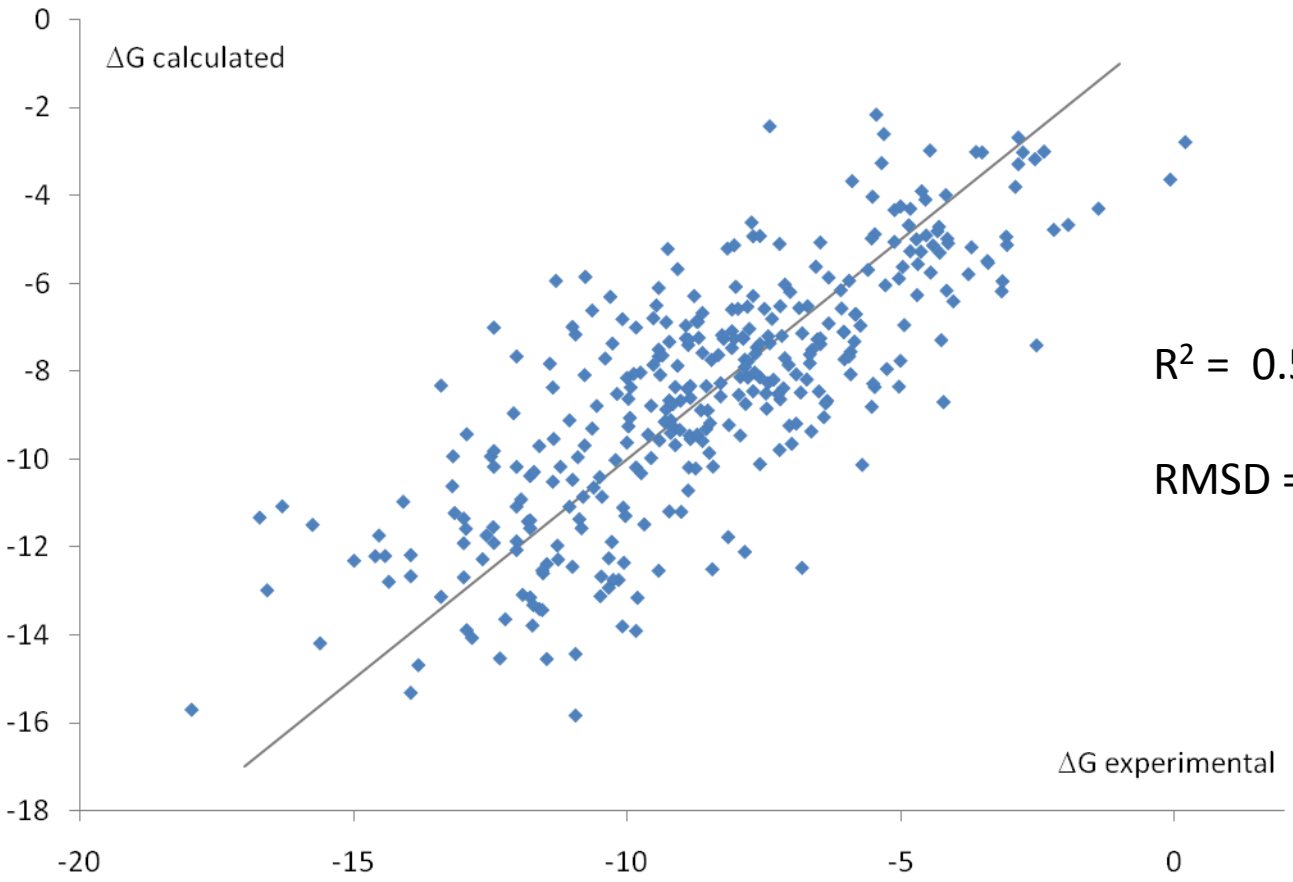
Lead Finder in CSAR scoring challenge

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 23 August 2010

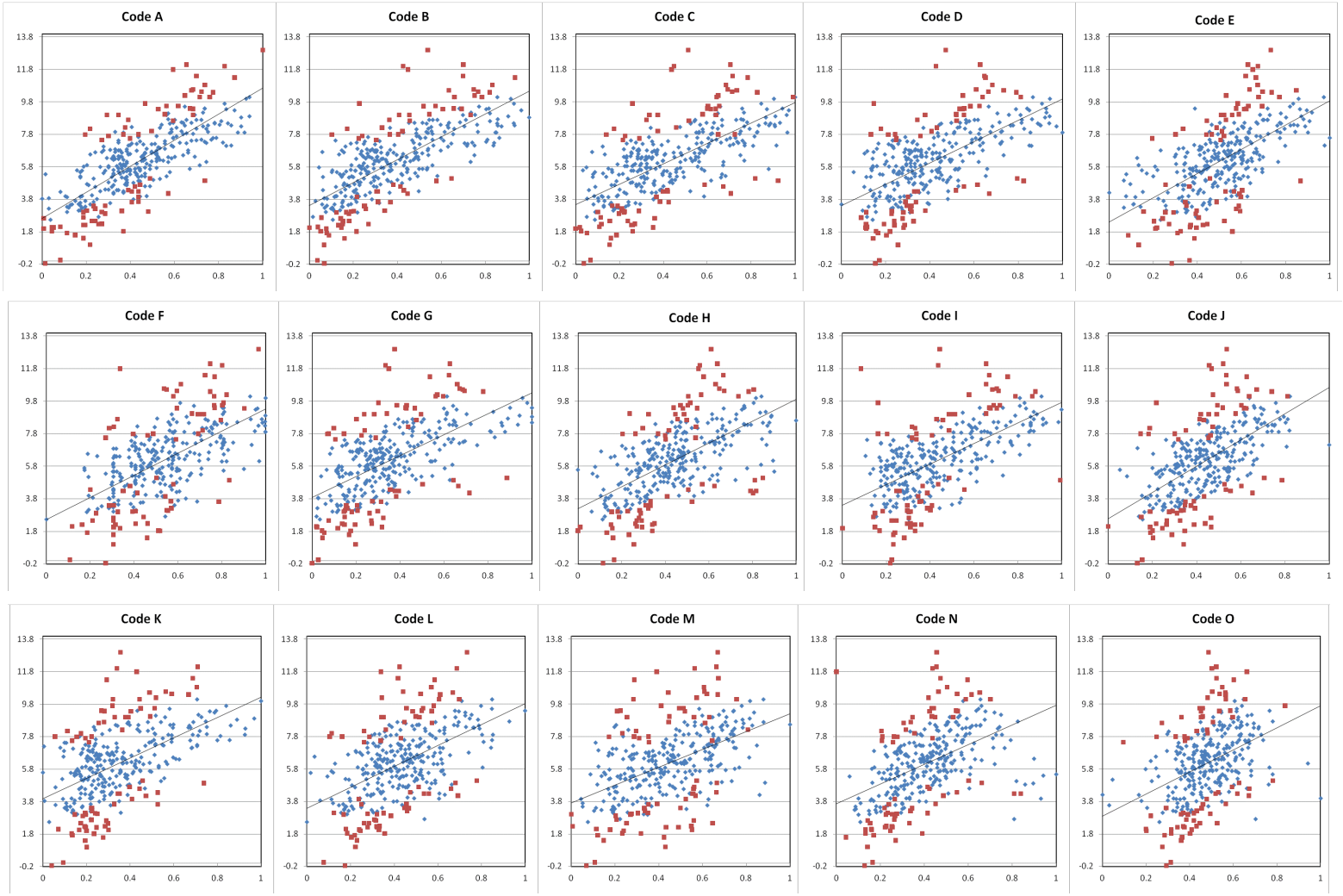
Lead Finder in CSAR scoring challenge



R^2 = 0.58

RMSD = 1.98 kcal/mol

Scoring performance in CSAR challenge



Outline of the presentation

- Basic ingredients
  - Van der Waals and solvation
  - Electrostatics
  - Hydrogen bonds
- Magic ingredients
- Where do we go from here?

