3nd Strasbourg Summer School on Chemoinformatics
Strasbourg, June 25-29, 2012

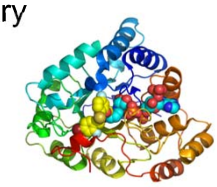
Scoring functions for

# of protein-ligand docking:

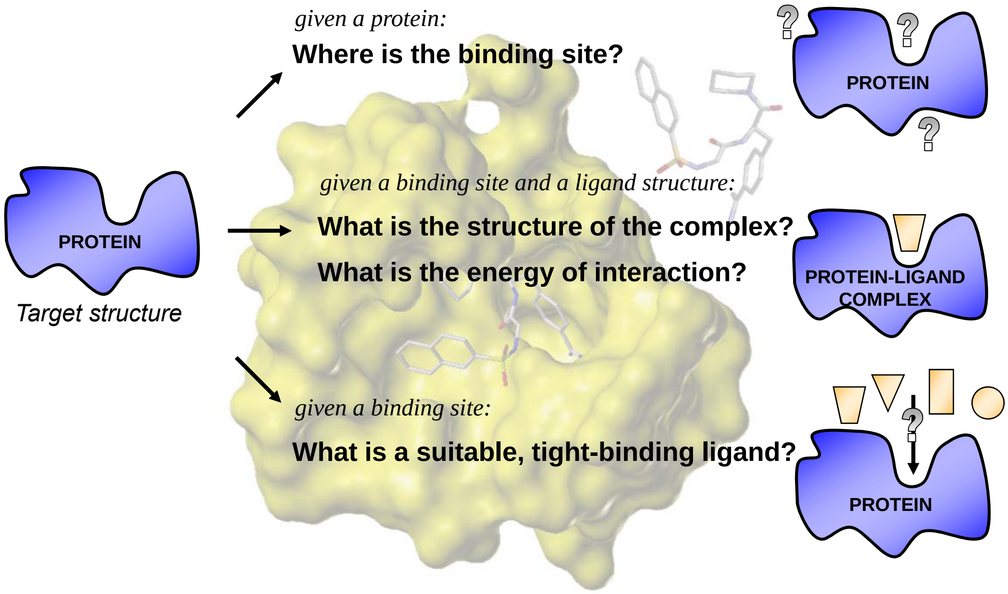
New routes towards old goals

Christoph Sotriffer

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



Key questions in structure-based drug design



given a protein:
Where is the binding site?

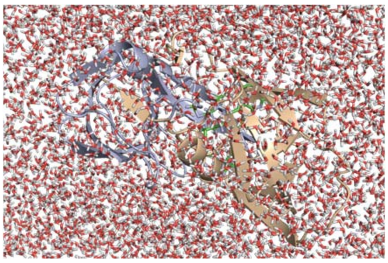
given a binding site and a ligand structure:
What is the structure of the complex?
What is the energy of interaction?

given a binding site:
What is a suitable, tight-binding ligand?

⇨ Required: some sort of **affinity prediction**

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

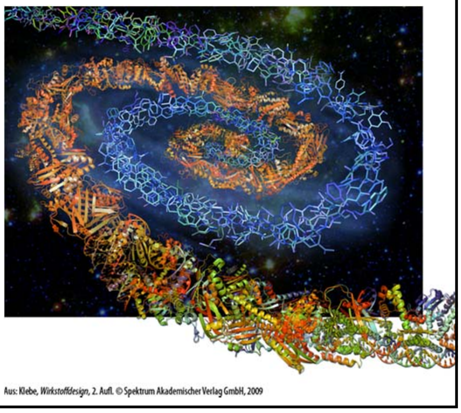
Docking runs: ~ 10^7 – 10^9 configurations need to be evaluated

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

Scoring functions: Goals

The ultimate goals of an ideal function:

- accurate within less than 1 pK_d unit (<1.4 kcal/mol)
- generally valid (not system specific; large affinity range)
- robust (tolerant with respect to small structural uncertainties)
- widely applicable (docking, virtual screening)
- physically meaningful (interpretable)
- fast and easy to compute



Aus: Klebe, *Wirkstoffdesign*, 2. Aufl. © Spektrum Akademischer Verlag GmbH, 2009

Scoring functions: Tasks and types

Application tasks:

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

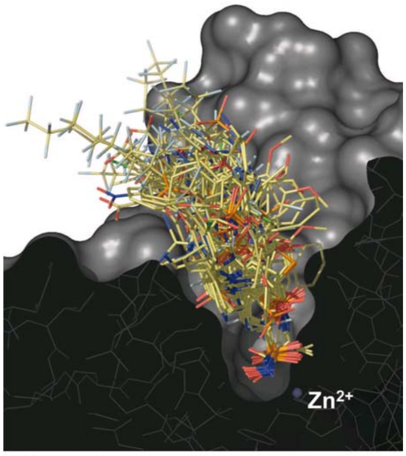
Pose prediction in docking

B) Identification of new active ligands

Virtual screening

C) Affinity ranking for compound series

Ligand design, lead optimization



Aus: Klebe, Wirkstoffdesign, 2. Aufl. © Spektrum Akademischer Verlag GmbH, 2009

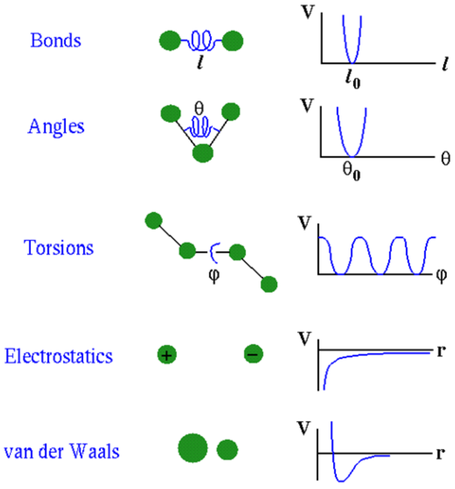
Available approaches:

- 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

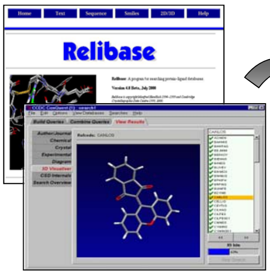
Derivation from crystal-structure data

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

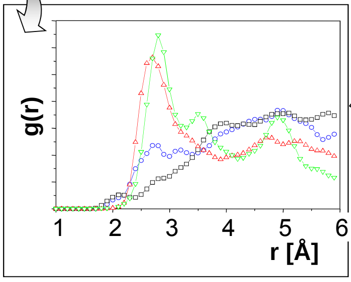
P_{ij}: distance-dependent pair potential

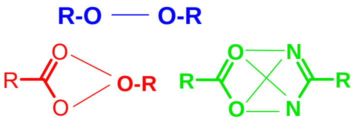
g_{ij}: frequency distribution of atom-atom contacts

g_{ref}: reference distribution



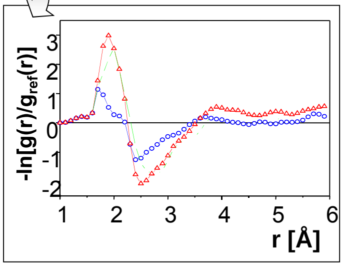
Frequency of occurrence





No experimental affinities used!

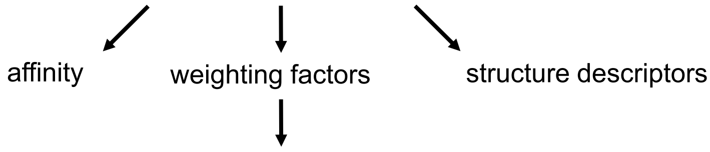
Statistical potential



Empirical scoring functions

Regression-based:

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



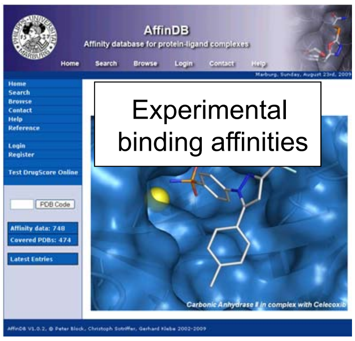
affinity

weighting factors

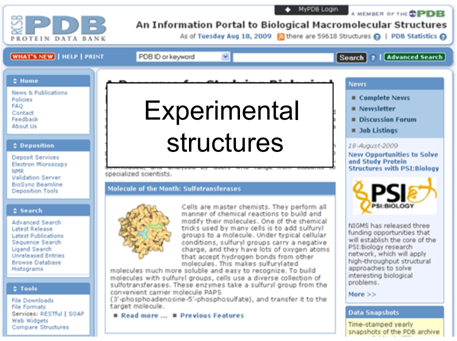
structure descriptors

determined via regression analysis (MLR, PLS)

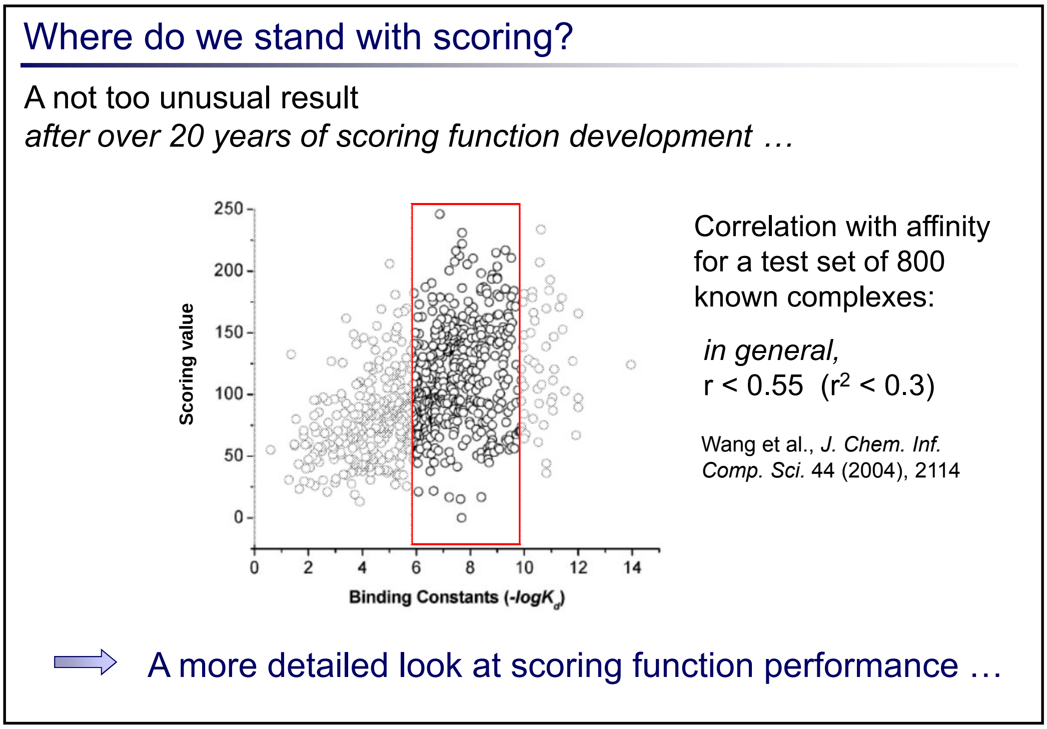
Data:



Experimental binding affinities


PDB
PROTEIN DATA BANK
An Information Portal to Biological Macromolecular Structures
Home
Deposition
Search
Tools
News
Molecule of the Month: Sulfotransferases
PSI

Experimental structures



Where do we stand with scoring?

