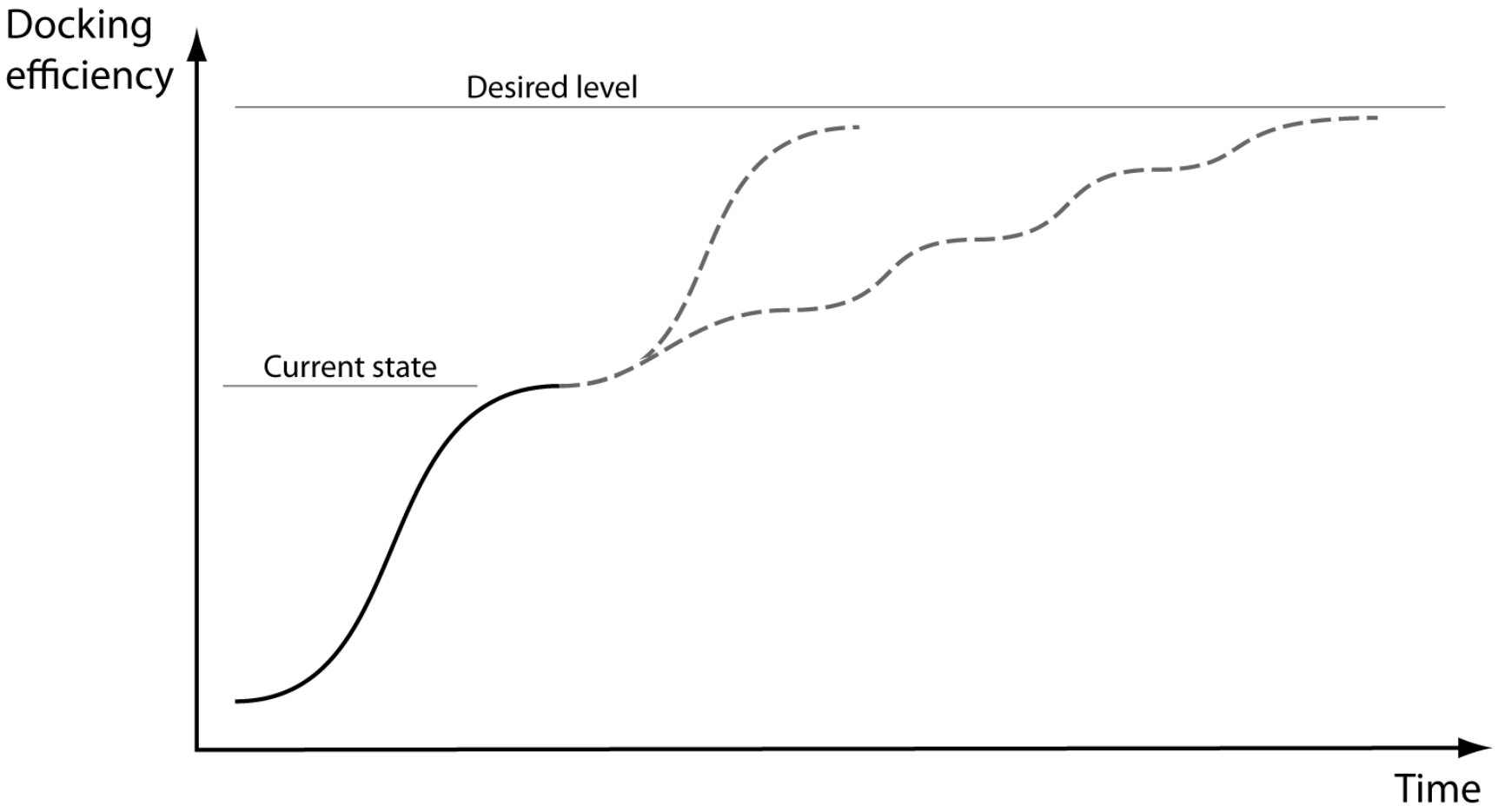
# Key Directions of Mastering Docking and Scoring Approaches

Oleg Stroganov[1], Fedor Novikov[1], Viktor Stroylov[1], Val Kulkov[2] and **Ghermes Chilov**[1]