The Forcefield scoring functions (in Lead Finder)

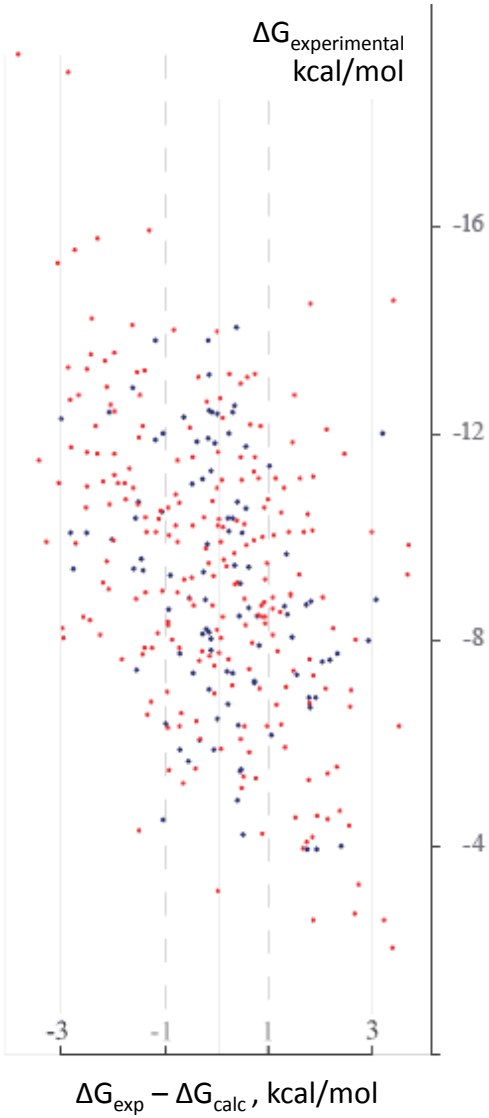
AMBER, OPLS, CHARMM etc.

- van der Waals energy,
- Electrostatics,
- Hydrogen bonds,
- Dihedrals energy,
- Bonds, Angles
- Solvation

+ Magic ingredient

= best scoring function ever!

How to brew a scoring function: step 2



$$\Delta G = k_{VdW} E_{VdW} + k_{Elec,i} E_{Elec,i} + k_{Hbonds,i} E_{Hbonds,i} + \ldots$$

1 for Van der Waals energy,
4 for electrostatics ,
5 for hydrogen bonds,
1 for interaction with metals,
5 for Solvation,
4 for internal energy

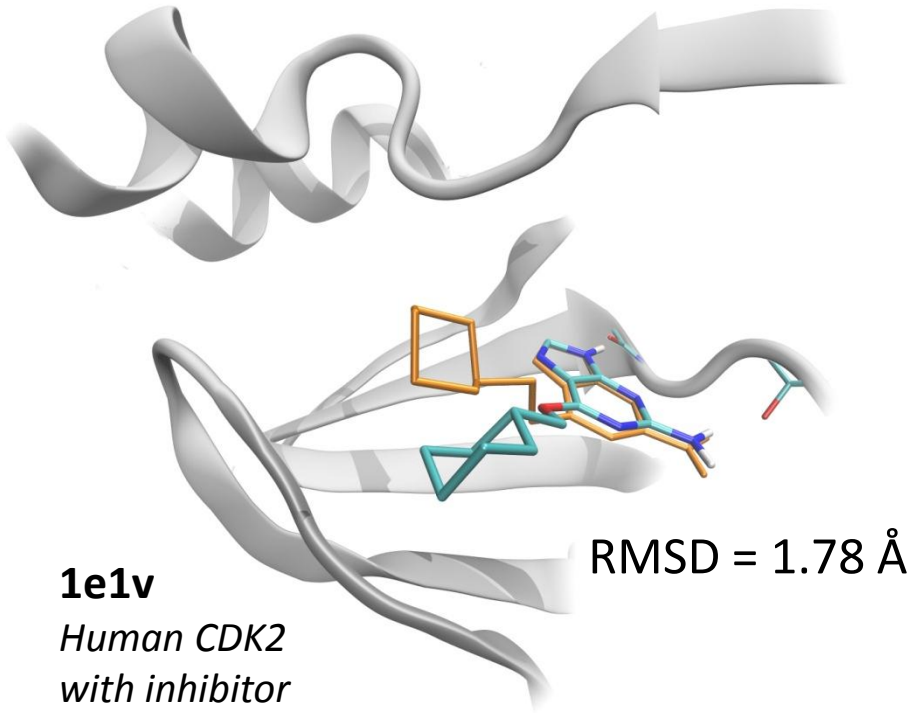
} 20 coefficients

Training set: 230 structures (blue dots)

Test set: 100 structures (red dots)

RMSD of ΔG = 1.75 kcal/mol

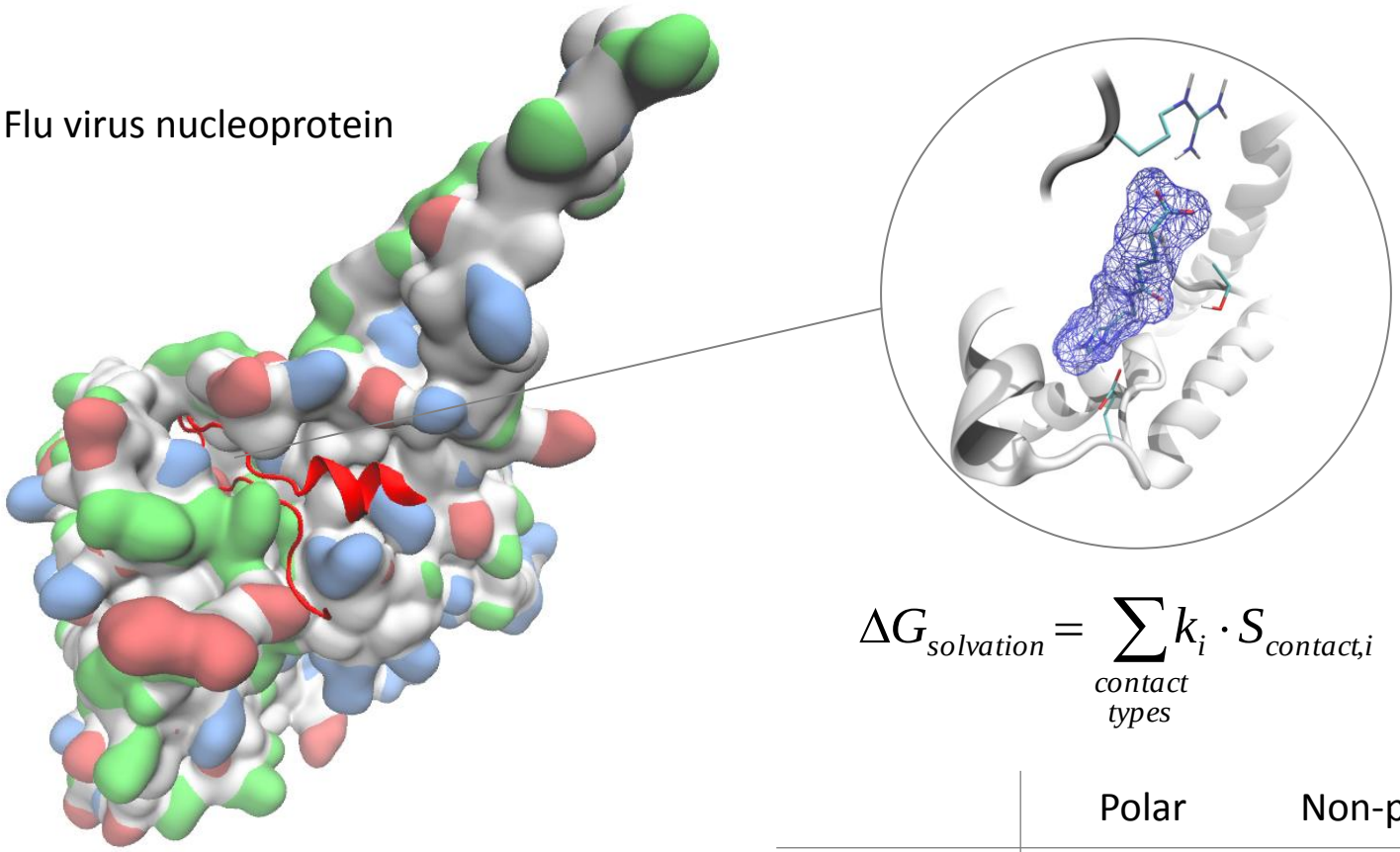
Why do we need van der Waals energy?



