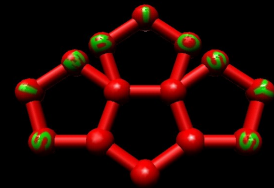


eHiTS scoring function



<http://www.simbiosys.ca/>



Zsolt Zsoldos,
Danni Harris

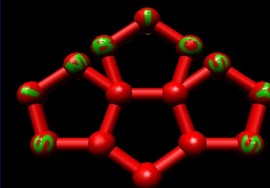
SimBioSys Inc., © 2009



Contents:

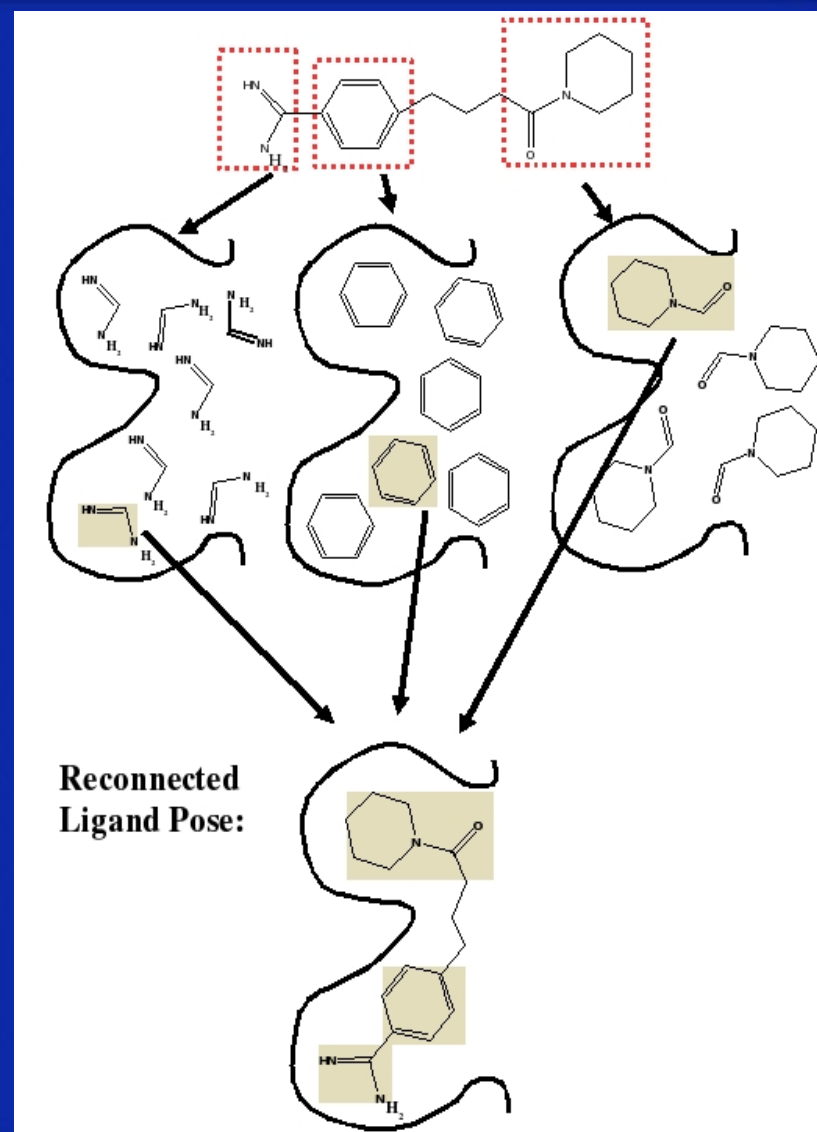
- Introduction: overview of various scoring models and their problems
- Interaction Surface Point (ISP) based scoring method
- Statistically derived parameters with empirical weights
- Protein family clustering and weight tuning
- Binding energy estimation and enrichment results

2. eHiTS algorithm overview: intelligent exhaustive search

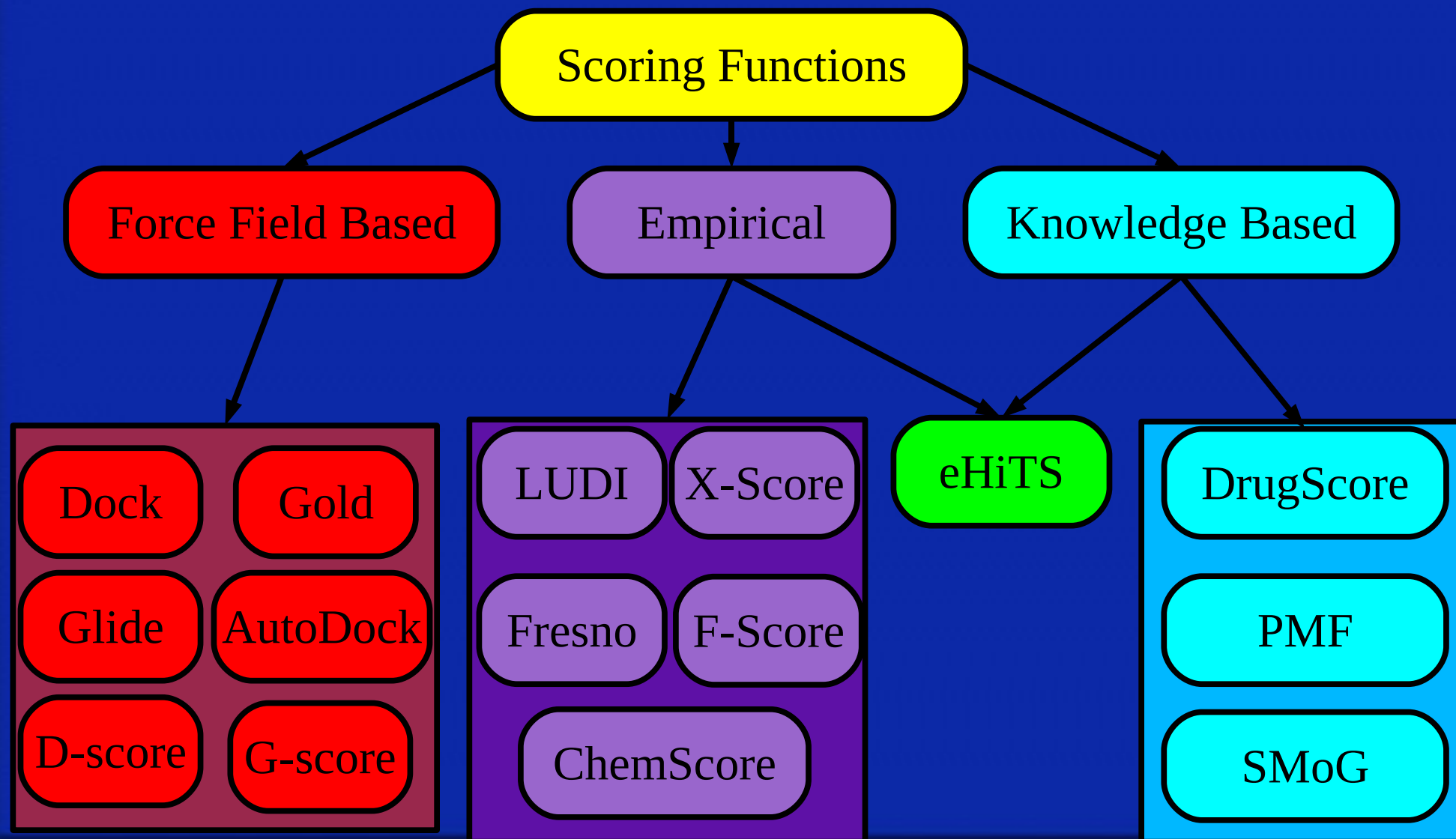
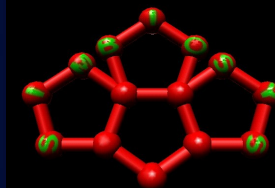


- Ligand is divided into rigid fragments, flexible chains
- All rigid fragments are docked **independently** (many poses)
- Pose matching (clique detection)
- Flexible chain fitting (continuous)
- Local energy minimization

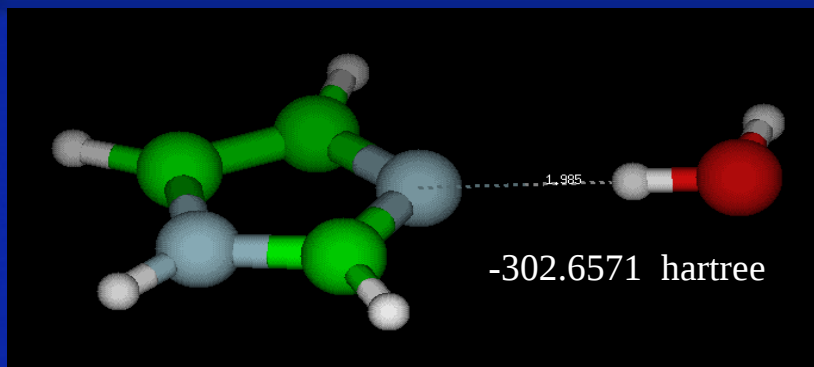
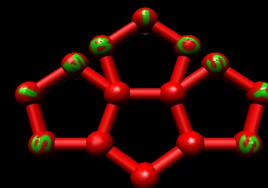
J.MGM (26) #1, July 2007, pp 198-212
doi: 10.1007/s10822-007-9164-5



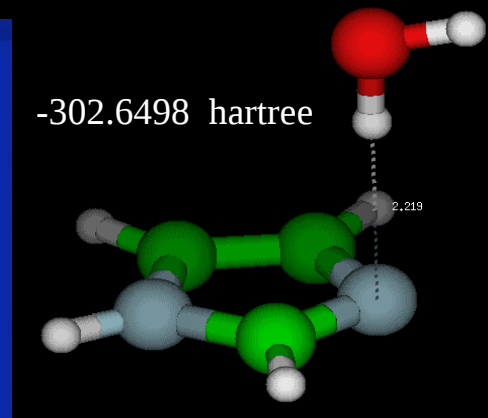
3. Various Scoring Models



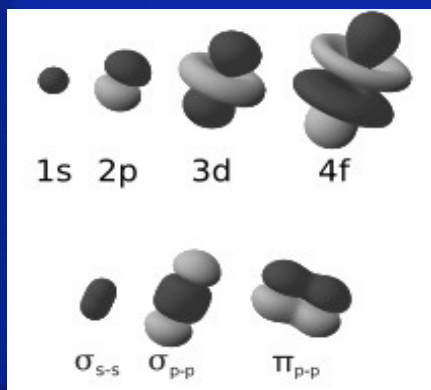
4. Problem with atom-centre based models



