TSAR — a new graph-theoretical approach to computational modeling of ionization properties of proteins

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

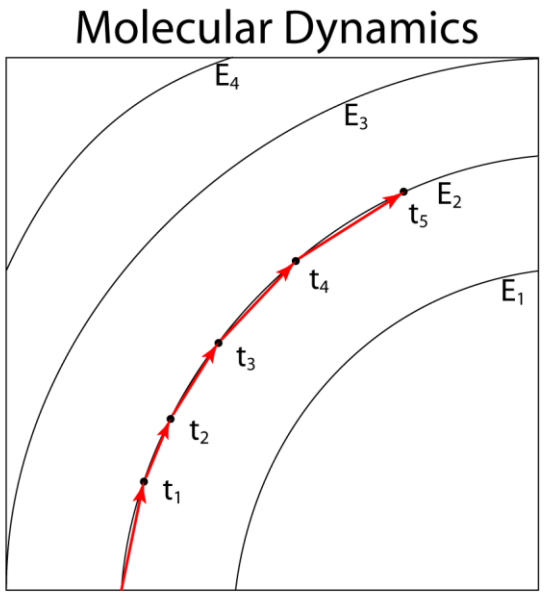
ACS 241st Meeting, Anaheim, USA 27 March 2011

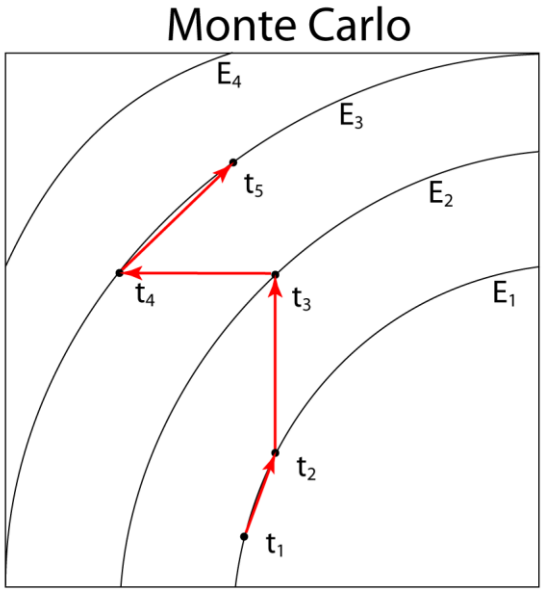
Knowing state space of macromolecule is crucial for understanding its properties

- Protein structure, ionization properties, stability
- Protein-ligand and protein-protein binding
- However… the number of different conformational states is astronomic even for small proteins
  - Protein of 150 residues, 20 ionizable residues
  - 10 conformations/residue
  - totally $\mathbf{10^{156}}$ conformational states
  - $\mathbf{10^{6}}$ distinct charged states

Current approaches to explore state space of molecular systems

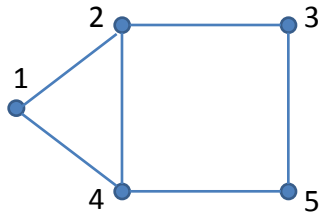
- Dynamical approaches (MD…)
- Stochastic approaches (Monte Carlo…)





What about direct enumeration of system states?

- Direct enumeration of system states
  - allows global sampling of system state space with a given (space) resolution
  - is practically limited to 10^6-10^{12} (molecular mechanical) energy evaluations



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

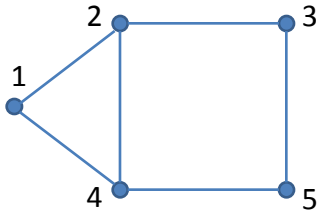
$$N_{states} = \prod_{i=1..5} N_i \propto N^5$$

Novel graph-theoretical approach for multistate calculations

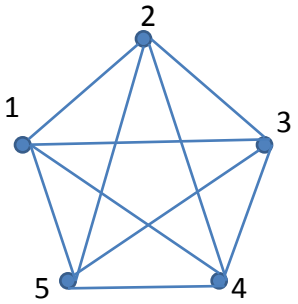
- Account of a system topology allows to make enumeration of system states optimal

Actual number of terms in statistical sum:

N^2



N^3



N^5

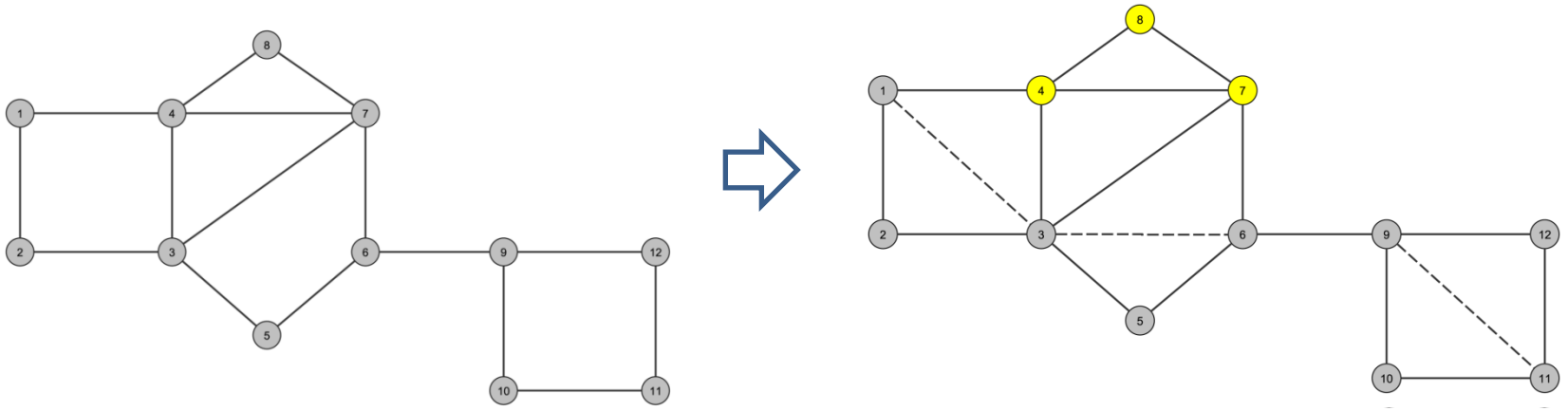
TSAR – a new algorithm for multistate calculations

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

- By accounting topology of a graph describing molecular system

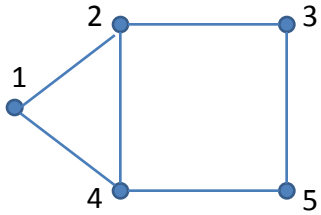
a graph of N nodes each having M states (complexity M^N)

is transformed to the clique of n nodes (n<<N) with complexity M^n



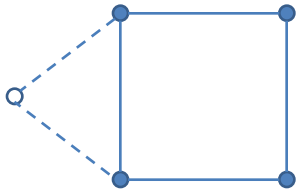
M^{12} **M^3**

TSAR – a new algorithm for multistate calculations

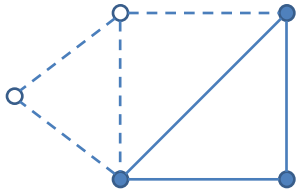


$$\sum_{N_{states}} e^{-E_{12345}}$$

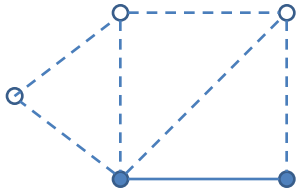
$$N_{states} = \prod_{i=1}^{5} N_i \sim N^5$$



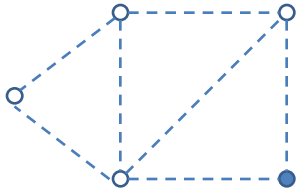
$$\sum_{N_{states}} e^{-E_{12345}} = \sum_{2345}\sum_{1} e^{-(E_1+E_{12}+E_{14})} \cdot e^{-E_{2345}} = \sum_{2345} W_{24} \cdot e^{-E_{2345}} \qquad N^3$$



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



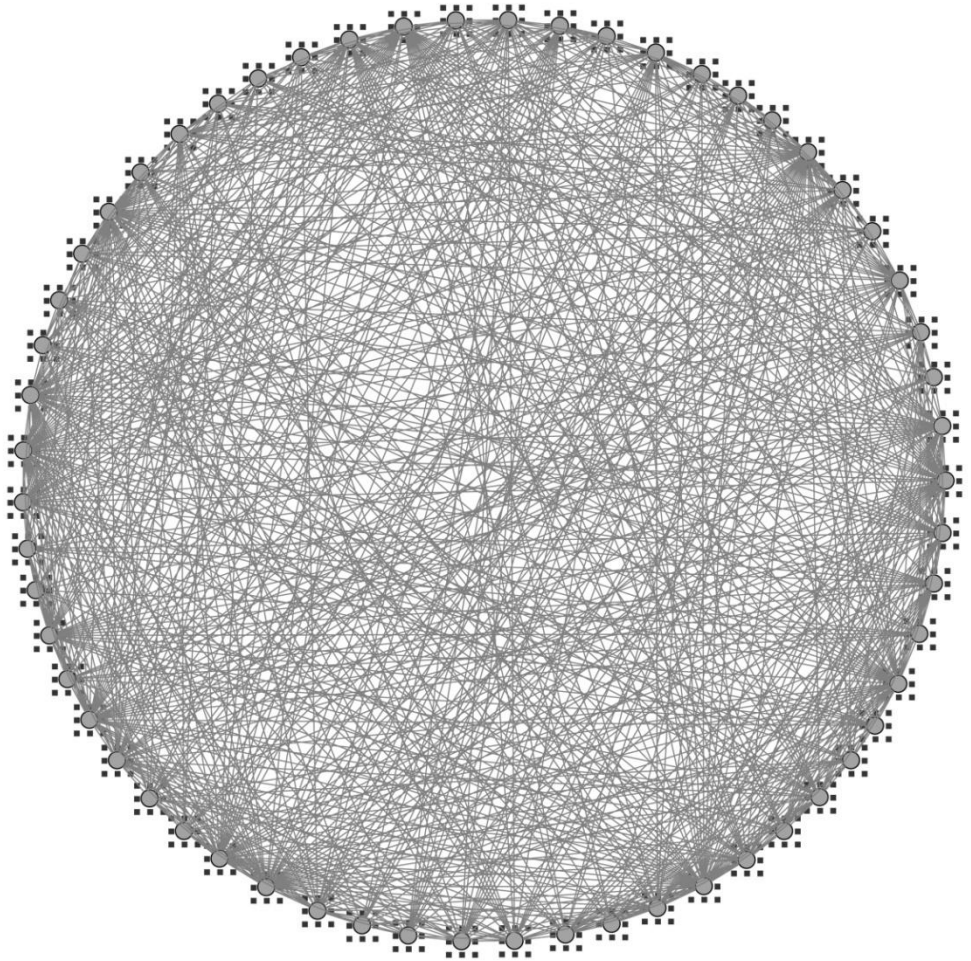
$$\sum_{345} W_{34} \cdot e^{-E_{345}} = \sum_{45}\left(\sum_{3} W_{34} \cdot e^{-(E_3+E_{35})}\right) \cdot e^{-E_{45}} = \sum_{45} W_{45} \cdot e^{-E_{45}} \qquad N^3$$



