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? What is a suitable, tight-binding ligand?

Required: ligand placement and affinity prediction
Docking problems & scoring tasks

„Single Docking“: 1 protein – 1 ligand
looking for: binding mode (and affinity) of the ligand

„Virtual Screening“: 1 protein – many (potential) ligands
looking for: ligands with high affinity for target protein

„Selectivity“: many proteins – one or more ligands
looking for: ligands with high selectivity for one target

Why is docking a „problem“?
• complex 3D jigsaw puzzle
• conformational flexibility
• mutual adaptations („induced fit“)
• solvation in aqueous media
• complexity of thermodynamic contributions
• no easy route to $\Delta G$ evaluation for scoring

Simplifications und heuristic approaches necessary

*Modelling and computer-aided drug design are frequently a quest for suitable simplifications*
Approaches to solve the docking problem

"Representing" — "Searching" — "Evaluating"

- Surface representations
- Physicochemical descriptors
- Grid-based approaches
- Rule-based
- Energy-driven
- Force fields
- Regression functions
- Knowledge-based potentials

"Representing": Molecular representations for docking

A.) Protein

0. Restricting the search space to the binding pocket

1. Geometric surface descriptors
   e.g., sphere representation of binding pockets
   (→ program DOCK)
2. Physicochemical descriptors
   Interaction points and vectors
   \[ \rightarrow \text{Programs LUDI, FlexX} \]

3. Grid representations
   Interaction potentials of probe atoms are mapped to grid points
   \[ \rightarrow \text{Program AutoDock} \]
   \[ \Delta P_{g,t} = \sum_{p \in P} \Delta W_{t,T(p)} (r) \]
   \( g: \) grid point 
   \( t: \) ligand atom type (probe) 
   \( T(p): \) atom type of protein atom \( p \) 
   \( r: \) distance

B.) Ligand
   major problem: conformational flexibility
   \[ \rightarrow \text{strategies for flexible ligand docking} \]
   1.) rigid docking of conformers
   2.) simultaneous optimization of orientation and conformation
   3.) placement of a base fragment followed by incremental construction
Searching: Search algorithms for docking procedures

1.) Rule-based: geometric-combinatorial methods

- FlexX
- Incremental construction algorithm

![Rule-based geometric-combinatorial methods diagram]

2.) Energy-driven: stochastic optimization methods

- General assumption: experimentally determined complex structure corresponds to global minimum of $\Delta G_{\text{bind}}$
- Docking = optimization problem
- Search for $\min(\Delta G_{\text{bind}})$-binding mode
- $\Delta G_{\text{bind}}$ approximated by scoring function
- "Rugged", multi-dimensional energy landscape
- Monte-Carlo methods, genetic algorithms

Examples: AutoDock, ICM, GOLD
Before docking ...

... take care of the setup!

- Protein structures:
  - Protonation states and H-bonding networks
  - Quality and completeness of structural data
  - Location of binding site
  - Experimental data about water molecules and flexible regions

- Ligand structures:
  - Protonation states (influenced by protein!)
  - Tautomers
  - Conformers

- Docking program:
  - Choose suitable parameters
  - Validate, validate, validate (in particular for your system!)

- Know you program!

- Check structures and setup visually!

- Critically assess the quality of automated setup routines!

„Evaluating“ / Scoring: 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 $\Delta G^\circ$
needs integration over entire phase space!

   Computationally very expensive!

2.) The prediction methods need to be fast

Database screens: $\sim 10^3 - 10^6$ molecules need to be compared
Docking runs: $\sim 10^7 - 10^9$ 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\textsubscript{D} unit (<1.4 kcal/mol)
- generally valid (not system specific; large affinity range)
- robust (tolerant with respect to structural uncertainties)
- widely applicable (docking, virtual screening)
- physically meaningful (interpretable)
- fast and easy to compute

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

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

\[ P_i (r) = - \ln \frac{g_{ij} (r)}{g_{ref}} \]

\( P_i \): distance-dependent pair potential
\( g_{ij} \): frequency distribution of atom-atom contacts
\( g_{ref} \): reference distribution

No experimental affinities used!
Empirical scoring functions

Regression-based:

\[ p_{Ki} = \sum p_{Ki_n} f_n(\text{structure}) \]

- affinity
- weighting factors
- structure descriptors

determined via regression analysis (MLR, PLS)

Data:

- Experimental binding affinities
- Experimental structures

Where do we stand with docking & scoring?