ΔE 4.5 kcal/mol

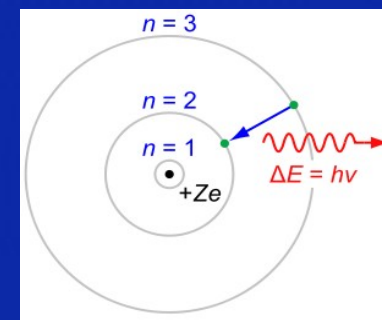


- Imidazole: 4.5 kcal/mol difference between lone-pair direction and above plane direction based on QM calculation
- Atom-center based QM-fitted point charge FF model => no difference!
- Fundamental contradiction between QM and FF models:

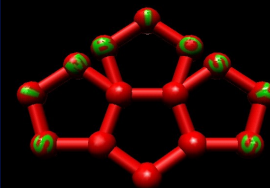


QM: all about electron density
(location probability)

FF: ignores electron density
~ century old Bohr model

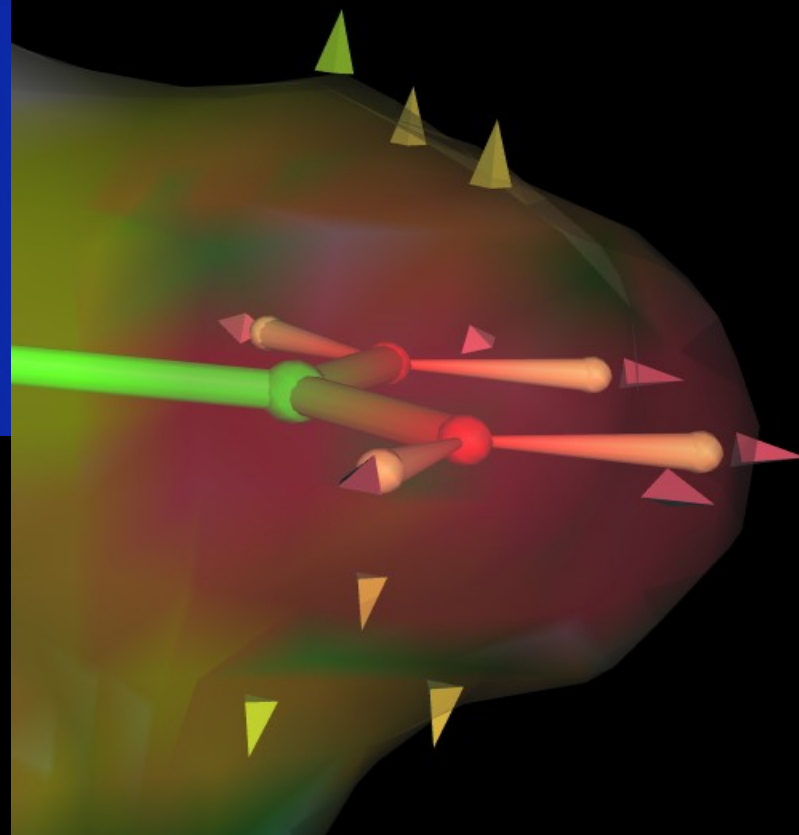
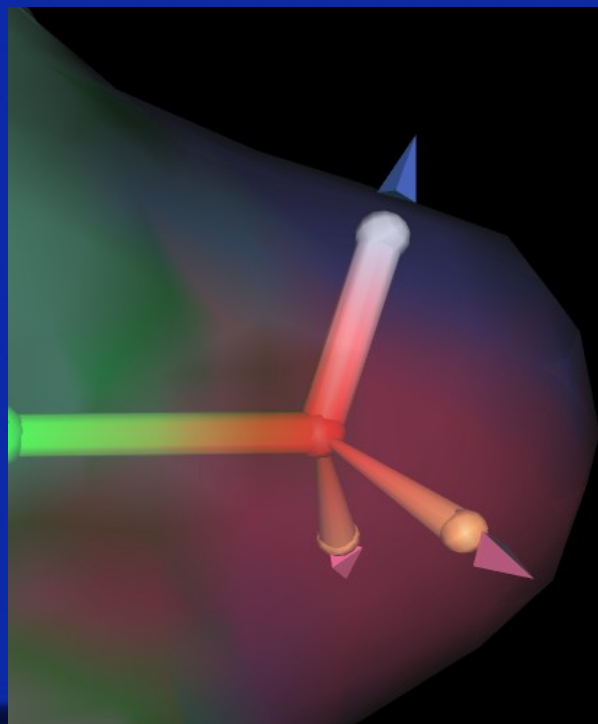


5. Interaction Surface Points (ISP)

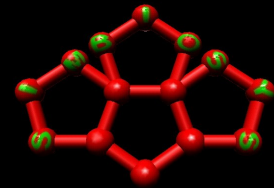


- eHiTS places directional surface points in specific locations on the surface of molecules to represent various interaction capabilities:

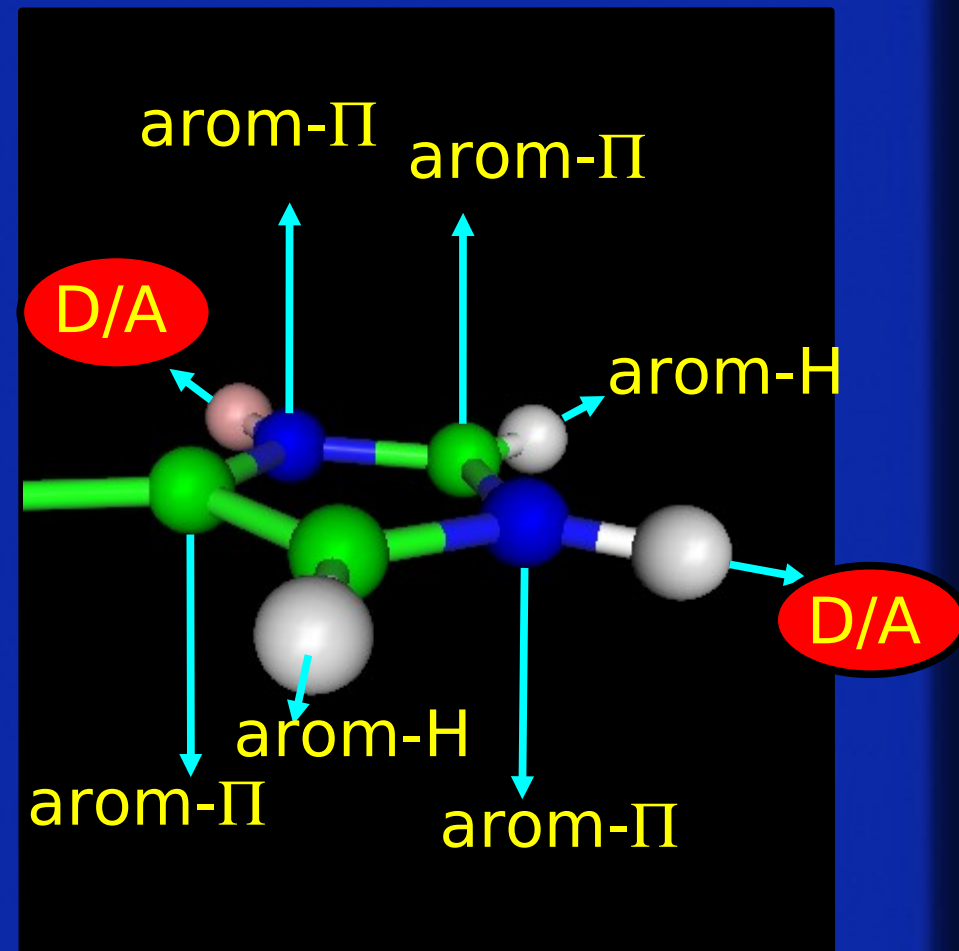
- H atoms,
- lone electron pairs,
- π electrons



