Lecture 6: Graph Theory Based Modeling and Analysis

Guowei Wei

Mathematics

Michigan State University

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

NSF-CBMS Conference on Mathematical Molecular Bioscience and Biophysics

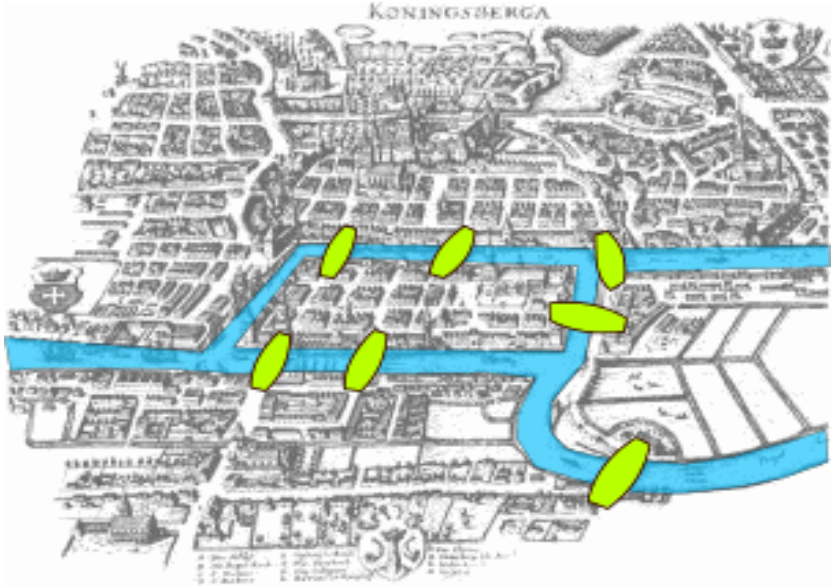
University of Alabama

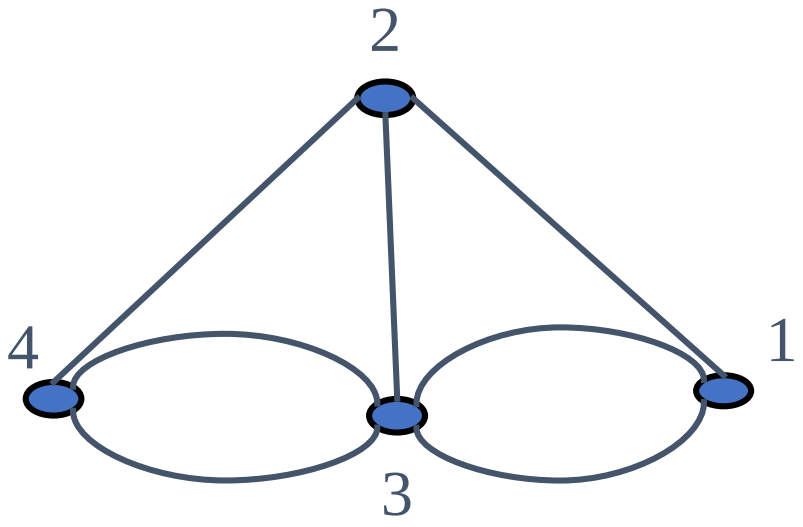
Tuscaloosa, May, 13-17, 2019

Grant support: NSF, NIH, MSU, BMS, and Pfizer

The Seven Bridges of Königsberg, Germany

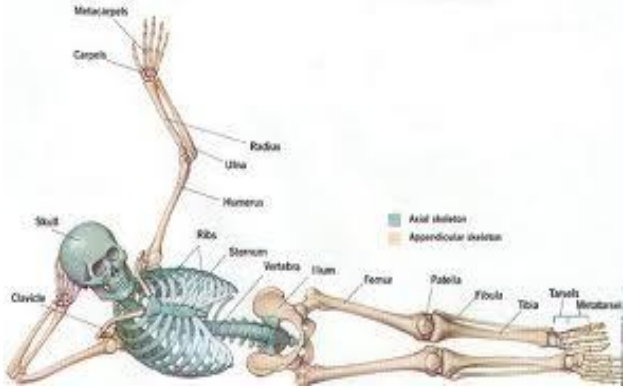
The residents of Königsberg, Germany, wondered if it was possible to take a walking tour of the town that crossed each of the seven bridges over the Presel river exactly once. Is it possible to start at some node and take a walk that uses each edge exactly once, and ends at the starting node?

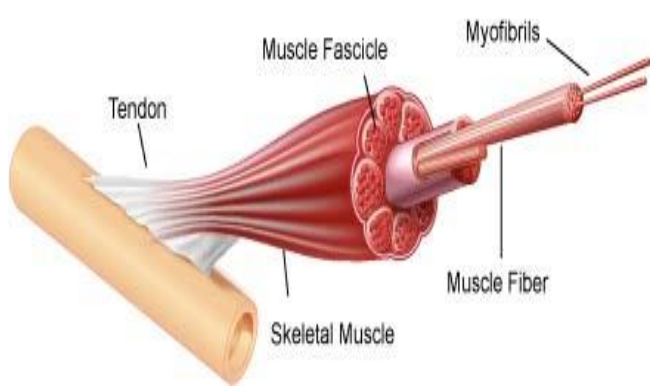


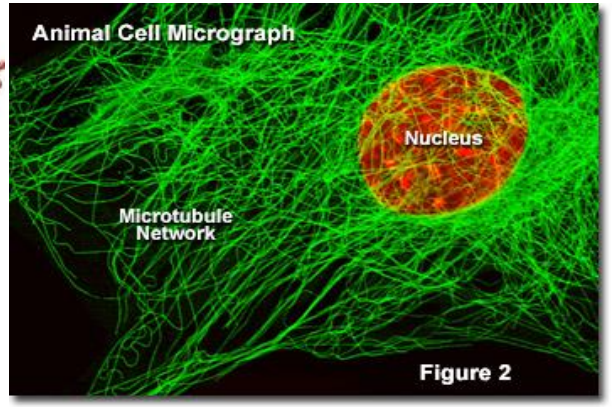


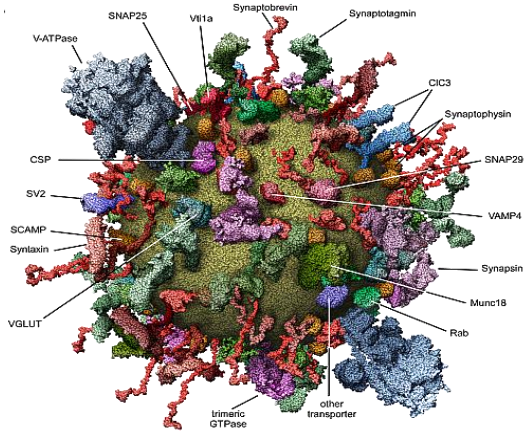
Structural stability and flexibility

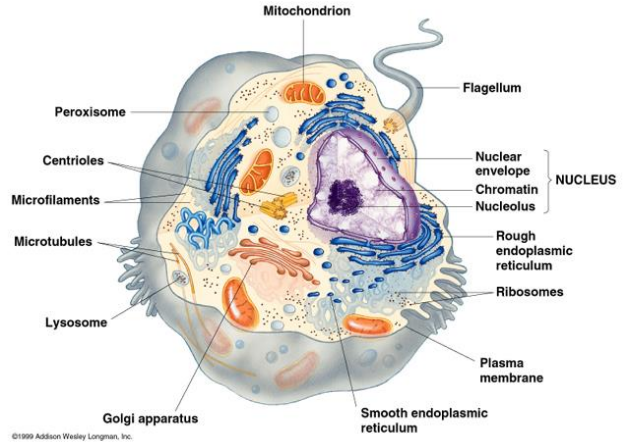
Image credits: Internet

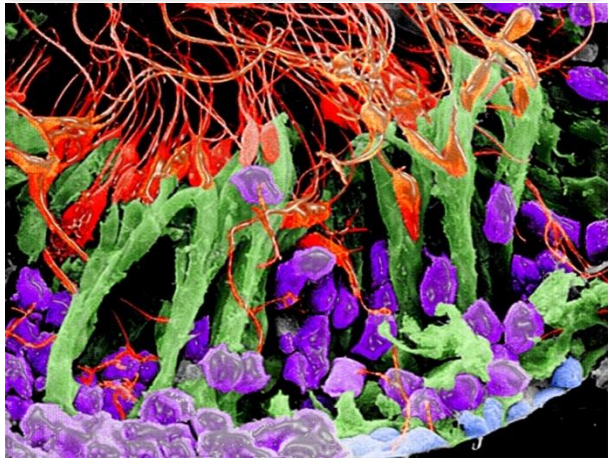












Graph theory for molecular bioscience

- Structural stability and flexibility analysis
- Surface modeling
- Visualization
- Biomolecular domain analysis and hinge detection
- Entropy estimation
- Modeling of a wide range of biomolecular interactions
- Prediction of a wide variety of chemical and biological properties, including binding affinity, solubility, participation coefficient, mutation impact, reaction rates, toxicity, ordered-disordered transition, …

(Balaban, Graovac, Gutman, Hosoya, Randić and Trinajstić, …)

Graph Theory

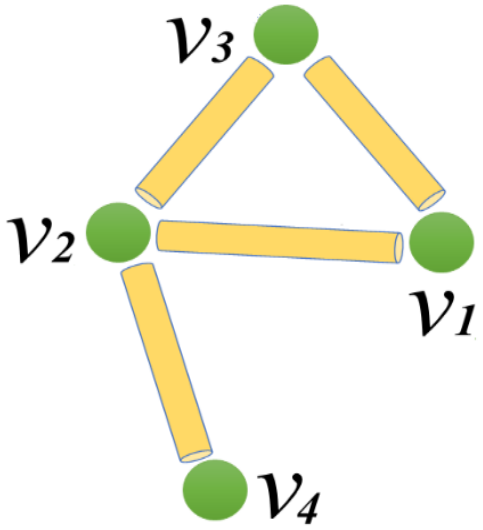
Definition: Graph

A generalization of the simple concept of a set of nodes and edges.

Representation: Graph $G = (V, E)$ *consists set of vertices denoted by* V*, or by* $V(G)$ *and set of edges* E*, or* $E(G)$*.*

Example: A simple (undirected) graph:

$G = (V, E),\ V = \{v_1, v_2, v_3, v_4\},\ E = \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\}, \{v_2, v_4\}\}$

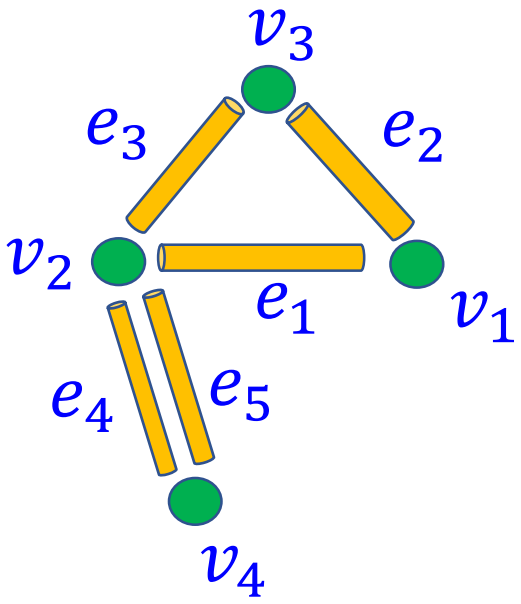


Graph Theory

Definition: Multigraph graph *Graph* $G = (V, E, f)$ *consists set of vertices denoted by* V*, set of edges* E*, and a function* f*:* $E \rightarrow \{\{v_2, v_4\} | v_2, v_4 \in V, v_2 \neq v_4\}$*. The edges* e_4 *and* e_5 *are called multiple or parallel edges if* $f(e_4) = f(e_5)$.

Example:

$G = (V, E, f)$, $V = \{v_1, v_2, v_3, v_4\}$, $E = \{e_1, e_2, e_3, e_4, e_5\}$



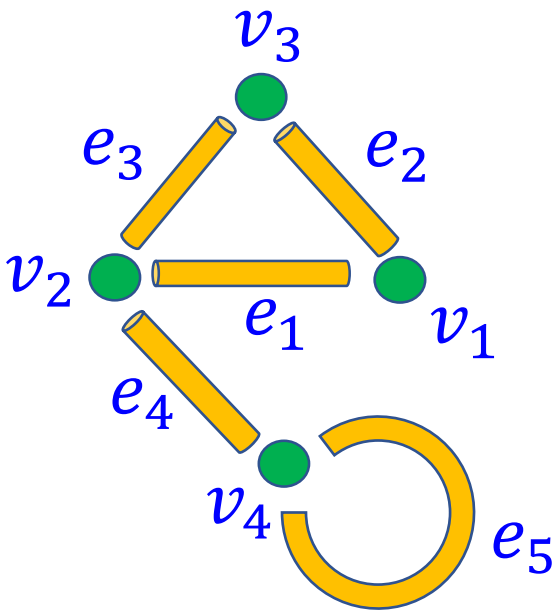
Graph Theory

Definition: Pseudograph graph *Graph* $G = (V, E, f)$ *consists set of vertices denoted by* V, *set of edges* E, *and a function* f: $E \rightarrow \{\{v_4\} | v_4 \in V, e_5 = \{v_4, v_4\}\}$. *Loop is allowed.*

Example:

$G = (V, E, f)$, $V = \{v_1, v_2, v_3, v_4\}$, $E = \{e_1, e_2, e_3, e_4, e_5\}$.