[1] MolTech Ltd, Russian Federation, [2] BioMolTech Corp, Canada

ACS Fall Meeting, Boston, USA 22 August 2010

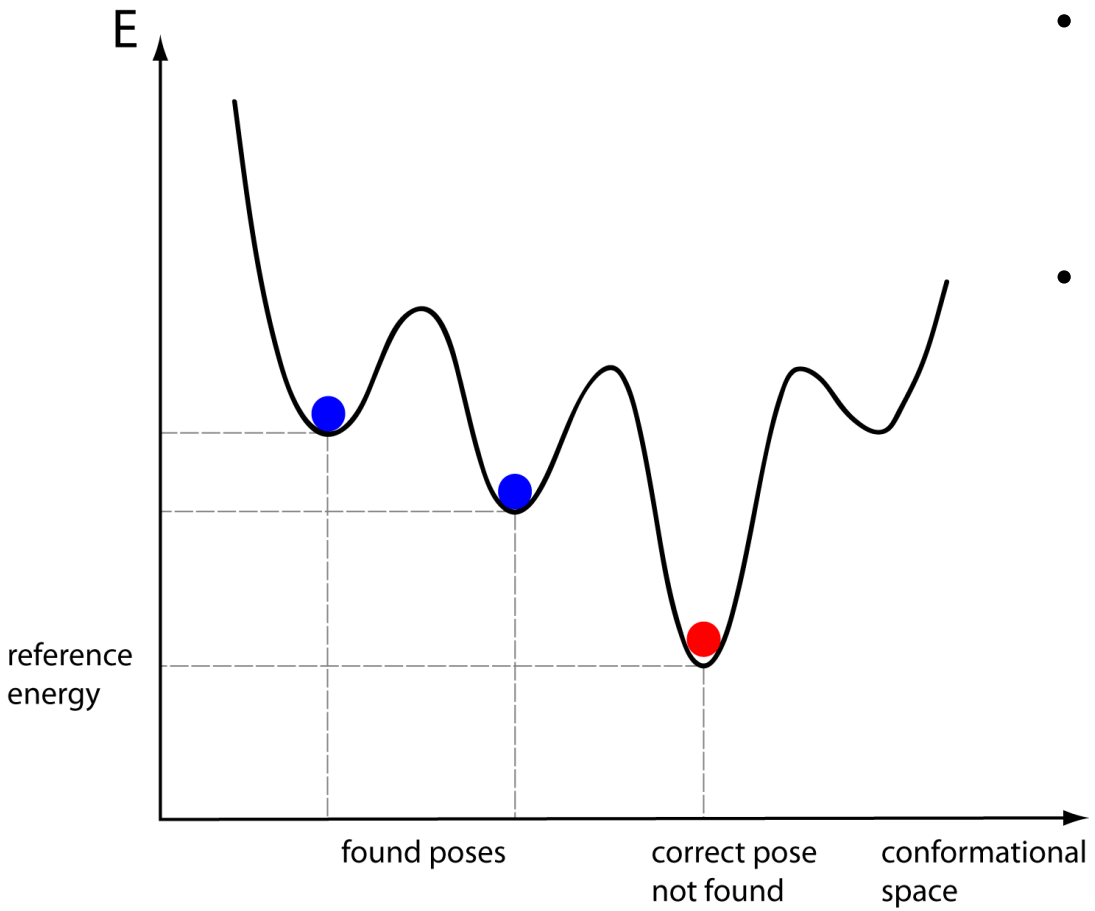
# Current state of docking approach



# Plan of the presentation

- Sampling errors
- Scoring errors
- Novel concepts in docking and scoring

# Sampling errors




- Account of ligand degrees of freedom
- Protein, solvent degrees of freedom…

# Statistics of sampling errors

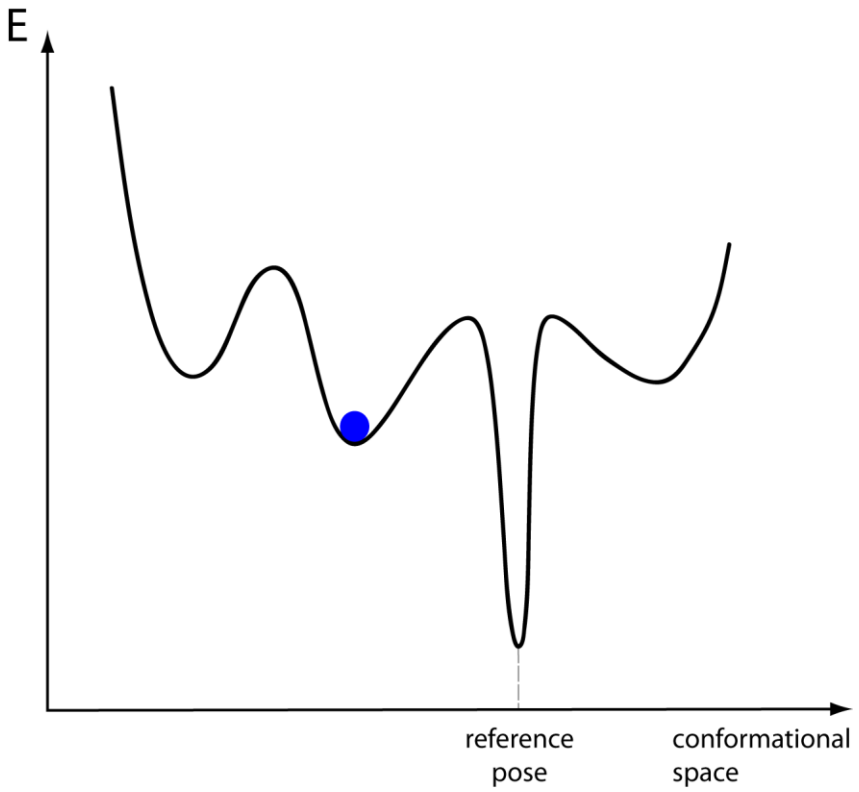
- 407 protein-ligand complexes were used in the docking success rate benchmark
- 32 out of 63 docking failures were attributed to the sampling errors*

| Ligand NFRB | N structures | % of errors |
|---|---|---|
| 00-05 | 246 | 2.4 |
| 06-10 | 103 | 6.8 |
| 11-15 | 42 | 26.2 |
| >15 | 16 | 43.8 |
| all | 407 | 7.6 |

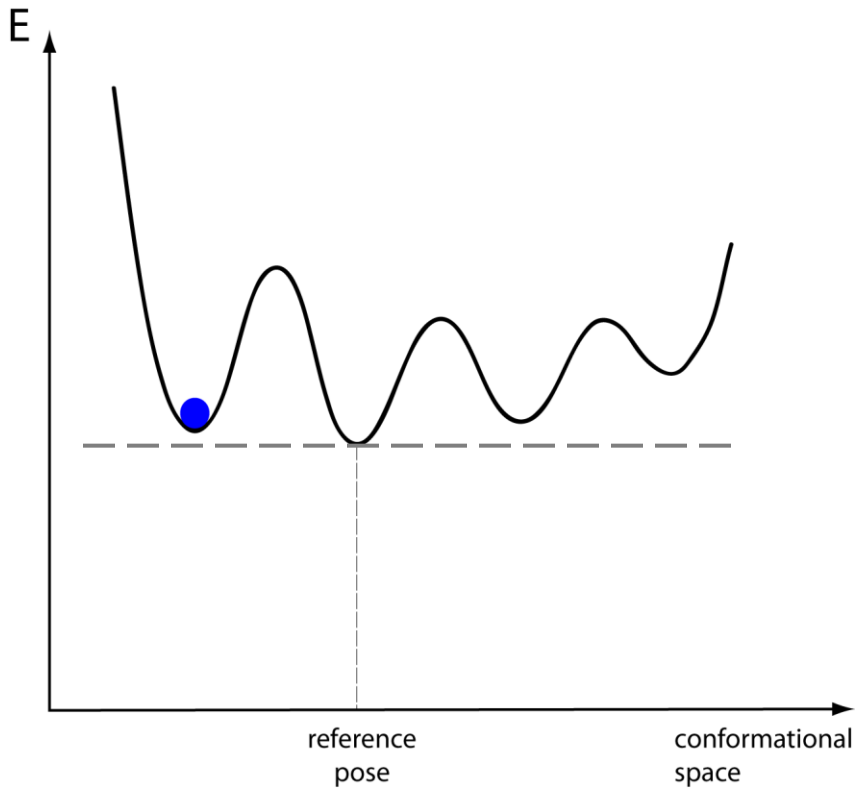
* J. Chem. Inf. Model., 2008, 48, 2371-2385

# Examples of sampling errors

"Narrow gorge" energy profile
(PDB 1hte)




"Flat" energy profile
(PDB 1tni)



