

Protein-Ligand Docking

Matthias Rarey



GMD - German National Research Center for Information Technology
Institute for Algorithms and Scientific Computing (SCAI)
53754 Sankt Augustin, Germany

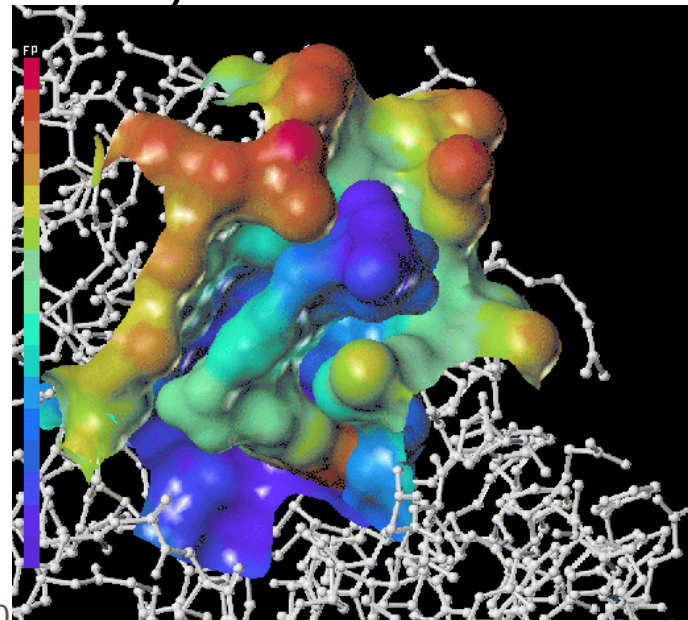
rarey@gmd.de

Outline of this lecture

- Introduction
 - The docking problem
 - Applications
 - Scoring functions
- Rigid-body protein-ligand docking
 - Clique-search-based methods
 - The CLIX approach
 - Geometric-hashing-based methods
- Flexible protein-ligand docking
 - Docking by simulation
 - Incremental construction algorithms
 - Genetic algorithms
- Protein-protein docking
 - next lecture by T.Lengauer

Introduction

- The molecular docking problem:
 - Given two molecules with 3D conformations in atomic detail
 - Do the molecules bind to each other? If yes:
 - How strong is the binding affinity?
 - How does the molecule-molecule complex look like?
- Docking problems in biochemistry:
 - Protein-Ligand docking
 - ◆ rigid-body docking
 - ◆ flexible docking
 - Protein-Protein docking
 - Protein-DNA docking
 - DNA-Ligand docking



Some basic principles...

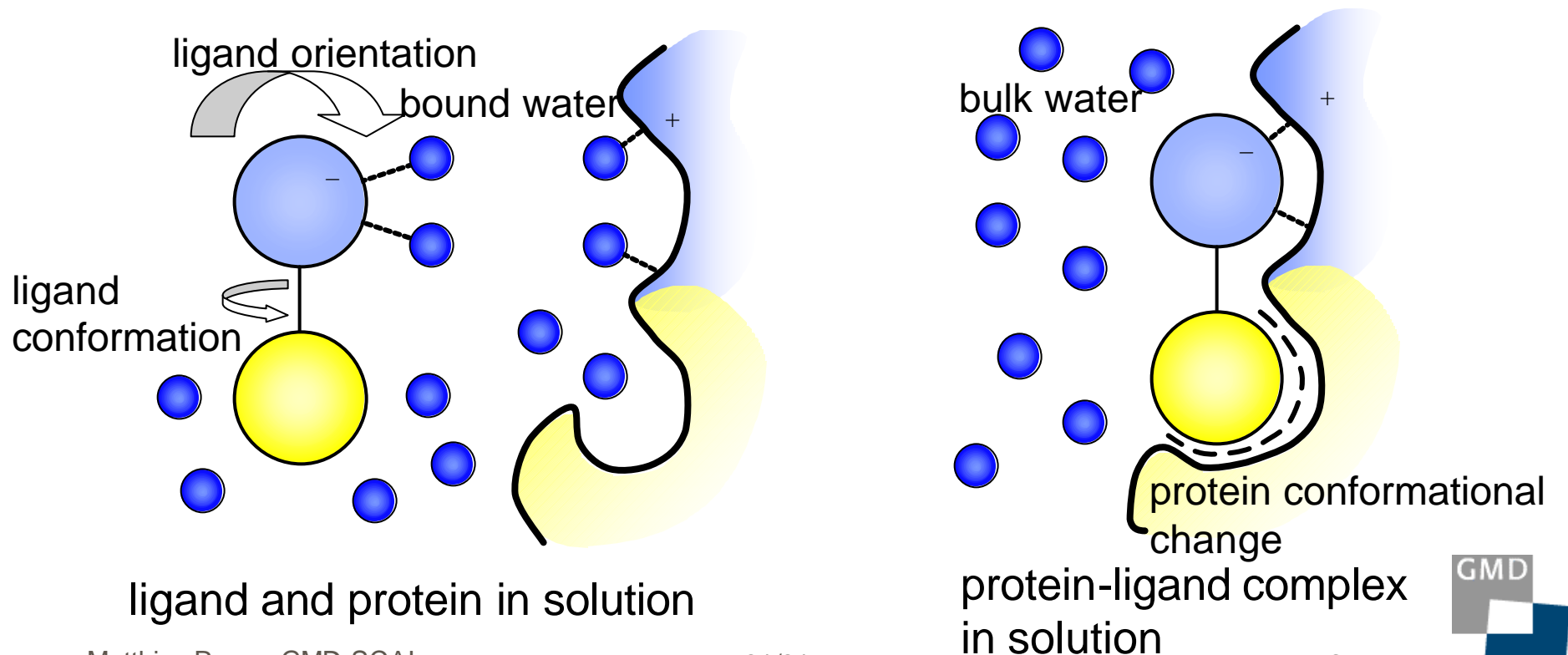
- The association of molecules is based on interactions:
 - hydrogen bonds, salt bridges, hydrophobic contacts
 - electrostatics
 - very strong repulsive interactions (van der Waals) on short distances
- The associative interactions are weak and short-range
=> tight binding implies surface complementarity
- Most molecules are flexible:
 - bond lengths > bond angles > torsion angles / ring conformations
 - macro molecules are restricted in conformational space in a complicated way

More basic principles...