A not too unusual result ...

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

\[ r < 0.55 \quad (r^2 < 0.3) \]


- So, what is possible and what is not?
I. Docking

Preface: „Comparing protein-ligand docking programs is difficult“
Cole et al., Proteins 60 (2005), 325

• Test sets needs to be carefully selected to
  - ensure sufficient diversity
  - provide good experimental reliability
  - avoid crystal packing effects
• Consider search complexity and timings
• RMSD values can be misleading
• Tests may cover different aspects, e.g.
  - redocking
  - crossdocking
  - blind docking
  - blind predictions

Comparative evaluations of docking programs

Compiled by Moitessier et al., Br. J. Pharmacol. 153 (2008), S7

• best approaches typically around 60%
• individual success rates up to 90%
• no approach consistently best
• highly target-dependent

Similar general conclusions by recent studies:
• Cross et al., J. Chem. Inf. Model. 49 (2009), 1455
  (68 complexes; DOCK, FlexX, Glide, ICM, PhDOCK, Surflex)
• Li et al., J. Comput. Chem. 31 (2010), 2109
  (195 complexes; Glide, GOLD, LigandFit, Surflex)
1. Docking

A critical issue: Conformational flexibility!

- Complex reconstruction from **rigid binding partners**: Essentially a solved problem!
  - e.g.: RosettaLigand, 85 complexes (Astex diverse): 99% success rate; av. RMSD <1Å
    
    Davis et al., *J.Mol.Biol.* 385 (2009), 381

- **Flexible ligand** – rigid protein docking: Standard, but not without problems
  - docking success rate drops for more flexible ligands (>7-8 rotatable bonds)
  - danger of insufficient sampling (correct conformation and pose is not generated)

- **Flexible ligand** – **flexible protein** docking: Active field of development

  Modeling of protein flexibility:
  - **before** ligand placement (e.g., ensemble docking)
  - **after** ligand placement (e.g., complex refinement)
  - **during** ligand placement (e.g., MC/MD techniques)

  But: even „simple“ conformational changes can be out of reach!

---

Example 1: TGT - Successful docking and ...

- X-ray confirmed a perfect binding-mode-prediction for a new virtual-screening hit!
I. Docking: What is possible and what is not?

... surprises out of reach for any docking program

Virtually impossible to predict
with current protein-flexibility
docking approaches
(unless alternative conformation
is experimentally known in advance)

Backbone flip at Leu231
and water molecule
mediate formation of new
H-bond interaction!
Example 2: Aldose Reductase - docking to multiple pocket conformers

Sorbinil pocket

Tolrestat pocket

IDD594 pocket

“In-situ” Cross-Docking of new pyridazinone inhibitor to multiple pocket conformations


Example 2: Aldose Reductase - docking to multiple pocket conformers

binder to IDD594 pocket!

Docking vs. X-ray
RMSD = 0.49 Å

Steuber et al., J. Mol. Biol. 356 (2006), 45
Example 2: Aldose Reductase - docking to multiple pocket conformers

Reason for successful prediction:

- ligand binds to protein conformer known from previous X-ray structures
- scoring function correctly scores the true binding mode much better than binding modes in alternative protein conformers

Docking vs. X-ray
RMSD = 0.49 Å
despite protein flexibility:
„easy task“ for common docking tools

Steuber et al., J. Mol. Biol. 356 (2006), 45

Design of new inhibitors: Tolrestat analogues

Do the new compounds adopt the same binding mode as tolrestat?

