

Global Free Energy Scoring Functions based on Distance-Dependent Atom-Type Pairs

Christian Kramer

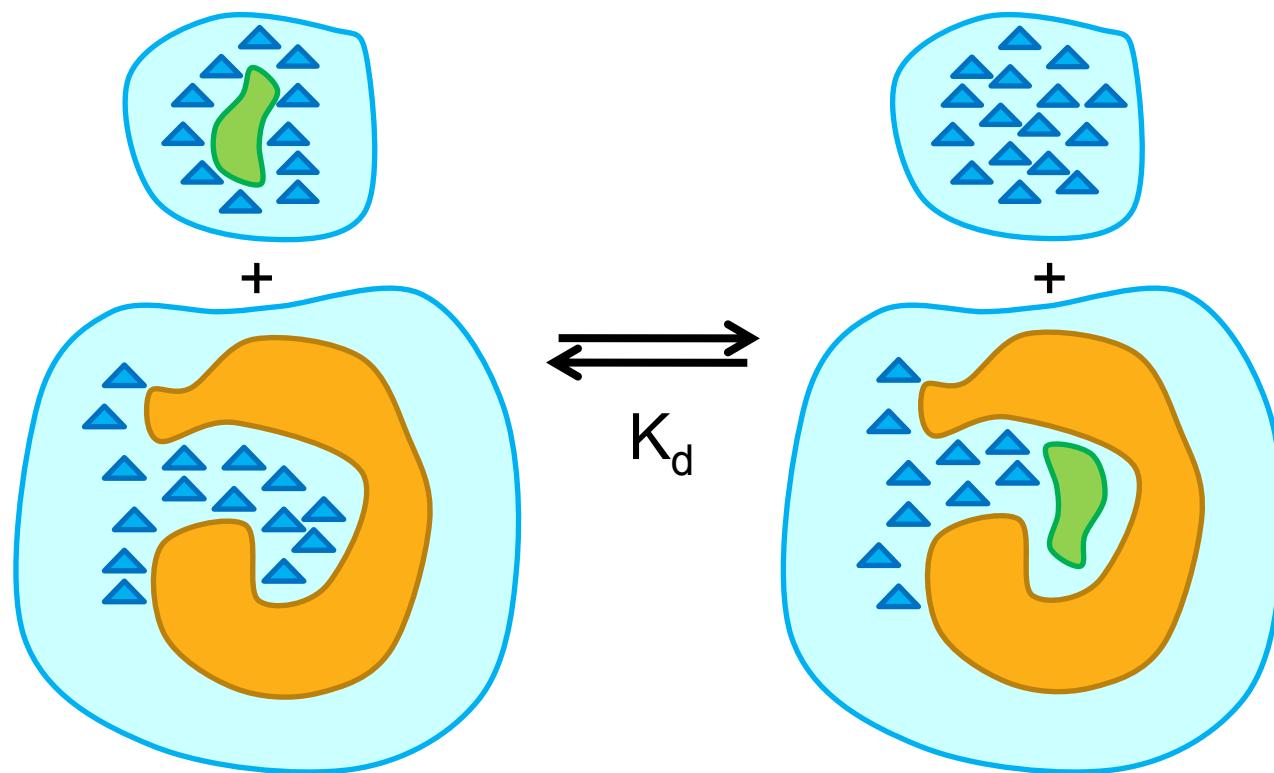
June 8th, 2011

International Conference on Chemical Structures,
Noordwijkerhout, The Netherlands

Overview

- Scoring – What's it all about
- State-of-the-art
- Distance-binned Atom-type Pair Descriptors
- Results on PDBbind09
- Results on the CSARdock HiQ benchmark set
- Summary & Conclusions

Docking & Scoring: The million \$\$\$ question



High pK_d ($-\log K_d$) is *conditio sine qua non* in drug discovery.

Estimation of pK_d is essential for structure-based drug design and all related tools (De-Novo design, Structure-based Virtual Screening, ...).

