

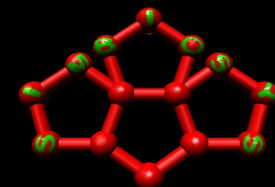
How eHiTS solves the docking and scoring problems

Zsolt Zsoldos

SimBioSys Inc., © 2010

Booth #945

Ten lessons learned during
ten years of eHiTS docking and scoring development

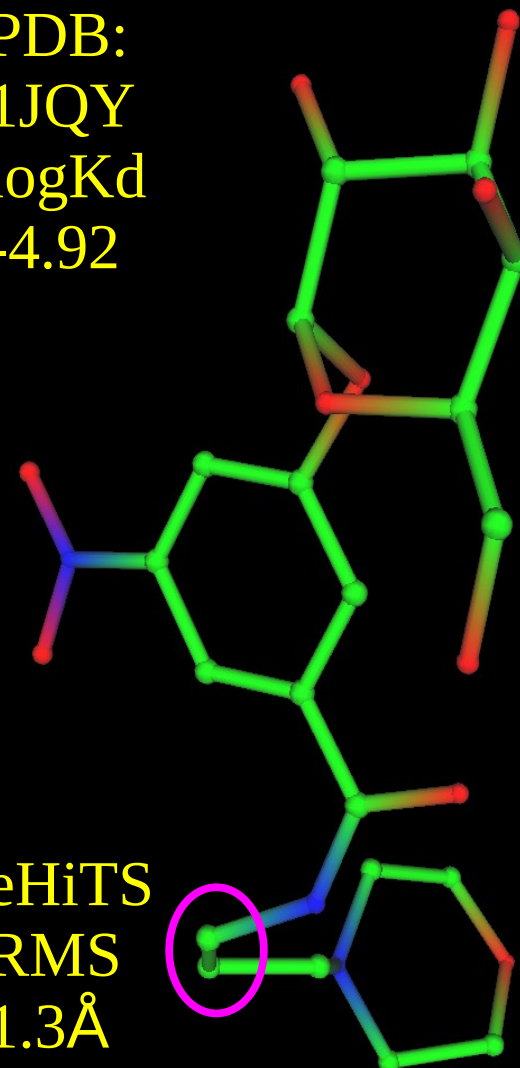


1. It is not sufficient to sample a few dozen low energy conformers

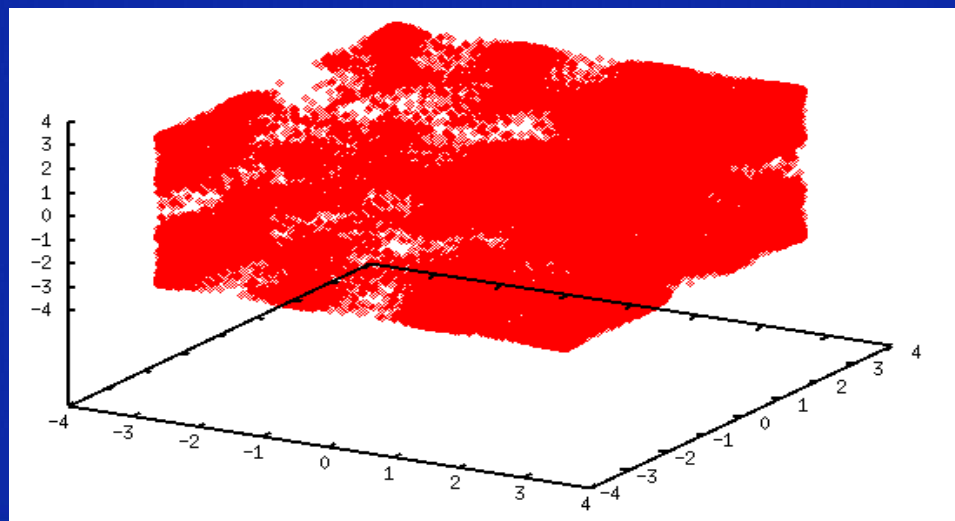
- High energy, strained torsions appear with high frequency (37%) in PDB
- Sampling every 60° for each rotatable bonds miss 97% of X-ray conformations by more than 5° error
- Dihedral angles are fully scattered and fill the whole range