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

TSAR – a new algorithm for multistate calculations

PDB ID	Complexity , 10^X	
	Straightforward	TSAR
1amm	5319	153
1bd8	4097	120
1c9o	3693	106
1ctj	1974	60
1eca	3722	110
1igd	1728	50
1nar	9228	264
1qlw	17569	527
2cpl	4608	129
5pti	1937	56



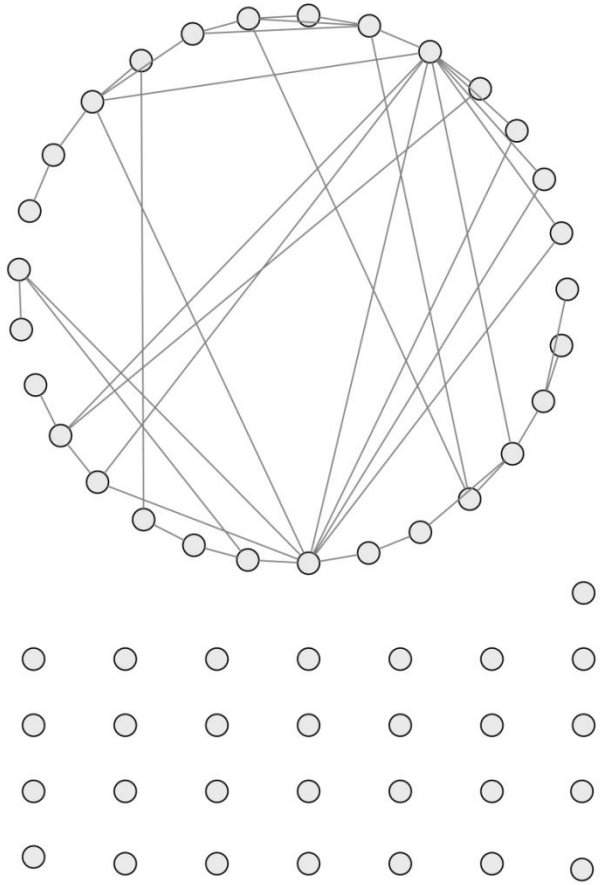
BPTI (5pti)

TSAR: approximations

- Deletion of energetically unfavorable states of graph nodes (side-chains, ligand, etc)
- Deletion of graph edges (bonds) which energy varies negligibly
- Additional sampling for “unstable” graph nodes
- Ability to discard all states of a system less energetically favorable than a given state

TSAR – a new algorithm for multistate calculations

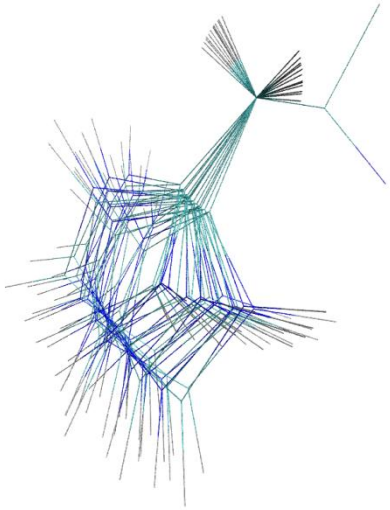
PDB ID	Complexity , 10^X	
	Straightforward	TSAR
1amm	622	6.0
1bd8	490	6.0
1c9o	418	5.9
1ctj	250	6.1
1eca	445	5.8
1igd	193	6.1
1nar	1031	5.9
1qlw	2047	6.1
2cpl	563	6.0
5pti	220	6.1