Da Settimo et al., J. Med. Chem. 48 (2005), 6897.
Naphtho[1,2-d]isothiazole acetic acid derivatives as a novel class of selective aldose reductase inhibitors.
Docking of \( \textbf{1} \) to three different binding pocket conformers, using AutoDock

I. Docking – Example 2: Aldose Reductase

Preferred binding mode of \( \textbf{1} \):

- sorbinil-like, not tolrestat-like
- closed specificity pocket
- 4-COO\(^-\) binds to catalytic site (!)

Zentgraf et al., Angew. Chem. Int. Ed. 46 (2007), 3575
Docking result of 1 in comparison with crystal structure

- specificity pocket closed
- 4-COO⁻ in catalytic site

**But:**

Unexpected conformational changes!

- Trp 20 rotated by 35°
- Lys 21 salt bridge broken
- Trp 219 disordered

Unpredictable with docking methods! (incl. FlexX, GOLD, Glide)

Docking of 2 to three different binding pocket conformers, using AutoDock

-12.0 kcal/mol
-10.4 kcal/mol
-11.4 kcal/mol

sorbinil
tolrestat
idd 594
Docking of 2 to three different binding pocket conformers, using AutoDock

Preferred binding mode of 2:
- sorbinil-like, not tolrestat-like
- closed specificity pocket
- \(2\text{-COO}^-\) binds to catalytic site

Zentgraf et al., Angew. Chem. Int. Ed. 46 (2007), 3575

Docking result of 2 in comparison with crystal structure

- specificity pocket closed
- \(2\text{-COO}^-\) in catalytic site
- no conformational changes!

**But:**

Water molecules immobilized in binding pocket!

3 very „similar“ ligands lead to
3 very different binding modes!
AutoDock results obtained when using the “correct” binding-site conformer

1. Docking – Example 2: Aldose Reductase

*Binding mode exactly reproduced in both cases!*
II. Scoring

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*

Available approaches:

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

Success rate for identifying best-scored ligand binding pose with

- rmsd < 1.0 Å
- rmsd < 2.0 Å
- rmsd < 3.0 Å

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

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

![Graph showing enrichment factors for different scoring functions with annotations]

II. Scoring

C) Affinity prediction
Correlation of scores with experimental binding affinities and ranking of compounds

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

![Graph showing Spearman correlation coefficients with annotations]
II. Scoring: What is possible and what is not?

*Since all methods are of empirical nature:*

Do more and „better“ experimental data lead to better functions?

<table>
<thead>
<tr>
<th>SFCscore empirical scoring functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SFC: Scoring Function Consortium</strong></td>
</tr>
<tr>
<td>Data collection from public &amp; industry sources</td>
</tr>
<tr>
<td>• affinity data from literature for PDB complexes</td>
</tr>
<tr>
<td>• „diversity“ from PDB, SAR series from industry</td>
</tr>
<tr>
<td>• unique data format and encoding for industry data</td>
</tr>
<tr>
<td>• up to 58 complexes per target, 28 series, mostly IC(_{50}) (!)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw data in total (public + industrial):</th>
</tr>
</thead>
<tbody>
<tr>
<td>complexes from PDB: 440 filtered: 290</td>
</tr>
<tr>
<td>complexes from industry: 618 filtered: 565</td>
</tr>
<tr>
<td>total: 1058 855</td>
</tr>
</tbody>
</table>
### II. Scoring: What is possible and what is not?