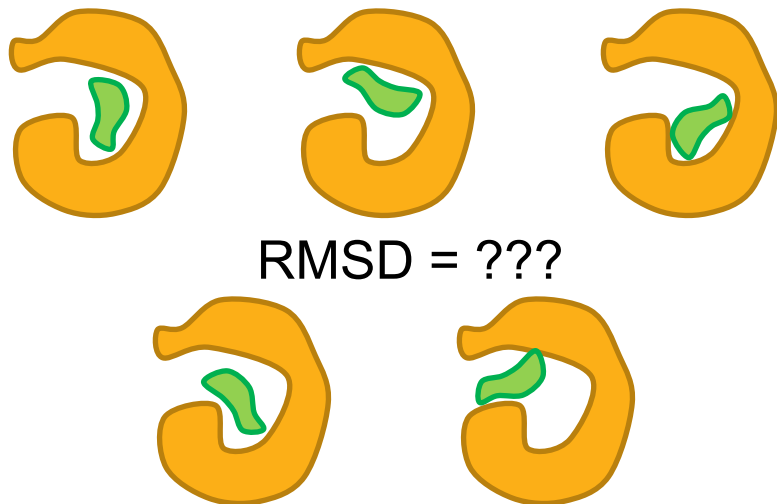
Separate the Problem – Pose Finding and Energy prediction

Community feel:

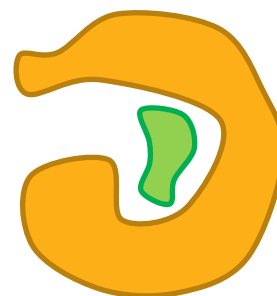
Pose Prediction works more or less (one out of ten predicted geometries with $\text{RMSD} < 2\text{\AA}$),
scoring only poorly.

→ Split problem in two pieces:

Pose Finding



Free Energy Estimation



$\text{pK}_d = ???$

Crystal Structure & Binding Data Database

PDBbind & PDBbindCN (R. Wang, S. Wang et al)

Probably largest collection of crystal structures complemented with binding data.

PDBbindCN 2009: 1741 crystal structures with K_d values, resolution $\leq 2.5 \text{ \AA}$, binding pocket & ligand completely resolved, no uncommon atoms in ligand, only non-covalent ligands.

(www.pdbbind-cn.org)

Ligand exclusions in this study:

MWt	>	900	(mainly polypeptides),
#P	>	1	(mainly ADP,ATP,NADH,NADPH)
#donors + #acceptors	>	20	(mainly polyglycosides)

→ 1387 complexes left

Standard commercially available Scoring Functions

Performance of Standard commercially available Scoring Functions on the 1387 subset of PDBbind2009 Refined Set:

Correlation of Predicted Scores with measured Free Energy of Binding

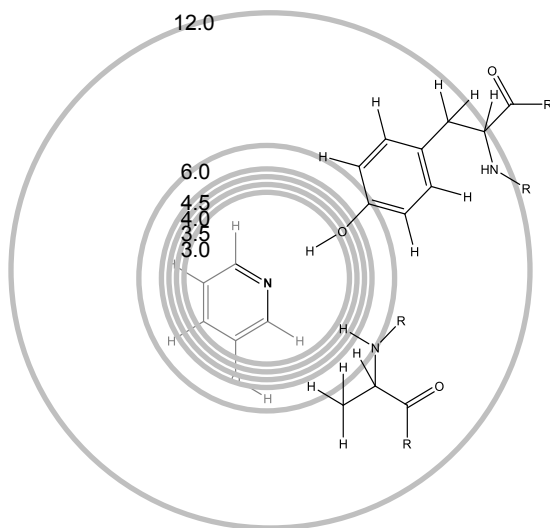
	ASP	Chem Score	GOLD Score	GlideXP	HMScore	HPScore	HSScore
R^2_{pearson}	0.25	0.25	0.18	0.10	0.36	0.37	0.37

All Scoring Functions have been trained on parts of this validation set, so the true performance is probably lower.

Refitting improves GOLDScore to $R^2_{\text{pearson}} = 0.35$.