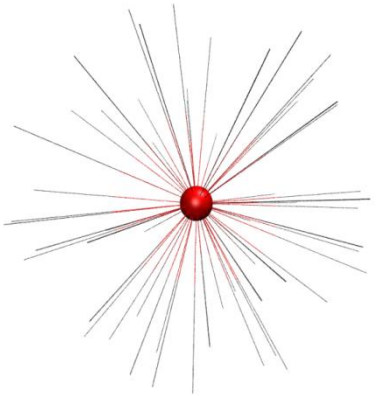


BPTI (5pti)

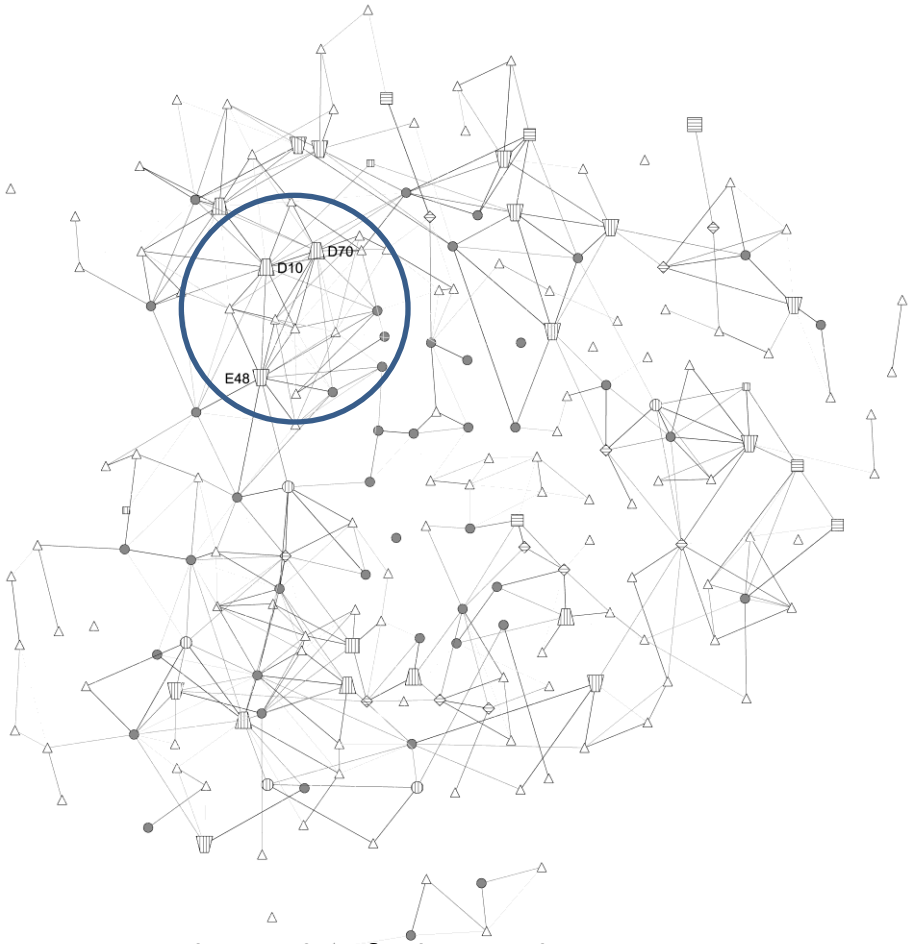
TSAR: application to protein ionization properties

- Graph nodes
  - Side-chains, water molecules
- Node states
  - Conformations, ionization forms
- Edges
  - Nodes within 4 Å from each other interact