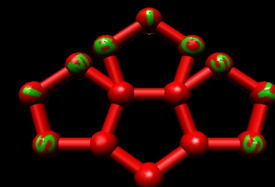


PDB:
1JQY
logKd
-4.92



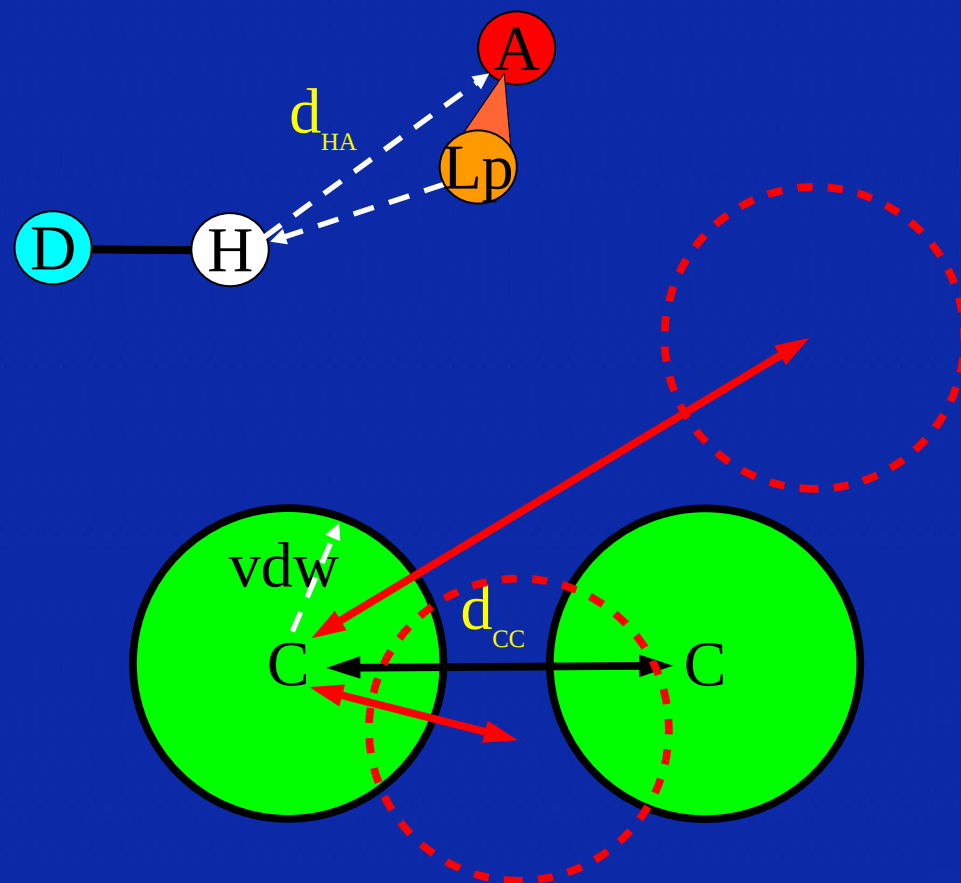
eHiTS
RMS
1.3Å

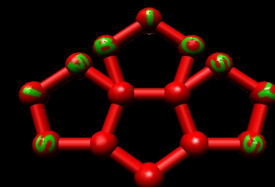




2. The search space is vast: sampling precision requirements

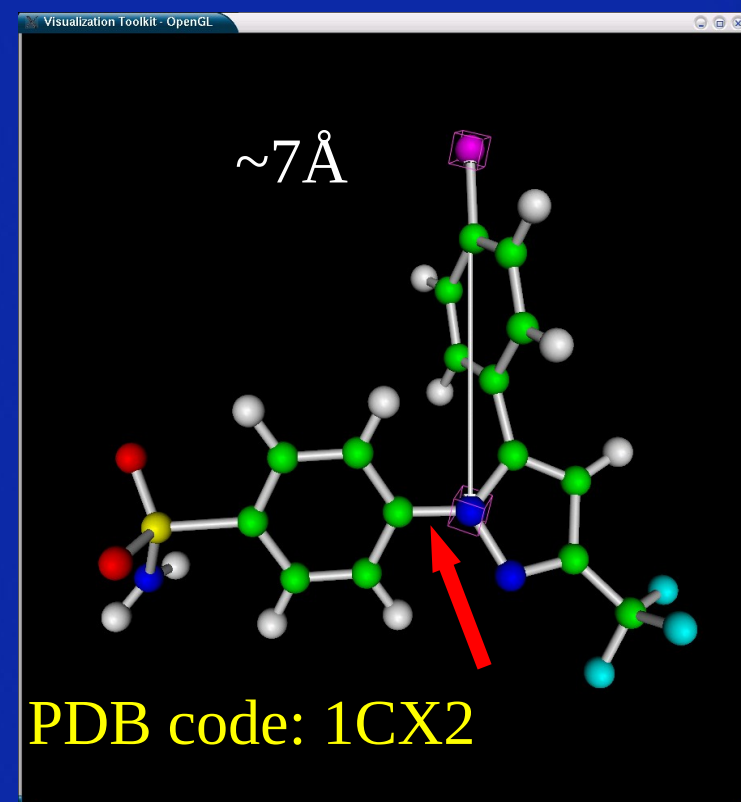
- H-bond geometry
H-acceptor distance range
 1.6\AA to 2.2\AA , i.e. $1.9\text{\AA}\pm 0.3\text{\AA}$
- Hydrophobic contact
carbon-carbon distance range
 3.2\AA to 4.2\AA , i.e. $3.7\text{\AA}\pm 0.5\text{\AA}$
- discretization must be fine enough to sample *atom* placements every **0.5\AA** or less





2. The search space is vast: rotation and dihedral sampling needs

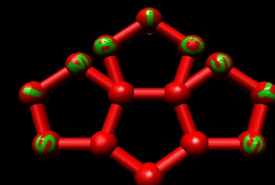
- Goal: sample atom placement every 0.5\AA or finer
- Drug-like molecules can have heavy atoms at 7\AA distance from a rotation axis (see figure)
- Simple trigonometric calculation: tangential movement of 0.5\AA is caused by rotation of about 5° at a rotation radius of 7\AA
- Consequence: orientation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





2. The search space is vast for exhaustive flexible ligand docking

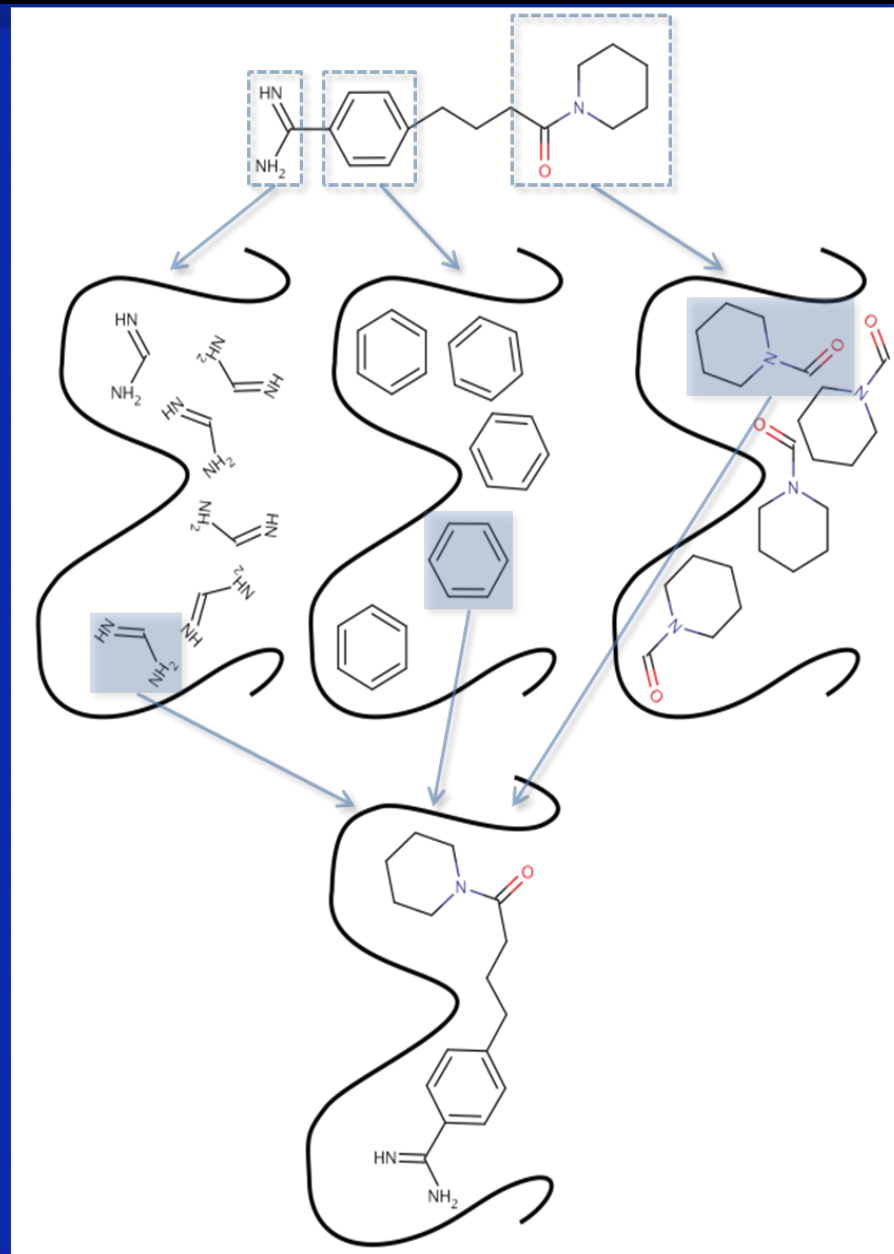
- Number of poses to examine with sampling defined:
Translations(0.5\AA) * Rotations(5°) * Dihedrals(5°) =
 $20^3 * 72^3 * 72^n$
for $n=6$ rot.bonds \Rightarrow **$2*10^{20}$ poses *per ligand***
- Brute force evaluation 2000/s \Rightarrow 3 billion years
- Stochastic methods can only explore a tiny 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

