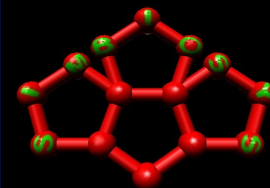


Scoring Performance of eHiTS on the CSAR dataset



Zsolt Zsoldos

SimBioSys Inc., © 2010

www.simbiosys.com

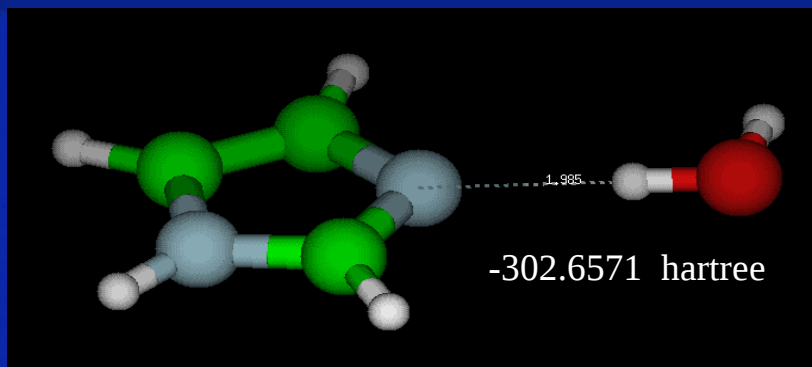
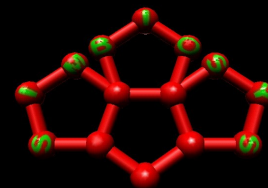
booth #945



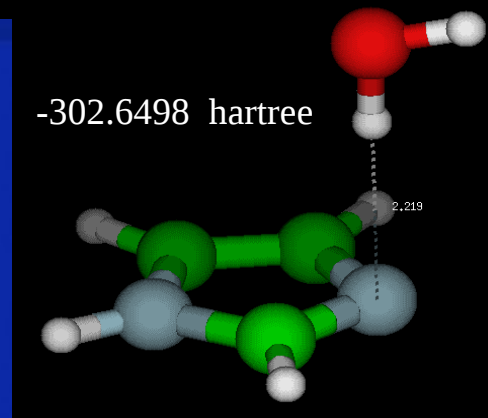
Contents:

- Interaction Surface Point (ISP) based scoring method
- Statistically derived parameters with empirical weights
- Protein family clustering and recognition
- Automated parameter tuning method
- Results on the CSAR dataset
- Effects of tuning on the CSAR dataset

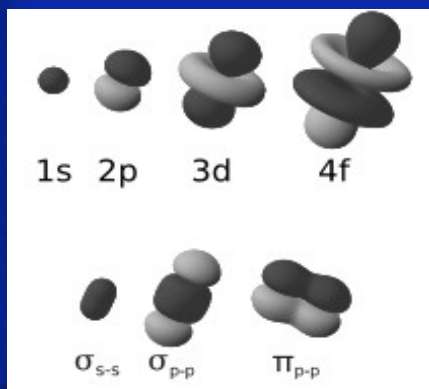
2. Problem with atom-centre based scoring models



ΔE 4.5 kcal/mol

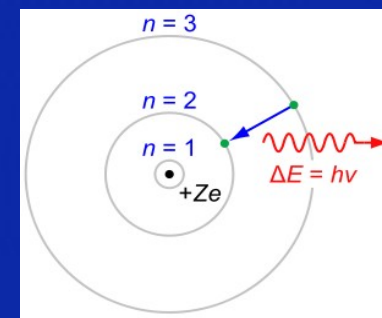


- Imidazole: 4.5 kcal/mol difference between lone-pair direction and above plane direction based on QM calculation
- Atom-center based QM-fitted point charge FF model => no difference!
- Fundamental contradiction between QM and FF models:

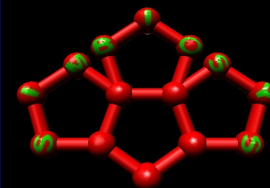


QM: all about electron density
(location probability)

FF: ignores electron density
~ century old Bohr model

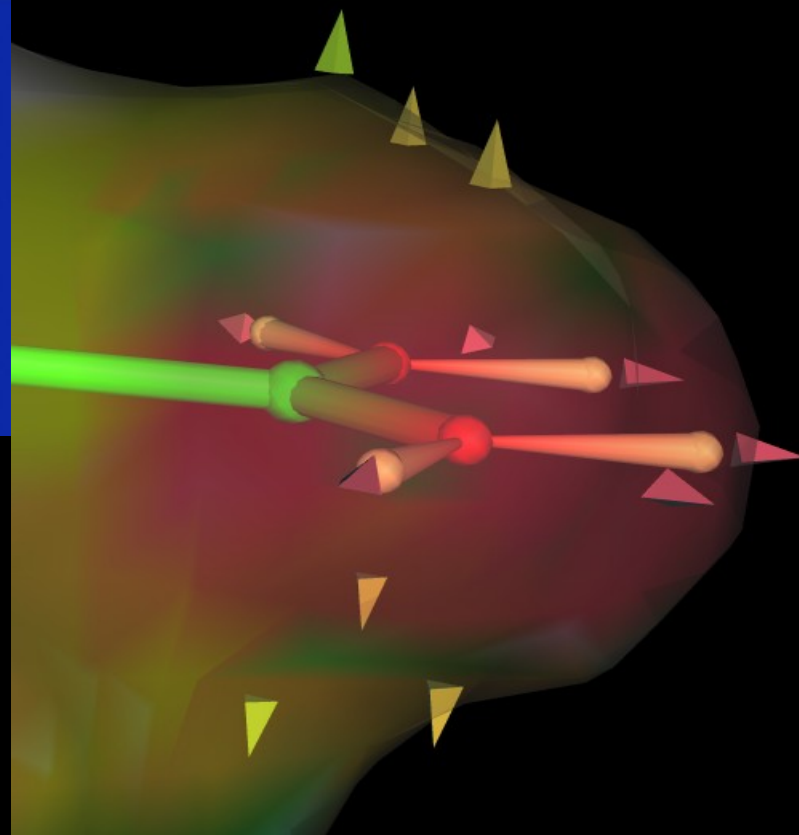
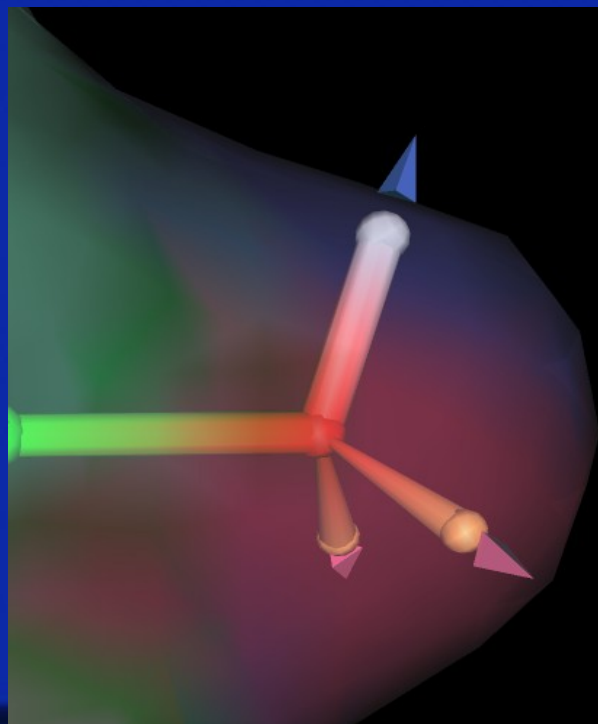