- The binding affinity is the energetic difference to the uncomplexed state:
 - the surrounding medium (water in most cases) plays an important role
 - entropy can have a significant impact to the binding energy
- The binding affinity describes an ensemble of complex structures, not a single one
 - tight binders often have a dominating binding mode ...
 - ... and weak binders?

Energetic Contributions

- weak short-range interactions imply complementarity
- ligand (and protein) are conformationally flexible
- energy estimation is difficult (solvent, electrostatics, entropic effects, etc.)



Binding affinities

Free Energy of Binding

$$\Delta G = \Delta H - T \Delta S$$

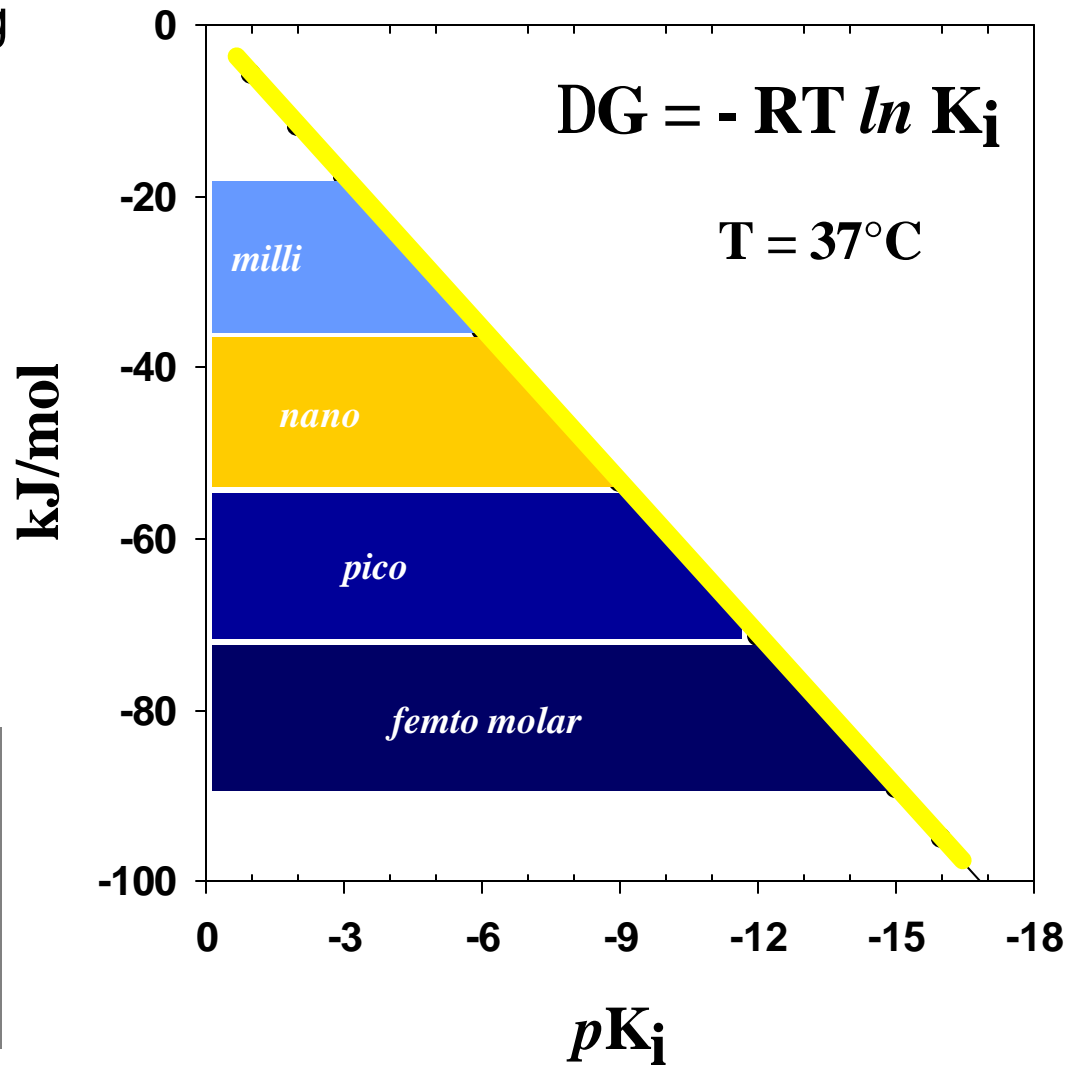
Equilibrium Constant

$$K_i = \frac{[P][L]}{[PL]}$$

~ 6 kJ/mol

@ 1 order in K_i

"1 -2 hydrogen bonds"



Applications

- Estimating the binding affinity
 - Searching for lead structures for protein targets
 - Comparing a set of inhibitors
 - Estimating the influence of modifications in lead structures
 - De Novo Ligand Design
 - Design of targeted combinatorial libraries

- Predicting the molecule complex
 - Understanding the binding mode / principle
 - Optimizing lead structures

Scoring functions

- Input: 3D structure of a protein-ligand complex
- Output: estimated binding energy ΔG (freie Enthalpie)
- Comments:
 - measured ΔG describes energetic difference between bound and unbound state based on a structure ensemble.
 - Assumption: measured ΔG is dominated by a single structure of minimal energy
 - $\Delta G = \Delta H - T \Delta S$ ΔH : enthalpic contributions, ΔS : entropic contr. ΔS is very difficult to approximate!
 - more about energy: Atkins, (Kurzlehrbuch) Physikalische Chemie, Spektrum Akademischer Verlag, 1992

Scoring functions

■ Force field:

- describes only enthalpic contributions ΔH , no estimate for ΔG
- conformation terms (bond lengths and angles) have a steep rise (sometimes not used in docking calculations)
- time consuming calculations (electrostatics)