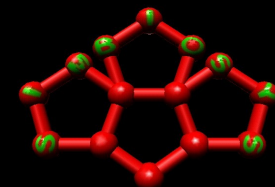


How the search engine of eHiTS covers the vast search space exhaustively (Re: lessons 1&2)

- 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

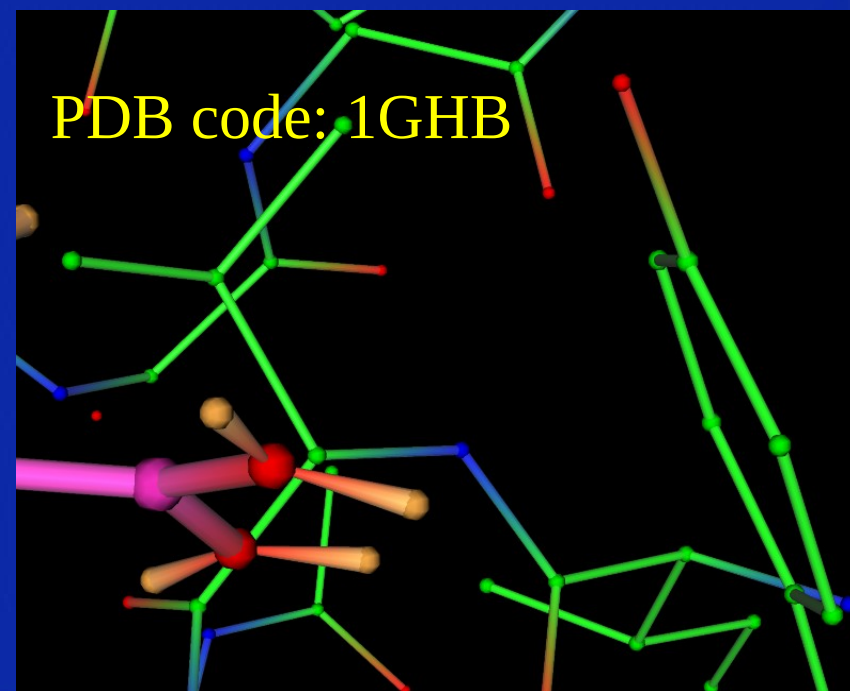
J.MGM (26) #1, July 2007, pp 198-212
doi: 10.1007/s10822-007-9164-5

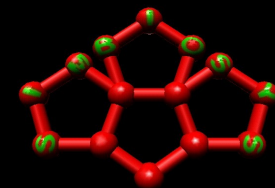




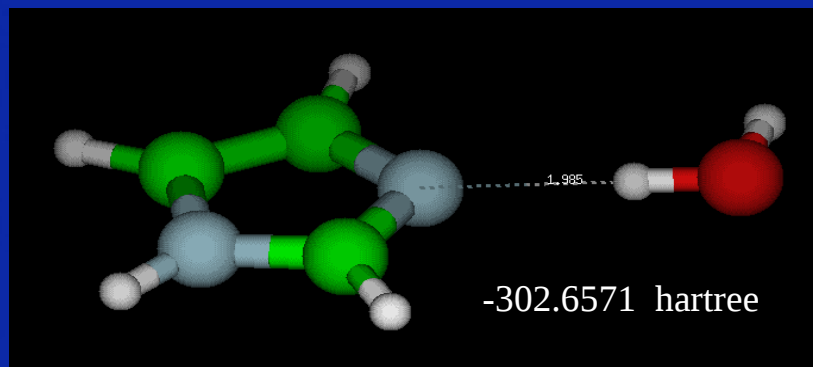
3. Need to hit all the good interactions while avoiding bad ones, BUT it is not always clear what is good...