- VdW-guided global search (docking)
- Optimization of given ligand poses
- Energy of the contact between ligand and protein

Solvation free energy

Flu virus nucleoprotein

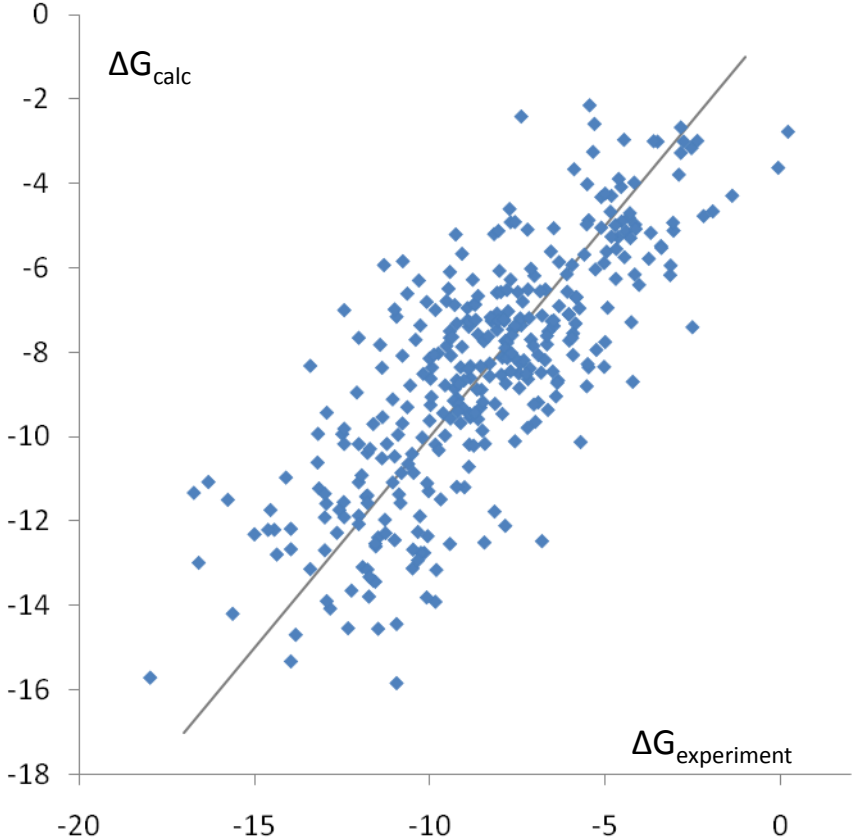


$$\Delta G_{solvation} = \sum_{\substack{contact \\ types}} k_i \cdot S_{contact,i}$$

	Polar	Non-polar
Protein	-0.25	-0.40
Solvent	0.30	-0.01

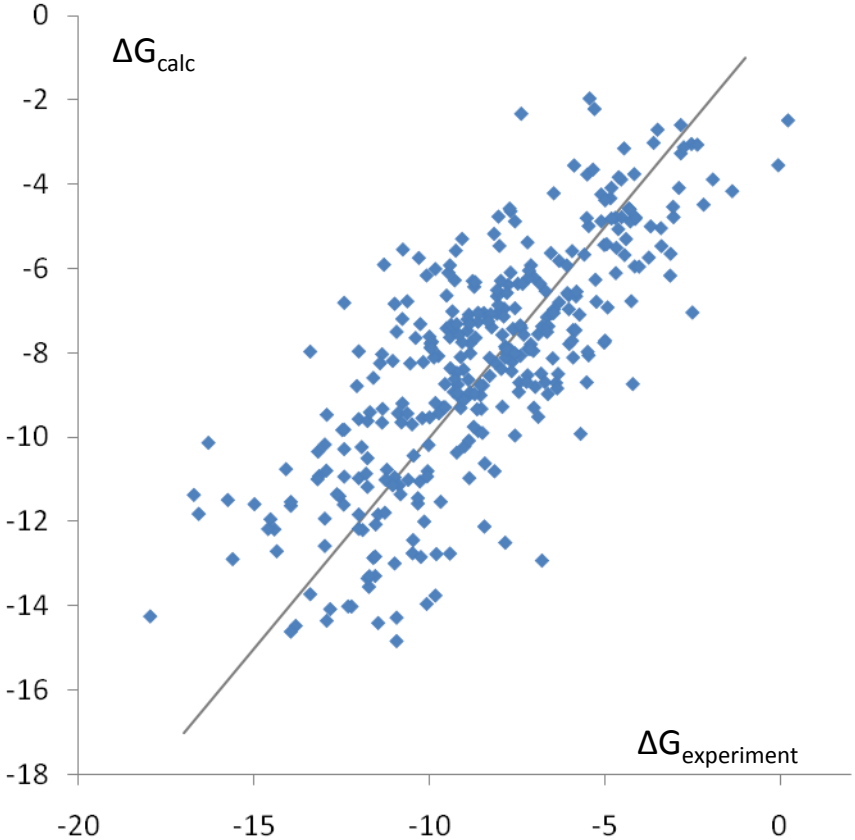
Solvation and VdW energy are interchangeable