# Approaches to minimize sampling errors

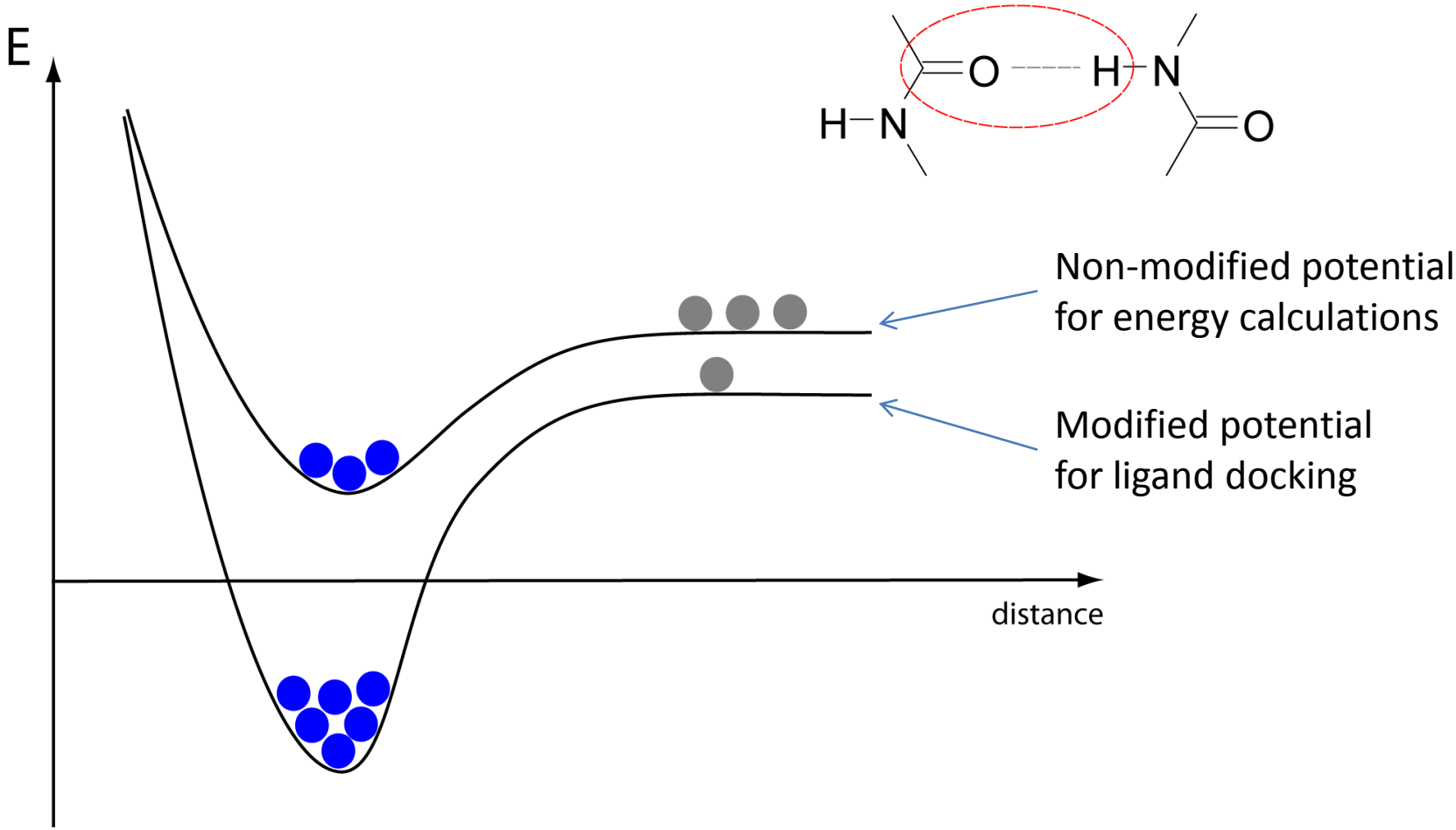
- Restraints on particular pairs of atoms
  - Rigid bond
  - Specification of protein and ligand atoms required
- Energy traps for particular interactions
  - Soft bond
  - Specification of ligand atom is not required
- Protein-specific optimization of docking settings
  - Dependence of the docking algorithm settings on the protein and ligand properties

# Geometric restraints during ligand docking*

| Name | PDB ID | NFRB | Constraints | | Docking success rate, % | | speedup |
|---|---|---|---|---|---|---|---|
| | | | N | Type | default | constrained | |
| uPA | 1gj8 | 5 | 1 | Q** | 75% | 100% | 0,84 |
| DHFR | 1dhf | 7 | 1 | Q | 85% | 100% | 1,17 |
| Coagulation factor X | 1g2l | 9 | 1 | Q | 70% | 100% | 1,18 |
| ACE | 2oc2 | 13 | 1/1 | Q/Hb*** | 50% | 95% | 1,67 |
| HIV-1 protease | 1b6p | 15 | 1 | Hb | 70% | 100% | 1,67 |
| Plasmepsin-2 | 1lf2 | 15 | 1 | Hb | 25% | 75% | 1,25 |
| HIV-1 protease | 2a4f | 15 | 2 | Hb | 80% | 100% | 1,05 |
| PPAR-delta | 2awh | 15 | 1/1 | Q/Hb | 30% | 60% | 1,15 |
| Protein D7 | 3dzt | 15 | 2 | Hb | 85% | 100% | 1,44 |
| HIV-2 protease | 1hii | 16 | 3 | Hb | 30% | 45% | 1,10 |
| Cytochrome P450 102 | 1jpz | 16 | 1 | Hb | 10% | 70% | 1,34 |
| HIV-1 protease | 1hxw | 18 | 1 | Hb | 40% | 85% | 1,43 |
| MMP-3 | 1hfs | 18 | 1 | Me**** | 25% | 95% | 1,57 |
| FTase-alpha | 1jcq | 19 | 1/1 | Me/Q | 40% | 95% | 1,22 |
| Trans-sialidase | 1s0i | 22 | 2 | Q | 35% | 75% | 1,25 |

* Lead Finder v. 1.1.14; ** charged H-bond interaction; *** neutral H-bond interaction; **** metal coordination

# Energy traps during ligand docking

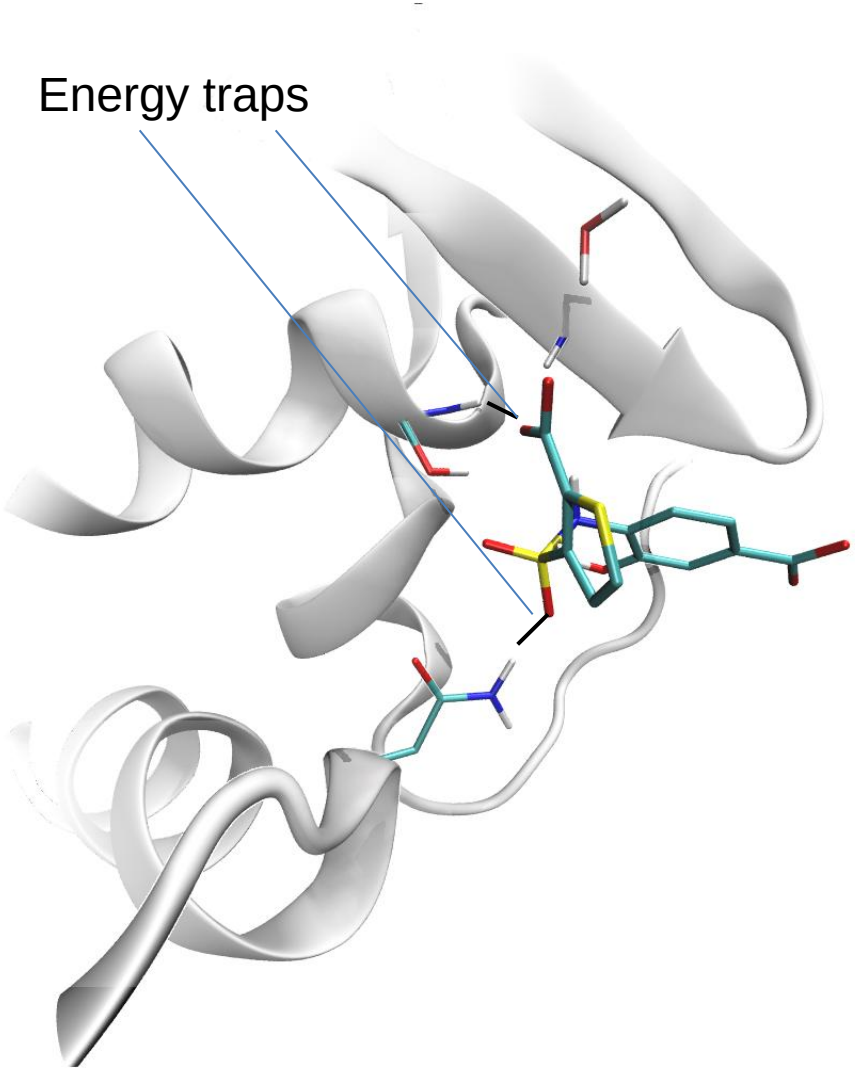


# Efficiency of energy traps in ligand docking

- 16 protein-ligand complexes with NFRB>14
- 1-2 energetic traps per protein with X-fold increased potential

| | N of docked structures | Average docking time, s |
|---|---|---|
| No constraints | 7 | 450 |
| 5-fold increase | 9 | 456 |
| 10-fold increase | 10 | 455 |
| 50-fold increase | 10 | 450 |