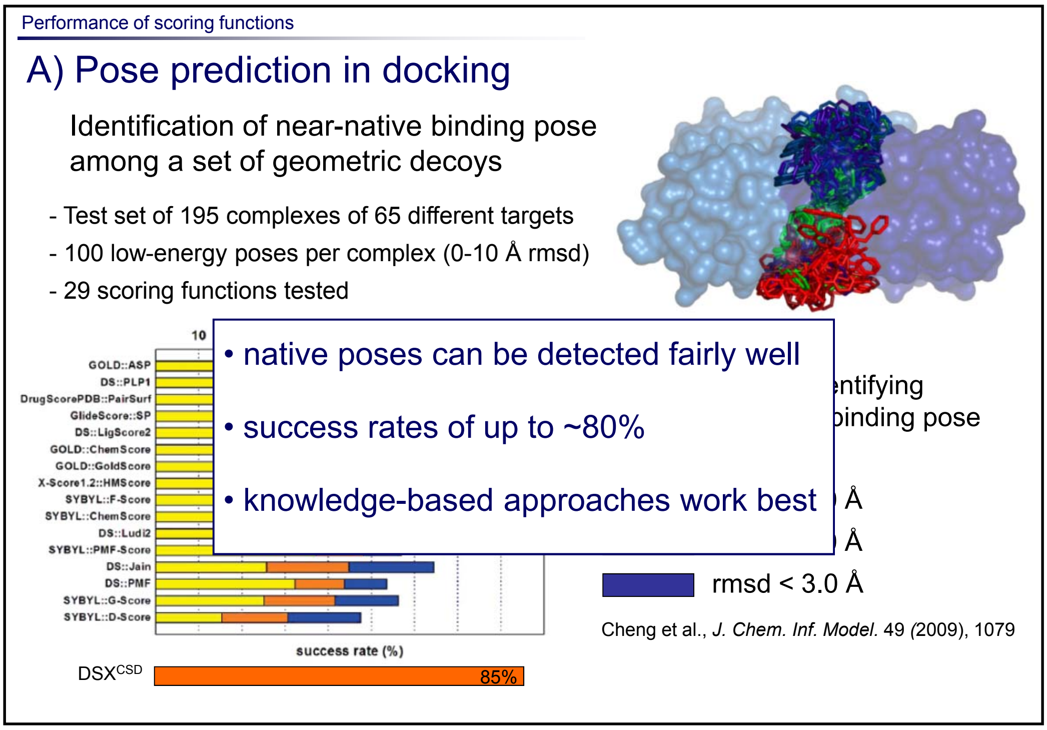
A not too unusual result
after over 20 years of scoring function development …

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

in general,
r < 0.55 ($r^2 < 0.3$)

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

→ A more detailed look at scoring function performance …



Performance of scoring functions

A) Pose prediction in docking

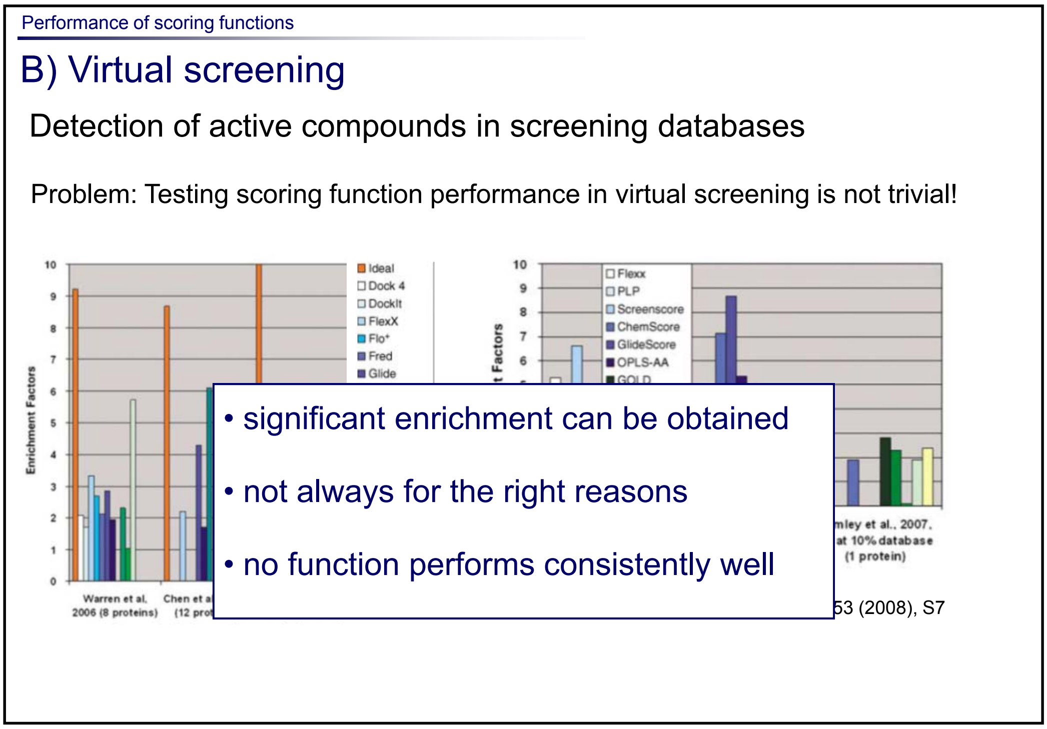
Identification of near-native binding pose among a set of geometric decoys

- Test set of 195 complexes of 65 different targets
- 100 low-energy poses per complex (0-10 Å rmsd)
- 29 scoring functions tested