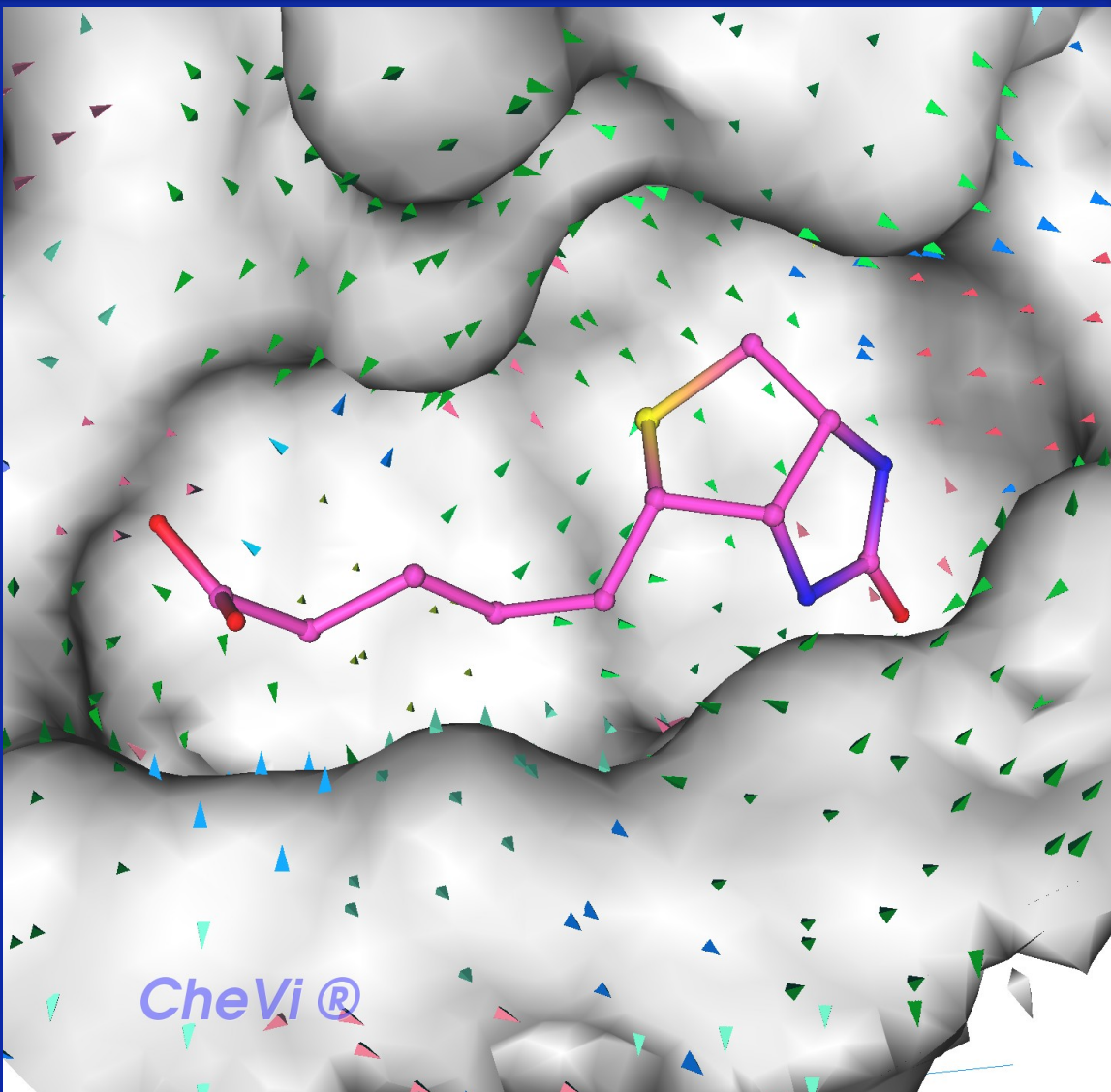
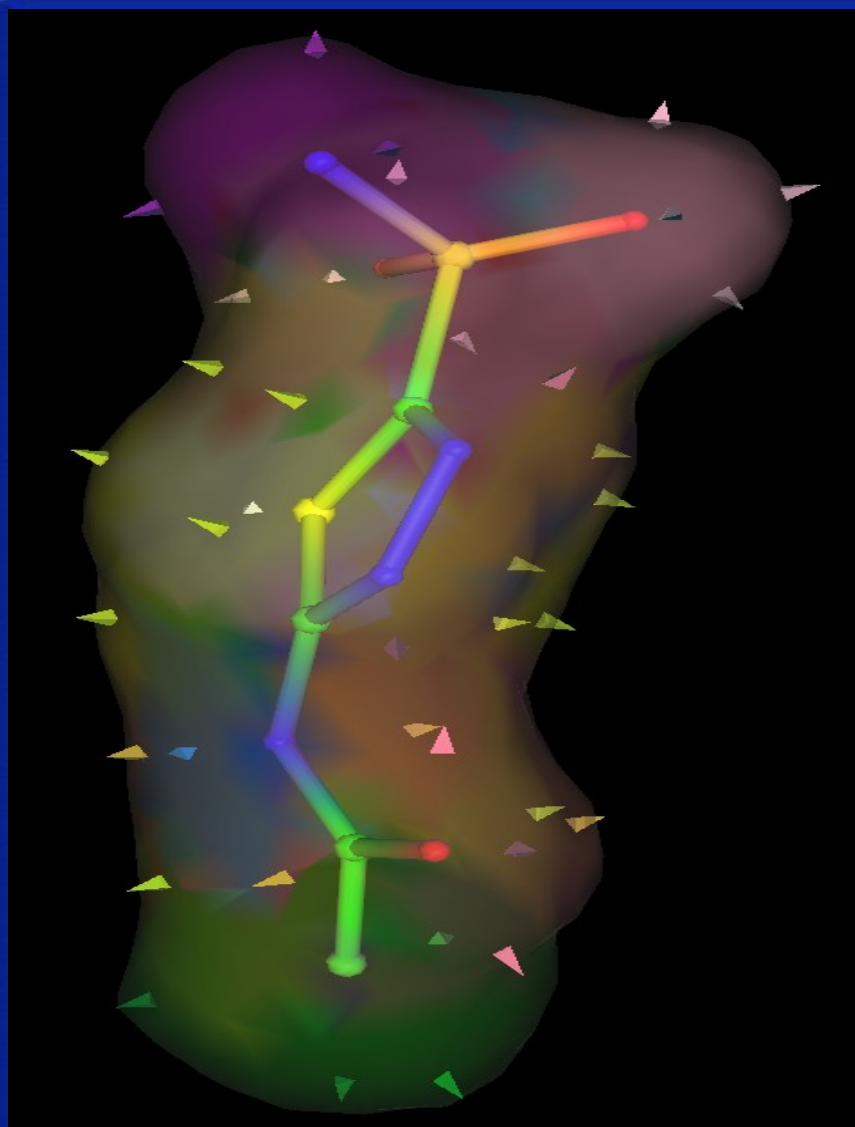
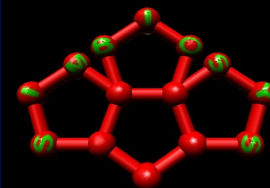
6. Interaction surface point (ISP) types



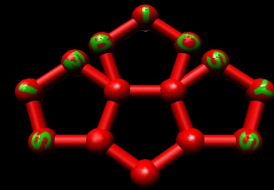
- **H-bond Donors:**
 - charged, amine, strong, weak, rotatable
- **H-bond Acceptors:**
 - charged, acid, strong, weak, rotatable
- **Ambivalent H donor/acceptor**
- **Aromatic Pi-stacking:**
 - carbon, polar, resonance, edge-H, arom- π
- **Hydrophobic:**
 - strong / weak lipophil, neutral
- **Metal ions**
- **Misc (Sulfur, Halogens)**



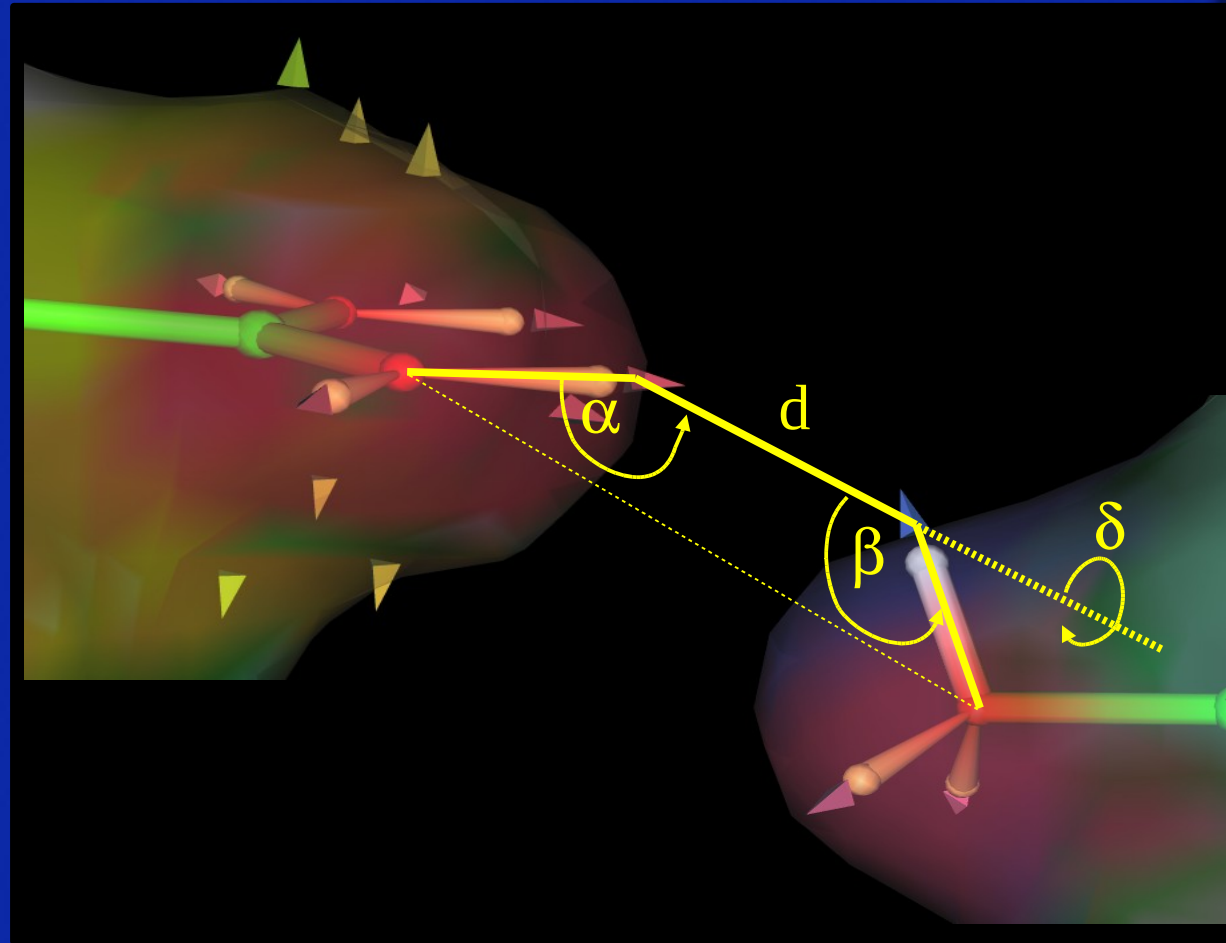
7. ISP set examples



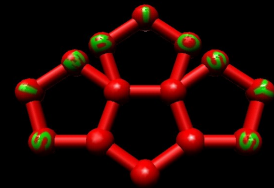
8. Interaction Geometry scoring



- Interactions can not be described by distance (d) alone, the angles between ISP directions and interaction directions (α, β) as well as the torsions (δ) between them must be considered:

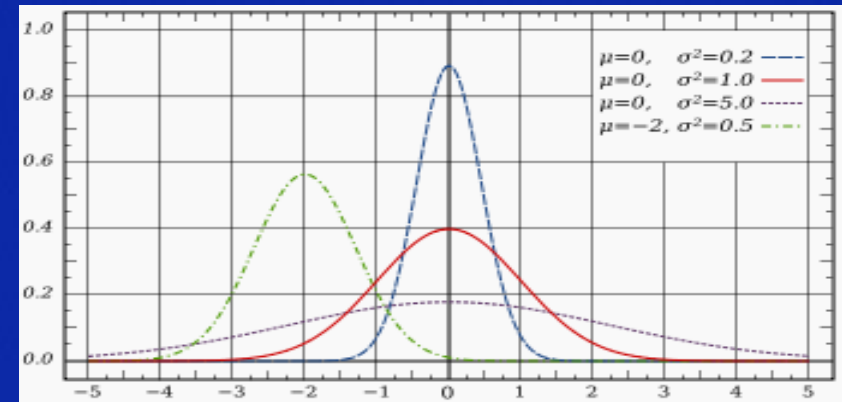


9. Statistical data collection



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

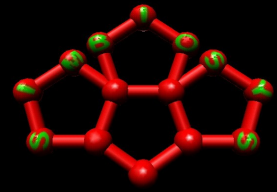
$$p(d) = \left(\frac{B}{4\pi}\right)^{-3/2} \int_0^\pi \int_0^{2\pi} \exp\left(\frac{-4\pi^2 r_{\alpha\beta}^2}{B}\right) d^2 \sin(\alpha) d\alpha d\beta$$



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

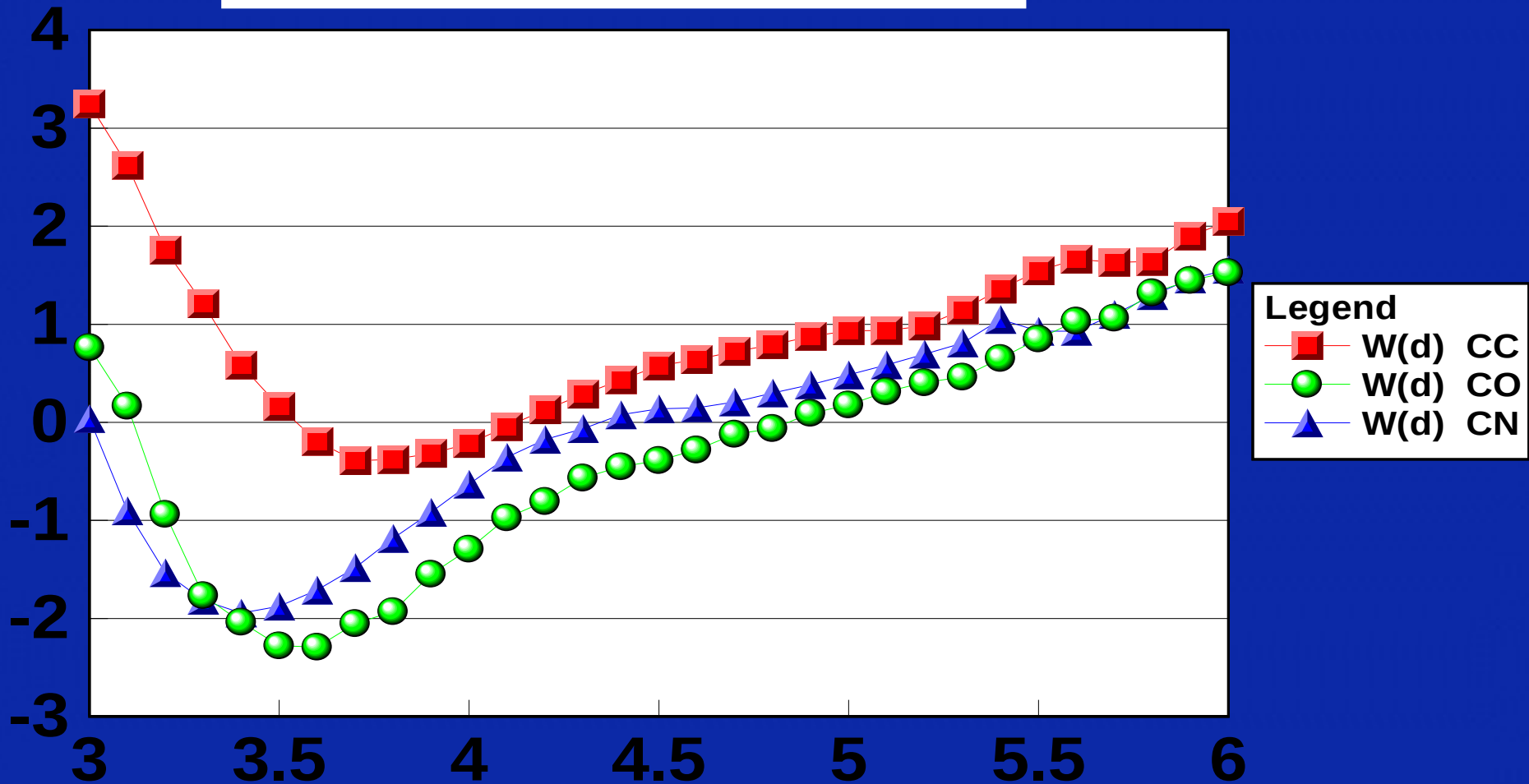
$$P(d) = \left(\frac{4\pi}{B_0 + B_1}\right)^{\frac{3}{2}} d^2 \int_0^\pi \int_0^{2\pi} \exp\left(\frac{-4\pi^2}{B_0 + B_1} \|P_0 - P_1 + P_s\|^2\right) \sin \alpha d\alpha d\beta$$

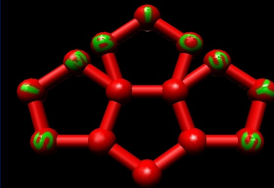
- Similar formulae for angle and torsional components
- 4D data collection using fine numerical integral sampling



10. Energy by Boltzmann distribution

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





11. Fitting empirical functions to the statistical data

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

$$g(d, \alpha, \beta, \delta) = e_0 s(d) l(\alpha) r(\beta) t(\delta)$$

$$s(d) = p_{10}d + p_{11}d^2 + p_{12}d^3 + p_{13}a(d) + p_{14}c(d) + p_{15}$$

$$a(d) = p_8(d - p_9)^2$$

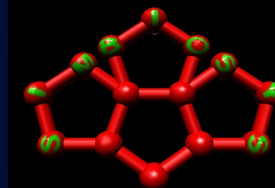
$$c(d) = \begin{cases} \cos(a(d)) & \text{if } a(d) > -\pi \wedge a(d) < \pi \\ -1 & \text{otherwise} \end{cases}$$

$$l(\alpha) = p_0 \cos \alpha + p_1 \cos^2 \alpha + p_2 \sqrt{\cos \alpha} + p_3$$

$$r(\beta) = p_4 \cos \beta + p_5 \cos^2 \beta + p_6 \sqrt{\cos \beta} + p_7$$

$$t(\delta) = p_{16} \cos \delta + p_{17} \cos^2 \delta + p_{18} \sqrt{\cos \delta} + p_{19}$$

12. The Scoring matrix



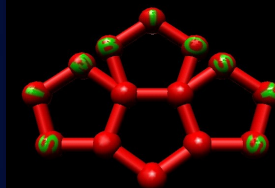
polar/H-bond π \rightarrow π interactions π /aromatic--cation/H-donor

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.	

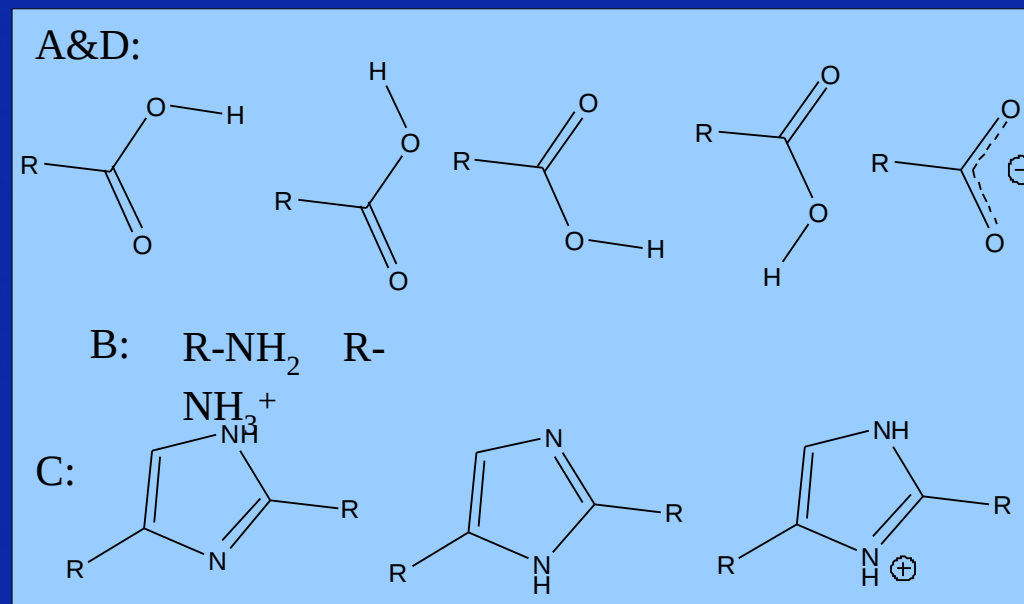
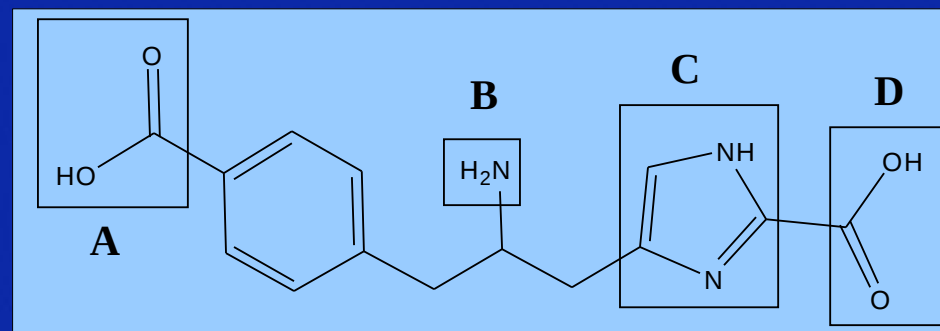
metal-ion

