Topological modelling and analysis for biomolecular data: 1 from graph to simplicial complex

Kelin Xia

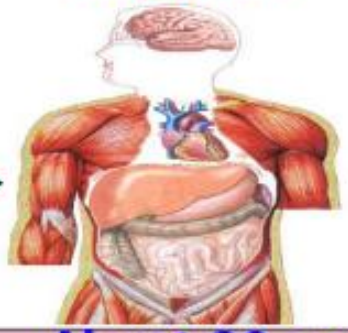
School of Physical and Mathematical Sciences, Nanyang Technological University

应用拓扑短期课程与报告会, 12月13-16, 2020

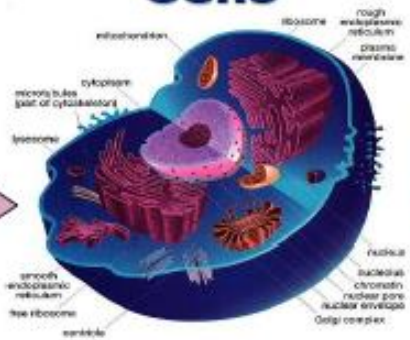
Fund: NTU-JSPS(2019), MOE-Tier 1(2018,2019), MOE-Tier 2(2018,2021), Alibaba-NTU(2020), Merlion(2020)

Species

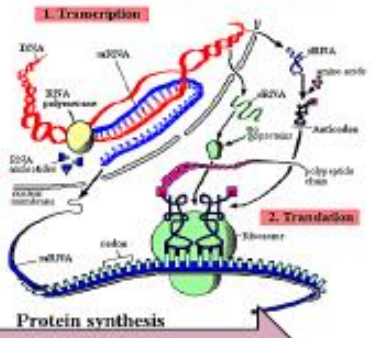
Organs



Cells



Molecules

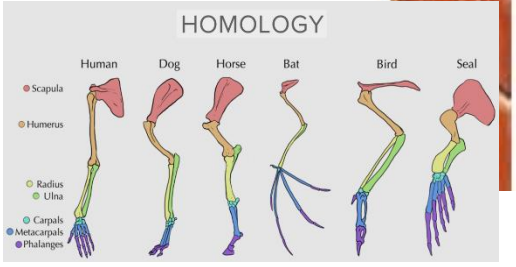


About 20 orders in spatial scales

About 20 orders in time scales

Evolutionary biology

Reaction diffusion
Stochastic models
Kinetic models
Delayed ODEs
Discrete models
Homology models
Machine learning



Developmental biology
Physiology
Biomechanics

Continuum models
Mechanical models
Navier-Stokes
(Non-) linear elasticity
Maxwell's equation
Thermal models
Rheological models
Hodgkin-Huxley model
Lattice models
Neural networks
Geometric models
Topological models

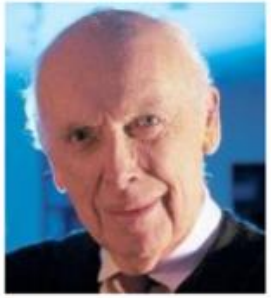
Cellular biology
Systems biology
Cellular mechanics

Chemical kinetics (ODEs)
Gene regulatory network
Protein network
Neural networks
Hodgkin-Huxley model
FitzHugh-Nagumo model
Mechanical models
Reaction diffusion
Phase field models
Stochastic models
Statistical models
Monte Carlo
Combinatory
Topological models
Machine learning

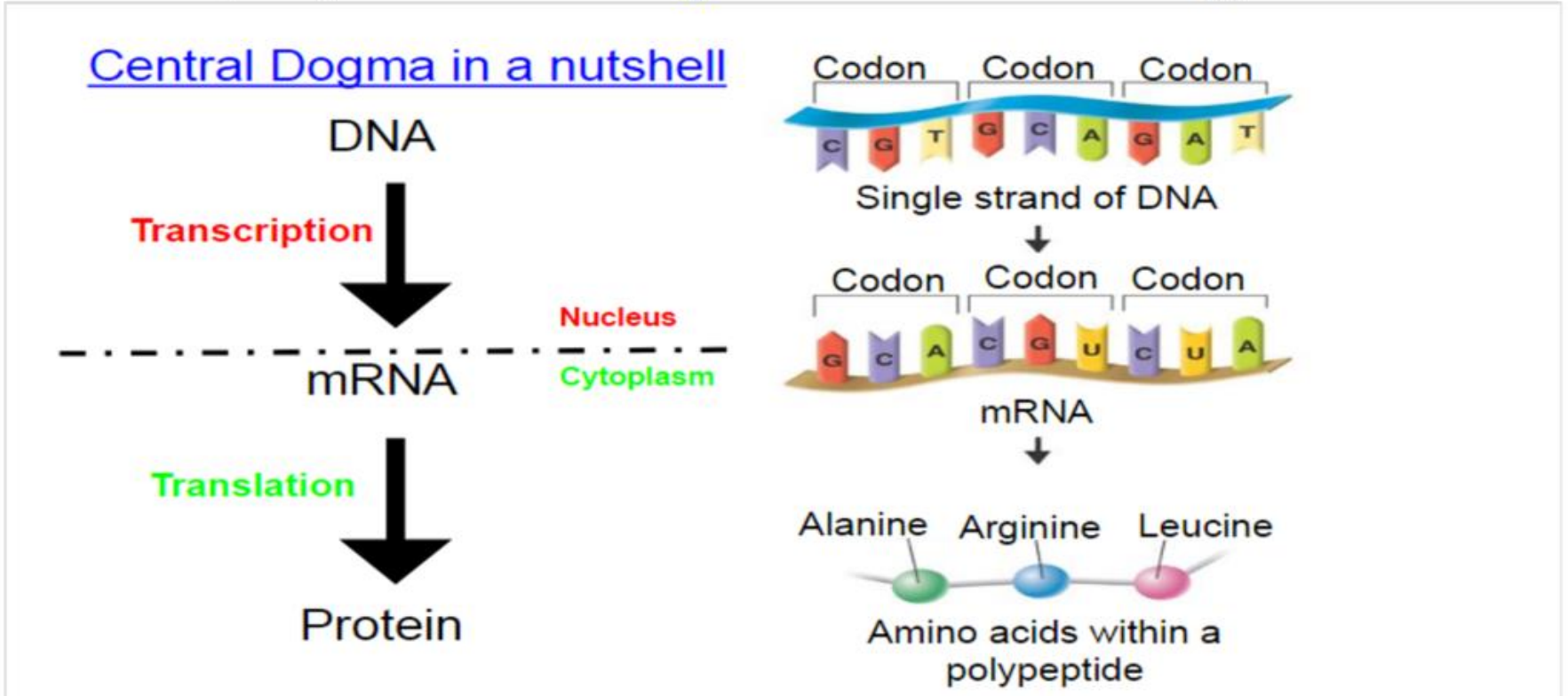
Molecular biology
Biochemistry
Biophysics

Molecular dynamics
Thermal dynamics
Brownian dynamics
Lagevin dynamics
Quantum models
QM/MM
Electrostatics
Implicit models
Boltzmann equation
Vlasov-Boltzmann
Fokker-Planck
Monte Carlo
Master equation
Homology models
Knot theory

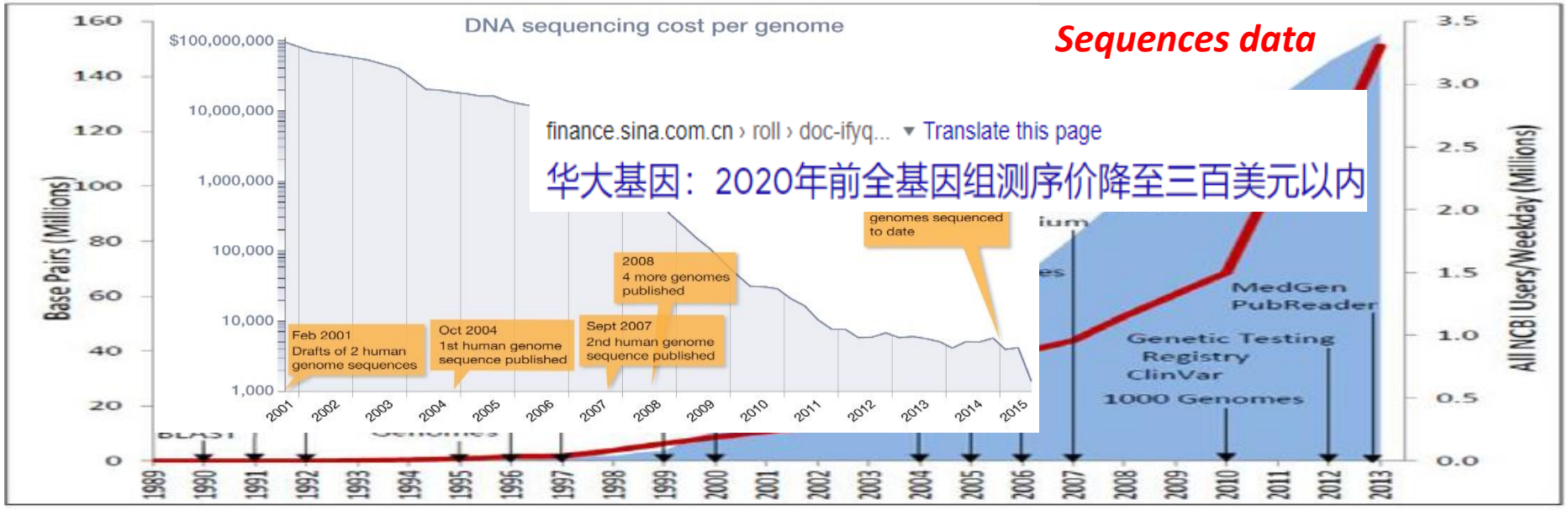
Molecular biology

Francis Crick (1916 - 2004) and James Watson (1928 –) Nobel Prize in Physiology or Medicine in 1962 "for discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material", i.e., the Central Dogma of molecular biology:



Biological data



Sequences data

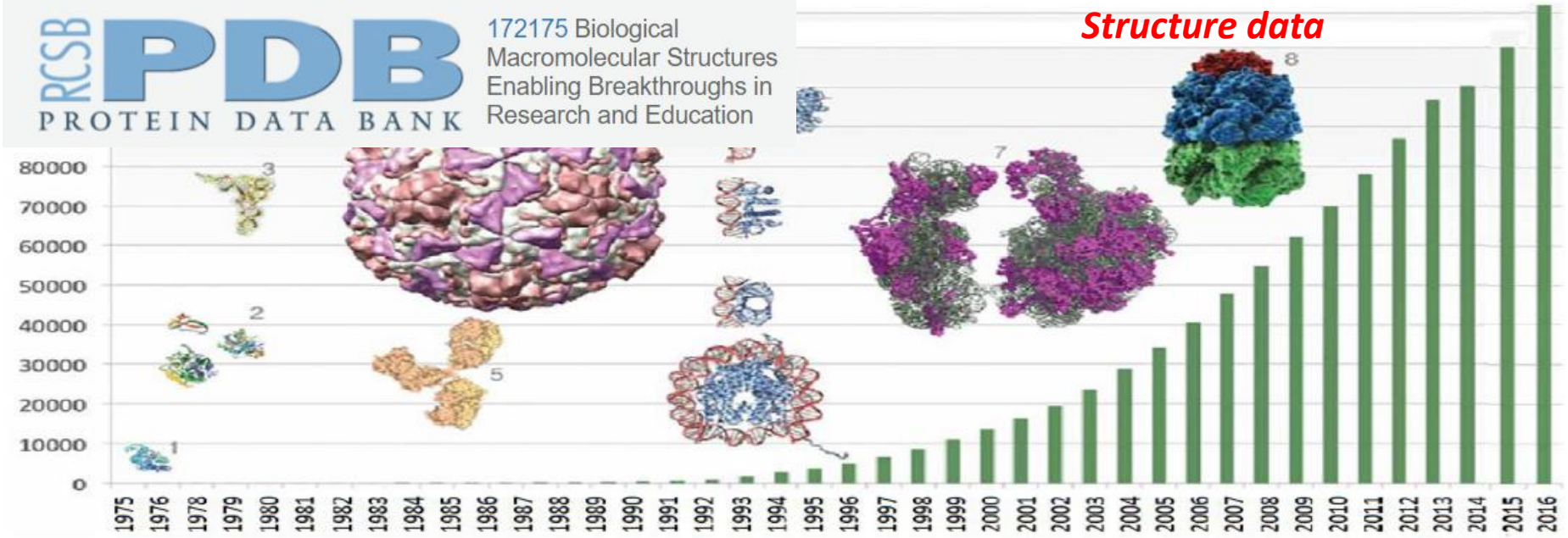
finance.sina.com.cn › roll › doc-ifyq... ▾ Translate this page