■ Potentials of mean force / Knowledge-based scoring

- Analysis of known low-energy complexes: frequent occurrence \rightarrow energetically favorable
- Pair potentials: $f(a,b,d)$ = relative frequency of observation *atom of type a and atom of type b occur with distance d in the database*
- Conversion into an energy term $g_{ab}(d)$ (inverse Boltzmann law)

total energy:
$$E(R, L) = \sum_{\text{Atome } r, l \in R, L} g_{a(r)a(l)}(d(r, l))$$

$d(r, l)$: distance between r and l $a(r)$: atom type of r

Scoring functions

- Empirical scoring functions
 - calibration of microscopic observations with measured macroscopic ΔG values
 - data: set of protein-ligand complexes with known 3D structure and binding affinity ΔG
- Example: Böhm-Function

(Böhm, J.Comput.-Aided Mol. Design, Vol. 8 (1994), pp 243)

- Scoring function:

$$\Delta G = \Delta G_0 + \Delta G_{rot} N_{rot} + \Delta G_{hb} \sum_{\text{neutral H-bonds}} f(\Delta R) f(\Delta \mathbf{a}) +$$

$$\Delta G_{io} \sum_{\text{ionic interactions}} f(\Delta R) f(\Delta \mathbf{a}) + \Delta G_{lipo} |A_{lipo}|$$

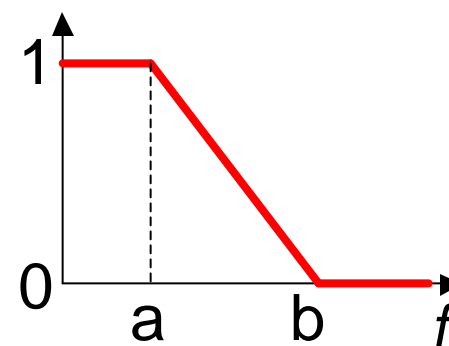
Scoring functions

■ Contributions:

- ΔG_0 : Lost of transformation entropy (?)
- ΔG_{rot} : Lost of conformational degrees of freedom (ligand entropy)
 $\Delta G_{\text{hb}} / \Delta G_{\text{io}}$: hydrogen bonds (neutral / charged)
- ΔG_{lipo} : lipophilic contact surface area

■ The function f penalizes deviations from the ideal interaction geometry:

■ ΔG values are determined by regression



Rigid-body protein-ligand docking

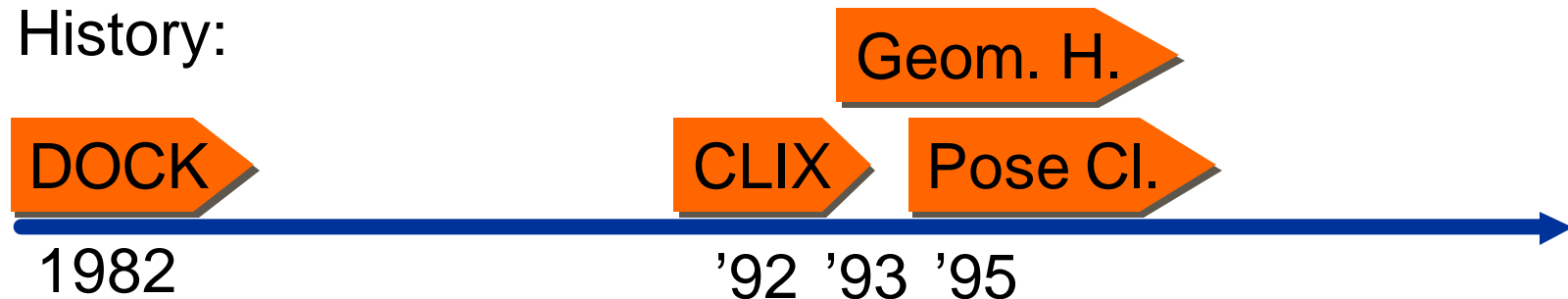
- Main assumptions:

- protein is considered as rigid
- ligand is considered as rigid

- Applications:

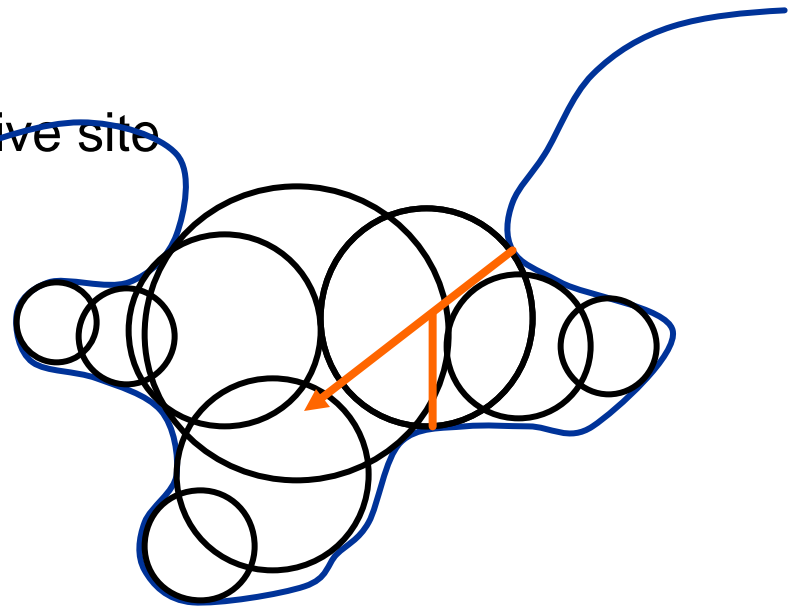
- docking of small or very rigid molecules
- docking of fragments (flexible docking, de novo design, combinatorial library design)
- docking of multi-conformer databases

- History:



DOCK

- Kuntz et al., J.Mol.Biol. Vol. 161, pp. 269
- Basic Idea: represent active site by set of spheres, perform sphere matching
- Algorithm 1: SPHGEN
 - calculate the molecular surface
 - generate spheres covering the active site
 - cluster spheres, remove
 - ◆ very similar ones
 - ◆ radius too large
 - select clusters defining the active site
 - color spheres by properties



DOCK

- Algorithm 2: MATCH (calculate a matching between ligand atoms L and protein spheres K)

- two matches (l_1, k_1) , (l_2, k_2) are *distance-compatible* if
$$|d(l_1, l_2) - d(k_1, k_2)| \leq \mathbf{e}$$