Coding essential interactions in distance-dependent atom-type pair descriptors

Atom-type pair frequencies in specified distances



Descriptors:

- $6 (\#dist) * 84 (\#L-atypes) * 39 (\#P-atypes)$
= 19656 descriptors (**ddPLATp**) , ~ 50% highly sparse
- $6 (\#dist) * 9 (\#L-NonH-atoms) * 4 (\#P- NonH-atoms)$
= 540 descriptors (**ddPLEp**)
- $84 (\#L-atypes) + 39 (\#P-atypes \text{ in Bdg Pocket})$
= 123 descriptors
- simple **MOE** ligand only descriptors (N=23)

Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors. *J. Chem. Inf. Model.*, **2011**, *51*, 707–720.

Crippen-like atom typing

Table 1. Atom Typing SMARTS List for Carbon^a

ID	SMARTS	logP incr	don/acc/neu
C1	[CH4]	0.1441	N
C1	[CH3]C	0.1441	N
C1	[CH2](C)C	0.1441	N
C2	[CH](C)(C)C	0	N
C2	C(C)(C)C	0	N
C2	[CH3][N,O,S,F,Cl,Br,I]	-0.2035	N
C3	[CH2 × 4][N,O,S,F,Cl,Br,I]	-0.2035	N
C28*	[C](=[O,N])[O,N]	-0.2051	N
C4	[CH1 × 4][N,O,S,F,Cl,Br,I]	-0.2051	N
C4	[CH0 × 4][N,O,S,F,Cl,Br,I]	-0.2051	N
C5	[C]=[!C;A;!#1]	-0.2783	N
C6	[C;A]=C	0.1551	N
C7	[CX2]#[A;!#1]	0.0017	N
C8	[CH3]c	0.08452	N
C9	[CH3]a	-0.1444	N
C10	[CH2 × 4]a	-0.0516	N
C11	[CHX4]a	0.1193	N
C12	[CH0 × 4]a	-0.0967	N
C13	[cH0]-[A;!C;!N;!O;!S;!F;!Cl;!Br;!I;!H]	-0.5443	N
C14	[c][#9]	0	N
C15	[c][#17]	0.245	N
C16	[c][#35]	0.198	N
C17	[c][#53]	0	N
C18	[cH]	0.1581	N
C19	[c](a):(a)a	0.2955	N
C20	[c](a):(a)-a	0.2713	N
C21	[c](a):(a)-C	0.136	N
C22	[c](a):(a)-N	0.4619	N
C23	[c](a):(a)-O	0.5437	N
C24	[c](a):(a)-S	0.1893	N
C25	[c](a):(a)=[C,N,O]	-0.8186	N
C26	[C](=C)(a)[A;!#1]	0.264	N
C26	[C](=C)(c)a	0.264	N
C26	[C](=C)a	0.264	N
C26	[C]=c	0.264	N
C27	[CX4][A;!C;!N;!O;!S;!F;!Cl;!Br;!I;!#1]	0.2148	N
CS	[#6]	0.08129	N

^aThe full SMARTS definition for all elements can be found in the Supporting Information. The star(*) denotes the atom types different from the standard Crippen types.

Model Building

- Stepwise Multiple Linear Regression (MLR)
- Bagging
(N = 50 independent models, 75% randomly picked training set
→ Each sample has 12.5 test set predictions on average.)
- Descriptor-pool size adjusted F-value as stopping criterion