$k_{VdW} = 1, \quad k_{Sol} = 1$



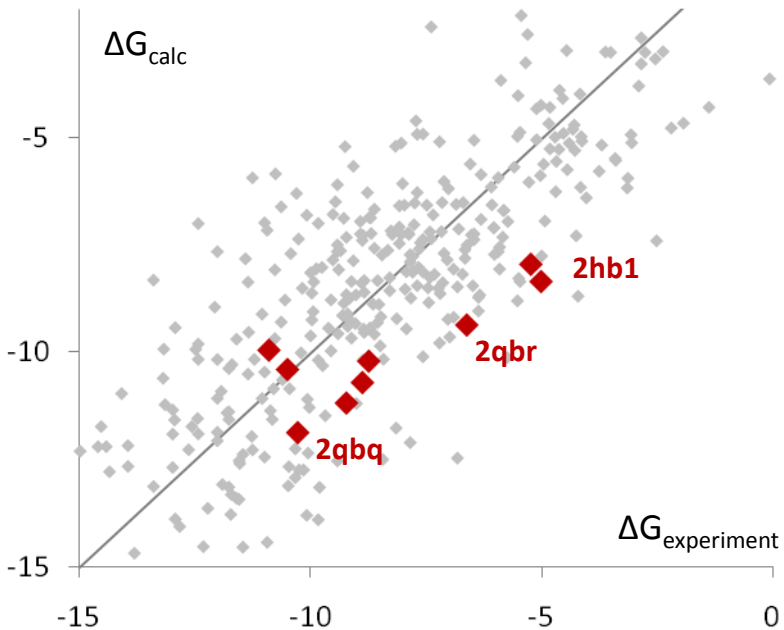
RMSD = 1.98 kcal/mol
R^2 = 0.580

$k_{VdW} = 0.5, \quad k_{Sol} = 1.7$



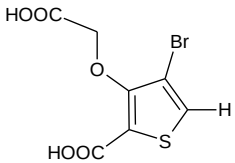
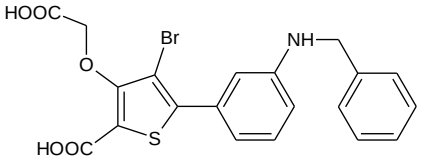
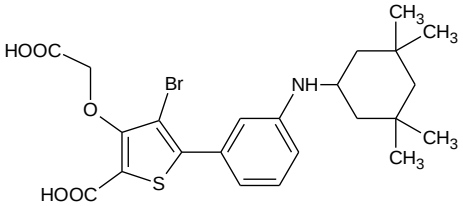
RMSD = 2.06 kcal/mol
R^2 = 0.57

Solvation works!

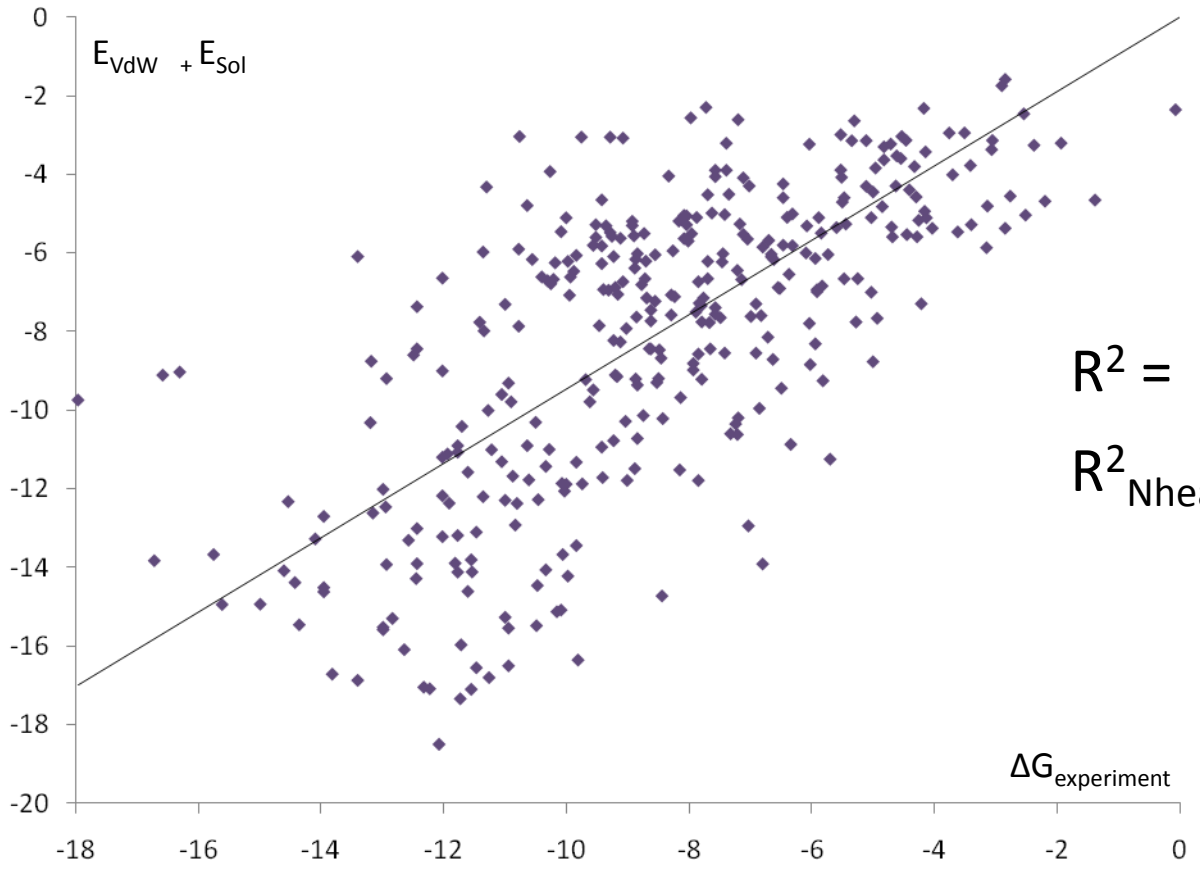


Model	R^2
Full model	0.68
Solvation + VdW	0.73
N(heavy)	0.80
N(all atoms)	0.87

Tyrosine protein phosphatase type 1

		ΔG_{exp} kcal/mol	$E_{sol} + E_{VdW}$
2hb1		-5.3	-6.7
2qbr		-8.7	-10.1
2qbq		-10.3	-11.0

Solvation explains almost everything?