- search for matchings $M = \{(l_i, k_i)\}$ with

$$\max_{i,j} |d(l_1, l_2) - d(k_1, k_2)| \leq \mathbf{e}$$

- Matching-Graph: nodes L x K, edges between distance-compatible nodes
- Matchings are cliques in the matching graph
(cliques = completely connected subgraphs)

DOCK

- Outline of MATCH:

- enumeration of all matchings of size 4
- orientation of molecule with RMSD fit routine
- filtering of orientations: protein-ligand overlap, stereo chemistry,...
- extension of matching
- optimizing the orientation (all matches fit)
- scoring and selection

- Extensions of DOCK:

- several scoring schemes
- ligand flexibility (fragment joining and incremental construction)
- chemical properties in matching phase

Algorithm: Superposition of point sets

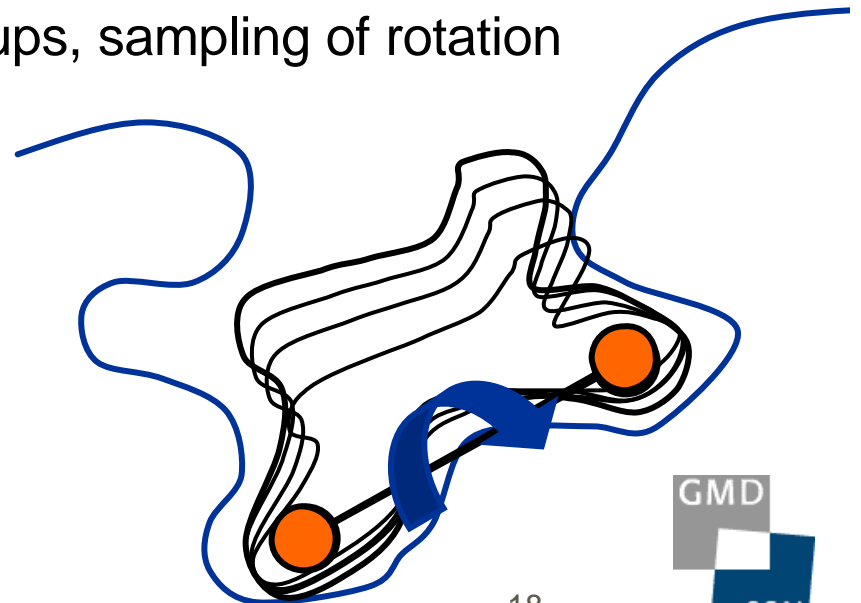
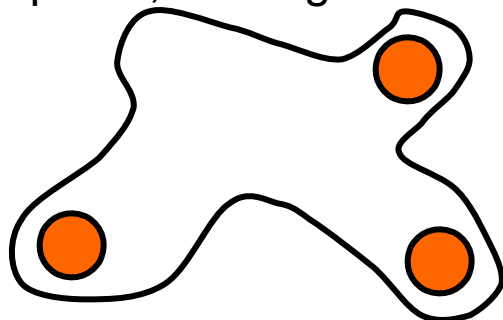
- Problem: given two vector sets $X=\{x_1, x_2, \dots, x_n\}$, $Y=\{y_1, y_2, \dots, y_n\}$, calculate transformation (t, Ω) minimizing

$$\text{RMSD}_{X,Y}(\Omega, t) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \Omega y_i - t)^2}$$

- see lecture 12.12 (protein structure alignment)

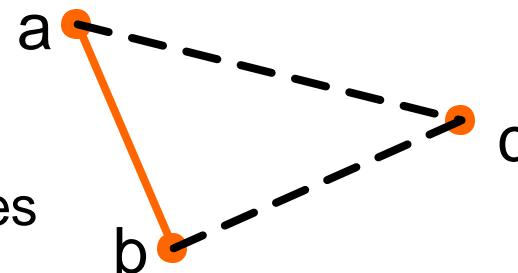
CLIX

- Lawrence et al, PROTEINS: Struct., Func., and Gen., Vol. 12 (1992), pp 31
- based on interaction maps calculated with GRID
- Algorithm:
 - identification of *interaction target points* in the maps ●
 - enumeration of all pairs of distance-compatible matches
 - superposition of two matches groups, sampling of rotation around common axis:
 - ◆ searching for additional matches
 - ◆ overlap test, scoring



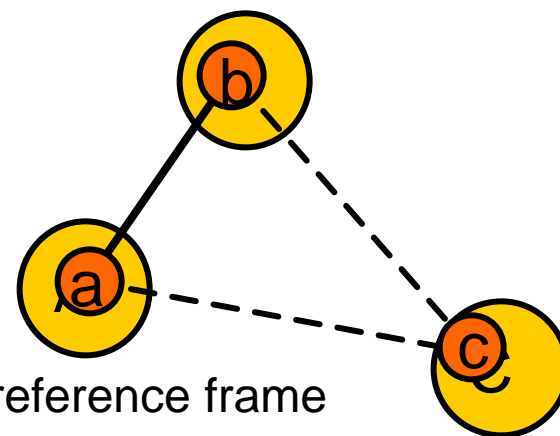
Geometric hashing

- Fischer et al, J.Mol.Biol., Vol. 248 (1995), pp. 459
- Key features
 - method from pattern recognition applied to docking
 - based on the dock sphere representation
 - allows direct application to database search
- Constructing the hash table for ligand atom triplets (a,b,c):
 - entries have address based on atom-atom distances
 - information stored: ligand id, basis (a,b)
- Basic search algorithm:
 - search for matching (two spheres, basis) allowing large number of third atom matches
 - extension and evaluation of matches