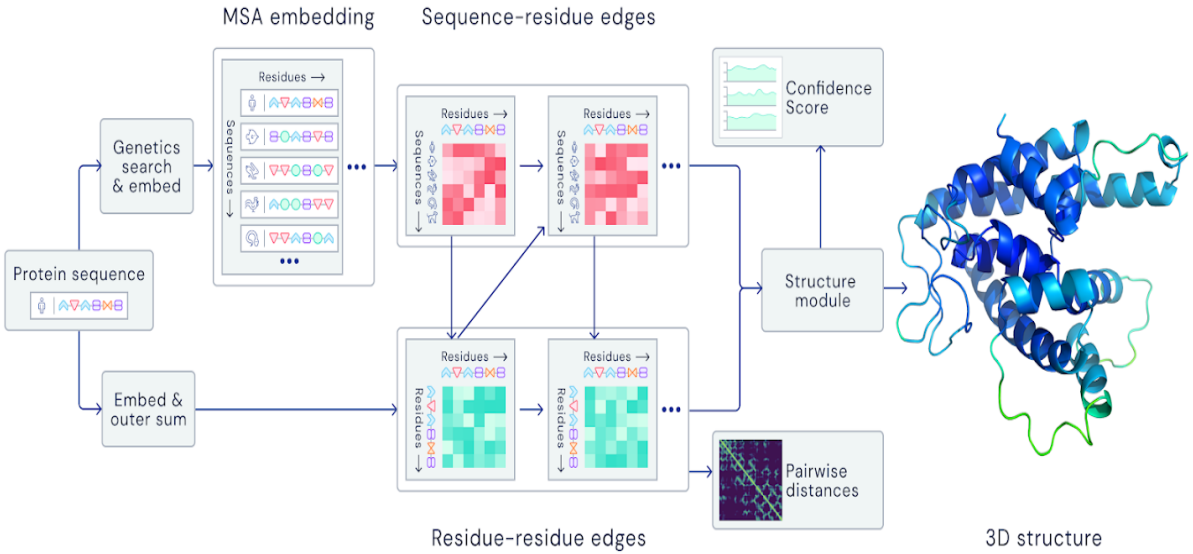
华大基因：2020年前全基因组测序价降至三百美元以内

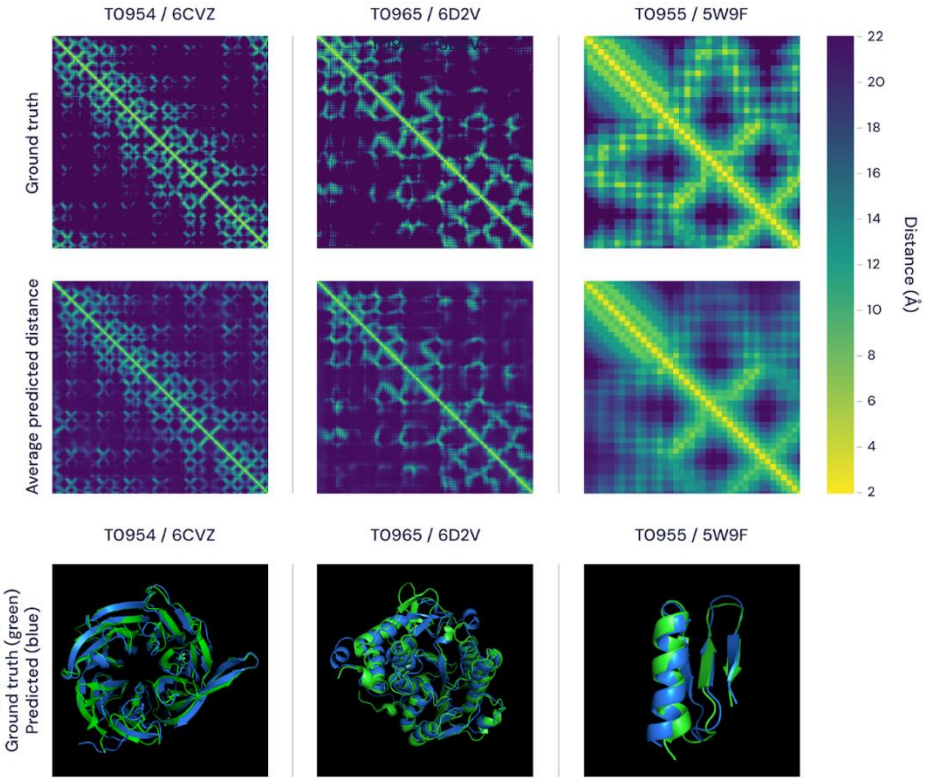
RCSB PDB
PROTEIN DATA BANK

172175 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Structure data

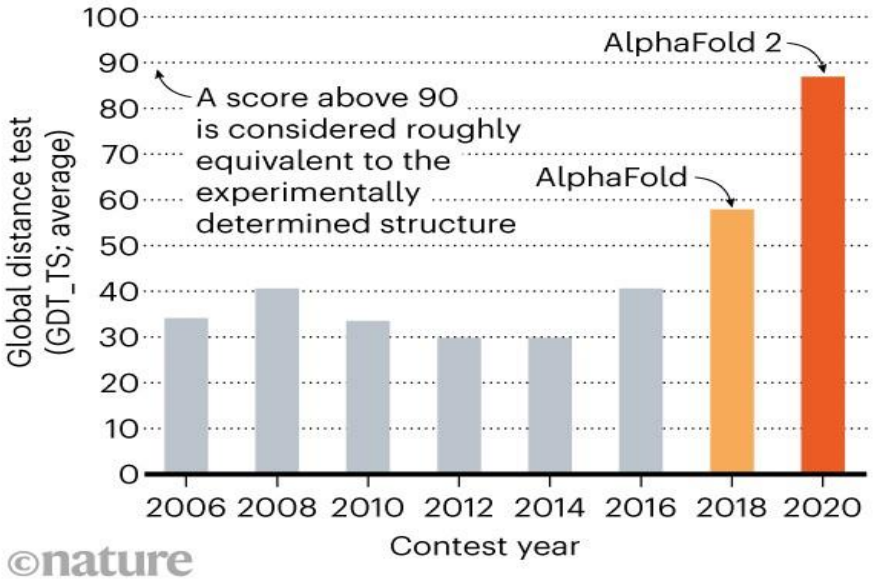


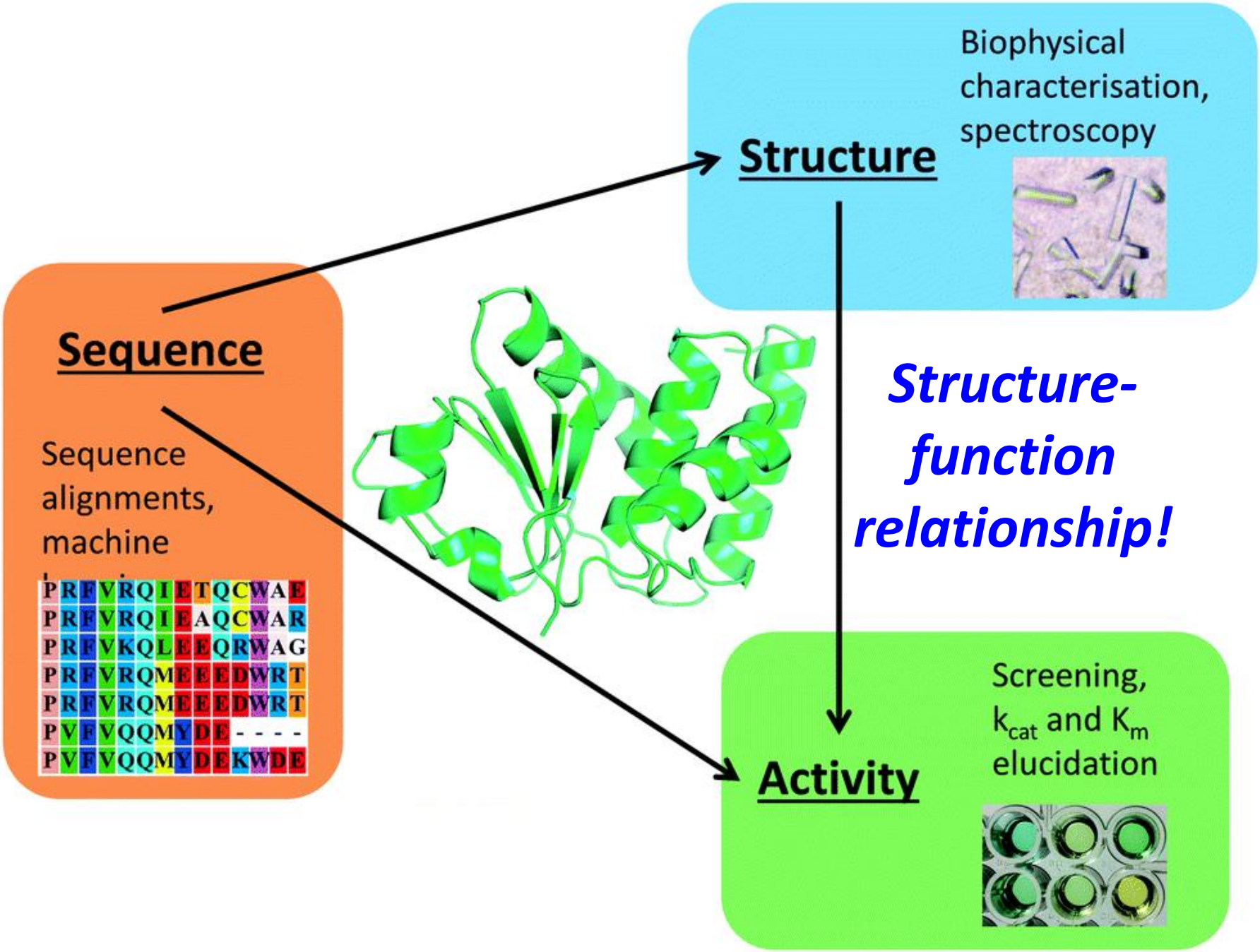




STRUCTURE SOLVER

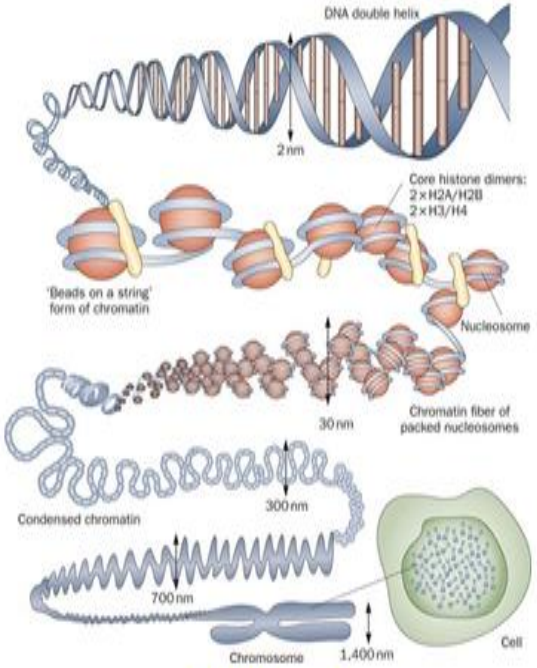
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



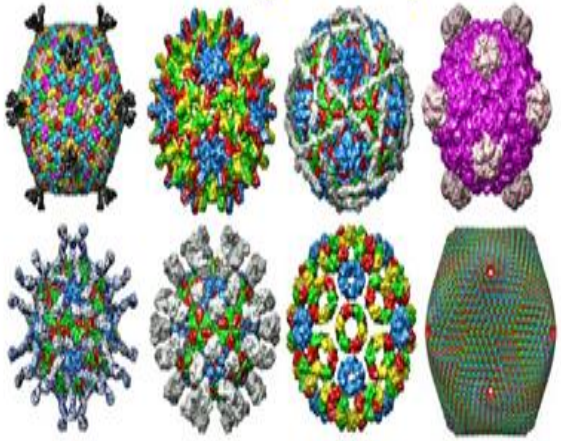

Structure
Biophysical characterisation, spectroscopy
Sequence
Sequence alignments, machine
Structure-function relationship!
Activity
Screening, k_{cat} and K_m elucidation