hydrophobic

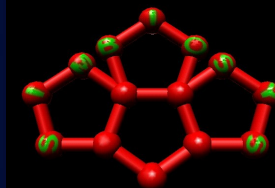
13. Universal protonation handling



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

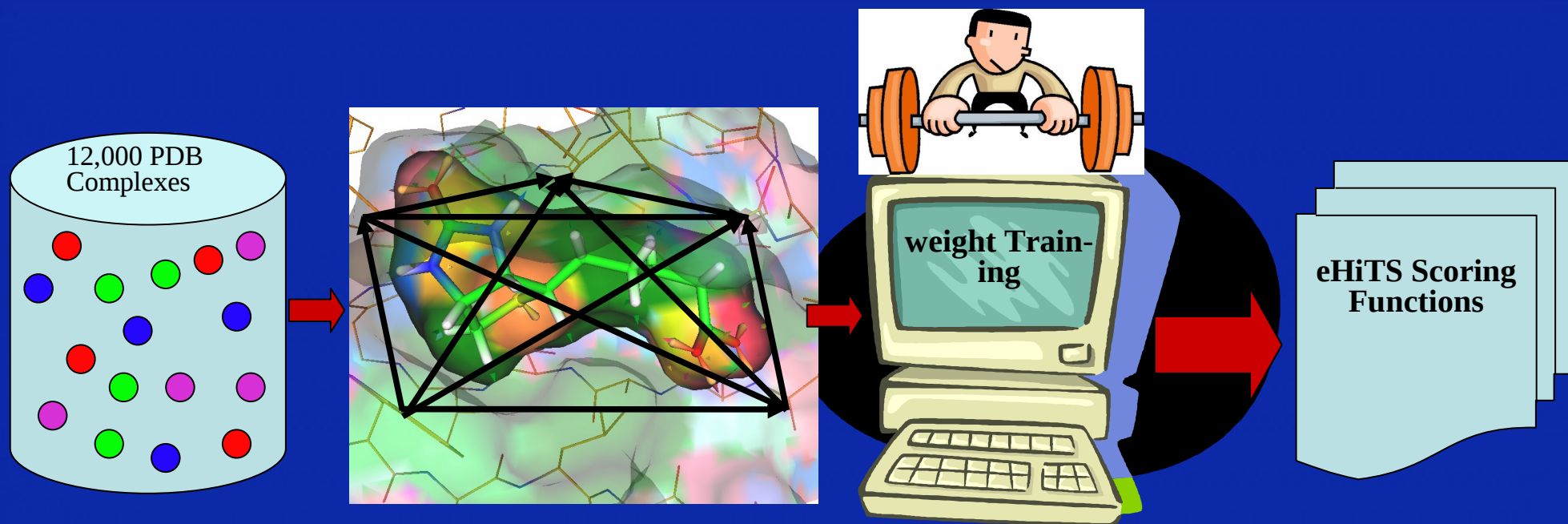
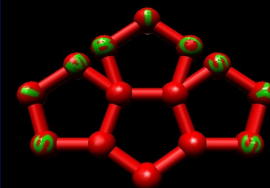


14. Additional scoring terms



- De-solvation: continuous model, ISP type dependent
- Steric clash penalty: distance-square from Connolly surface
- Pocket depth: signed distance of atoms from convex hull
- Coulomb electrostatic term (all atom pairs, no cut-off)
- Hydrophobic term based on $\log(D)$ value pairs
- Ligand strain energy (torsional probability + vdw LJ 6-12)
- Ligand intra-molecular interaction score (ISP pair ~ receptor)
- Ligand entropy loss (frozen rotatable bonds)

15. Protein family based weight tuning



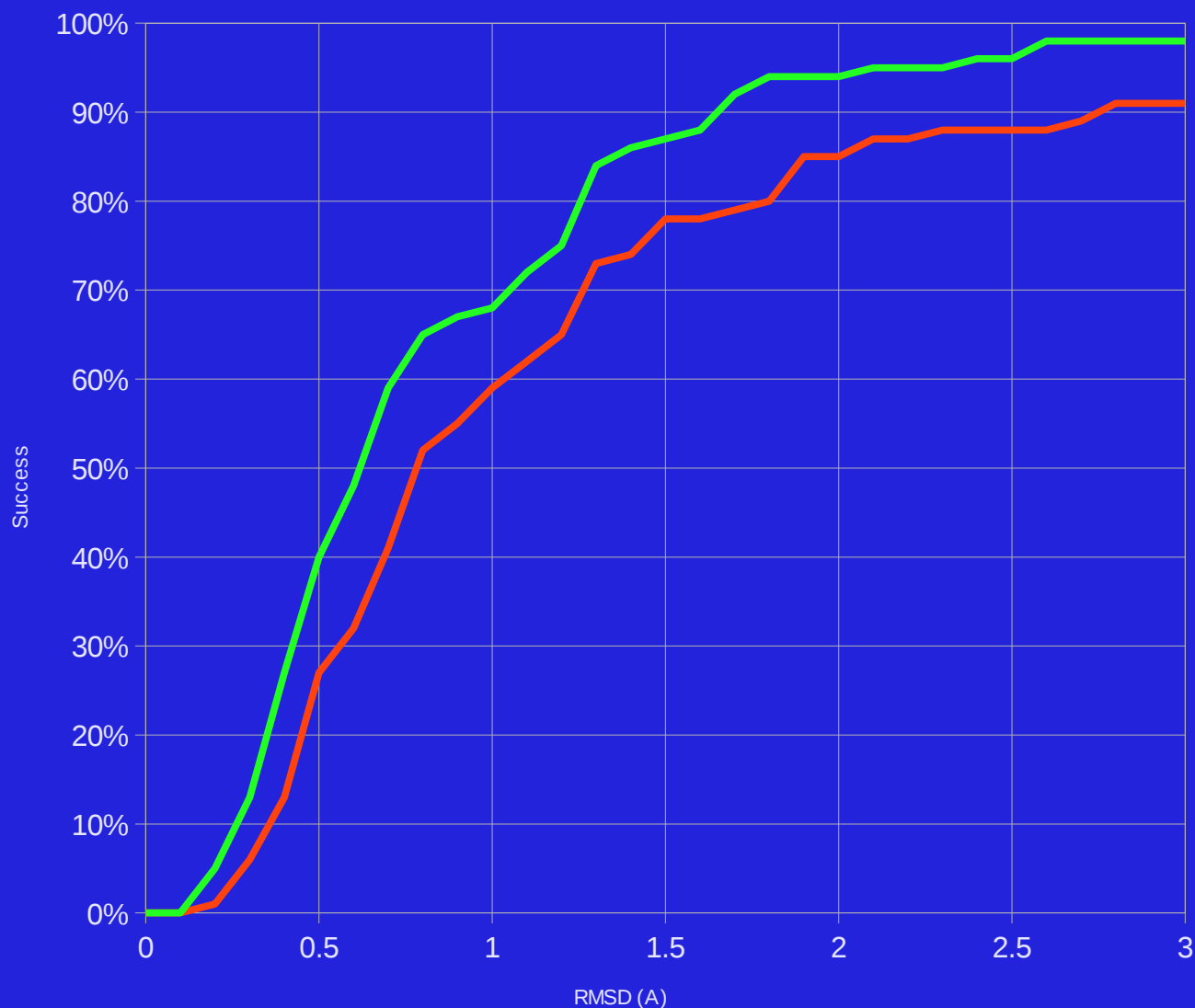
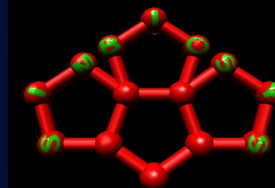
1. 12,000 PDB complexes chosen to represent a wide range of protein families

2. Complexes are clustered automatically into ~500 protein families plus one default, global set based on amino acid ligand-component distance metrics.

3. eHiTS training utility optimizes the scoring term weights for each family

4. Scoring functions for each family are output and used as default scoring functions of eHiTS

16. Validation results on the Astex diverse benchmark set (85 cases)



	TopRank	Closest
Average	1.41Å	0.91Å
Mean	0.84Å	0.67Å
<1Å	59%	68%
<2Å	85%	94%
<3Å	91%	98%

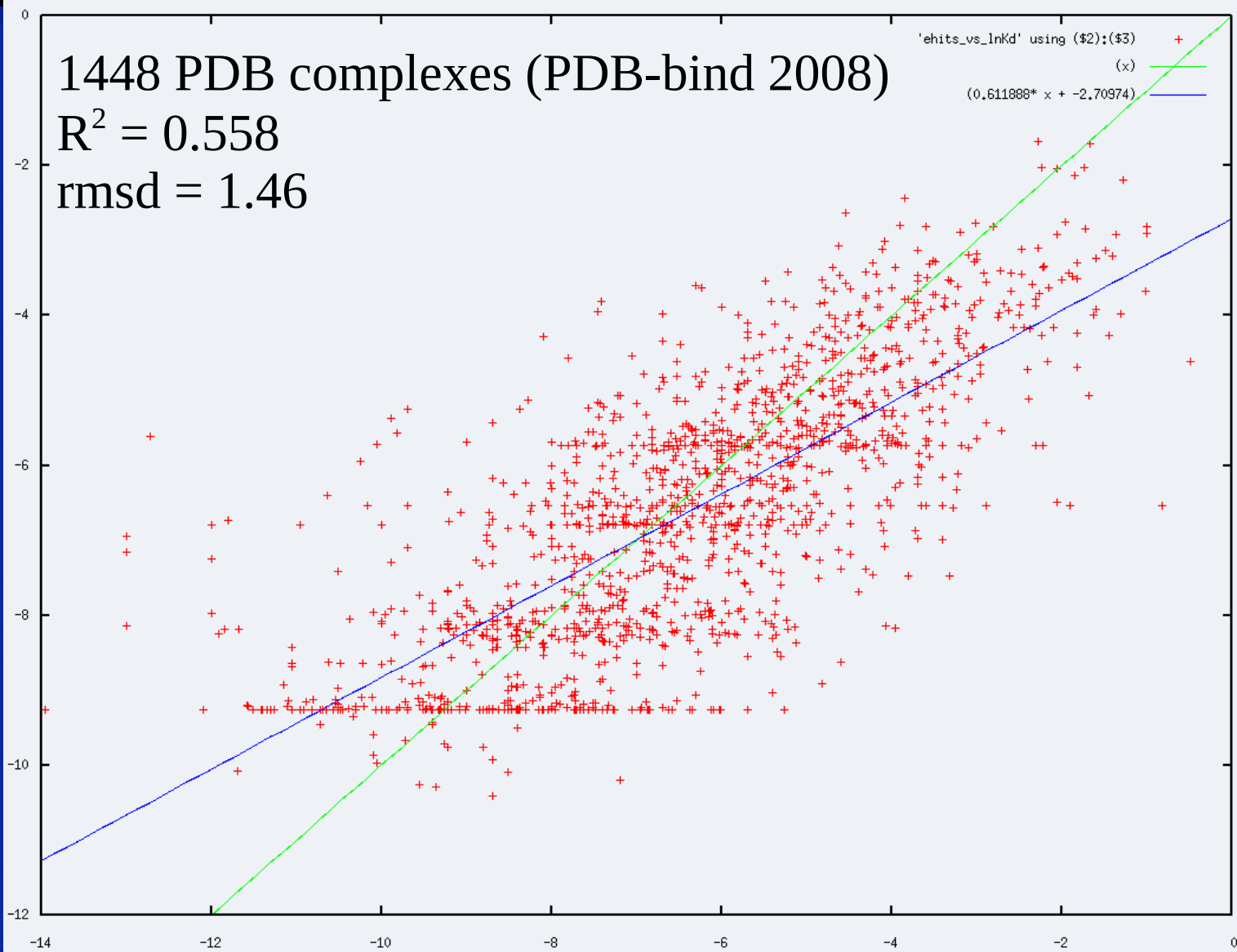
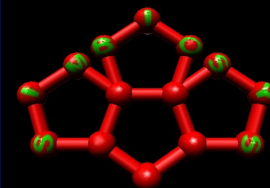
Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance.