3. Interaction Surface Points (ISP)

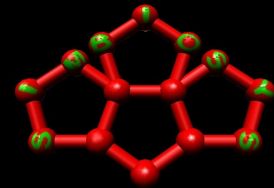


- eHiTS places directional surface points in specific locations on the surface of molecules to represent various interaction capabilities:

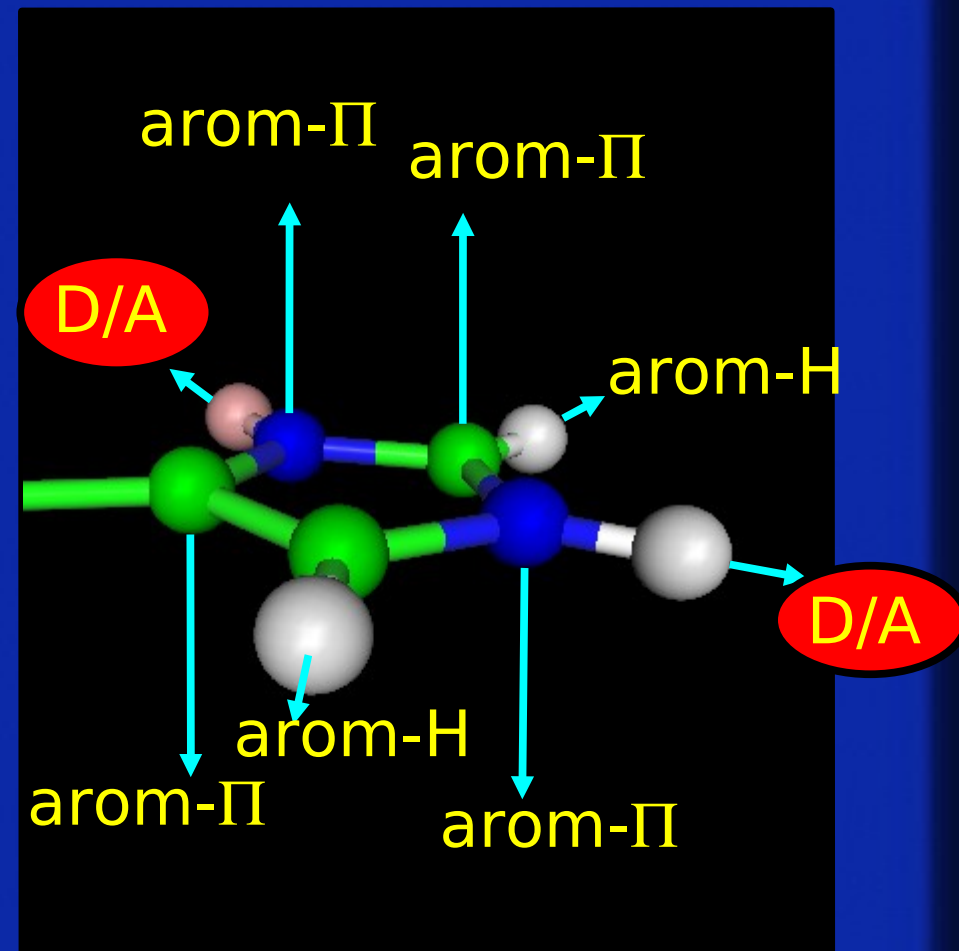
- H atoms,
- lone electron pairs,
- π electrons