(n = number of samples, k = number of descriptors available, p = number of descriptors in equation)

$$F = \frac{R_{q+1}^2 - R_q^2}{(1 - R_{q+1}^2)/(n - q - 2)}$$

$$F_{max,step}(n, k, p = 1) = \exp\left(\sum_{j=1}^7 \beta_j y_j\right)$$

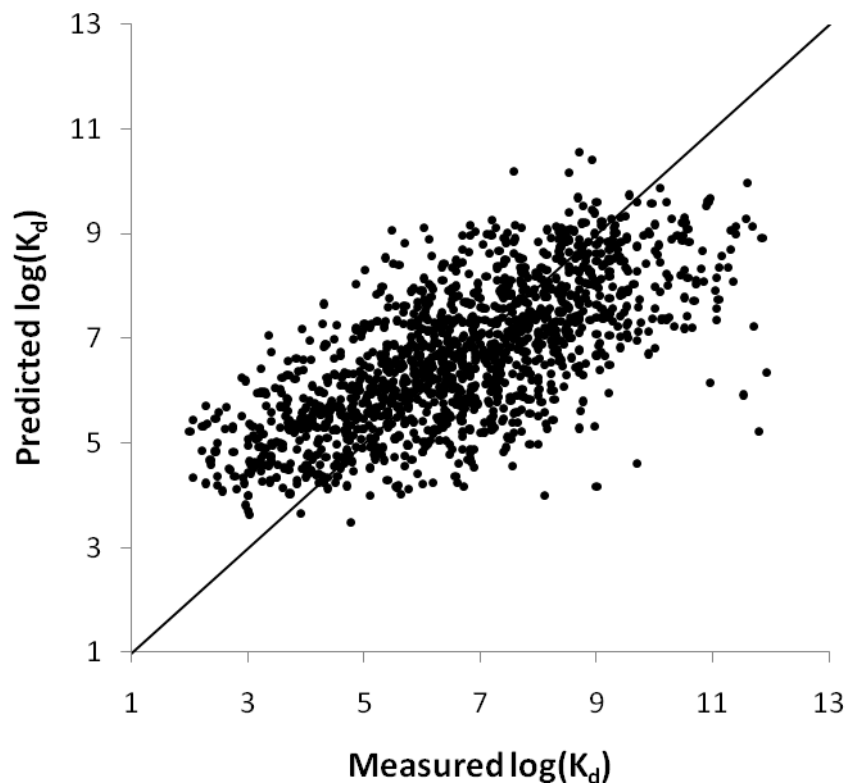
Table 3. Coefficients for the Equation for Approximating $\ln(F_{max,step})$ with $p = 1$

element	coefficient	P = 0.90	P = 0.95	P = 0.99
$y_1 = 1$	β_1	2.92371	2.99124	3.12901
$y_2 = k^{-0.5}$	β_2	-7.43273	-6.83193	-5.72839
$y_3 = n^{-1}$	β_3	12.3075	13.3999	15.9427
$y_4 = k^{-1}$	β_4	20.6095	19.3445	16.6906
$y_5 = n^{-1}k^{-0.5}$	β_5	-35.7882	-35.2874	-34.8941
$y_6 = k^{-1.5}$	β_6	-23.1200	-21.6866	-18.6847
$y_7 = n^{-1}k^{-1}$	β_7	42.6362	40.9349	39.3152

➔ Correct F-value and crossvalidation prevent overfitting.

C. Kramer, C.S. Tautermann, D.J. Livingstone, D.W. Salt, D.C. Whitley, B. Beck, T. Clark. Sharpening the Toolbox of Computational Chemistry: A New Approximation of Critical F-Values for Multiple Linear Regression *J. Chem. Inf. Model.*, **2009**, 49 (1), pp 28–34.

Performance on PDBbind2009 druglike Refined Set



Descriptor Set	Bagged MLR		
	R^2	RMSE	MUE
ddPLATp	0.45	1.47	1.18
ddPLEp	0.41	1.52	1.22
MOE	0.34	1.62	1.29
ddPLATp+MOEcounts	0.48	1.44	1.14

Results for independent out-of-bag test set

CSARdock benchmark

CSARdock HiQ benchmark set

(Community Structure-Activity Resource, H. Carlson et al)

Aims to provide very high quality crystal structure complex data with hydrogens/ protonation states optimized and complex minimized (AMBER force field).

Current size: 343 reliable complexes with reliable K_d data.