- Good interactions:
 - H-bonds – geometry limits ? Strength vs. solvent ?
 - hydrophobic contacts – lack of specificity ?
 - complementary charges – protonation state ?
 - π -stacking – geometry ?
- Bad interactions:
 - Like charges – protonation ?
 - Polar vs. hydrophobic – see the example on the right

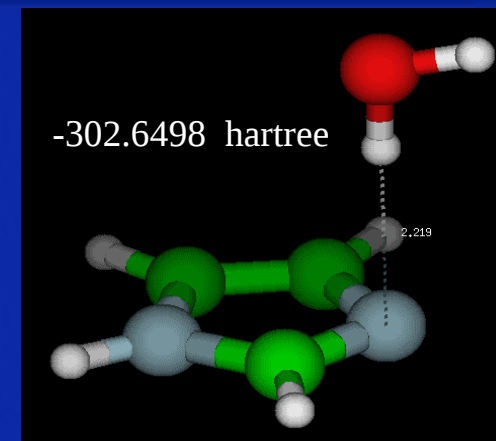




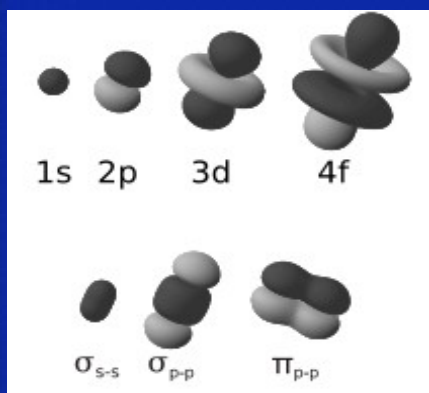
3. The problem of atom-centre based scoring functions to distinguish good and bad interactions



ΔE 4.5 kcal/mol

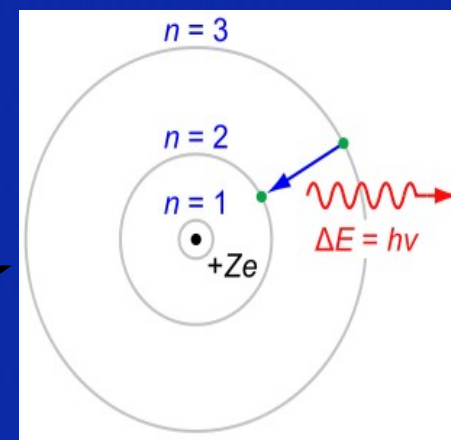


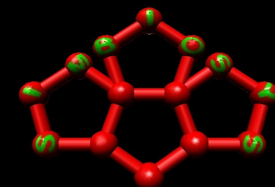
- Imidazole: 4.5 kcal/mol difference between lone-pair direction and above plane direction based on QM calculation
- Atom-center based QM-fitted point charge FF model => no difference!
- Fundamental contradiction between QM and FF models:



QM: all about electron density
(location probability)

FF: ignores electron density
~ century old Bohr model

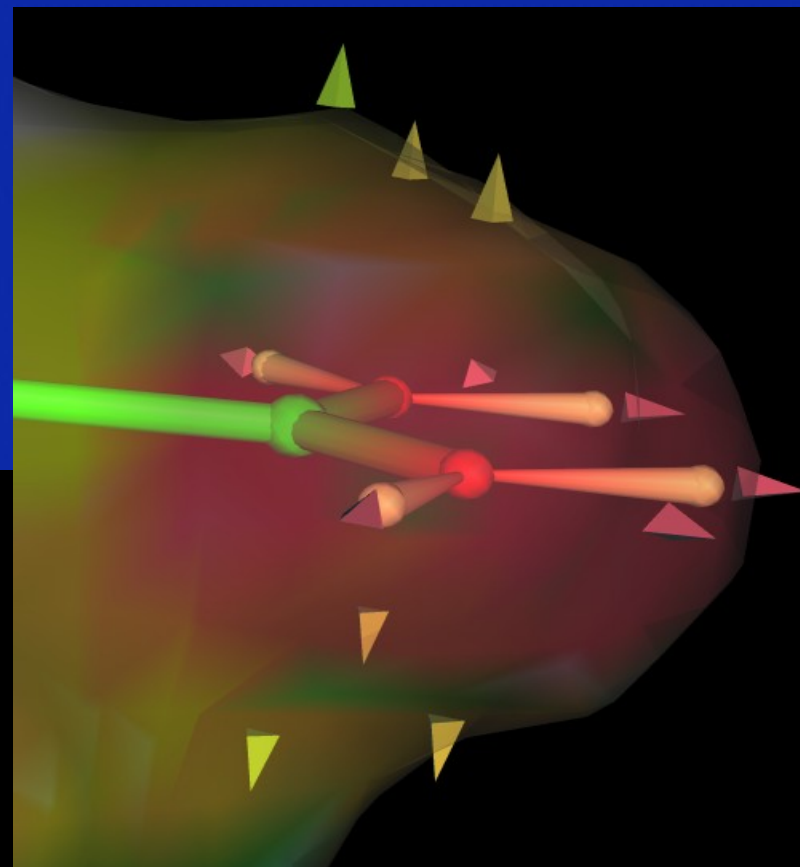
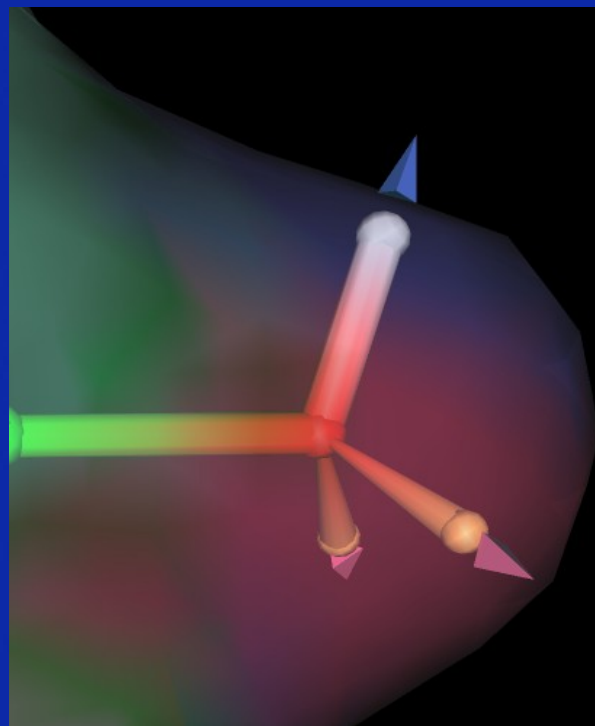




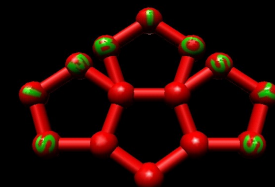
The eHiTS scoring function is based on Interaction Surface Points (Re: lesson 3)