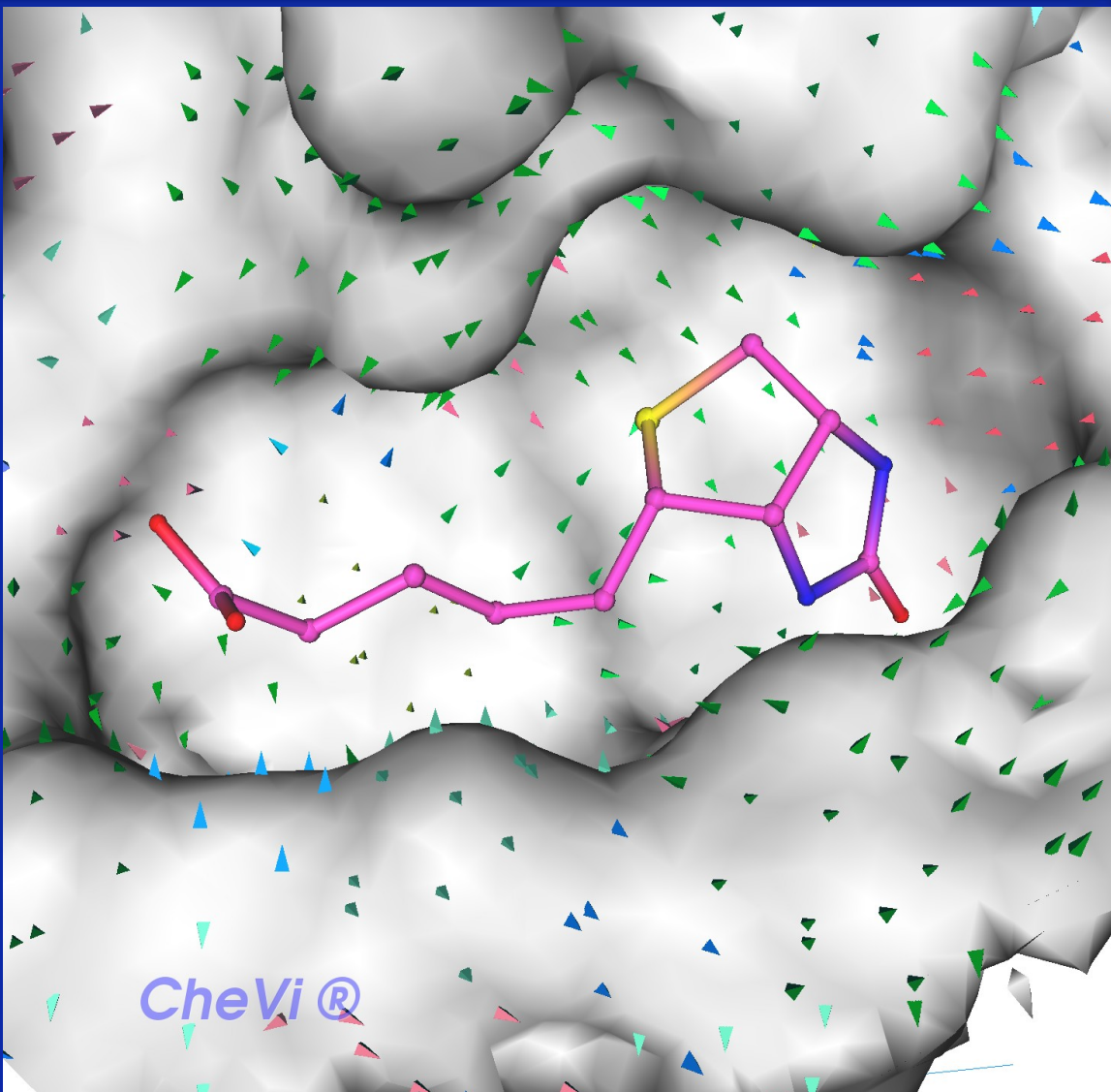
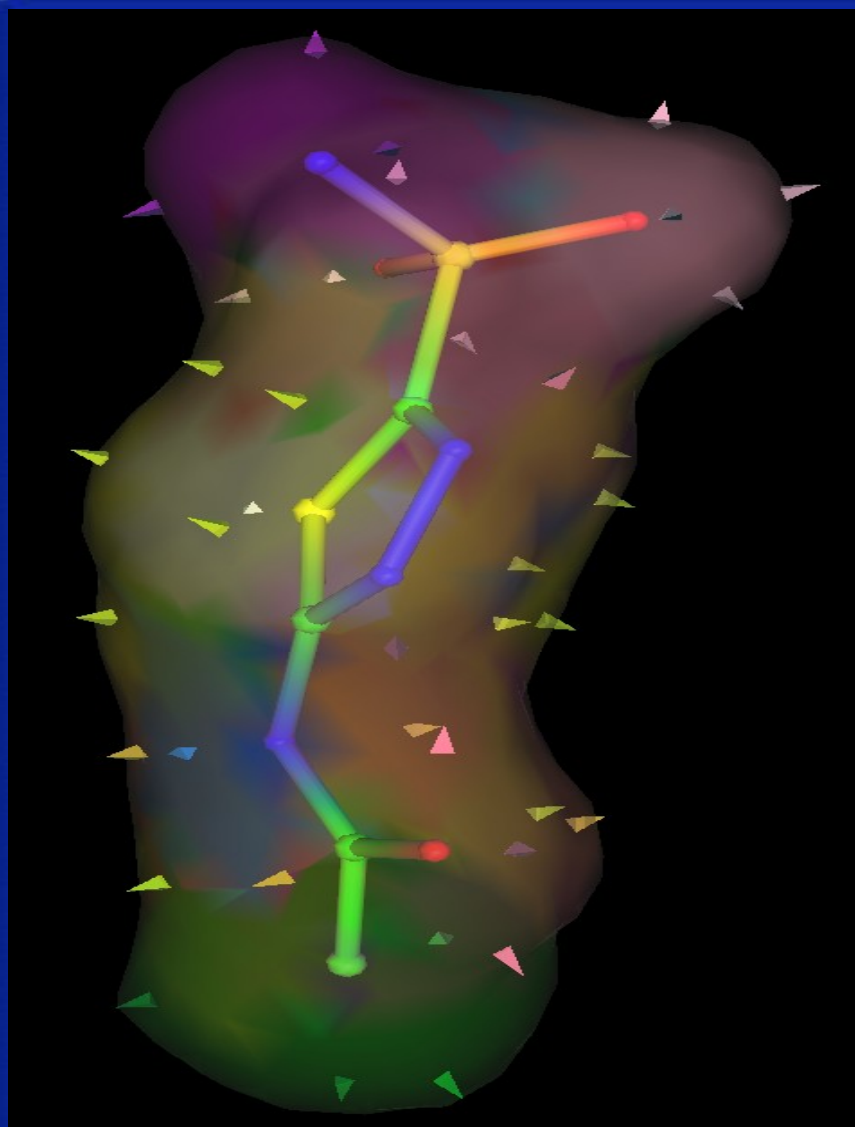
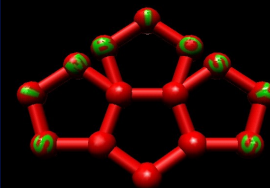
4. Interaction surface point (ISP) types



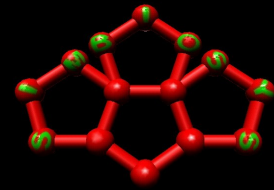
- **H-bond Donors:**
 - charged, amine, strong, weak, rotatable
- **H-bond Acceptors:**
 - charged, acid, strong, weak, rotatable
- **Ambivalent H donor/acceptor**
- **Aromatic Pi-stacking:**
 - carbon, polar, resonance, edge-H, arom- π
- **Hydrophobic:**
 - strong / weak lipophil, neutral
- **Metal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