Application: Quantum theory (see later)



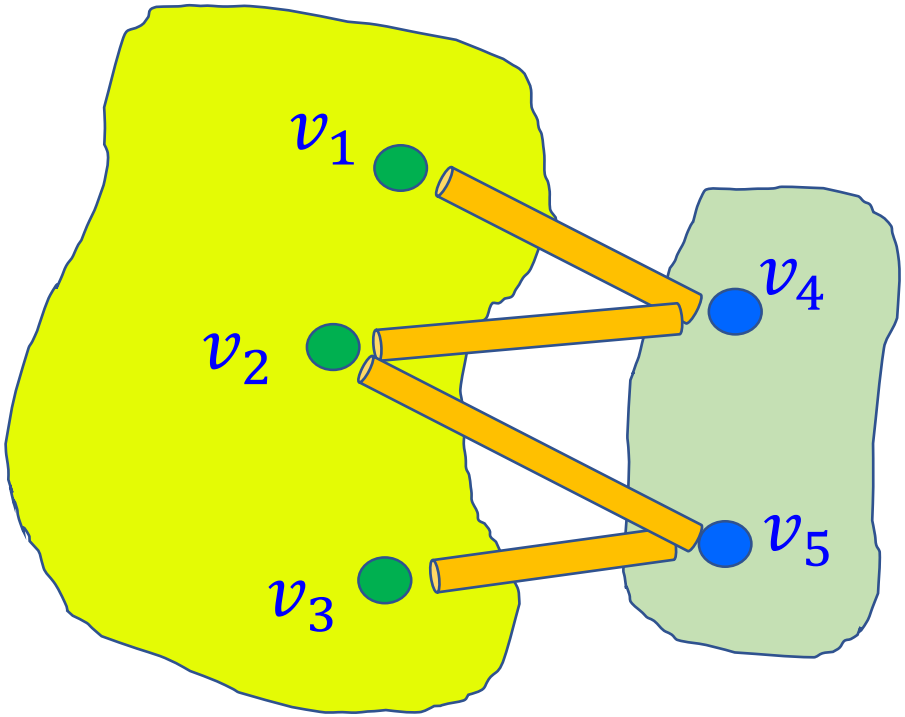
Graph Theory

Bipartite graph Graph $G = (V = \{V_1, V_2\}, E)$

Example (A protein-ligand binding complex):

$G = (V = \{V_1, V_2\}, E), V_1 = \{v_1, v_2, v_3\}, V_2 = \{v_4, v_5\},$

$E = \{\{v_1, v_4\}, \{v_2, v_4\}, \{v_2, v_5\}, \{v_3, v_5\}\}$



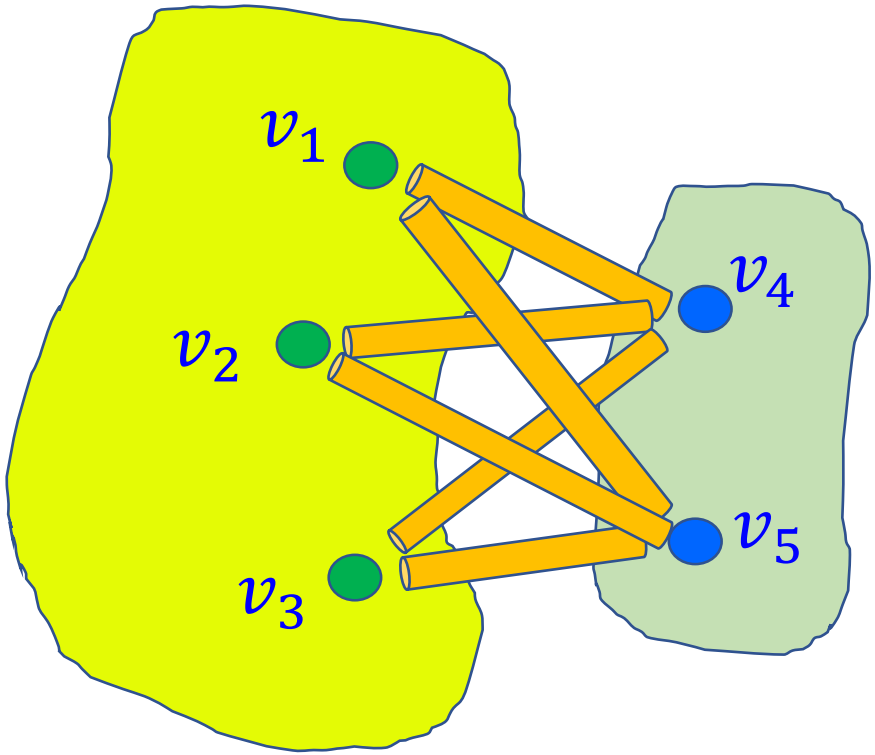
Graph Theory

Complete bipartite graph Graph $G = (V = \{V_1, V_2\}, E)$

Example (A protein-ligand binding complex):

$G = (V = \{V_1, V_2\}, E), V_1 = \{v_1, v_2, v_3\}, V_2 = \{v_4, v_5\},$

$E = \{\{v_1, v_4\}, \{v_1, v_5\}, \{v_2, v_4\}, \{v_2, v_5\}, \{v_3, v_4\}, \{v_3, v_5\}\}$

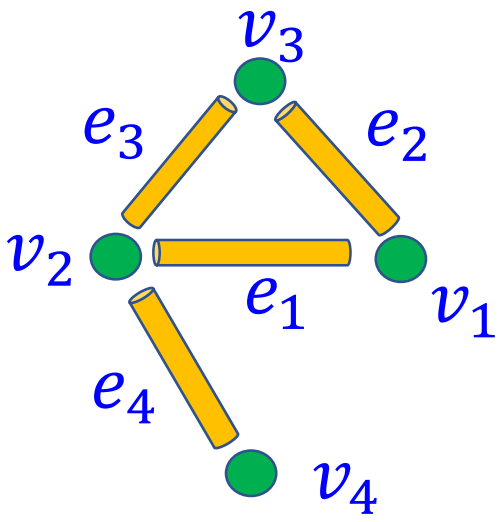


Graph Theory

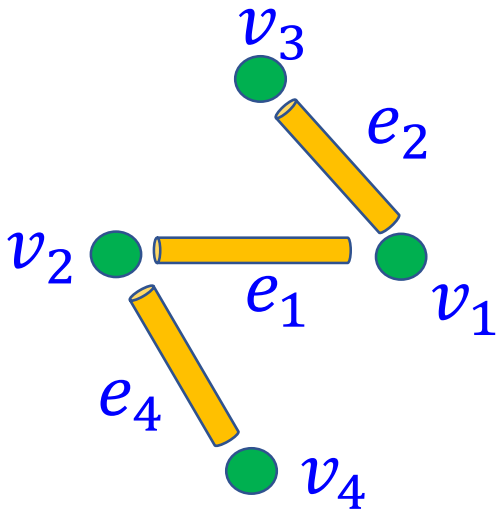
Subgraphs: *A subgraph of a graph* $G = (V, E)$ *is a graph* $F = (V', E')$ *where* $V' \subseteq V$, *and* $E' \subseteq E$.

Example: $G = (V, E), V = \{v_1, v_2, v_3, v_4\},$

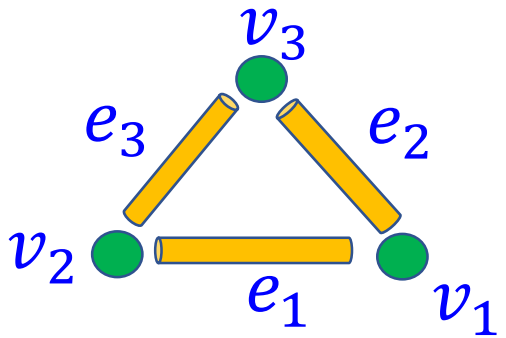
$$E = \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\}, \{v_2, v_4\}\}$$



G



$G_1 = (V_1, E_1)$



$G_2 = (V_2, E_2)$

$G = G_1 \cup G_2$, where $E = E_1 \cup E_2, V = V_1 \cup V_2$

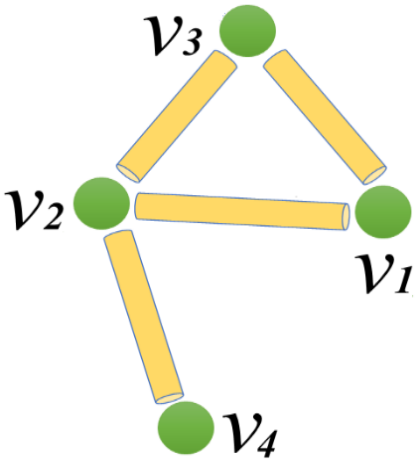
Algebraic graph Theory

Graph representations: Degree matrix (D), Laplacian matrix (L) and Adjacency matrix (A).

Example: A simple (undirected) graph:

$G = (V, E)$, $V = \{v_1, v_2, v_3, v_4\}$,

$E\ \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\}, \{v_2, v_4\}\}$



$$D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad L = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

$$L = D - A$$

$$\sum_i d_i = 2 + 3 + 2 + 1 = 2|E| = 8$$

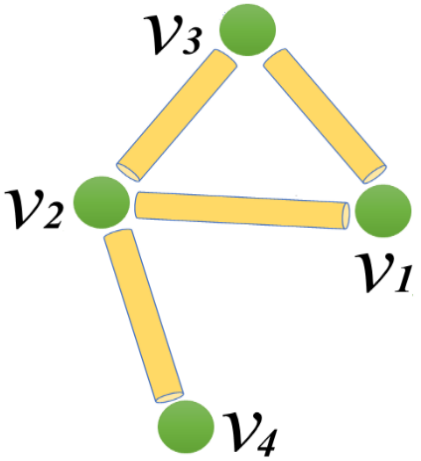
Degree of node i

Total # of edges

Algebraic Graph Theory

Laplacian matrix (L_G) is symmetric and has real valued entries. So it is self-adjoint and thus has real, non-negative eigenvalues:

- $0 \leq \lambda_1^L \leq \lambda_2^L \cdots 0 \leq \lambda_{\text{Max}}^L$
- $\lambda_1^L = 0$ for L_G
- $\lambda_2^L > 0$ if G is connected
- Multiplicity of 0 as an eigenvalue of L_G is equal to the number of connected components of G (the topology).



Let L_G be a symmetric matrix with eigenvalues $\lambda_1^L \leq \lambda_2^L \cdots 0 \leq \lambda_{\text{Max}}^L$. Then

- $\lambda_1^L = \min_{x \neq 0} \frac{x^T L_G x}{x^T x}$
- $\lambda_2^L = \min_{x \neq 0, x \perp x_1^L} \frac{x^T L_G x}{x^T x}$
- $\lambda_{\text{Max}}^L = \max_{x \neq 0} \frac{x^T L_G x}{x^T x}$

$$L = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

Graph Theory

Hypergraph: A hypergraph (H) is a generalization of a graph in which an edge can join any number of vertices.

$H = (X, E)$, where X is a set of nodes or vertices and E is a set of hyperedges, which are a subset of the power set such that .

$E \subseteq \wp(S) \setminus \{\emptyset\}$.

A power set $\wp(S)$ of set S is the set of all subsets of S, including the empty set $\{\emptyset\}$ and S itself.

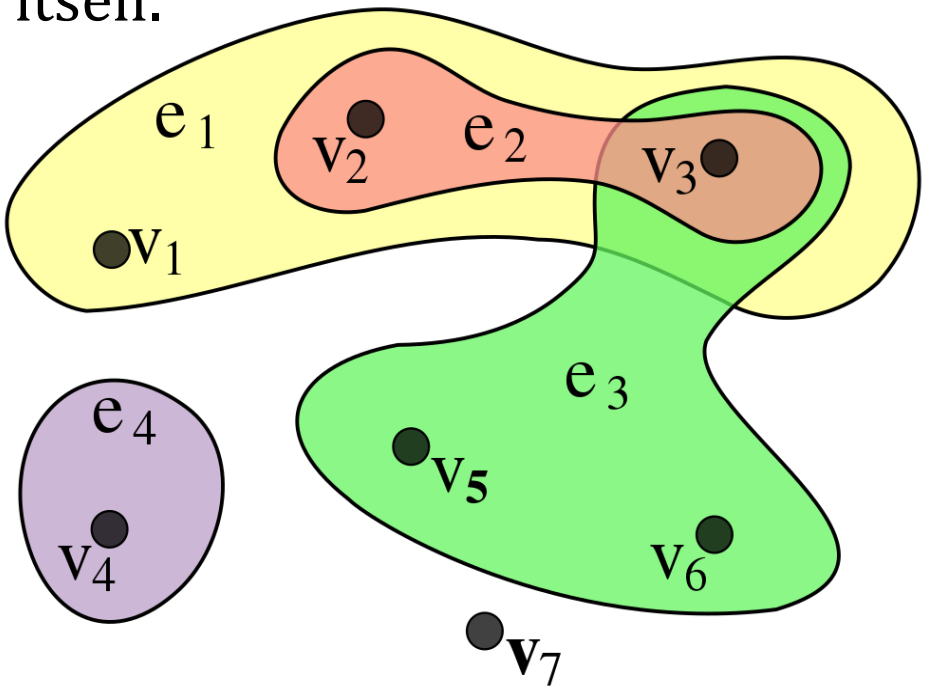
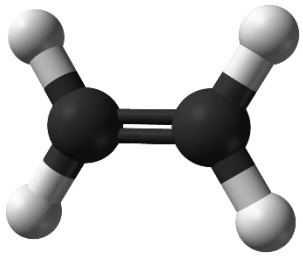


Image credit: Kilom691

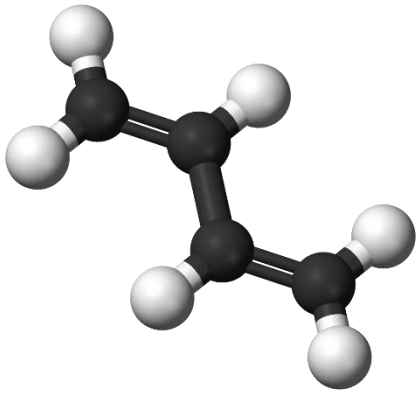
Hückel molecular orbital theory in QM



The Huckel's rule for conjugated π system:

Molecular orbitals are constructed by linear combination of atomic orbitals:

$$\Psi = \sum_j c_j \phi_j \,, \qquad \sum_j c_j^2 = 1$$

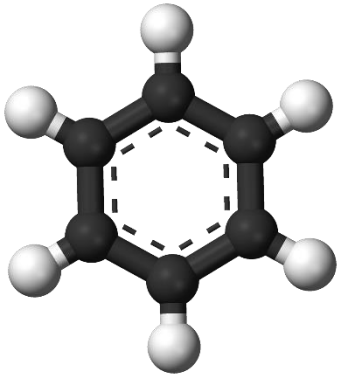


The Schrodinger equation:

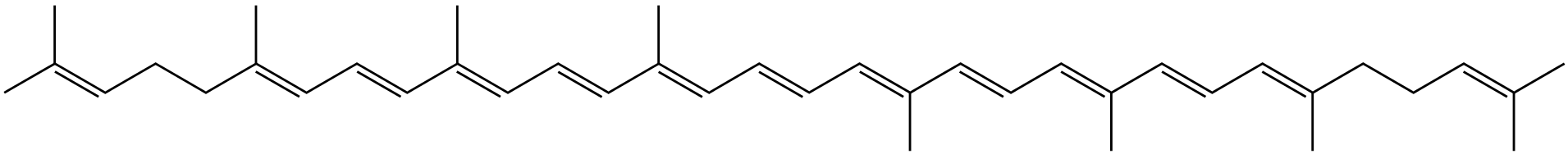
$$\hat{H}\,\Psi^i = E^i \Psi^i$$

where

$$H_{ij} = \int \phi_i \hat{H}\, \phi_j dr = \begin{cases} \alpha, & i = j \\ \beta, & i \neq j \end{cases}$$

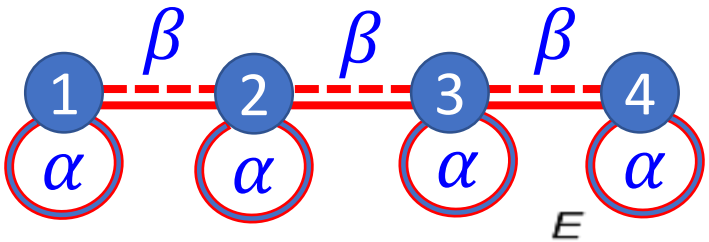


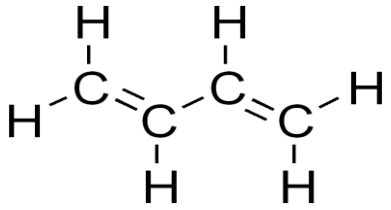
$$S_{ij} = \int \phi_i\, \phi_j dr = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$



