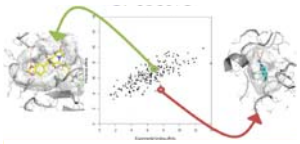
Strasbourg Summer School on Chemoinformatics
Strasbourg, June 23-27, 2014

Empirical scoring functions for docking and virtual screening

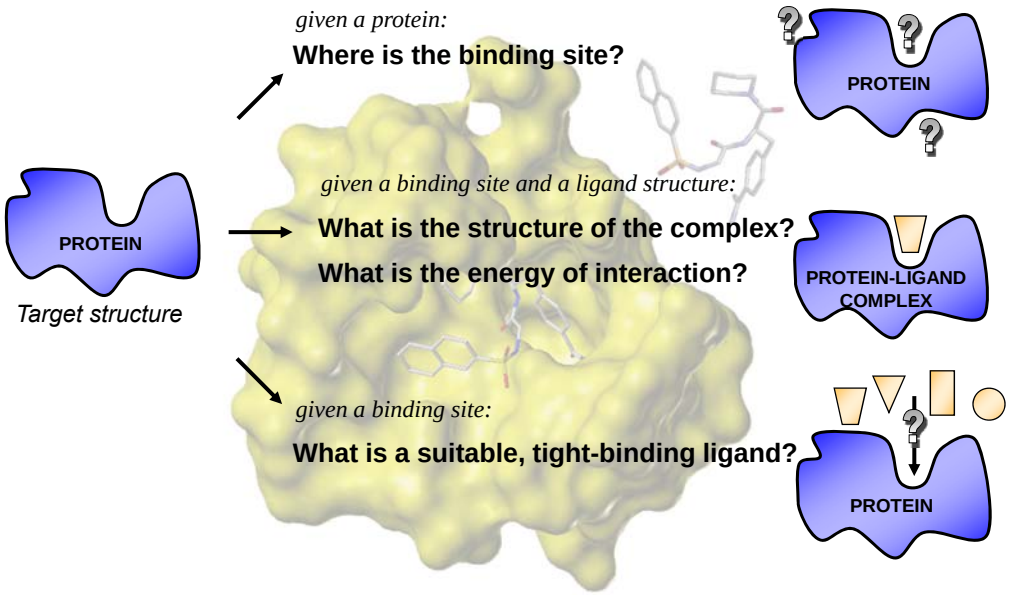
Fundamentals, challenges and trends

Christoph Sotriffer

Institute of Pharmacy and Food Chemistry
University of Würzburg
Am Hubland
D – 97074 Würzburg



Key questions in structure-based drug design



⇨ requires some sort of **affinity prediction** or **scoring**

Scoring functions: Tasks and types

Application tasks:

A) Determination of the correct binding mode for a given ligand

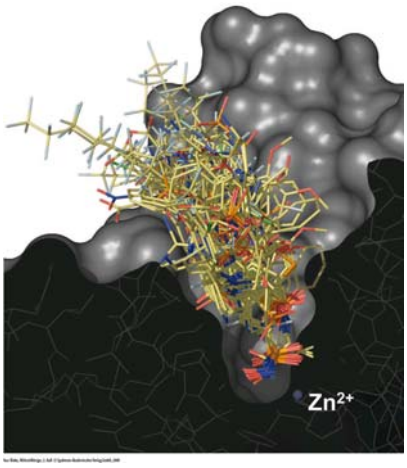
Pose prediction in docking

B) Identification and ranking of new ligands

Virtual screening

C) Affinity prediction for compound series

Ligand design, lead optimization



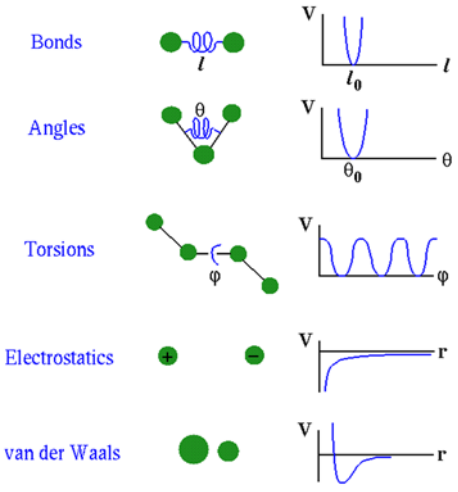
Available approaches:

- Force field-based methods
- Knowledge-based scoring functions
- **Empirical scoring functions**

Force field-based methods

Molecular Mechanics (MM):

- atoms → charged spheres
- bonds → springs
- classical potentials
- no electrons → no bond formation / cleavage
- typically parameterized to reproduce molecular potential energy surface (→ conformational ΔH in the gas phase!)