# Efficiency of energy traps



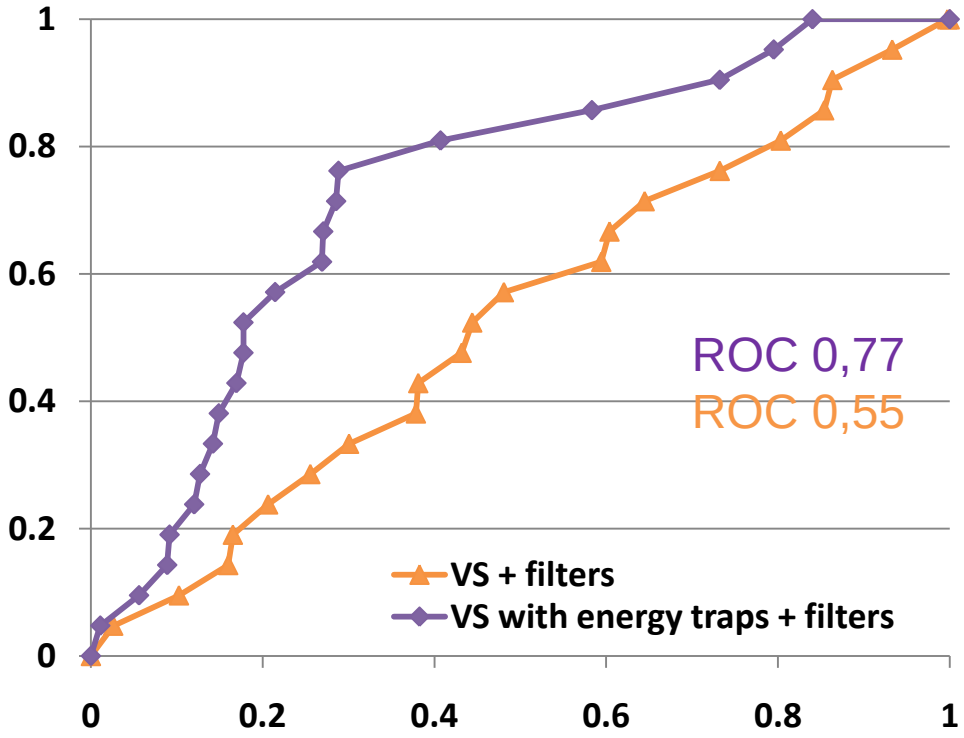

- Beta-lactamase – tough target of the DUD test set*
- ROC ~ 0,5 for 6 well-known docking programs**
- 2 cognate ligands out of 21 are docked correctly by Lead Finder

* J. Med. Chem. 2006, 49, 6789-6801

** J. Chem. Inf. Model. 2009, 49, 1455–1474

# Efficiency of energy traps in virtual screening

18 out of 21 ligands are correctly docked using energy traps




| Ligand | Docking success rate | |
|---|---|---|
| | default | energy traps |
| 1 | 33% | 11% |
| 2 | 11% | 56% |
| 3 | 0% | 22% |
| 4 | 22% | 67% |
| 5 | 0% | 89% |
| 6 | 11% | 56% |
| 7 | 22% | 100% |
| 8 | 0% | 0% |
| 9 | 0% | 100% |
| 10 | 22% | 78% |
| 11 | 33% | 100% |
| 12 | 22% | 89% |
| 13 | 33% | 67% |
| 14 | 11% | 89% |
| 15 | 11% | 78% |
| 16 | 11% | 89% |
| 17 | 33% | 89% |
| 18 | 100% | 89% |
| 19 | 22% | 78% |
| 20 | 33% | 78% |
| 21 | 67% | 100% |
| **>50%** | **2** | **18** |

# Protein-specific docking algorithm

- Adjustable settings of the docking algorithm
  - Initial pool of individuals
  - Number of generations
  - Number of individuals in the generation
  - Niche size and thickness

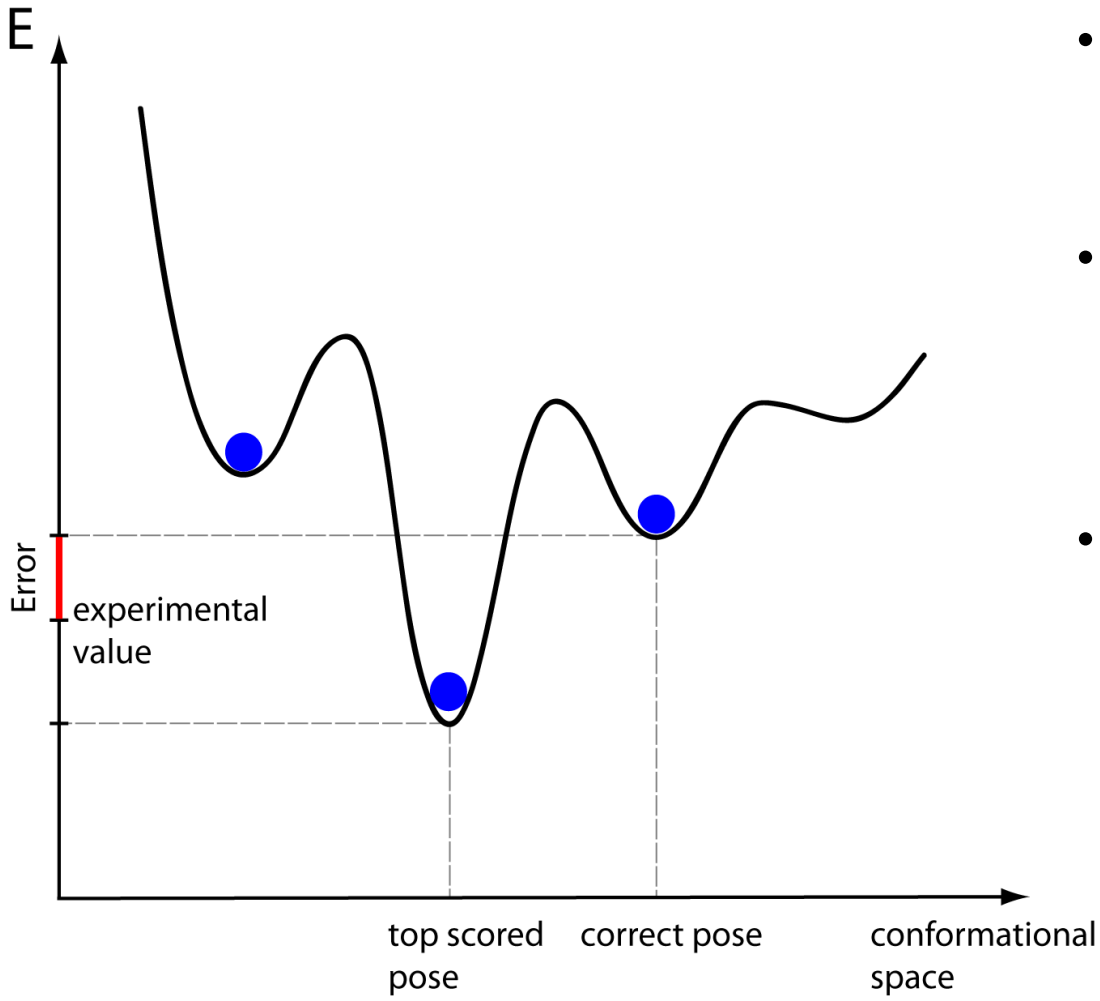
- Parameters influencing settings of the docking algorithm
  - Number of ligand freely rotatable bonds
  - Active site size
  - Number of H-bond donors and acceptors in ligand

