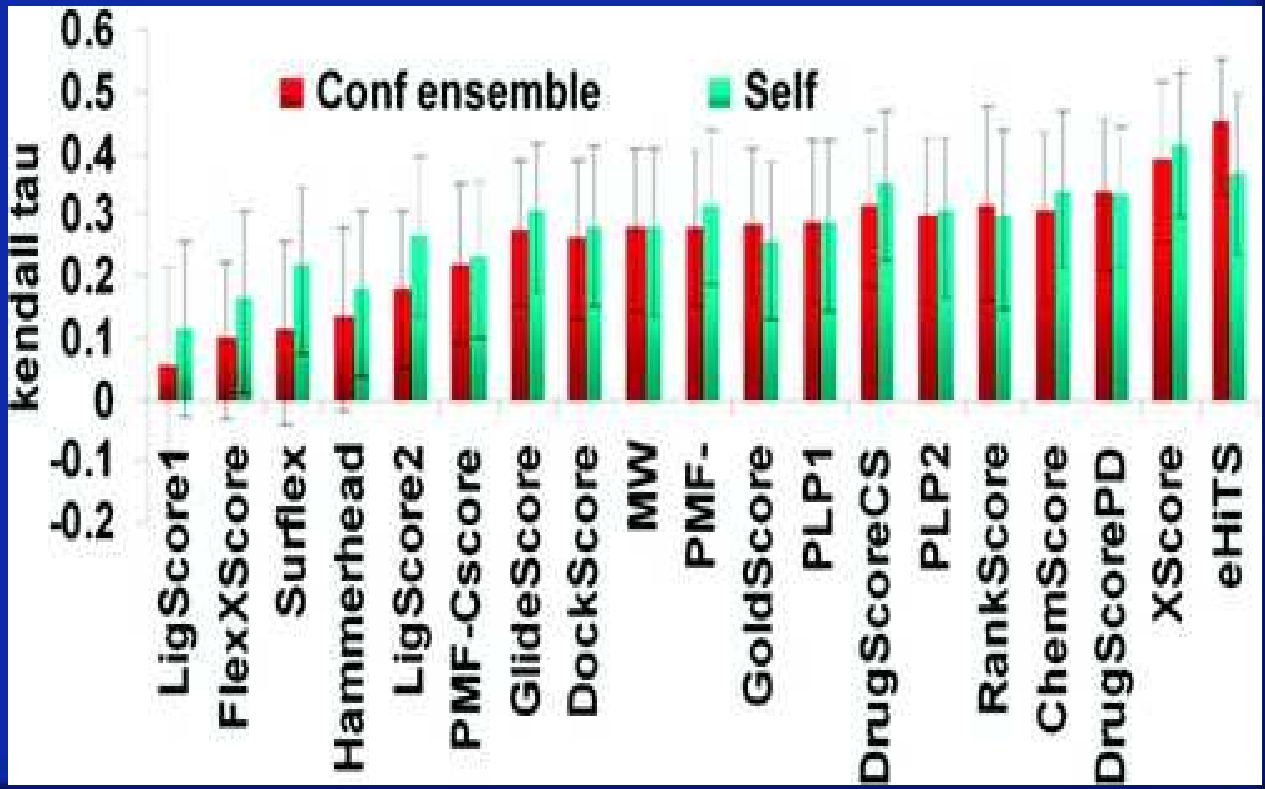
Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins?

Pablo Englebienne and Nicolas Moitessier

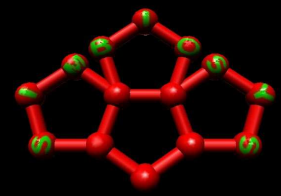
Publication Date (Web): May 15, 2009

DOI: 10.1021/ci8004308

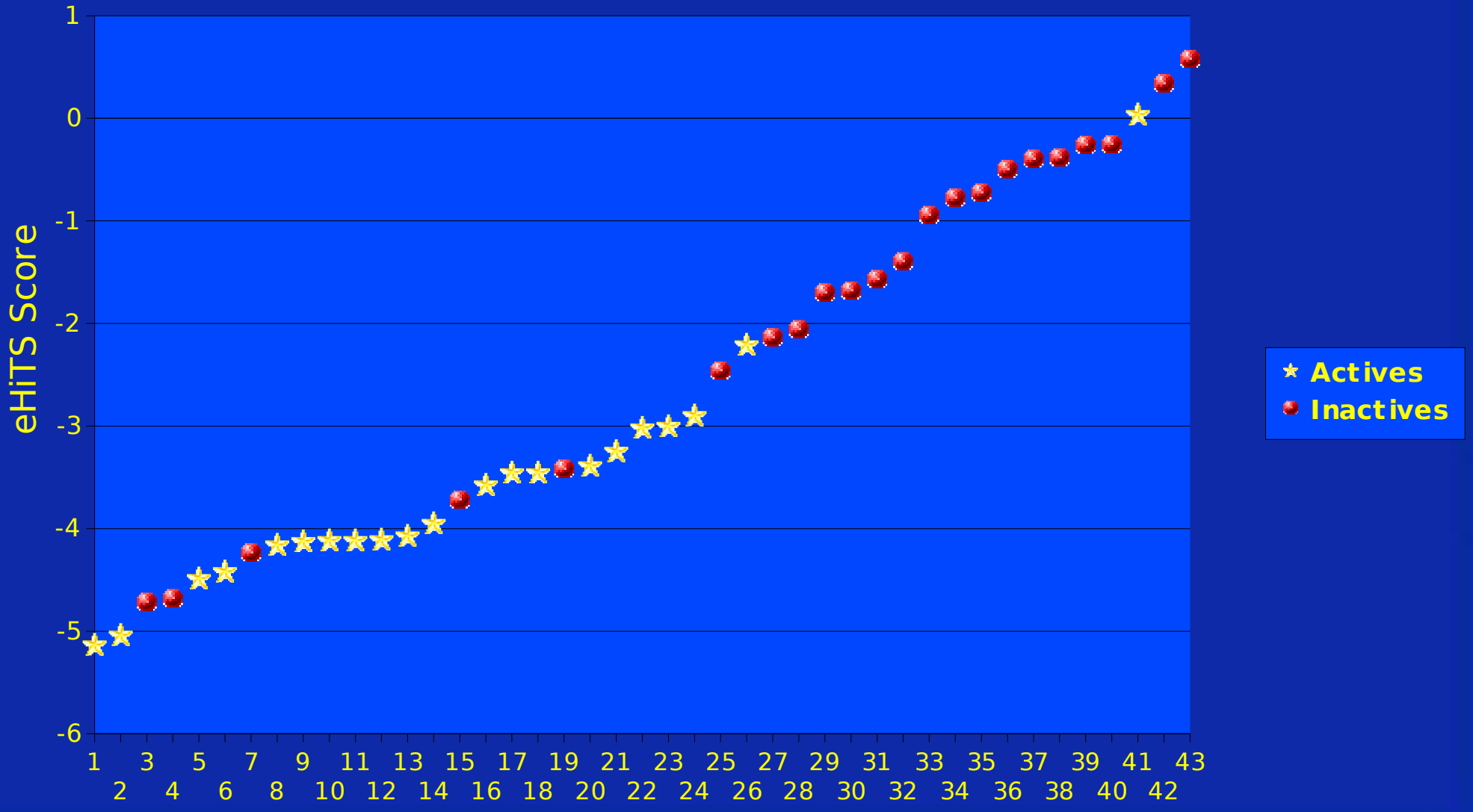
Note: the study used eHiTS v6.2 that was tuned for less families (90) and and on non-curated ~3K PDB complexes