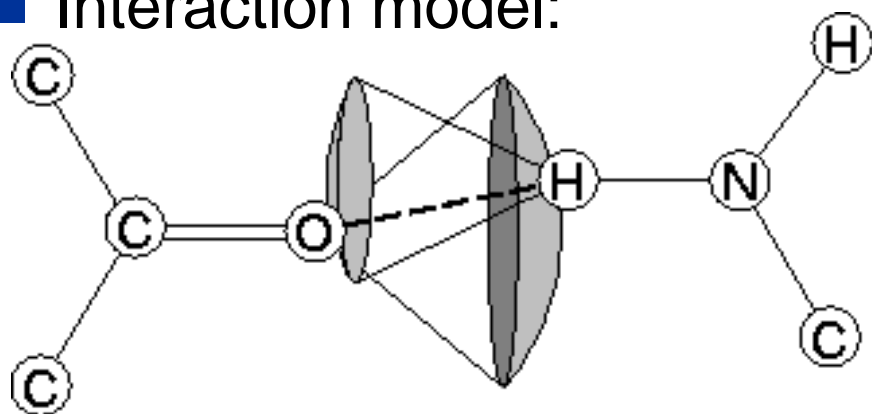
Geometric hashing

- Search for *seed-matchings*: (voting scheme)
- \forall pairs of spheres (A,B) // search for matching bases
 - \forall spheres C: // sphere who gives the vote
 - ◆ \forall entries (ligand,basis) from hash table with matching distances:
 - ◆ increase vote for (ligand,basis)
 - ◆ insert (C,c) into matchlist of (ligand,basis)
 - \forall (ligand,basis) with $> T$ votes:
 - ◆ check all pairwise distances
 - ◆ enter into seed matching list
- Method in pattern recognition:
 - basis is d-dimensional and defines a coordinate reference frame
- here:
 - due to complexity, basis is only 2-dimensional
 - \Rightarrow matches spheres/atoms may not be superimposable

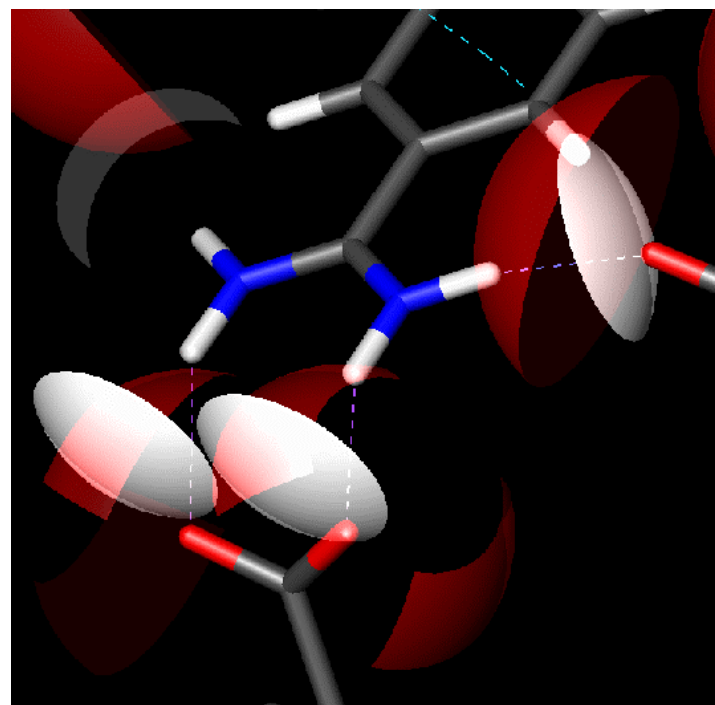


Pose clustering

- Rarey et al, J.Comput.-Aided Mol. Des., Vol 10 (1996), pp. 41
- Method from pattern recognition applied to ligand orientation based on physico-chemical interactions
- Interaction model:

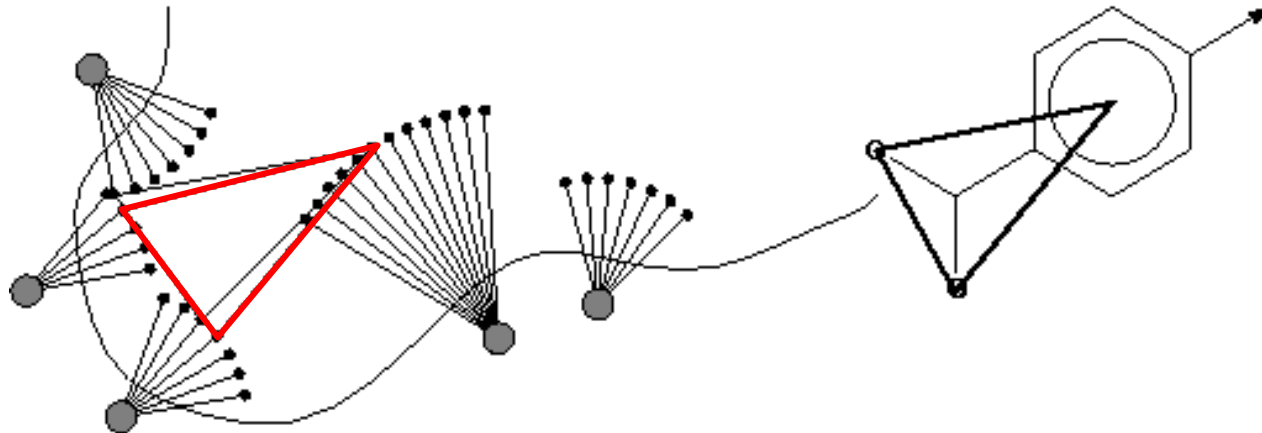
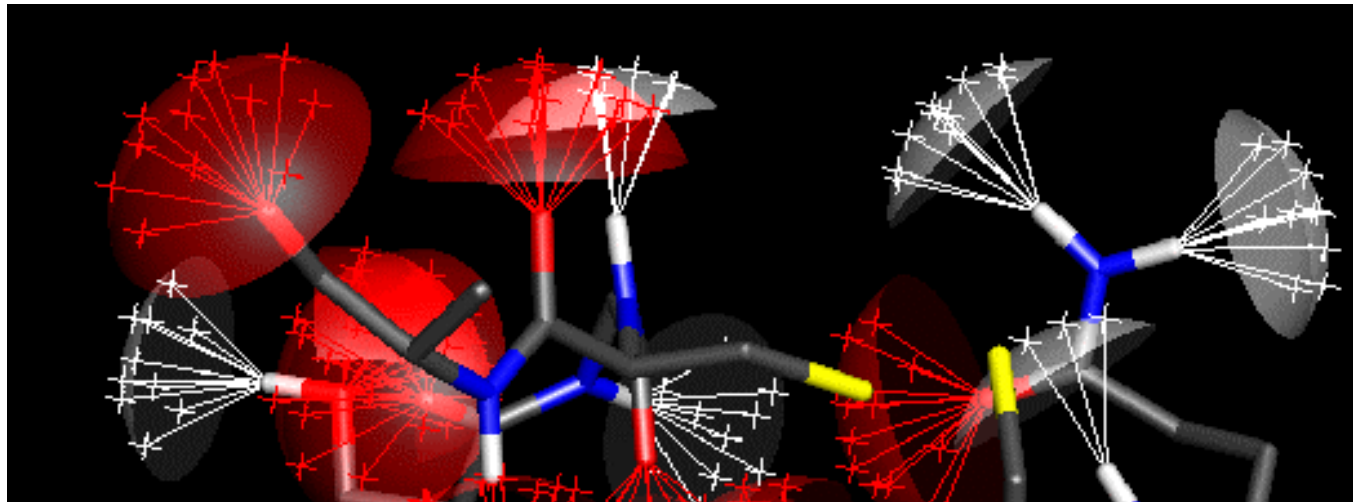