# Protein-specific docking algorithm

- The main contribution to the customized docking algorithm comes from the number of H-bond donors and acceptors in ligand

| | Minimum settings | Account of NFRB | Account of H-bonds | Account of active site size | Default settings | Maximum settings |
|---|---|---|---|---|---|---|
| Docking success rate, % | 32.7 | 44.2 | 58.9 | 64.5 | 66.0 | 68.2 |
| N of correctly docked structures | 25 | 39 | 60 | 68 | 68 | 71 |
| Average docking time, s | 18 | 42 | 40 | 45 | 79 | 206 |

# Scoring errors




- Incorrect pose out-score correct ones
- Energy of correct pose differs from the experimental value
- Error in small molecule energy calculation ~1 kcal/mol

* Mol Phys, 2008, 106, 2107-2143

# Statistics of scoring errors

- Accuracy of binding error estimation

| | N of structures | RMSD, kcal/mol | $R^2$ |
|---|---|---|---|
| CSAR set | 345 | 1.98 | 0.57 |
| Lead Finder set | 285 | 1.80 | 0.49 |
| All | 630 | 1.90 | 0.53 |

- Statistics of scoring errors in docking (407 protein-ligand complexes)

| ddG, kcal/mol | N of structures | % of errors |
|---|---|---|
| <0.5 | 18 | 58 |
| 0.5-1.0 | 10 | 32 |
| >1.0 | 3 | 10 |

# Approaches to minimize scoring errors

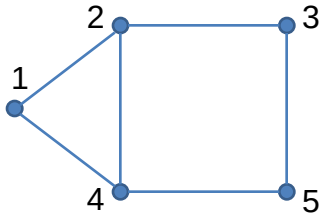
- Force field parameterization and refinement
  - *ab initio* approaches
- Scoring function customization
  - Protein- and/or ligand-based customization
  - Regime-based (docking, screening,…) customization
- Explicit treatment of the system
  - Protein flexibility
  - Explicit water
  - Free energy calculations

# Repairing scoring errors not related to scoring function

- Dynamical approaches (FEP, TI…)
  - Thermodynamic averaging along dynamic trajectory
- Stochastic approaches (Monte Carlo…)
  - Thermodynamic averaging along stochastic trajectory
- Graph-theoretical approach
  - Direct asymptotic evaluation of partition functions and thermodynamic averages

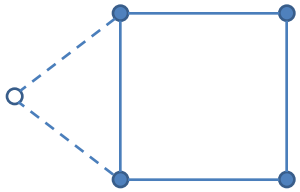
# TSAR – a new algorithm for multistate calculations

**T**hermodynamic **S**ampling of **A**mino acid **R**esidues

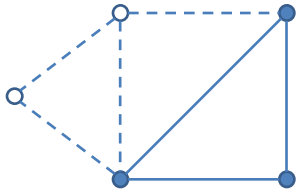



$$\sum_{N_{states}} e^{-E_{12345}}$$

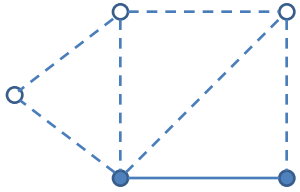
$$N_{states} = \prod_{i=1}^{5} N_i \sim N^5$$



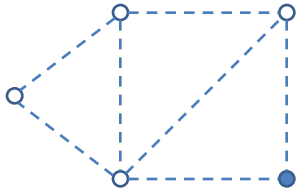
$$\sum_{N_{states}} e^{-E_{12345}} = \sum_{2345}\sum_{1} e^{-(E_1+E_{12}+E_{14})} \cdot e^{-E_{2345}} = \sum_{2345} W_{24} \cdot e^{-E_{2345}} \qquad N^3$$



$$\sum_{2345} W_{24} \cdot e^{-E_{2345}} = \sum_{345}(\sum_{2} W_{24} \cdot e^{-(E_2+E_{23}+E_{24})}) \cdot e^{-E_{345}} = \sum_{345} W_{34} \cdot e^{-E_{345}} \qquad N^3$$

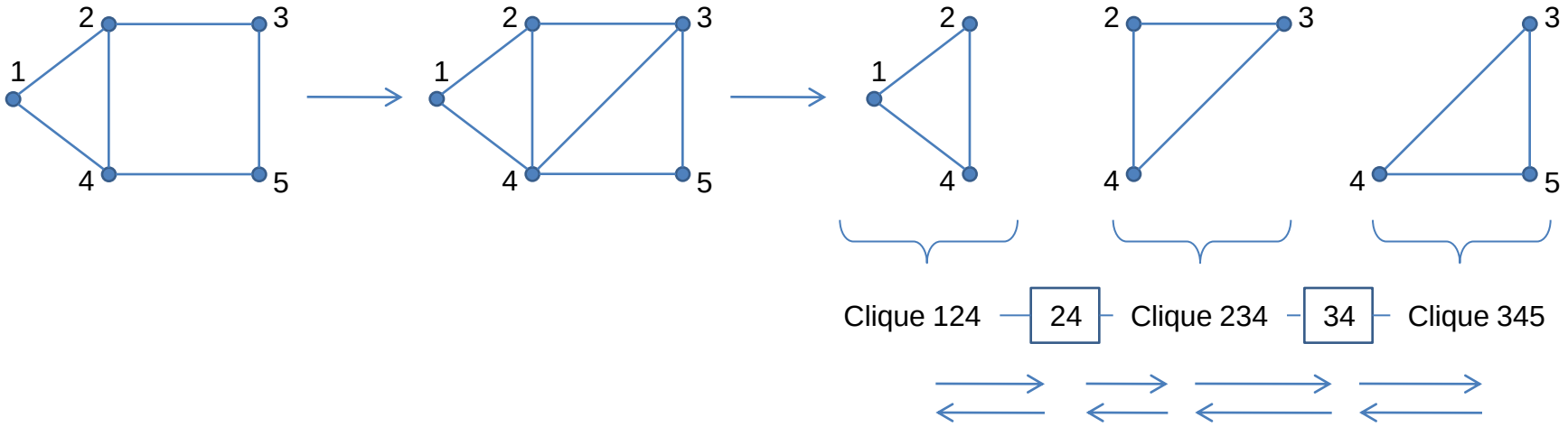


$$\sum_{345} W_{34} \cdot e^{-E_{345}} = \sum_{45}(\sum_{3} W_{34} \cdot e^{-(E_3+E_{35})}) \cdot e^{-E_{45}} = \sum_{45} W_{45} \cdot e^{-E_{45}} \qquad N^3$$



$$\sum_{45} W_{45} \cdot e^{-E_{45}} = \sum_{5}(\sum_{4} W_{45} \cdot e^{-(E_4+E_{45})}) \cdot e^{-E_5} = \sum_{5} W_5 \cdot e^{-E_5} \qquad N^2$$