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



More details on scoring:
23rd, 10:05 AM, Room: 157B



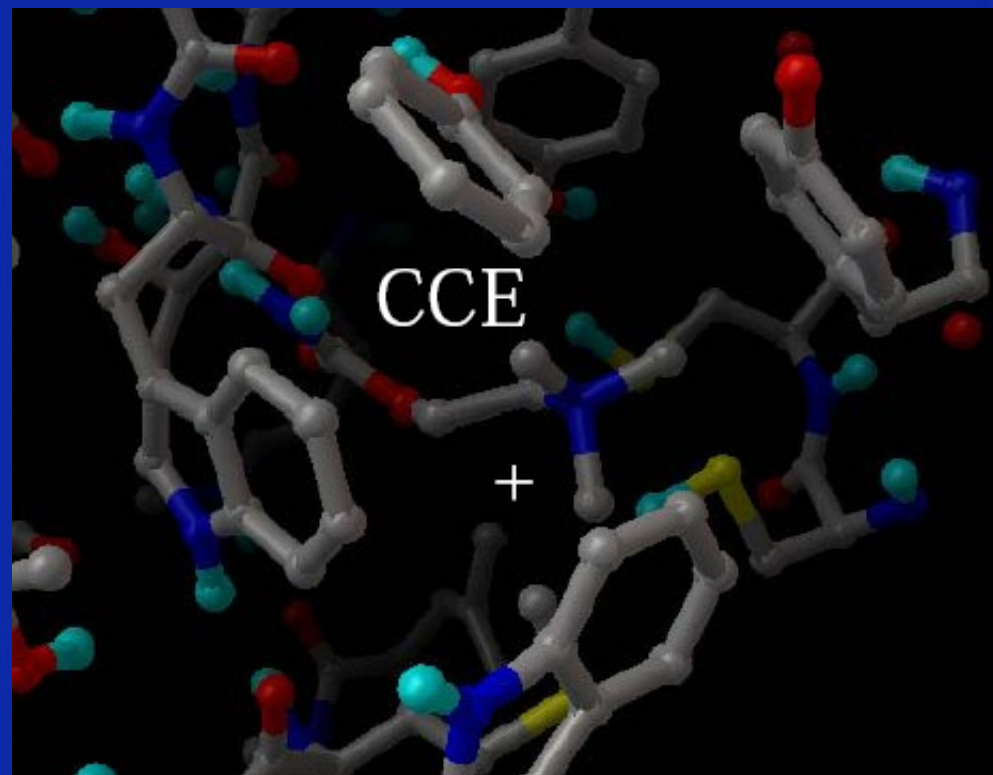
4. A weak interaction is better than none: π -cation, C-H...O and others

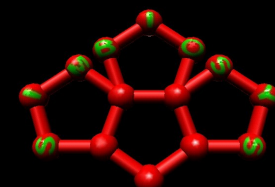
- π -cation interactions play a key role in acetylcholine binding proteins (AChBP), as well as in Epibatidine

- Pierce *et. al.* (Vertex) found many examples of aromatic C-H to O hydrogen-bonds in kinases:

Proteins 2002 **49**:567-76

- Gergely Toth *et. al.* lists a variety of weak interaction types observed in various biological receptors, *Curr.Pharm.Design* 2007. **13**:3476-93





The eHiTS Scoring matrix

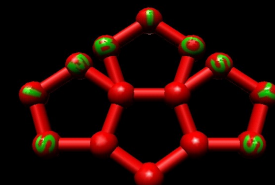
(Re: lesson 4)

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.	

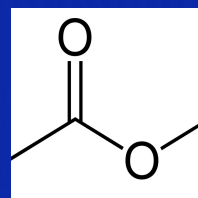
metal-ion

hydrophobic

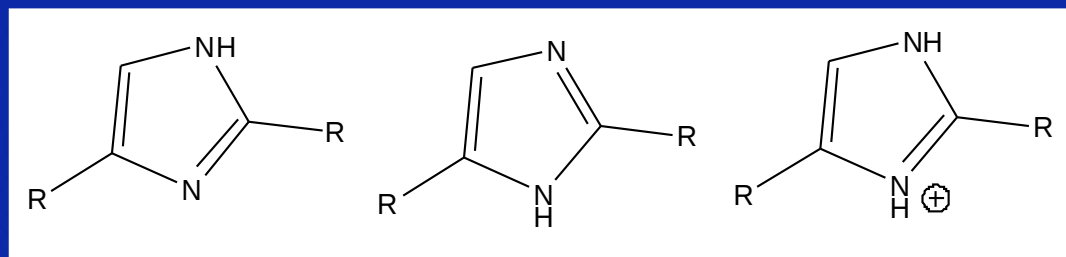


5. Not all H-bonds are created equal: the use of functional group (FG) knowledge base

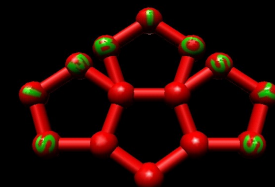
- The strength and energy of an H-bond depends on the functional group of the donor H as well as the acceptor lone-pair and even the specific atom if the FG is not symmetrical, e.g. ester



