Mathematics is the champion of biomolecular data challenges

Zixuan Cang, Duc Nguyen, Bao Wang, Chengzhang Wang, Menglun Wang, Guo-Wei Wei, Kedi Wu, and Zhixiong Zhao
Department of Mathematics, Michigan State University

Feature functional theory (FFT)

Goal: To prediction microscopic and macroscopic relationships in biomolecular data

Basic assumptions:

- **Representability assumption**: there exists a microscopic feature vector that can uniquely characterize, and distinguish one molecule from another

$$\mathbf{v}_i = (\mathbf{x}_i;\mathbf{o}_i) = (x_{i1}, x_{i2},\ldots, x_{in}; o_{i1}, o_{i2},\ldots, o_{il})$$

microscopic features; macroscopic features

- **Similarity assumption**: molecules with similar microscopic features have similar macroscopic features.
- **Feature-function relationship assumption**: the macroscopic features, i.e., solvation and binding free energies, of molecule A are functionals of microscopic feature vectors:

$$\Delta G_{\mathrm{A}} = f_{\mathrm{A}}(\mathbf{x}_{\mathrm{A}}, \mathbf{v}_1, \mathbf{v}_2,\ldots, \mathbf{v}_n)$$

Feature functional theory (FFT)

Microscopic features:

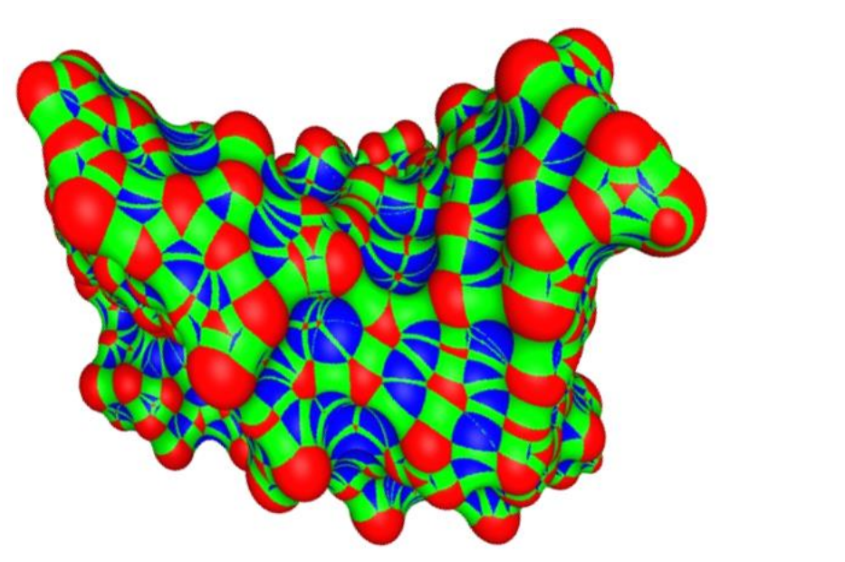
- Geometric: atomic surface areas, volume & curvatures
- Topological: Betti numbers
- Graph theory: discrete Laplacian, rigidity & flexibility
- Electrostatic: atomic charges, dipoles, quadrupoles & reaction filed energies (Poisson-Boltzmann equation)
- van der Waals: Lennard-Jones potentials

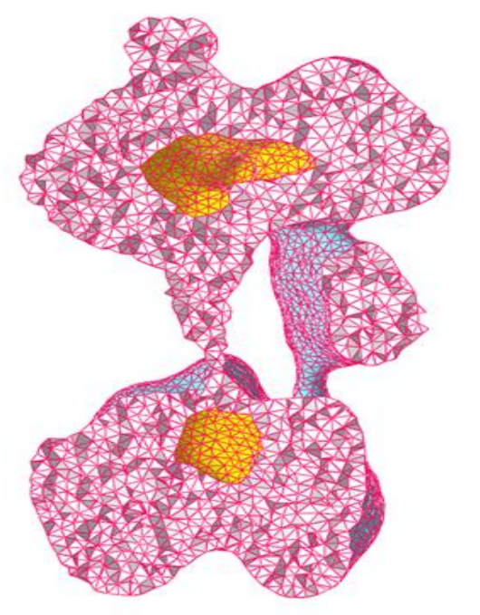
Macroscopic features:

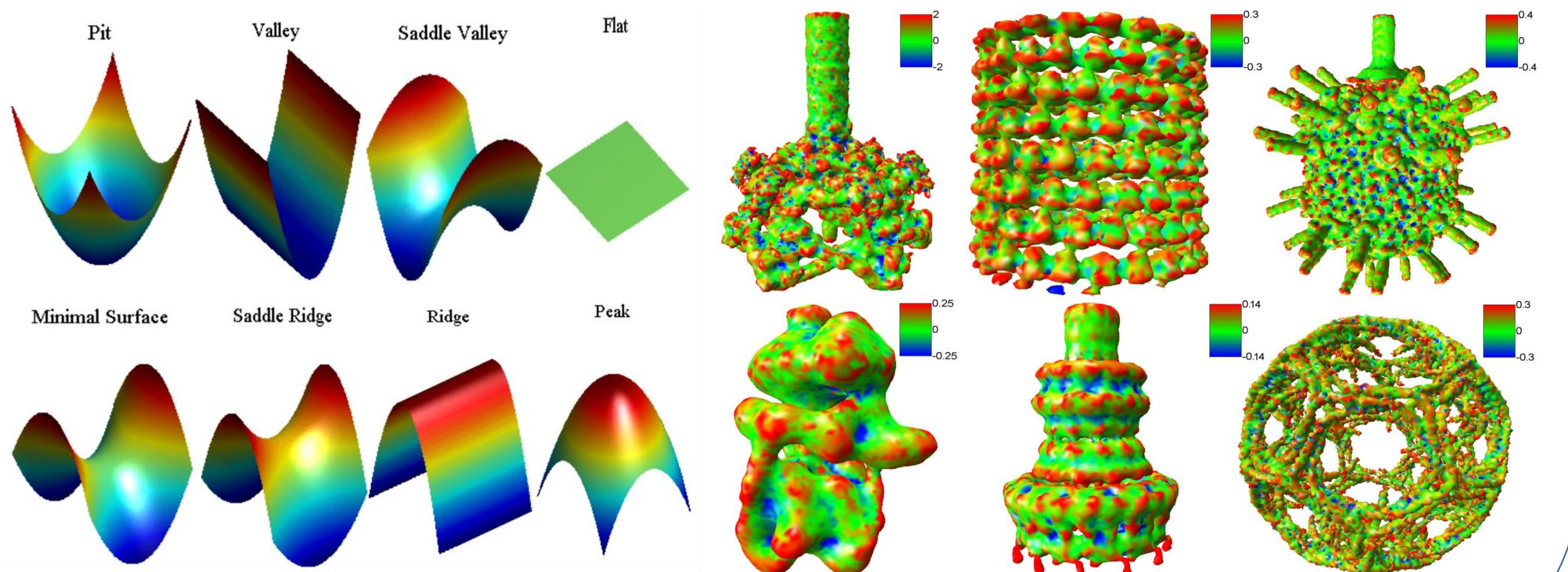
- Protein-ligand binding affinities
- Protein mutation energy changes (stability changes)
- Drug partition coefficients
- Drug solvation free energies
- Protein-DNA/RNA binding energies
- Protein-protein binding affinities