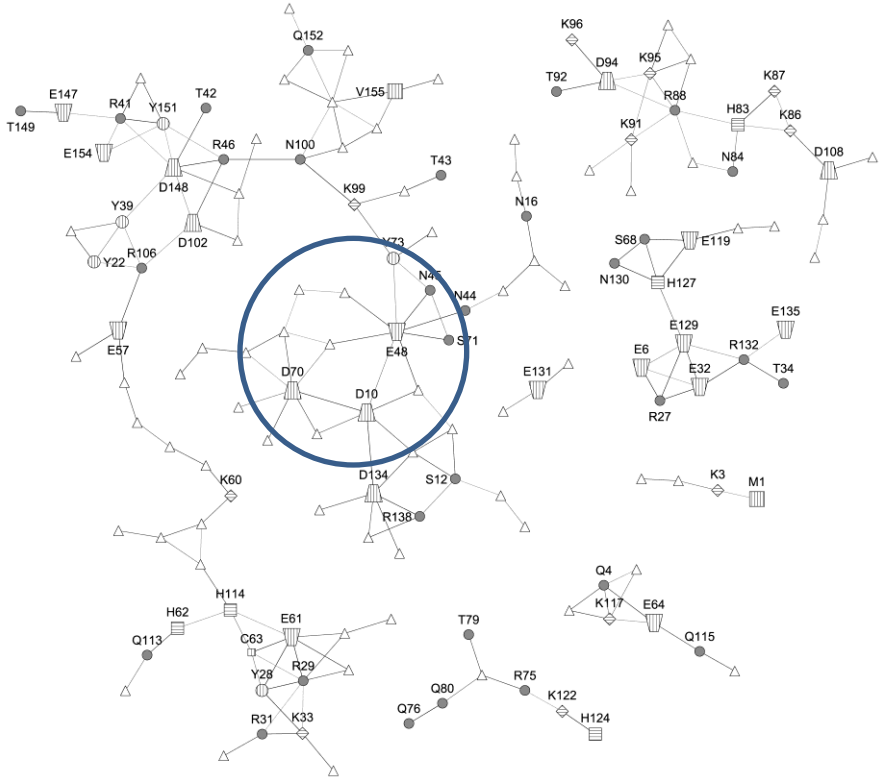


How TSAR works



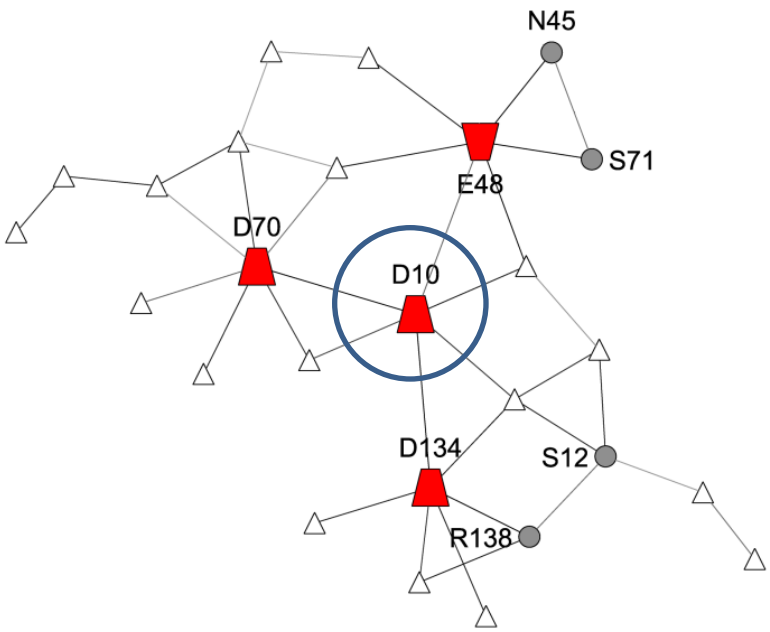
Initial graph of ribonuclease H,
complexity $\mathbf{10^{166}}$
2.4 edges/node

Maximum clique size = 8

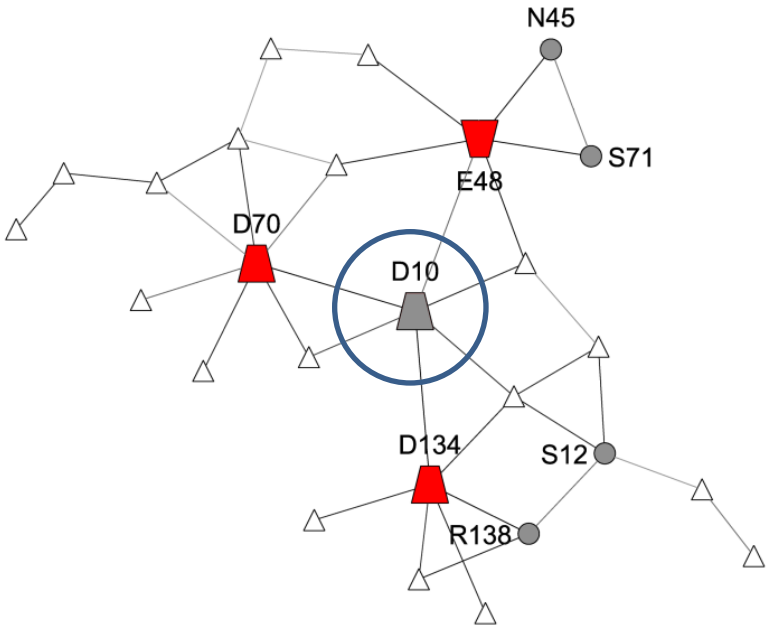


Final graph of ribonuclease H,
complexity $\mathbf{10^{6}}$
0,54 edges/node

TSAR: application to model protein ionization properties



States with D10 deprotonated



States with D10 protonated

$$\Delta G_i(pH) = RT \cdot \ln \sum_{States\_i\_protonated} e^{-E_{State}} \Big/ \sum_{States\_i\_deprotonated} e^{-E_{State}}$$