1. Protein and DNA structures

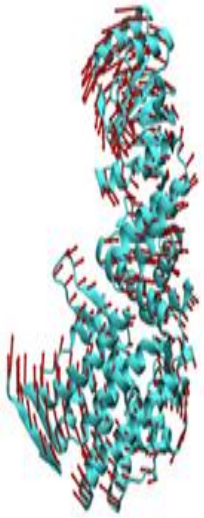
Chromosome structure



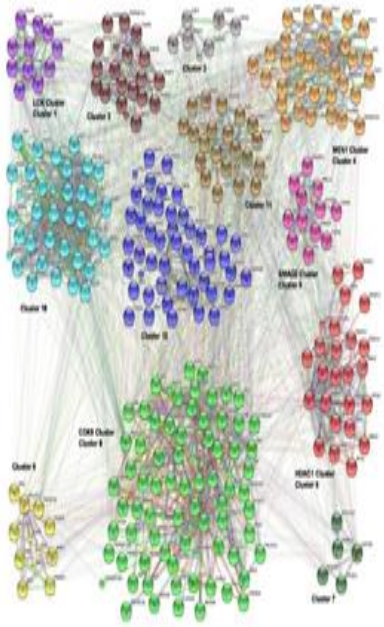
Virus self-assembly



Molecular dynamics

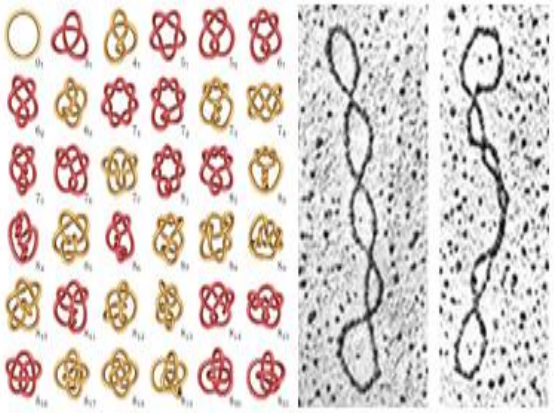


Protein-protein Interactions

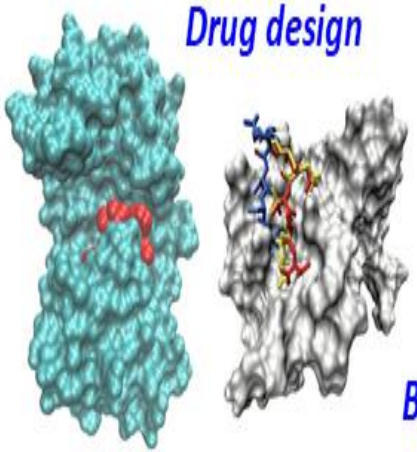


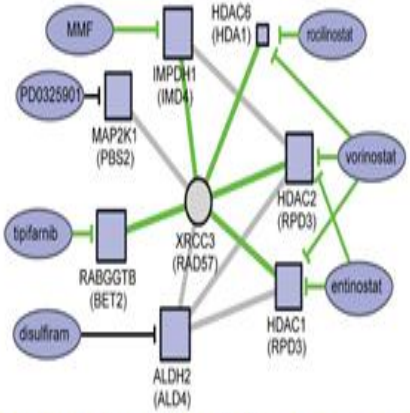
Biomolecular Topology

DNA knots



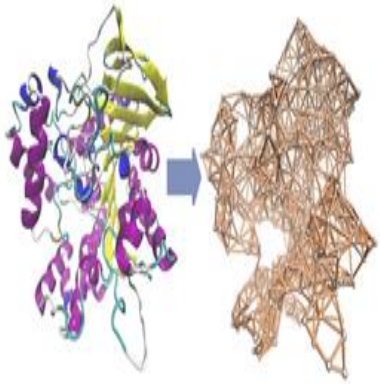
Drug design





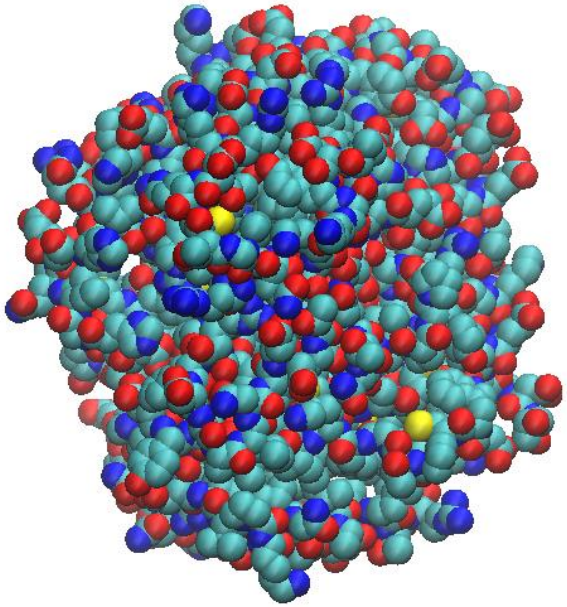
Biomolecular regulatory network

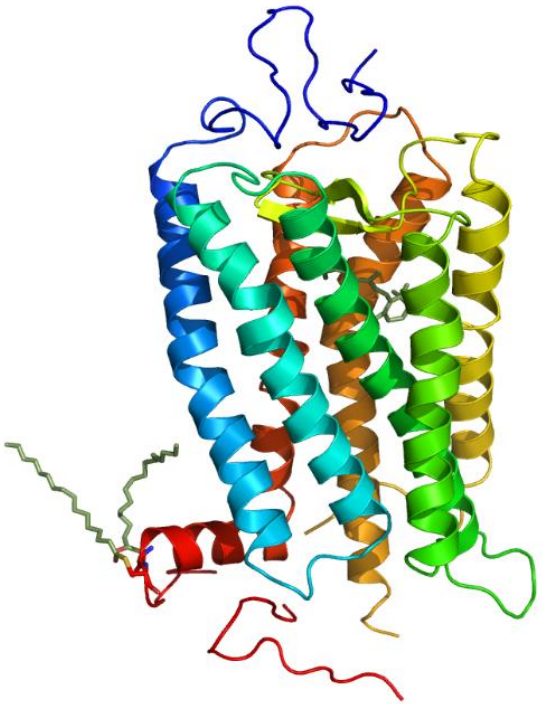
Protein structure network

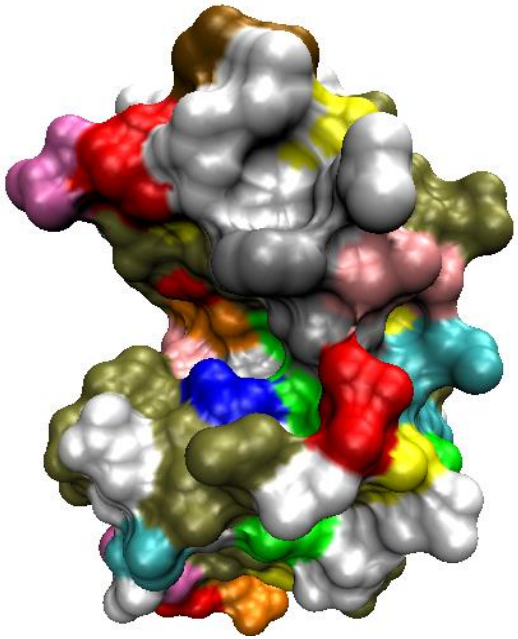


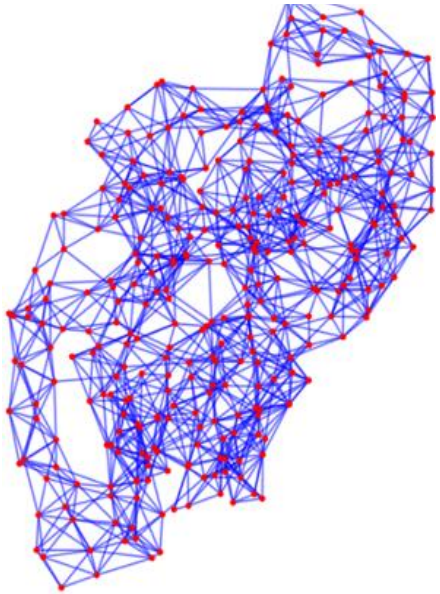
Biomolecular structure-function relationship!

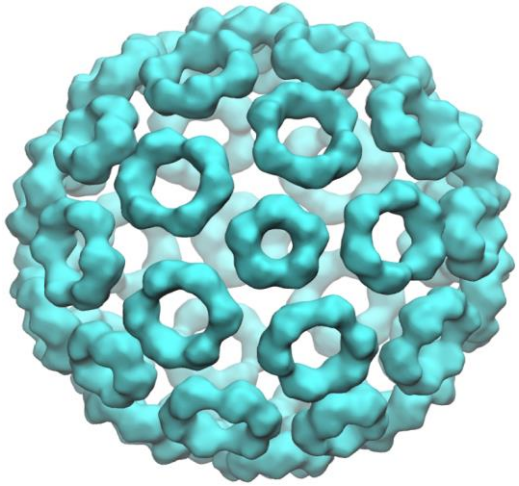
All proteins?

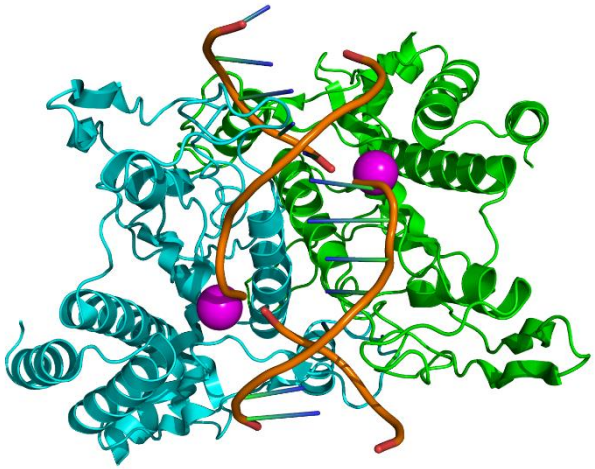




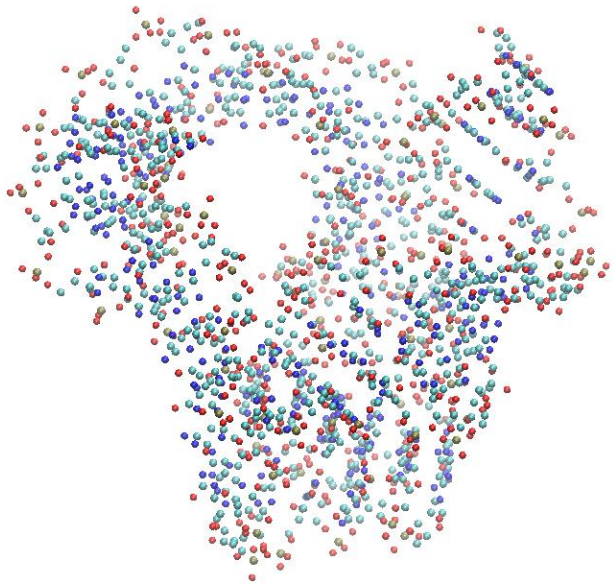
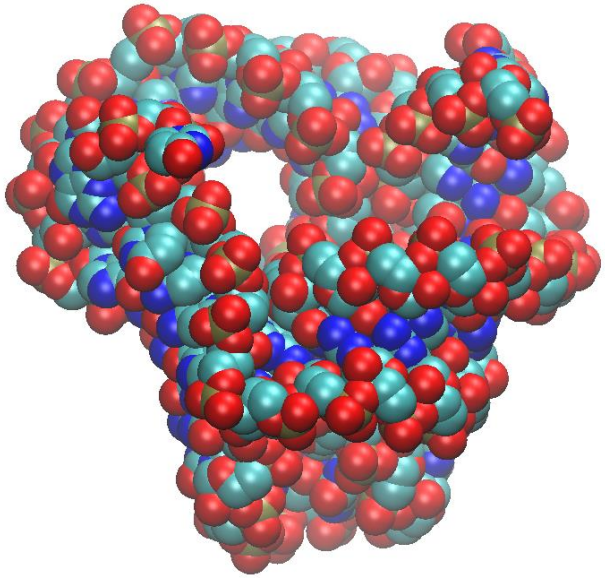
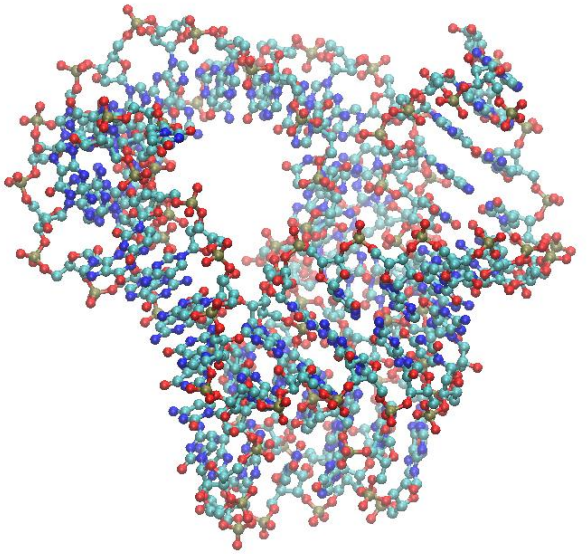
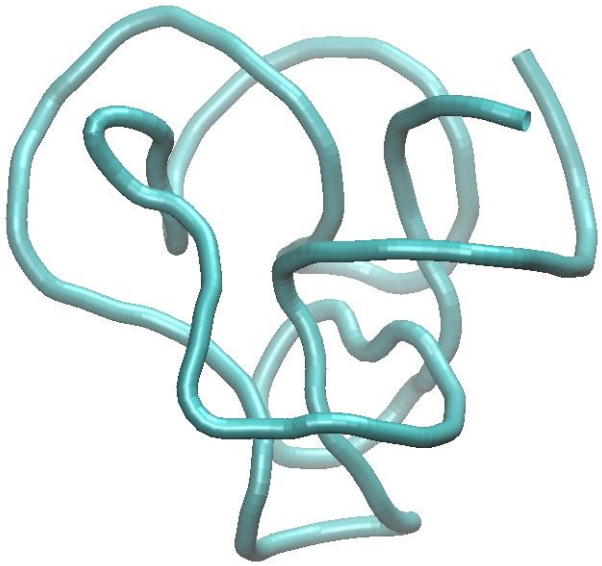
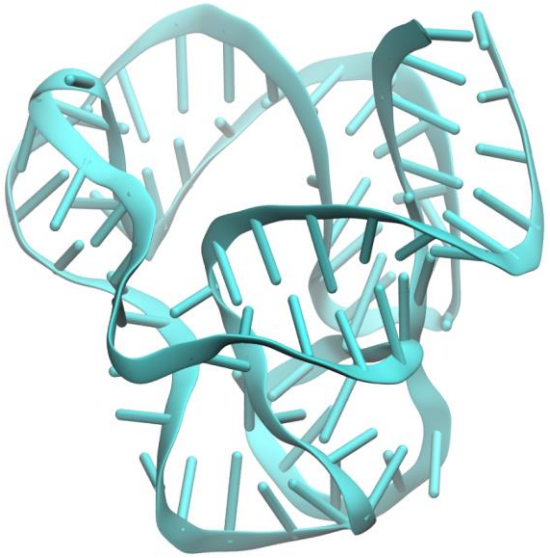
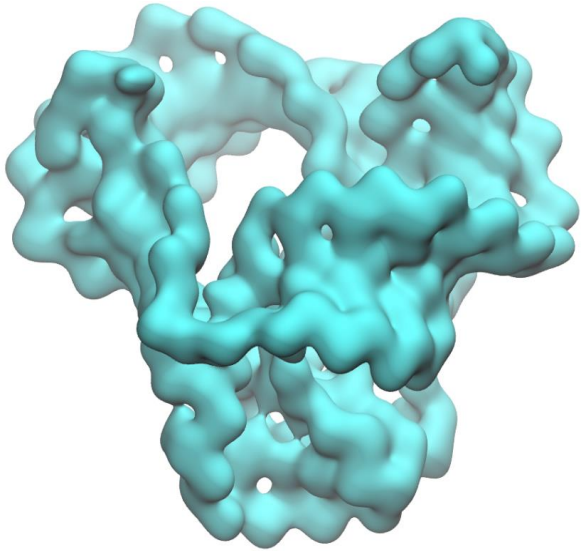






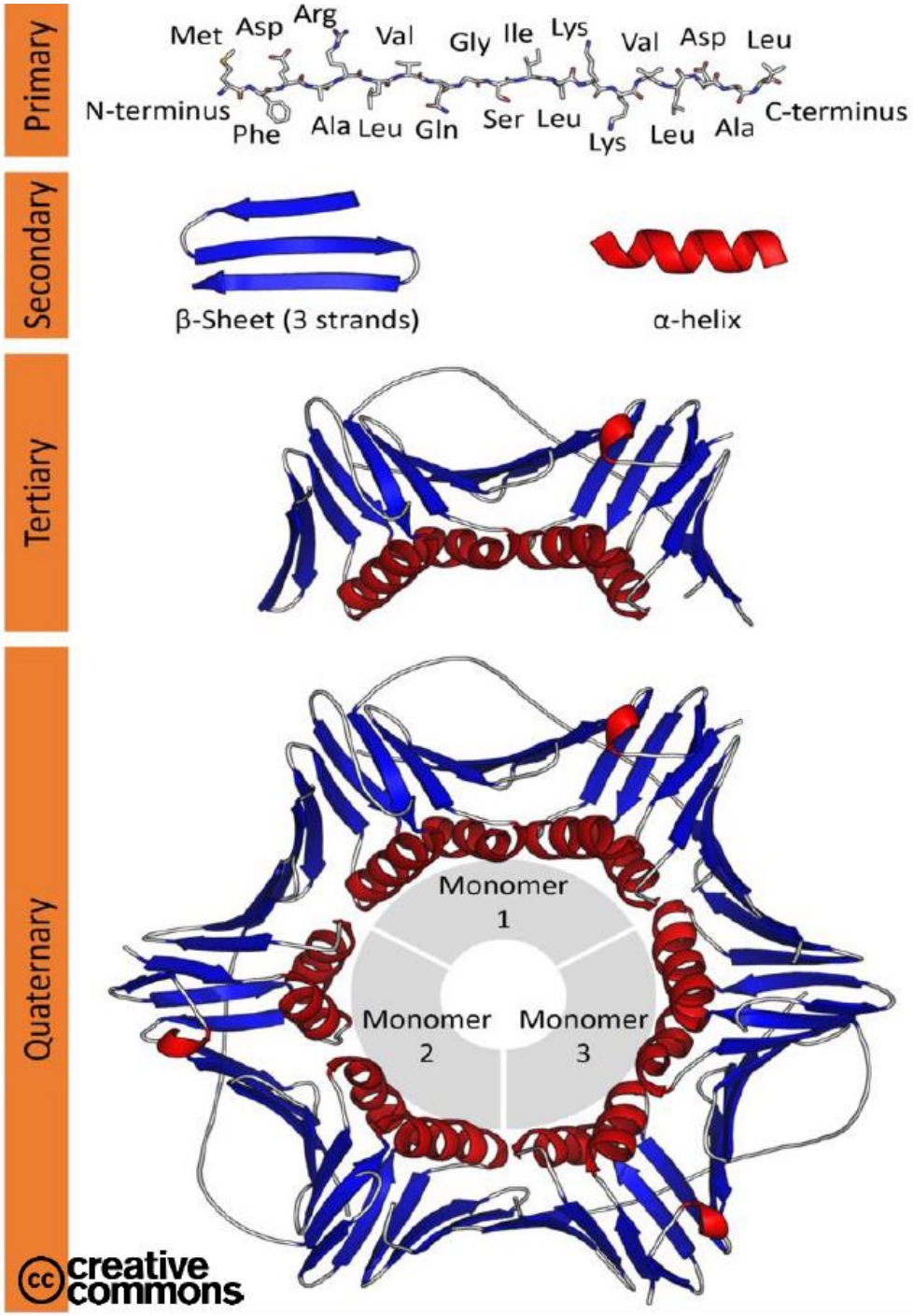


Same biomolecule?