⇨ Scoring protein-ligand complexes:

\+ for pose prediction in docking

− for ligand ranking by affinity

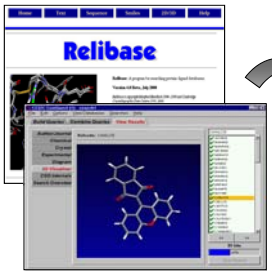
⇨ Terms accounting for (de)solvation & entropic factors required (cf. MM-PBSA)

Knowledge-based scoring functions

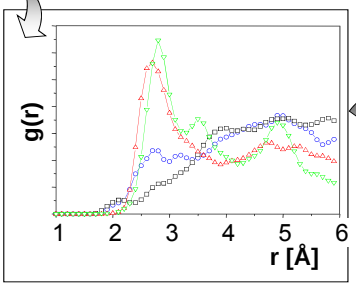
Derivation from crystal-structure data

$$P_{ij}(r) = -\ln \frac{g_{ij}(r)}{g_{ref}}$$

P_{ij}: distance-dependent pair potential
g_{ij}: frequency distribution of atom-atom contacts
g_{ref}: reference distribution

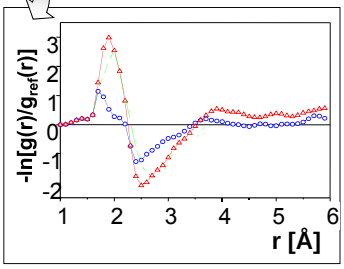


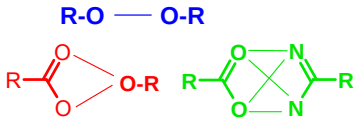
Frequency of occurrence



No experimental affinities used!

Statistical potential





Empirical scoring functions

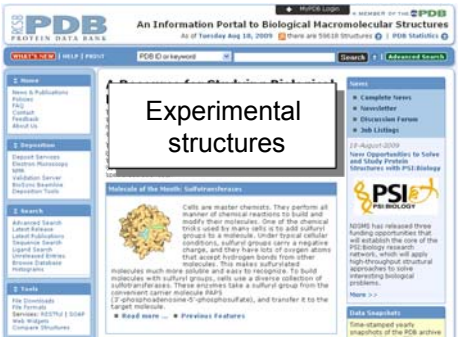
Regression-based:

$$pKi = \Sigma\, pKi_n\, f_n(\text{structure})$$

affinity | weighting factors | structure descriptors

determined via regression analysis (MLR, PLS)

Data:



The prototype: SCORE1 (Böhm, 1994)

Journal of Computer-Aided Molecular Design, 8 (1994) 243–256
ESCOM

J-CAMD 247

The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure

Hans-Joachim Böhm

BASF AG, Central Research, D-67056 Ludwigshafen, Germany

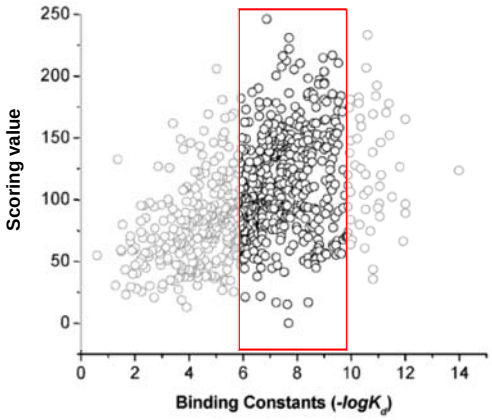
$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hb} \Sigma_{h\text{-bonds}} f(\Delta R,\Delta\alpha) + \Delta G_{ionic} \Sigma_{ionic\ int.} f(\Delta R,\Delta\alpha) + \Delta G_{lipo} |A_{lipo}| + \Delta G_{rot} NROT$$

SUMMARY

A new simple empirical function has been developed that estimates the free energy of binding for a given protein–ligand complex of known 3D structure. The function takes into account hydrogen bonds, ionic interactions, the lipophilic protein–ligand contact surface and the number of rotatable bonds in the ligand. The dataset for the calibration of the function consists of 45 protein–ligand complexes. The new energy function reproduces the binding constants (ranging from [illegible] corresponding to binding energies between –9 and –76 kJ/mol) of the dataset with a standard deviation of 7.9 kJ/mol, corresponding to 1.4 orders of magnitude in binding affinity. The individual contributions to protein–ligand binding obtained from the scoring function are: ideal neutral hydrogen bond: –4.7 kJ/mol; ideal ionic interaction: –8.3 kJ/mol; lipophilic contact: –0.17 kJ/mol Å^2; one rotatable bond in the ligand: +1.4 kJ/mol. The function also contains a constant contribution (+5.4 kJ/mol) which may be rationalized as loss of translational and rotational entropy. The function can be evaluated very fast and is therefore also suitable for application in a 3D database search or de novo ligand design program such as LUDI.

Affinity prediction on generic data sets

Scoring function performance **2004**
or: ***The „large-test-set“ shock*** *…*



Correlation with affinity for a test set of 800 known complexes:

for most functions
$r < 0.50$ ($r^2 < 0.25$)

Wang et al., *J. Chem. Inf. Comp. Sci.* 44 (2004), 2114

Affinity prediction on generic data sets

Scoring function performance **2004**
or: ***The „large-test-set“ shock*** *…*

Table 2. Correlation Evaluation of 14 Scoring Functions on the Entire Test Set[a]

scoring function	N^b	R_p	SD	ME	a	b
Cerius2::PLP2	*800*	*0.455*	*1.96*	*1.53*	*2.6×10^{-2}*	*3.93*
Cerius2::PMF	795	0.253	2.13	1.71	1.1×10^{-2}	5.37
Cerius2::LUDI1	790	0.334	2.08	1.66	2.6×10^{-3}	4.88
Cerius2::LUDI2	799	0.379	2.04	1.62	4.2×10^{-3}	4.28
Cerius2::LUDI3	800	0.331	2.08	1.67	3.2×10^{-3}	4.68
GOLD::GoldScore	694	0.285	2.16	1.72	2.4×10^{-2}	5.33
GOLD::GoldScore_opt	772	0.365	2.06	1.63	3.0×10^{-2}	4.70
GOLD::ChemScore	741	0.423	2.00	1.56	8.5×10^{-2}	4.65
GOLD::ChemScore_opt	762	0.449	1.96	1.52	8.6×10^{-2}	4.41
HINT	800	0.330	2.08	1.65	0.20	6.36

- poor correlation for generic data sets
- hardly possible to obtain correct ranking
- of limited use for ligand optimization

Correlation with affinity for a test set of 800 known complexes:

for most functions
$r < 0.50$ ($r^2 < 0.25$)

Wang et al., *J. Chem. Inf. Comp. Sci.* 44 (2004), 2114

How to improve empirical scoring functions?