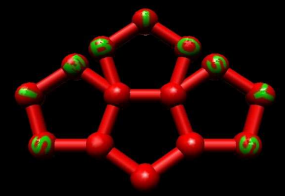
57. Screening success on CDK5

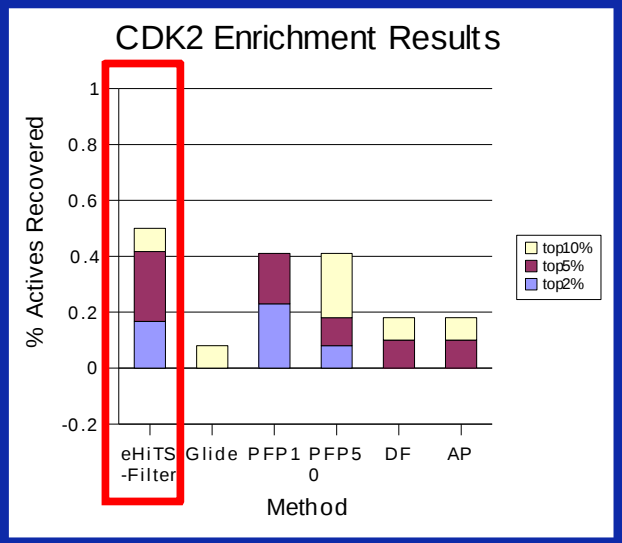


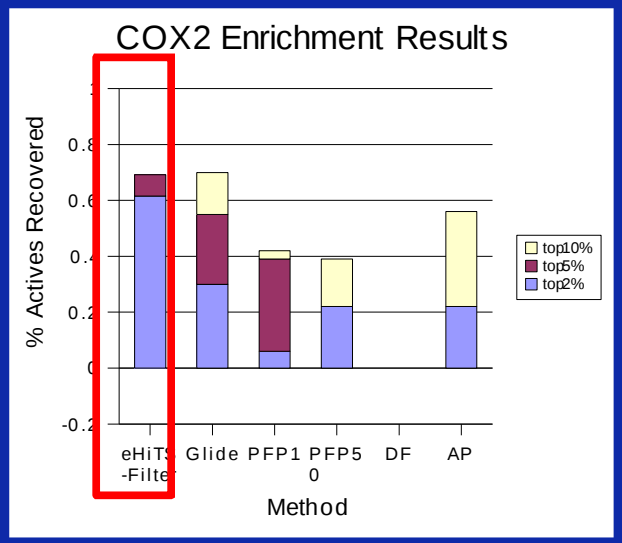
Kenneth R. Auerbach, Center for Neurologic Diseases, Harvard, Cambridge, MA, USA

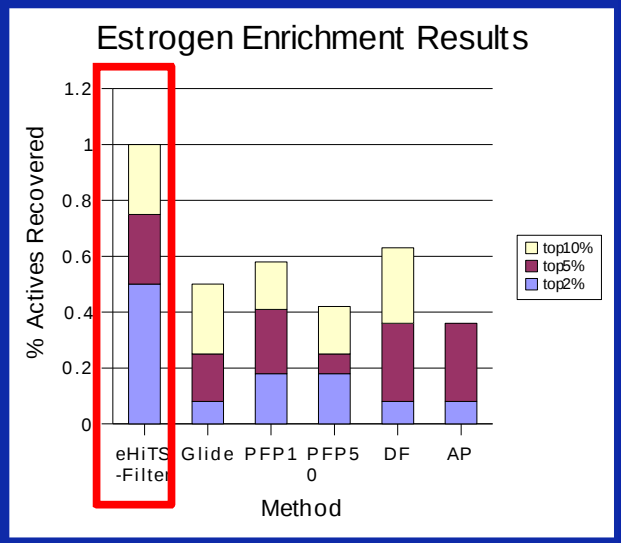


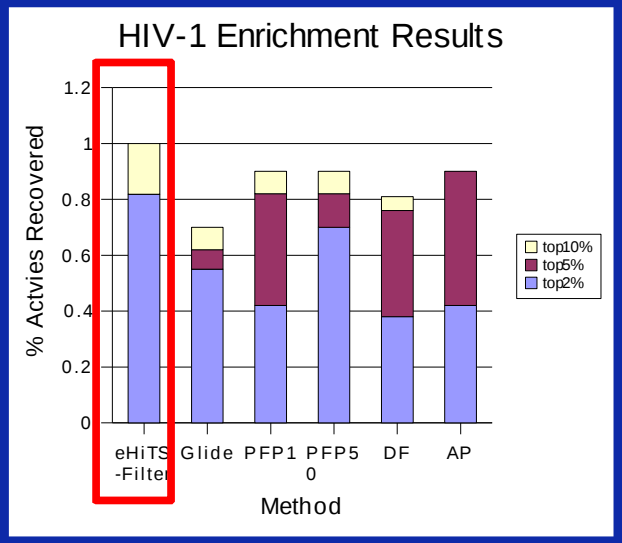
58. Enrichment result comparison on the Roche test set