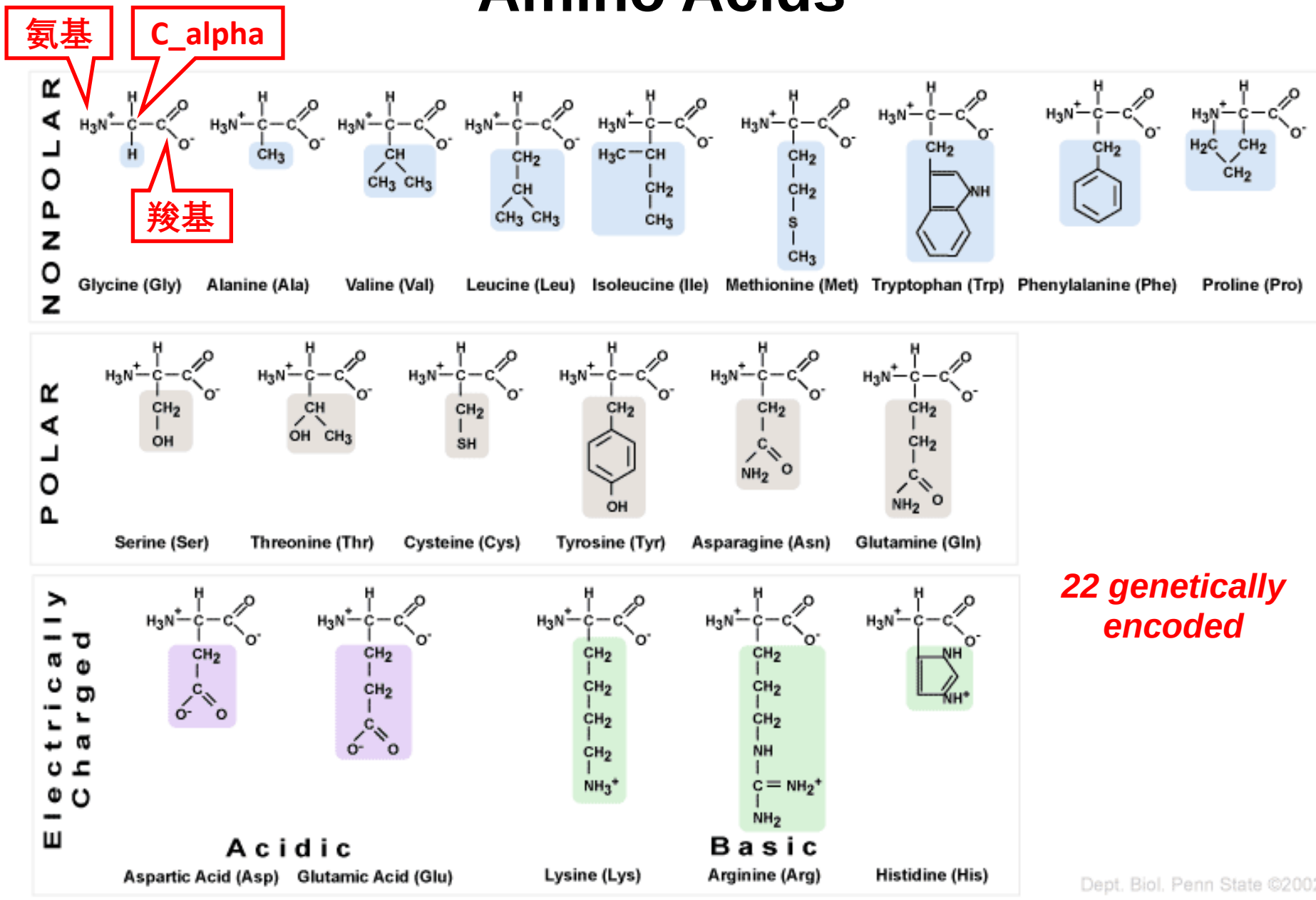







Protein Structure

Primary structure
Secondary structure
Tertiary structure
Quaternary structure

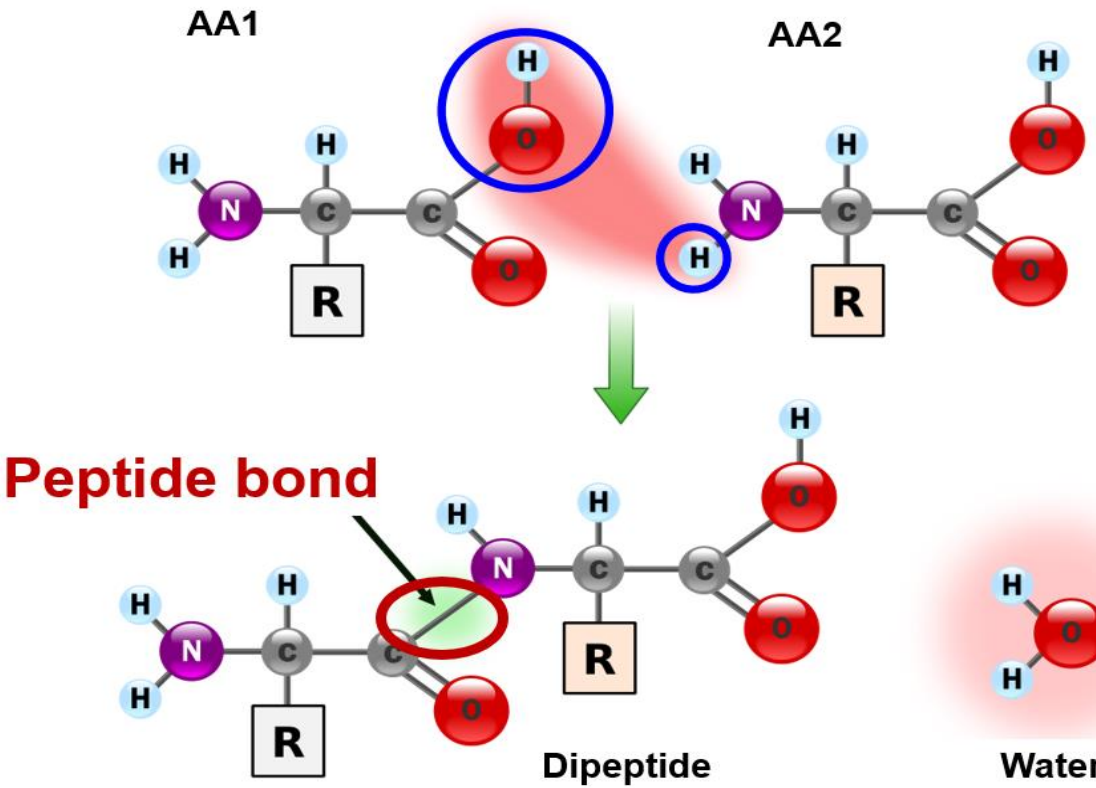


Amino Acids

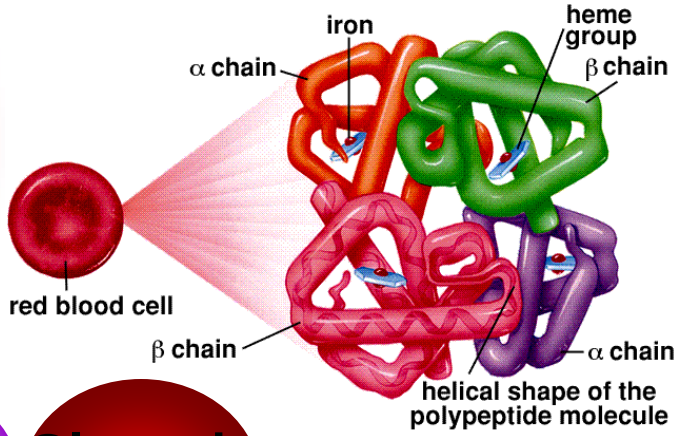


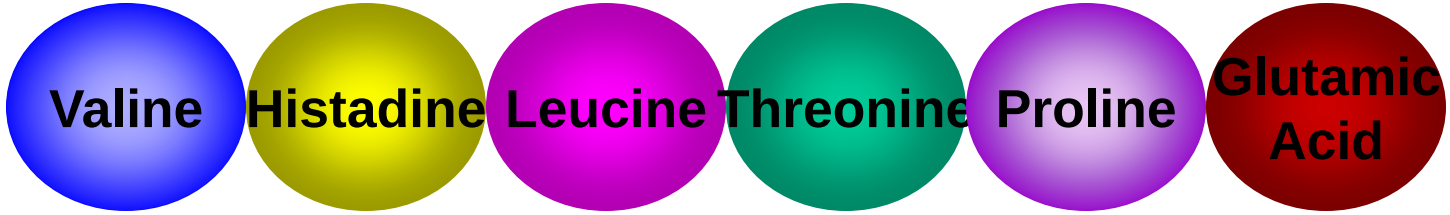
22 genetically encoded

Peptide chain



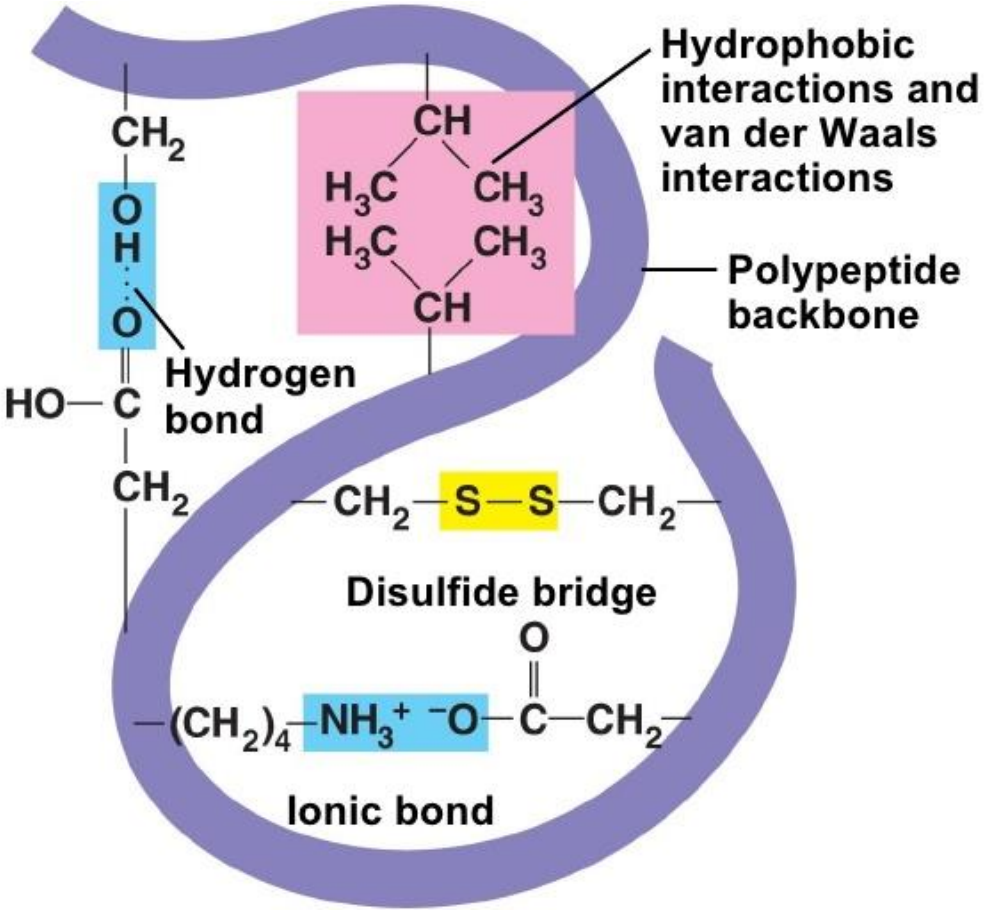
Peptide bond:
C-N ~100kcal/mol





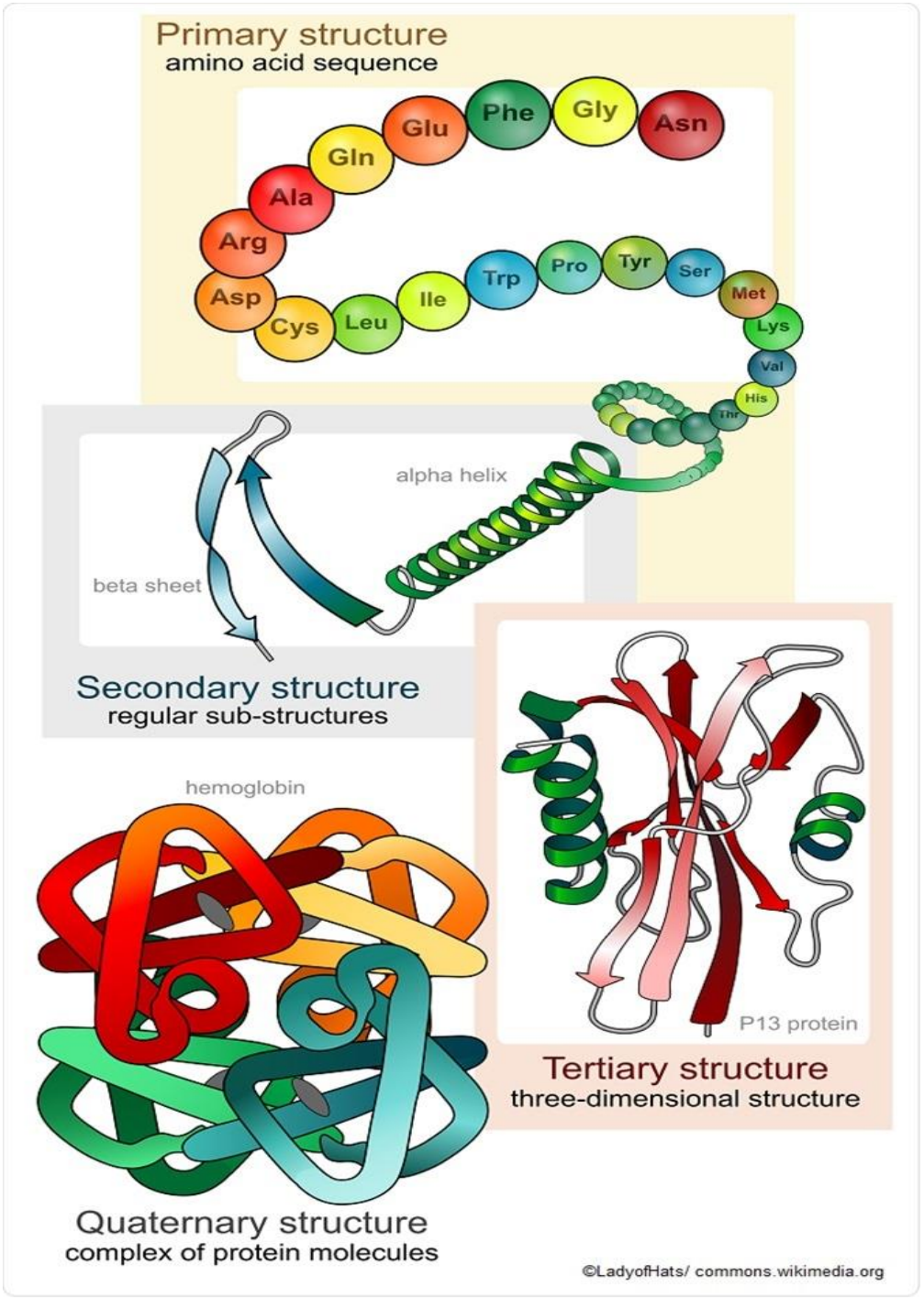
hemoglobin

Protein bond types



- **Disulfide bond ~60kcal/mol**
- **Salt (Ionic) bond ~20kcal/mol**
- **Hydrogen bond ~10kcal/mol**