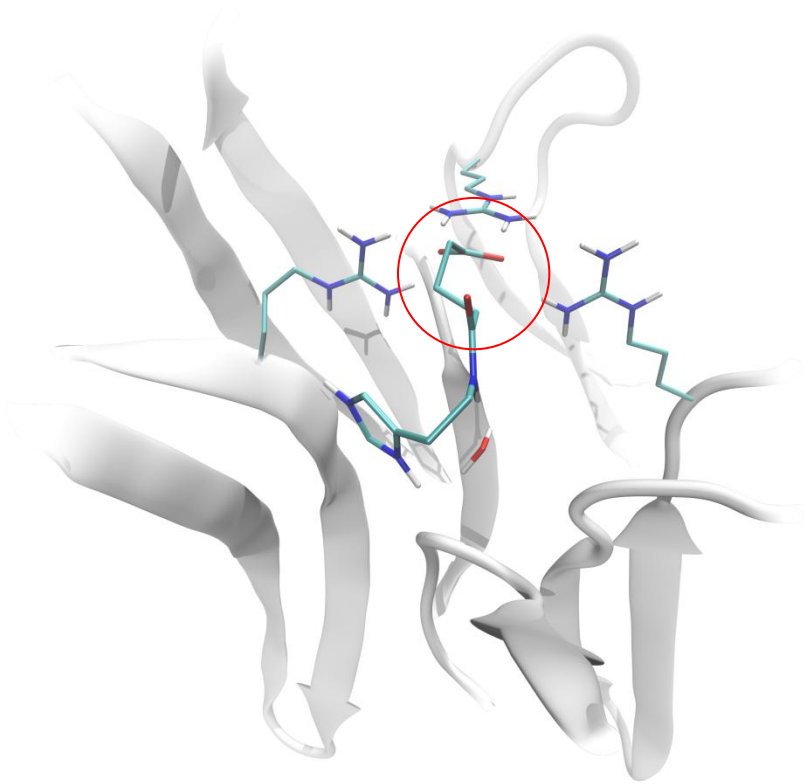
$R^2 = 0.47$

$R^2_{Nheavy} = 0.26$

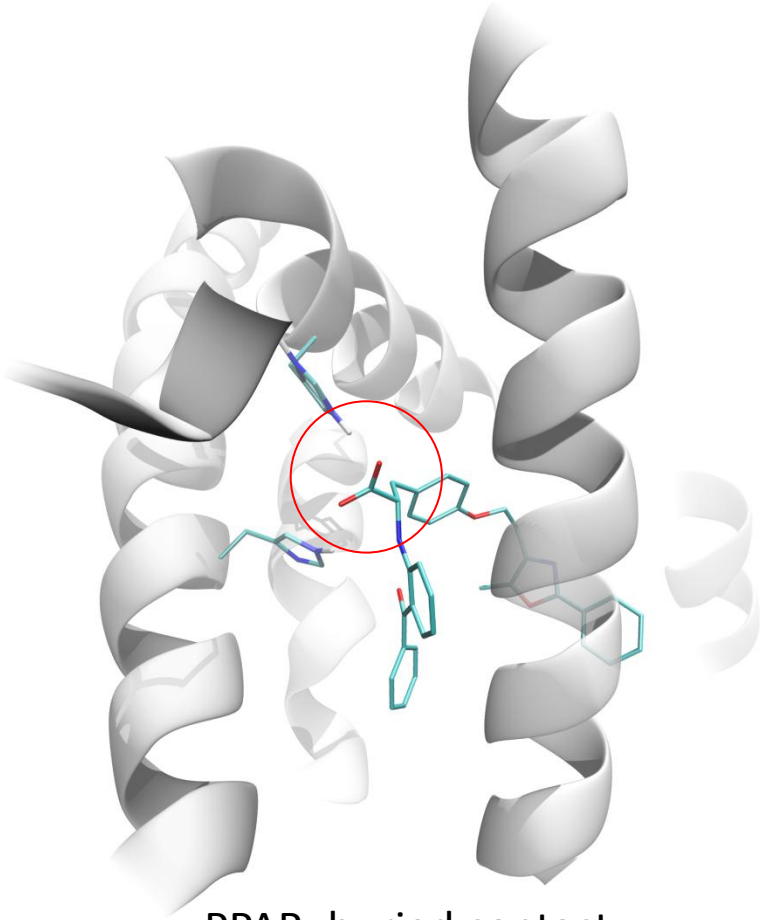
Electrostatics pitfalls

- Long range interactions
  - Slowness of interaction energy decrease
  - Dependence of dielectric permittivity on (micro)environment
- Short range interactions
  - Calculations of atomic charges on ligand and protein
  - Polarization of interacting atoms
  - Competition between electrostatics and explicit interactions (h-bonds)
- Common pitfalls
  - Sampling of spatial distribution of charges
  - Sampling of ionization states of protein and ligand

Electrostatics in Lead Finder

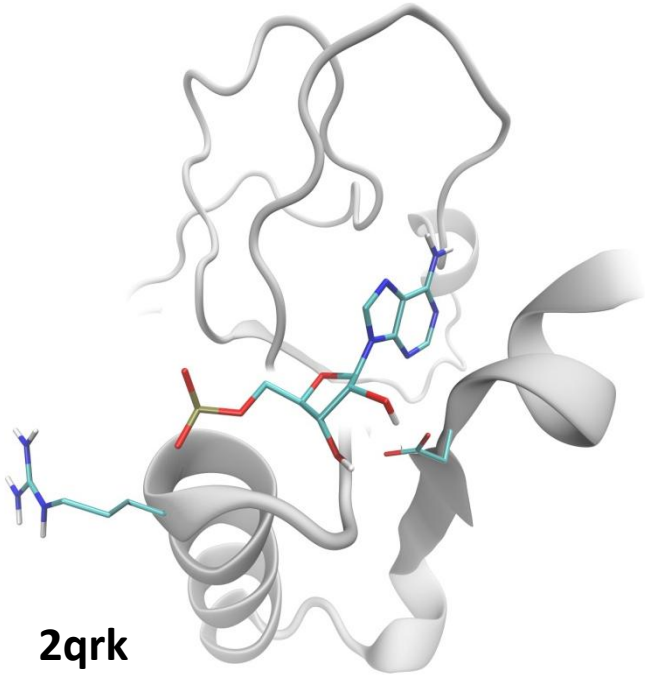


Neuraminidase: surface contact



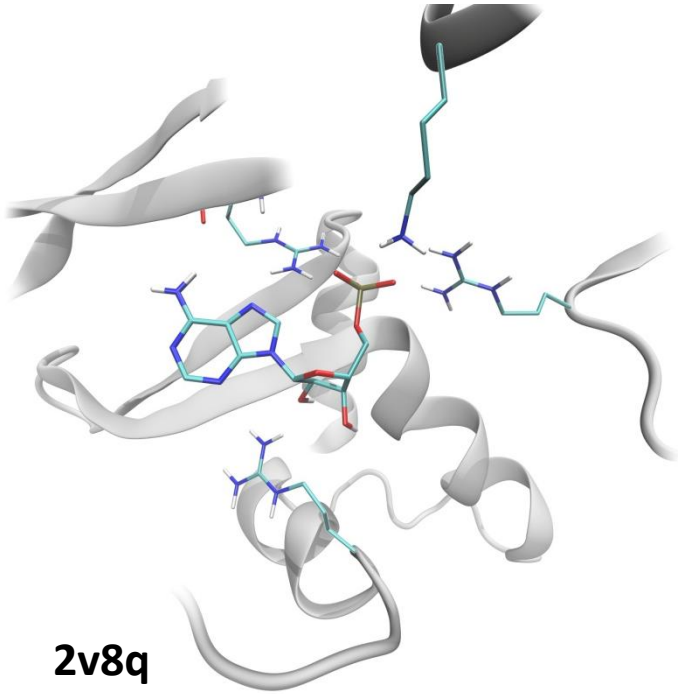
PPAR: buried contact

Electrostatics doesn't always work…



2qrk
Saccharopine dehydrogenase

ΔG_{exp} = -5.9 kcal/mol
ΔG_{calc} = -3.7 kcal/mol



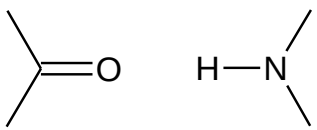
2v8q
Glycogen-binding domain of AMP-activated kinase beta2

ΔG_{exp} = -6.4 kcal/mol
ΔG_{calc} = -8.7 kcal/mol

H-bonds penalties and rewards

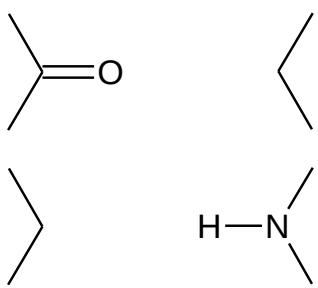
$$\Delta G_{HB} = \Delta G_{HB,complex} - \Delta G_{HB,solution}$$