Geometric modeling

Besides surface area and volume, curvature matters



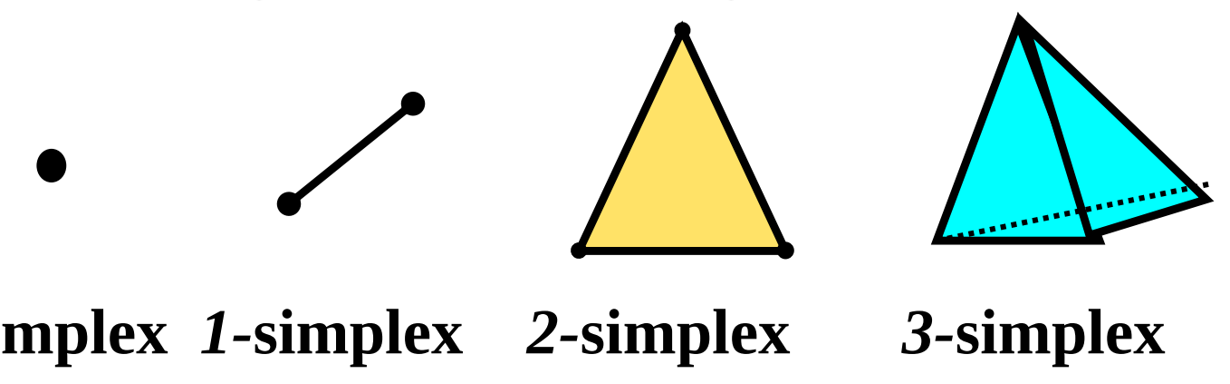




Feng, Xia, Tong and Wei, IJNMBI,2012; JCP, 2013

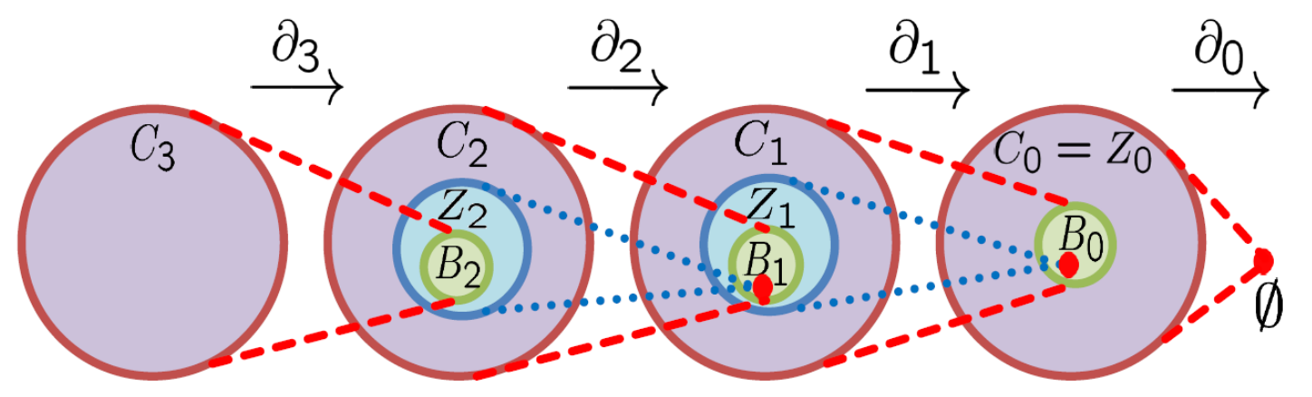
Persistent topology modeling

Simplicial complex:



k-chain: $\sum_i c_i \sigma_i^k$

Chain group: $C_k(K,Z_2)$



Boundary operator:

Xia, Feng, Tong, Wei, JCC, 2015; Liu, Wang, Zhao, Tong, Wei, 2016; Xia, Wei 2016

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i \{v_0, v_1,\ldots,\hat{v}_i,\ldots,v_k\}$$