- The H-bond strength also depends on the protonation state of the FG

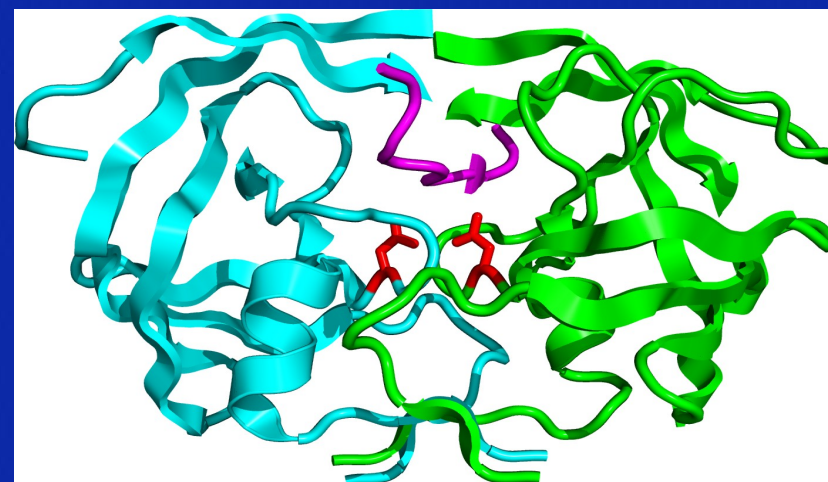
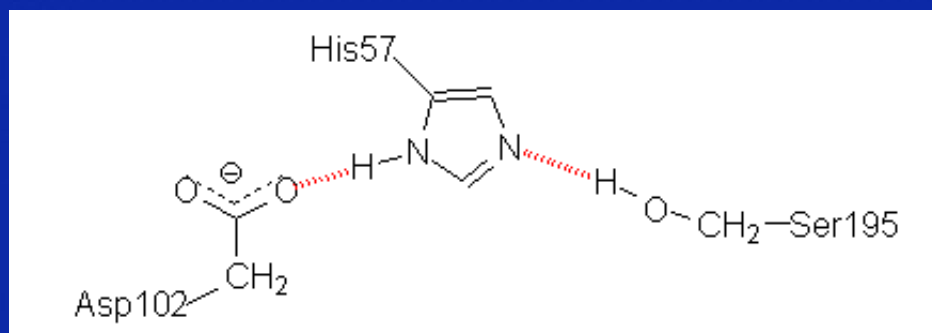


- Similarly, the strength of hydrophobic interactions depend on the atomic logP (or logD) contributions which are also stored for each atom of each FG

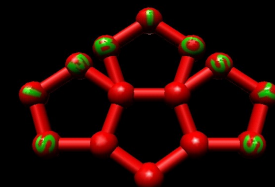


6. The importance of protonation states, induced changes upon binding

- Input files to docking may not contain H protons (PDB) or maybe protonated to neutral forms on all FG (not the best for carboxylic acid)
- Protonating for a target pH is better, but not good enough, e.g. HIV-1
- Asp-His-Ser catalytic triad presents a reactive deprotonated alcohol

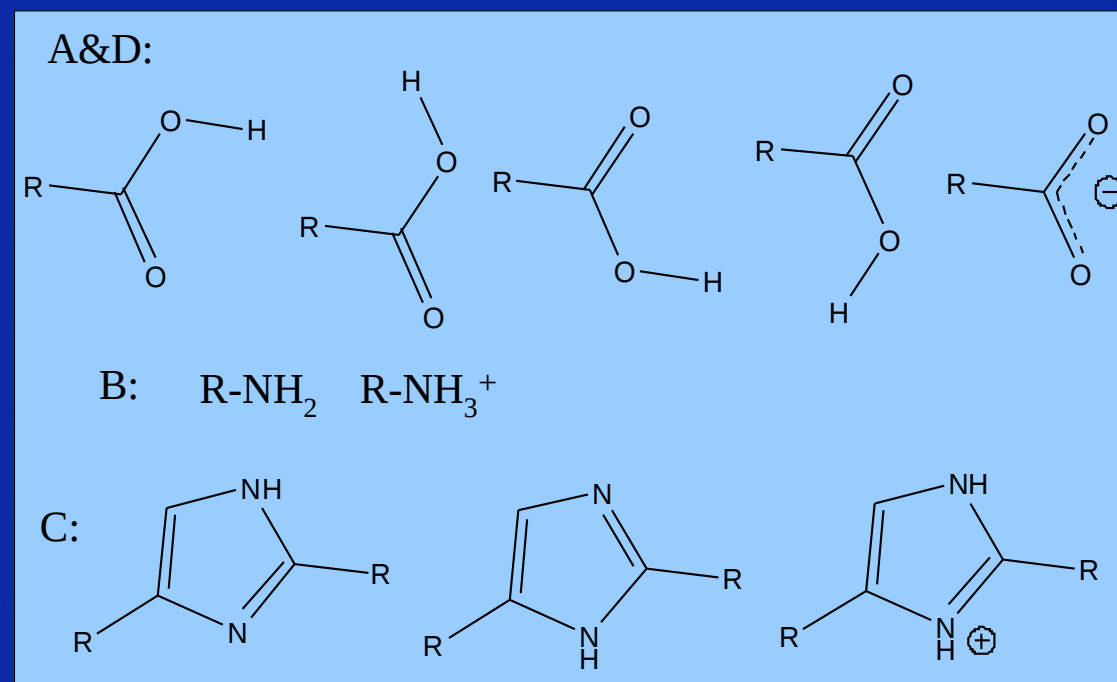
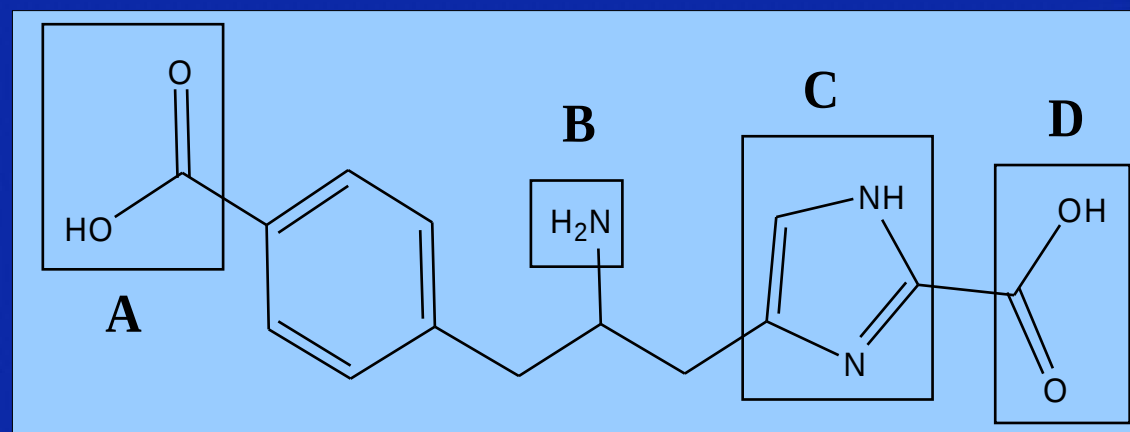


- Boris Aguliar *et.al.* *Biophysical Journal*, 2010 **98**:872-880 concludes that in most of the protein-ligand complexes at least 1 residue changes protonation state upon binding