- compatible interaction types
- interaction center of first group lies approximately on interaction surface of second group ...
- ... and vice versa



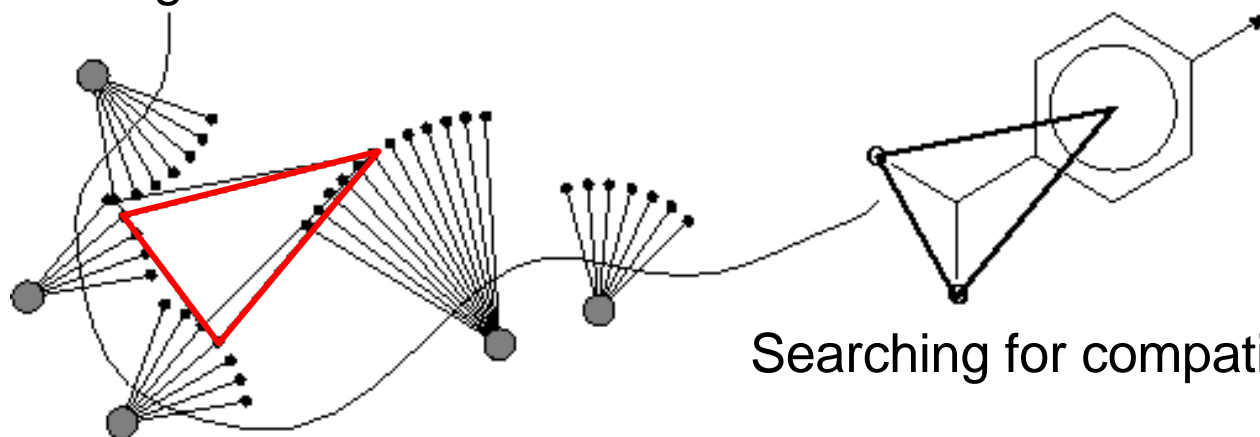
Pose clustering

- Interaction surfaces are approximated by discrete points:

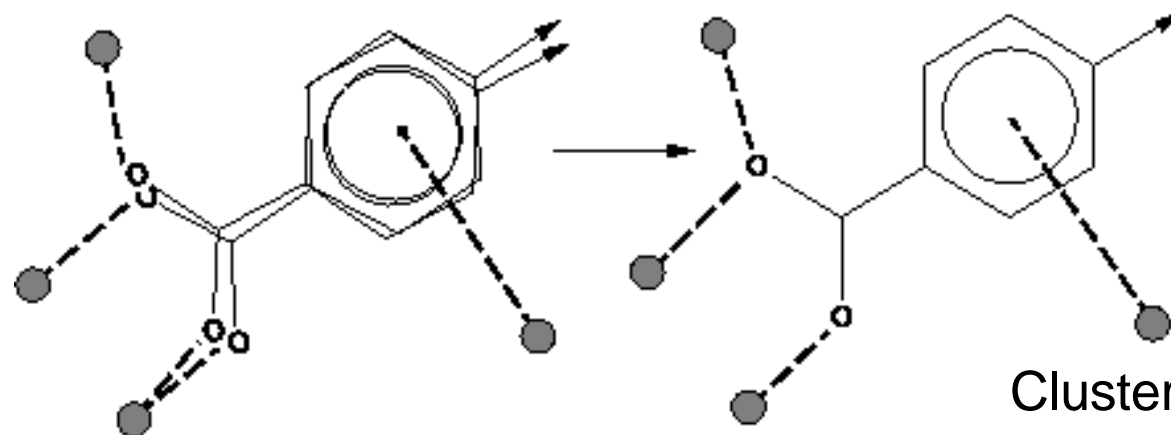


Pose clustering

Pose clustering:



Searching for compatible triangles



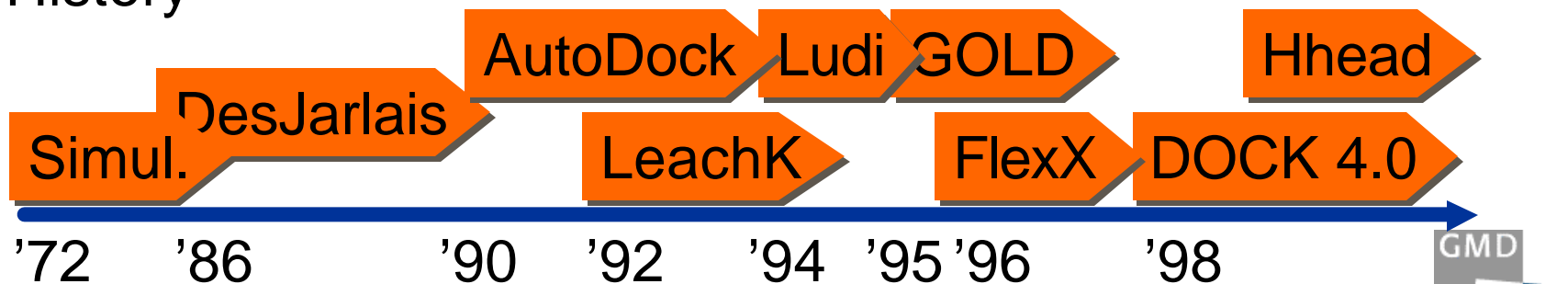
Clustering of transformations

Pose clustering

- Preprocessing: construct hash table for all interaction type pairs a,b :
 - store all pairs of interaction points p,q with address $d(p,q)$
 - chain lists twice, sorted by point id of p and q
- Search of initial ligand orientations:
 - \forall triplets (a,b,c) of ligand interaction centers:
 - ◆ generate a list of all type- and distance-compatible pairs of interaction points for (a,b) and (a,c)
 - ◆ construct all distance-compatible triangles (p,q,r) by list merging
 - ◆ \forall triangles (p,q,r) : generate ligand transformation, overlap test
- Cluster orientations by pairwise RMSD
- \forall remaining ligand orientations:
 - ◆ extend matching, overlap test, scoring

Flexible protein-ligand docking

- Main assumptions (not valid for simulations)
 - ligand flexibility is limited to torsion angles (+ ring conformations)
 - protein is considered as (nearly) rigid
 - discrete models for conformations and interactions
 - “binding-pathway” is not considered
- Application
 - Analyzing complexes, searching for possible binding modes
 - Virtual screening of small molecule databases
- History



Docking by simulation

■ Method:

- generate (random) start orientations
- MD simulation / energy minimization for all start orientations

■ Pros/Cons:

- can handle protein flexibility to an arbitrary extend
- very time consuming
- more a local minimization (large structural changes are difficult)

■ Applications:

- Di Nola et al, PROTEINS: Struct., Func., and Gen., Vol. 19 (1994), pp 174
- Luty et al, J. Comp. Chem., Vol. 16 (1995), pp 454

Hybrid methods

- Method:
 - use fast algorithms for placement, MD for refinement
- Applications:
 - Wang et al, PROTEINS: Struct., Func., and Gen., Vol. 36 (1999), pp 1
 - Hoffmann et al, J. Med. Chem., in press
- Wang's procedure:
 - generate low energy conformations
 - rigid-body docking (soft van der Waals potentials)
 - minimization in the active site (amber force field, rigid protein)
 - torsion angle refinement routine (scanning alternative torsions)
 - simulated annealing (minimization, all degrees of freedom)

Simulated annealing: AutoDOCK

- Goodsell et al., PROTEINS: Struct., Func., and Gen. Vol. 8 (1990), pp. 195
- Simulated annealing:
 - random change in configuration is accepted with probability

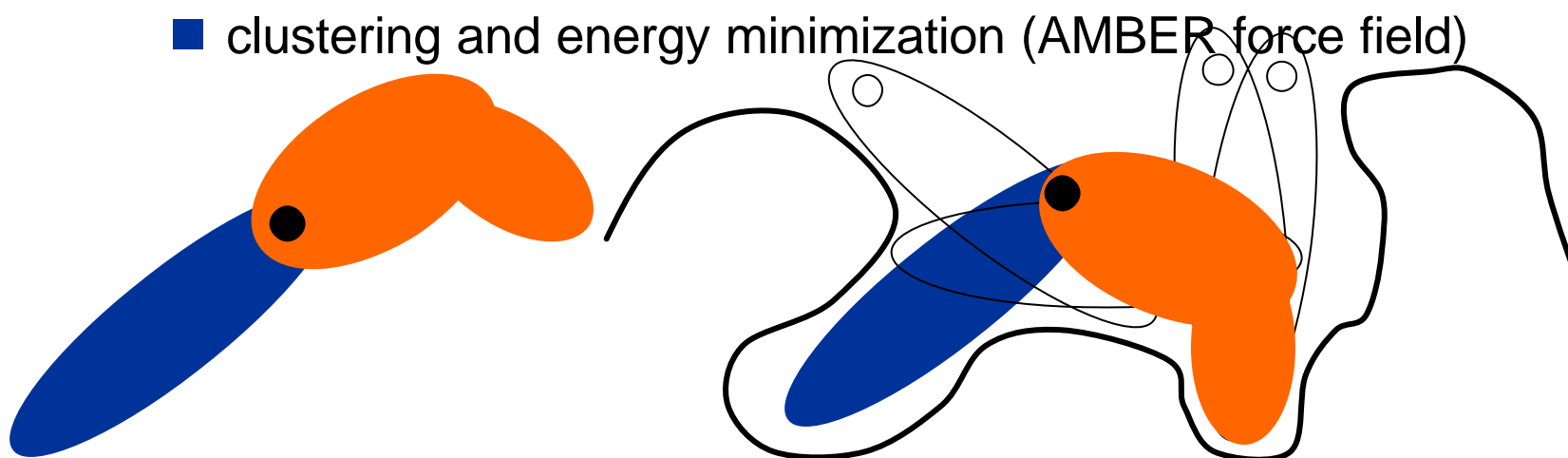
$$P(\Delta E) = e^{\frac{-\Delta E}{k_B T}}$$

ΔE : energy difference of change
 k_B : Boltzmann's constant
 T : user defined temperature

- cooling schedule reduces T over time (for example $T \leftarrow cT$)
makes energetically unfavorable moves more unlikely
- Application specific:
 - move: small random displacement of all degrees of freedom
 - calculation of E : affinity potentials as in GRID

Place & join algorithms

- DesJarlais et al, J.Med.Chem. Vol. 29 (1986), pp 2149
- Algorithm:
 - cut the ligand into few fragments (one overlapping atom (linker))
 - place all fragments with the DOCK algorithm
 - for a specific sequence of fragments:
 - ◆ join two fragments in all placement combinations with close location of the linker atom
 - clustering and energy minimization (AMBER force field)