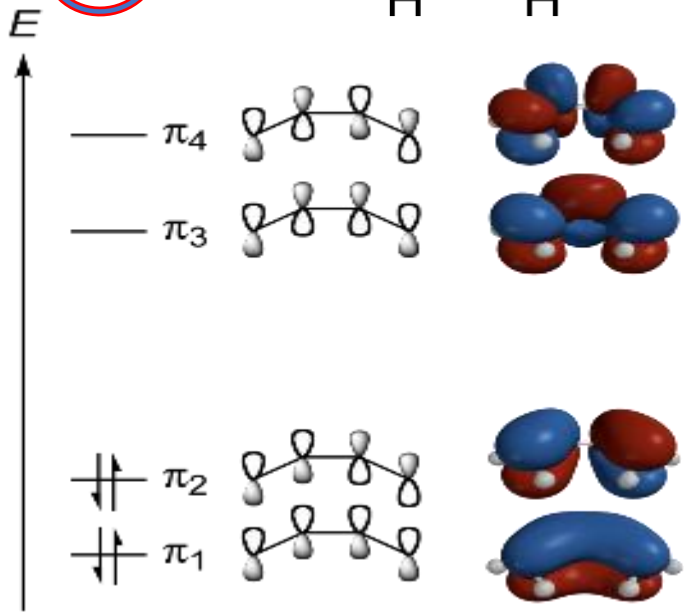
Graph representation of Hückel molecular orbital theory

The conjugated π system is modeled as a seudograph:





$$A = \begin{bmatrix} \alpha & \beta & 0 & 0 \\ \beta & \alpha & \beta & 0 \\ 0 & \beta & \alpha & \beta \\ 0 & 0 & \beta & \alpha \end{bmatrix} = H \text{ Hamiltonian}$$



Eigenvalues:

$$\lambda^A_{1,2,3,4} = \alpha + \frac{\sqrt{5} \pm 1}{2}\beta, \alpha - \frac{\sqrt{5} \mp 1}{2}\beta$$

Eigenvectors:

$$\Psi_1 \approx 0.372\phi_1 + 0.602\phi_2 + 0.602\phi_3 + 0.372\phi_4$$
$$\Psi_2 \approx 0.602\phi_1 + 0.372\phi_2 - 0.372\phi_3 + 0.602\phi_4$$
$$\Psi_3 \approx 0.602\phi_1 - 0.372\phi_2 - 0.372\phi_3 + 0.602\phi_4 \text{ and}$$
$$\Psi_4 \approx 0.372\phi_1 - 0.602\phi_2 + 0.602\phi_3 - 0.372\phi_4$$

The HOMO–LUMO gap: $\Delta E = \lambda^A_2 - \lambda^A_3 \cong 1.236\beta$

Image credit: Ben Mills

Graph Theory

Graph partition: Partition a graph $G = (V, E)$ into smaller components with certain properties.

Fiedler eigenvalue and eigenvector: the second smallest eigenvalue (λ_2^L) provides a lower bound on ratio-cut partition: $c \geq \frac{\lambda_2^L}{|E|}$. The associated eigenvector, Fiedler vector bisects the graph into two sections based on the sign of the eigenvector (i.e., spectral bisection based on algebraic connectivity).

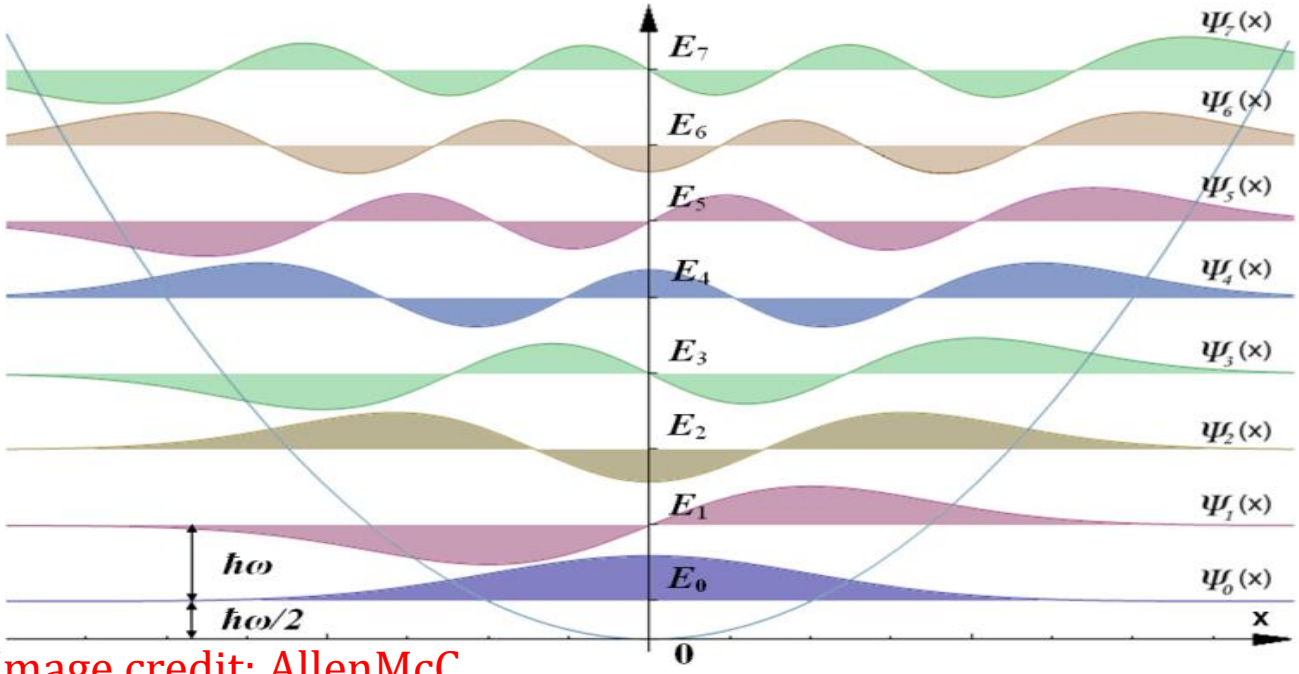
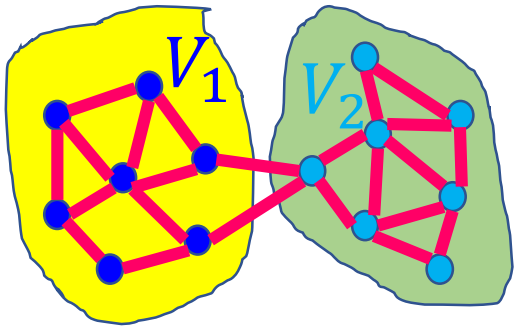
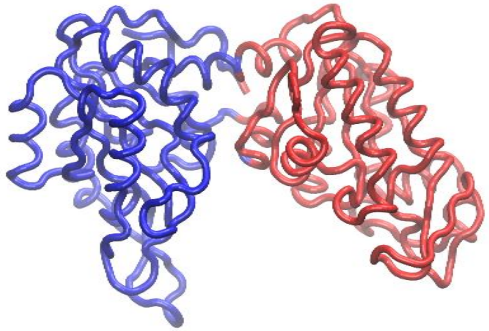


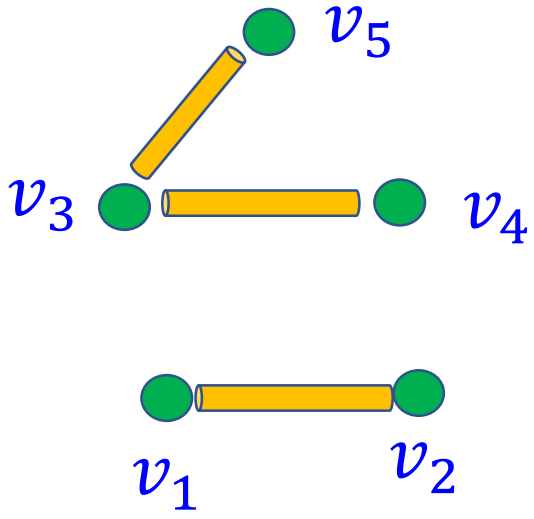
Image credit: AllenMcC.





Spectral bisection

Example I:



$$\text{Vec} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & -\frac{1}{\sqrt{3}} & 0 & 0 & -\frac{2}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{6}} \end{bmatrix} \quad L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

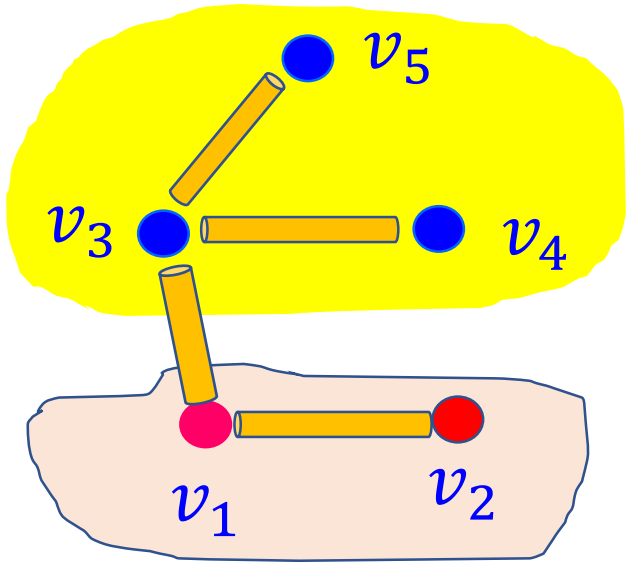
$$\text{Eig} = [\ 0,\quad 0,\quad 1,\quad 2,\quad 3]$$

Two disconnected components (harmonic part due to the topology)

Spectral bisection

Example II:

$$L = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 3 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$



$$\text{Vec} = \begin{bmatrix} 0.45 & 0.34 & 0 & -0.70 & 0.44 \\ 0.45 & 0.70 & 0 & 0.54 & -0.14 \\ 0.45 & -0.20 & 0 & -0.32 & -0.81 \\ 0.45 & -0.42 & -0.70 & 0.24 & 0.26 \\ 0.45 & -0.42 & 0.70 & 0.24 & 0.26 \end{bmatrix}$$

$$\text{Eig} = [\,0,\quad 0.52,\quad 1.00,\quad 2.31, 4.17]$$

Graph modularity for domain classification

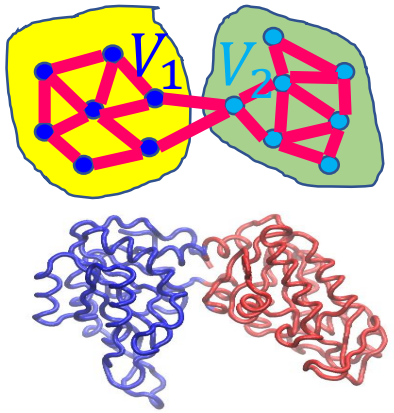
Graph modularity (Q): The fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random for a given connectivity:

$$Q = \frac{1}{2|E|}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2|E|}\right)\frac{s_i s_j + 1}{2}$$

where s_i is a membership variable,

$$s_i = \begin{cases} 1 & v_i \in V_1 \\ -1 & v_i \in V_2 \end{cases}$$

Expected # of edges between nodes i and j



Properties:

- $-1 \leq Q \leq 1$
- $Q = 0 \Rightarrow$ all nodes in one group
- $Q > 0$ more edges in V_1
- $Q < 0$ more edges in V_2
- Selecting s_i to maximize Q. When Q is optimized, the modularity matrix $\left(B_{ij} = A_{ij} - \frac{d_i d_j}{2|E|}\right)$ no positive eigenvalue.

Macromolecular flexibility analysis

Flexibility:

- Flexibility is a measure of biomolecular thermal stability.
- It correlates with biomolecular reactivity.
- It correlates with allosteric activation, inhibition, modulation, etc.
- It is associated with protein ordered-disordered transition.
- It can be measured by X-ray scattering in terms of atomic isotropic displacement (u), i.e., the B-factor:

$$B = 8\pi^2\langle u^2\rangle$$

PDB file:

```
ATOM 1 N   THR A 1 17.047 14.099 3.625 1.00 13.79
ATOM 2 CA  THR A 1 16.967 12.784 4.338 1.00 10.80
ATOM 3 C   THR A 1 15.685 12.755 5.133 1.00 9.19
ATOM 4 O   THR A 1 15.268 13.825 5.594 1.00 9.85
ATOM 5 CB  THR A 1 18.170 12.703 5.337 1.00 13.02
ATOM 6 OG1 THR A 1 19.334 12.829 4.463 1.00 15.06
ATOM 7 CG2 THR A 1 18.150 11.546 6.304 1.00 14.23
```

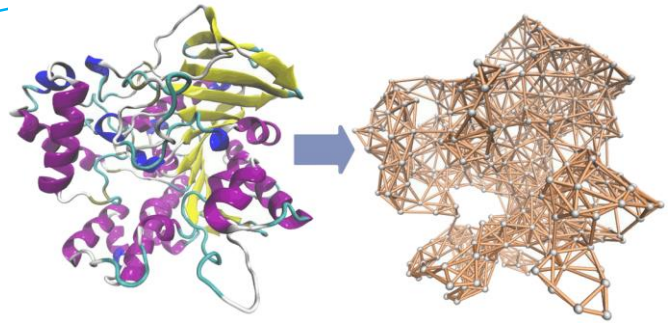
Gaussian Network Model (GNM)—Laplacian model

(Bahar, Atilgan and Erman, FD, 1997)

$$B_j^{\mathrm{GNM}} = \alpha \left(L^{-1}\right)_{jj},$$

$$L_{ij} = \begin{cases} -1, & i \neq j,\ r_{ij} \leq r_c \\ 0, & i \neq j,\ r_{ij} > r_c \\ -\sum_{j,j\neq i}^{N} L_{ij}, & i = j \end{cases}$$

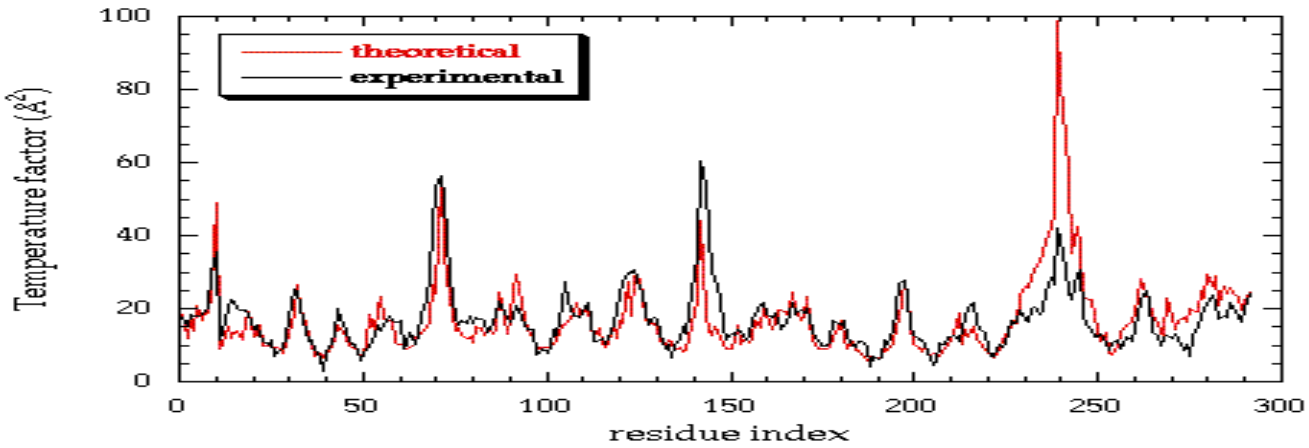
L is a Laplacian matrix, $O(N^3)$ method



$$\left(L^{-1}\right)_{jj} = \sum_{k=2}^{N} \frac{1}{\lambda_k} \left[u_k u_k^T\right]_{jj}$$

Moore–Penrose pseudoinverse

where α is fitting parameter, r_c a cutoff distance (7 Å is often used for C_α networks), λ_k the kth eigenvalue and u_k the kth eigenvector.