Regression-based: $pKi = \Sigma\ pKi_n\ f_n(\text{structure})$

affinity — weighting factors — structure descriptors

determined via regression analysis (MLR, PLS)

Development options:

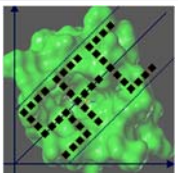
- **training sets**
- **descriptors**
- **regression methods**

The SFCscore approach

- Training sets: SFC: Scoring Function Consortium

  ⇒ Data collection from public & industry sources

  up to 855 complexes with affinity data



- Descriptors:

larger training set ⇕ additional descriptors

No.	Abbreviation	Description	
1	MW	Molecular weight	
2	NAtoms	Number of atoms	
3	NRotBonds	Number of rotatable bonds	
4	n_rot	Number of rotatable bonds (only sp3-sp3 and sp3-sp2)	rotatable bonds
5	RBScore	Rotatable bond score	
6	NHBonds	Number of H-bonds	
7	c_hb	Charged H-bond score	
8	n_hb	Neutral H-bond score	
9	HBScore	Total H-bond score	H-bonds and metal interactions
10	met	Boehm-type metal score	
11	AHPDI	Atom hydrophobicity difference	hydrophobic interactions
12	BURCP	Buried carbon percentage	
13	RRScore	Ring-ring interaction score	ring interactions
14	RMScore	Ring-metal interaction score	
15	RKScore	Ring-Lys (NZ) interaction score	
16	TotLigSurf	Total ligand surface area	
17	HydLigSurf	Hydrophobic ligand surface area	
18	PolLigSurf	Polar ligand surface area	surface areas
19	AroLigSurf	Aromatic ligand surface area	
20–23	[Tot,Hyd,Pol,Aro]BurSurf	[Total, hydrophobic, polar, aromatic] buried LSA	
24–27	[Tot,Hyd,Pol,Aro]ExpSurf	[Total, hydrophobic, polar, aromatic] exposed LSA	
28	BurdTotLigSurf	Ratio of buried to total LSA	
29–31	Bur2Tot[Hyd,Pol,Aro]Surf	Ratio of buried [hyd, pol, aro] to total [hyd, pol, aro] LSA	surface ratios I
32	Exp2BurLigSurf	Ratio of exposed to buried LSA	
33–35	Exp2Bur[Hyd,Pol,Aro]Surf	Ratio of exposed [hyd, pol, aro] to buried [hyd, pol, aro] LSA	
36–38	Bur[Hyd,Pol,Aro]2TotLigSurf	Ratio of buried [hyd, pol, aro] to total LSA	
39	BurHydAro2TotLigSurf	Ratio of buried hyd and aro surface to total LSA	
40–42	Exp[Hyd,Pol,Aro]2TotLigSurf	Ratio of exposed [hyd, pol, aro] to total LSA	
43	ExpHydAro2TotLigSurf	Ratio of exposed hyd and aro surface to total LSA	
44–52	SURFC-[H,PA]-[H,PA]	Contact surfaces for all 9 combinations of LSA and PSA types, E.g., SURFC-PP = contact surface between pol LSA and pol PSA	contact surfaces
53	HH_HA_AH_surfc	Sum of HH, HA, and AH contact surfaces	
54	HH_HA_AH_AA_surfc	Sum of HH, HA, AH, and AA contact surfaces	
55	PH_HP_surfc	Sum of PH and HP contact surfaces	
56	PH_HP_PA_AP_surfc	Sum of PH, HP, PA, and AP contact surfaces	
57	HX_surfc2HydBurSurf	Ratio of HX contact surface to buried hydrophobic surface	
58	PP_surfc2PolBurSurf	Ratio of PP contact surface to buried polar surface	
59	AA_surfc2AroBurSurf	Ratio of AA contact surface to buried aromatic surface	
60	HX_surfc2HydLigSurf	Ratio of HX contact surface to hydrophobic ligand surface	
61	PP_surfc2PolLigSurf	Ratio of PP contact surface to polar ligand surface	
62	AA_surfc2AroLigSurf	Ratio of AA contact surface to aromatic ligand surface	surface ratios II
63	HHAAPP_surfc2TotLigSurf	Fraction of the ligand surface which is ideally buried	
64	HHHAAHAA_surfc2HydAroLigSurf	Ratio of HH + HA + AH + AA contact surface to aro and hyd LSA	
65	PH_HP_surfc2TotLigSurf	Fraction of the ligand surface which is unfavourably buried	
66	Buried-PH_HP_surfc2TotLigSurf	Fraction of the ligand surface which is favourably buried (i.e, buried ligand surface minus PH_HP_surfc)	