(www.csardock.org)

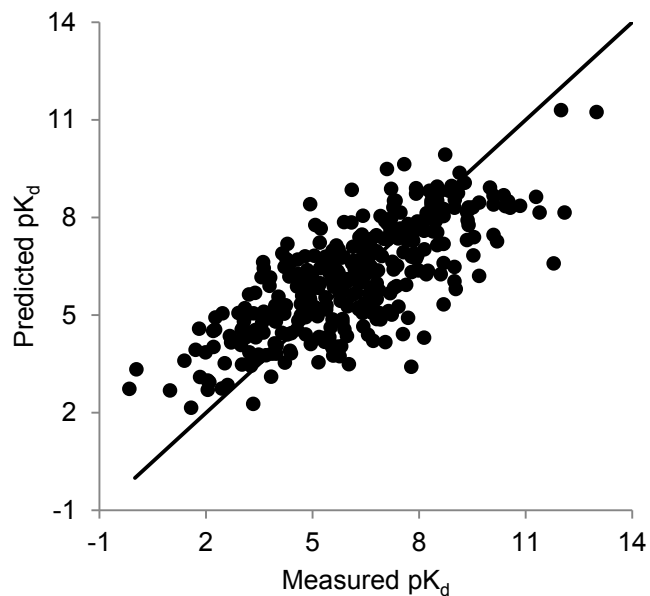
Collection of high-quality crystal structure & K_d data from public and private sources to provide a benchmark for current scoring functions and identify “problem” complexes that no current scoring function can handle.

Benchmark results presented at the ACS 2010 in Boston. Best scoring function result (Leadfinder): $R^2 = 0.62$.

Complexity reduction of the final model

- Pool-size adjusted F-value ensures that correlation is higher than correlation of 95% of random descriptors.
- Large number of submodels and descriptors/model: Some noise descriptors might enter nevertheless.
- Training set is different in every single model: Noise descriptors differ from model to model.
 - ➔ seldomly appearing descriptors might represent noise and can thus be removed.
 - ➔ Here: Only 3 descriptors appear in more than 50 % of the models (out of 4-6 descriptors/model).

CSARdock benchmark



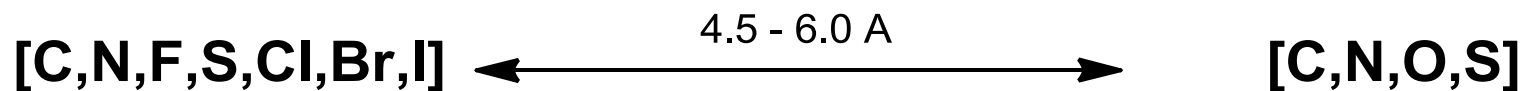
	Bagged MLR model		
	R^2	RMSE	MUE
Training Set	0.62	1.38	1.11
Test Set	0.52	1.54	1.22

	Pruned bagged MLR model		
	R^2	RMSE	MUE
Training Set	0.57	1.47	1.17
Test Set	0.55	1.49	1.19

Descriptor	Coefficient standardized	Coefficient direct
LX.PX.6.0	1.59	0.383
LH4.PC3.12.0	-0.77	-0.038
LS1.PC19.6.0	0.47	1.232
constant		-0.695

New descriptor: LX.PX: 4.5-6.0 Ångstrom distance

- LC.PC: 4.5 - 6.0 count has $R^2=0.41$ with free energy of binding.
- Adding the counts for [N,F,S,Cl,Br,I] for the ligand and all heavy atoms of the protein increases the correlation to 0.44.
- Adding O and P on the ligand decreases the correlation.