Weighted graph Laplacian (WGL) – Generalized GNM

$$B_j^{\mathrm{GGNM}} = \alpha\,(L^{-1})_{jj},$$

$$L_{ij} = \begin{cases} -\Phi\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|;\eta\right), & i \neq j \\ -\sum_{j,j\neq i}^{N} L_{ij}, & i = j \end{cases}$$

$$(L^{-1})_{jj} = \sum_{k=2}^{N} \frac{1}{\lambda_k}\left[u_k u_k^T\right]_{jj},$$

where the monotonic function satisfies

$\Phi\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|;\eta\right)=1$, as $\|\boldsymbol{r}_i - \boldsymbol{r}_j\| \to 0$,

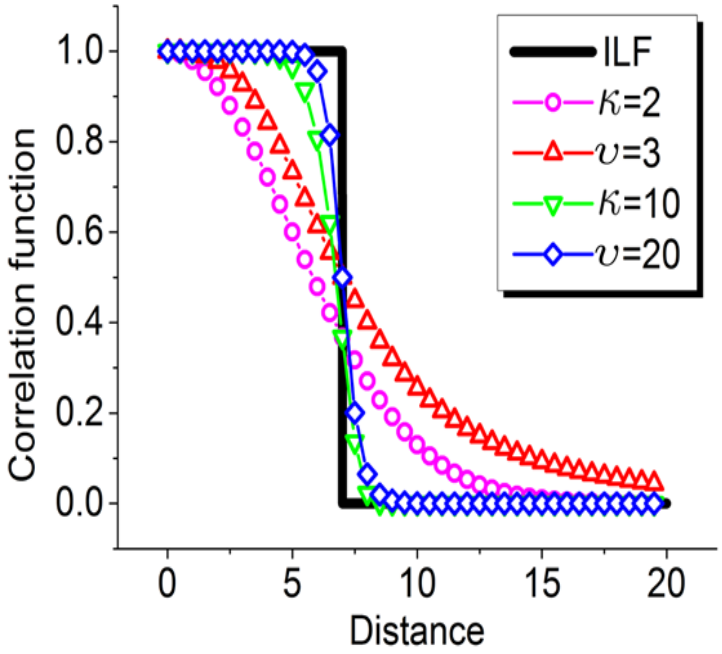
$\Phi\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|;\eta\right)=0$, as $\|\boldsymbol{r}_i - \boldsymbol{r}_j\| \to \infty$.

GGNM becomes GNM (ILF) as $\eta \to 0$.

Examples:

$$\Phi\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|;\eta\right) = e^{-\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|/\eta\right)^{\kappa}},$$

$$\Phi\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|;\eta\right) = \frac{1}{1+\left(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|/\eta\right)^{\upsilon}}$$



(Xia, Opron and Wei, JCP, 2015)

Multiscale GNM (mGNM)

$$B_j^{\mathrm{mGNM}} = \alpha \left(L^{-1}\right)_{jj},$$

$$L_{ij} = \sum_{k=1}^{n} c_k L_{ij}\left(\Phi\left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|; \eta^k\right)\right)$$

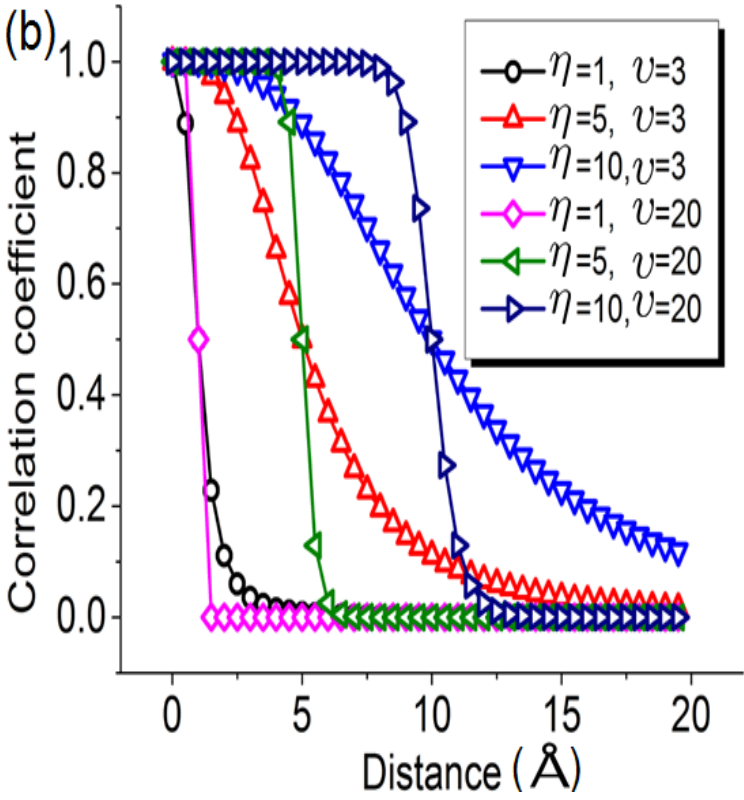
where c_k will be optimized.

Here the kernel is a monotonic radial basis function

$$\Phi\left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|; \eta^k\right) = e^{-\left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|/\eta^k\right)^{\kappa}},$$

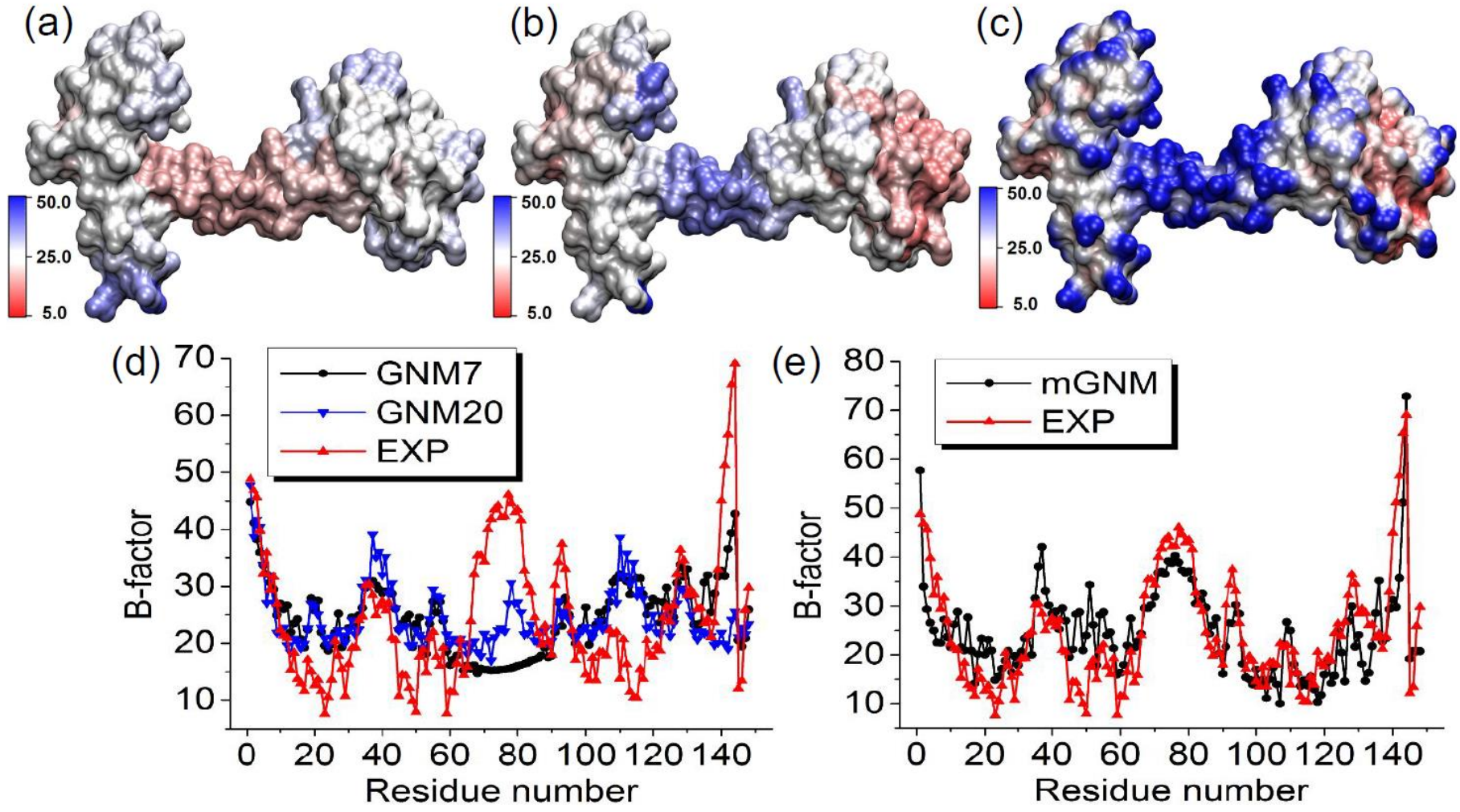
$$\Phi\left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|; \eta^k\right) = \frac{1}{1 + \left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|/\eta^k\right)^{\nu}}$$

We typical use two or three scales for macromolecules.



(Xia, Opron and Wei, JCP, 2016)

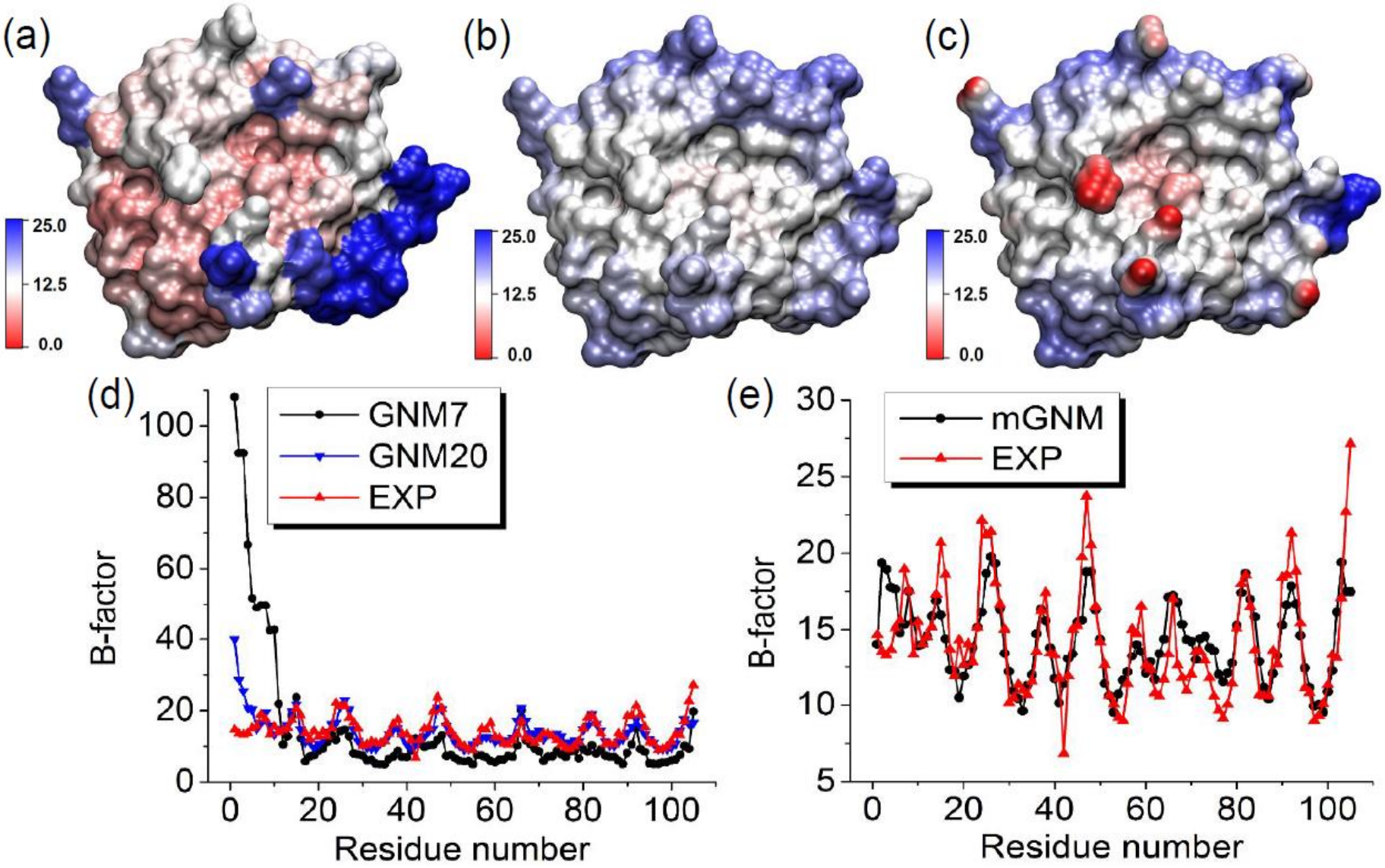
Comparison of GNM and Multiscale GNM



1V70

(Xia, Opron and Wei, JCP, 2016)

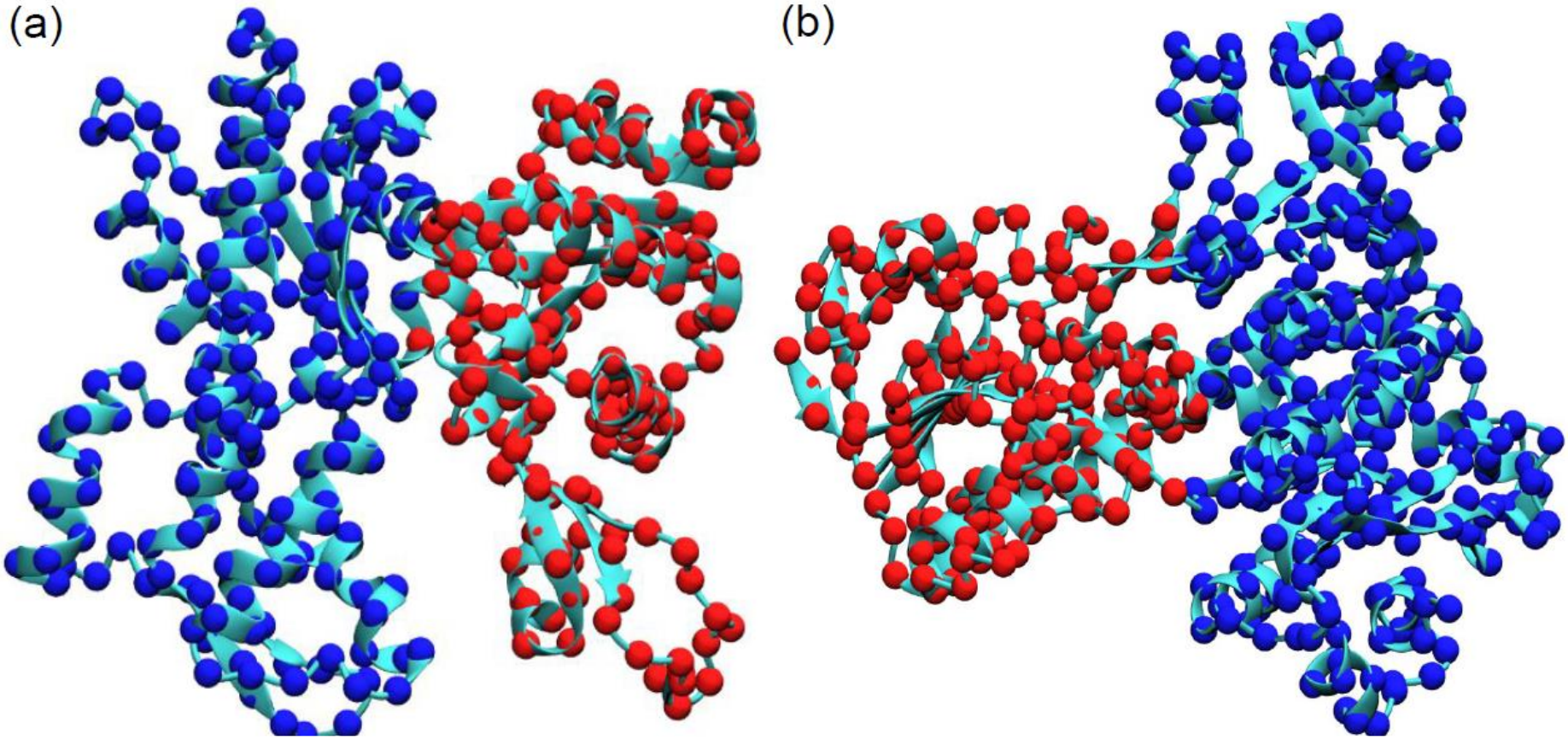
Comparison of GNM and Multiscale GNM



2HQK

(Xia, Opron and Wei, JCP, 2016)

Multiscale GNM based domain analysis



Protein domain decomposition with Type-1 mGNM. The first non-zero eigenvector (Fiedler vector) is used to decompose the protein into two domains. (a) Protein 1ATN (chain A); (b) protein 3GRS.