**SFCscore Training sets: Regression statistics**

<table>
<thead>
<tr>
<th>Function</th>
<th>Method</th>
<th>$N$</th>
<th>$k$</th>
<th>$r$</th>
<th>$r^2$</th>
<th>$s$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sfc_290m</td>
<td>MLR</td>
<td>290</td>
<td>7</td>
<td>0.843</td>
<td>0.711</td>
<td>1.085</td>
<td>99.2</td>
</tr>
<tr>
<td>sfc_229m</td>
<td>MLR</td>
<td>229</td>
<td>7</td>
<td>0.842</td>
<td>0.709</td>
<td>1.098</td>
<td>76.9</td>
</tr>
<tr>
<td>sfc_frag</td>
<td>MLR</td>
<td>130</td>
<td>4</td>
<td>0.810</td>
<td>0.656</td>
<td>0.973</td>
<td>58.8</td>
</tr>
<tr>
<td>sfc_855</td>
<td>PLS</td>
<td>855</td>
<td>6</td>
<td>0.770</td>
<td>0.593</td>
<td>0.994</td>
<td>205.9</td>
</tr>
<tr>
<td>sfc_ser</td>
<td>PLS</td>
<td>466</td>
<td>4</td>
<td>0.843</td>
<td>0.711</td>
<td>0.952</td>
<td>284.0</td>
</tr>
<tr>
<td>sfc_met</td>
<td>PLS</td>
<td>341</td>
<td>4</td>
<td>0.844</td>
<td>0.713</td>
<td>1.046</td>
<td>208.9</td>
</tr>
<tr>
<td>sfc_290p</td>
<td>PLS</td>
<td>290</td>
<td>5</td>
<td>0.867</td>
<td>0.751</td>
<td>1.005</td>
<td>171.3</td>
</tr>
<tr>
<td>sfc_229p</td>
<td>PLS</td>
<td>229</td>
<td>6</td>
<td>0.875</td>
<td>0.766</td>
<td>0.982</td>
<td>121.2</td>
</tr>
</tbody>
</table>

$N$, number of complexes in the training set; $k$, number of components for PLS functions, number of variables for MLR functions; $r$ and $r^2$, correlation coefficient and its square; $s$, standard error; $F$, $F$-value.

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

---

**SFCscore Training sets: Internal cross validation**

<table>
<thead>
<tr>
<th>Function</th>
<th>$Q^2$</th>
<th>$s_{PRESS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sfc_290m</td>
<td>0.692</td>
<td>1.121</td>
</tr>
<tr>
<td>sfc_229m</td>
<td>0.683</td>
<td>1.147</td>
</tr>
<tr>
<td>sfc_frag</td>
<td>0.627</td>
<td>1.015</td>
</tr>
<tr>
<td>sfc_855</td>
<td>0.572</td>
<td>1.033</td>
</tr>
<tr>
<td>sfc_ser</td>
<td>0.692</td>
<td>1.028</td>
</tr>
<tr>
<td>sfc_met</td>
<td>0.688</td>
<td>1.135</td>
</tr>
<tr>
<td>sfc_290p</td>
<td>0.722</td>
<td>1.080</td>
</tr>
<tr>
<td>sfc_229p</td>
<td>0.723</td>
<td>1.086</td>
</tr>
</tbody>
</table>

For the functions derived by MLR, leave-one-out (LOO) cross-validation was used (lines highlighted in italics); for PLS functions, 10-fold cross-validation (20 runs) was applied and the average $Q^2$ and $s_{PRESS}$ of the 20 runs are reported.

Sotriffer et al., Proteins 73 (2008), 395
### II. Scoring: What is possible and what is not?

#### Comparison with other scoring functions

<table>
<thead>
<tr>
<th>Function</th>
<th>$R$</th>
<th>$R^2$</th>
<th>$s$</th>
<th>$F$</th>
<th>$Q^2$</th>
<th>$s_{PRESS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFCscore: sfc_290m</td>
<td>0.843</td>
<td>0.711</td>
<td>1.09</td>
<td>99.2</td>
<td>0.692</td>
<td>1.12</td>
</tr>
<tr>
<td>(k = 7, n = 290)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-CScore eq3 (Wang 2002):</td>
<td>0.756</td>
<td>0.571</td>
<td>1.41</td>
<td>70.4</td>
<td>0.551</td>
<td>1.47</td>
</tr>
<tr>
<td>(k = 4, n = 200)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemscore (Eldridge 1997):</td>
<td>0.843</td>
<td>0.710</td>
<td>1.40</td>
<td>47.1</td>
<td>0.658</td>
<td>1.52</td>
</tr>
<tr>
<td>(k = 4, n = 82)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score2 (Böhm 1998):</td>
<td>0.890</td>
<td>0.792</td>
<td>1.27</td>
<td>40.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(k = 7, n = 82)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score1 (Böhm 1994):</td>
<td>0.873</td>
<td>0.762</td>
<td>1.38</td>
<td>32.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(k = 4, n = 45)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Testing on external data set and comparison with other functions