Ligand Protein



$$E_{H-bond} = E_0(r_{HA}) \cdot k_{DHA} \cdot k_{LP}$$

For the most cases $\Delta N_{hbonds} \approx 0$

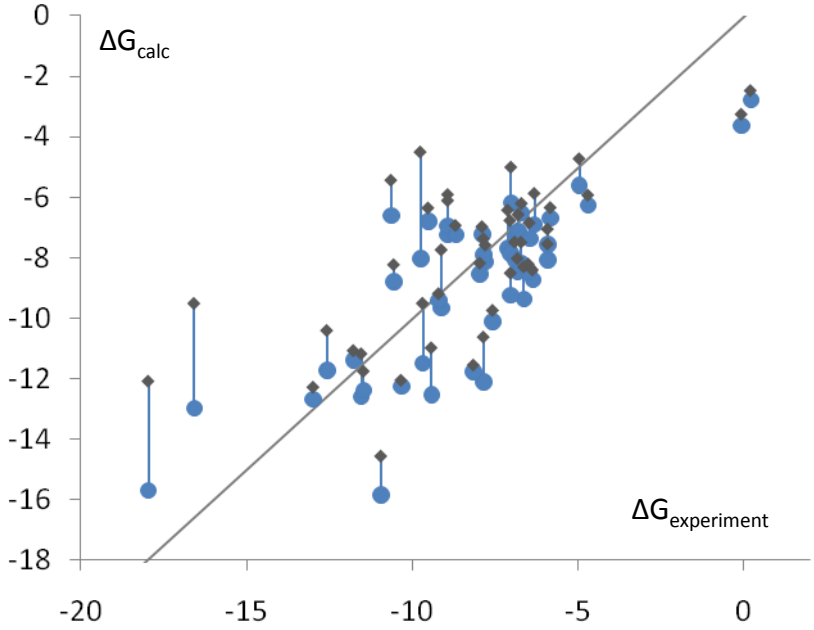


$$E_{penalty} = k \cdot N_{lost,ligand}$$

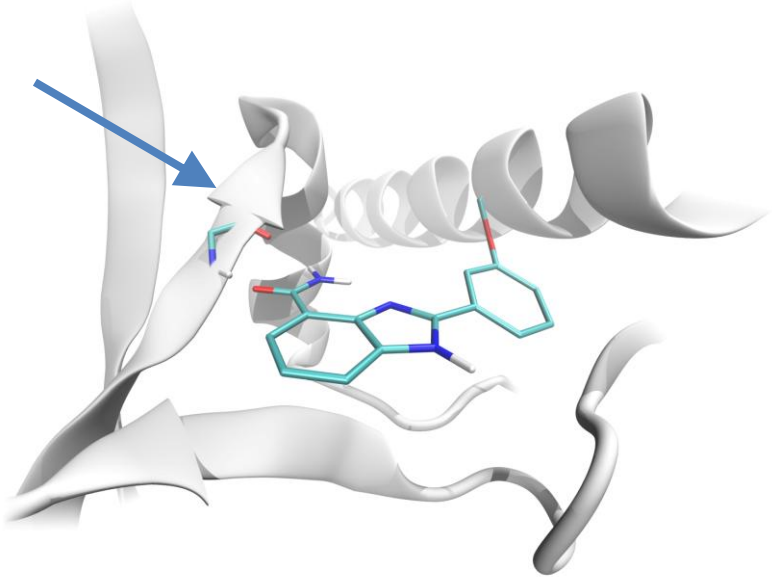
$$E_{penalty} = k \cdot N_{lost,protein}$$

H-bonds penalties serve to sieve out bad poses and poor binders

H-bonds extra energy



PARP with inhibitor



R^2 without extra H-bonds = 0.47
R^2 with extra H-bonds = 0.62

on CSAR subset of 48 structures, where systems of correlated H-bonds were found

Quest for new molecular interactions

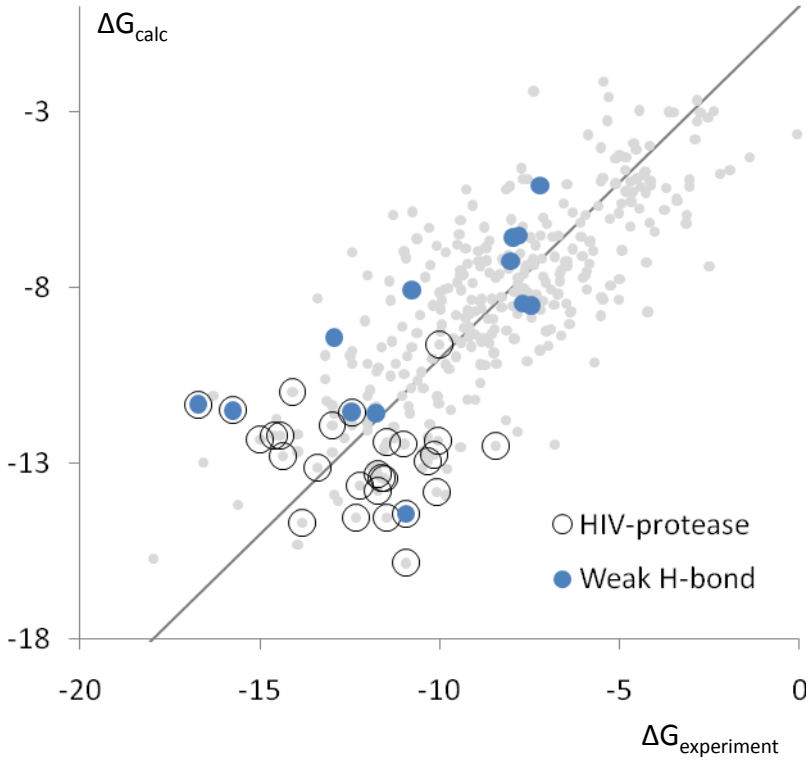
- Thoroughly inspect complexes with discrepancies between experimental and calculated free energies
- Point out “interaction X”
- Estimate energy of the interaction
- **Add interaction to the program, avoiding overfitting and false positives**

Weak & rare interactions

- Weak hydrogen bonds
  - Aromatic rings as hydrogen bonds acceptors
  - Polarized C-H bond (Cα)
  - F as acceptor: CF ··· HX (O,N)
- Specific halogen interactions
  - Orthogonal multipolar interactions (C-X ··· C=O)
  - Interactions of halogens with nuclephils and electrophils
- Specific aromatic contacts
  - π-cationic interactions
  - Specific orientations