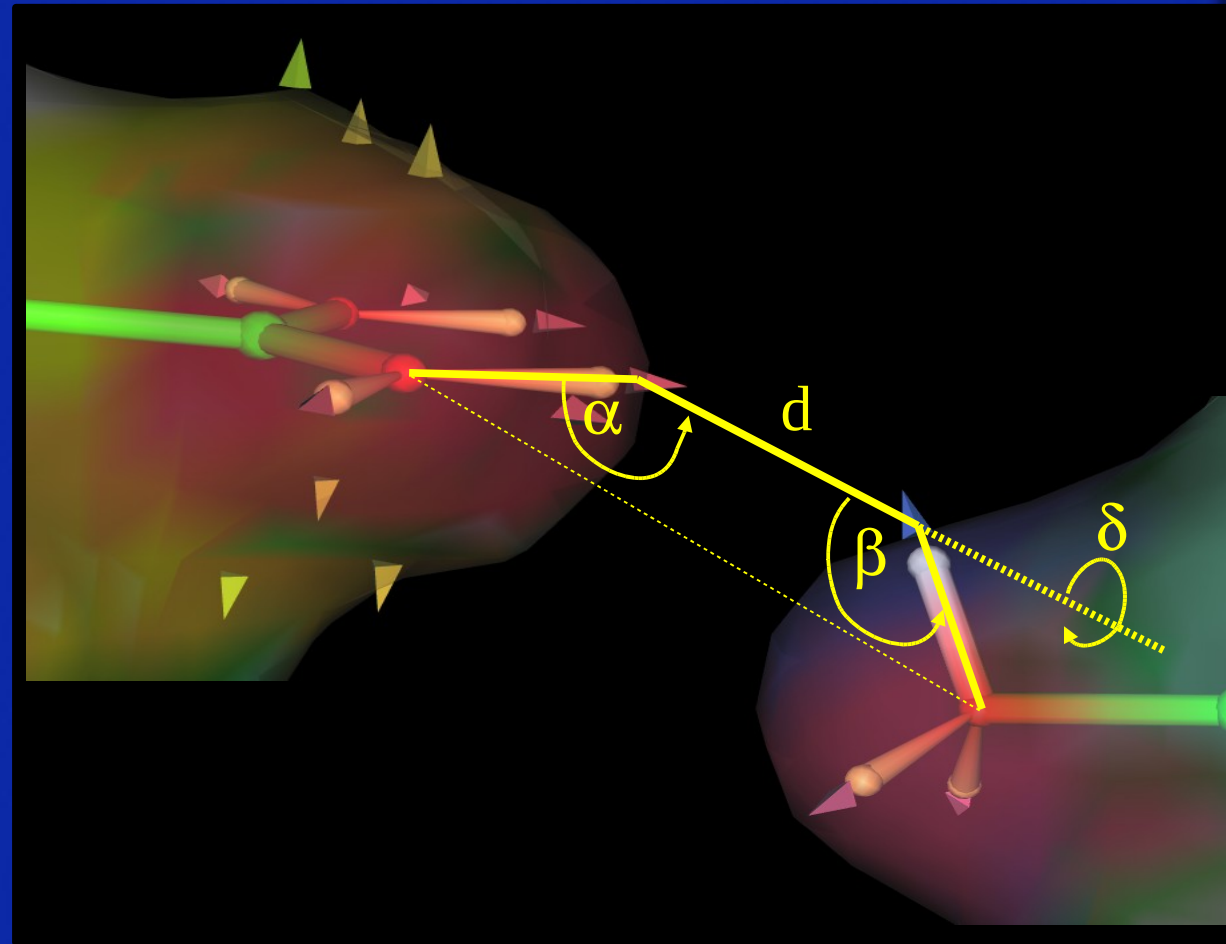
5. ISP set examples



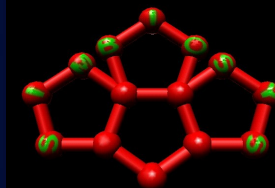
6. Interaction Geometry scoring



- 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:

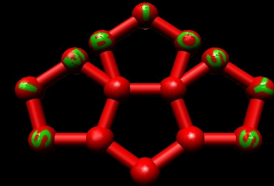


7. PDB file filtering and curation



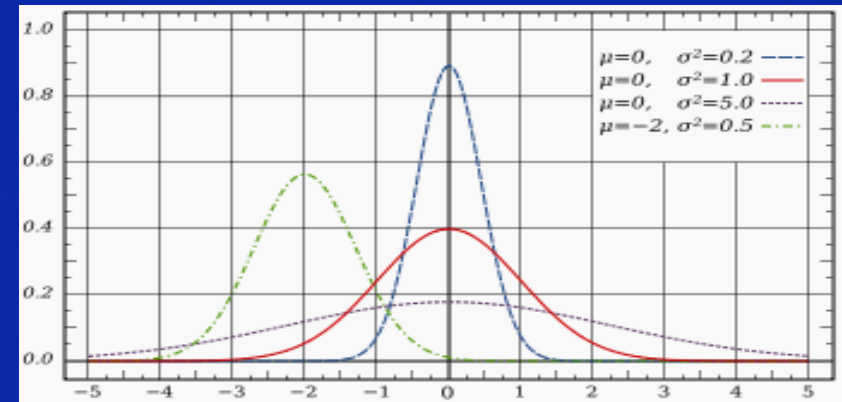
1. Protein-ligand complexes from the PDB, xray resolution 2.5Å or better: ~21,000
2. The PDB-report created by the WHAT_CHECK software was used for filtering:
Errors in protein structures. R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Nature (1996) 381, 272-272
 - Major modeling errors
 - High bond 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

8. Statistical data collection



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

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

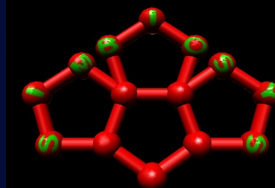


- Probability of distance d to occur between two heavy atoms:

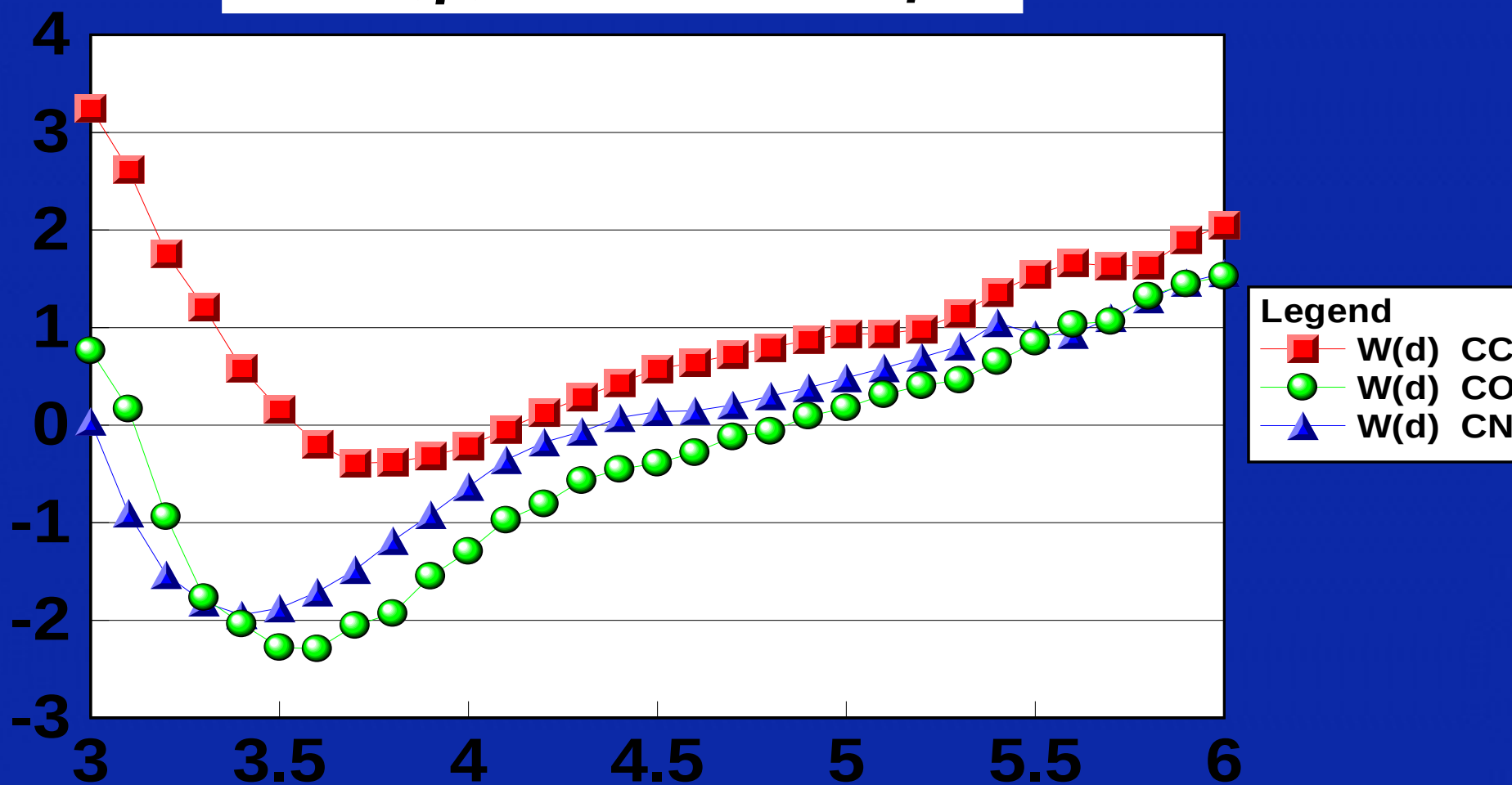
$$P(d) = \left(\frac{4\pi}{B_0 + B_1}\right)^{\frac{3}{2}} d^2 \int_0^\pi \int_0^{2\pi} \exp\left(\frac{-4\pi^2}{B_0 + B_1} \|P_0 - P_1 + P_s\|^2\right) \sin \alpha d\alpha d\beta$$

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

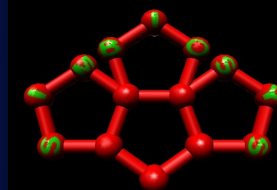
9. Energy by Boltzmann distribution



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



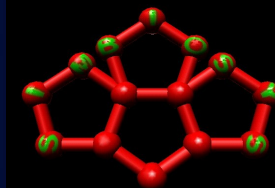
10. The Scoring matrix



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

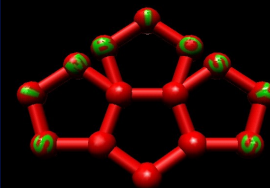
Re\Li	DonH+	Amine	Don-H	PO3--	AcidL	AccLp	WS-Lp	Ambiv	Rot-H	RotLp	CLipo	AromH	WSlip	Neutr	AromP	Res+	Res_C	Sp2+	Sp2_C	Halog	Join.	
METAL	-9.99	1.78	0.02	2.18	2.12	0.57	0.43	5.54	0.32	0.12	-4.8	-4.18	-3.19	0.27	-4.11	0.15	-1.49	0.98	-1.71	-6.46	4.39	METAL 0
DonH+	-5.15	-6.13	-5.38	2.57	3.8	1.14	-0.49	-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.	

11. 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)

12. Protein “family” clustering



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

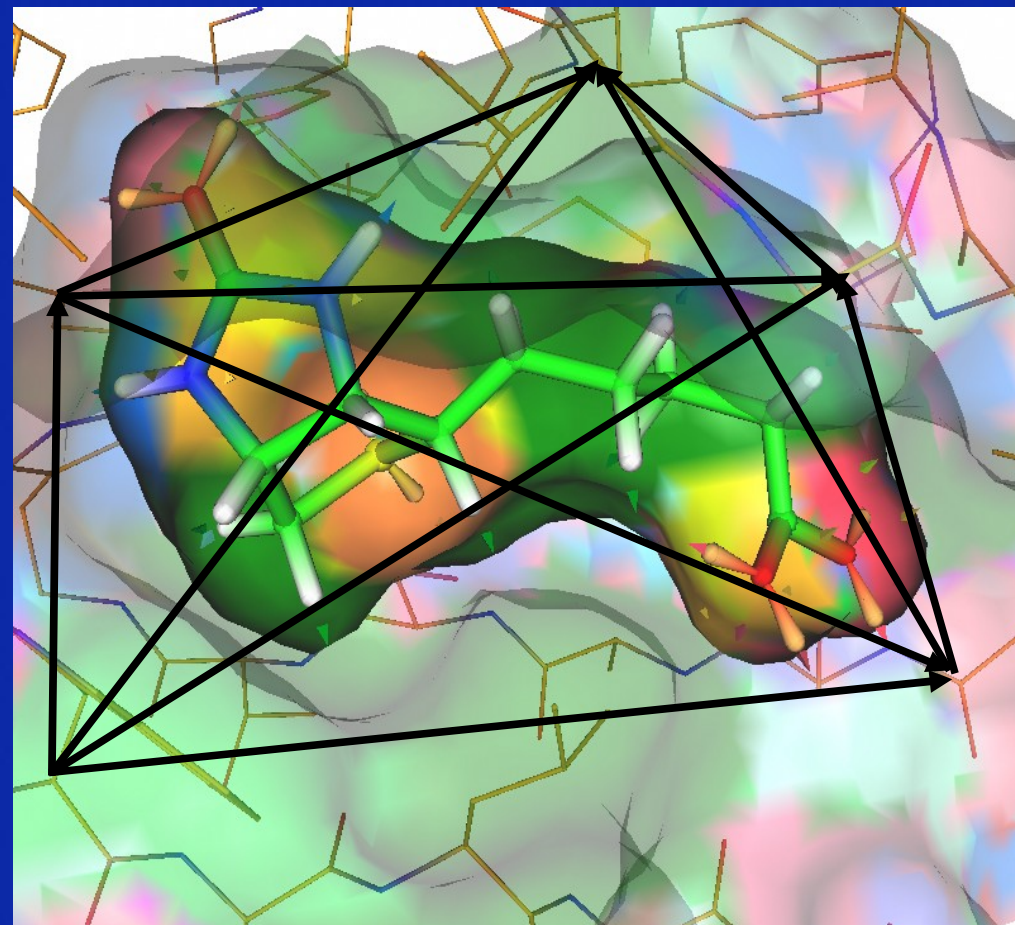
Geometric clustering is based on binding site residue C α distance matrix.