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

▇ rmsd < 3.0 Å

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



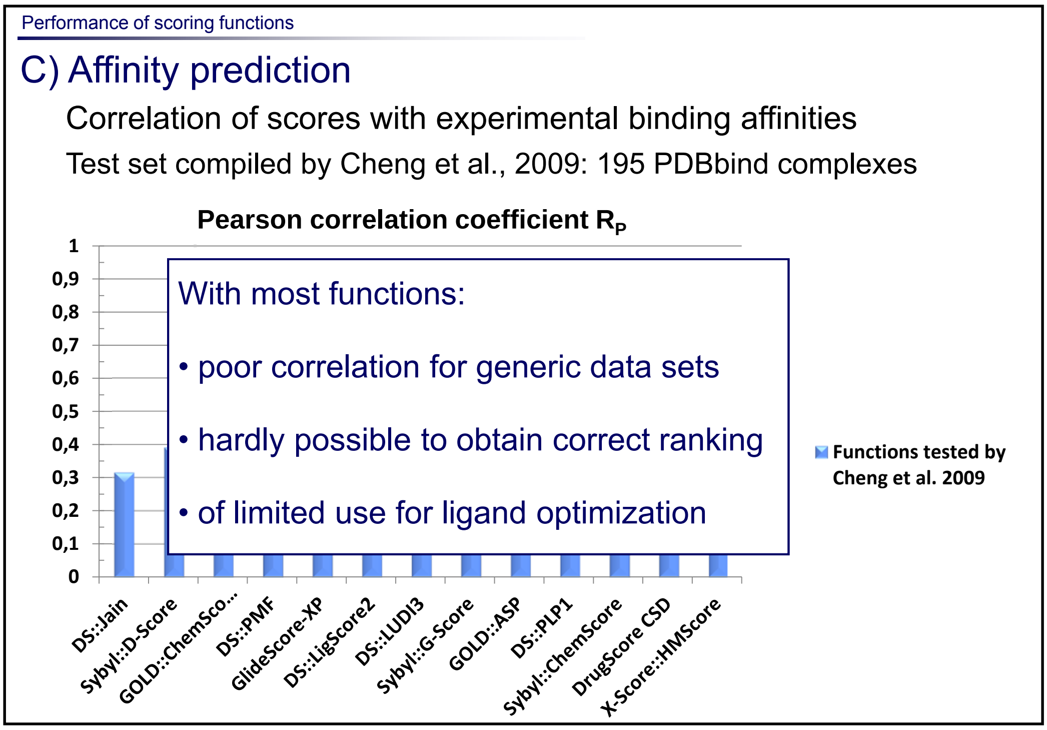
B) Virtual screening

Detection of active compounds in screening databases

Problem: Testing scoring function performance in virtual screening is not trivial!

- significant enrichment can be obtained
- not always for the right reasons
- no function performs consistently well

53 (2008), S7



C) Affinity prediction

Correlation of scores with experimental binding affinities

Test set compiled by Cheng et al., 2009: 195 PDBbind complexes

Pearson correlation coefficient R_P

With most functions:

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

C) Affinity prediction

Correlation of scores with experimental binding affinities

CSAR-NRC HiQ evaluation set: 343 (332) complexes

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

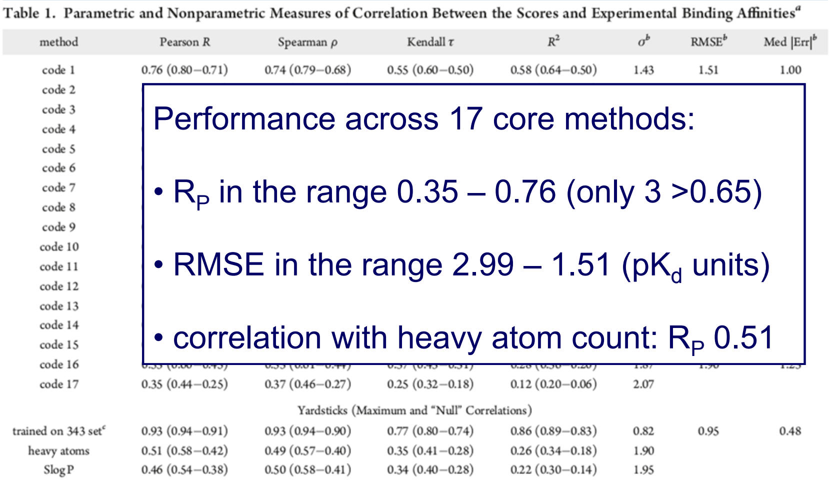


Table 1. Parametric and Nonparametric Measures of Correlation Between the Scores and Experimental Binding Affinities[a]

method	Pearson R	Spearman ρ	Kendall τ	R^2	σ[b]	RMSE[b]	Med |Err|[b]
code 1	0.76 (0.80–0.71)	0.74 (0.79–0.68)	0.55 (0.60–0.50)	0.58 (0.64–0.50)	1.43	1.51	1.00
code 2							
code 3							
code 4							
code 5							
code 6							
code 7							
code 8							
code 9							
code 10							
code 11							
code 12							
code 13							
code 14							
code 15							
code 16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
code 17	0.35 (0.44–0.25)	0.37 (0.46–0.27)	0.25 (0.32–0.18)	0.12 (0.20–0.06)	2.07		
Yardsticks (Maximum and "Null" Correlations)							
trained on 343 set[c]	0.93 (0.94–0.91)	0.93 (0.94–0.90)	0.77 (0.80–0.74)	0.86 (0.89–0.83)	0.82	0.95	0.48
heavy atoms	0.51 (0.58–0.42)	0.49 (0.57–0.40)	0.35 (0.41–0.28)	0.26 (0.34–0.18)	1.90		
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

How to improve current scoring functions?

Empirical scoring functions

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

affinity

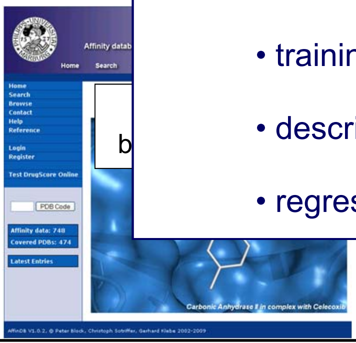
weighting factors

structure descriptors

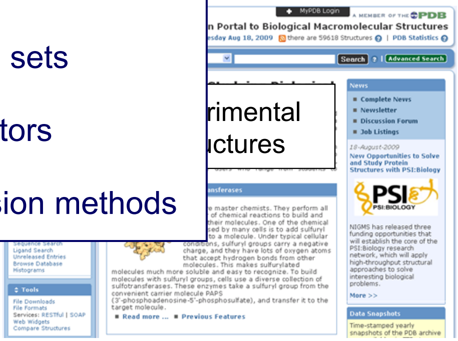
dete

PLS)

Data:



b

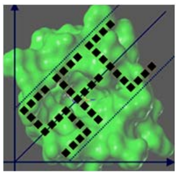


rimental
uctures

Development options:

- training sets
- descriptors
- regression methods

The SFCscore approach

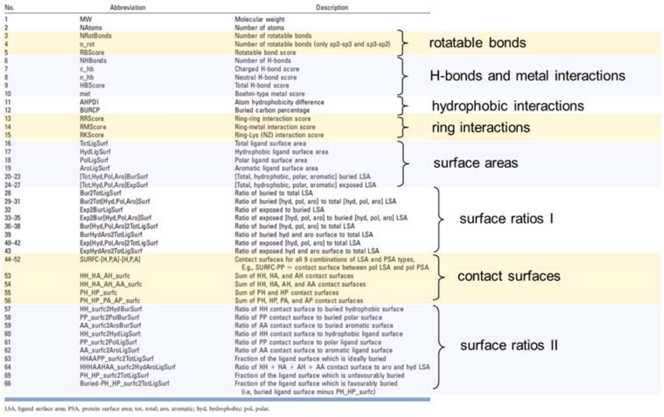


- Training sets: SFC: Scoring Function Consortium

⇨ Data collection from public & industry sources

up to 855 complexes with affinity data

- Descriptors:



No.	Abbreviation	Description	
1	MW	Molecular weight	
2	NAtoms	Number of atoms	
3	NRotBonds	Number of rotatable bonds	rotatable bonds
4	n_rot	Number of rotatable bonds (only sp3-sp3 and sp3-sp2)	
5	RBScore	Rotatable bond score	
6	NHBonds	Number of H-bonds	H-bonds and metal interactions
7	c_hb	Charged H-bond score	
8	n_hb	Neutral H-bond score	
9	HBScore	Total H-bond score	
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	surface areas
17	HydLigSurf	Hydrophobic ligand surface area	
18	PolLigSurf	Polar ligand surface area	
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	Bur2TotLigSurf	Ratio of buried to total LSA	surface ratios I
29–31	Bur2Tot[Hyd,Pol,Aro]Surf	Ratio of buried [hyd, pol, aro] to total [hyd, pol, aro] LSA	
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,P,A]-[H,P,A]	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	HH_surfc2HydBurSurf	Ratio of HH contact surface to buried hydrophobic surface	surface ratios II
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	HH_surfc2HydLigSurf	Ratio of HH 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	
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)	

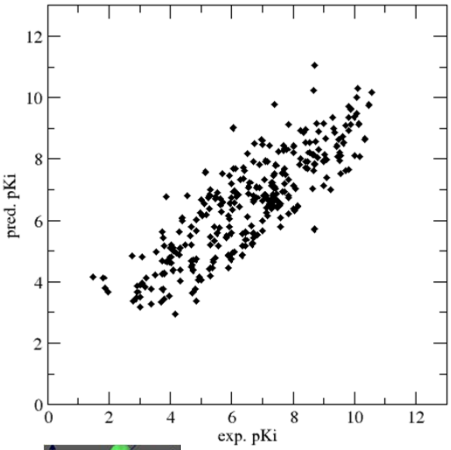
LSA, ligand surface area; PSA, protein surface area; tot, total; aro, aromatic; hyd, hydrophobic; pol, polar.

- Regression method: MLR + PLS

SFCscore

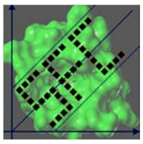
Example: SFCscore function „sfc_290m"

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



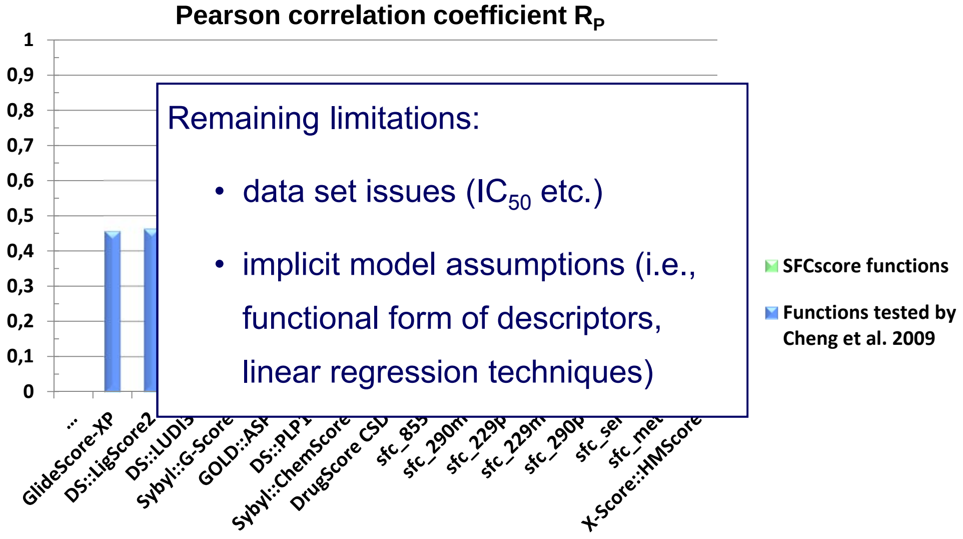
Statistical parameters for training set (n = 290).

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



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

Performance of SFCscore functions: Cheng test set (195 complexes)



Remaining limitations:

- data set issues (IC_{50} etc.)
- implicit model assumptions (i.e., functional form of descriptors, linear regression techniques)

Overcoming the limitations

- Training sets:

  growth of PDBbind → 1105 complexes with K_i data

  (not overlapping with Cheng test set)

- Regression methods:

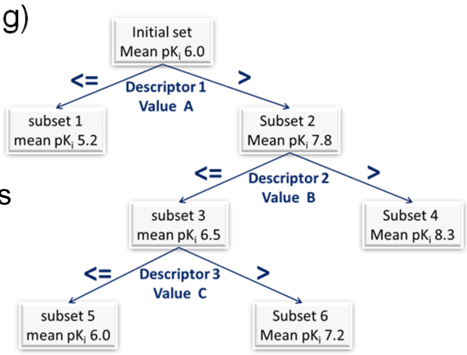
  Non-parametric machine-learning methods:

  (not imposing any particular functional form)

in particular : **Random Forest**

Random Forest

Decision Tree (or Recursive Partitioning)



Advantages:

- handles high-dimensional data well
- has ability to ignore irrelevant descriptors
- handles multiple mechanisms of action
- is amenable to model interpretation

Disadvantage:

- Relatively low prediction accuracy

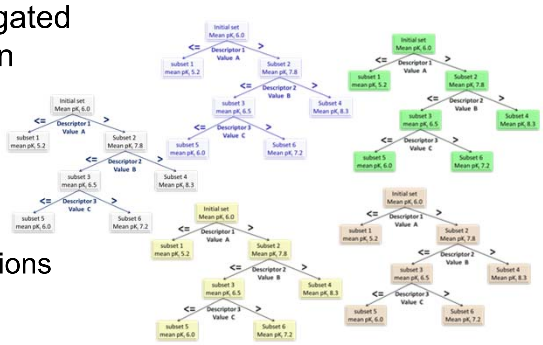
⇨ can be overcome by using ensembles of trees

⇨ one ensemble method: Random Forest (RF)

Svetnik et al., *JCICS* 43 (2003), 1947

Random Forest

RF: outputs of all trees are aggregated
to produce one final prediction



for **classification**:
class predicted by majority of trees

for **regression**:
average of the individual tree predictions

Training of a *Random* Forest:

1) Draw a random sample of the training data

2) For each sample, grow a tree to maximum size (no pruning) as follows:
at each node choose the best split among a randomly selected subset of m_{try} descriptors

3) Repeat the above steps until a sufficiently large number of trees are grown

Svetnik et al., *JCICS* 43 (2003), 1947

Random Forest for scoring functions

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?*

Random Forest for scoring functions: SFCscore^{RF}

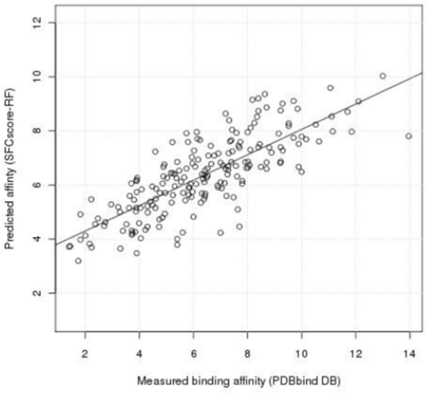
⇨ use SFCscore descriptors to train Random Forest model!

⇨ **SFCscore^{RF}**

- Training set: 1105 PDBbind complexes
- Descriptors: 63 SFCscore descriptors

Test set (Cheng)