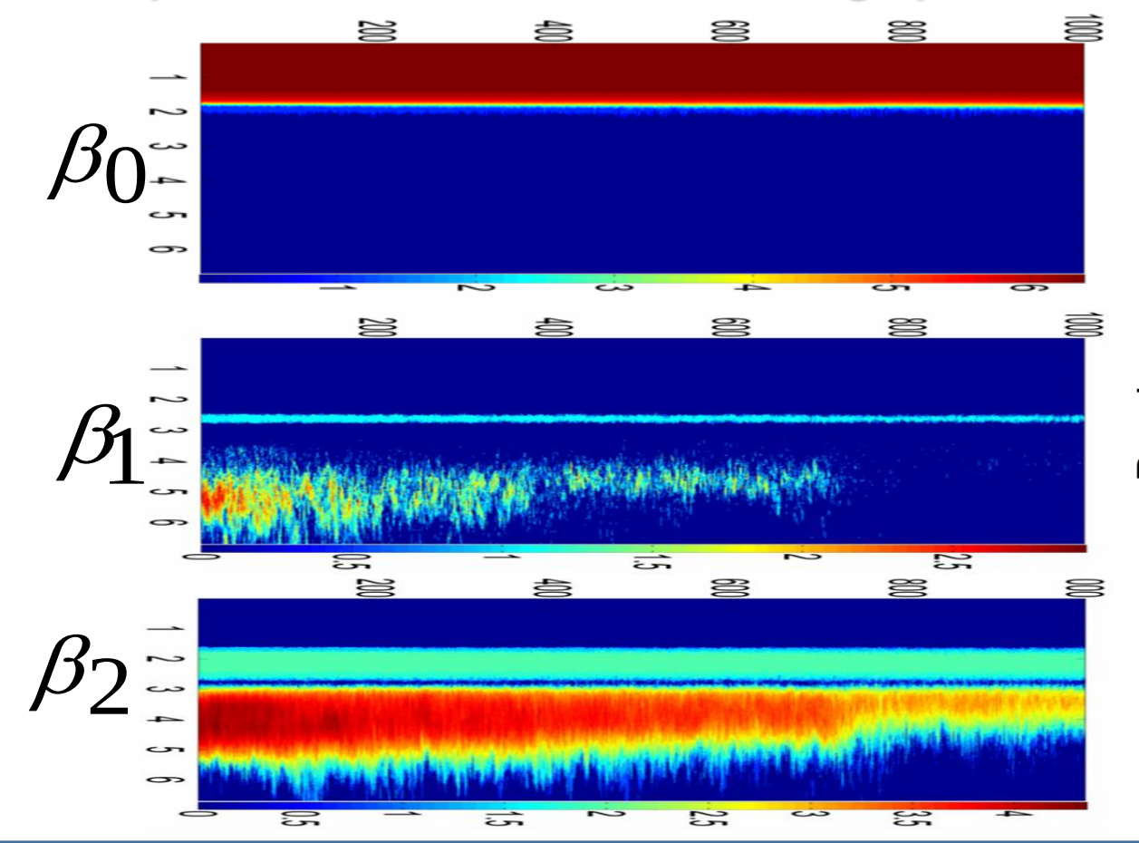
$$Z_k = \operatorname{Ker} \partial_k$$

$$B_k = \operatorname{Im} \partial_{k+1}$$

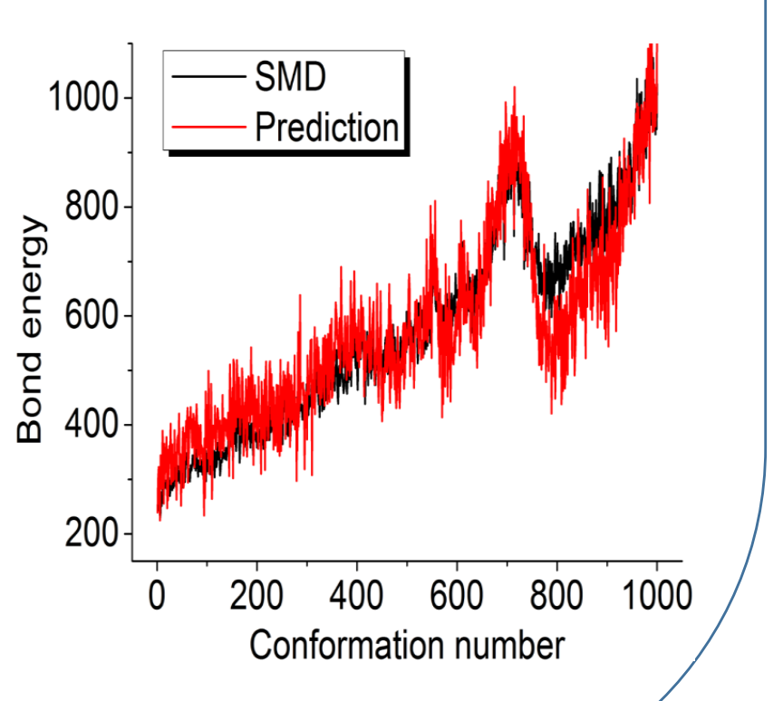
$$H_k = Z_k / B_k$$

$$\beta_k = \operatorname{Rank}(H_k)$$

2D persistence of a unfolding protein:



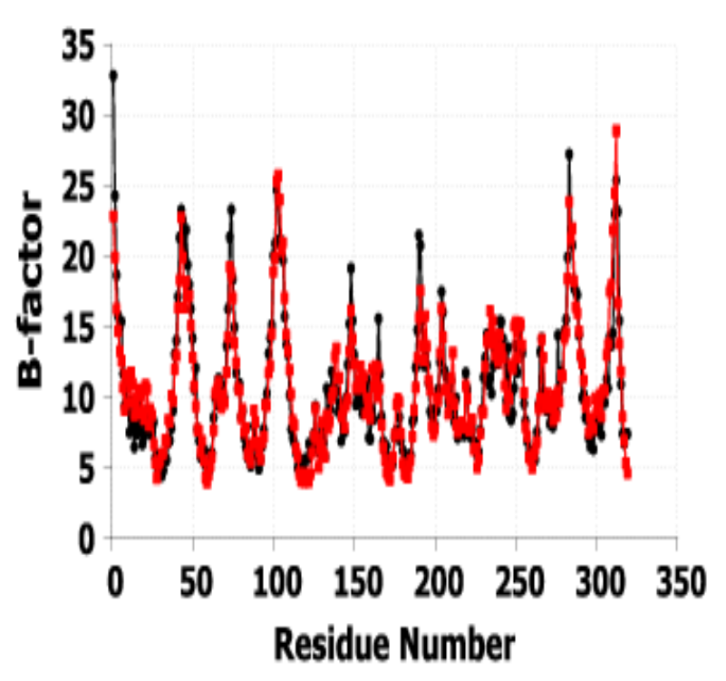
Quantitative topology:

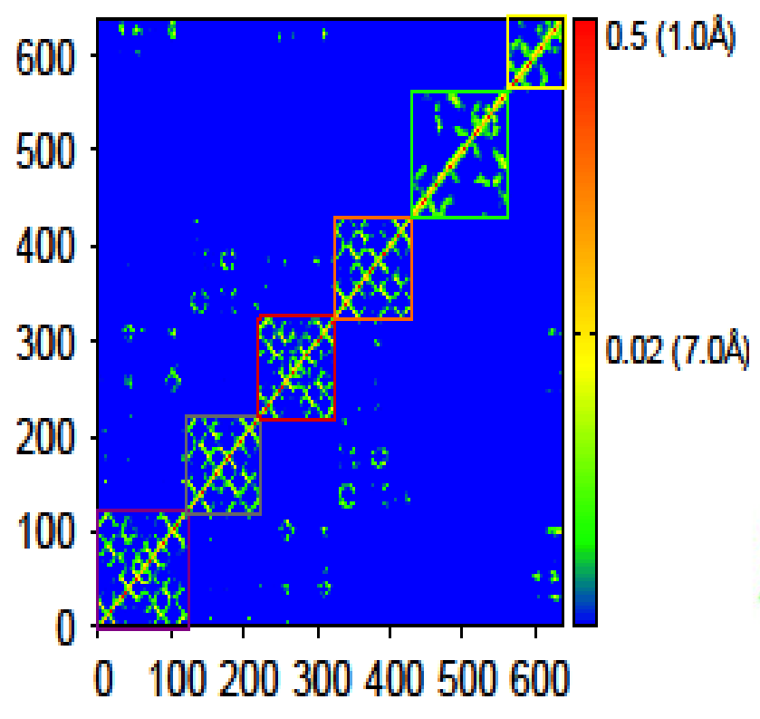


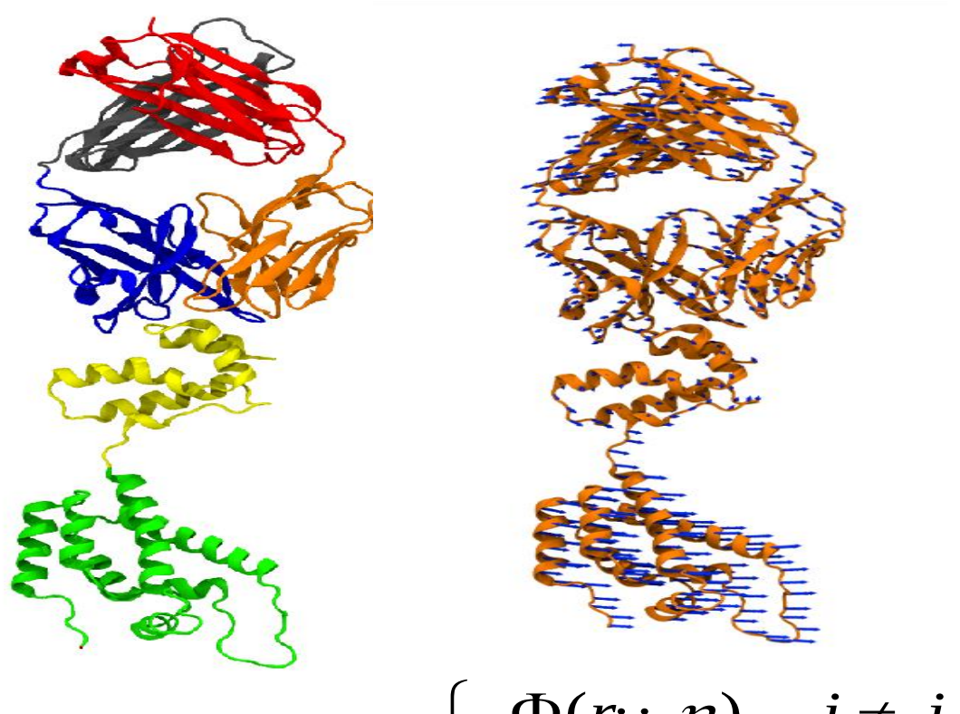
Graph theory modeling

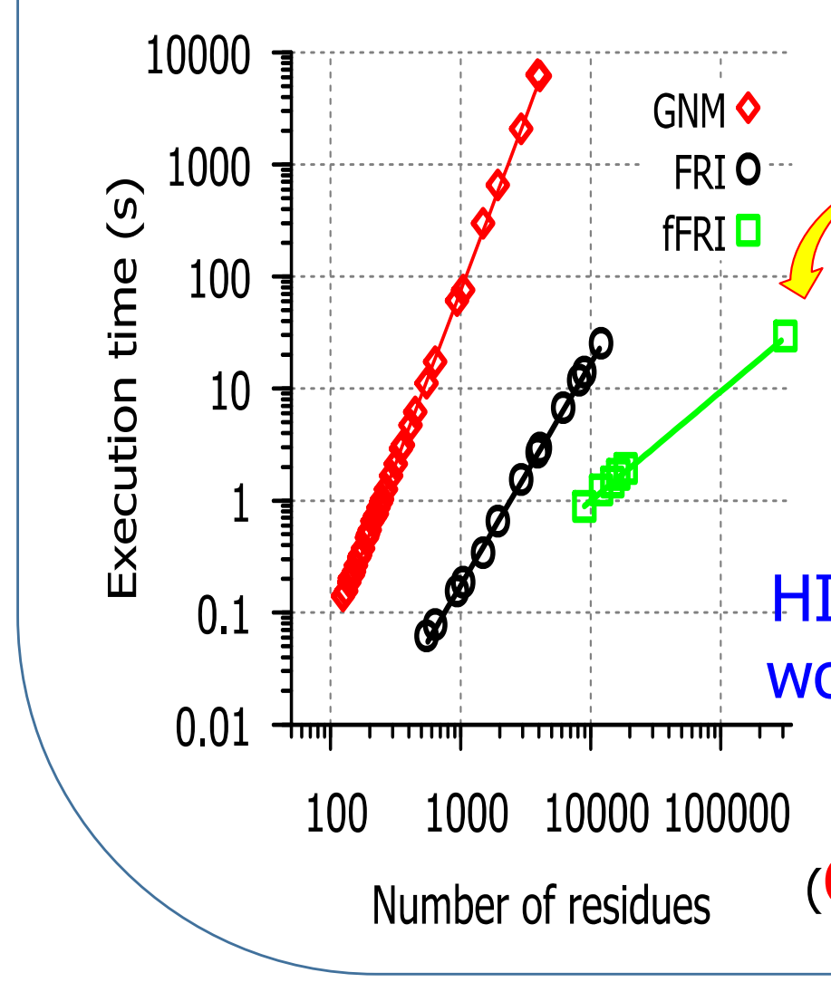
Weighted graph Laplacian, Flexibility rigidity index (FRI)

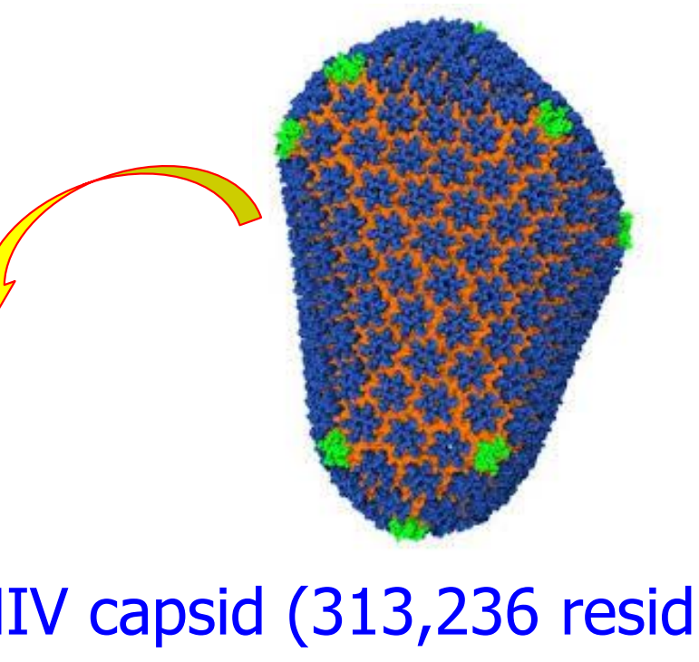
FRI is about 20% more accurate than Gaussian network model (GNM) in B-factor prediction, based on 364 proteins.











HIV capsid (313,236 residues) would takes GNM 120 years to compute!

$$\boldsymbol{\Gamma}_{ij}(\Phi) = \begin{cases} -\Phi(r_{ij},\eta), & i \neq j, \\ -\sum_{j,j\neq i}^{N} \boldsymbol{\Gamma}_{ij}, & i = j \end{cases}$$

$$\Phi(r_{ij},\eta) = 1, \quad r_{ij} \to 0$$

$$\Phi(r_{ij},\eta) = 0 \quad r_{ij} \to \infty$$

$$\Phi(r_{ij},\eta) = e^{-(r_{ij}/\eta)^{\kappa}}$$

$$B_i^{\mathrm{FRI}} = a(\boldsymbol{\Gamma}_{ii}(\Phi))^{-1}$$

(Opron, Xia and Wei, JCP, 2013; JCP 2014; JCP, 2015)

Physical modeling

Explicit solvent models (*Molecular dynamics, QM/MM, MC*)

❑Atomistic modeling of both solvent and solute molecules.
❑Accurate but time consuming and subjects to force field errors.

Integral equation models (*Ornstein-Zernike, Percus-Yevick and hypernetted-chain equations, RISM, LDFT, etc.*)

❑Continuous function modeling of solvent molecules, while atomistic modeling of the solute.
❑Still accurate but less time consuming.