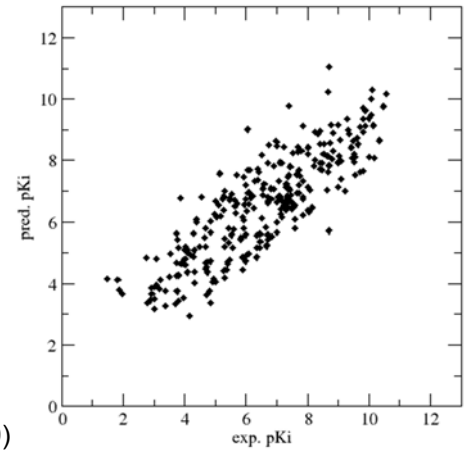
LSA, ligand surface area; PSA, protein surface area; tot, total; aro, aromatic; hyd, hydrophobic; pol, polar.

- Regression method: MLR + PLS

SFCscore

Example: SFCscore function „sfc_290m“

$pKi = - pKi_1 \times n\_rot\_bonds$
$+ pKi_2 \times neutral\_H\_bonds$
$+ pKi_3 \times metal\_interaction$
$+ pKi_4 \times AHPDI$
$+ pKi_5 \times ring\text{-}ring\_interaction$
$+ pKi_6 \times ring\text{-}metal\_interaction$
$+ pKi_7 \times total\_buried\_surface$
$+ pKi_8$



Statistical parameters for training set (n = 290)

R	R^2	s	F	Q^2	s_{PRESS}
0.843	0.711	1.09	99.2	0.692	1.12

Comparison with SCORE1 (n = 45):

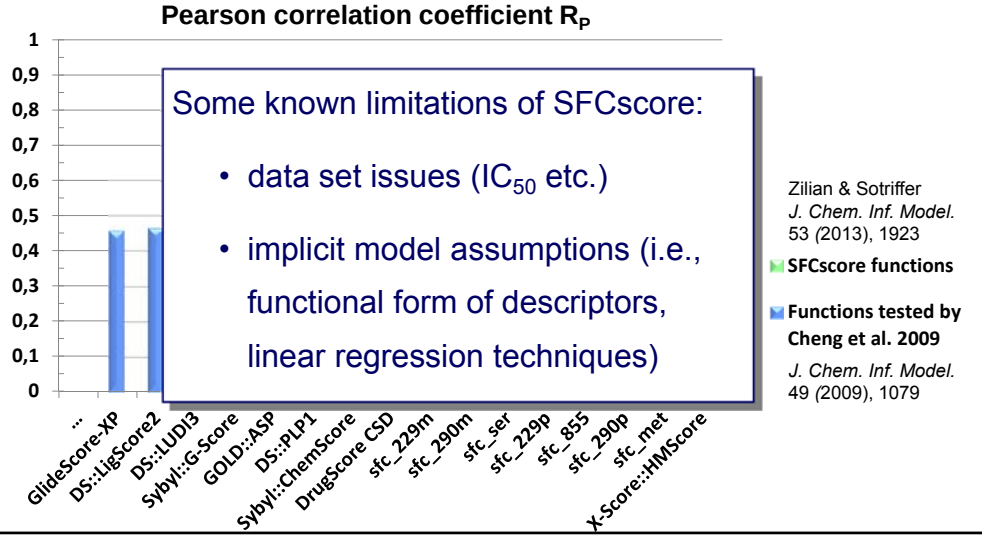
R	R^2	s	F	Q^2	s_{PRESS}
0.873	0.762	1.40	32.1	0.696	1.67

Sotriffer et al., *Proteins* 73 (2008), 395

2009 benchmark

Correlation of scores with experimental binding affinities

Test set compiled by Cheng et al., 2009: 195 PDBbind complexes (65 targets)



Addressing these limitations …

- Training sets:

  growth of PDBbind → 1005 complexes with K_i data

  (not overlapping with Cheng & CSAR test sets)

- Regression methods:

  Non-parametric machine-learning methods:

  (not imposing any particular functional form)

in particular : **Random Forest**

First scoring function trained with Random Forest:

RF-Score (Ballester & Mitchell, *Bioinformatics* 2010)

- Training set: 1105 PDBbind complexes
- Descriptors: count of protein-ligand atom type pair contacts withing 12 Å

  9 atom types (C, N, O, S, P, F, Cl, Br, I) → 36 pairs

  → each complex characterised by vector of 36 contact counts

⇨ RF-Score yields much higher R_p for Cheng test set!

BUT: *Do the pure contact counts sufficiently well capture the physicochemical interaction features?*

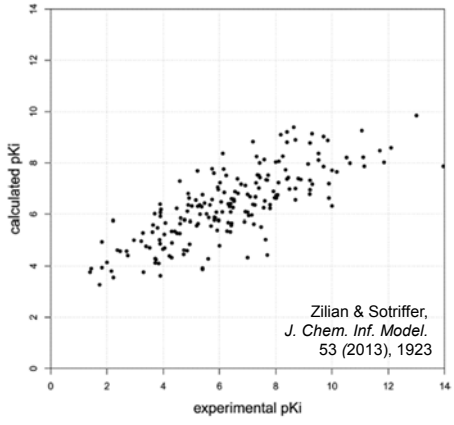
⇨ use SFCscore descriptors to train Random Forest model!