M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, C. W. Murray

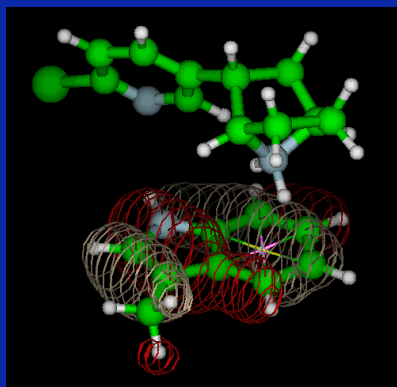
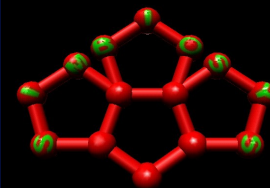
J. Med. Chem., 50, 726-741, 2007.

[DOI:10.1021/jm061277y]

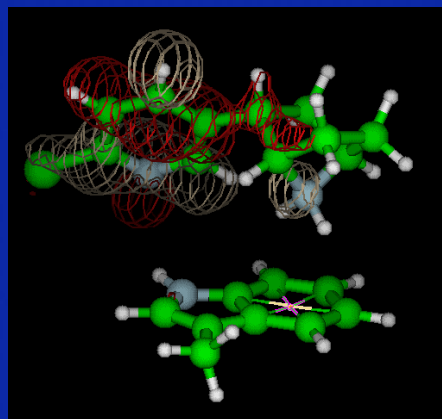
17. Correlation of eHiTS-score with experimental binding data



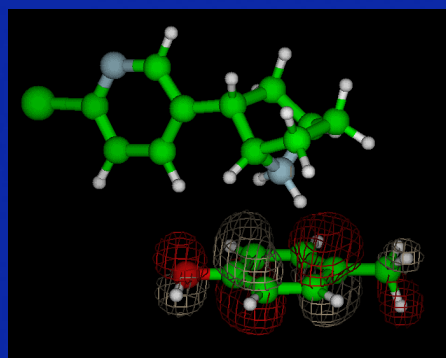
18. Cation- π interactions of Epibatidine revealed from QM analysis at eHiTS docked geometries.



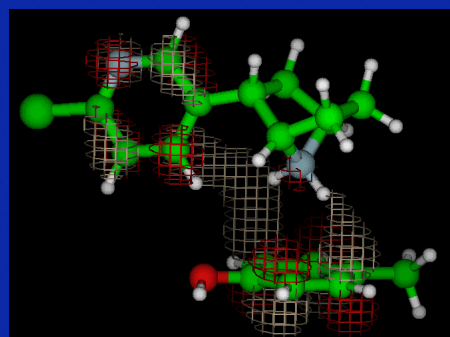
EPI-Trp(HOMO)



EPI-Trp(LUMO)



EPI-TYR(HOMO)

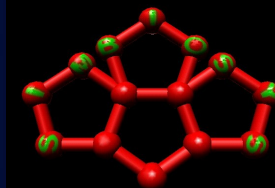


EPI-TYR(LUMO)

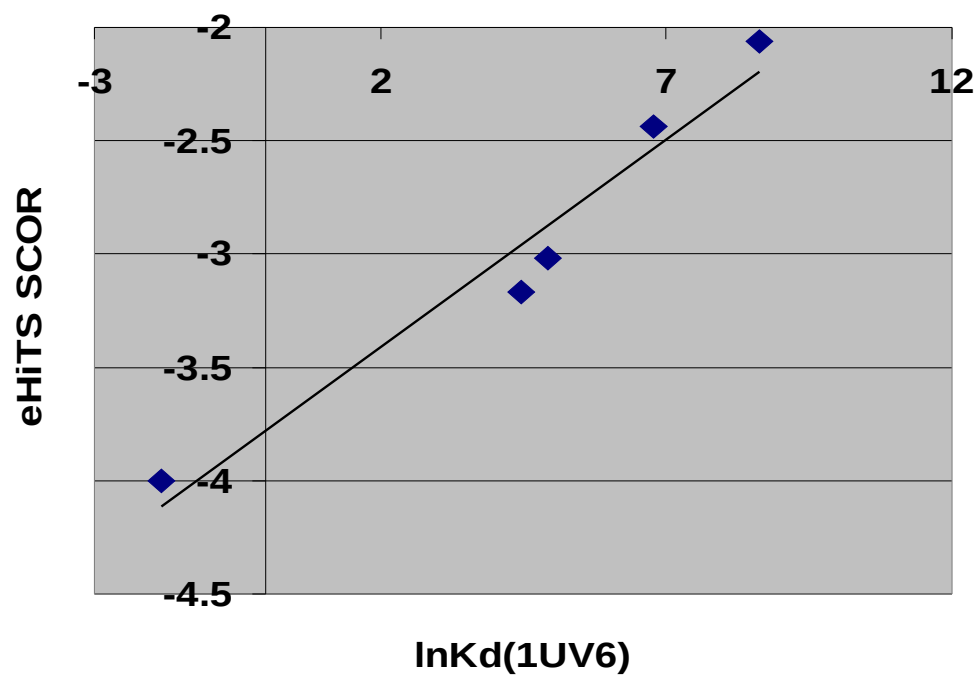
The electron density plots for the highest occupied and lowest unoccupied orbitals (HOMO and LUMO) show the intimate interactions between the cationic center of nicotine and epadipine and 'models' of the Trp and Tyr centers.

Note that a portion of the LUMO's exist on the 'cationic' ligands and these orbitals obtain charge density from the polarized Trp/Tyr centers.

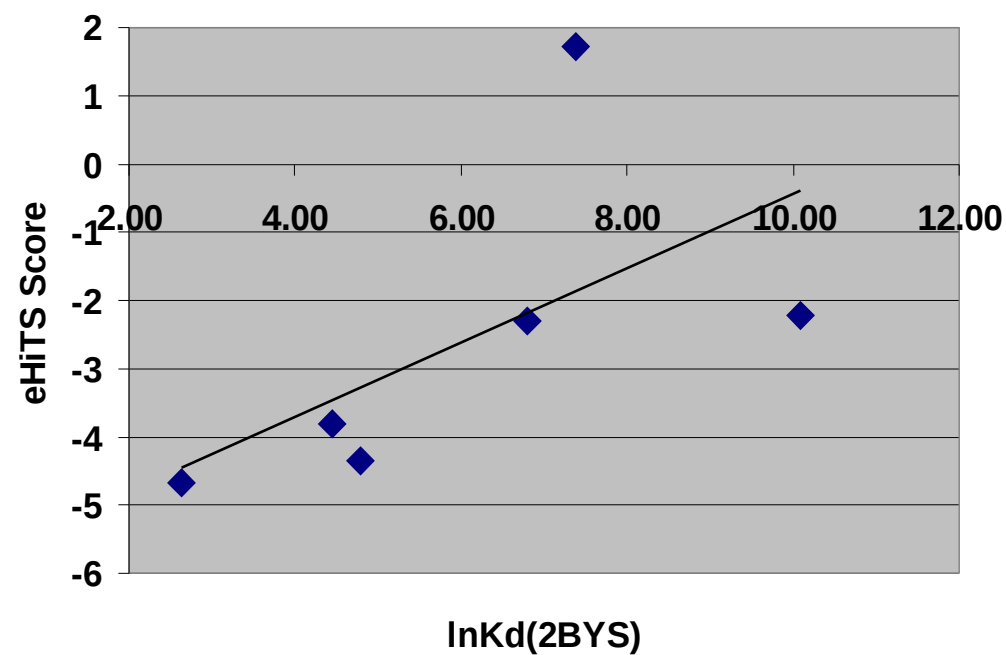
19. Score vs. binding affinity correlation



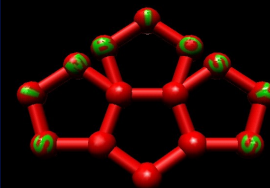
1UV6(Lymnaea)