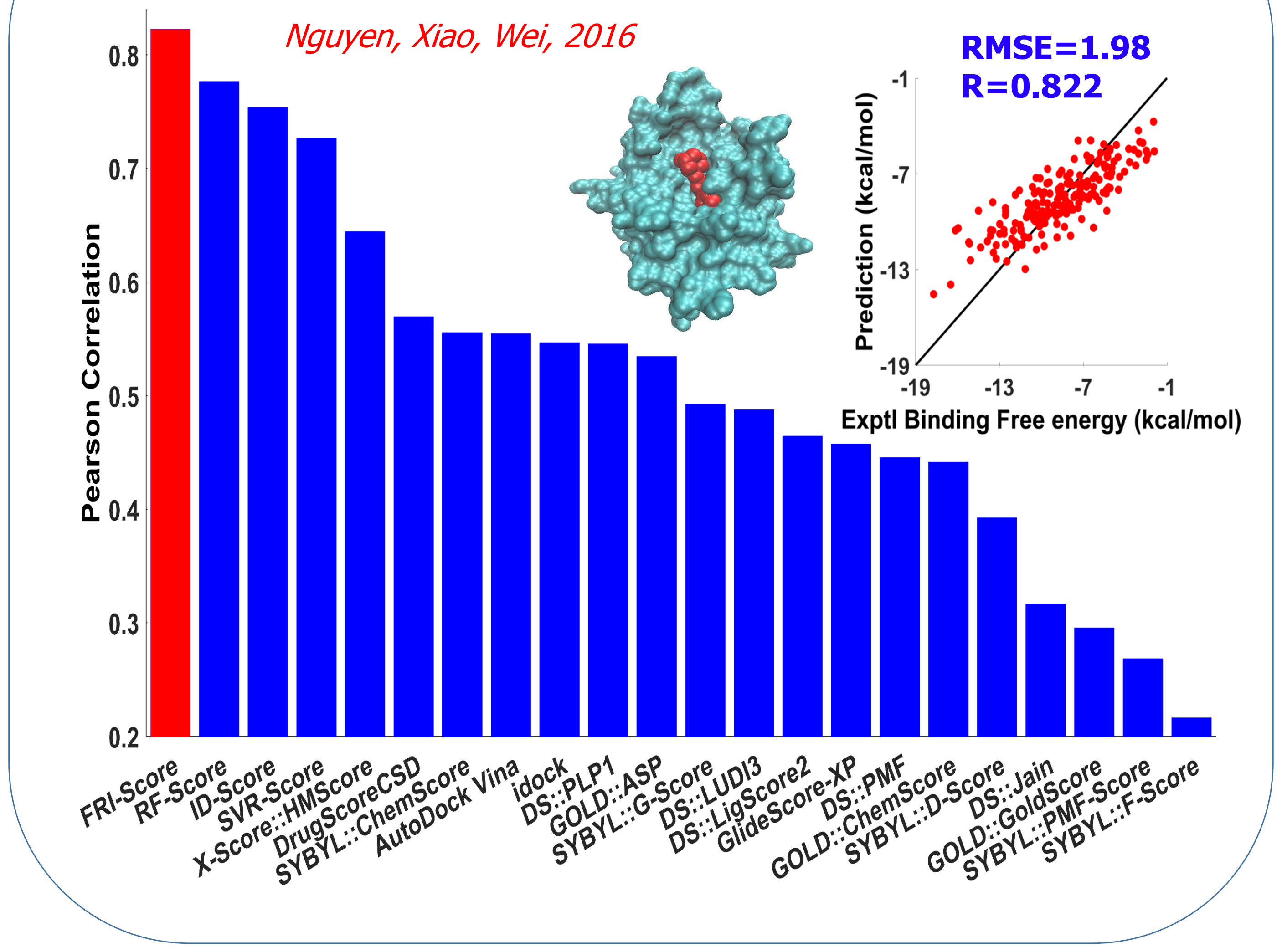
Implicit solvent models (*Image charge, Generalized Born, Poisson-Boltzmann, Polarizable Continuum*)

❑Dielectric continuum modeling of solvent molecules, while atomistic modeling of the solute.
❑A good trade off between accuracy and efficiency.

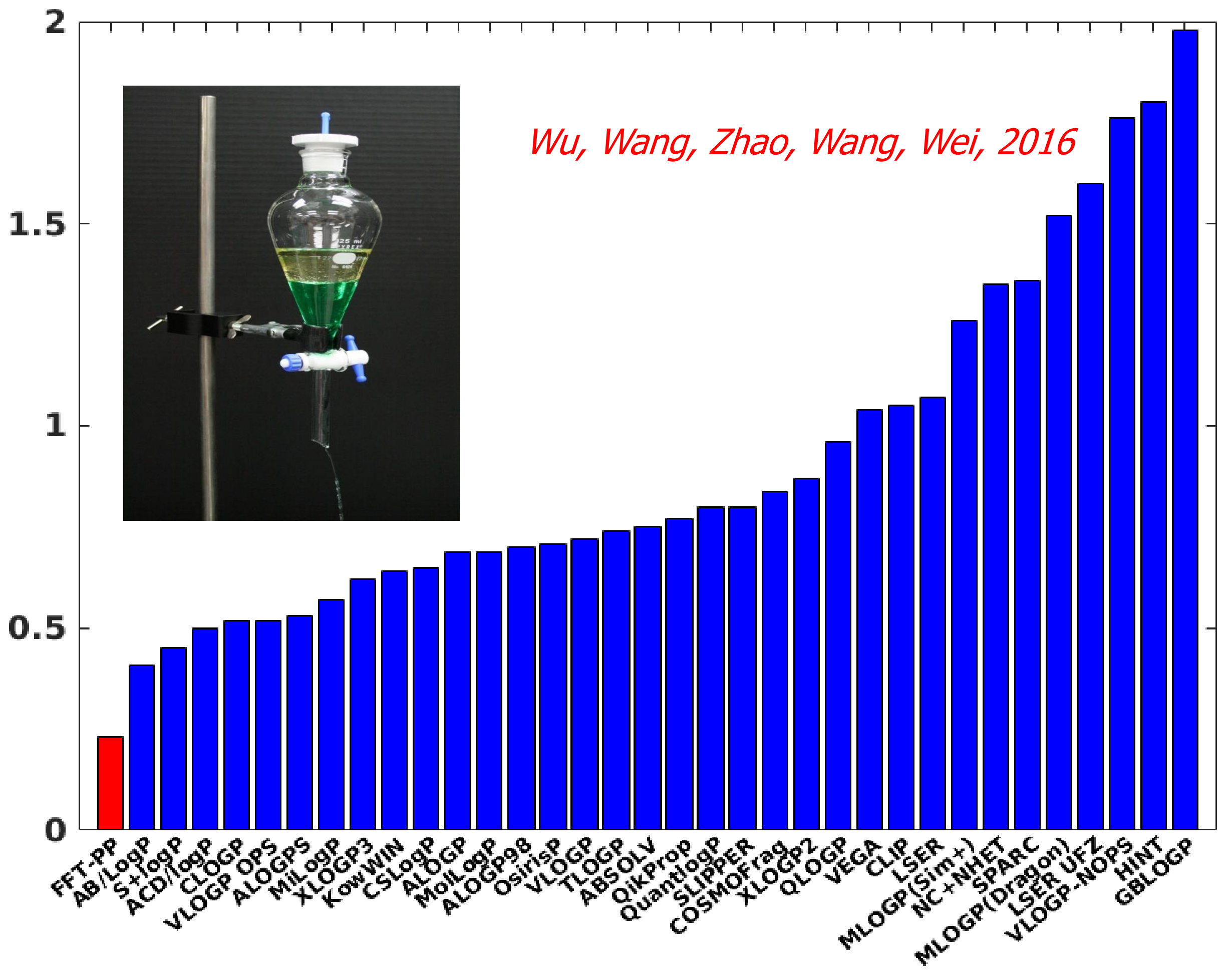
Variational multiscale models (*nonpolar, polar and QM*)