Place & join algorithms

- Sandak et al., CABIOS Vol. 11 (1995), pp. 87
- Hinge Bending: extending geometric hashing
 - Hinge: Ligand with two adjacent, flexible bonds or protein domain movement
 - Hash table for ligand data set:
 - ◆ store ligand fragment, hinge location
 - Matching phase: \forall receptor sphere triplets:
 - ◆ search for ligand atom triplets in hash table
 - ◆ perform a voting for a hinge location
 - Join phase: \forall hinges with high votes
 - ◆ combine collision free placements of fragments
 - ◆ scoring and selection

Incremental construction algorithms

■ Overall strategy:

- divide the molecule into fragments
- place one (several) fragment(s) into the active disregarding the rest of the molecule
- add remaining fragments incrementally:
 - ◆ explore conformation space, clash test
 - ◆ search for new interactions, scoring
 - ◆ select new set of extended placements

■ Application to the docking problem:

- Moon et al., PROTEINS: Struct., Func., and Gen., Vol. 11 (1991), pp 314
- Leach et al., J. Comp. Chem., Vol. 13 (1992), pp 730
- Rarey et al., J. Mol. Biol., Vol. 261 (1996), pp 470
- Welch et al., Chem. & Biol., Vol. 3 (1996), pp 449
- Makino et al., J. Comp. Chem., Vol. 18 (1997), pp 1812

Incremental construction algorithms

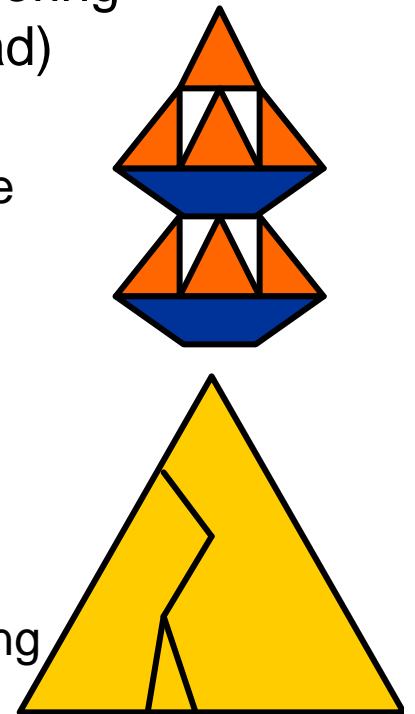
■ Search Strategies:

- GREEDY: after adding a fragment, select the high scoring ones and reject the rest (GROW, FlexX, Hammerhead)

- ◆ scales linear with the number of fragments
- ◆ optimal solution may be sub-optimal during buildup (the larger the considered set and the lower the number of fragments, the lower is the risk of missing the optimal placement)

- BACKTRACKING: performs a recursive (depth first) search through the whole configuration tree (Leach)

- ◆ scales exponentially with the number of fragments
- ◆ no risk of losing the optimal solution due to tree pruning

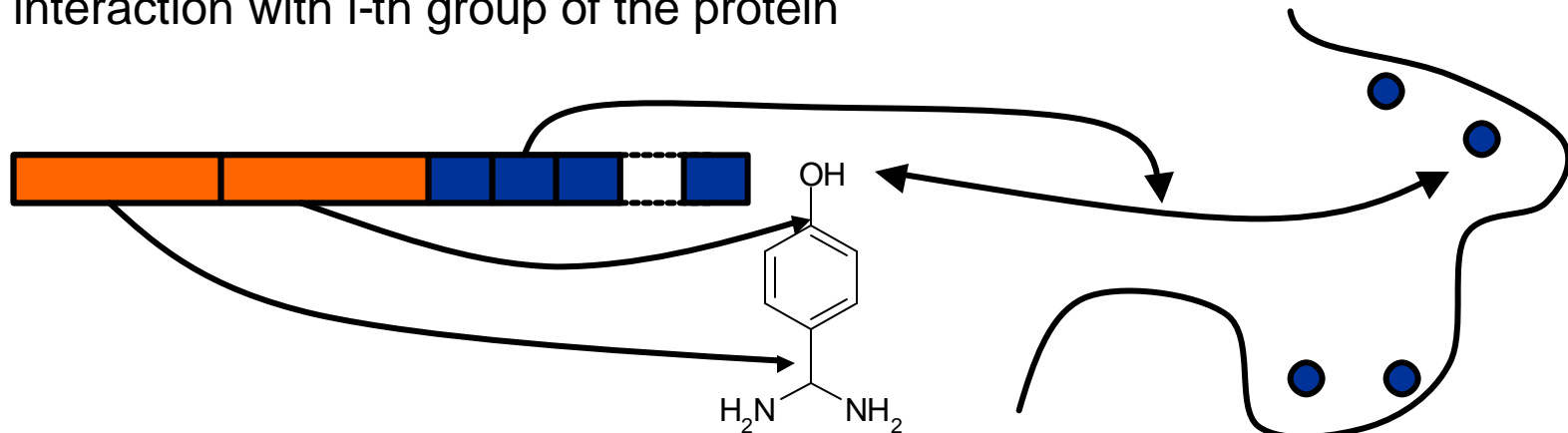


■ Additional steps:

- Score estimation
- Placement optimization
- Solution clustering

GOLD

- Molecule representation (N rotatable bonds)
 - conformation string (N bytes), one byte each coding a torsion angle
 - a matching string (integer), defines mapping between hydrogen bond donors/acceptors: $M(k)=l$ if k-th interaction group of ligand forms interaction with l-th group of the protein



- Fitness evaluation of individual with chromosome c:
 - build conformation according to c
 - superimpose matched interacting groups
 - calculate docking score: $-E_{\text{hydrogen bond}} - (E_{\text{internal}} + E_{\text{complex}})$

GOLD

- Population:
 - 5 sub-populations of 100 individuals each
 - about 20-50 runs, each up to 100000 genetic operations
- Genetic Operators:
 - crossover: two-point crossover between two parent individuals
 - mutation: one-point mutation
 - migration: one individual moves between sub-populations
- operators are randomly selected

Concluding remarks

- docking performance

- correct structure can be predicted in about 70% of the test cases

- prediction of binding affinity is very difficult:

1. ranking protein-ligand complex geometries → good, not perfect
2. ranking different ligands with respect to binding → weak correlations
3. free energy estimation of protein-ligand complexes → more or less unsolved

- challenges

- handling protein flexibility

- improving reliability of structure and affinity prediction