⇨ **SFCscore^{RF}**

- Training set: 1005 PDBbind complexes
- Descriptors: 63 SFCscore descriptors

Test set (Cheng)

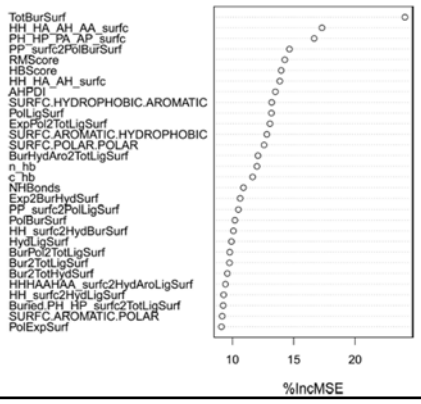
R_P = 0.779 RMSE = 1.56



Relative descriptor importance

Increase of the mean squared error

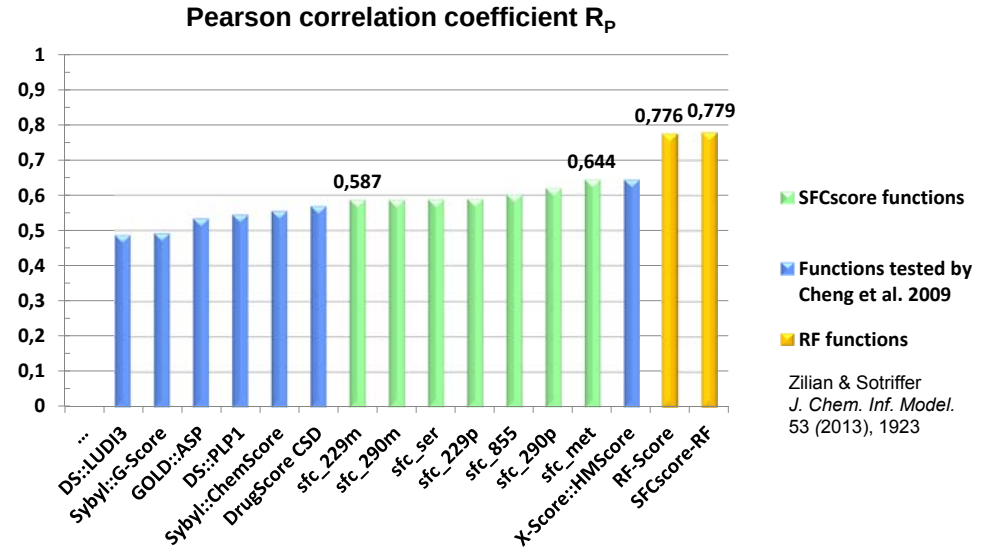
when randomly permuting the descriptor values



Scoring function performance

Correlation of scores with experimental binding affinities

Test set compiled by Cheng et al., 2009: 195 PDBbind complexes (65 targets)

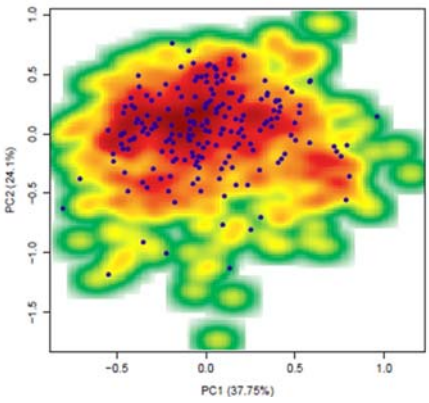


Zilian & Sotriffer
J. Chem. Inf. Model.
53 (2013), 1923

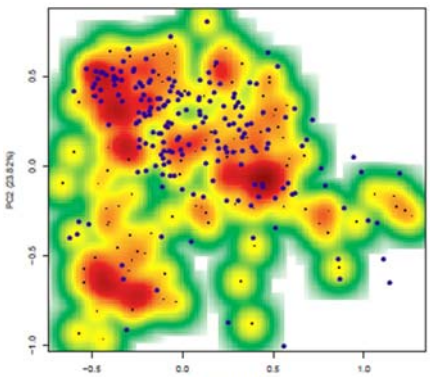
Applicability domain of SFCscore^{RF}

Why does SFCscore^{RF} outperform the other SFCscore functions?

SFCscore^{RF} training data



sfc_229m training data



● Cheng test set complexes

⇨ better coverage of training-set region

Knowing in advance the best SFCscore function for each individual complex would lead to

R_P = 0.93 RMSE = 1.03

One more generic test set: CSAR-NRC HiQ (2010)

Correlation of scores with experimental binding affinities

CSAR-NRC HiQ evaluation set: 332 complexes

Dunbar et al., *J. Chem. Inf. Model.* 51 (2011), 2036; Smith et al., *J. Chem. Inf. Model.* 51 (2011), 2115

Table 1. Parametric

method							
code 1							
code 2							
code 3							
code 4							
code 5							
code 6							
code 7							
code 8							
code 9							
code 10							
code 11							
code 12	0.57 (0.63−0.49)	0.57 (0.65−0.49)	0.41 (0.47−0.35)	0.32 (0.40−0.24)	1.82	2.18	1.28
code 13							
code 14							
code 15							
code 16							
code 17							
trained on 343 set[c]							
heavy atoms							
SlogP	0.46 (0.54−0.38)	0.50 (0.58−0.41)	0.34 (0.40−0.28)	0.22 (0.30−0.14)	1.95		

Performance across 17 core methods:

- R_P in the range 0.35 – 0.76 (only 3 >0.65)
- RMSE in the range 2.99 – 1.51 (pK_d units)
- correlation with heavy atom count: R_P 0.51

SFCscore^{RF}:

R_P = 0.73 RMSE = 1.53 (pK_d units)

Scoring function performance

One more generic test set: CSAR-NRC HiQ (2010)

Where are the limits?