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

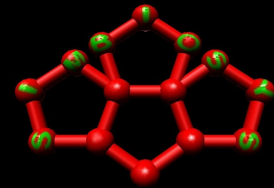
Correspondence to biological activity family is not exact, e.g. Kinase DFG-in DFG-out is separate, but thrombin and trypsin in same set.

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



13. 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

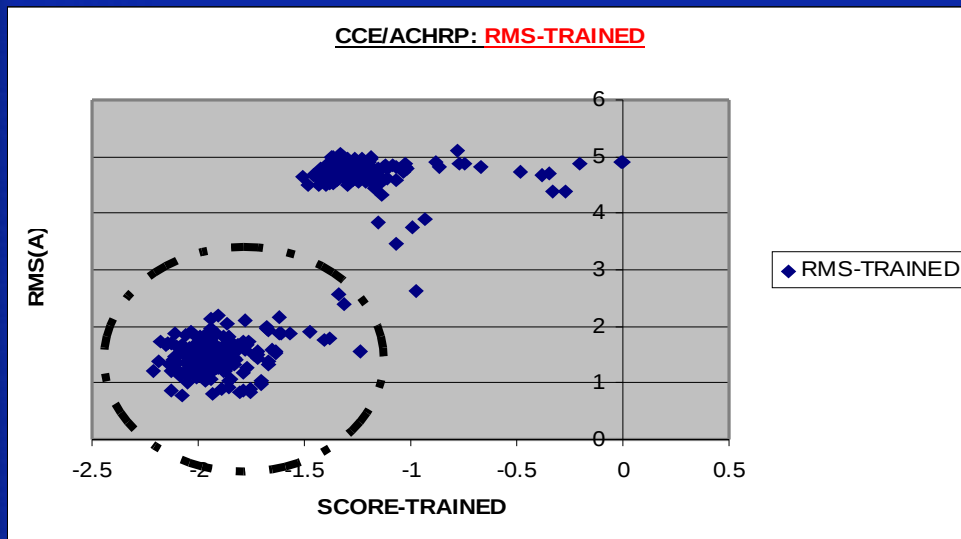
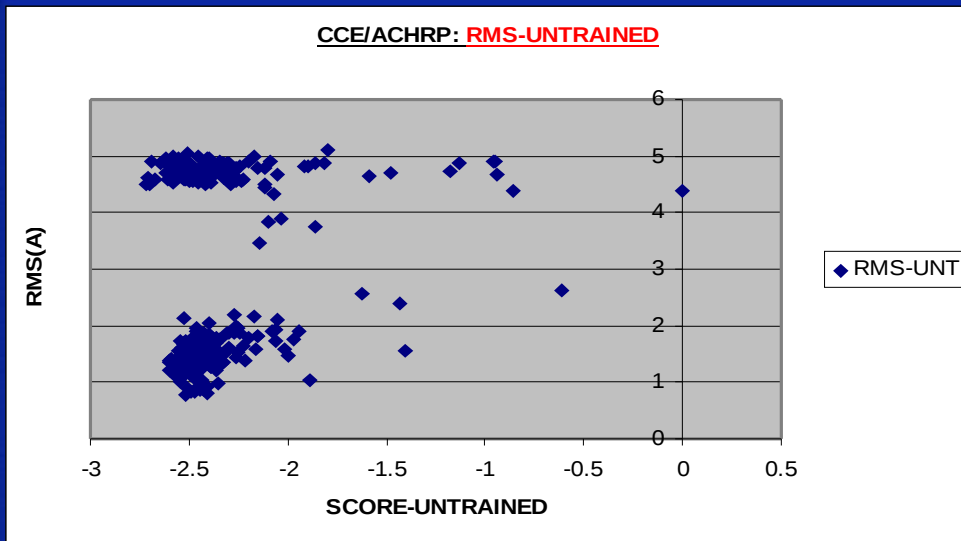
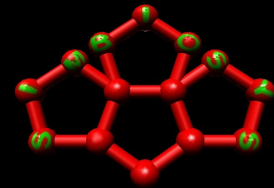
Along with the additional terms, we have 20 weight parameters to tune

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)