Energy calculations

- Simplified molecular mechanical functional:
  - Steric clashes $k_{clashes}\sum_{i,j} E_{clashes,ij}\left(r_{ij}\right)$
  - Hydrogen bonds $k_{hbonds}\sum_{i,j} k_{ij}\, f_H\, f_{LP} E_{LJ,ij}\left(r_{ij}\right) + k_{penalty}\min(0;2N_c - N_{hb})$
  - Electrostatic interactions $k_{elec}\sum_{i,j}\frac{q_i q_j}{R'_{ij}}\frac{1}{\varepsilon(R'_{ij})}$
  - Interaction with metal ions $k_{metal}\sum_{i,j} k_{ij}\, f_{Me}\, f_{LP} E_{LJ,ij}\left(r_{ij}\right)$

$$E_{State} = \sum_i E_i + \sum_{i,j} E_{i,j}$$

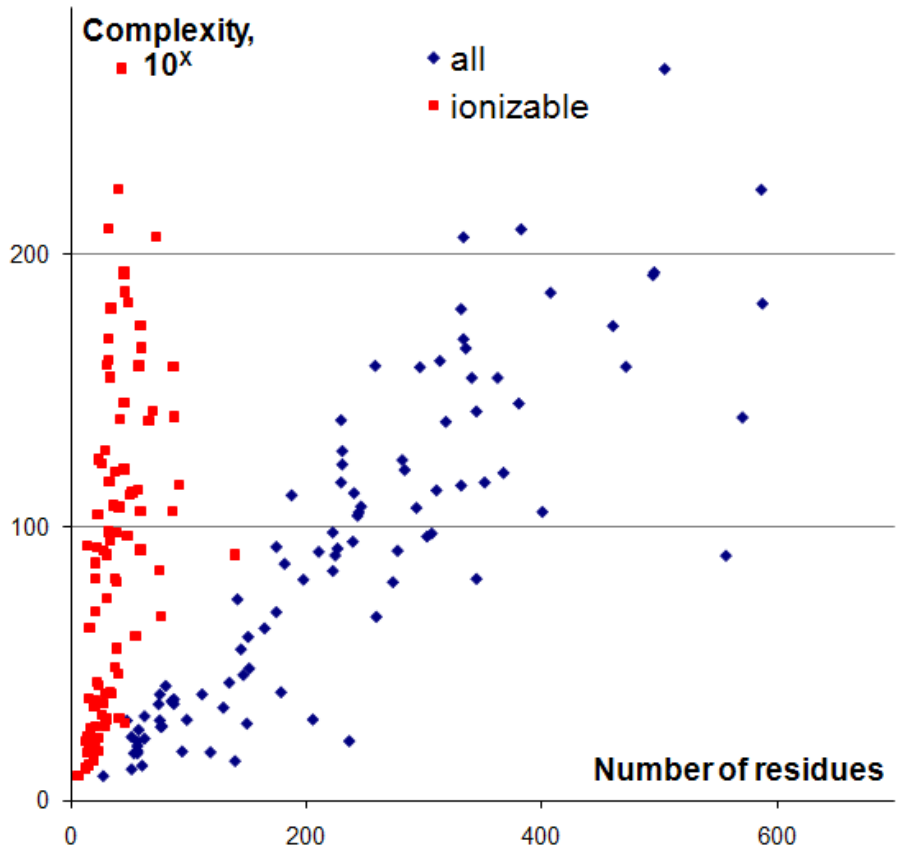
Test set of protein pKa values

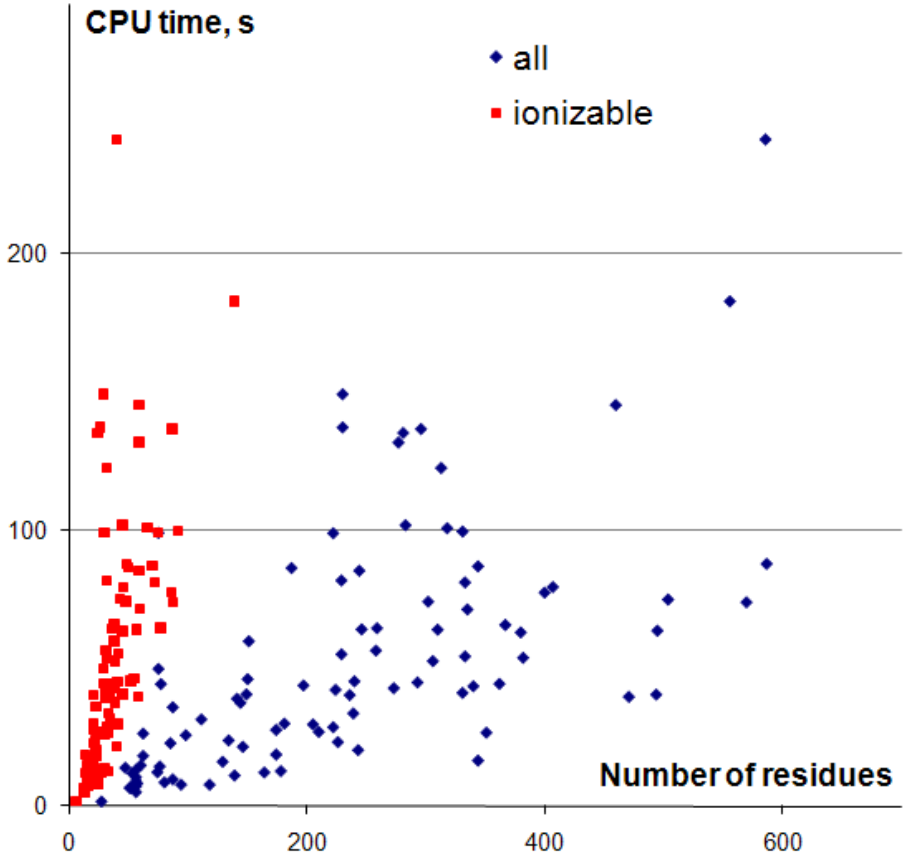
484 pKa values from 96 proteins

pH*	Asp	Glu	His	Tyr	Lys	TerC	TerN
1-2	2						
2-3	33	5	1			6	
3-4	75	34				8	
5-6	17	74	3				
6-7	1	5	30				1
7-8	2	4	60	1			1
8-9		2	23				6
9-10		2	5	3			
10-11			3	1	6		1
11-12				5	37		
12-13				8	18		
13-14				1			
All	130	126	125	19	61	14	9

* pH interval in which pKa of amino acids lie

Computational performance of TSAR





Accuracy of pKa calculations with TSAR

	RMSD	R^2	% of cases with $\lvert pKa_{exp} - pKa_{calc} \rvert >$ 1.0	% of cases with $\lvert pKa_{exp} - pKa_{calc} \rvert >$ 2.0