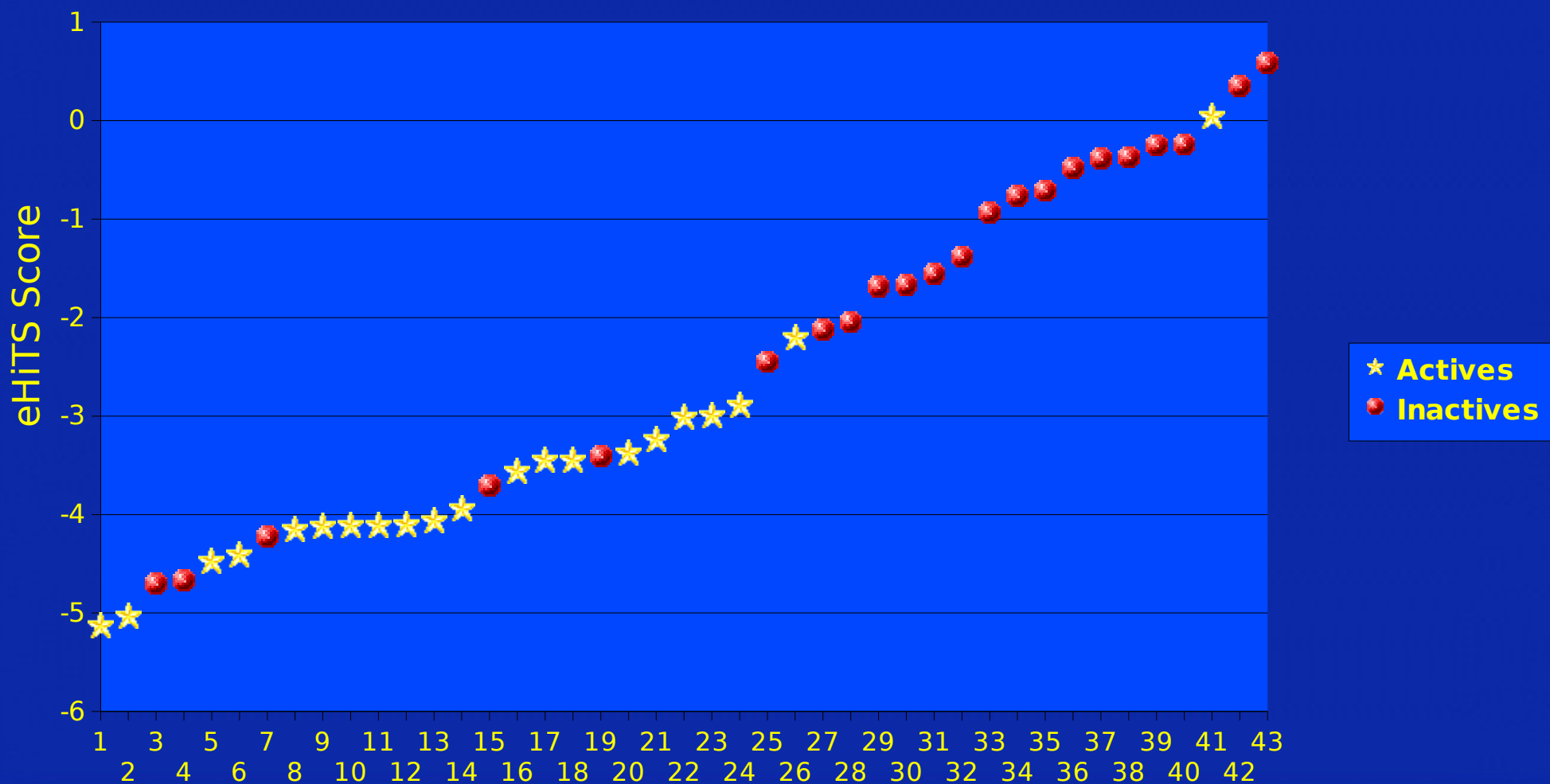
2BYS(Aplysia)

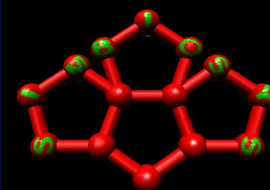


20. Screening success on CDK5



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

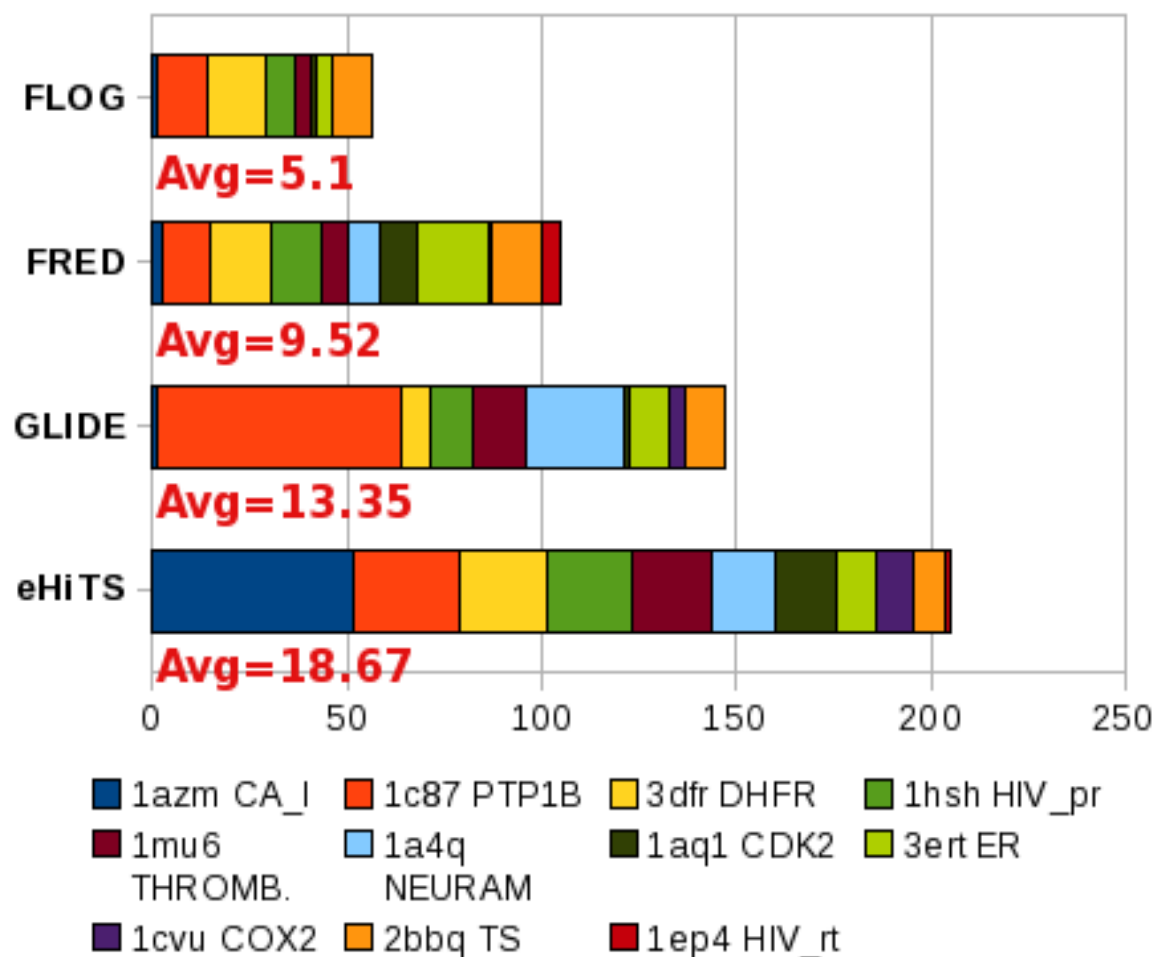




21. Screening results with eHiTS

Merck's comparison of 4 dockers on 11 test cases

Enrichment results with: Flog, Fred, Glide and eHiTS

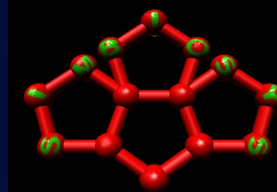


Data and competitor results from the paper:

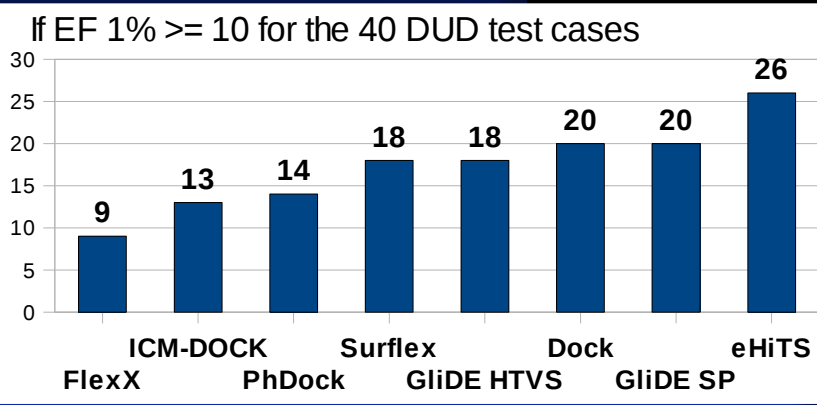
J. Chem. Inf. Model. 2007; 47(4), pp 1504 - 19
DOI: 10.1021/ci700052x

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

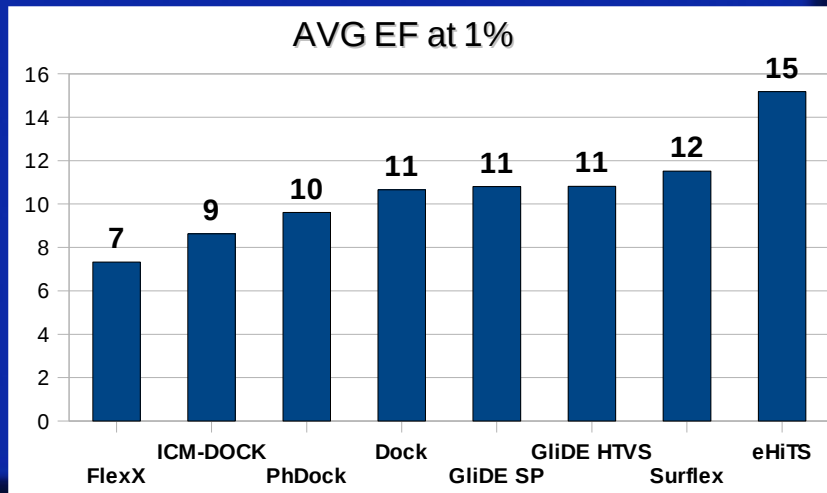
22. Enrichment results on DUD set (EF at 1% of database)



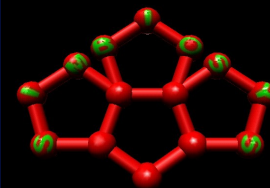
	DOCK	FlexX	GLIDE HTVS	GLIDE SP	ICM-DOCK	PhDOCK	Surflex	eHiTS
AR	3.9	6.4	18.0	12.8	15.4	24.4	7.7	7.7
Erag	7.6	18.2	19.7	7.6	25.7	22.7	9.1	7.6
Erant	13.6	16.3	1.6	16.3	19.1	10.9	16.3	23.7
GR	3.9	2.6	2.6	3.9	14.2	16.8	16.8	5.19
MR	0.0	7.2	7.2	21.7	36.2	43.5	36.2	14.29
PARP	30.7	30.5	6.1	15.4	5.9	0.0	21.3	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