- **Hydrophobic interaction and van der Waals ~1kcal/mol**

Covalent bond: C-C ~100kcal/mol



Protein structure

Nuclei Acid Structure

Nitrogenous base (Adenine, Guanine, Cytosine, Thymine (in DNA), Uracil (in RNA))
5-carbon sugar called deoxyribose (found in DNA) and ribose (found in RNA).
One or more phosphate groups.

Primary structure
Secondary structure
Tertiary structure
Quaternary structure

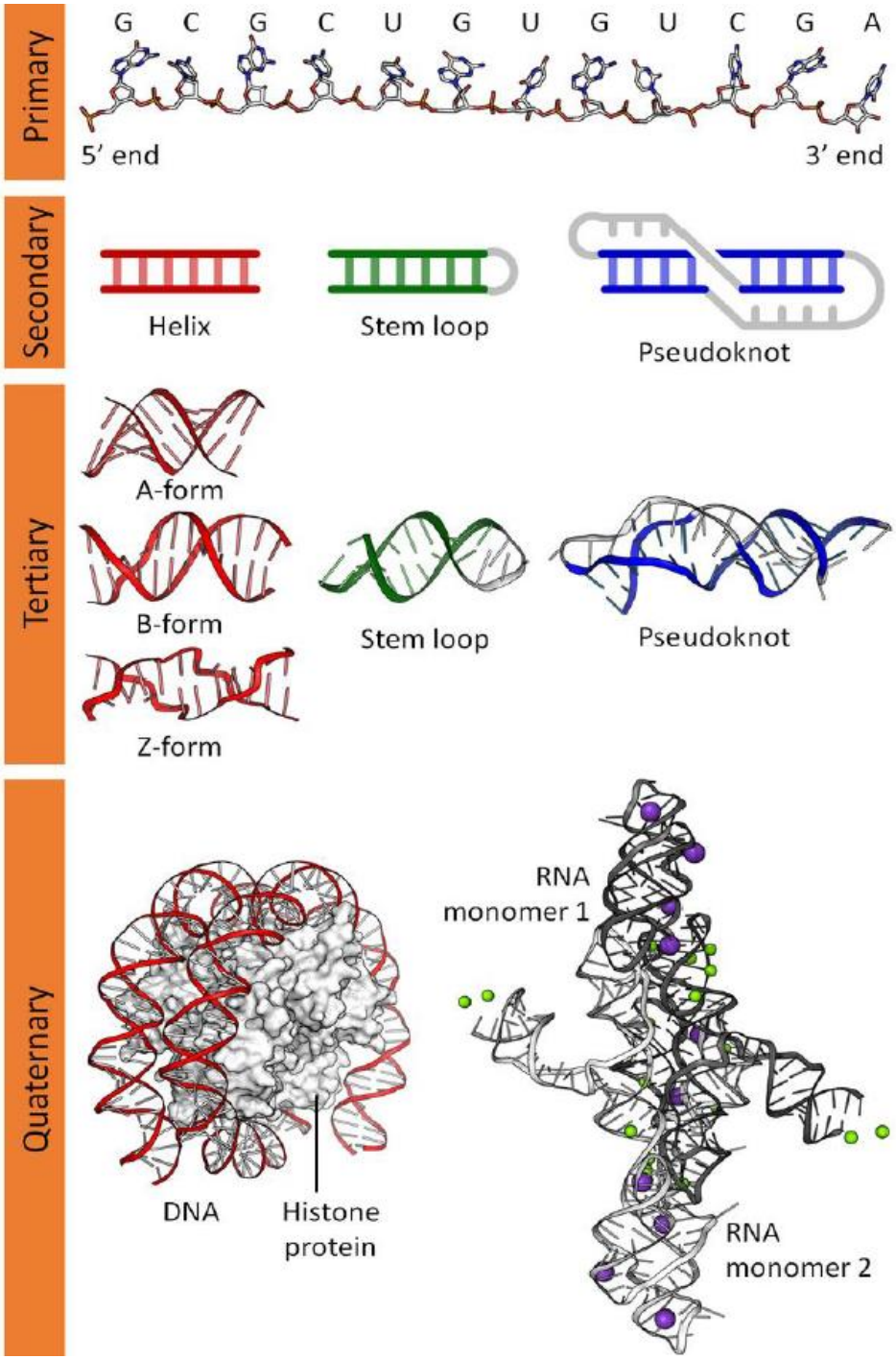
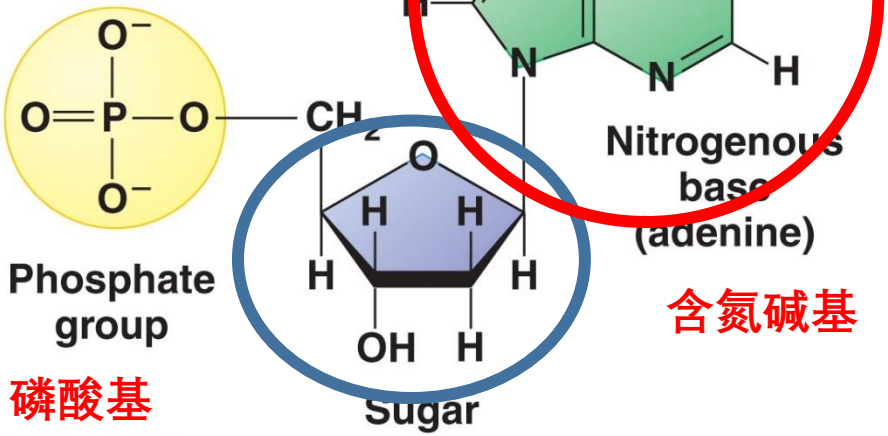
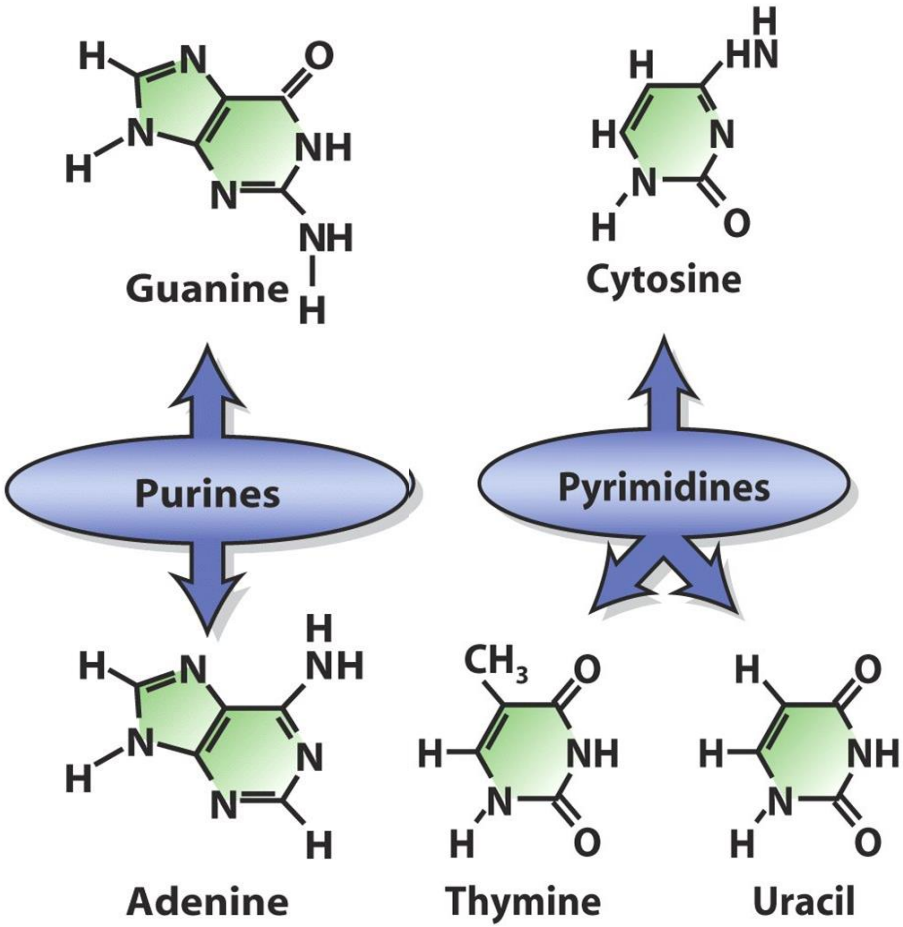
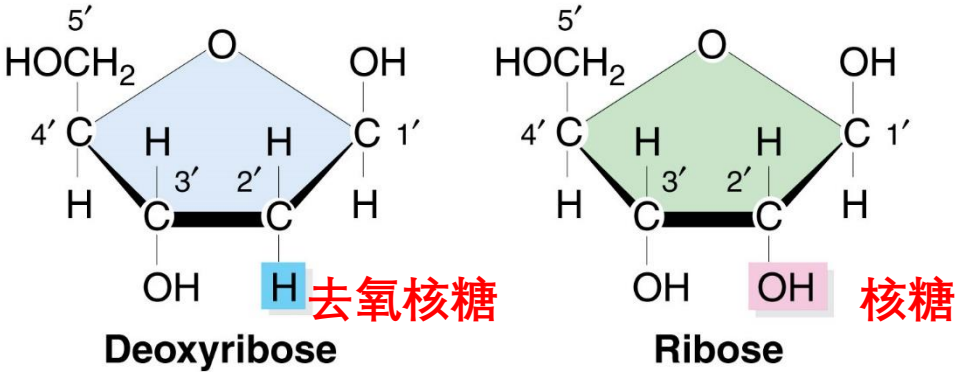