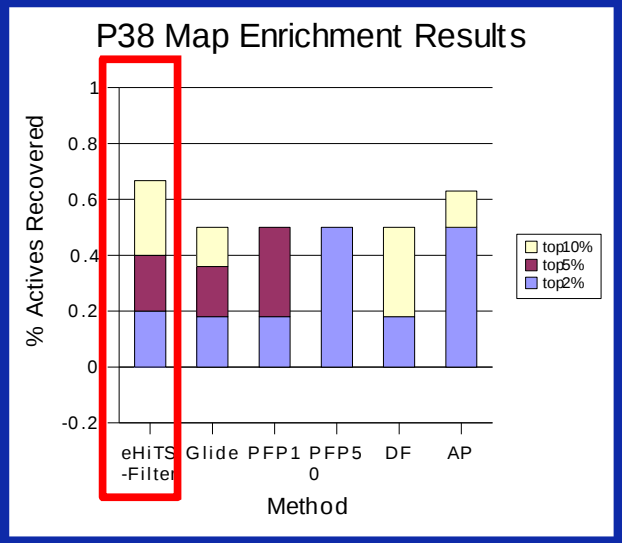


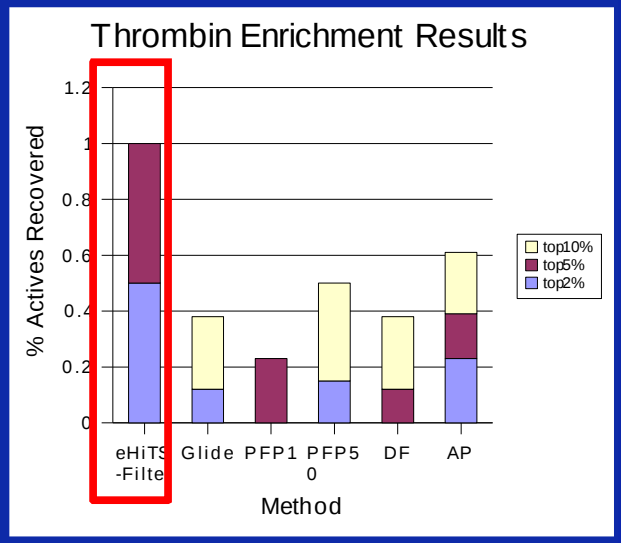




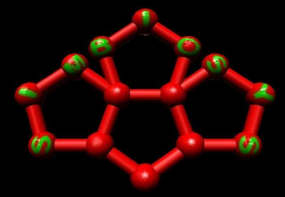






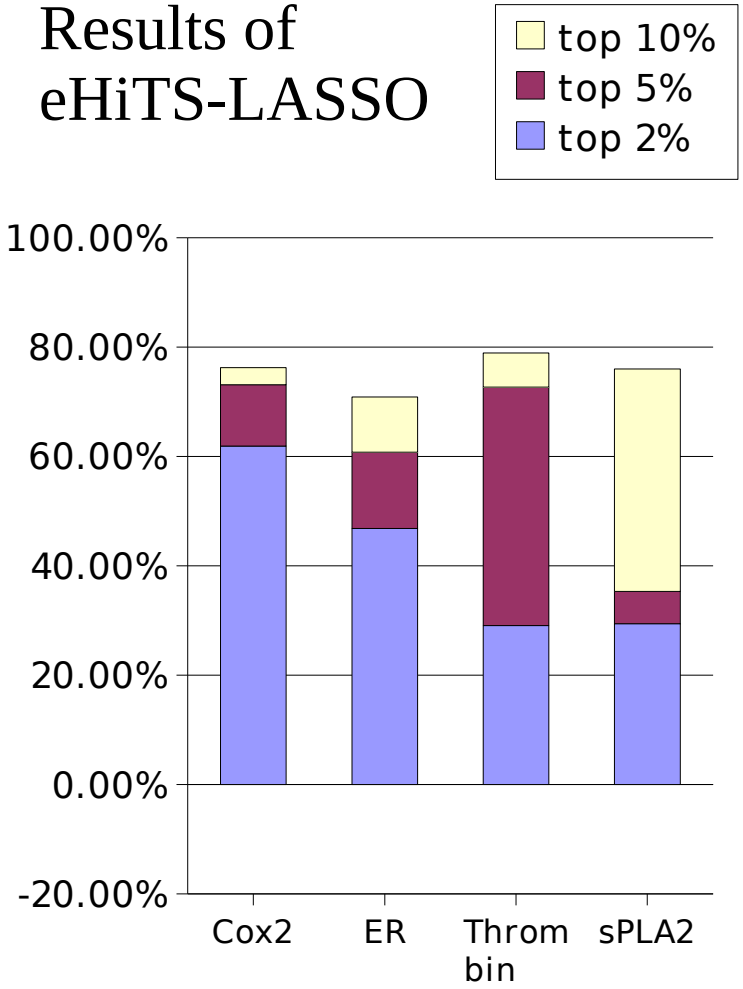


59. Enrichment results on the AstraZeneca data set

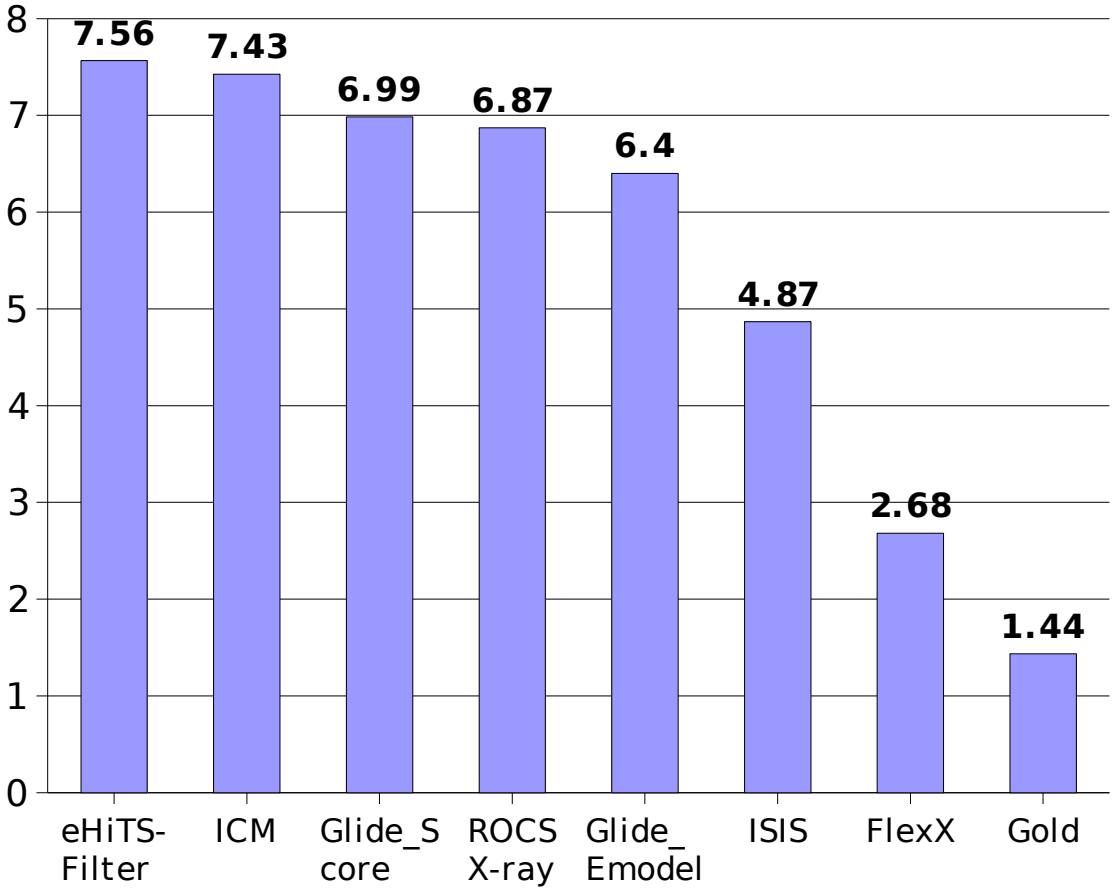


Hongming Chen et.al. , J. Chem. Inf. Model.; 2006; 46(1) pp 401 - 415

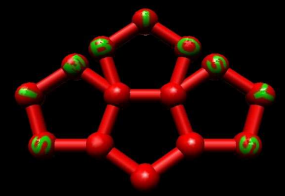
Results of eHiTS-LASSO



Average enrichment factor comparison:

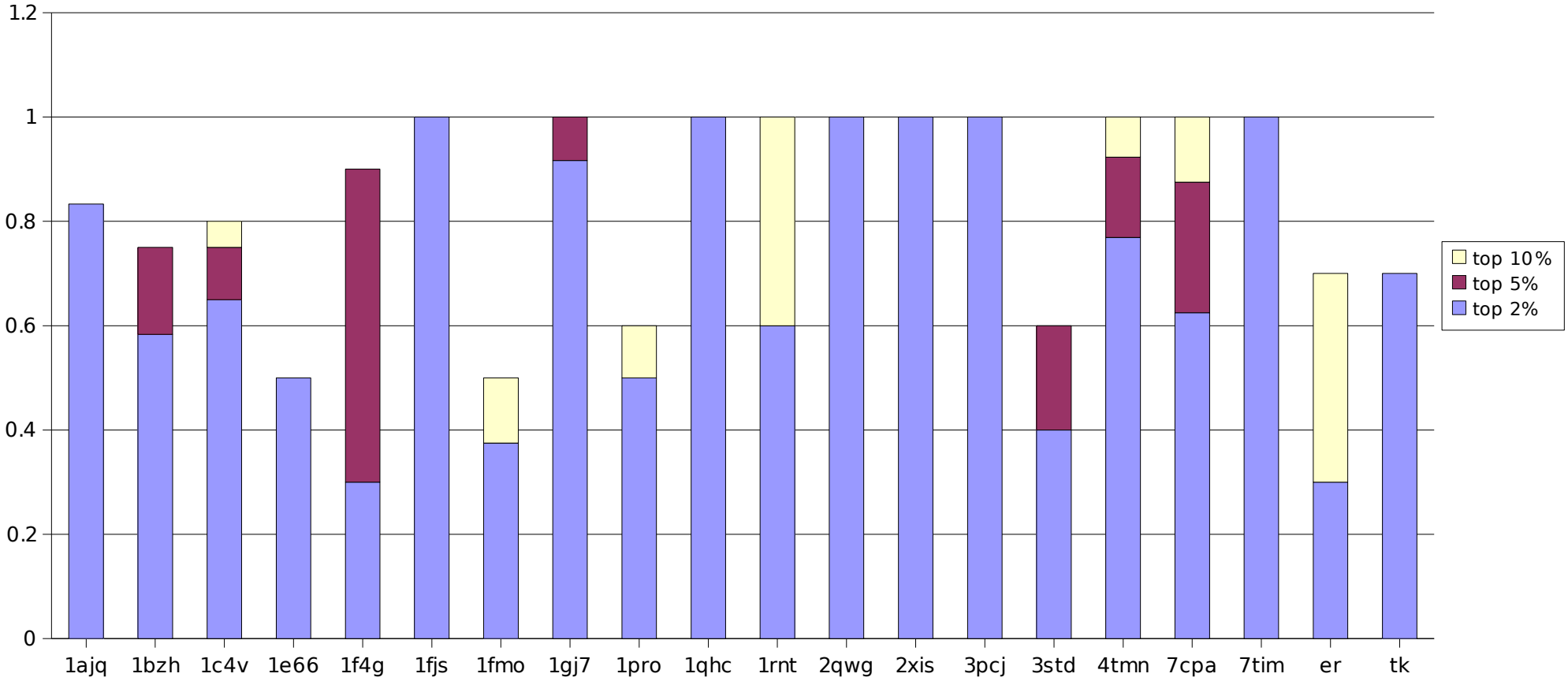


60. LASSO results on the Surflex set

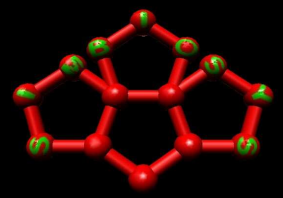


Test data taken from:

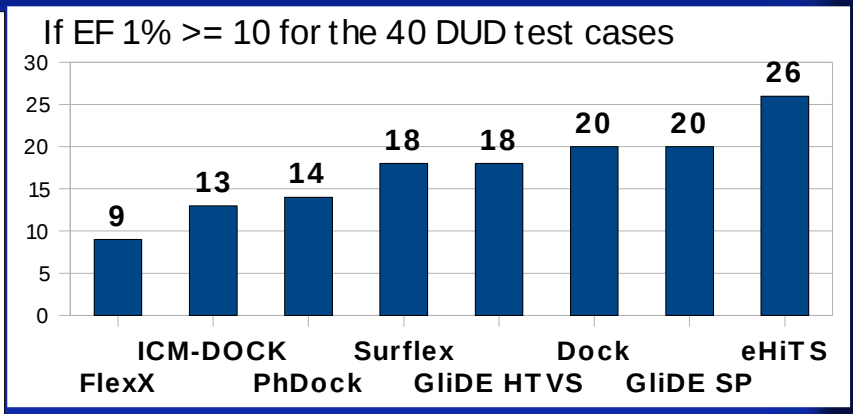
Pham, T.A. and Jain, A.N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data J. Med. Chem., 2005, 10.1021



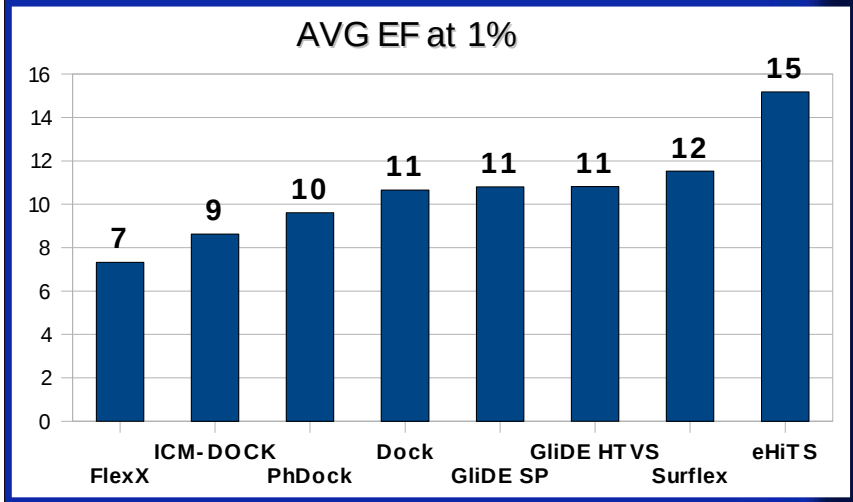
61. Enrichment results on DUD set (EF at 1% of database)