# TSAR – a new algorithm for multistate calculations




$$W_{24} = \sum_{1} W_{124} = \sum_{1} e^{-(E_1+E_{12}+E_{14})} \cdot e^{-(E_2+E_4+E_{24})}$$

$$W_{234} = W_{24} \cdot e^{-(E_3+E_{32}+E_{34})}$$

$$W_{34} = \sum_{2} W_{234}$$

$$W_{345} = W_{34} \cdot e^{-(E_5+E_{53}+E_{54})} \rightarrow W_3, W_4, W_5$$

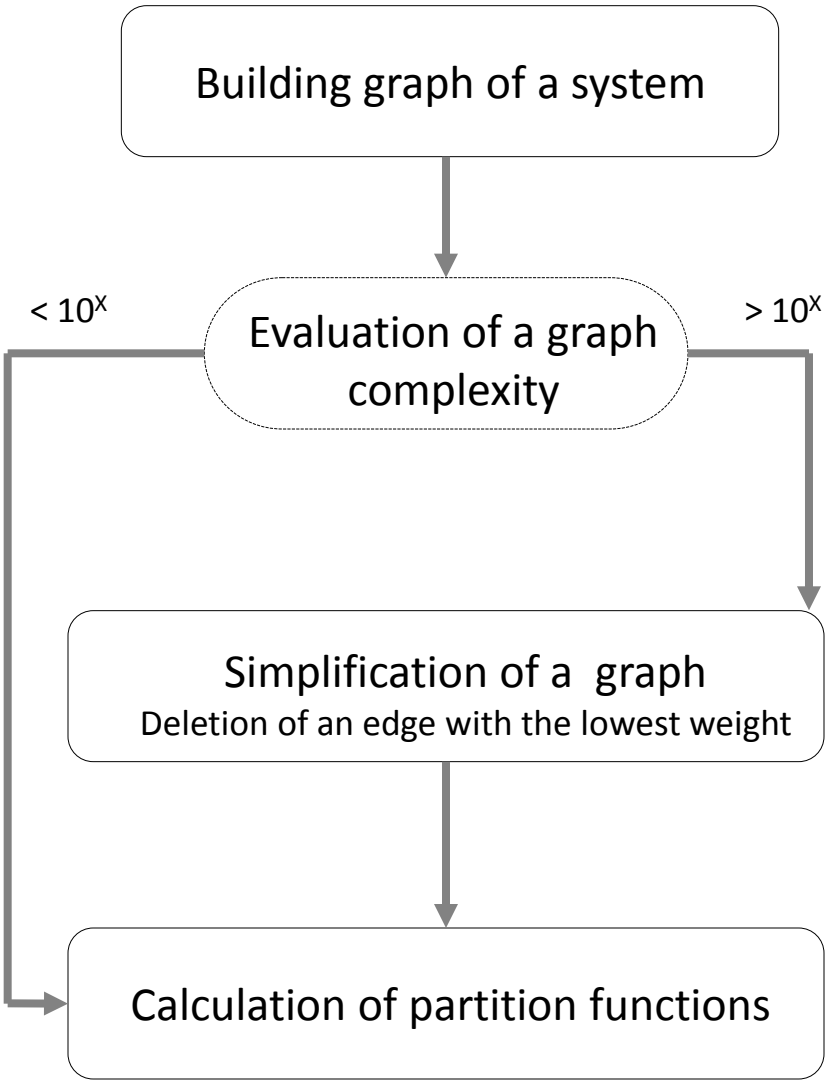
$$W'_{34} = \sum_{5} W_{345}$$

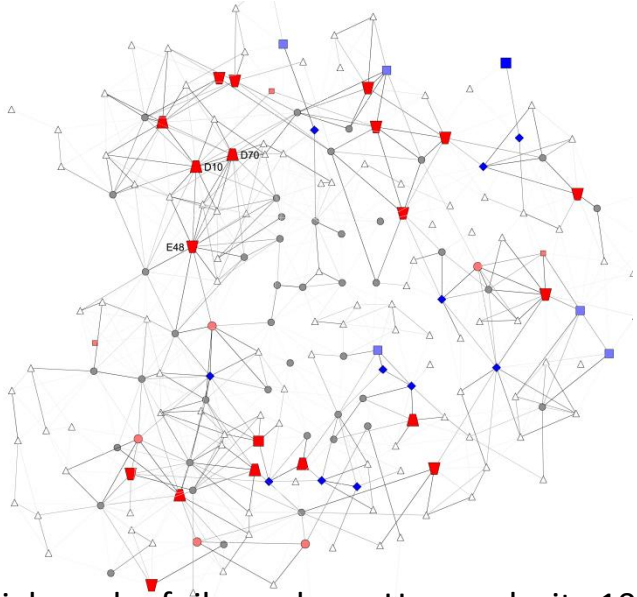
$$W'_{234} = W'_{34} \cdot e^{-(E_2+E_{23}+E_{24})} \rightarrow W_2$$

$$W'_{24} = \sum_{3} W'_{234}$$

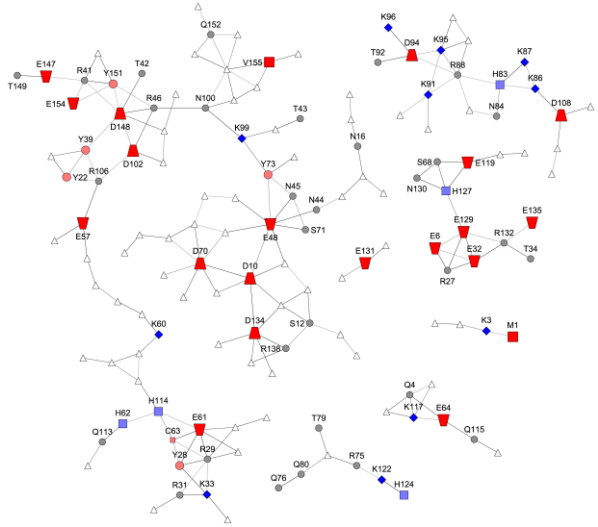
$$W'_{124} = W'_{24} \cdot e^{-(E_1+E_{12}+E_{14})} \rightarrow W_1$$