Image credit: Thomas Shafee

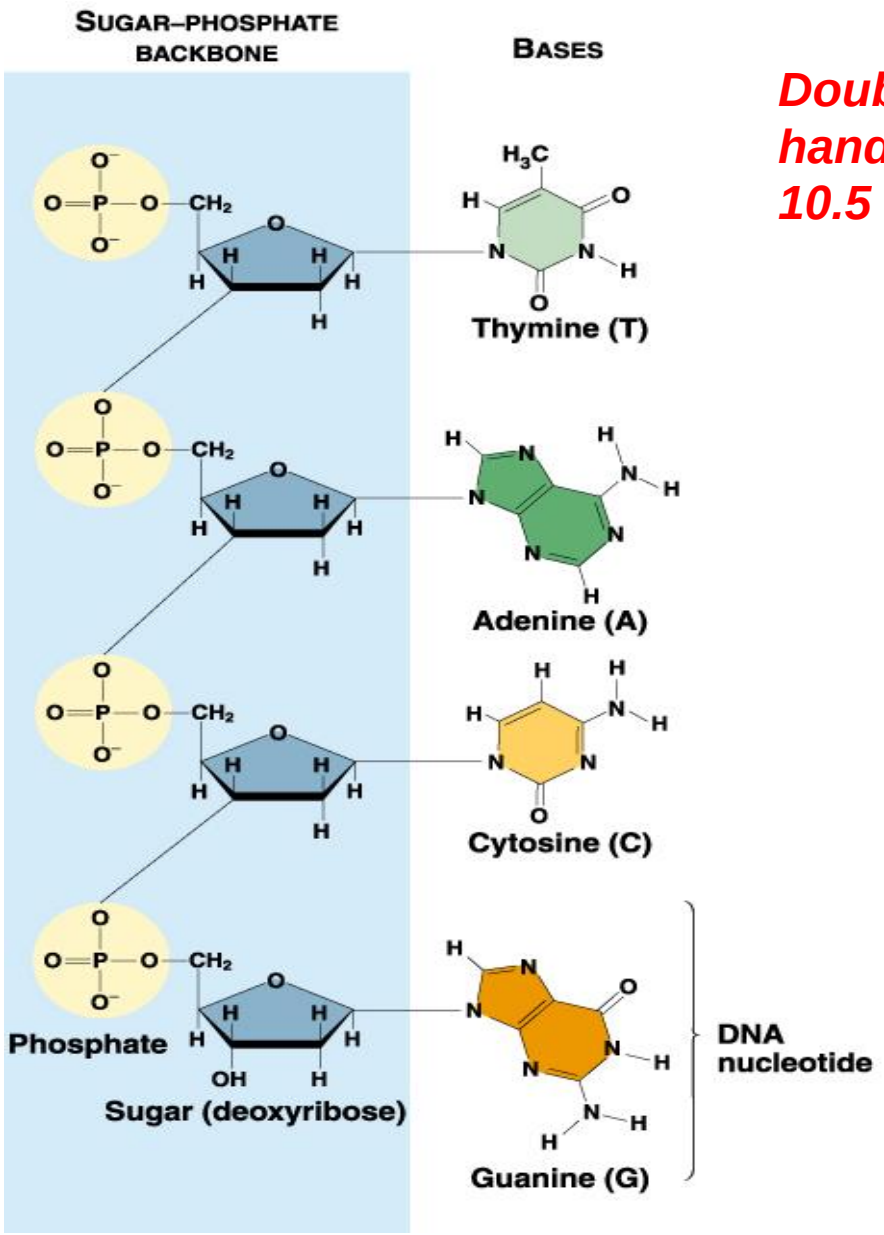
Nucleic acid structure



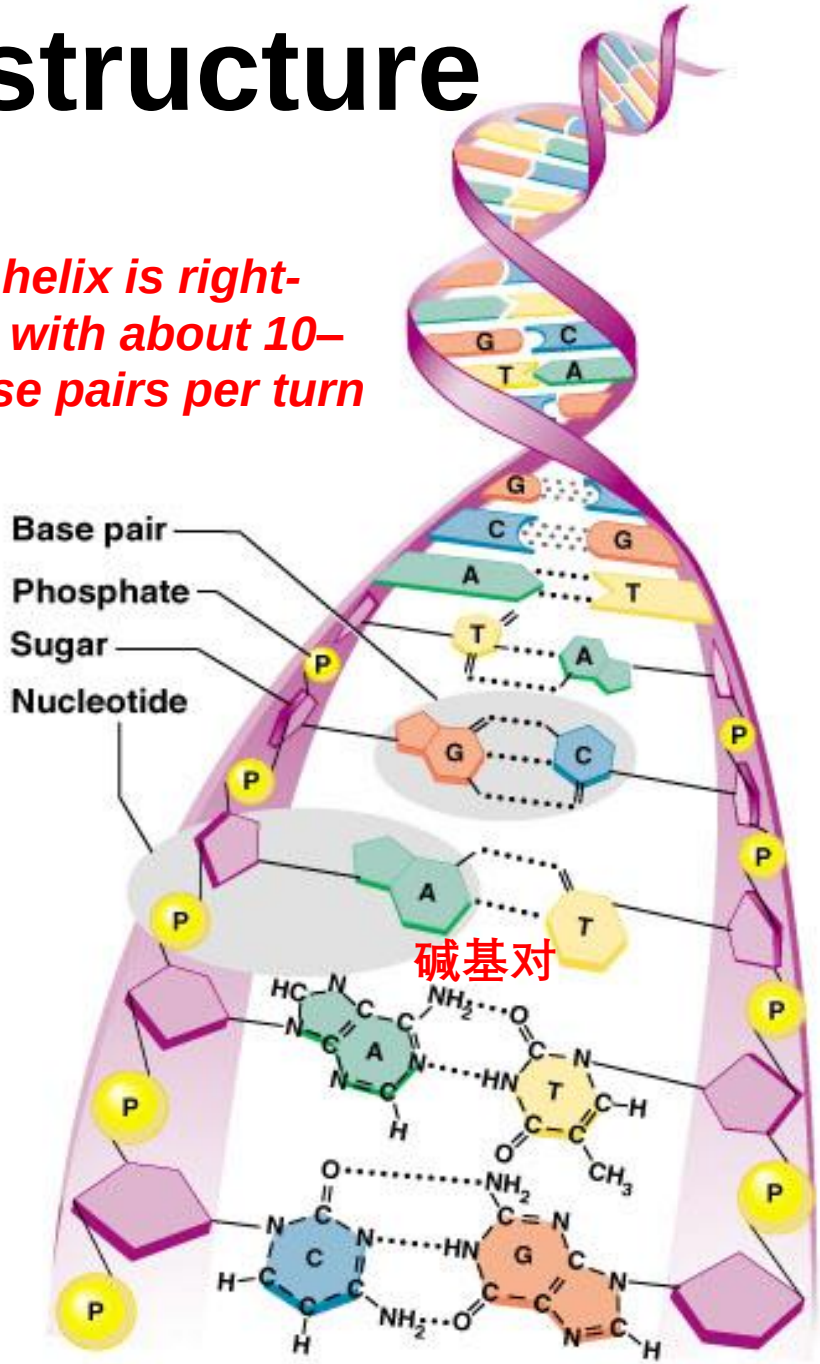




Nucleic acid structure



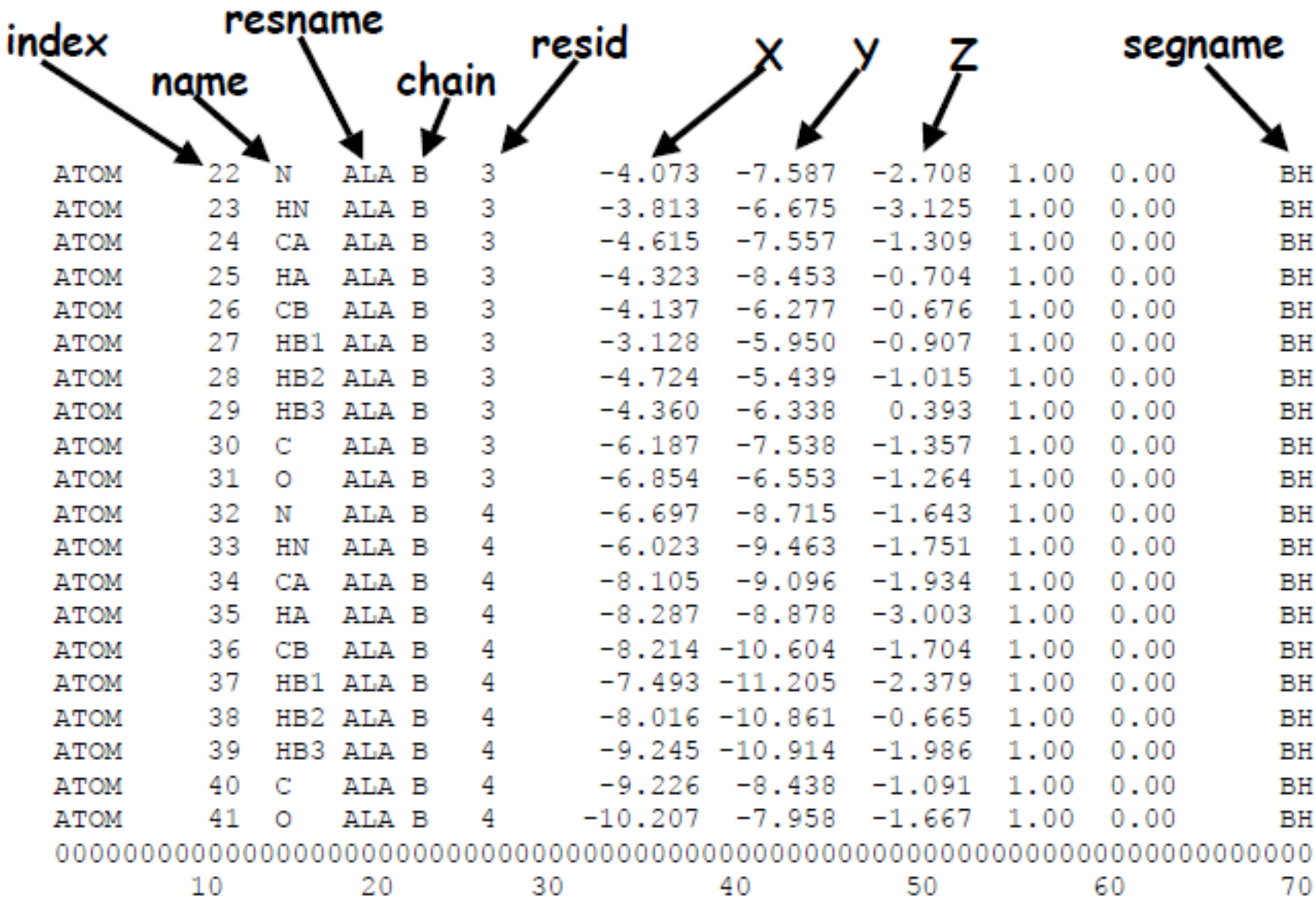
Double helix is right-handed with about 10–10.5 base pairs per turn



Experimental tools and datasets

- **Experimental tools**
  - X-Ray Crystallography
  - NMR Spectroscopy
  - *Cryogenic electron microscopy (Cryo-EM)*
- **Repositories**
  - Protein Data Bank
  - Cryo-Electron Microscopy Databank
- **Classification Data Bank**
  - CATH (Class, Architecture, Topology, Homologous superfamily)
  - SCOP (Structural Classification Of Proteins)
  - FSSP (Fold classification based on Structure-Structure alignment of Proteins)

Structure of a PDB file



2. Graph representation for biomolecules

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, ...

❖ ***Molecular descriptors***
❖ ***Molecular fingerprints***
❖ ***Graph neural network***

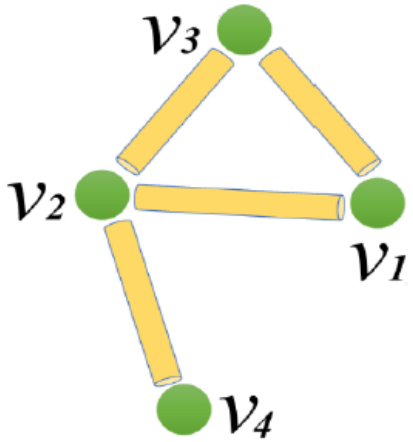
Algebraic graph Theory

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

Example: A simple (undirected) graph:

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

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



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

$$L = D - A$$

$L = B * B^T$

B:boundary operator!

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

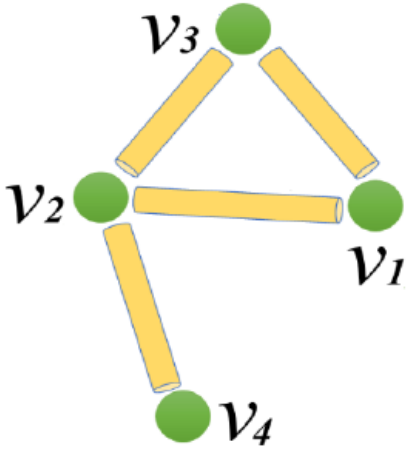
Degree of node i

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

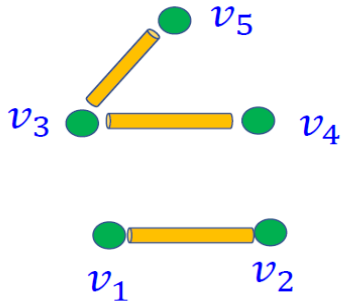
Graph partition: Partition a graph $G = (V, E)$ into smaller components with certain properties.

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).

Example I:

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

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



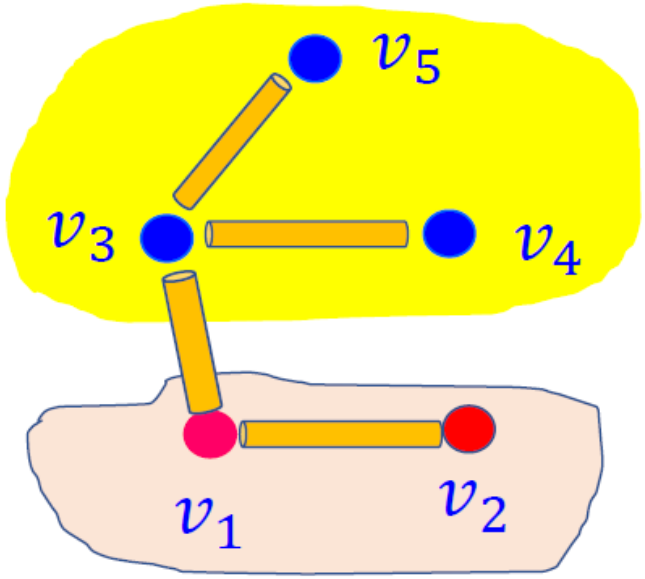
$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Two disconnected components (harmonic part due to the topology)

Spectral bisection

Example II:

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



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

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

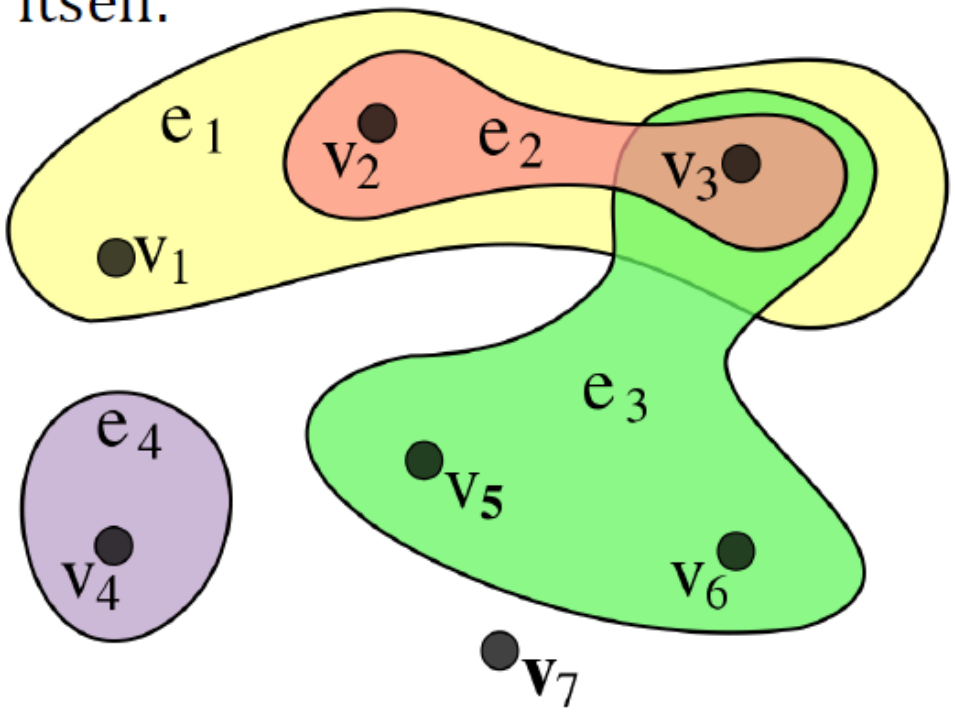
Graph Theory

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.