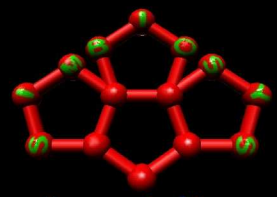
	DOCK	FlexX	GliDE HTVS	GliDE SP	ICM-DOCK	PhDOCK	Surflex	eHiTS
AR	3.9	6.4	18.0	12.8	15.4	24.4	7.7	7.7
Erag	7.6	18.2	19.7	7.6	25.7	22.7	9.1	7.6
Erant	13.6	16.3	1.6	16.3	19.1	10.9	16.3	23.7
GR	3.9	2.6	2.6	3.9	14.2	16.8	16.8	5.19
MR	0.0	7.2	7.2	21.7	36.2	43.5	36.2	14.29
PARP	30.7	30.5	6.1	15.4	5.9	0.0	21.3	29.41
PR	4.0	0.0	0.0	0.0	27.7	7.9	4.0	23.08
RXR	33.0	5.5	33.0	5.5	33.0	33.0	11.0	26.32
CDK2	11.4	8.5	14.2	8.5	0.0	9.9	11.4	14.1
EGFr	22.2	12.5	19.7	16.5	0.6	3.4	4.9	6.8
FGFr1	15.2	0.9	0.9	3.4	11.9	3.4	4.2	2.5
HSH90	5.5	0.0	0.0	0.0	0.0	2.8	2.8	8.3
P38	12.0	2.0	1.3	2.9	4.2	2.0	5.6	18.32
PDGFrb	10.7	5.9	0.0	5.9	3.6	3.0	1.2	14.79
SRC	24.8	7.0	14.6	23.6	13.4	5.7	6.4	27.9
TK	0.0	0.0	4.7	0.0	0.0	23.4	9.3	0.0
VEGFr2	17.6	9.4	16.4	9.4	8.2	3.5	10.6	4.6
FXA	13.2	24.4	15.3	20.2	5.6	10.4	17.4	6.9
Thrombin	9.8	8.4	19.7	14.1	1.4	14.1	2.8	32.39
Trypsin	12.4	0.0	18.5	20.9	6.2	0.0	8.2	10.42
ACE	12.6	8.4	4.2	12.6	2.1	0.0	6.3	12.5
ADA	0.0	0.0	8.3	16.6	2.8	0.0	11.0	18.4
COMT	21.9	0.0	21.9	11.0	0.0	11.0	0.0	10.0
PDE5	15.3	1.2	7.1	3.5	8.2	0.0	11.8	22.99
DHFR	8.4	16.8	8.1	14.5	17.8	2.2	18.2	17.6
GART	2.6	5.2	15.5	20.6	0.0	0.0	10.3	0.0
Ache	3.7	1.9	3.7	0.9	0.0	2.8	2.8	9.4
ALR2	27.5	0.0	3.9	7.9	7.9	3.9	19.6	28.0
AmpC	14.4	0.0	0.0	4.8	0.0	0.0	0.0	5.0
COX-1	0.0	12.5	12.5	8.3	16.6	25.0	4.2	0.0
CoX-2	0.7	1.7	29.5	29.4	3.3	1.2	16.5	24.7
GPB	0.0	0.0	21.2	3.9	1.9	5.8	9.7	29.4
HIVPR	6.5	0.0	4.8	14.5	6.5	0.0	8.1	9.84
HIVRT	2.4	0.0	14.5	12.1	14.5	4.8	14.5	14.29
HMGA	17.4	14.5	23.2	23.2	0.0	37.7	40.6	26.47
InhA	24.8	27.2	4.7	13.0	1.2	1.2	4.7	12.9
NA	4.1	4.1	14.5	0.0	26.9	22.8	26.9	10.4
PARP	12.2	30.5	6.1	6.1	3.1	0.0	21.3	27.38
PNP	0.0	0.0	8.8	4.4	0.0	8.8	17.5	12.24
SAHH	0.0	3.3	6.5	16.2	0.0	16.2	9.7	31.25
AVG	**10.65**	**7.33**	**10.81**	**10.8**	**8.63**	**9.61**	**11.52**	**15.18**