Multiscale Anisotropic Network Model

ANM Hessian matrix (Doruker, Atilgan, & Bahar, 2000):

$$H_{ij} = \frac{-\gamma}{r_{ij}^2}\begin{bmatrix} (x_j - x_i)(x_j - x_i) & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)(y_j - y_i) & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)(z_j - z_i) \end{bmatrix}$$

Hessian matrix of multiscale ANM (mANM): $H_{ij} = \sum_n c_n H_{ij}^n$

$$H_{ij}^n = \frac{-\Phi_{ij}^n}{r_{ij}^2}\begin{bmatrix} (x_j - x_i)(x_j - x_i) & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)(y_j - y_i) & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)(z_j - z_i) \end{bmatrix}$$

where $\Phi_{ij}^n = \Phi\left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|; \eta^n\right)$

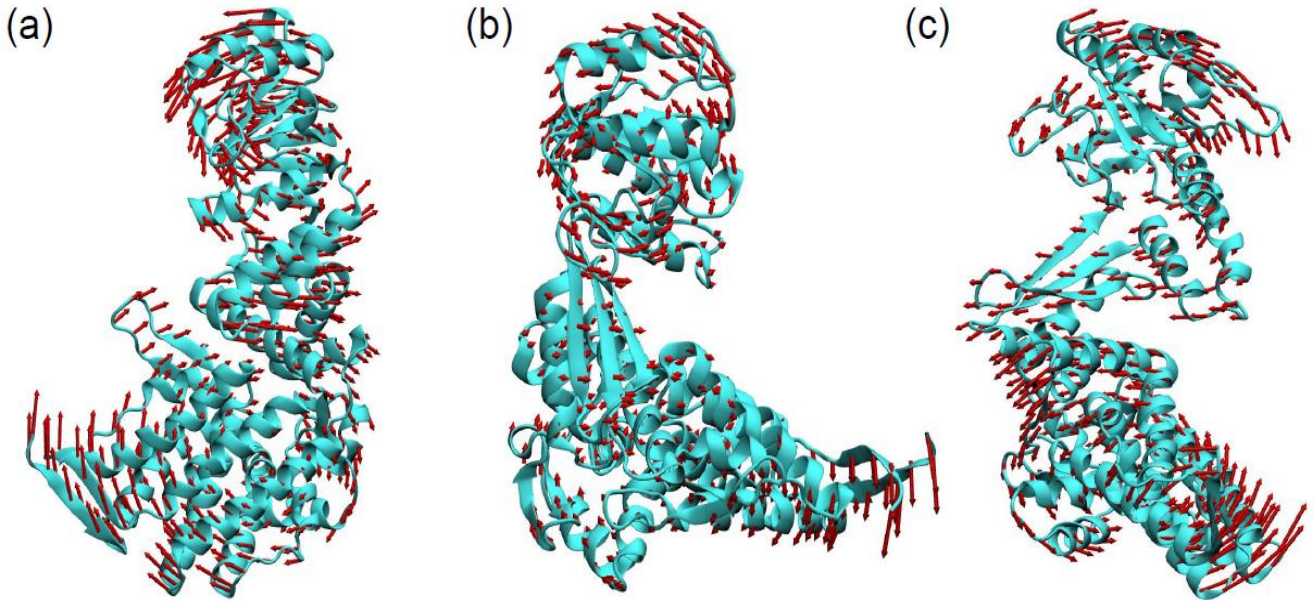
The rigidity index of mANM:

$$\mu_i = \sum_{j\neq i} \frac{\Phi_{ij}^n}{r_{ij}^2}\left[\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2 + \left(z_j - z_i\right)^2\right] = \sum_{j\neq i} \Phi_{ij}^n$$

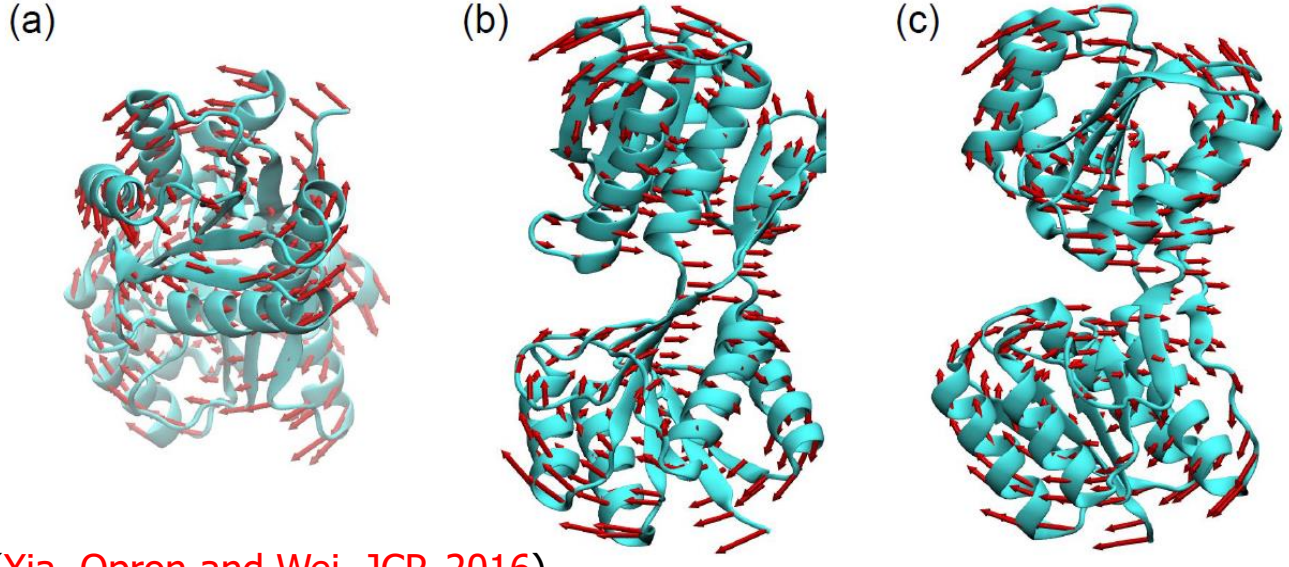
$$\mathrm{Min}_{c_n}\left\{\sum_i \left|\sum_n c_n \mu_i^n - \frac{1}{B_i^{\mathrm{Exp}}}\right|^2\right\}$$

(Xia, Opron and Wei, JCP, 2015)

Multiscale Anisotropic Network Model



The motions of 1GRU (chain A). The 7th, 8th, and 9th nANM modes are demonstrated in (a)–(c), respectively



The motions of 1URP (chain A). The 7th, 8th, and 9th mANM modes are demonstrated in (a)–(c), respectively.

(Xia, Opron and Wei, JCP, 2016)

Summary of Algebraic Graph Methods

Advantages:

GNM, ANM, and their generalizations are efficient approach for

- Protein flexibility (B-factor) regression and analysis.
- Protein or multi protein domain and hinge detection.
- Protein collective motion analysis.

Disadvantages:

- The accuracy of the B-factor prediction is relatively low.
- The computational complexity is relatively high, i.e., $O(N^3)$.
- They do not work well heterogeneous elements, i.e., the mixture of C, N, O, S, P, etc.
- They do not describe chemical, physical and biological interactions, i.e., hydrogen bonds, electrostatic effects, polarization, hydrophilicity and hydrophobicity.

Geometric graphs – Flexibility and rigidity index (FRI)

Motivation: To provide an atomistic shear modulus for biomolecules (Continuous elasticity with atomistic rigidity):

$$\rho\ddot{\boldsymbol{w}} = \nabla\lambda\nabla\cdot\boldsymbol{w} + \nabla\mu\cdot\left[\nabla\boldsymbol{w} + (\nabla\boldsymbol{w})^T\right] + (\lambda+\mu)\nabla\nabla\cdot\boldsymbol{w} + \mu\nabla^2\boldsymbol{w} + \boldsymbol{f}$$

where $\mu(\boldsymbol{r}) = \sum_j \Phi\left(\left\|\boldsymbol{r}-\boldsymbol{r}_j\right\|;\eta_j\right)$ is an atomistic shear modulus, describing protein density and can be used for surface modeling.

Rigidity index: $\mu_i = \mu(\boldsymbol{r}_i) = \sum_j \Phi\left(\left\|\boldsymbol{r}_i-\boldsymbol{r}_j\right\|;\eta_j\right) \approx L_{ii}$

Flexibility index: $f_i = \frac{1}{\mu_i} = \frac{1}{\sum_j \Phi\left(\left\|\boldsymbol{r}_i-\boldsymbol{r}_j\right\|;\eta_j\right)} \approx (L_{ii})^{-1} = B_i^{\mathrm{FRI}} \neq (L^{-1})_{ii}$

B-factor analysis: $\mathrm{Min}_{c,b} \sum_i \left| cB_i^{\mathrm{FRI}} + b - B_i^{\mathrm{Exp}} \right|^2$

Complexity: $O(N^2)$ and with **cell lists** $O(N)$.

Multiscale FRI (mFRI): $L_{ii}^n = \sum_{j,j\neq i} \Phi\left(\left\|\boldsymbol{r}_i-\boldsymbol{r}_j\right\|;\eta_j\right)$

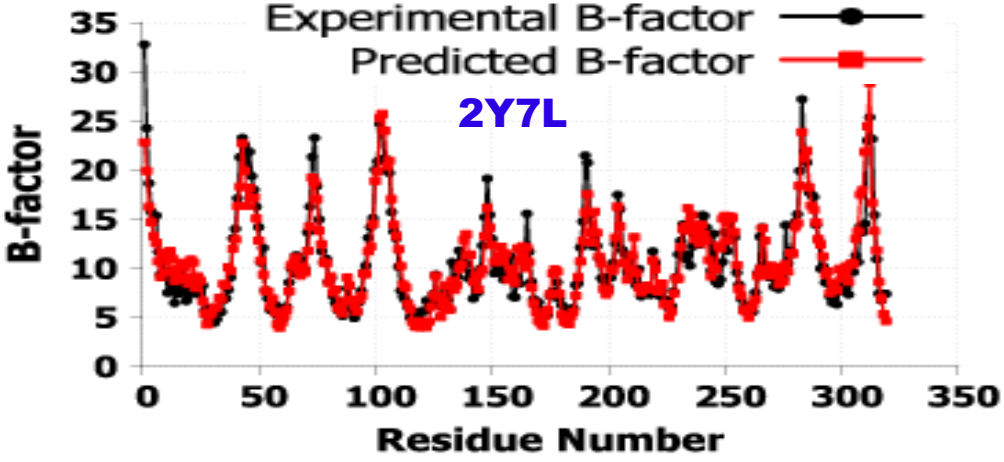
B-factor analysis: $\mathrm{Min}_{c_n} \sum_i \left| \sum_n c_n \left(L_{ii}^n\right)^{-1} + c_0 - B_i^{\mathrm{Exp}} \right|^2$

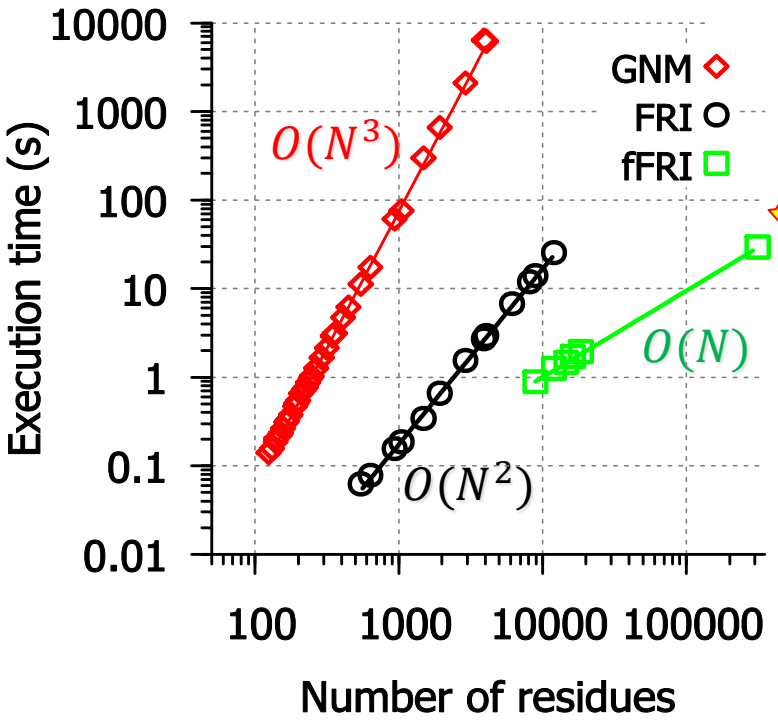
(Xia, Opron and Wei, JCP, 2013; JCP 2014; Opron, Xia and Wei, JCP, 2013; JCP 2014)

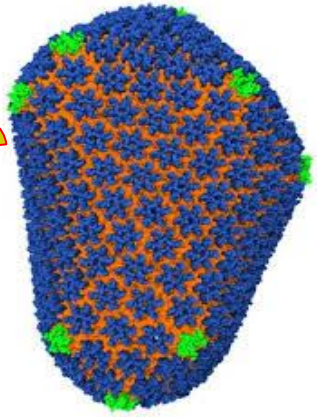
Weighted graph Laplacian, Flexibility rigidity index (FRI)