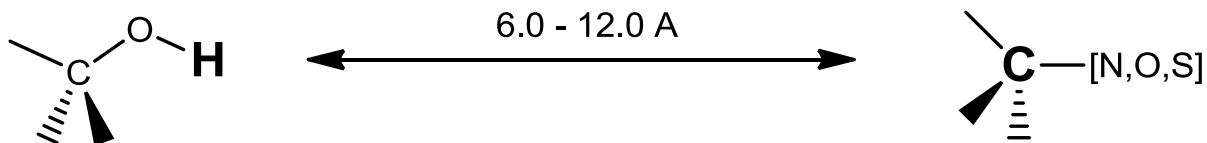


- Measure of Heavy Atoms times Buriedness
- Highly correlated with the solvent excluded surface area

LH4.PC3: 6.0 - 12.0 Ångstrom distance

H4: aliphatic hydroxy hydrogen

C3: aliphatic carbon, bound to at least one non-carbon heavy atom (occurs in methionine, proline, glycine, serine, cysteine, lysine and arginine)

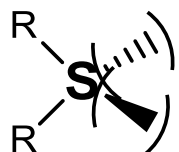


Surrogate for LH4 count?

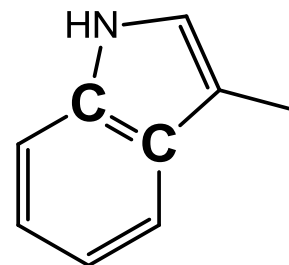
- Descriptor can indeed be replaced with count of ligand hydroxy hydrogens and oxygens, while nearly keeping the performance ($R^2 = 0.53$ vs. $R^2 = 0.55$)
- Complement to the LX.PX: 4.5-6.0 descriptor

LS1.PC19: 4.5 - 6.0 Ångstroem distance

This descriptor is nonzero only in 12 compounds, but nevertheless highly significant.

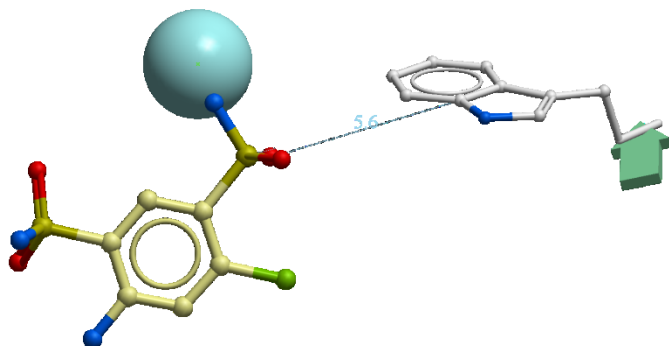


4.5 - 6.0 Å

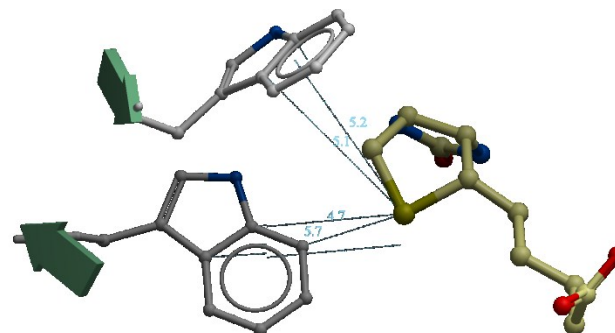


S1: Uncharged non-aromatic Sulfur

c neighbors
surs in Tryptophan



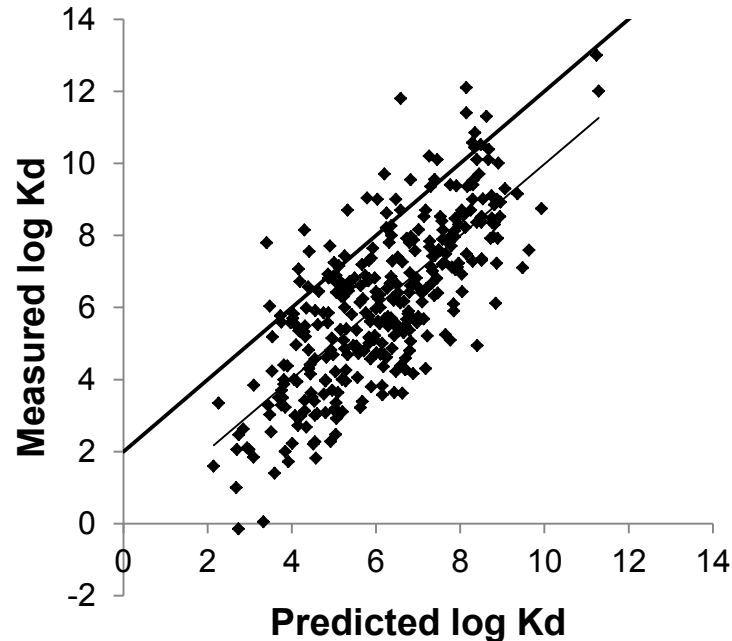
Interaction of Sulfonamide with
Tryptophan in 2pov.