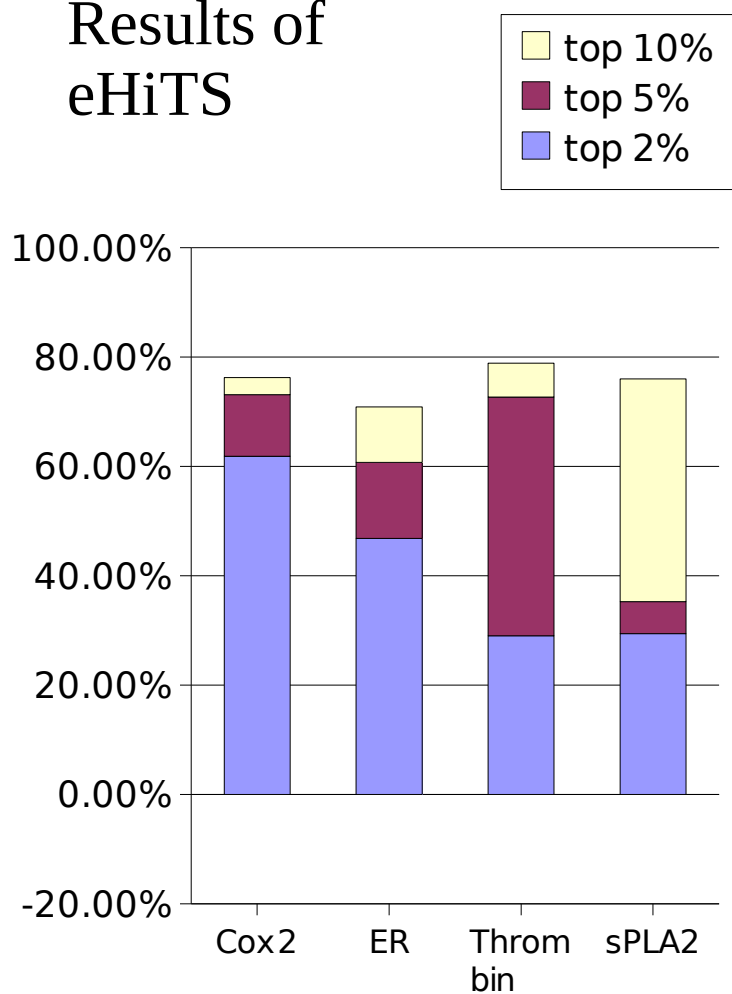


23. Enrichment results on the AstraZeneca data set

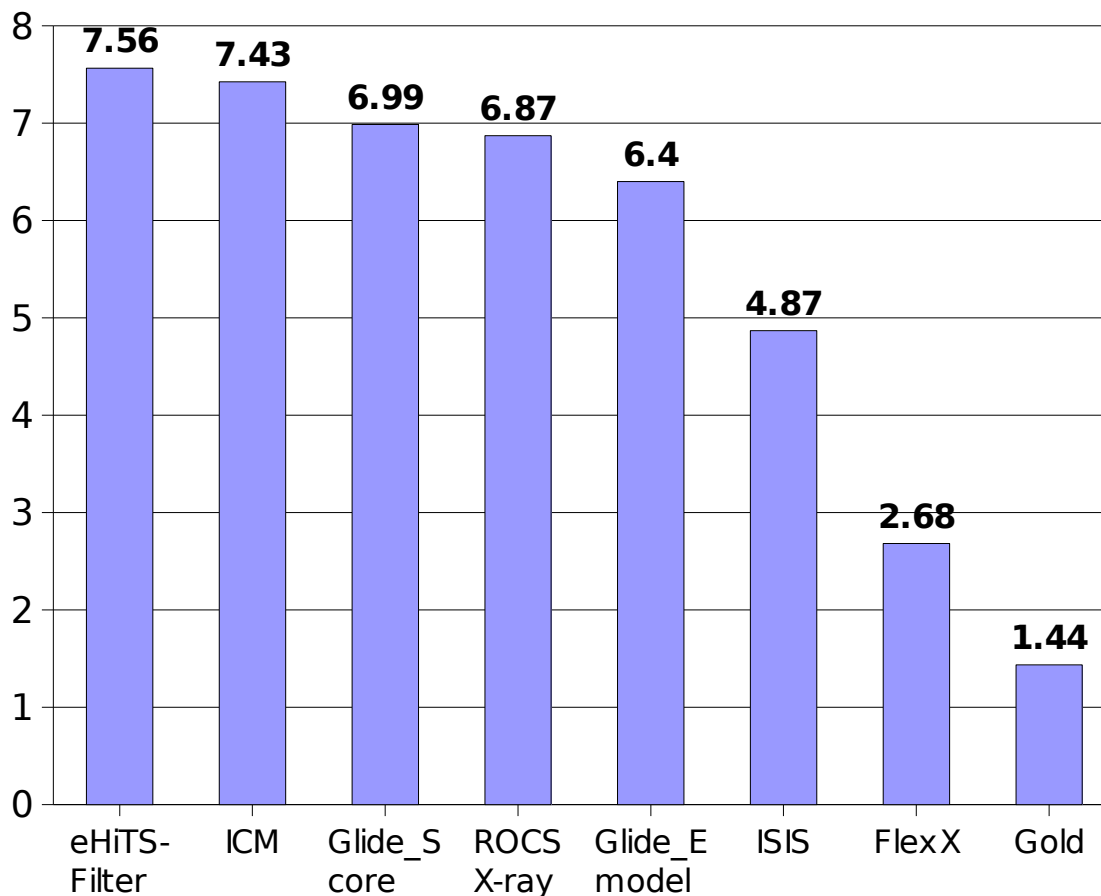


Hongming Chen et.al. , J. Chem. Inf. Model.; 2006; 46(1) pp 401 - 415

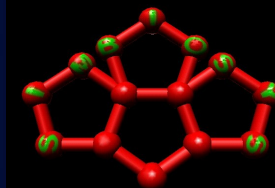
Results of eHiTS



Average enrichment factor comparison:

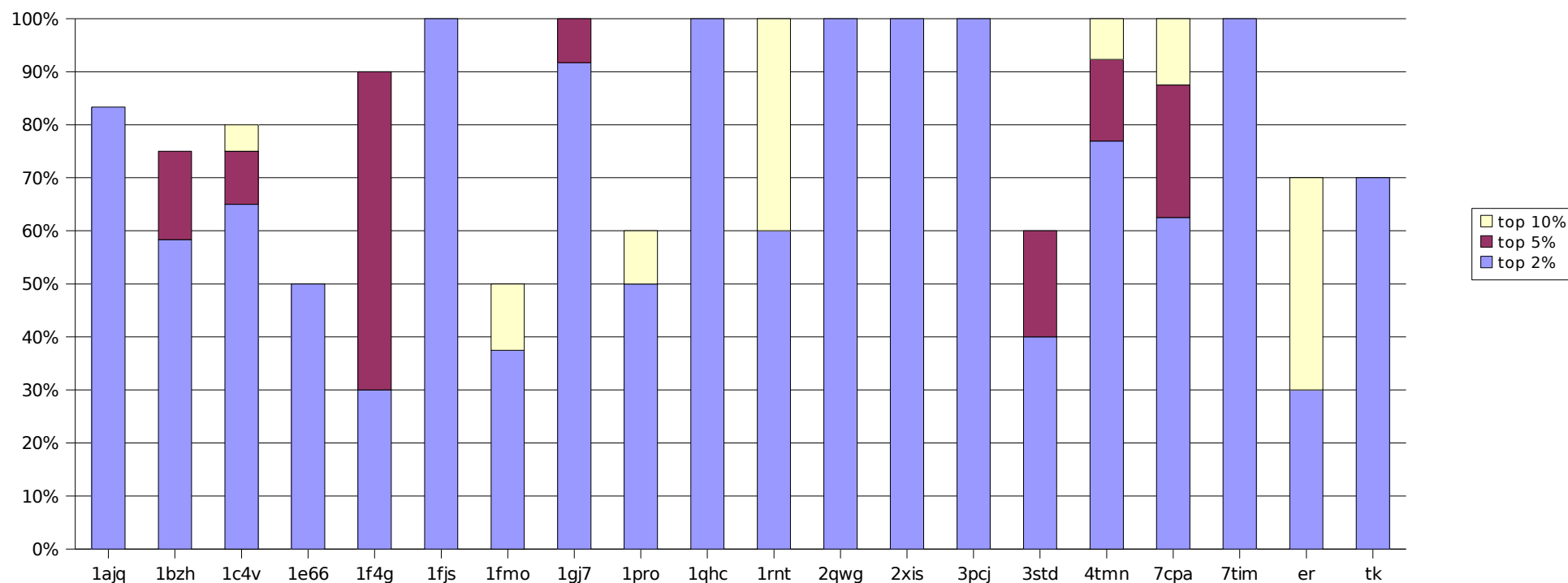


24. Enrichment Results on Surflex set

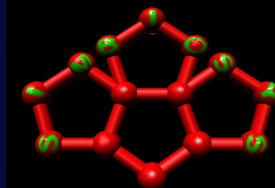


eHiTS was used to screen a dataset of 869 decoys plus actives (ranging from 5 to 20). The results show remarkable enrichment across a wide range of receptor families, with the average enrichment of ~80% of the actives recovered in the top 10% of the ranked database.

20 Codes out of the 29 Surflex set - screened with eHiTS_Filter



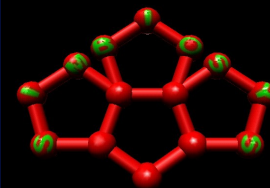
25. Future work



- PDB data error filtering based on matching to raw electron density map from Uppsala electron density server
- Hydrogen bond strength parameters based on QM calculations
- Mediating water placement
- Improved solvation based on explicit water molecules
- Automated detection of common pharmacophore patterns in protein families to generate constraints
- Functional group and fragment patterns matching pharmacophores
- Entropy estimate based on ensembles of poses



26. Summary of Features



- Interaction Surface Point based statistical interactions scoring
- Additional scoring terms combined with empirical weight set
- Automated protein family clustering and specialized weight tuning
- Pre-optimized weight sets for over 500 families included
- Automated tuning tool to customize the scoring for in-house data
- Very fast – eHiTS = **e**lectronic **H**igh **T**hroughput **S**creening (parallel/cluster support: SMP, linux clusters, PBS, LSF, SGE)
- eHiTS Lightning = **e**xtrremely **H**igh **T**hroughput **S**creening on IBM's Cell B.E. Supercomputer-in-a-chip

Visit us at Booth #316 at the exhibition

Request a free evaluation online: <http://www.simbiosys.ca/>