Inherent experimental error

limits the possible correlation between scores and measured affinity.

R_P is limited to:

~0.91
when fitting to the data set without overparameterizing

~0.83
when scoring the data set with a method trained on outside data

(estimate based on error with σ = 1.0 log K)

Dunbar et al., *J. Chem. Inf. Model.* 51 (2011), 2146

SFCscore^{RF}:

R_P = 0.73 RMSE = 1.53 (pK_d units)

What about individual targets?

Leave-Cluster-Out (LCO) Validation: Target-dependent performance

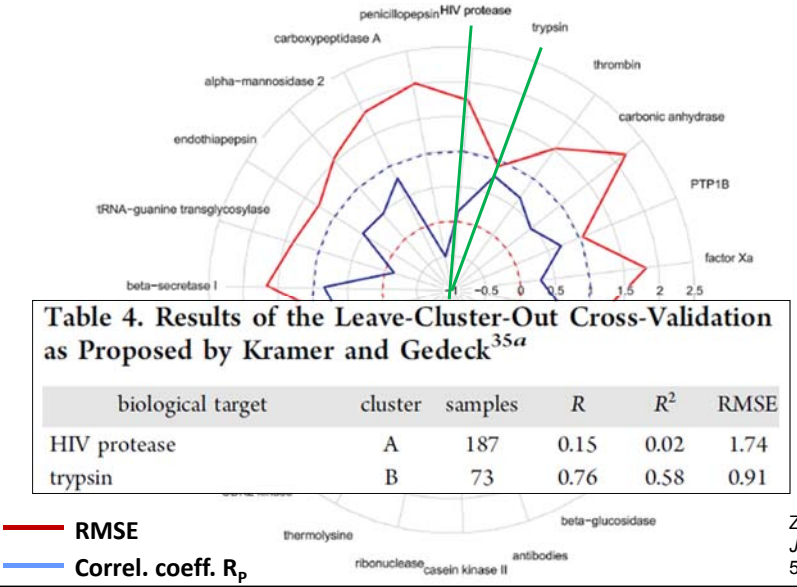


Table 4. Results of the Leave-Cluster-Out Cross-Validation as Proposed by Kramer and Gedeck[35a]

biological target	cluster	samples	R	R^2	RMSE
HIV protease	A	187	0.15	0.02	1.74
trypsin	B	73	0.76	0.58	0.91

RMSE

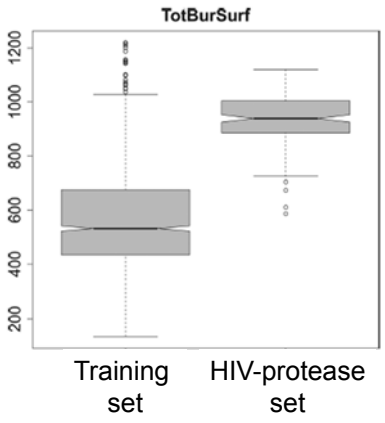
Correl. coeff. R_P

Zilian & Sotriffer
J. Chem. Inf. Model.
53 (2013), 1923

What about individual targets?

Leave-Cluster-Out (LCO) Validation: Target-dependent performance

BUT: Somewhat artificial setup …



Out-of-bag (OOB) predictions for HIV-protease class (n=97):

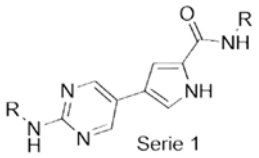
R_P = 0.60 **RMSE = 1.26**

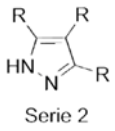
What about individual targets and docked ligands?

The CSAR 2012 challenge

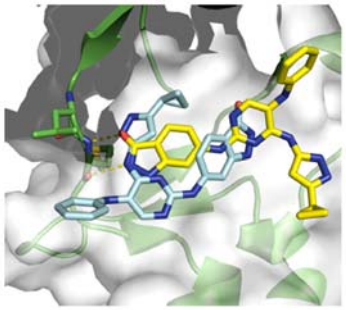
Example: **ERK2** test set

~40 compounds for docking and affinity ranking





⇨ rather poor results for most groups:
median R_p = 0.37 best: 0.66 SFCscore^{RF}: 0.49



Major problem:

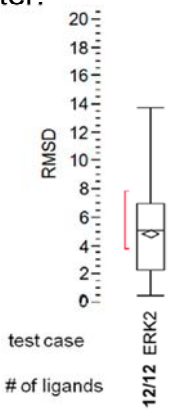
binding-mode prediction!

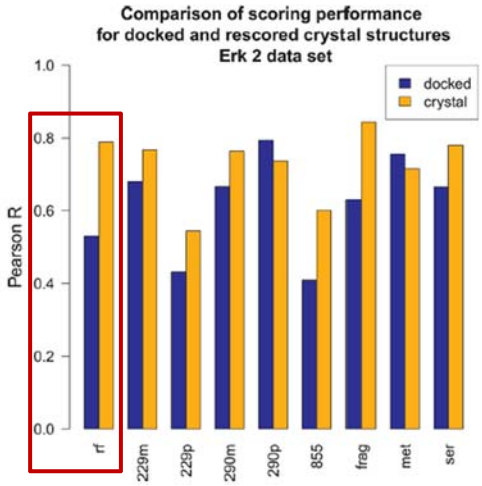
What about individual targets and docked ligands?