# TSAR – a new algorithm for multistate calculations



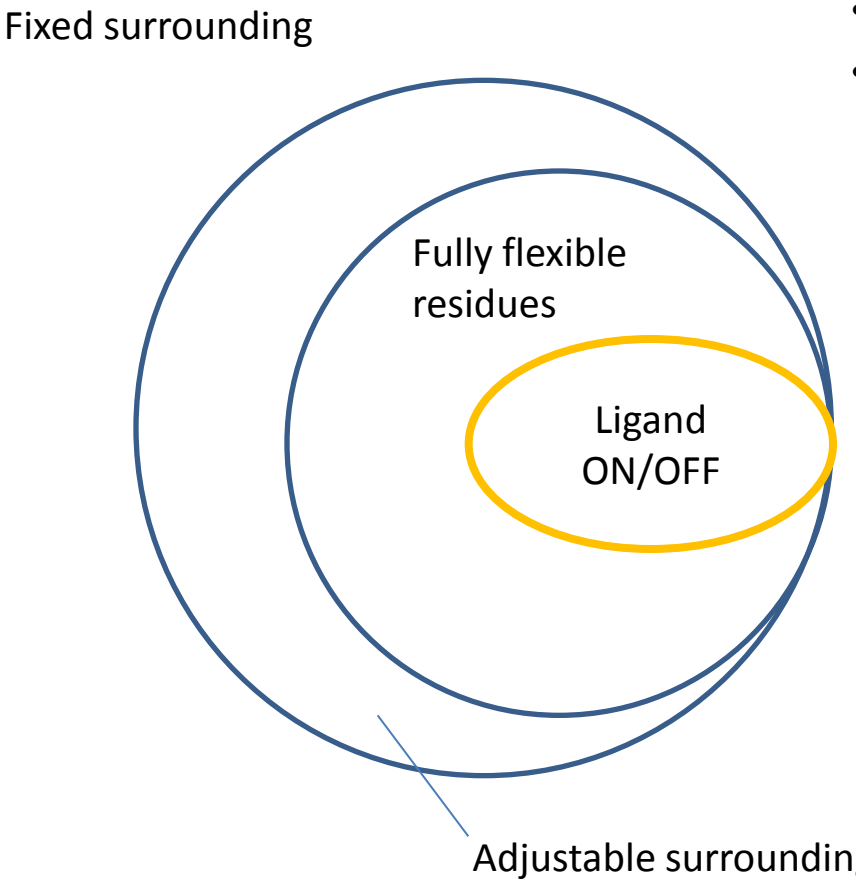

Initial graph of ribonuclease H, complexity $10^{166}$



Final graph, complexity $10^{7}$

# Application of TSAR to ΔG estimations

- ~$10^4$ Ligand states, local/global sampling
- ~$10^1$-$10^3$ States per Residue, local/global sampling
- Simplified scoring function for energy calculations



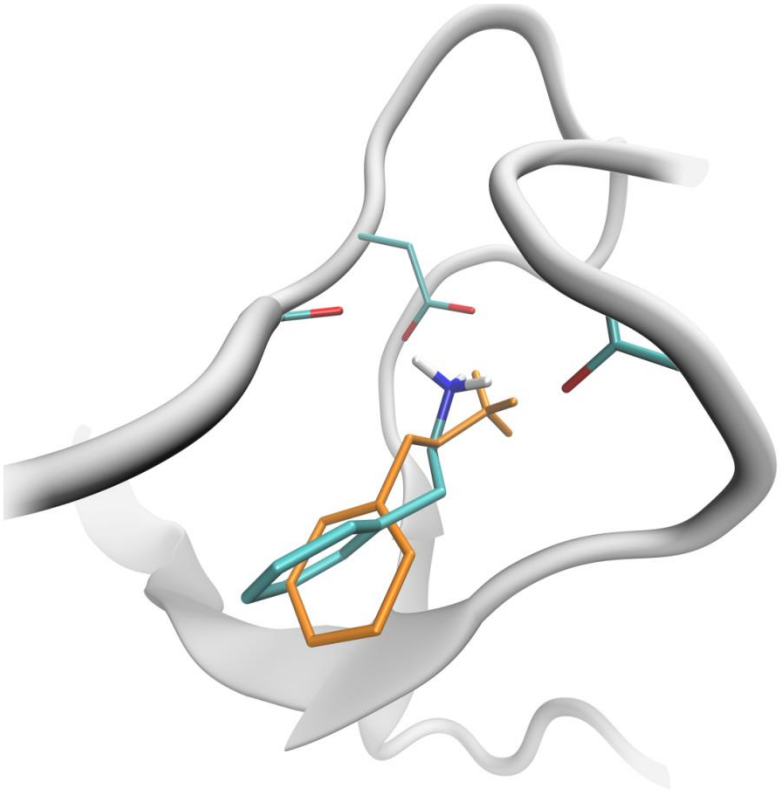

$$\Delta\Delta G = RT \ln \frac{e^{-E_{Optimum\_State}}}{\sum_{Ligand\_States} e^{-E_{State}}}$$

$$\Delta\Delta G = RT \ln \frac{\sum_{States\_Ligand\_ON} e^{-E_{State}}}{\sum_{States\_Ligand\_OFF} e^{-E_{State}}} - E_{Optimum\_State}$$

$$\Delta G = E_{Docking} + \alpha \cdot \Delta\Delta G$$

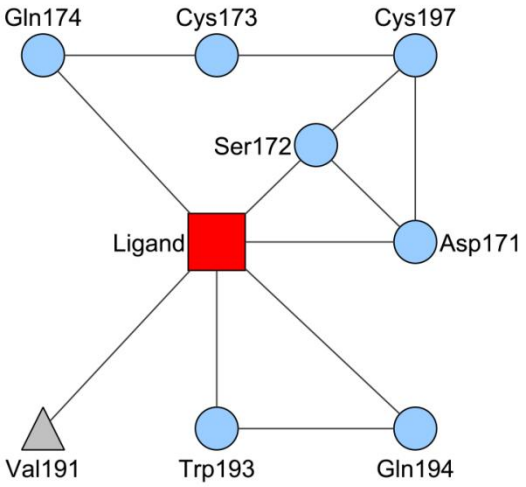
# Application of TSAR to ΔG estimations

Trypsin inhibitor (PDB 1tnj)



Protein-Ligand ΔG of binding, (kcal/mol):

- Experimental -2,7*
- Calculated -5,4**
- TSAR ΔΔG = 0,22 (ligand flexibility)
  = 0,7 (active site flexibility)



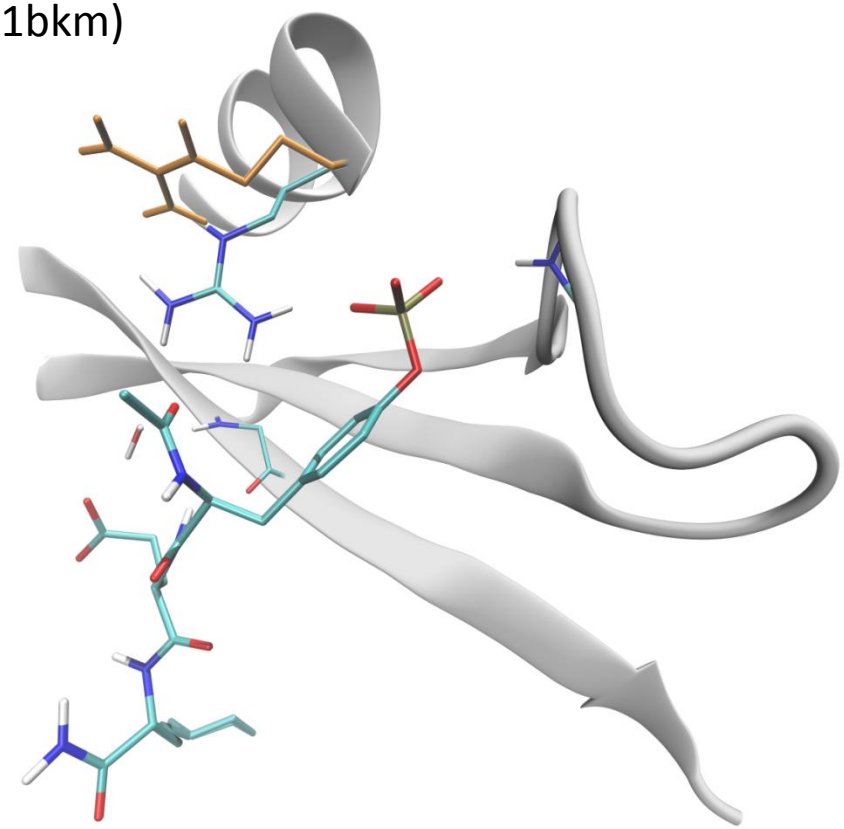

Complexity of a graph ~$10^7$

* Nat. Struct. Biol. 1994, 1(10), 735-743

** Lead Finder v. 1.1.14

# Application of TSAR to ΔG estimations

Ligand bound to SH2 domain of v-Src (PDB 1bkm)