Universal protonation handling by eHiTS (Re: lesson 6)

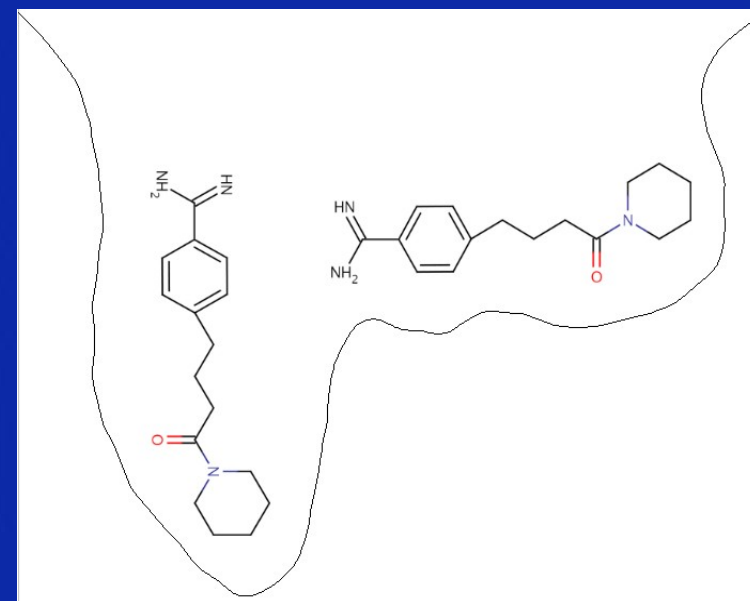
- Generic form using alternative flags (H/Lp)
- Scoring picks better one for each atom
- Example:
 - 150 states enumerated
 - 11 independent H/Lp

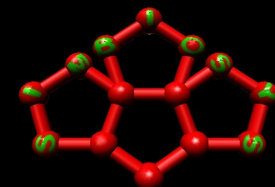


7. Location, location, location: buried versus exposed surfaces, pocket depth



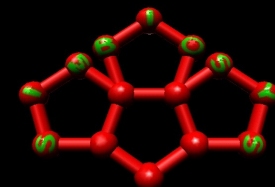
- A hydrogen bond deep in the pocket can be more valuable than one on the open surface of the protein even if the latter is formed with better geometry
- The cost of an unsatisfied H-site on the protein depends on whether it is exposed to solvent or buried by the ligand
- Covering a hydrophobic surface patch by the ligand gains score, but in addition to the area, the score should also depend on the location
- Hydrophobic contacts in an exposed open area need to be tighter than in deep buried regions – the goal is to prevent water access to the surface





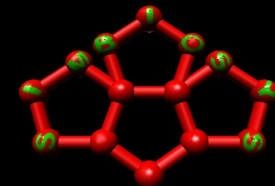
8. Beware of data errors: curation of PDB structures, binding data inconsistencies

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 Wahlby & 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

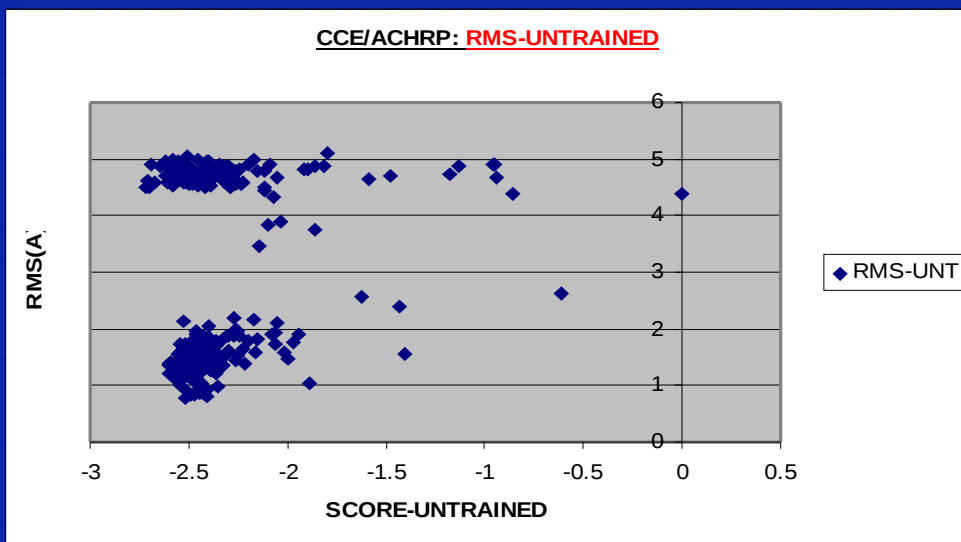


9. Diversify 3 aspects of scoring: pose ranking, enrichment and binding energy estimation

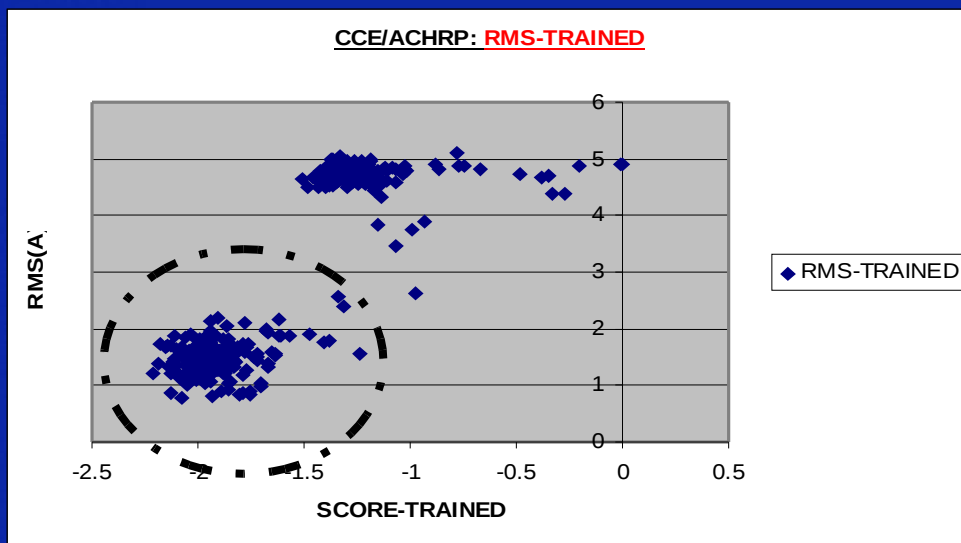
- Certain scoring terms (e.g. steric clash penalty) are very important during the pose generation and ranking process, but irrelevant for enrichment and binding energy estimation
- Estimation of binding free energy is the most difficult aspect – if we could do that perfectly, it would have to work well for enrichment too
- We can get much better enrichment score by focusing on key interactions of the given protein family target
- The scoring terms of eHiTS are combined using different weight sets for:
 - Ranking of the generated docking poses of the same ligand
 - Enrichment score for differentiating actives and decoys
 - Binding energy estimation, i.e. rank ordering actives
- More in scoring talk: Aug. 23rd 10:00 am, room 157B