The CSAR 2012 challenge

Example: **ERK2** test set

Based on 12 crystal structures released later:





Damm-Ganamet et al., *J. Chem. Inf. Model.* 53 (2013), 1853

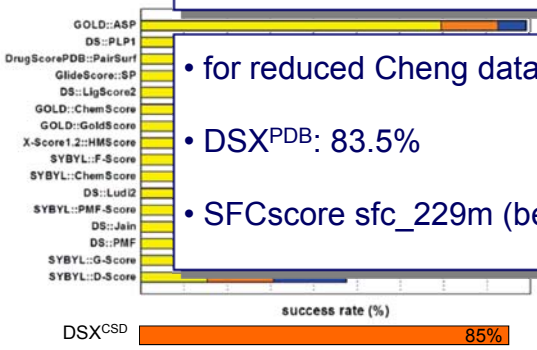
Scoring function performance (II)

Pose prediction in docking

Identifica
among a

- Test set of 1
- 100 low-ene
- 29 scoring f

- native poses can be detected fairly well
- success rates of up to ~80%
- knowledge-based approaches work best



- for reduced Cheng data set (n=176), rmsd 2 Å:
- DSX^{PDB}: 83.5%
- SFCscore sfc_229m (best): **38.1% !?!**

DSX^{CSD} 85%

Cheng et al., *J. Chem. Inf. Model.* 49 (2009), 1079

SFCscore for docking pose prediction

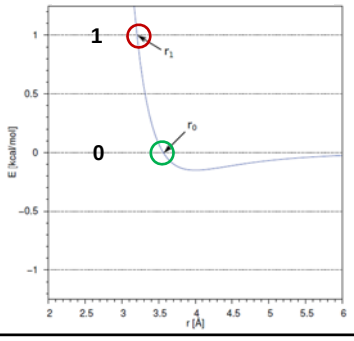
SFCscore functions: trained on crystal structures for affinity prediction

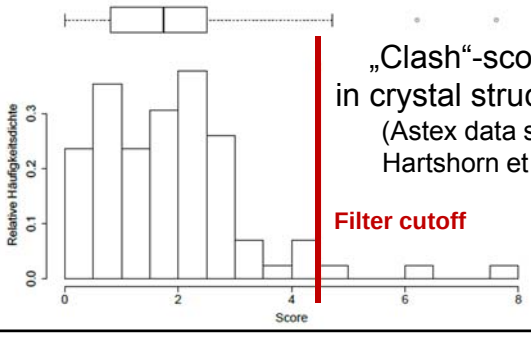
⇨ - insufficient information on unfavorable interactions
- no knowledge about decoy poses

In particular: penalties on bad contacts lacking

⇨ Using a simple „clash-descriptor“ as filter

Lennard-Jones potentials



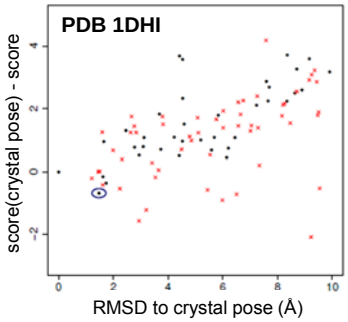


„Clash“-scores
in crystal structures
(Astex data set,
Hartshorn et al.)

Filter cutoff

Prefiltering poses with „clash-descriptor“ improves pose prediction with SFCscore

Scoring function	≤ 2.0 Å Pose (incl. crystal pose) Top-1-Pose	Top-2-Pose	Top-3-Pose
DSX^{PDB}::PAIR	81.2 (-2.3)	88.1 (-1.1)	89.2 (-3.4)
SFCscore::229m	67.0 (28.9)	75.6 (18.8)	81.2 (18.1)
SFCscore::229p	53.4 (26.1)	67.0 (25.5)	69.3 (19.9)
SFCscore::290m	66.5 (29.0)	75.0 (18.2)	82.4 (17.6)
SFCscore::290p	58.0 (27.9)	70.5 (27.3)	76.1 (22.1)
SFCscore::855	52.3 (27.9)	62.5 (26.1)	69.9 (23.3)
SFCscore::frag	69.9 (33.5)	83.0 (31.3)	86.9 (25.5)
SFCscore::met	54.5 (25.0)	66.5 (24.5)	73.3 (22.2)
SFCscore::ser	57.4 (29.0)	69.3 (26.1)	79.0 (29.6)



BUT: Success rates of DSX not reached

How to improve further?