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

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



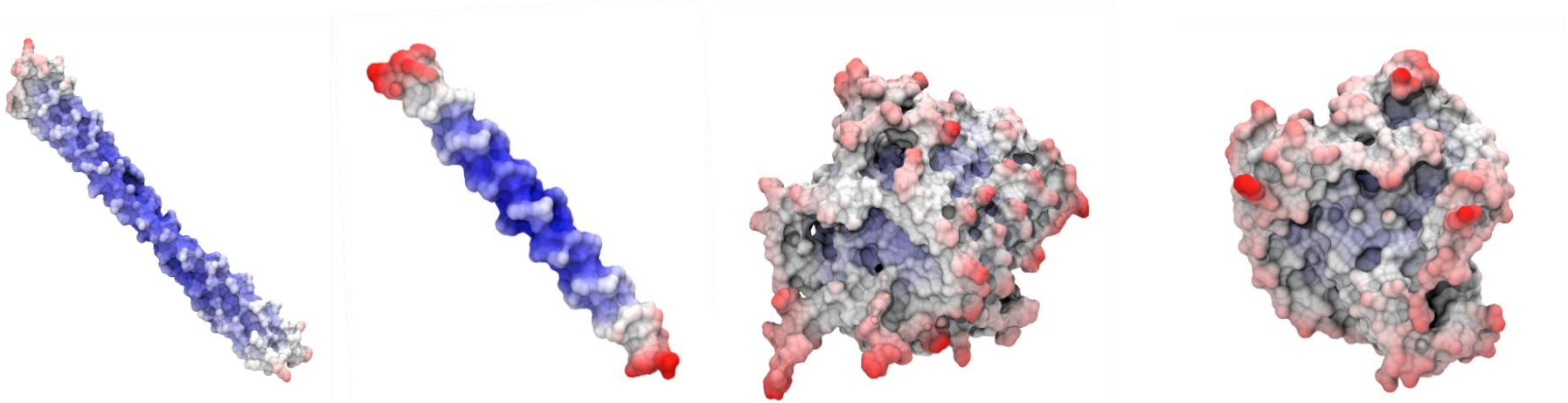




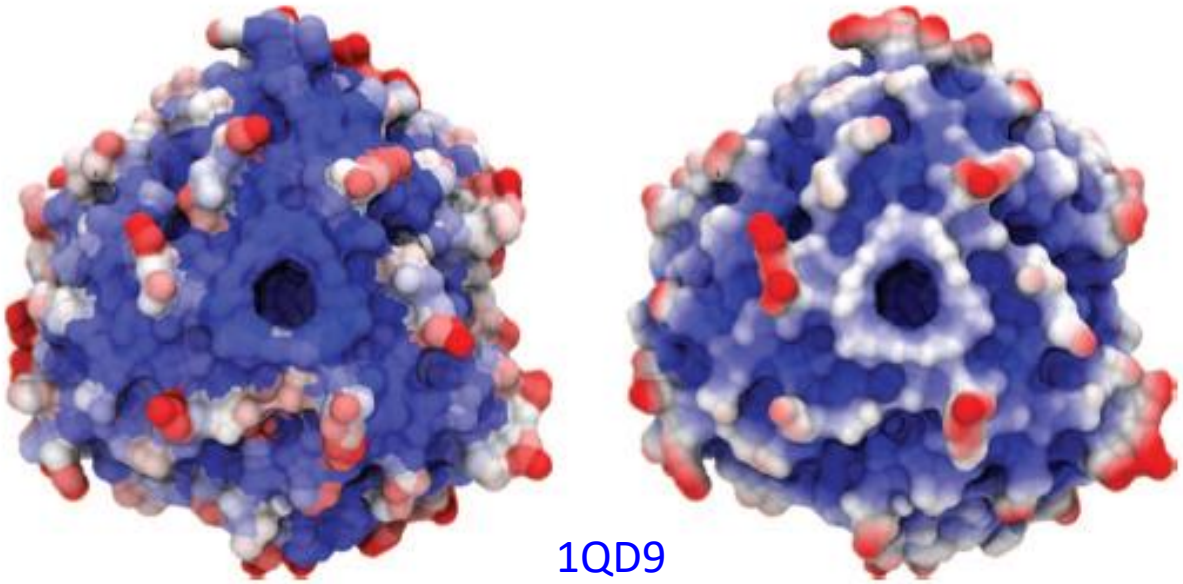
HIV capsid (313,236 residues) would takes GNM 120 years to compute!

Weighted graph Laplacian, Flexibility rigidity index (FRI)

Flexibility visualization



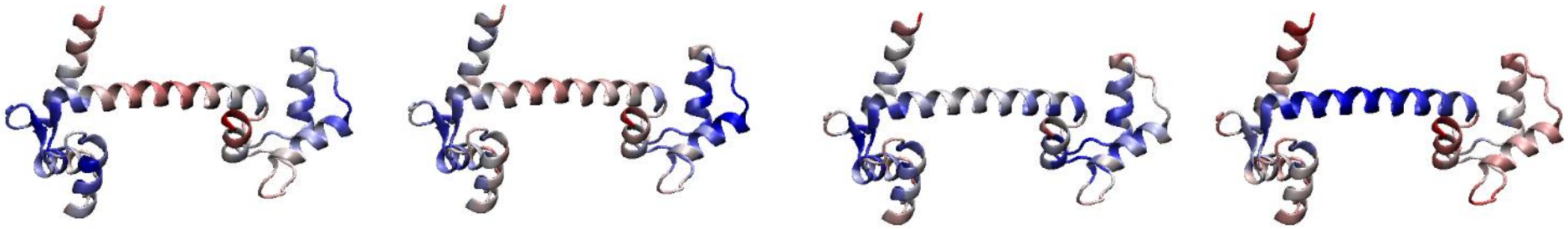
Keratin (3NTU), Collagen (1CAG), Fibroin (3UA0) and Amyloid fibrils (2RNM).



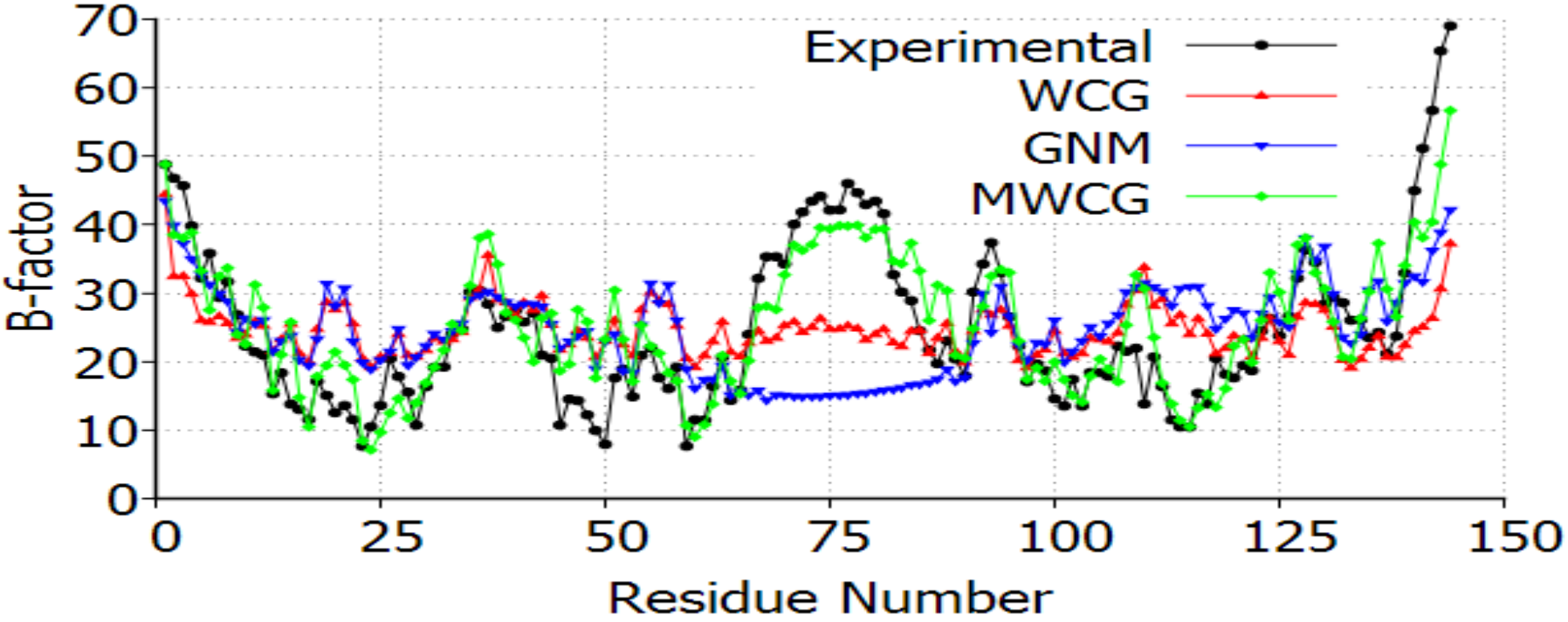
1QD9

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

Protein hinge detection



Experimental MWCG WCG GNM

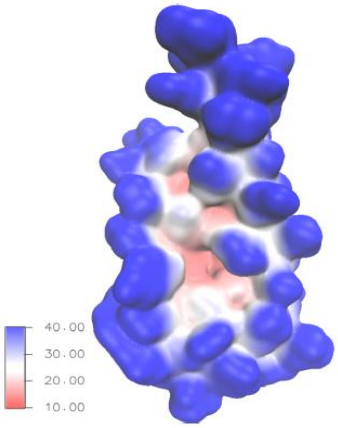


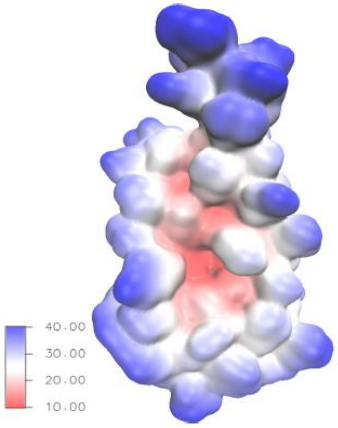
(Bramer and Wei, JCP, 2018)

Weighted graph Laplacian, Flexibility rigidity index (FRI)

1DF4 surface.

FRI based surface modeling



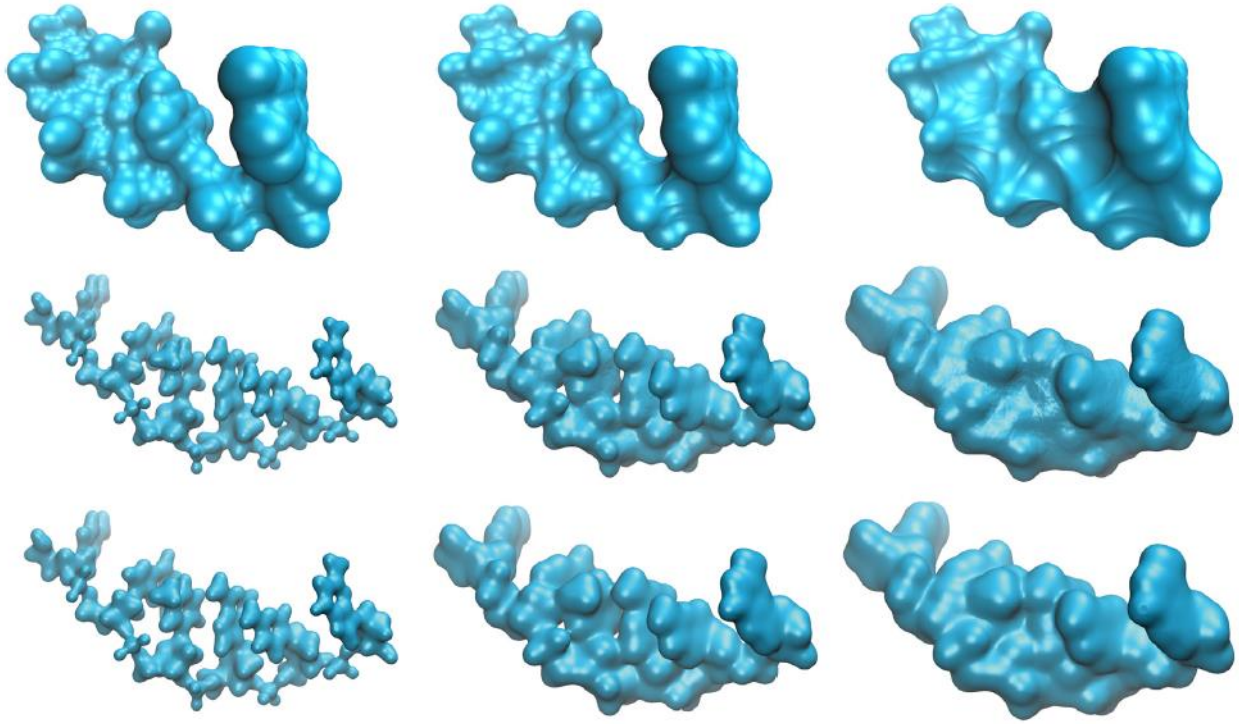


$$\mu(\boldsymbol{r}) = \sum_j \Phi(\|\boldsymbol{r} - \boldsymbol{r}_j\|; \eta_j)$$

(Nguyen, Xia and Wei, JCP, 2016)

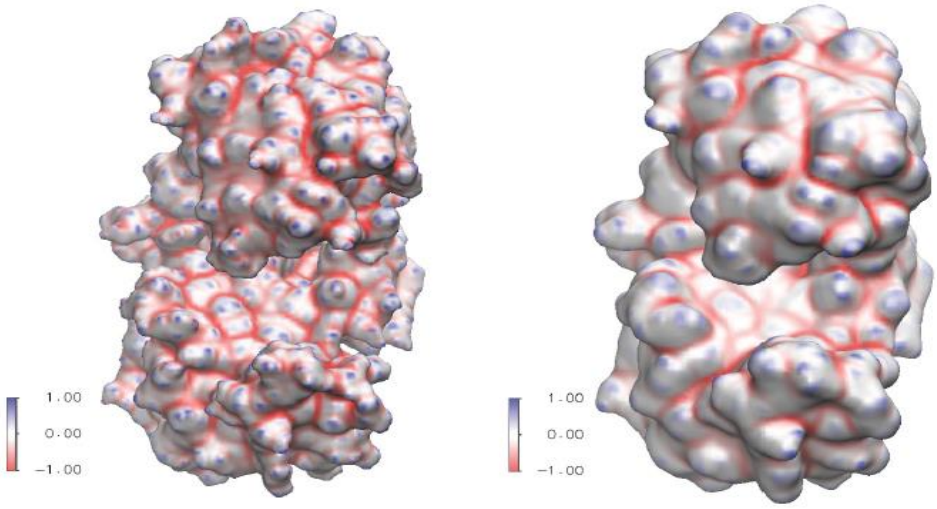
(Mu, Xia and Wei, JCAM, 2016)

Cross linking of DNA duplexes, PDBID:2GJB. Top row: MSMS. Middle row: FRI-Lorentz. Bottom row: FRI-Exponential.