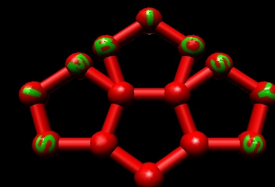
Effect of rank tuning



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-regime (circled) at low scores from poor scoring –high RMS results



10. Different protein families need different weighting schemes

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

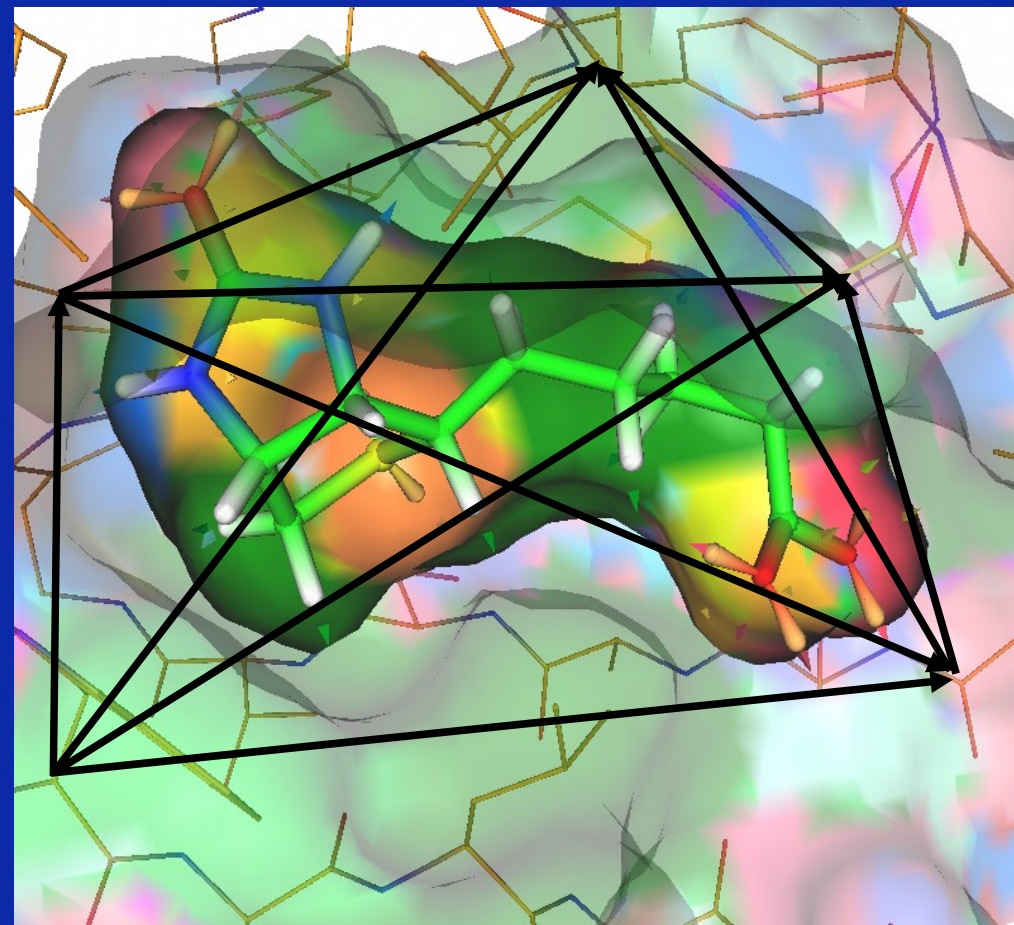
Geometric clustering is based on binding site residue C α distance matrix.

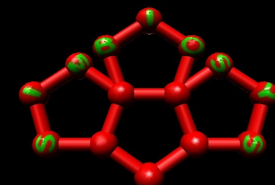
- distance tolerance (default 3Å)
- matching subset size minimum (5)
- minimum set-size (5 entries)

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

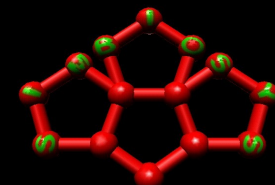
The same matching criteria is used to find the “family” of the target protein in the preprocessing step of a docking run





Summary: the 10 lessons

- It is not sufficient to sample few dozen low energy conformers
- The search space is vast: brute force or random searches fail
- Need to hit all good interactions and avoid bad ones, BUT which?
- Location, location: buried versus exposed surface, pocket depth
- A weak interaction is better than none: π -cation, C-H...O etc.
- Not all H-bonds are created equal: functional group dependence
- Importance of protonation states, induced changes upon binding
- Beware of data errors in PDB structures and binding energy data
- Diversify the 3 aspects of scoring: ranking, enrichment, energy
- Different protein families need different weight schemes



Summary: how eHiTS benefited from the lessons

- eHiTS exhaustively samples the ligand conformations
- Search engine: smart systematic sampling of the vast space
- Statistical interaction pattern scoring, full 23x23 matrix (good/bad)
- Surface coverage term (buried/exposed), pocket depth term
- Weak interactions (π -cation, C-H...O etc.) are handled by the matrix
- Functional group knowledge base with pKa, logD values per atom
- Ambivalent protonation states, enumeration in FGKB, optimized
- PDB structures are filtered and curated prior to statistics collection
- Different scoring weight set for ranking, enrichment and energy
- Protein family knowledge base with interaction patterns, weights

Scoring talk: Aug. 23rd 10:00 am, room 157B, Booth #945