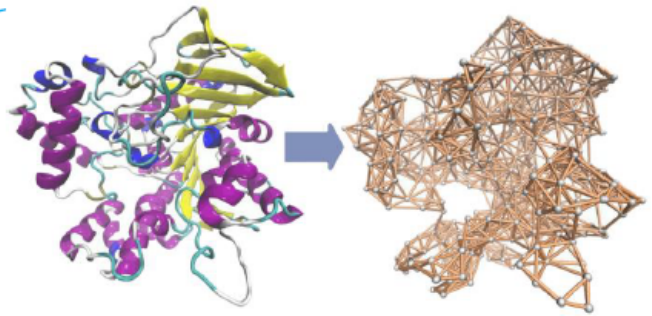
Gaussian Network Model (GNM)—Laplacian model

(Bahar, Atilgan and Erman, FD, 1997)

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

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

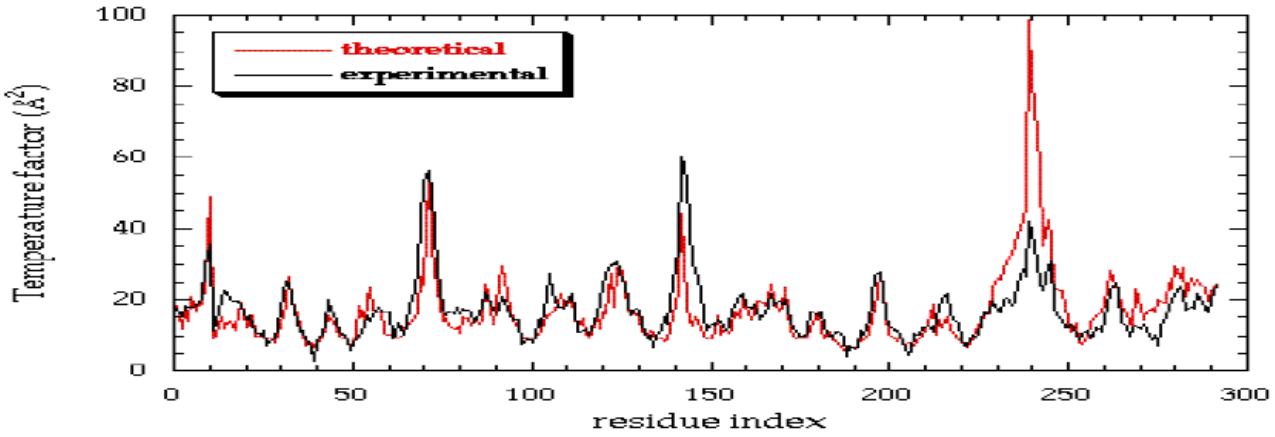
L is a Laplacian matrix, $O(N^3)$ method



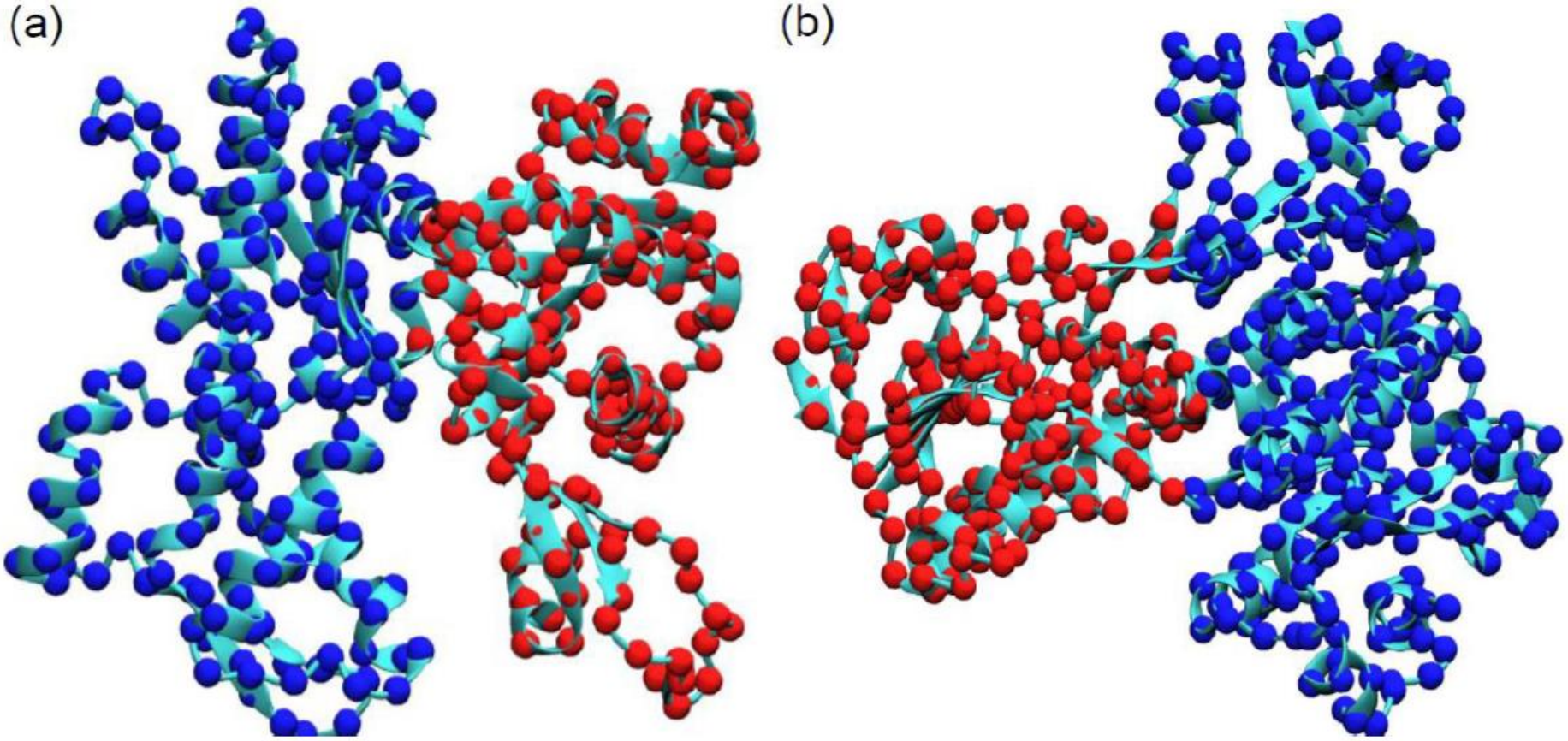
$$\left(L^{-1}\right)_{jj} = \sum_{k=2}^{N} \frac{1}{\lambda_k} \left[u_k u_k^T\right]_{jj}$$

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.



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.

Anisotropic network model

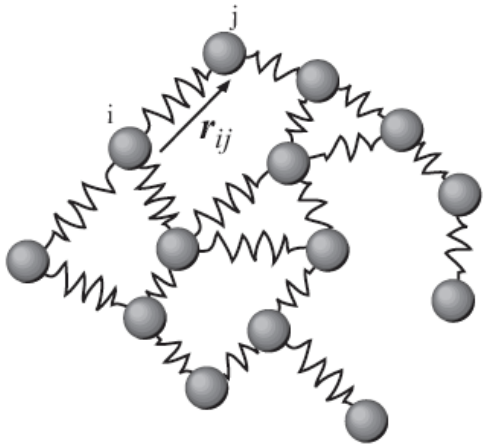
Potential function:

$r_{ij}^d = |\mathbf{r}_{ij}^d|$ $\quad r_{ij} = |\mathbf{r}_{ij}|$ $\quad \Delta\mathbf{R} = \{\Delta x_1, \Delta y_1, \Delta z_1, ..., \Delta x_N, \Delta y_N, \Delta z_N\}$

$$V^{\mathrm{ANM}} = \gamma \sum_{i,j}^{N} (r_{ij}^d - r_{ij})^2 f(r_{ij}) = \frac{\gamma}{2} \Delta\mathbf{R}^T H \Delta\mathbf{R}$$

$$H_{ij} = -\frac{1}{r_{ij}^2} \begin{bmatrix} (x_j - x_i)(x_j - x_i) & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)(y_j - y_i) & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)(z_j - z_i) \end{bmatrix} \quad i, j = 1, 2, \cdots, N, i \neq j \text{ and } r_{ij} \leq r_c$$

$$H_{ii} = -\sum_{i \neq j} H_{ij}, \ \forall i = 1, 2, \cdots, N$$



Moore–Penrose pseudoinverse:

eigenvector

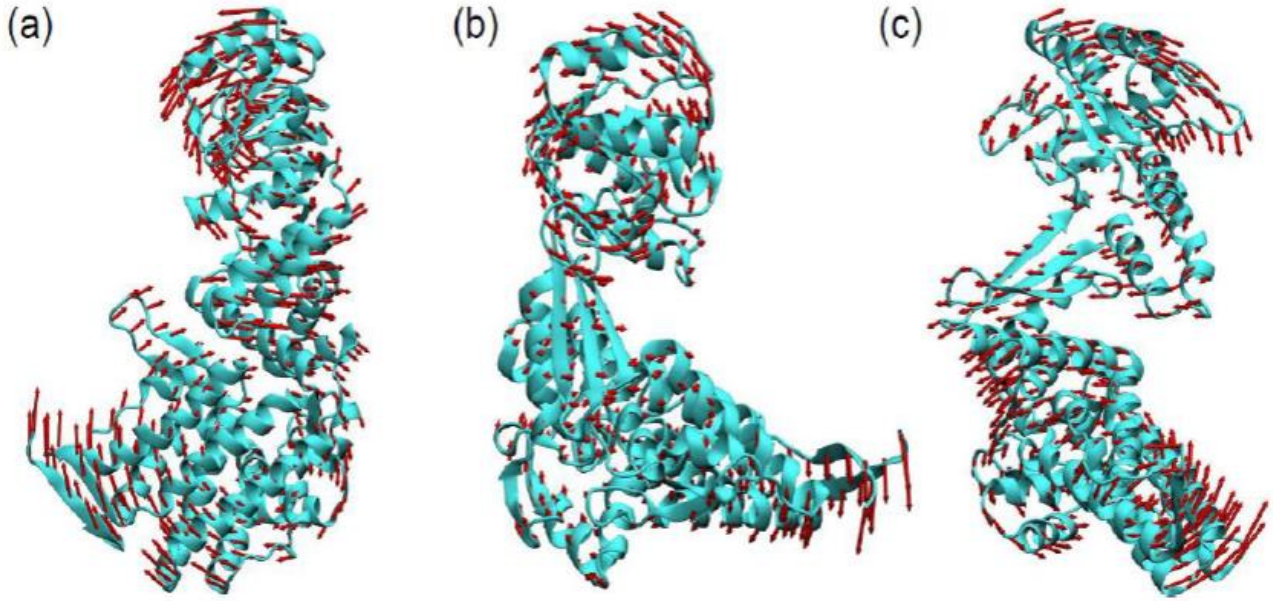
$$(H^{-1})_{ii} = \sum_{k=7}^{3N} \lambda_k^{-1} \left[\mathbf{v}_k \mathbf{v}_k^T\right]_{ii}$$

eigenvalue

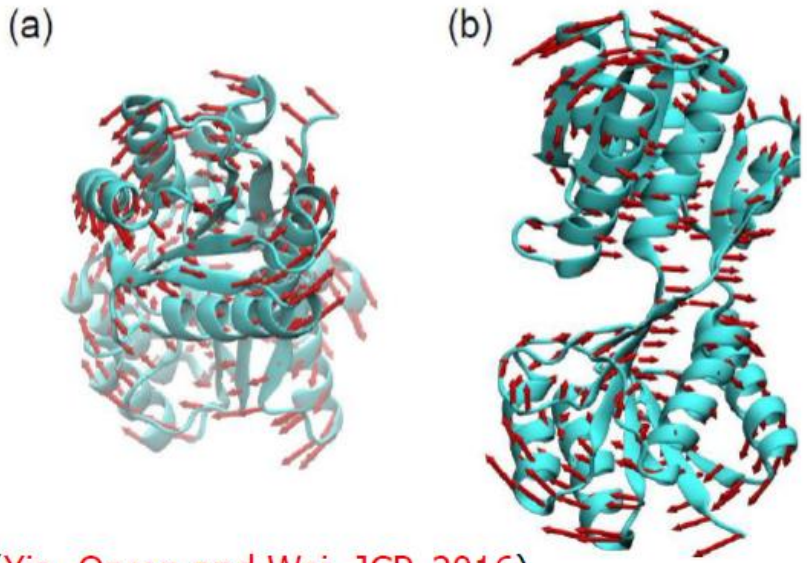
Predicted b-factor:

$$B_i^{\mathrm{ANM}} = \frac{8\pi^2}{3} \sum_{j=3i-2}^{3i} < \Delta\mathbf{R}_j \cdot \Delta\mathbf{R}_j >, \ \forall i = 1, 2, \cdots, N.$$

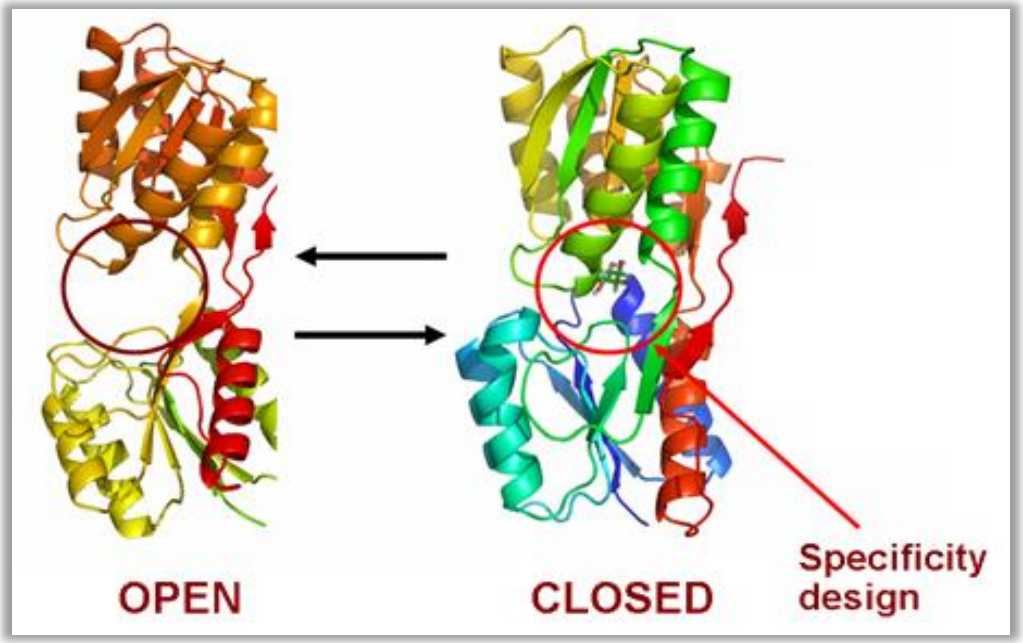
$$< \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j > = \frac{3k_B T}{\gamma} (H^{-1})_{ij}, \ \forall i, j = 1, 2, \cdots, 3N.$$