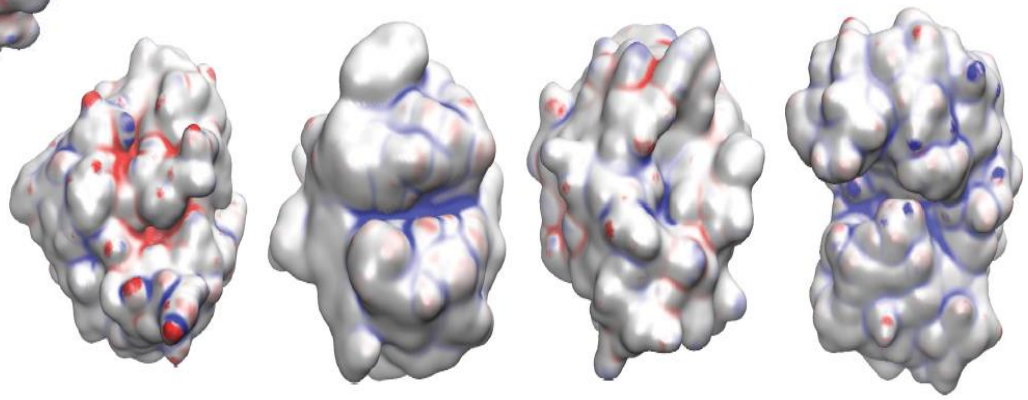
Weighted graph Laplacian, Flexibility rigidity index (FRI)

FRI based surface, curvature and electrostatic modeling

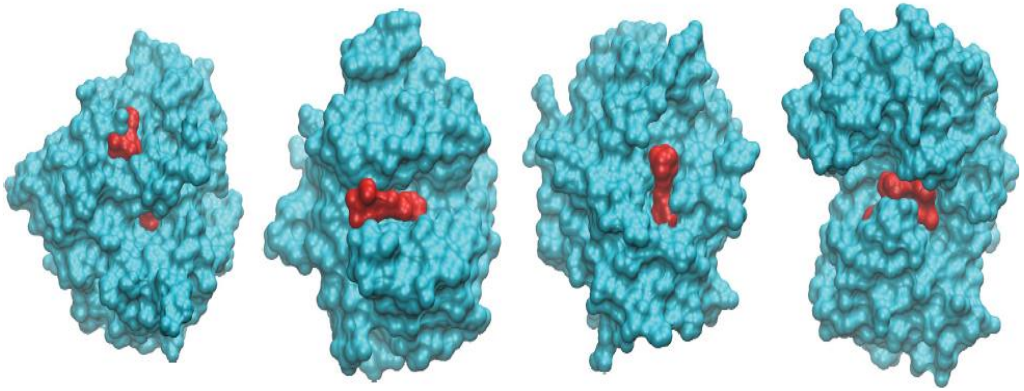


Minimal curvature of protein1PPL based on FRI representation.

(Mu, Xia and Wei, JCAM, 2016)



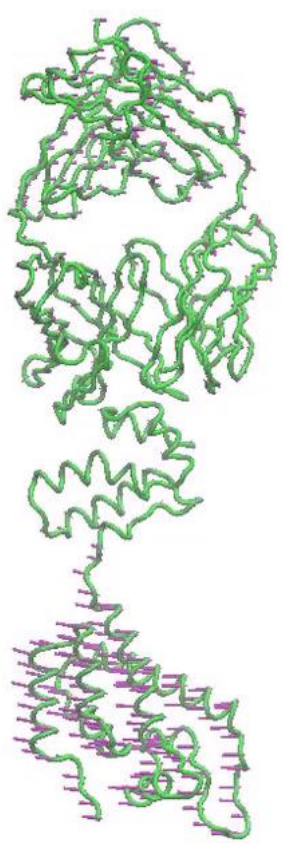
Protein binding site prediction using polarized curvatures. From left to right, 1ADS, 1BYH, 1EJN, and 2WEB. Top row: predicted binding sites. Bottom row: experimental binding sites

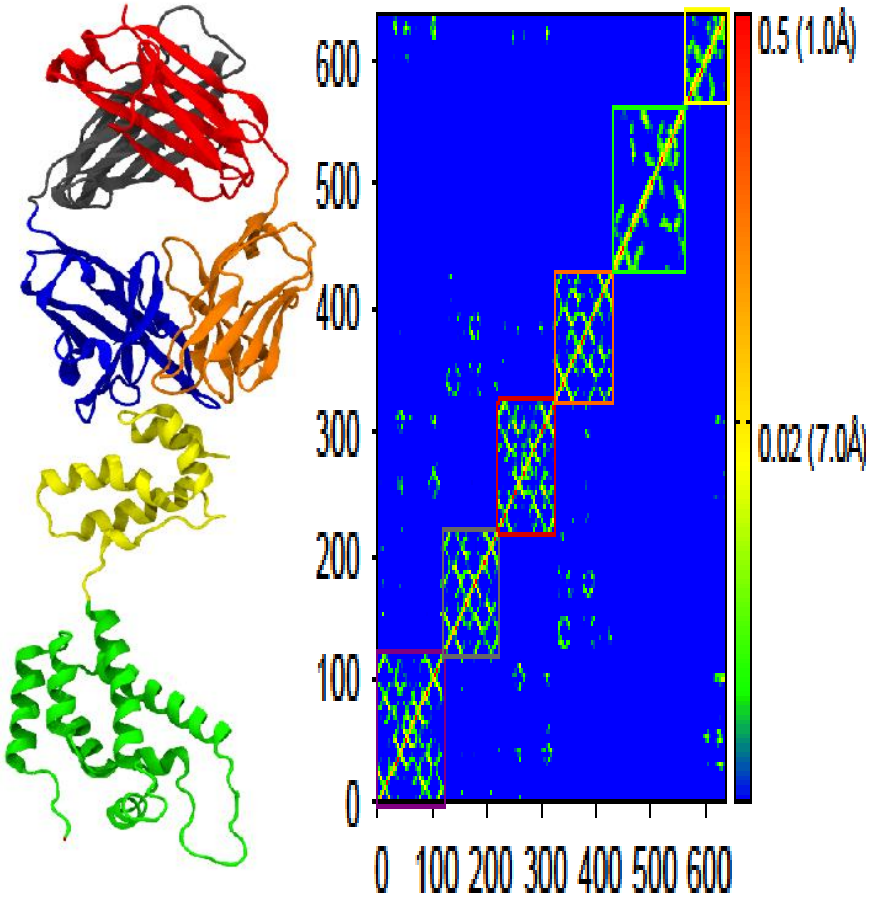


Weighted graph Laplacian, Flexibility rigidity index (FRI)

FRI based protein domain analysis

HIV (ID:1E6J)



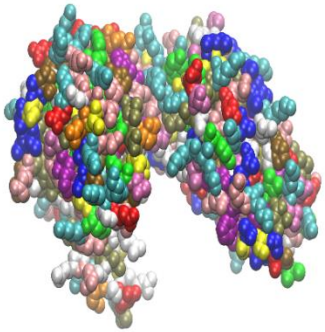


FRI

(Xia, Opron, Wei, JCP 2015)

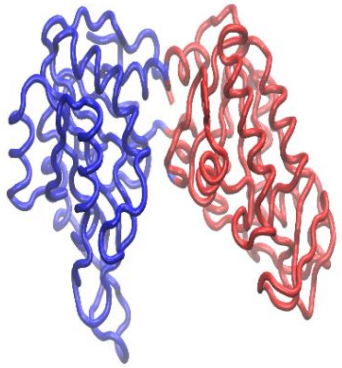
Weighted graph Laplacian, Flexibility rigidity index (FRI)

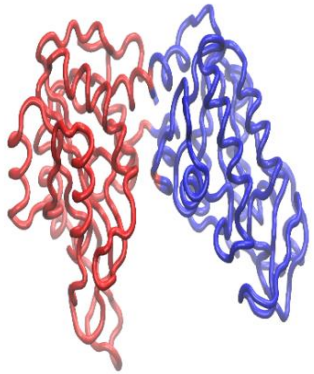
FRI based protein domain analysis

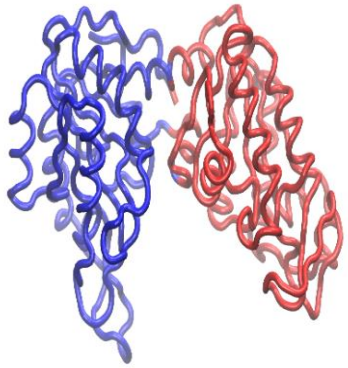


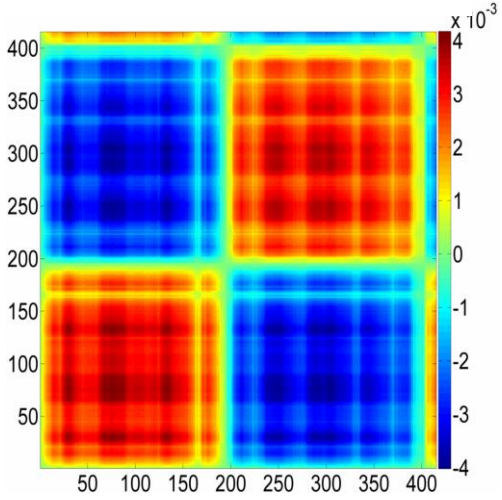
ID:3PGK

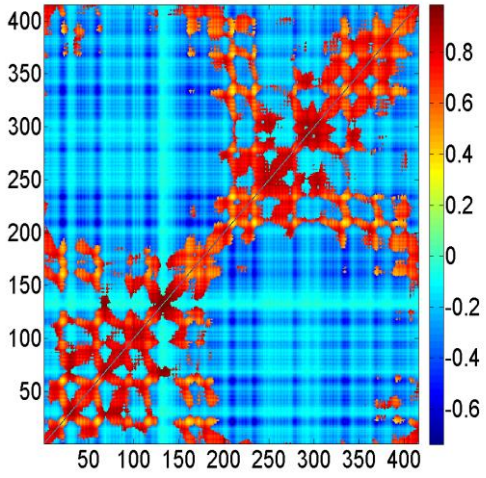
(Xia, Opron, Wei, JCP 2015)

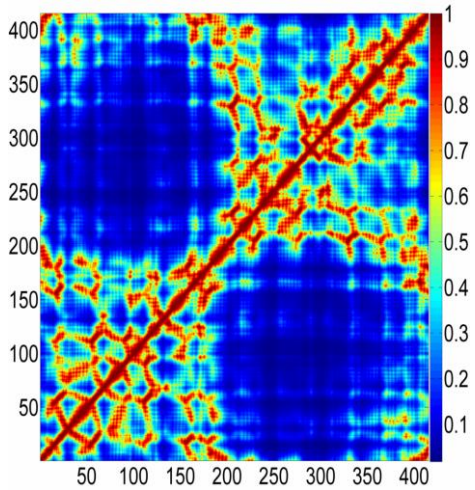












GGNM $O(N^3)$ **(Laplacian)**

Modularity $O(N^3)$ **(Graph theory)**

FRI $O(N)$

Vector Laplacian, Anisotropic flexibility rigidity index

FRI based protein motion analysis

$$\Phi_{uv}^{ij} \equiv \frac{\partial}{\partial u_i}\frac{\partial}{\partial v_j}\Phi\left(\left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|;\eta^n\right)$$

$$\Phi^{ij} \equiv \begin{pmatrix} \Phi_{xx}^{ij} & \Phi_{xy}^{ij} & \Phi_{xz}^{ij} \\ \Phi_{yx}^{ij} & \Phi_{yy}^{ij} & \Phi_{zz}^{ij} \\ \Phi_{zx}^{ij} & \Phi_{zy}^{ij} & \Phi_{zz}^{ij} \end{pmatrix}$$

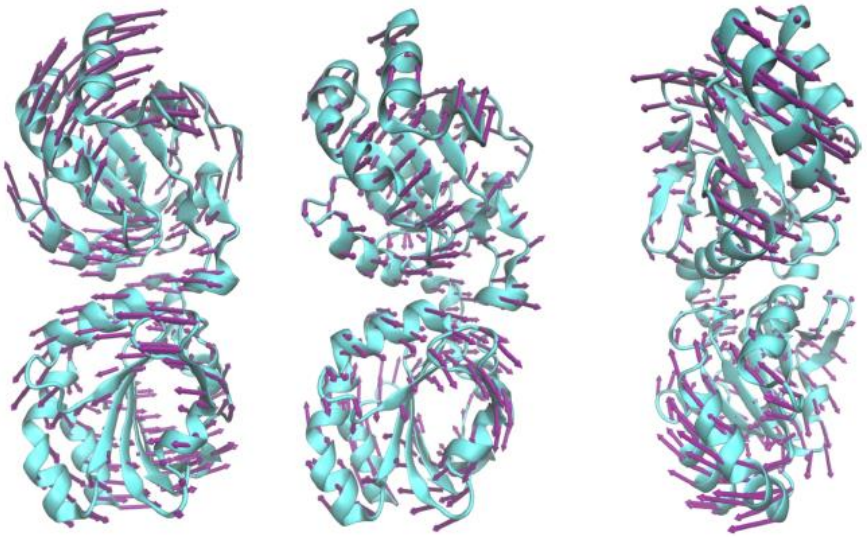
$$\mathcal{L} = \begin{cases} -\left(\Phi^{ij}\right)^{-1}, & i \neq j \\ -\sum_j \mathcal{L}_{ij}, & i = j \end{cases}$$

where $\Phi^{ij}\left(\Phi^{ij}\right)^{-1} = \left|\Phi^{ij}\right|$

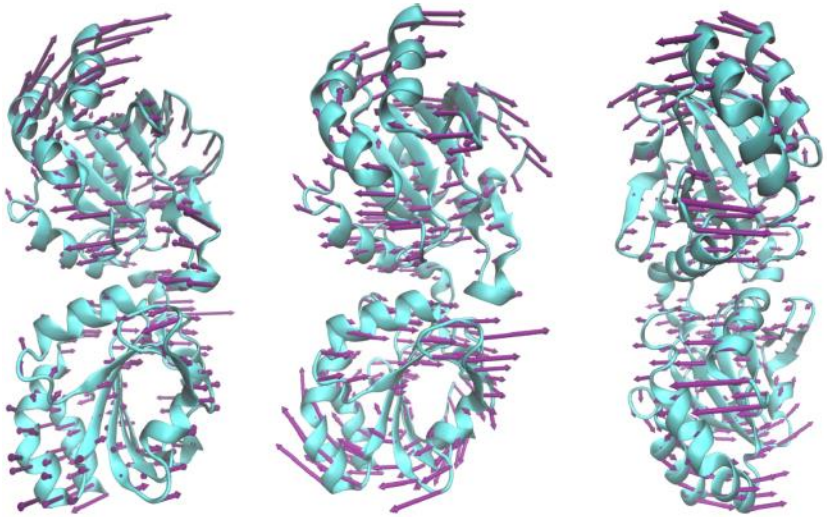
$$B_i^{\mathrm{AFRI}} = (\mathcal{L})_{xx}^{ii} + (\mathcal{L})_{yy}^{ii} + (\mathcal{L})_{zz}^{ii}$$

(Opron, Xia and Wei, JCP, JCP 2014; Nguyen, Xia and Wei, JCP, 2016)