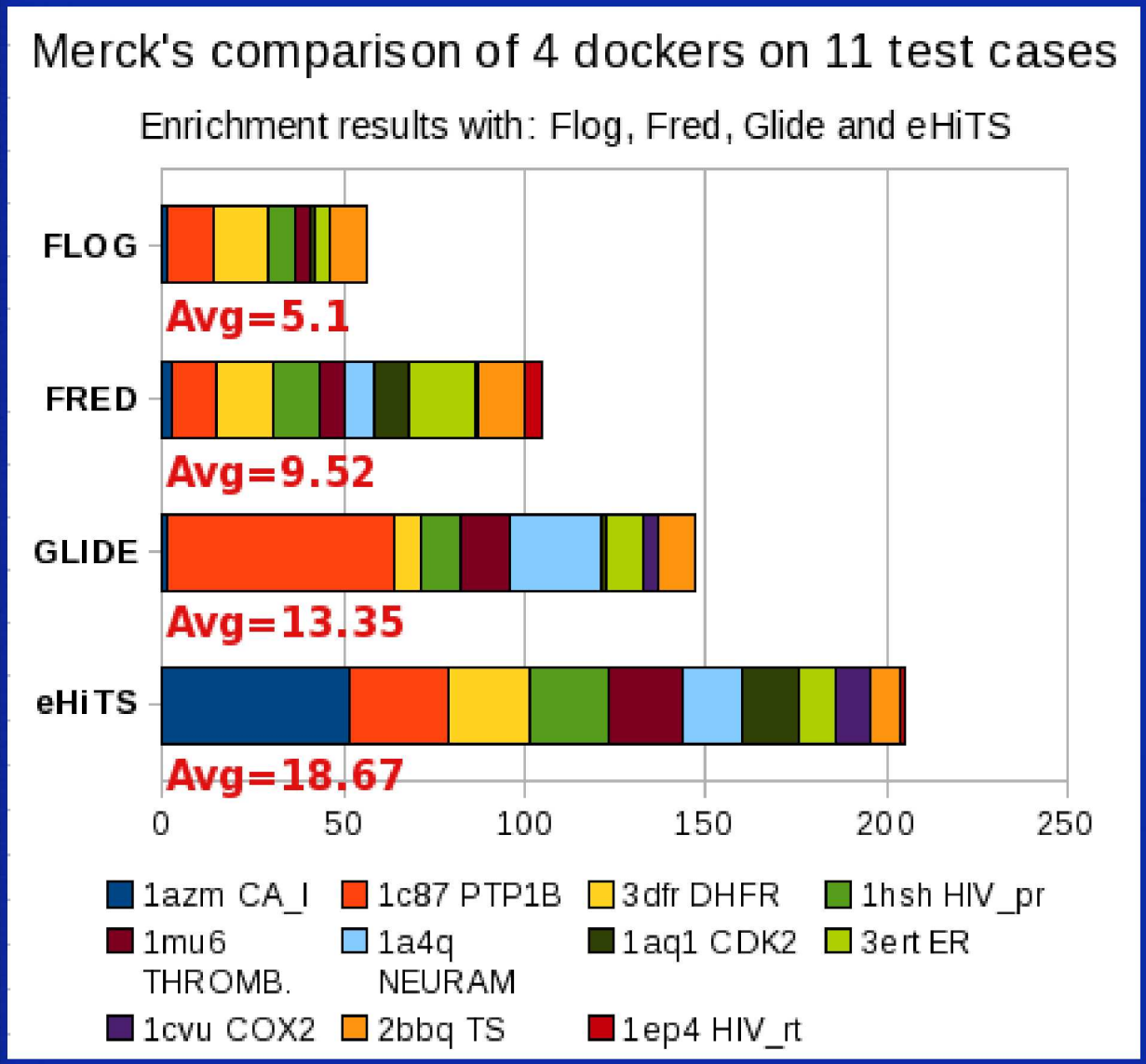


235th ACS New Orleans, COMP 144: "Comparison of pose generation and virtual screening accuracy for several molecular docking programs", Jason B. Cross, David C. Thompson, Brajesh K. Rai, J. Christian Baber, Kristi Yi Fan, Yongbo Hu, and Christine Humblet. Wyeth



62. Screening comparison by Merck

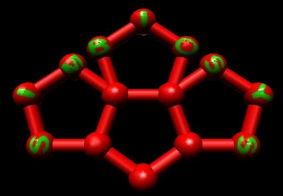




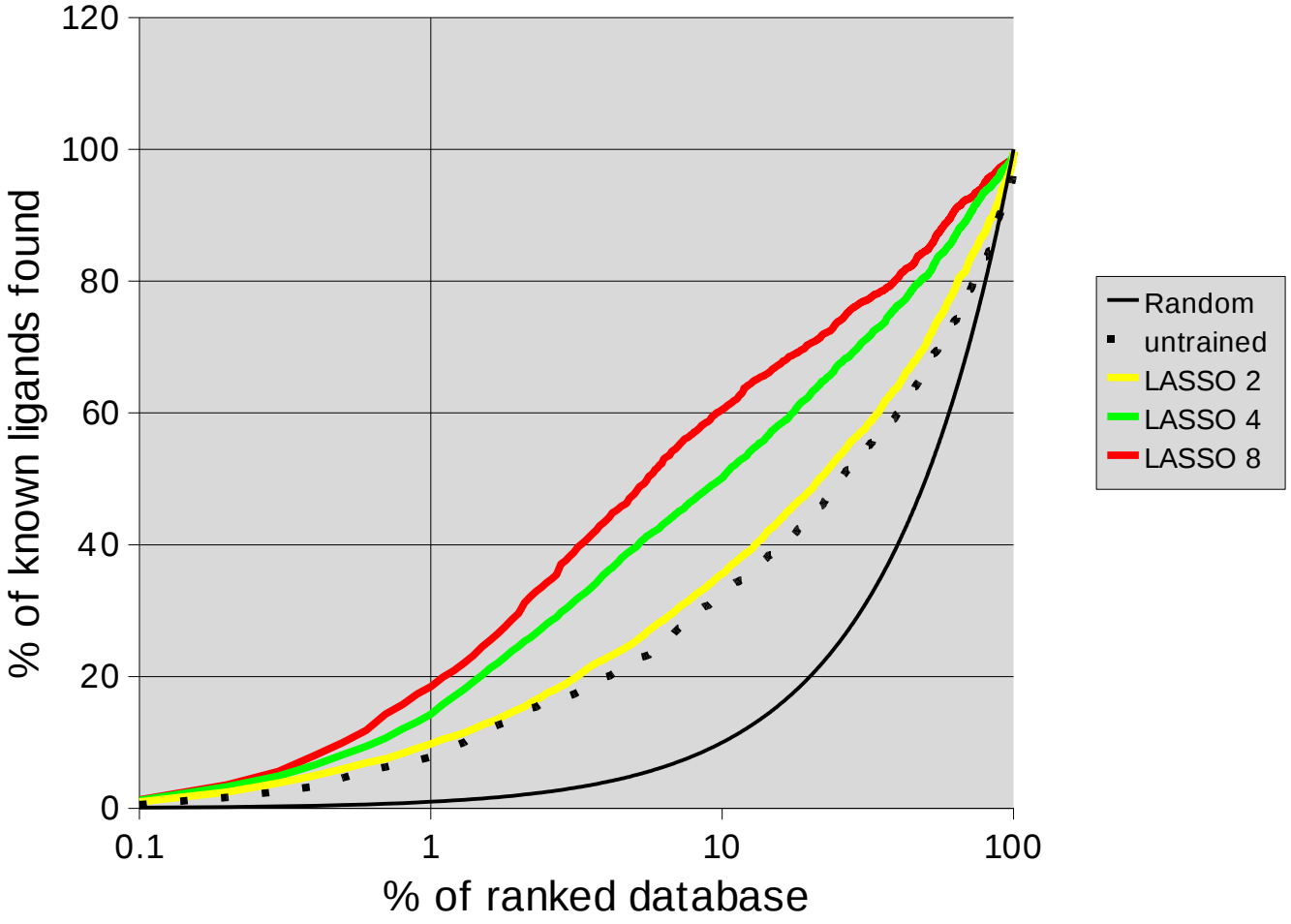
Data and competitor results from the paper:

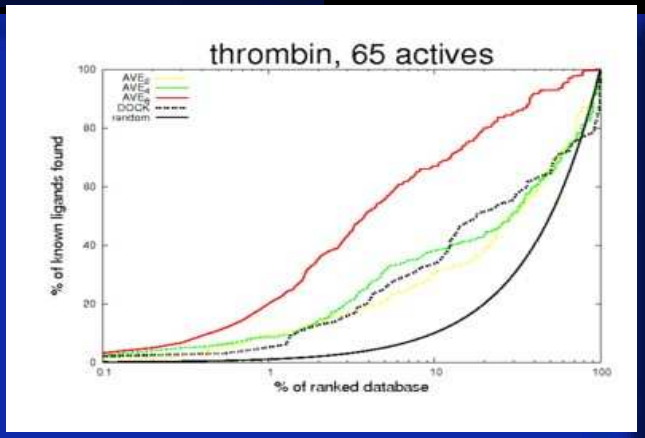
J. Chem. Inf. Model. 2007; 47(4), pp 1504 - 19
DOI: 10.1021/ci700052x