❑Couple polar and nonpolar components by variational surfaces.
❑Self-consistent surface, charge, polarization and energy.

Blind binding affinity prediction of PDBBind v2007 core set of 195 protein-ligand complexes

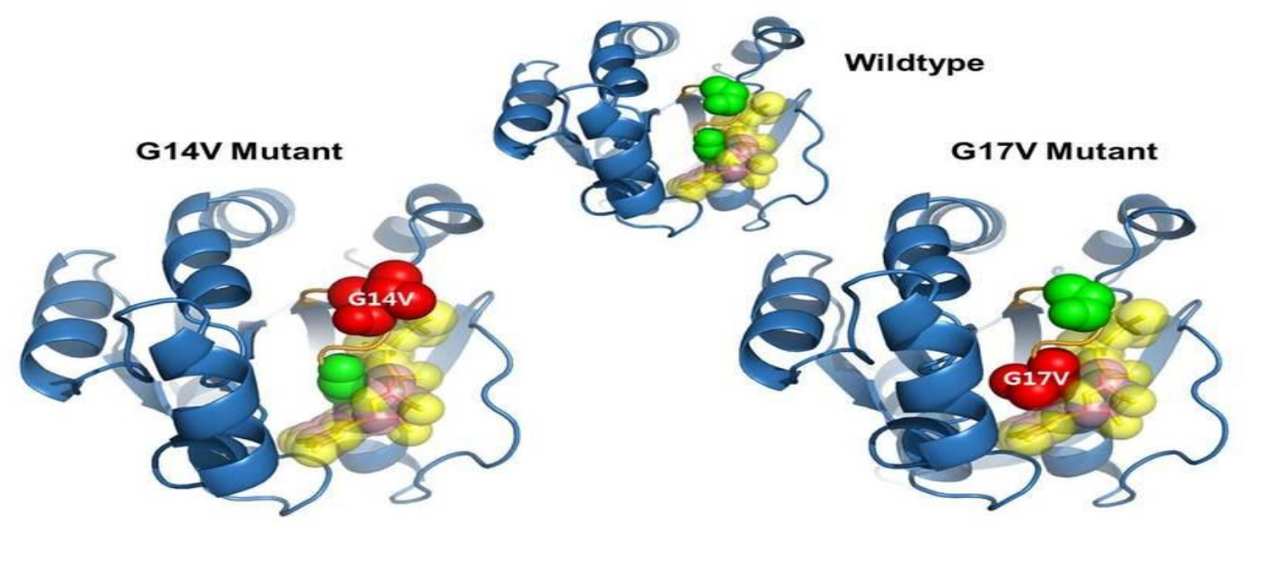


Prediction of partition coefficients: Star Set (223 molecules)