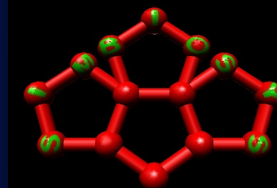
14. Effect of rank tuning



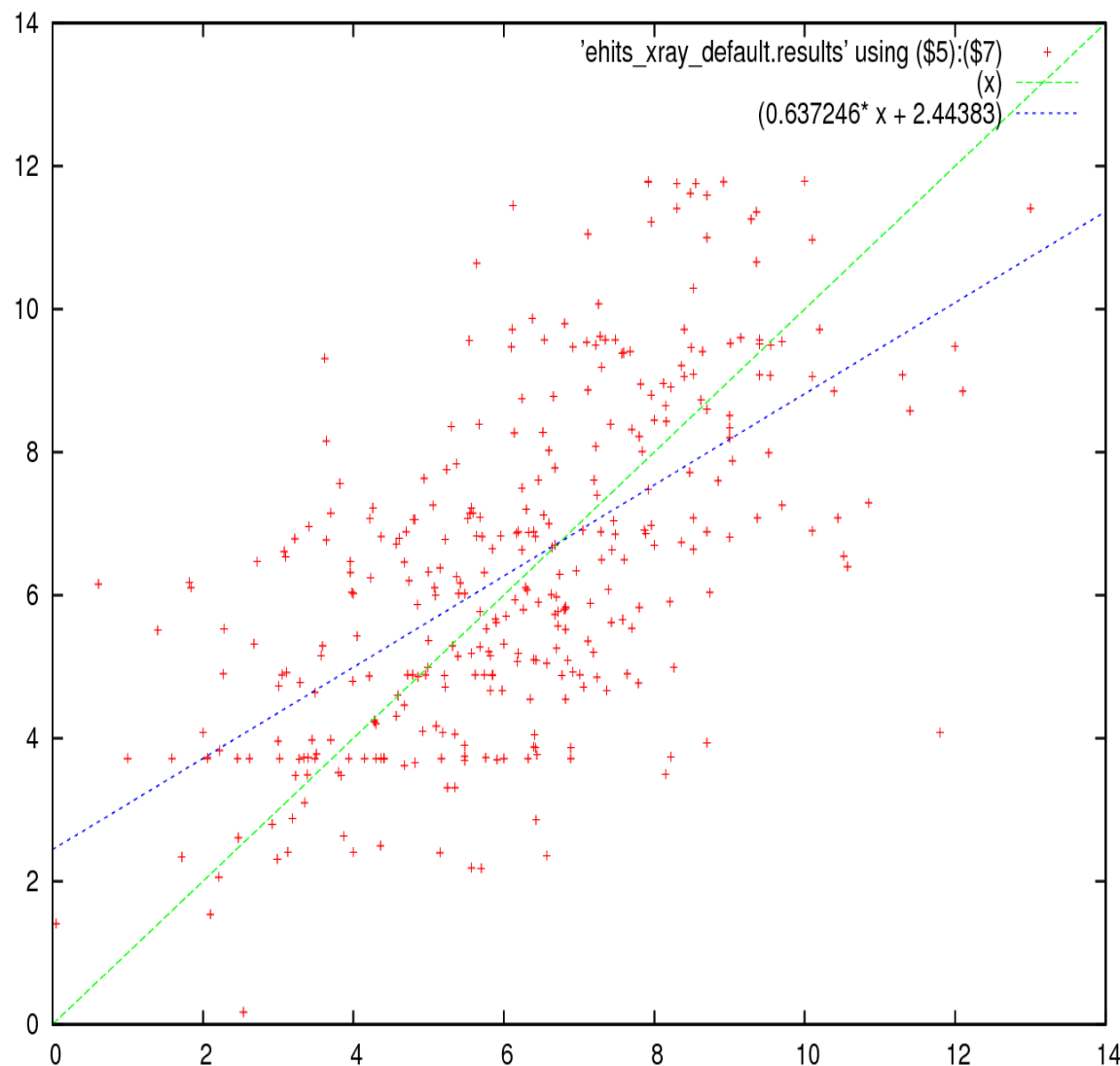
Untrained Scoring: Note that while there are many low RMS solutions in the good score regime there are also high RMS solutions with same score range

• **Trained Scoring:** There is dramatic score-separation of the 'correct' pose RMS-regime (circled) at low scores from poor scoring –high RMS results

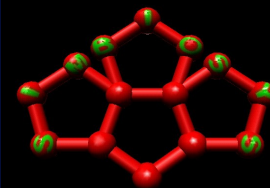
15. Results on the CSAR set using the default weight sets of eHiTS



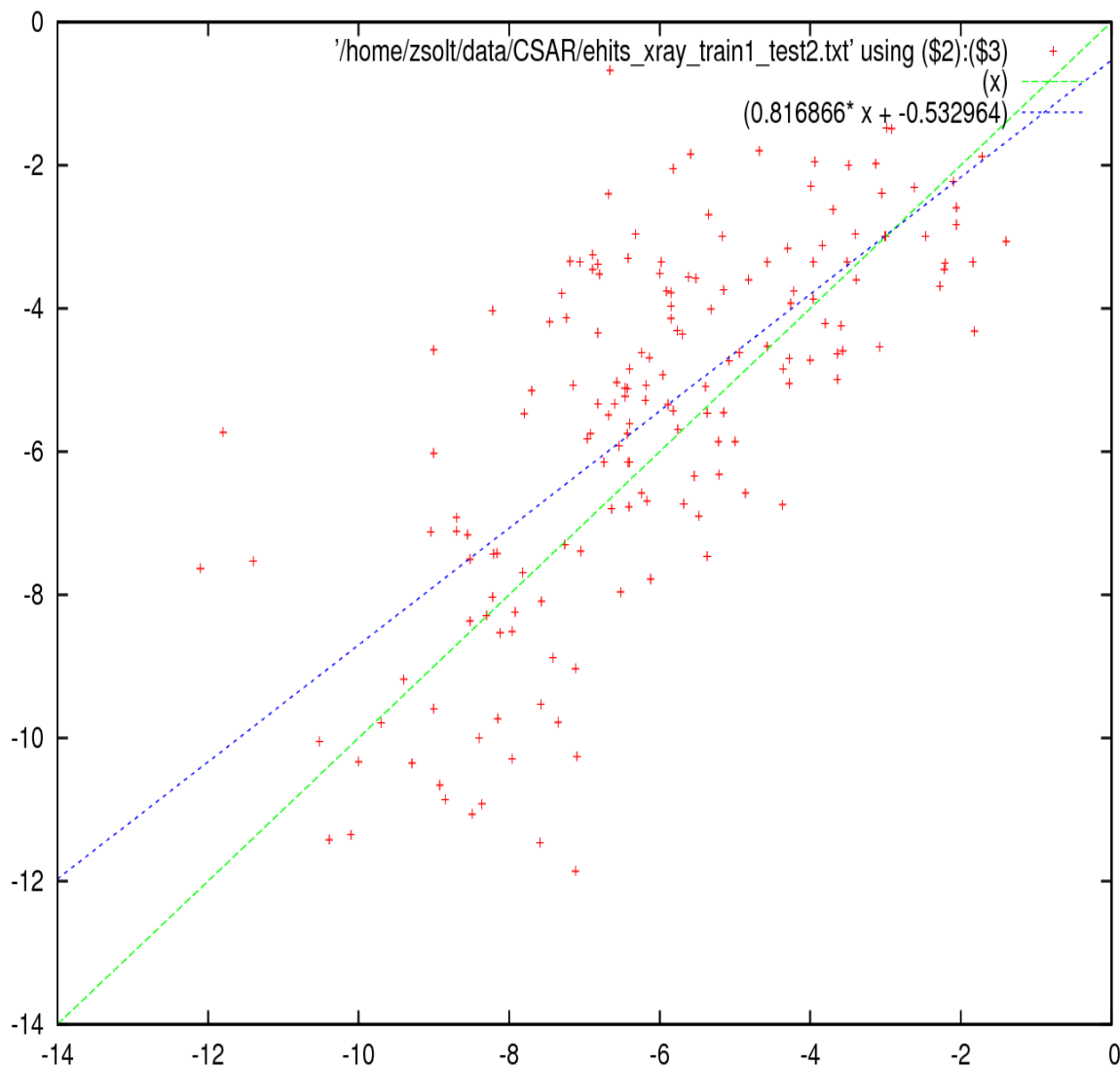
- $R^2=0.37$, $R=0.61$
- $Rmsd=2.02$
- 183 PDB complex codes are common between the CSAR set and the set of ~13,000 used for eHiTS scoring development, i.e.
 - 53% of CSAR set
 - 1.4% of the tuning set
- EhiTS scoring is tuned for pose recognition and enrichment, not for binding energy estimation