Interaction of Biotin with Tryptophan
of Streptavidin in 1swk.

Structure-based Ligand Efficiency

- Classic Ligand Efficiency: ~ 0.36 kcal mol⁻¹ /heavy atom (Hopkins 2004)
- Refined (Structure-Based) Ligand Efficiency using Binding Site Model:



→ Takes important structural features like buriedness (PPI “pockets” vs. deep pockets) and ligand properties (atom types → hydrophobicity) into account

Summary

- Few rather simple descriptors can give surprisingly good models for the free energy of binding.
- Descriptor-based scoring functions can be built in a QSAR style and are very competitive in terms of prediction of free energy of binding.
- Bagging stepwise MLR with descriptor-pool size adjusted F-value and complexity reduction is a very useful tool for QSAR-like fitting problems in very high-dimensional spaces.
- There should be lots of space for improvements in scoring functions – the situation is not as bad as its reputation.
- The models presented themselves are not useful for real life applications (only positive examples in the training), but can be used to estimate an upper limit of binding affinity.

Consequences & Outlook

- In contrast to QSAR data sets the Crystal Structure & Activity databases will grow continuously – QSAR-like fitted scoring functions can and will easily be improved based on a larger fitting dataset.
- For full use as Scoring Function: Decoys (negative examples, usually measured as “> X μM ”) and fitting methods to work with mixed target data are needed. (in preparation)
- Various surface-based terms as descriptors in scoring functions are currently investigated (and promising).
- Estimate of experimental uncertainty in biological K_d measurements is necessary for serious interpretation: Current values in the CSARdock database (0.045 median of standard deviations given) are unrealistically small (for example, measurement pH 3 – 8.6 shift induces differences much larger than the uncertainties given).

Acknowledgments

- Peter Gedeck (Novartis)
- Novartis Institutes for BioMedical Research (NIBR) for a Presidential PostDoc Fellowship

Backups

Descriptor importance

Descriptor	Occurrence (out of 50)	Coefficient	Meaning
Ptype#.H1	50	+++	Number of H1 hydrogens (aliphatic) in the binding pocket
Weight	47	+++	Molecular Weight
LO3.PH3.6.0	46	--	Number of Ligand-O3 (ether oxygen) : Protein H3 (hydrogen connected to aromatic carbon) atom pairs between 4.5 and 6.0 Ångstroem
Ptype#.C18	37	+++	Number of C18 carbons (aromatic carbon carrying one hydrogen) in the binding pocket
LCl.PC23.4.0	36	++	Number of Ligand-Chlorine : Protein C23 (Tyrosine aromatic carbon to hydroxy Oxygen) atom pairs between 3.5 and 4.0 Ångstroem
b_rotN	36	---	Number of rotatable bonds in the ligand
LC25.PN10.3.5	29	++	Number of Ligand-C25 (aromatic carbon connected to C,N,O): Protein N10 (charged histidine guanidinium nitrogen) atom pairs between 3.0 and 3.5 Ångstroem
LC24.PN16.6.0	28	++	Number of Ligand-C24 (aromatic carbon connected to nonaromatic sulfur) : Protein N16 (tryptophan aromatic nitrogen) atom pairs between 4.5 and 6.0 Ångstroem
LN1.PH6.6.0	26	++	Number of Ligand-N1 (aliphatic uncharged nitrogen carrying at least one hydrogen) : Protein H6 (Tyrosine phenoxy hydrogen) atom pairs between 4.5 and 6.0 Ångstroem
...			

Largest Outliers