Asp	0.54 (0.65)	0.48	**3 (12)**	**0 (2)**
Glu	0.70 (0.70)	0.52	**15 (17)**	**1 (6)**
His	0.64 (0.70)	0.62	**9 (30)**	**1 (5)**
Tyr	0.47 (0.46)	0.90	**0 (37)**	**0 (16)**
Lys	0.46 (0.45)	0.37	**0 (8)**	**0 (0)**
TerC	0.19 (0.23)	0.86	**0 (0)**	**0 (0)**
TerN	0.45 (0.51)	0.83	**0 (22)**	**0 (0)**

The influence of sampling on the accuracy of pKa calculations

	R^2			% of cases with $\lvert pKa_{exp} - pKa_{calc} \rvert >$ 1.0			% of cases with $\lvert pKa_{exp} - pKa_{calc} \rvert >$ 2.0		
		I	II		I	II		I	II
Asp	0.48	(0.30)	(0.41)	**3**	**(9)**	**(8)**	**0**	**(0)**	**(0)**
Glu	0.52	(0.53)	(0.53)	**15**	**(15)**	**(16)**	**1**	**(2)**	**(2)**
His	0.62	(0.58)	(0.59)	**9**	**(12)**	**(11)**	**1**	**(1)**	**(1)**
Tyr	0.90	(0.92)	(0.92)	**0**	**(0)**	**(0)**	**0**	**(0)**	**(0)**
Lys	0.37	(0.40)	(0.40)	**0**	**(2)**	**(3)**	**0**	**(0)**	**(0)**
TerC	0.86	(0.79)	(0.85)	**0**	**(0)**	**(0)**	**0**	**(0)**	**(0)**
TerN	0.83	(0.77)	(0.83)	**0**	**(0)**	**(0)**	**0**	**(0)**	**(0)**

I — the number of states per distinct ionized form of amino acid is reduced to 2
II — additional sampling of residues is switched off

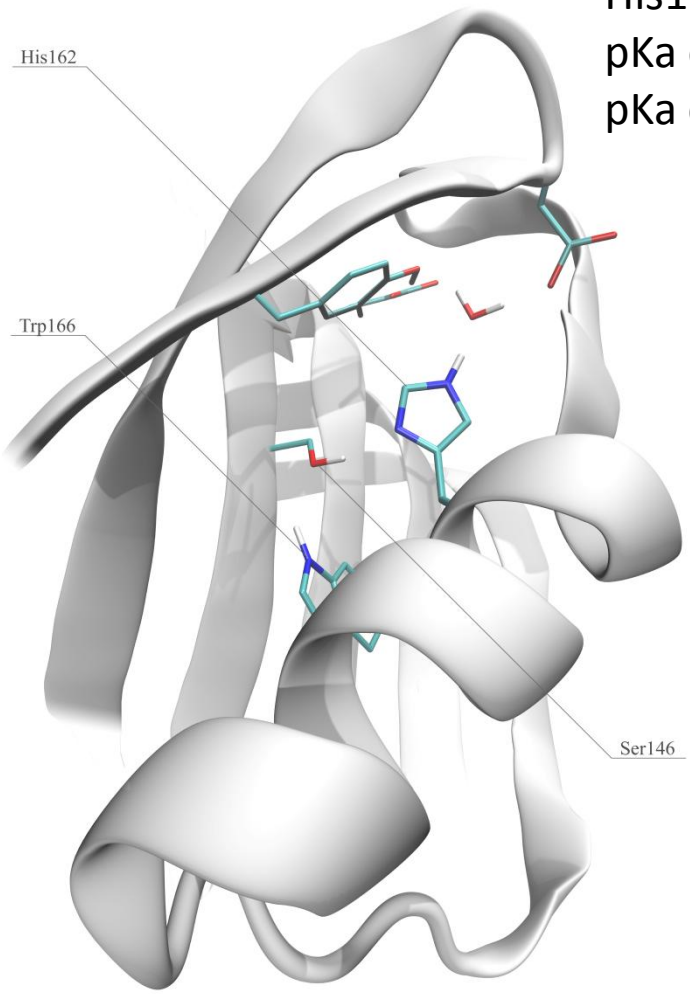
The contributions of electrostatics and H-bonds