J. Med. Chem., 2010, 53, 5061-5084

Weak hydrogen bonds



CSAR has 13 cases of weak H-bonds (Hα),
Average O-Hα distance is 2.15 Å

<ΔΔG> = -1.3 kcal/mol

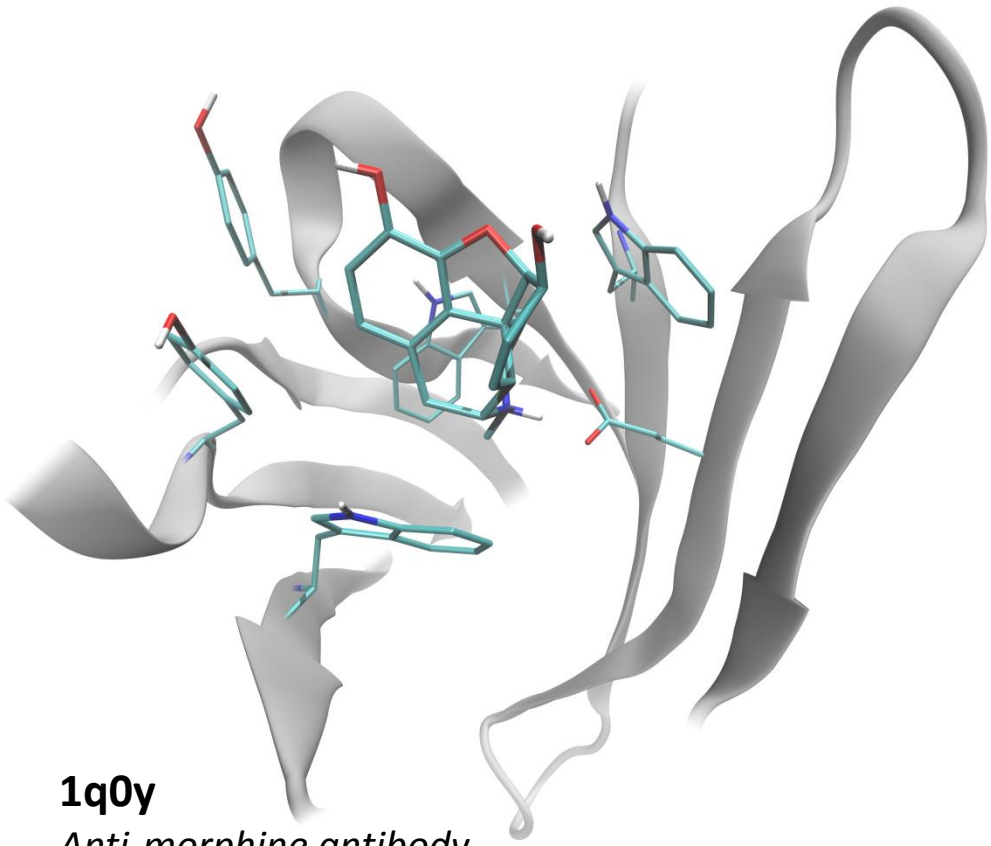
Halogen interactions

Halogen	N structures	Error, kcal/mol
F	29	+0.6
Cl	21*	+0.1
Br	7**	+1.3
I	1	+2.6

* 10 of 21 structures with Cl are coagulation factor X with inhibitors. R^2 within this subset is 0.8

** 6 of 7 structures with Br are tyrosine protein phosphatase type 1 with inhibitors

Stacking and π-cationic interactions

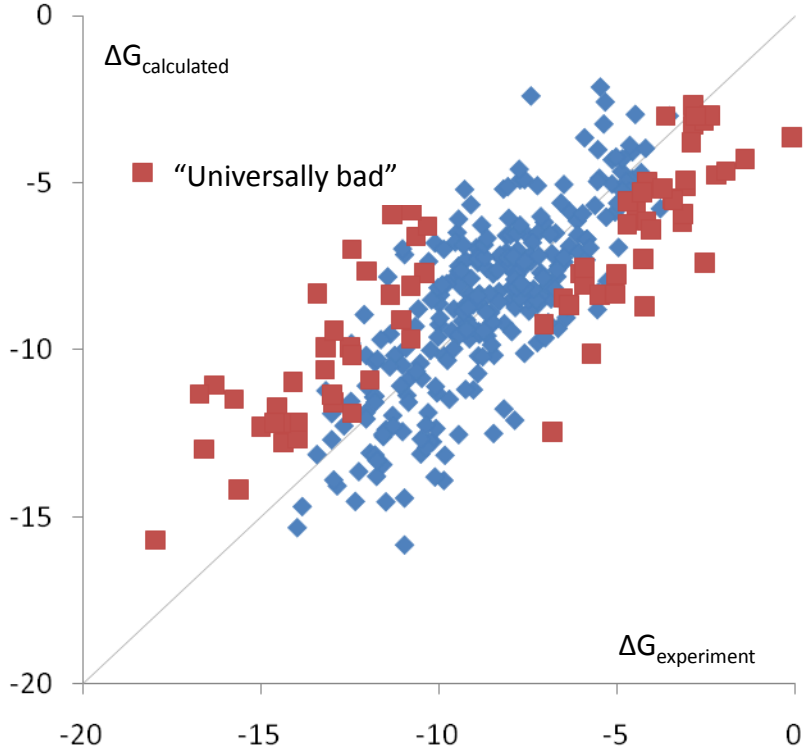


1q0y
Anti-morphine antibody complexed with morphine

ΔG_{exp} = -12.4 kcal/mol

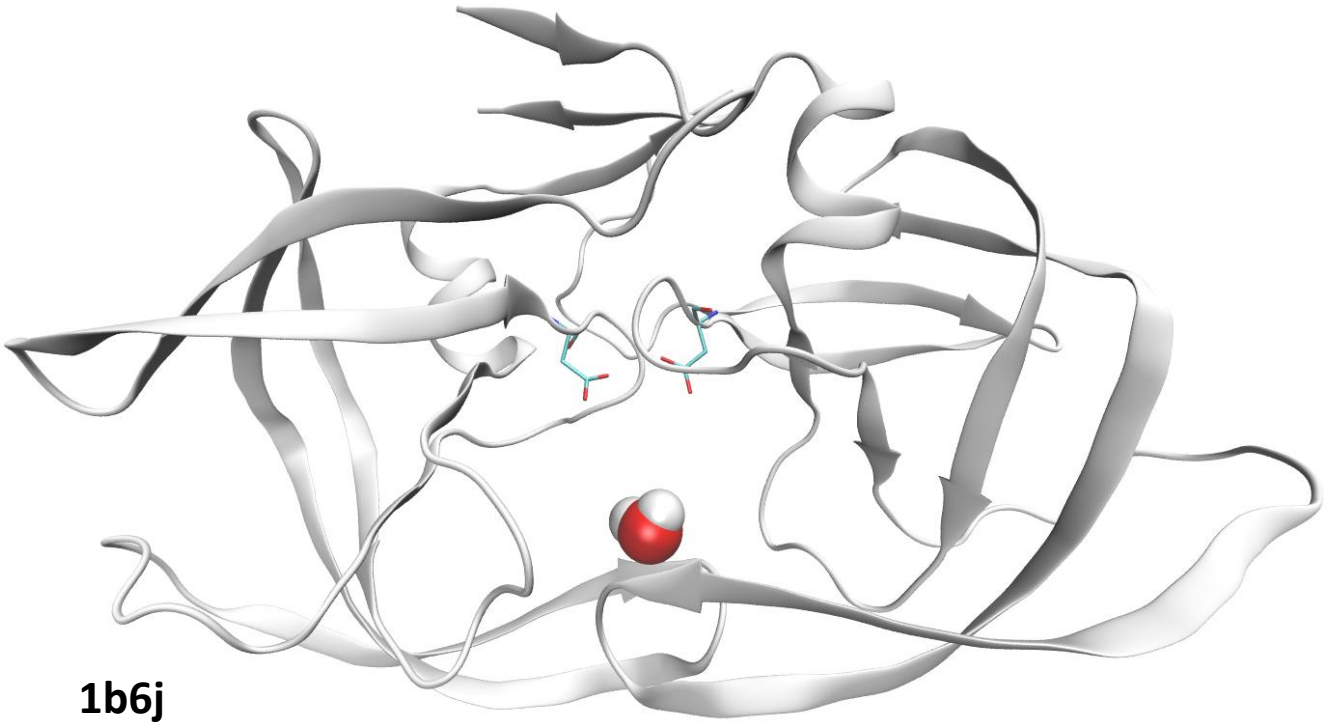
ΔG_{calc} = -7.0 kcal/mol

Are we missing something?