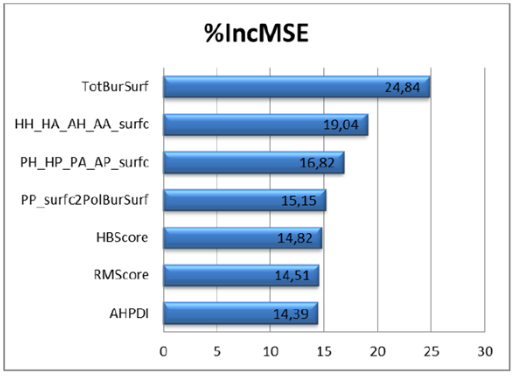
R_P = 0.787 RMSE = 1.53

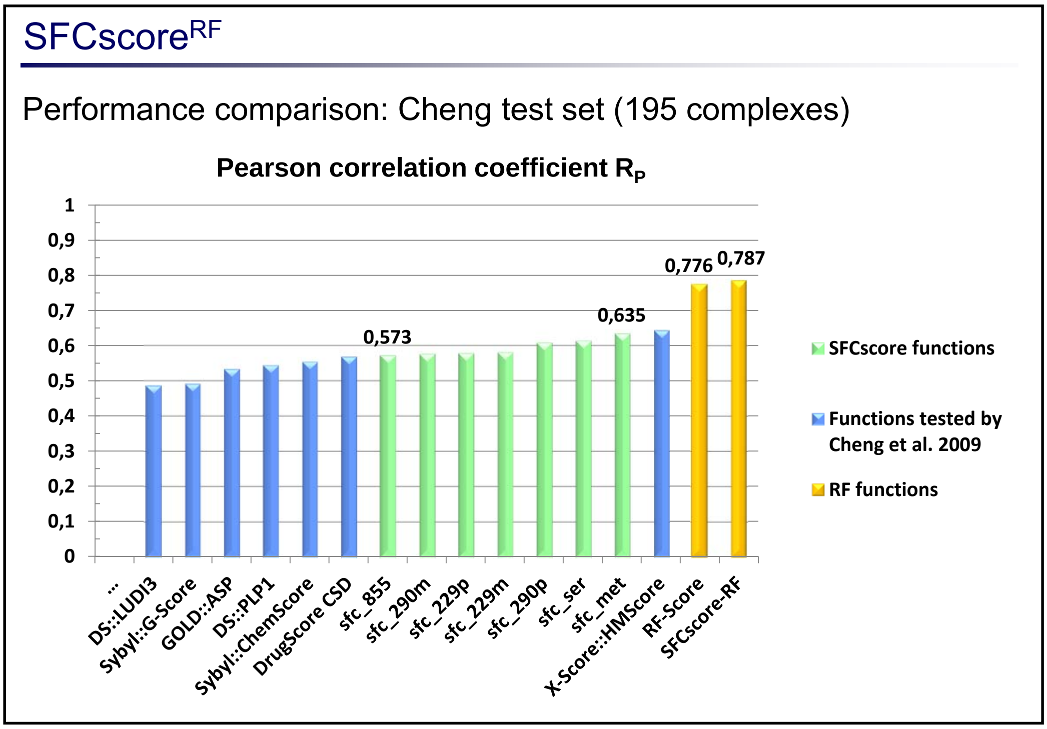


Relative descriptor importance

Increase of the mean squared error

when randomly permuting the descriptor values

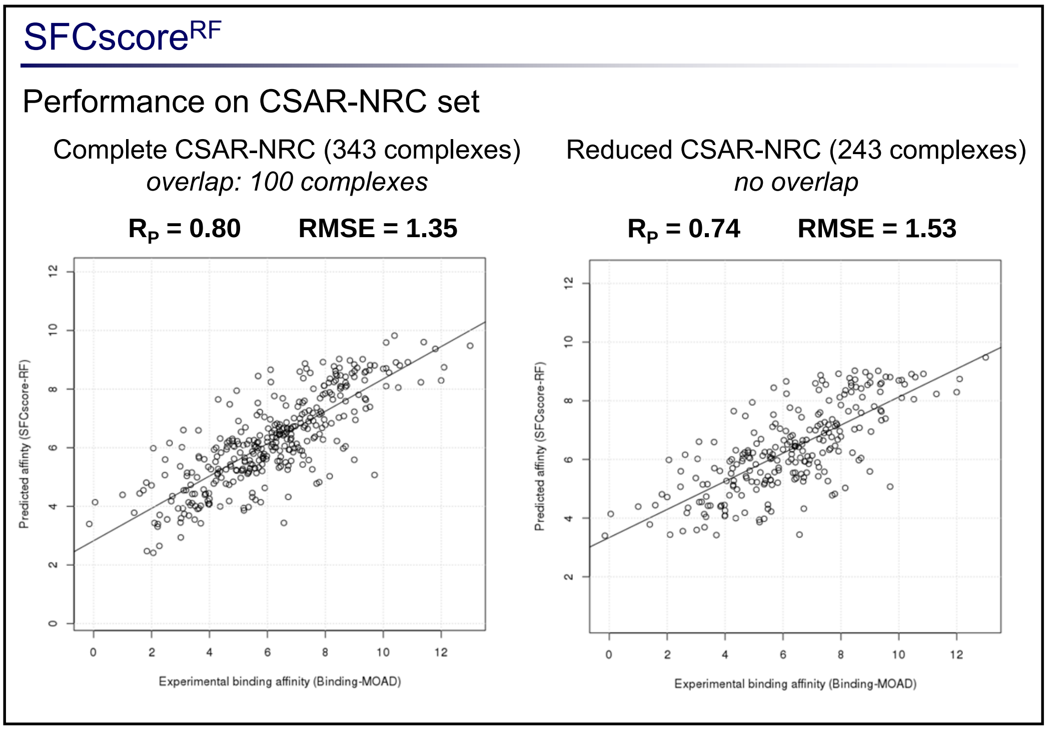




SFCscore^{RF}

Performance comparison: Cheng test set (195 complexes)

Pearson correlation coefficient R_P



SFCscore^{RF}

Performance on CSAR-NRC set

Complete CSAR-NRC (343 complexes)
overlap: 100 complexes

R_P = 0.80 RMSE = 1.35

Reduced CSAR-NRC (243 complexes)
no overlap

R_P = 0.74 RMSE = 1.53

SFCscore^{RF}

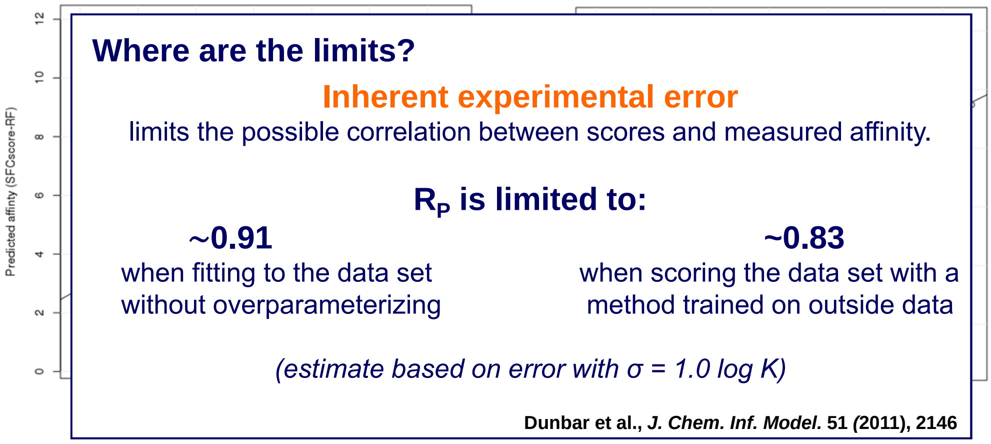
Performance on CSAR-NRC set

Complete CSAR-NRC (343 complexes)
overlap: 100 complexes

R_P = 0.80 RMSE = 1.35

Reduced CSAR-NRC (243 complexes)
no overlap

R_P = 0.74 RMSE = 1.53



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

Fundamental limitations of scoring functions (I)

- Accuracy of experimental data!

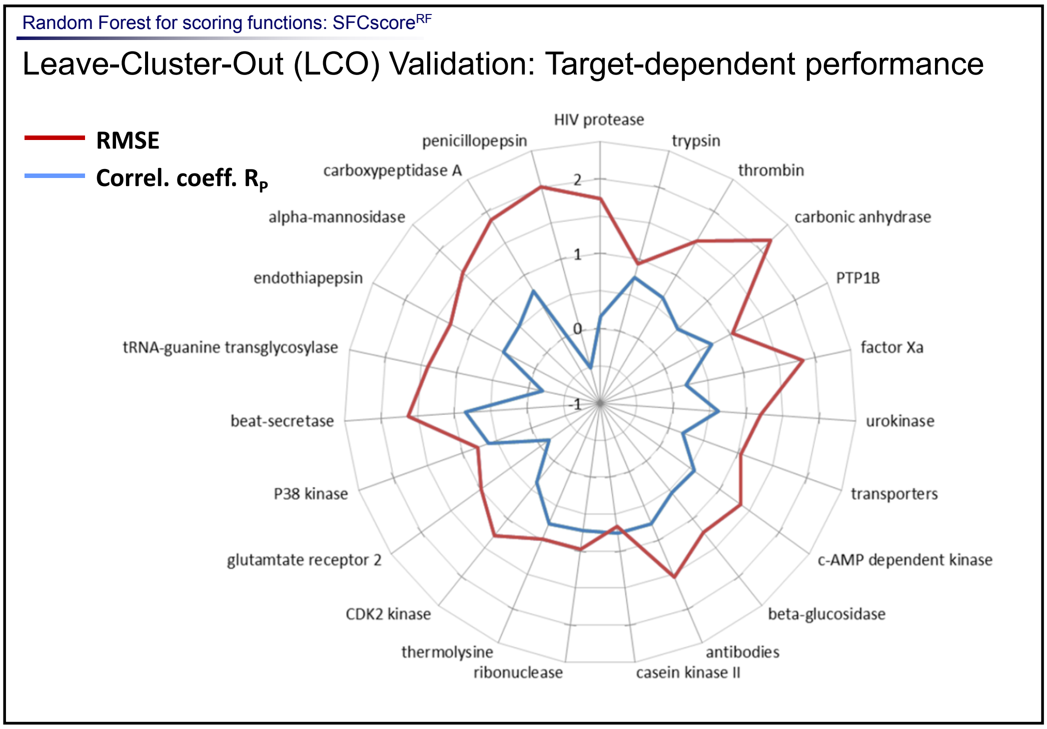
> Structural data (mainly X-ray) of protein-ligand complexes