AFRI for 2XHF

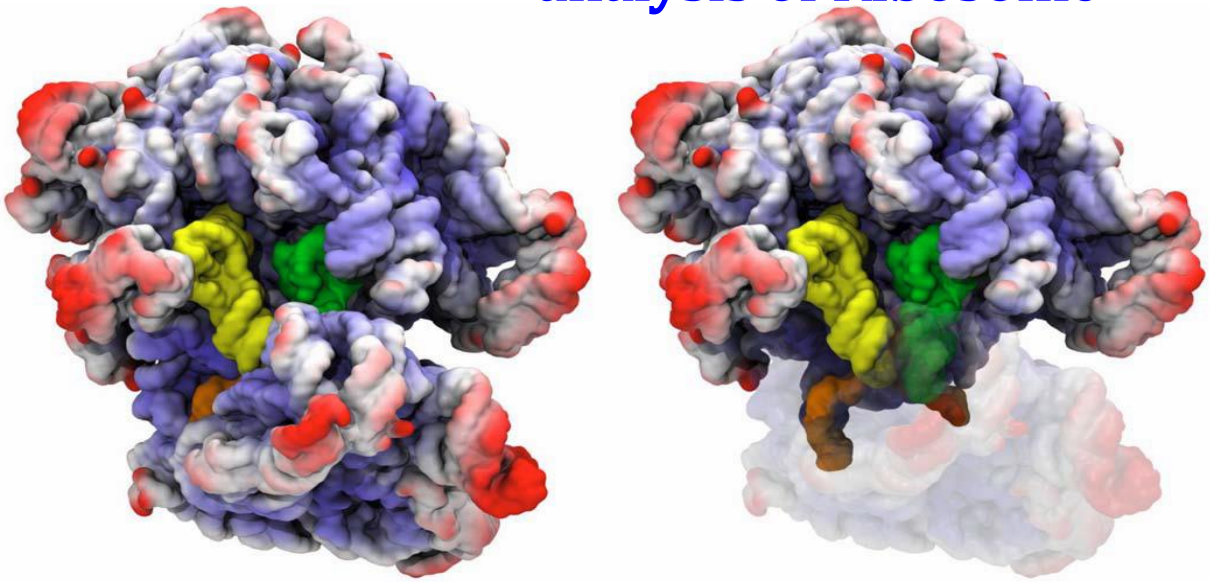


ANM for 2XHF



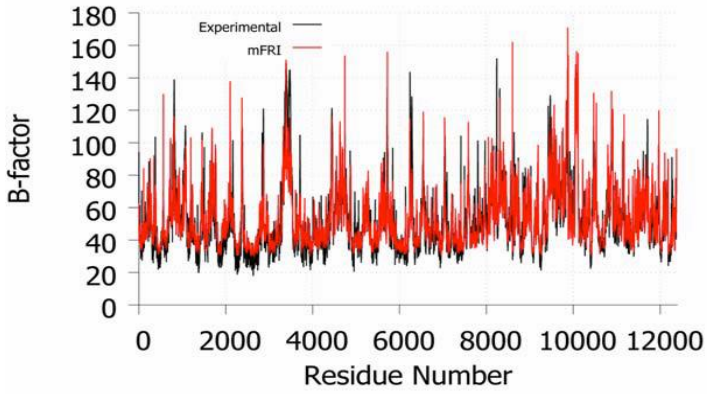
Vector Laplacian, Anisotropic flexibility rigidity index

analysis of Ribosome

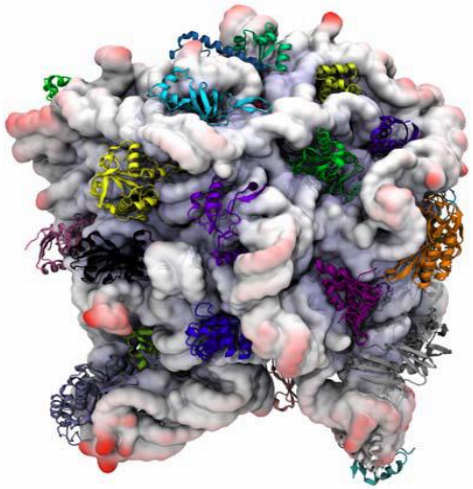


(a) Complete ribosome with bound tRNAs PDB ID: 4V4J.

(Opron, Xia and Wei, JCP, 2016)



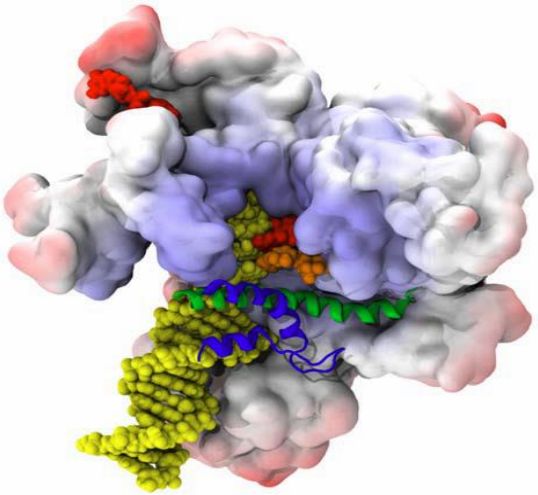
(b) Ribosome 50S subunit PDB ID: 1YIJ B factors



(c) Ribosome 50S subunit PDB ID: 1YIJ

Vector Laplacian, Anisotropic flexibility rigidity index

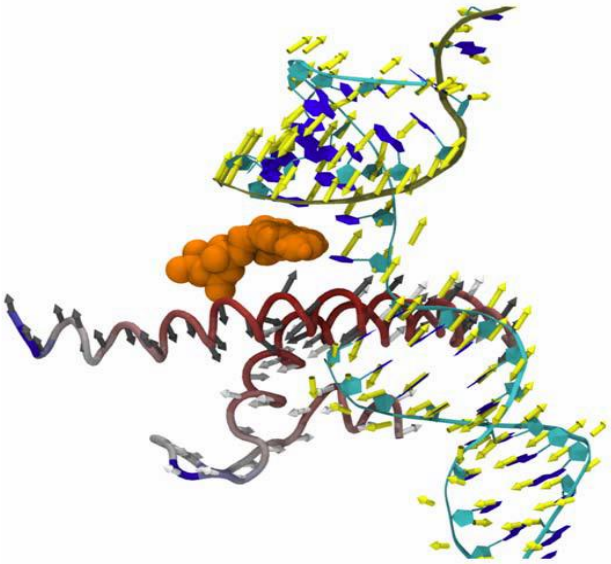
analysis of RNA polymerase



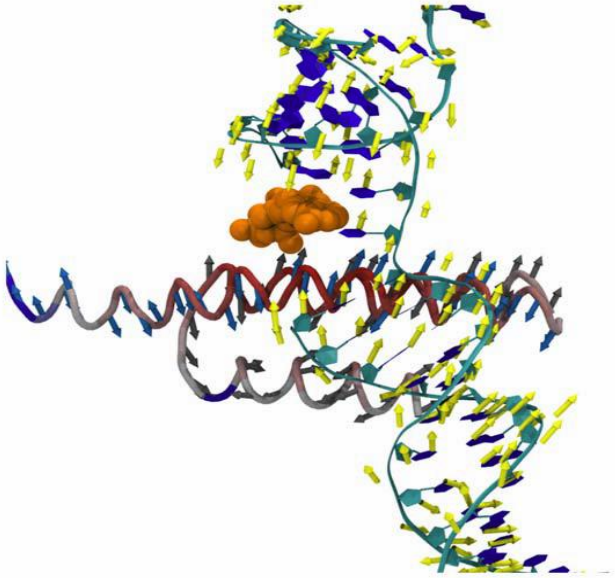
(a) RNA Polymerase with closed trigger loop

(b) Correlated motion near active site

(Opron, Xia and Wei, JCP, 2016)



(c) aFRI mode 1 - Open TL



(d) aFRI mode 1 - Closed TL

Multiscale weighted colored subgraph (MWCS)

Motivation:

- There is a need to discriminate heterogeneous elements, i.e., the mixture of C, N, O, S, P, etc.
- There is a need to represent chemical, physical and biological interactions, i.e., hydrogen bonds, electrostatic effects, polarization, hydrophilicity and hydrophobicity.

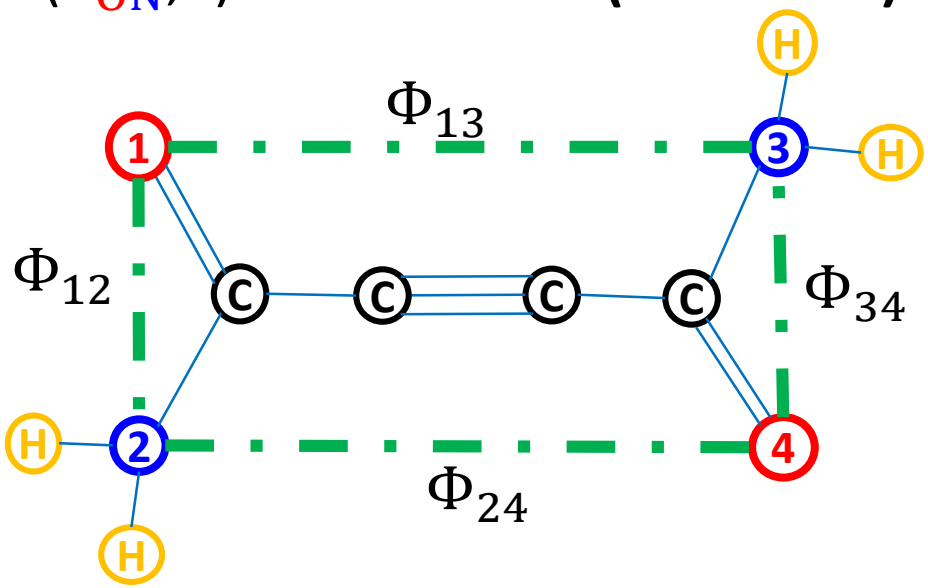
Approaches:

- We construct labeled or colored subgraph based atoms' element types, such as C, N, O, S, P, H, F, Cl, Br, I, ….
- We exclude covalent bond to emphasize non-covalent interactions in data with non-reactive interactions.
- We use multiscale to target different types of interactions.
- Mathematically, MWCSs can be formulated as a hypergraph.
- MWCSs are combined with machine learning for a wide range of chemical and biological predictions.

(Bramer and Wei, JCP, 2018;
Nguyen, Xiao and Wei, JCIM, 2017)

Multiscale weighted colored subgraphs

Weighted colored subgraph $G(V_{ON},E)$ for cellocidin ($C_4H_4N_2O_2$)



Adjacency matrix of $G(V_{ON},E)$

$$\begin{pmatrix} 0 & \Phi_{12} & \Phi_{13} & 0 \\ \Phi_{12} & 0 & 0 & \Phi_{24} \\ \Phi_{13} & 0 & 0 & \Phi_{34} \\ 0 & \Phi_{24} & \Phi_{34} & 0 \end{pmatrix}$$

Eigenvalues: $\lambda_1^A, \lambda_2^A, \ldots$

Laplacian matrix of $G(V_{ON},E)$

$$\begin{pmatrix} \Phi_{12}+\Phi_{13} & -\Phi_{12} & -\Phi_{13} & 0 \\ -\Phi_{12} & \Phi_{12}+\Phi_{24} & 0 & -\Phi_{24} \\ -\Phi_{13} & 0 & \Phi_{13}+\Phi_{34} & -\Phi_{34} \\ 0 & -\Phi_{24} & -\Phi_{34} & \Phi_{24}+\Phi_{34} \end{pmatrix}$$

Eigenvalues: $\lambda_1^L, \lambda_2^L, \ldots$

Addional sugraphs:
$G(V_{OC},E)$, $G(V_{OH},E)$,
$G(V_{NC},E)$, $G(V_{NH},E)$,
$G(V_{CH},E)$, … ,

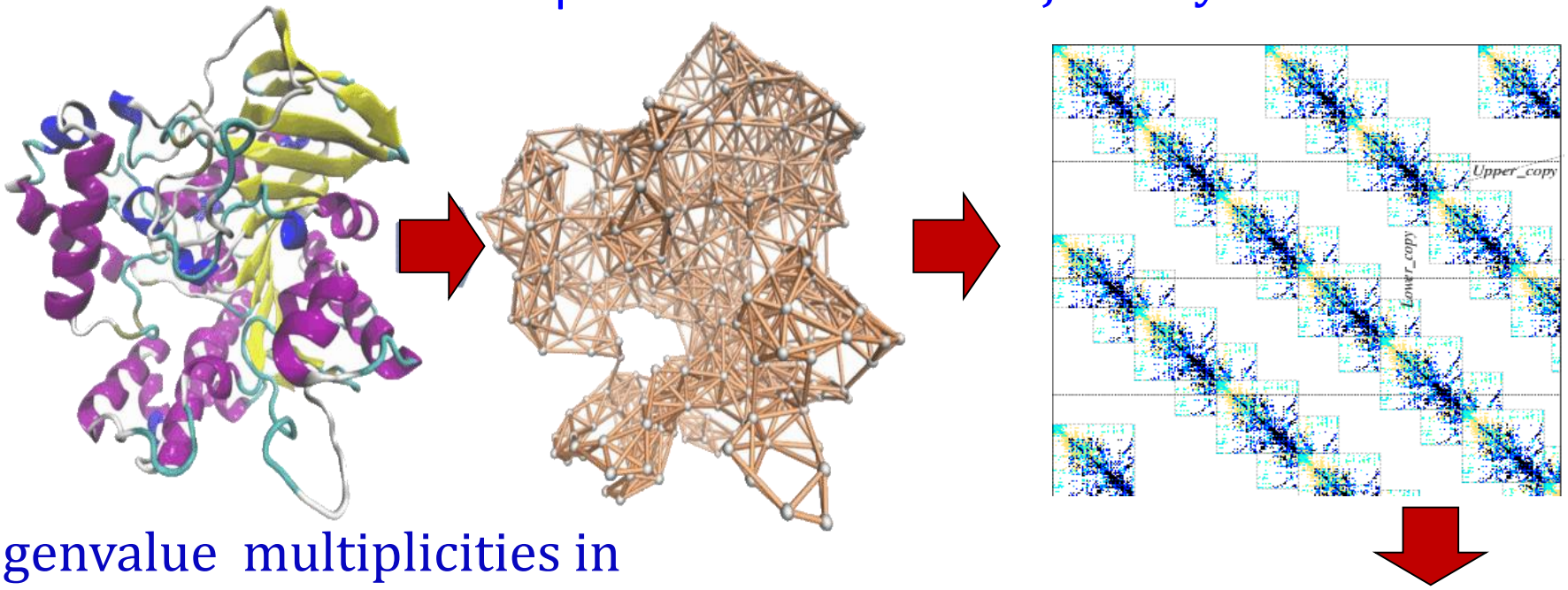
(Nguyen and Wei, JCMI, 2019)

Algebraic graph theory for biomolecules

Protein

Hypergraph representation

Laplacian matrices and adjacency matrices



Eigenvalue multiplicities in Laplacian and adjacency matrices are associated with structural self-similarity, stability, flexibility and activity and hotspots, etc.

Corresponding eigenvalues $\lambda_1^L, \lambda_2^L, \ldots$

$\lambda_1^A, \lambda_2^A, \ldots$

Mark Kac: Can one hear the shape of a drum?
Can one hear protein-drug binding?

(Image credit: Bahar et al.)

Further topics and future directions

- Graph analysis of DNA packaging in chromosomes.
- Graph analysis of Hi-C data.
- Graph and autamata theory analysis of self-assembled DNA complexes.
- Complete graph and bipartite graph of protein-protein interactions.
- Seudograph representation of biomolecules.
- Group theory analysis of viral capsids.
- Manifold convergence analysis of allosteric effects.
- Atom specific and/or element specific graphs for molecules.
- Combinatorics based modeling of biomolecules, such as RNA structures.
- Topological graph based modeling of biomolecules.
- Mathematical modeling of supramolecular assembly.
- Poset analysis of biomolecules.
- Hypergragh based modeling of biomolecules.
- Gromov-Hausdorff metric based graphs for biomolecules.


thank you