PDB id	Error, kcal/mol	
1duv	5.2	parameterization
2c1q	2.3	biotin
2i0d	5.4	HIV-protease
2qi5	2.7	HIV-protease
2qi6	2.4	HIV-protease
2fv5	2.8	??
1swk	3.6	biotin
1y1m	-4.9	protein conformation?
1y1z	-3.0	protein conformation?
…		

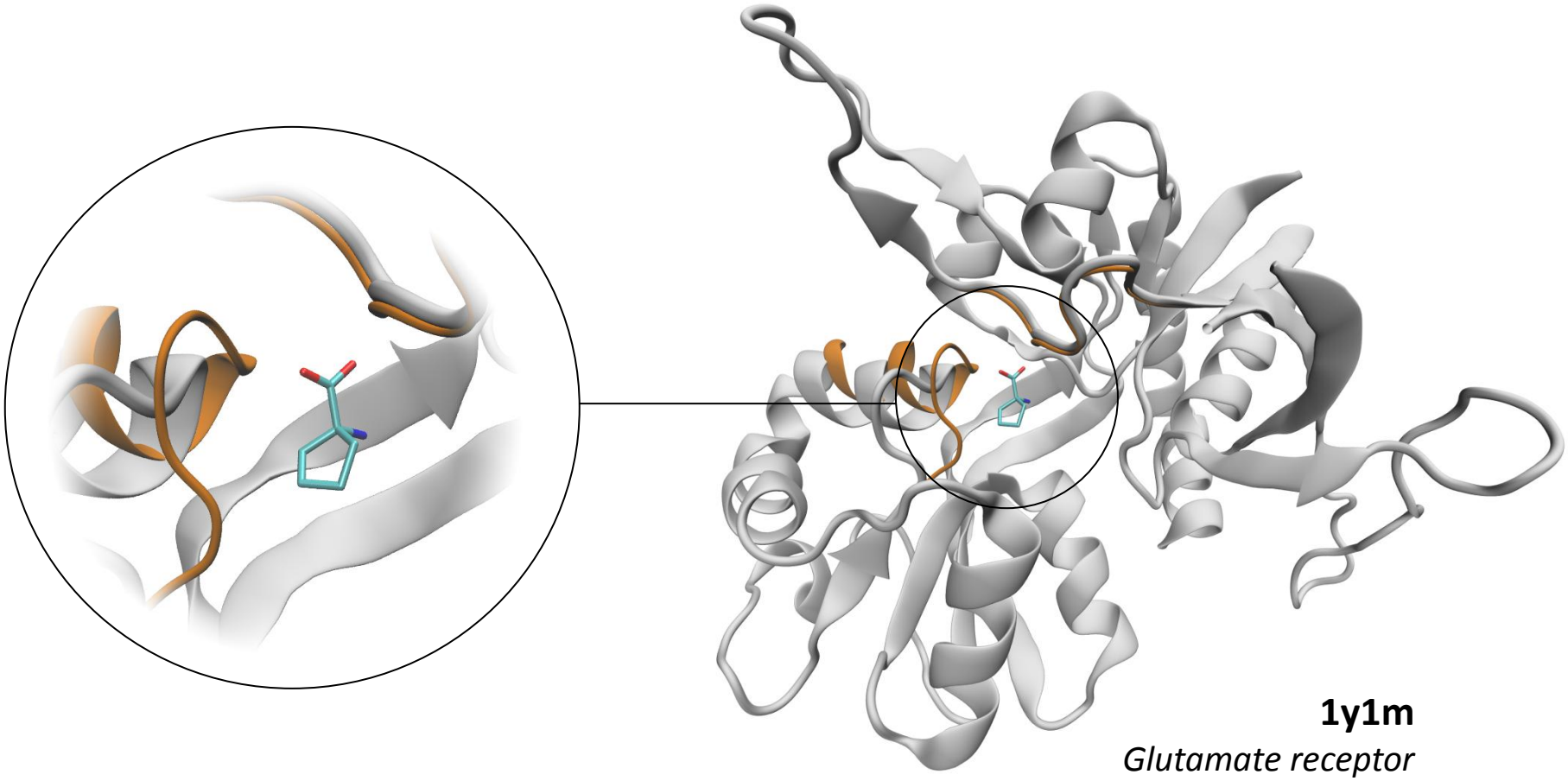
Explicit water



1b6j
HIV protease

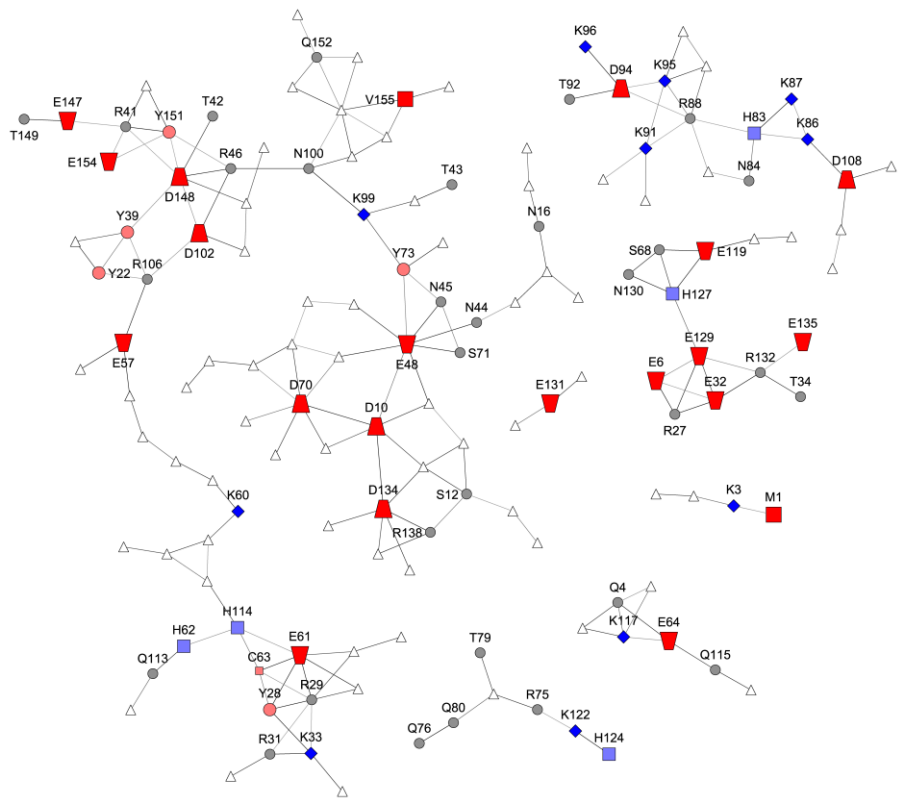
26 of 28 HIV protease inhibitors from CSAR set interact with conservative water molecule

Loops and sidechains flexibility



TSAR – a new algorithm for multistate calculations

Thermodynamic **S**ampling of **A**mino acid **R**esidues



Simplified Interactions graph for ribonuclease H (blue – Lys, His; red – Asp, Glu)

- Represent interactions between residues as graph
- Invoke belief networks theory to reduce complexity of graph
- Find global minima using Dead-End Elimination technique

Or

- Calculate energy difference between states

$$\Delta G = RT \ln \frac{\sum_{ligandenabled} e^{-E_i/RT}}{\sum_{liganddisabled} e^{-E_i/RT}}$$

Future directions of mastering scoring

- Improvements of sampling
  - Thermodynamic integration over ligand and protein conformations
  - Sampling of flexible loops
- Explicit treatment of water
  - Conservative water molecules
  - Replaced by ligand
  - Water networks rearrangements energy evaluation