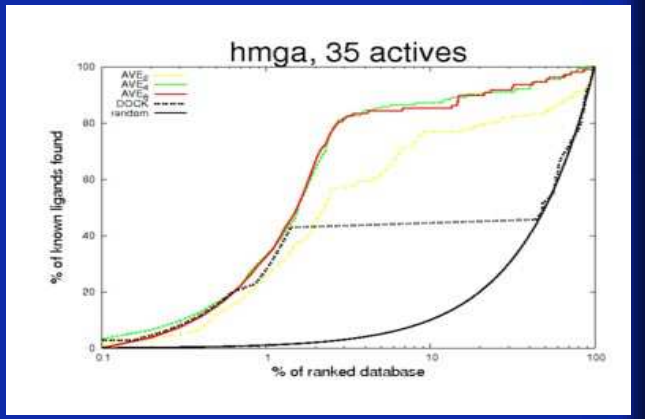
eHiTS results were not included in the publication. They were generated in collaboration with Merck after the paper submission

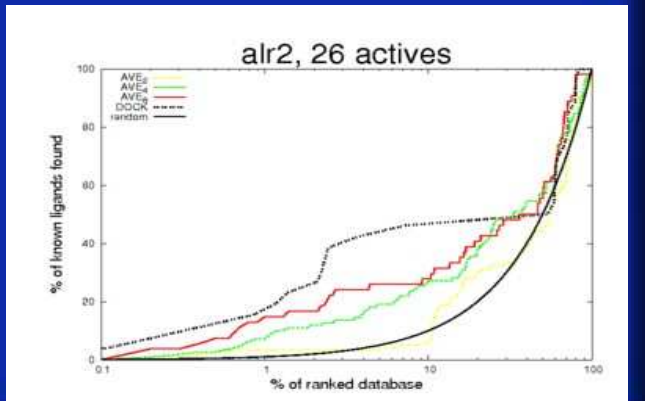


63. Effect of enrichment training (NN)

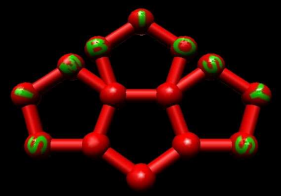








64. eHiTS speed

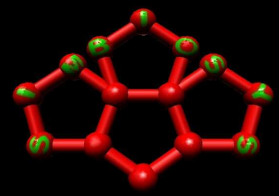


500 drug-like ligands from the ZINC DB were docked into 1err (estrogen) receptor. Table shows average CPU time in seconds per ligand on a 1.5GHz Pentium4 Linux PC

Accuracy:	1	2	3	4	5	6
Default	120	140	162	173	198	240
SQL DB	65	82	106	120	150	183
HTS param.	16	20	23	27	29	35

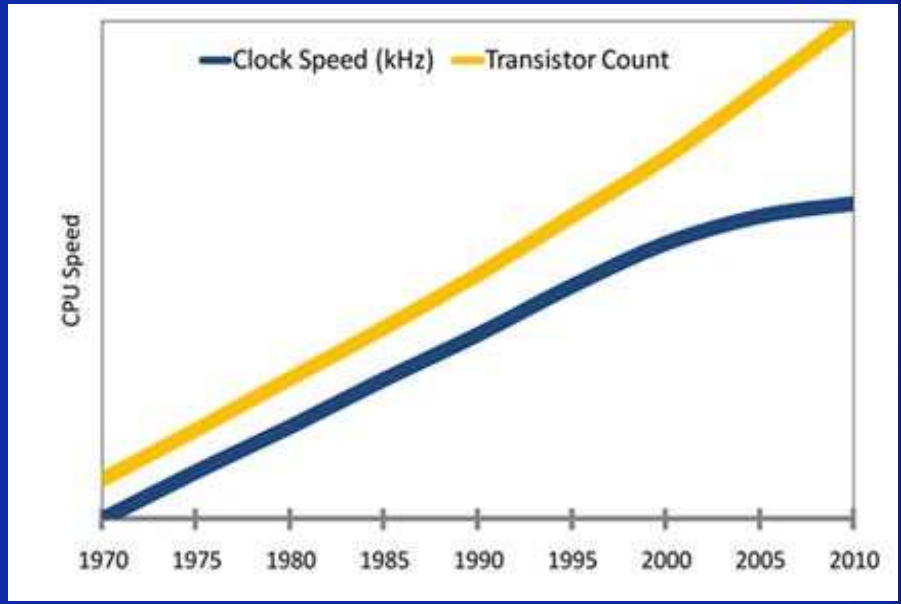
SQL DB: Dock-Table functionality, stored rigid fragment poses
HTS: special parameter set supplied with eHiTS for screening runs

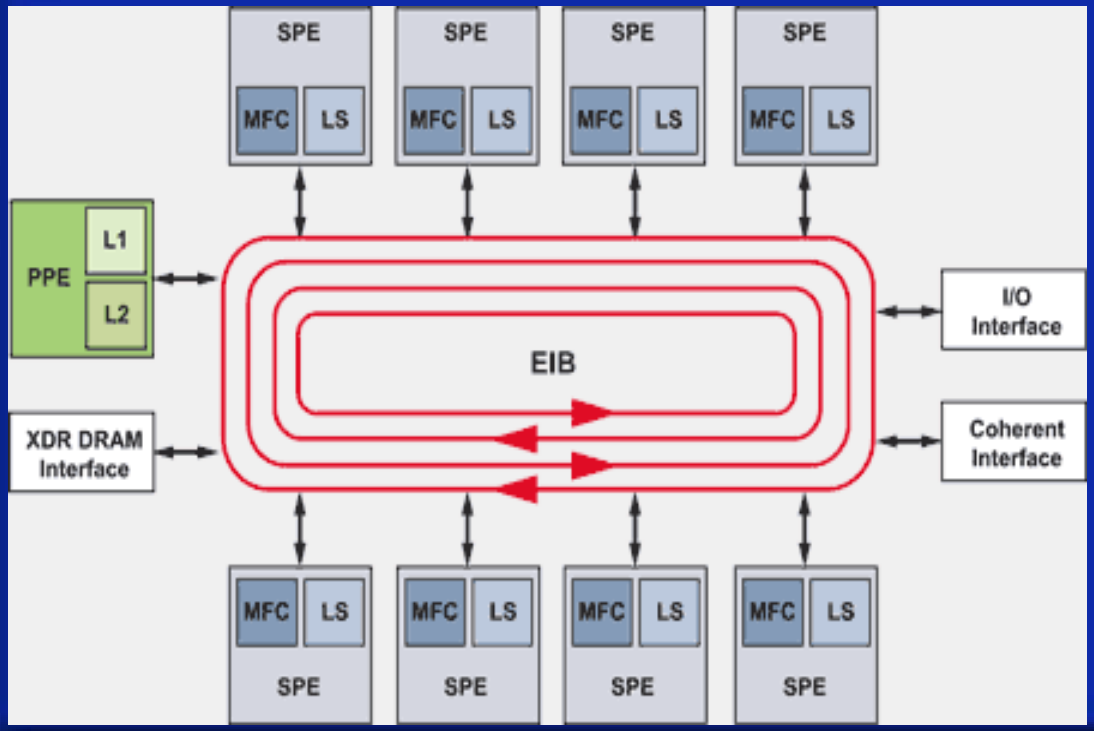
LASSO can filter tens of thousands of ligands per second