Learning from decoy poses

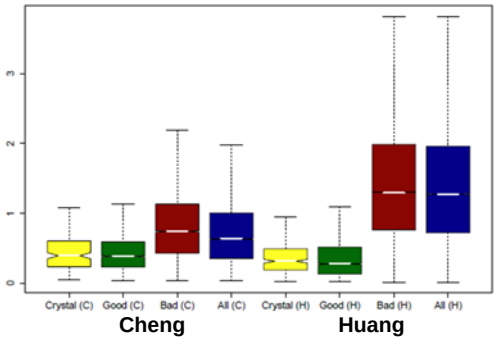
Data sets:

	Training		Test		
	Cheng	Huang	Cheng	Huang	CSAR-2012
C	120	-	56	318	58
H	-	120	176	198	58
C&H	60	60	374		58
C&H2	117	212	165		58

Huang: based on CSAR 2010
318 complexes (no overlap with Cheng)
500 poses/complex from Mdock & DOCK
0-18 Å RMSD (incl. native pose)

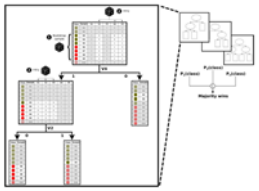
CSAR 2012:
58 complexes of 5 targets
199 decoy poses/complex from DOCK
(2-22 Å rmsd) + 1 near-native pose (<1 Å)

Exposed/buried ligand surface



SFCcsore for docking pose prediction

Random Forest classification model using SFCscore descriptors based on combined C&H2 training set



For pose prediction / docking power calc.:

Used in combination with DSX !

For each complex:
classification with RF-model

„near-native pose(s)“

if multiple poses:
rank with DSX, take top pose

„only decoys“

take top-ranked DSX pose

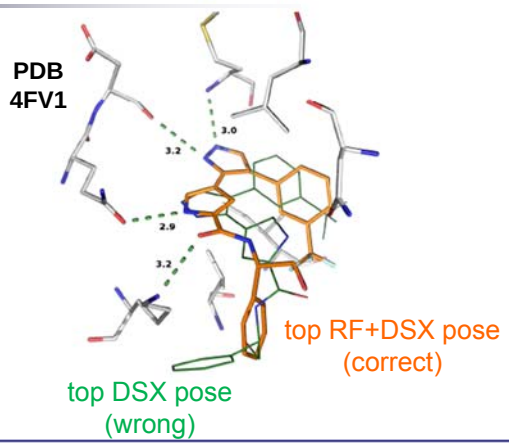
- Cheng/Huang test set (165 complexes):

  improving from 84.2% (DSX) → to 87.3% (RF+DSX)

- CSAR-2012 test set (58 complexes):

  improving from 87.9% (DSX) → to 91.4% (RF+DSX)

SFCcsore for docking pose prediction



- Cheng/Huang test set (165 complexes):

  improving from 84.2% (DSX) → to 87.3% (RF+DSX)

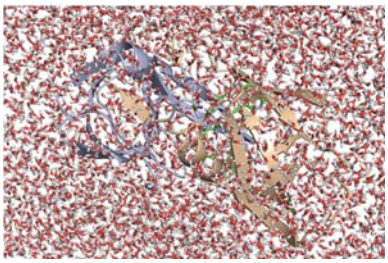
- CSAR-2012 test set (58 complexes):

  improving from 87.9% (DSX) → to 91.4% (RF+DSX)

Why is affinity prediction a challenge?

1.) Protein-ligand complexes are dynamic systems in aqueous solution

- huge number of particles
- simultaneous, unperiodic, continuously changing interactions



⇨ Simulation methods required!

Statistical thermodynamics: Calculation of ΔG° needs integration over entire phase space!

⇨ Computationally very expensive!

2.) The prediction methods need to be fast

Database screens: ~ 10^3 – 10^6 molecules need to be compared

Docking runs: ~ 10^7 – 10^9 configurations need to be evaluated

⇨ „Scoring functions“ required:
Fast, simplified, heuristic methods for prediction of binding strength

Fundamental limitations of empirical scoring functions (I)

- $\Delta G^0 = RT \ln K_D = \Delta H^0 - T\Delta S^0$

depending on the entire accessible phase space

referring to an equilibrium observable

difference between two states (bound/unbound)

yet scoring functions in general …
… consider only the complexed state
… consider only a single (or very few) configurations
… attempt to provide ΔG^0 also for arbitrary non-equilibrium states (poses)

⇨ Overall, the simplistic scoring functions work surprisingly well!

And: More sophisticated approaches start appearing …

e.g.: „Blurring“; Ucisik et al., *J. Chem. Theor. Comput.* 10 (2014), 1314
force-field based; ensemble generation, consideration of unbound state

Fundamental limitations of empirical scoring functions (II)

- Accuracy of experimental data!

  > Structural data (mainly X-ray) of protein-ligand complexes

  - multiple conformations (highly dynamic systems)
  - hydrogen atom positions (protonation states) not observable

The Experimental Uncertainty of Heterogeneous Public K_i Data

Christian Kramer,*,† Tuomo Kalliokoski,*,† Peter Gedeck, and Anna Vulpetti

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland

J. Med. Chem. 55 (2012), 5165

Exp. uncertainty in K_i for heterogeneous data: **MUE 0.44-0.48 pK_i units**

Upper limit of performance for all affinity prediction models!

max. performance of model with same uncertainty as exp. uncert.: **R_p = 0.81**

max. performance of a perfect model: **R_p = 0.90**

Acknowledgement

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

David Zilian

Michael Hein

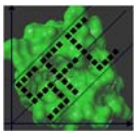
Manuel Krug

Benjamin Merget

Johannes Schiebel

Martin Sippel

Scoring
Function
Consortium



Astra Aventis
BASF Boehringer
Glaxo Novo Nordisk
Pfizer Agouron
Roche Schering CCDC

Hans Matter (Sanofi-Aventis)

Gerhard Klebe (Univ. of Marburg)
Paul Sanschagrin
Gerd Neudert

DFG (SFB 630, KFO 216)