- multiple conformations (highly dynamic systems)
- hydrogen atom positions (protonation states) not observable
- side-chain orientation may be ambiguous (Asn, Gln, His)
- water molecules are only partially observable
- binding modes may depend on crystallization conditions and crystal packing

> Affinity data of protein-ligand complexes

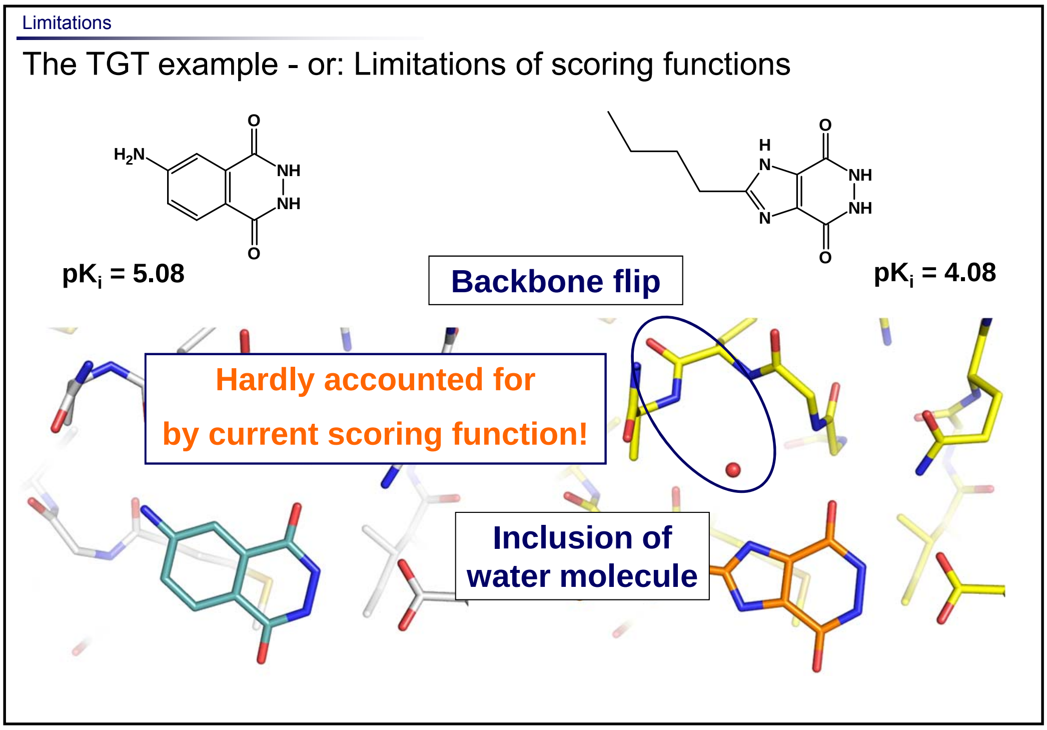
- **depend highly on pH, buffer, salt concentration, temperature**
- enyzme kinetics: inhibition mechanism must be known
- $IC_{50} \leftrightarrow K_i \leftrightarrow K_d$

Knowledge-based and empirical scoring methods cannot be better than the exp. data they are based on!



Random Forest for scoring functions: SFCscore^{RF}

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



Limitations

The TGT example - or: Limitations of scoring functions

pK$_i$ = 5.08

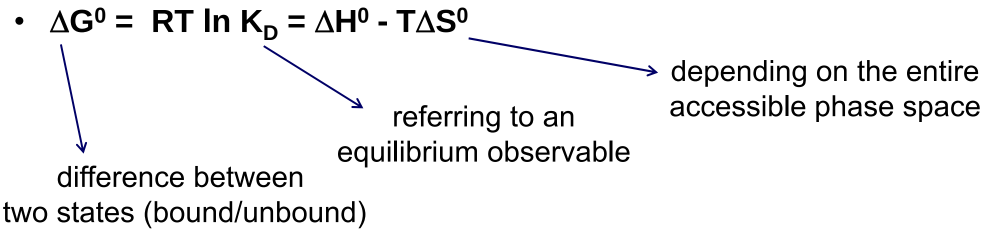
Backbone flip

pK$_i$ = 4.08

Hardly accounted for by current scoring function!

Inclusion of water molecule

Fundamental limitations of scoring functions (II)



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

difference between two states (bound/unbound)

referring to an equilibrium observable

depending on the entire accessible phase space

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)

„Dynamics – Water – Entropy“

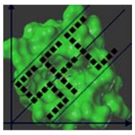
⇨ Overall, the simplistic scoring functions work surprisingly well!!

Acknowledgement

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

David Zilian
Daniel Cappel
Michael Hein
Manuel Krug
Monika Nocker
Ulrich Peinz
Benjamin Schaefer
Johannes Schiebel
Martin Sippel
Constanze Waltenberger
Armin Welker

Scoring
Function
Consortium

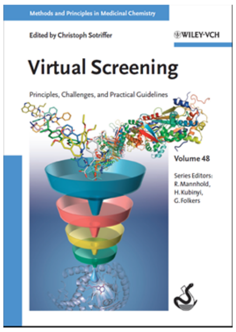


Astra	Aventis	
BASF	Boehringer	
Glaxo	Novo Nordisk	
Pfizer	Agouron	
Roche	Schering	CCDC

Hans Matter (Sanofi-Aventis)

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

DFG (SFB 630, KFO 216)


Methods and Principles in Medicinal Chemistry
Edited by Christoph Sotriffer
WILEY-VCH
Virtual Screening
Principles, Challenges, and Practical Guidelines
Volume 48
Series Editors: R. Mannhold, H. Kubinyi, G. Folkers