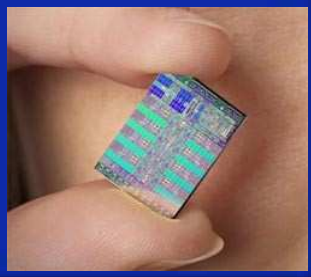
65. Hardware acceleration: Cell/BE

- Moore's law: CPU performance has doubled every 18 months
- In the last 5 years clock speed stuck at ~3GHz, increase # cores
- 8 cores provide 8X speed-factor
- Dual pipe 128bit SIMD => another 8X
- **8 * 8 * 3.2GHz =204.8GFlops**

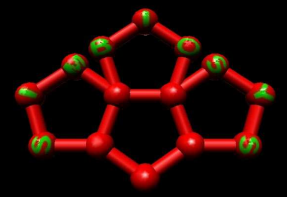




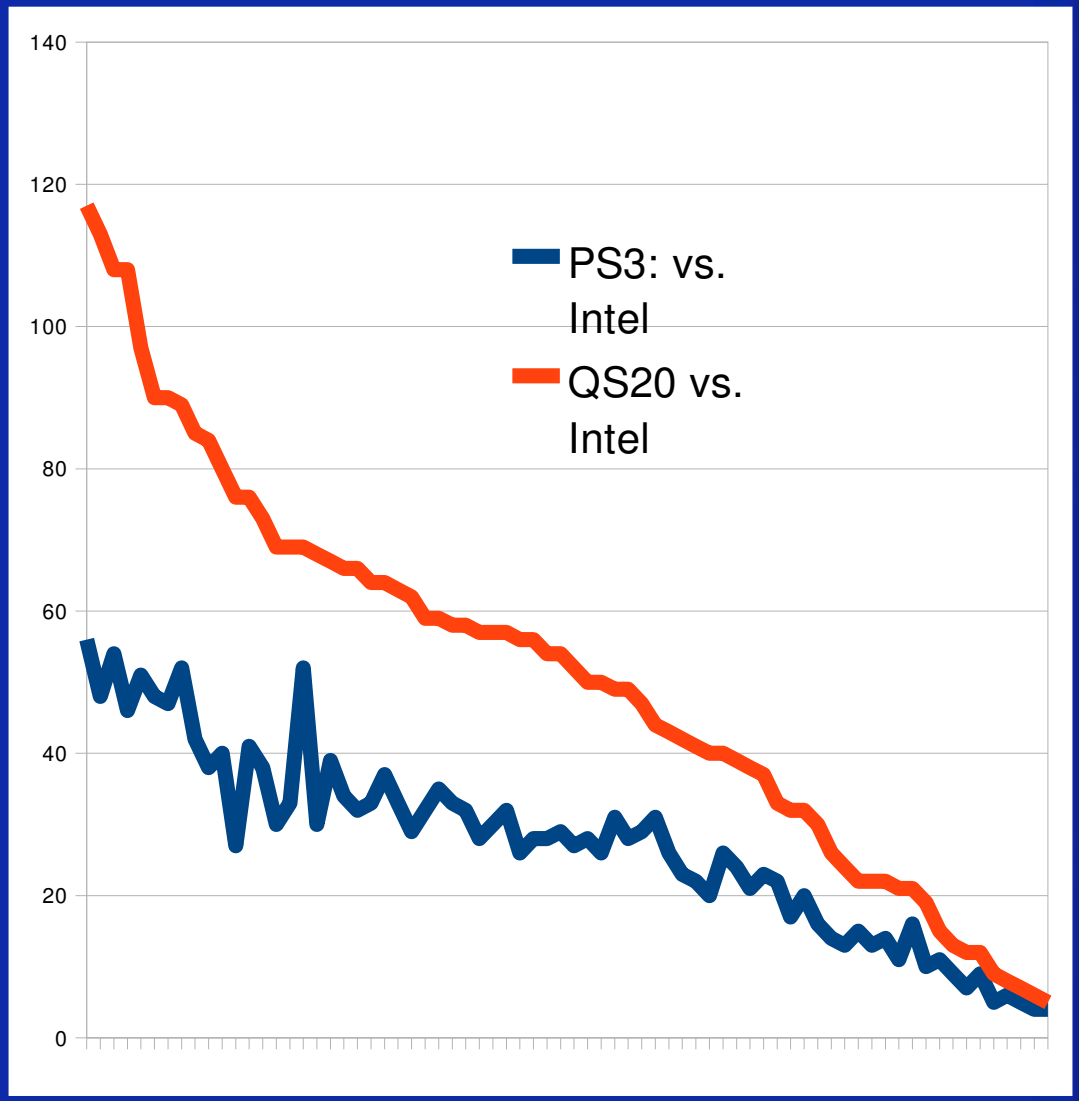
It was designed for gaming,
but we can use it for Science !



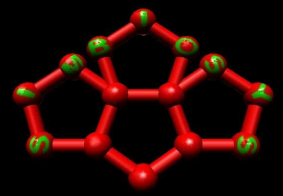
66. eHiTS Lightning
Speed-up Results: up to 117X



	run time	speed up
Intel avg.	221s	n/a
PS3 avg.	7s	32x
QS20 avg.	4s	62x
Intel max.	913s	n/a
PS3 max.	31s	56x
QS20 max.	16s	117x



67. Summary



- Exhaustive, high accuracy - don't miss a potential drug!
- Deterministic – no need to wait for a lucky day
- Fully automated - no manual setup, protonation, charges
- Integrated LASSO VHTS filter (>1 million ligands/minute/CPU)
- Interaction Surface Point based statistical interactions scoring
- Additional scoring terms combined with empirical weight set
- Automated protein family clustering and specialized weight tuning
- Per-optimized weight sets for over 500 families included
- Automated tuning tool to customize the scoring for in-house data
- Very fast – electronic High Throughput Screening with parallel/cluster/accelerator support: PBS, LSF, SGE, and available on the Cell B.E. supercomputer in a chip