Protein-Ligand ΔG of binding, (kcal/mol):

- Experimental -7,9*
- Calculated -10,8**
- TSAR ΔΔG =

-11,2 (ΔG with simplified scoring function) –

-18,2 ( $E_{optimum}$ with simplified scoring function)

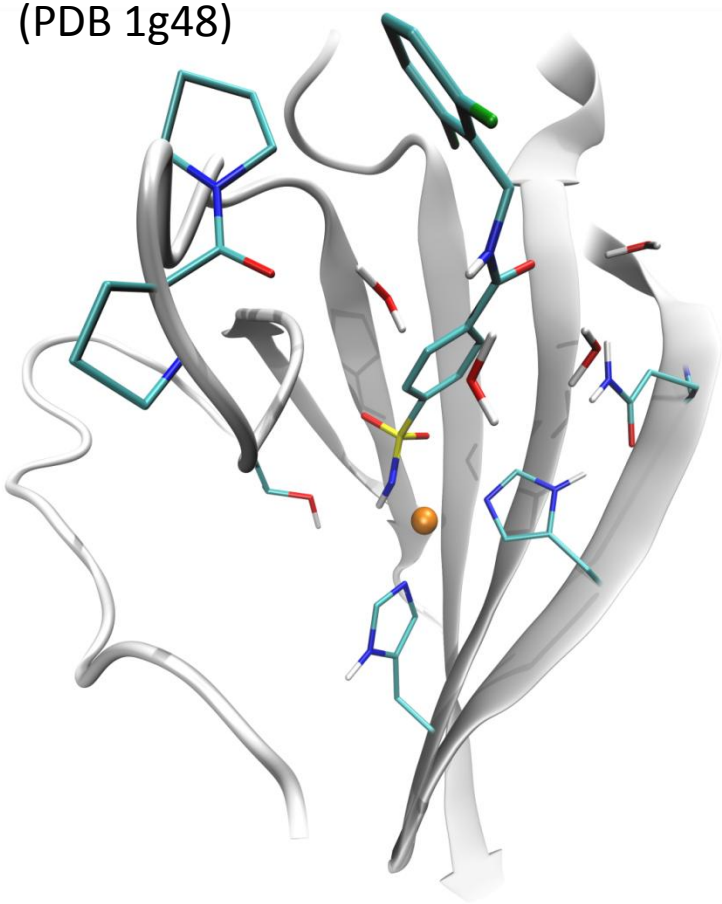
= 7 (~40% of $E_{optimum}$)

* J. Am. Chem. Soc. 1997, 119, 12471-12476

** Lead Finder v. 1.1.14

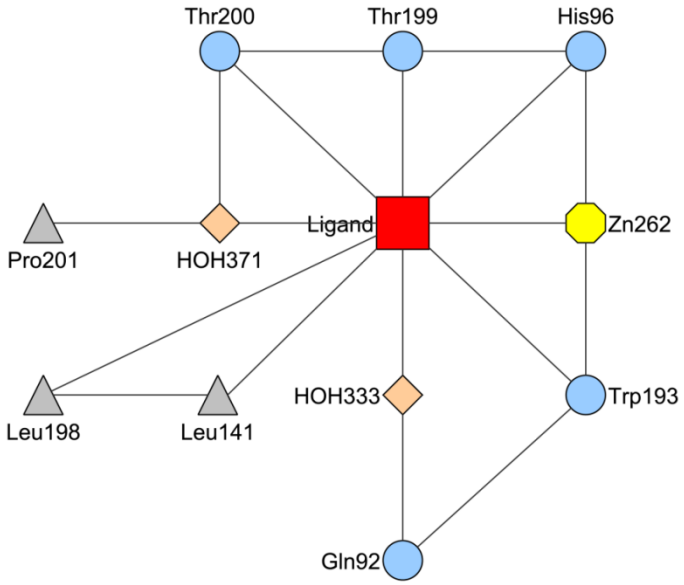
# Application of TSAR to ΔG estimations

Carbonic anhydrase II with bound inhibitor (PDB 1g48)



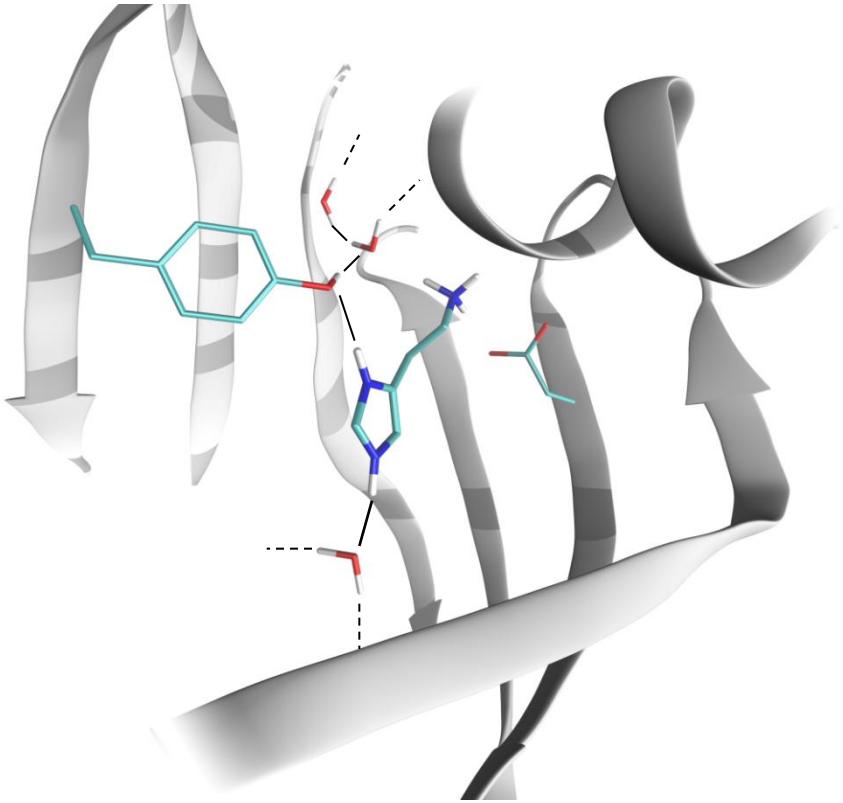
Complexity of a graph:

- before reduction ~$10^{16}$
- after reduction ~$10^{7}$



# What about energy underscoring?

Histamine bound to lipocalin (PDB 1bu1)



Protein-Ligand $\Delta G$ of binding, (kcal/mol):

- Experimental -11,3*
- Calculated**

  -5,9 (no explicit water)

  -9,1 (with explicit water)

* J. Biol. Chem. 2008, 283(27), 18721-18733

** Lead Finder v. 1.1.14

# Further improvements of thermodynamic averaging approach

- More accurate scoring function
- Explicit treatment of water
  - mediating ligand binding
  - displaced by ligand
- Loss of ligand's degrees of freedom