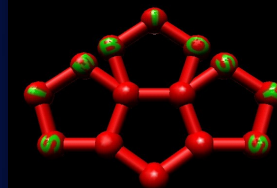
16. Results on the CSAR set2 using weights tuned based on CSAR set1



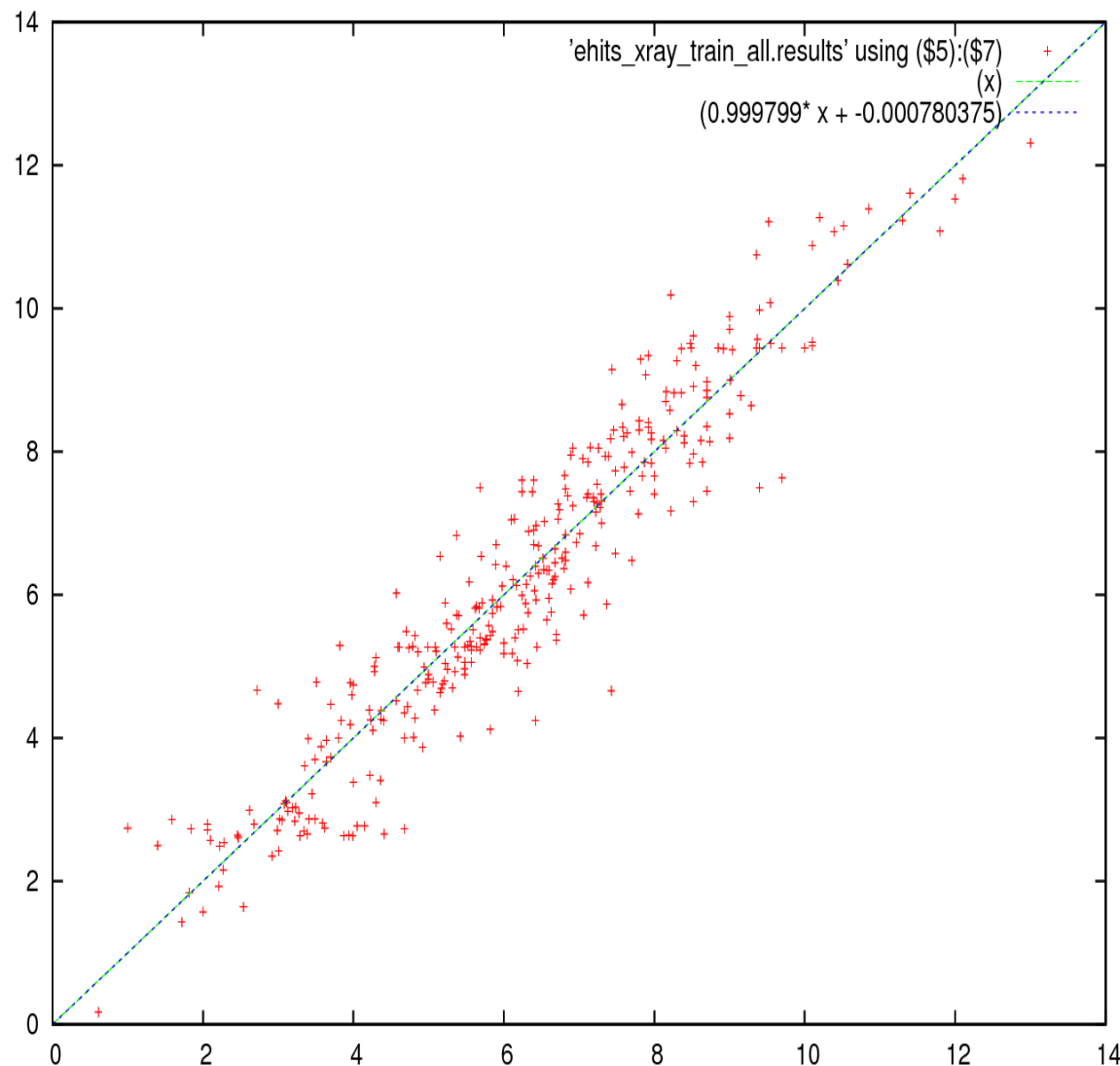
- $R^2=0.48$, $R=0.70$
- $Rmsd=1.94$
- Common PDB codes with eHiTS default tuning set:
 - 69 with set 1 of CSAR
 - 115 with set 2 of CSAR
- But, there is no code in common between set1 and set2, and the default eHiTS weight set is irrelevant for this study



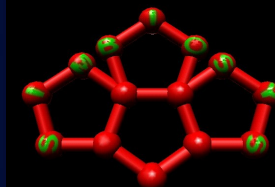
17. Results on the full CSAR set using weights tuned on the full CSAR set



- **WARNING: overtrained!**
- $R^2=0.90$, $R=0.95$
- $Rmsd=0.74$
- Not an indication of predictive power!
- Interesting note: same type of full-training on the PDB-bind 2008 set gives only $R^2=0.6$ value, which indicates that the CSAR set is significantly “cleaner”



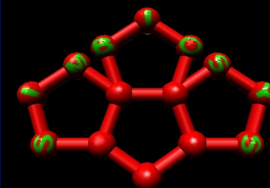
18. Changes in the weights upon tuning for CSAR set



name-of-the-term	all	set-1	default
METAL_ENERGY	1.22	2.05	1.03
HYDROGEN_BONDS	5.86	4.45	7.48
LIPOPHILIC	3.08	2.34	2.27
PI_STACK_ENERGY	1.97	1.52	1.99
OTHER_CONTACT	0	0.46	0.09
SOLVATION	2.27	2.78	2.41
STERIC_CLASH	0	0	0
FAMILY_COVERAGE	3.62	5.9	4.51
DEPTH	0	0.49	1.34
LIG_INTERNAL	9.43	0.13	9.79
RL_CHARGE	0.1	0.13	0.05
RL_SHAPE	0.35	0.64	0
RL_LOGD	0	0	0
REC_COVER	0.02	0	0
LL_CHARGE	0	0.15	0.87
STRAIN_ENERGY	0.03	0.15	0.13
LL_LOGD	0	0	0
LIG_COVER	0.46	0.02	0.55
COULOMB	0.18	0.29	1.14
ENTROPY	2.08	0.72	3.68



19. 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
- Per-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

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