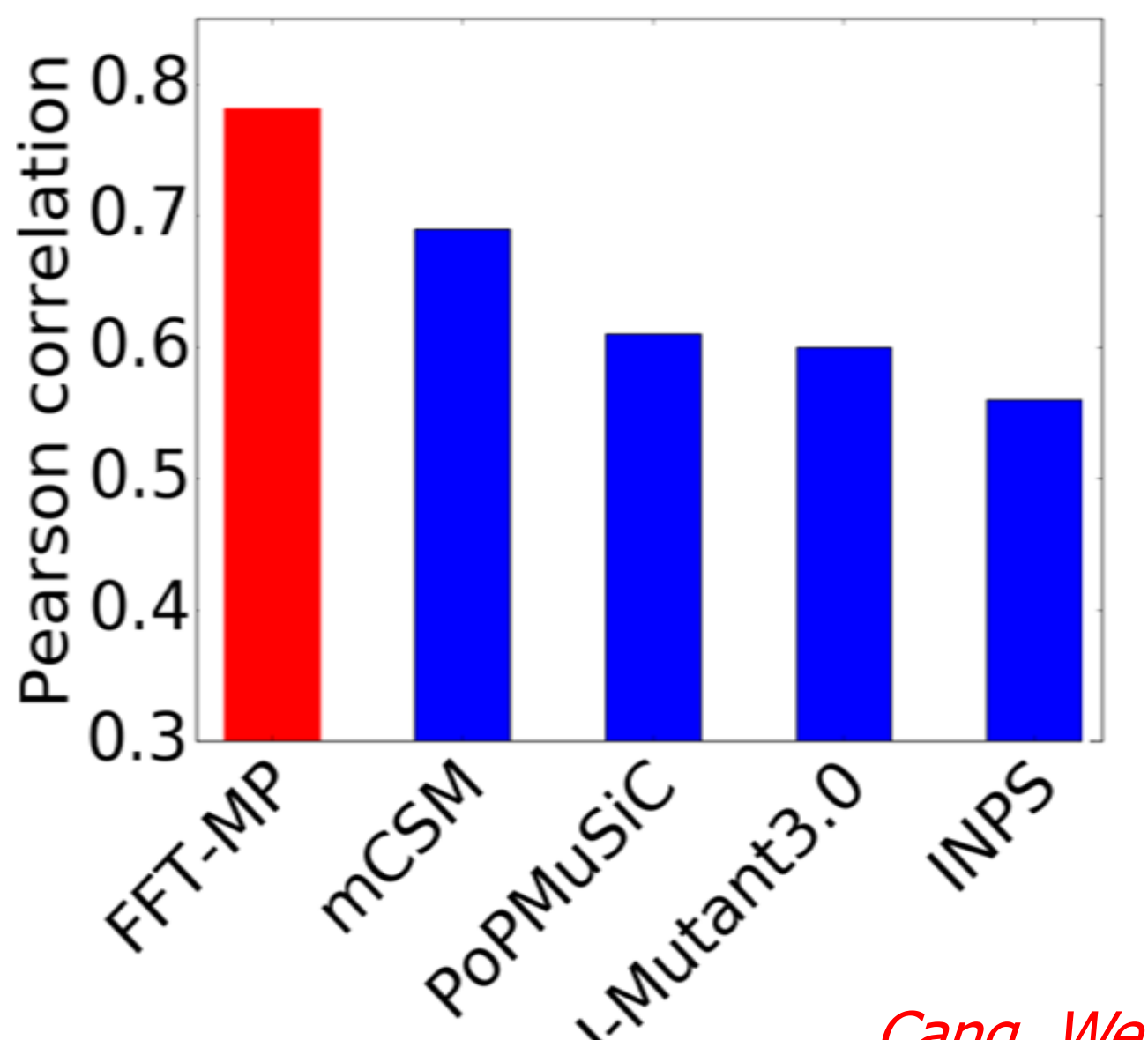


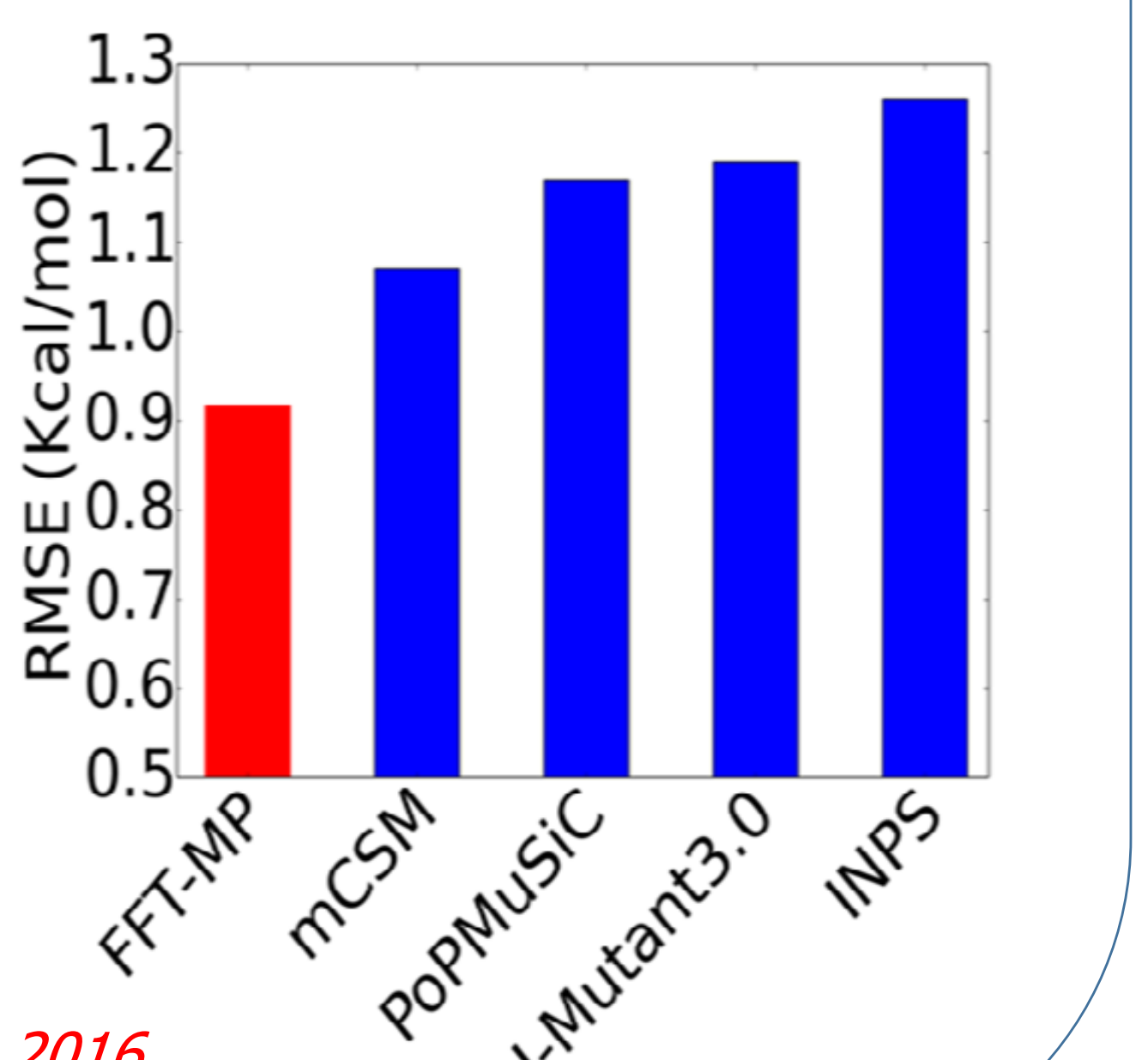
Blind prediction of mutation energies of 2648 dataset

Mutation process



Mutation induced disease





Cang, Wei, 2016

Acknowledgement:

This work was supported in part by NSF, NIH, MSU and MBI.

http://users.math.msu.edu/users/wei/