800 PDB complexes with exp. pKᵢ:

Wang et al.,


New, carefully compiled test set of 195 PDB complexes with exp. pKᵢ:

Cheng et al.,

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

Best functions:

<table>
<thead>
<tr>
<th>Function</th>
<th>$R_o$</th>
<th>SD</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFCscore: sfc_met</td>
<td>0.646</td>
<td>1.82</td>
<td></td>
</tr>
<tr>
<td>X-Score:HMScore</td>
<td>0.644</td>
<td>1.83</td>
<td></td>
</tr>
</tbody>
</table>

Improvement, but still only moderate correlation
Why did many functions in the past appear more successful than they are?

Very small external test sets of limited diversity

\[ \text{cf. how many of the now available complexes are well predicted by SFCscore!} \]

### Table: Residual < 1.5

<table>
<thead>
<tr>
<th>Function</th>
<th>N</th>
<th>( r_p )</th>
<th>( r_{pred}^2 )</th>
<th>SE_{pred}</th>
</tr>
</thead>
<tbody>
<tr>
<td>sfc_290m</td>
<td>551</td>
<td>0.874</td>
<td>0.763</td>
<td>0.809</td>
</tr>
<tr>
<td>sfc_223m</td>
<td>546</td>
<td>0.879</td>
<td>0.769</td>
<td>0.803</td>
</tr>
<tr>
<td>sfc_frag</td>
<td>417</td>
<td>0.915</td>
<td>0.818</td>
<td>0.835</td>
</tr>
<tr>
<td>sfc_855</td>
<td>555</td>
<td>0.850</td>
<td>0.720</td>
<td>0.820</td>
</tr>
<tr>
<td>sfc_ser</td>
<td>558</td>
<td>0.876</td>
<td>0.765</td>
<td>0.806</td>
</tr>
<tr>
<td>sfc_met</td>
<td>553</td>
<td>0.872</td>
<td>0.759</td>
<td>0.790</td>
</tr>
<tr>
<td>sfc_290p</td>
<td>559</td>
<td>0.875</td>
<td>0.765</td>
<td>0.796</td>
</tr>
<tr>
<td>sfc_223p</td>
<td>531</td>
<td>0.887</td>
<td>0.784</td>
<td>0.790</td>
</tr>
</tbody>
</table>

CAVE with any conclusions derived from too small test sets!

### II. Scoring: What is possible and what is not?

Have the limits of empirical approaches been reached?

Consider quality and comparability 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
  - may highly depend on pH, buffer, salt concentration, temperature
  - enzyme 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!
II. Scoring: What is possible and what is not?

Have the limits of empirical approaches been reached?

For the development of generic scoring functions:

Problems difficult to overcome, even by concerted efforts!

Focus on target- or target-class-specific functions!

- Target-specific adaptation of existing functions
- Better comparability of experimental data
- Definition of standards for acquisition of new affinity data possible

Recommendations ...

... for approaching the scoring problem:

1) Validate the scoring function for your system of interest

2) Train the scoring function for your system („Tailored scoring function“)

3) Try applying multiple scoring functions („Consensus Scoring“)

4) Tackle the problem with additional pre- and postfiltering steps
Further developments required …

... to overcome the most serious simplifications in scoring functions:

- single configuration of the binding partners in the complex
- no consideration of the unbound state
- no or simplified consideration of the solvent
- focused on enthalpic contributions and interaction descriptors
- additivity of interaction terms

_A single model may not be sufficient to capture the complex interplay of residual mobility, desolvation, and interaction quality in protein-ligand complexes!_

The wrong conclusion …
Acknowledgement

David Zilian
Daniel Cappel
Monika Nocker
Ulrich Peinz
Benjamin Schaefer
Martin Sippel
Christine Topf
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

Methods and Principles in Medicinal Chemistry
Edited by Christoph Santi

Virtual Screening
Principles, Challenges, and Practical Guidelines
Volume 48
Series Editors: E. Klimov, H. Kuty, G. Felters

available in autumn 2010