	R^2			% of cases with $\lvert pKa_{exp} - pKa_{calc} \rvert >$ 1.0			% of cases with $\lvert pKa_{exp} - pKa_{calc} \rvert >$ 2.0		
		I	II		I	II		I	II
Asp	0.48	(0.35)	(0.18)	**3**	**(12)**	**(12)**	**0**	**(0)**	**(2)**
Glu	0.52	(0.52)	(0.16)	**15**	**(15)**	**(15)**	**1**	**(2)**	**(5)**
His	0.62	(0.54)	(0.35)	**9**	**(17)**	**(19)**	**1**	**(1)**	**(4)**
Tyr	0.90	(0.90)	(0.41)	**0**	**(0)**	**(26)**	**0**	**(0)**	**(16)**
Lys	0.37	(0.16)	(0.28)	**0**	**(10)**	**(2)**	**0**	**(0)**	**(0)**
TerC	0.86	(0.77)	(0.77)	**0**	**(0)**	**(36)**	**0**	**(0)**	**(0)**
TerN	0.83	(0.81)	(0.86)	**0**	**(0)**	**(0)**	**0**	**(0)**	**(0)**

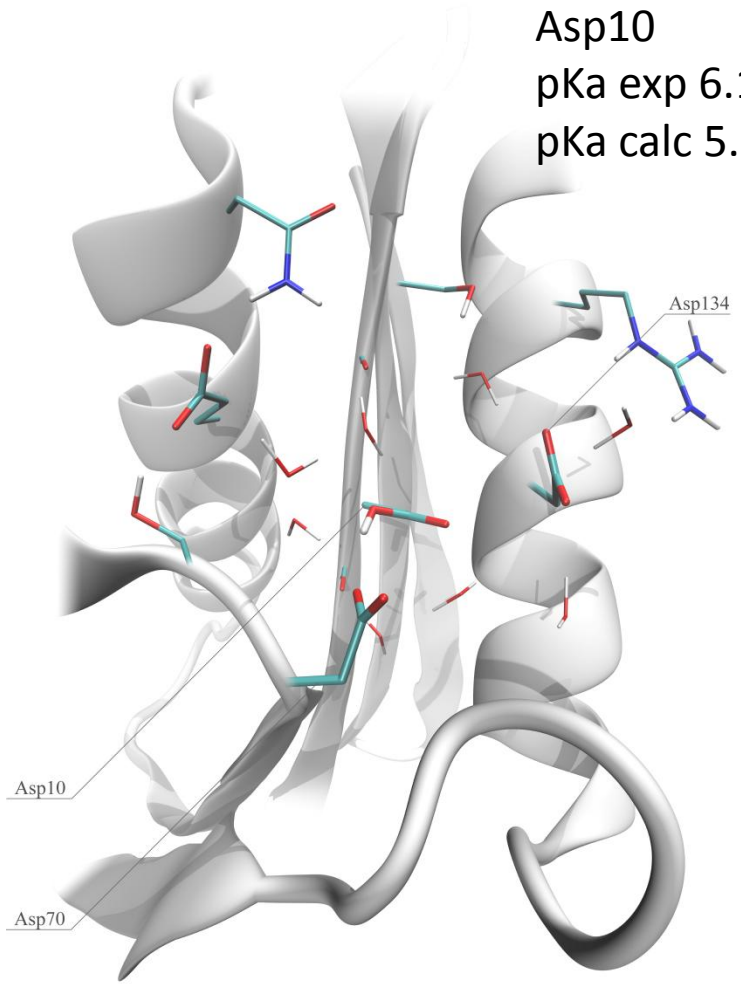
I — electrostatic interactions are switched off
II — H-bonds are switched off

Importance of sampling H-bond networks

His162
pKa exp 2.7
pKa calc 3.2



Asp10
pKa exp 6.1
pKa calc 5.7



Current limitations of TSAR and perspectives

Limitations:

- Fixed protein backbone
- Computational limitations on maximum clique size and number of node states

Novel applications:

- Protein structure modeling
- Protein design
- Fully flexible protein-ligand docking with explicit water treatment
- Free energy calculations