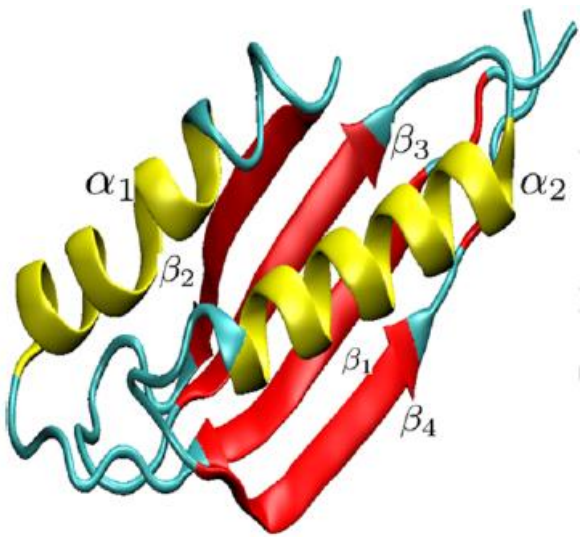
The motions of 1GRU (chain A). The 7th, 8th, and 9th nANM modes are demonstrated in (a)–(c), respectively

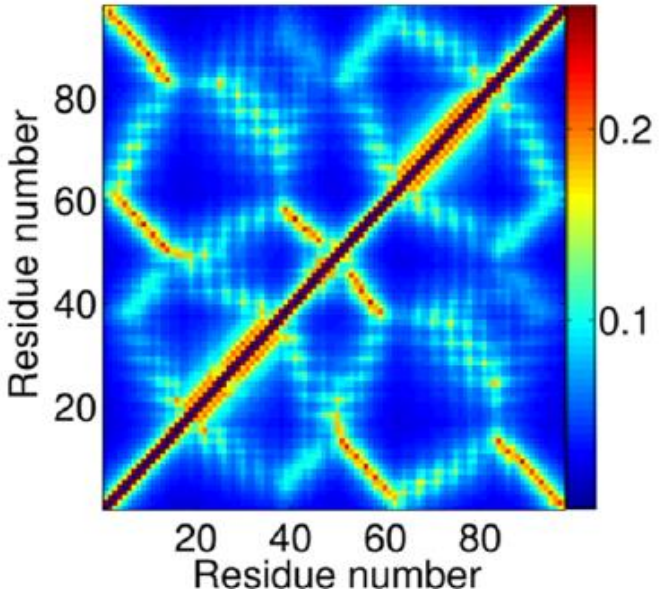


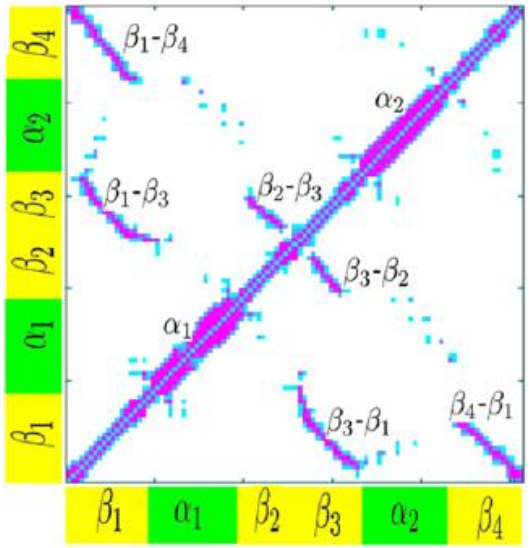
(Xia, Opron and Wei, JCP, 2016)



Flexibility rigidity index







Connectivity matrix:

$$A_{ij} = \begin{cases} \phi(\| r_i - r_j \|;\eta), & i \neq j; \\ -\sum_{i \neq j} A_{ij}, & i = j. \end{cases}$$

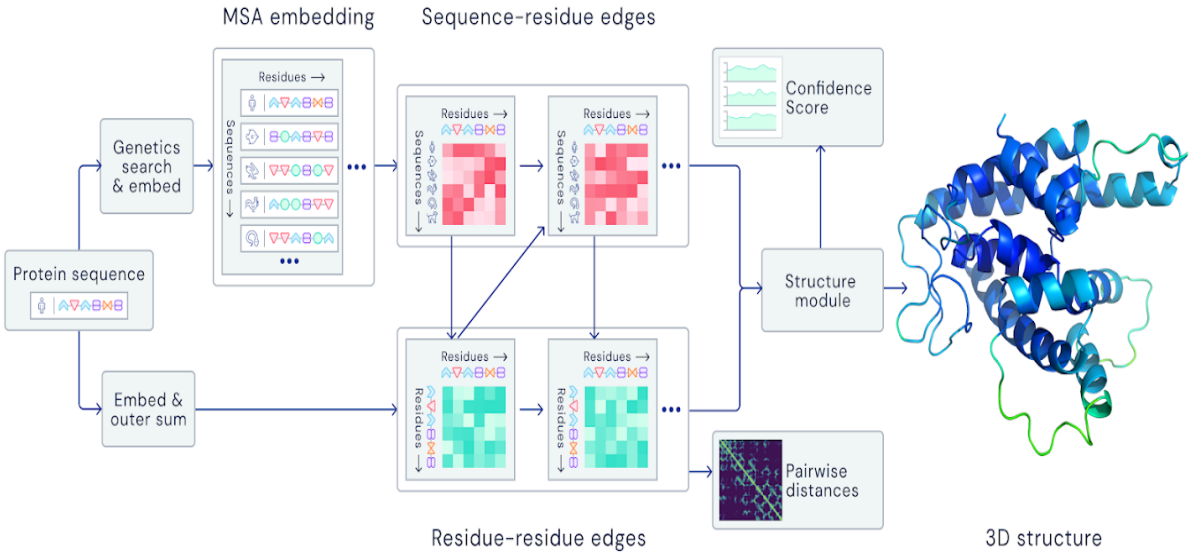
Kernel function:

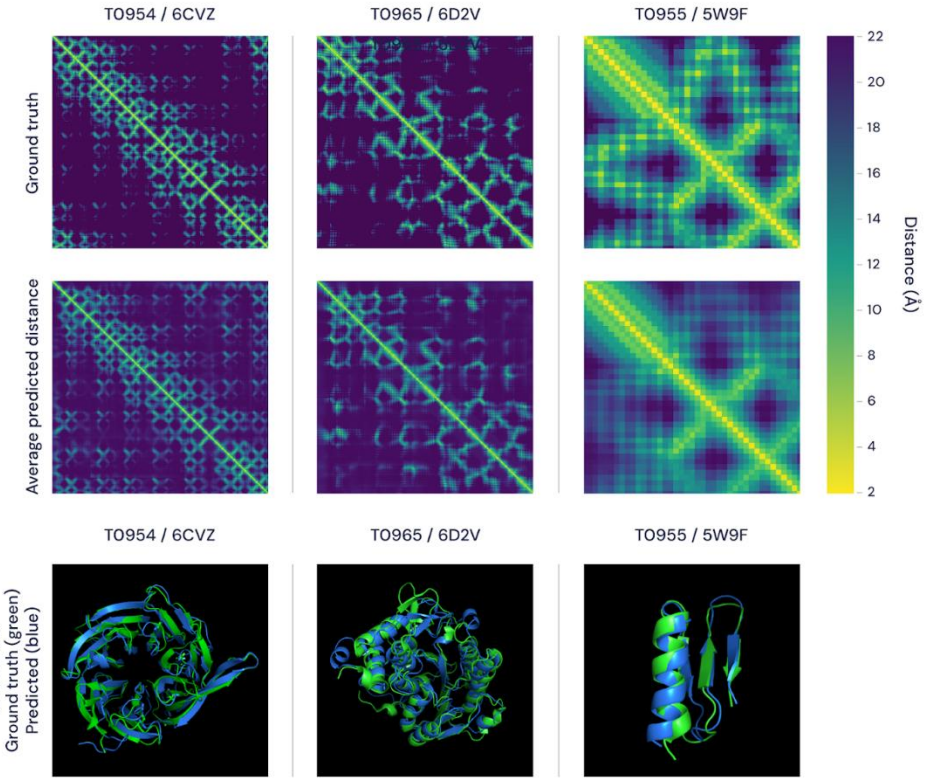
$$\phi(\| r_i - r_j \|;\eta) = 1, \ as \ \| r_i - r_j \| \to 0$$

$$\phi(\| r_i - r_j \|;\eta) = 0, \ as \ \| r_i - r_j \| \to \infty$$

Generalized exponential : $\phi(\| r_i - r_j \|;\eta) = e^{-(\|r_i - r_j\|/\eta)^{\kappa}}$

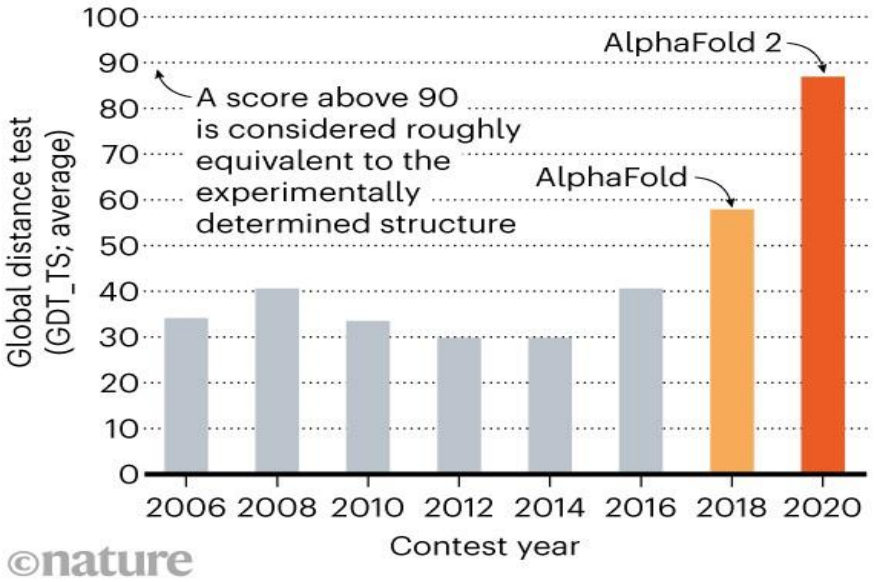
Generalized Lorentz: $\phi(\| r_i - r_j \|;\eta) = \dfrac{1}{1+(\| r_i - r_j \| / \eta)^{\nu}}$





STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



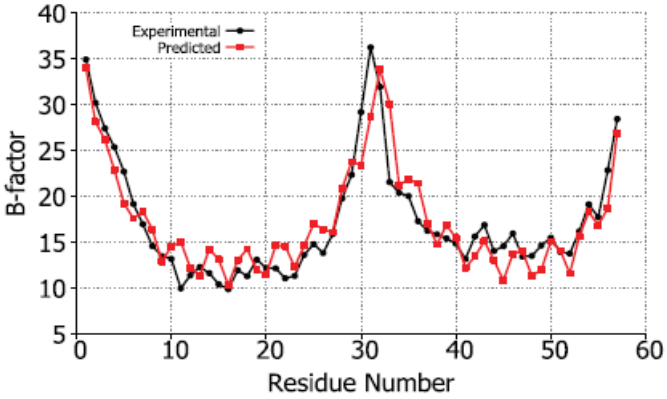
Flexibility rigidity index (FRI)

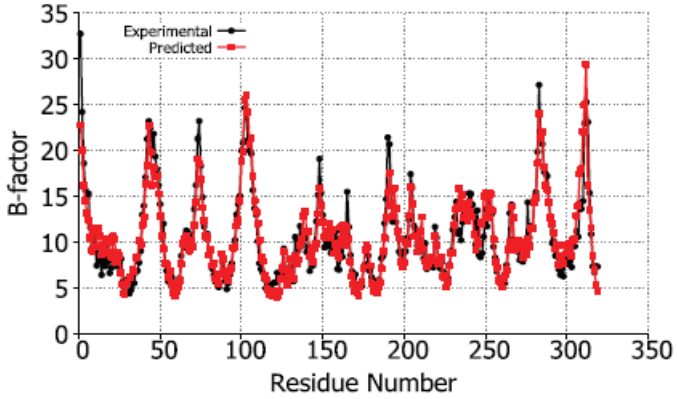
Rigidity index: $$\mu_i = \sum_{j \neq i} w_j \phi(\| r_i - r_j \|; \eta)$$

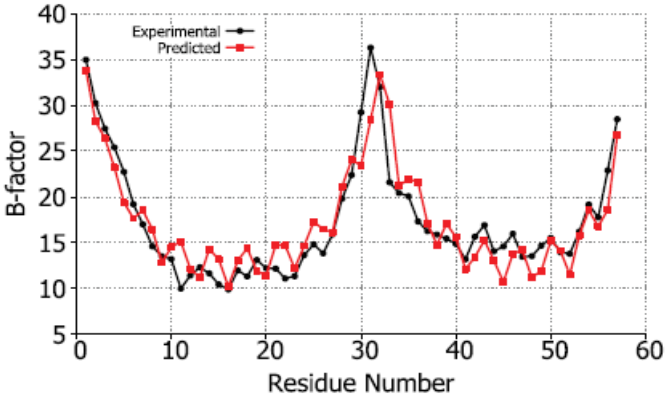
Flexibility index: $$f_i = \frac{1}{\mu_i} = \frac{1}{\sum_{j \neq i} w_j \phi(\| r_i - r_j \|; \eta)}$$

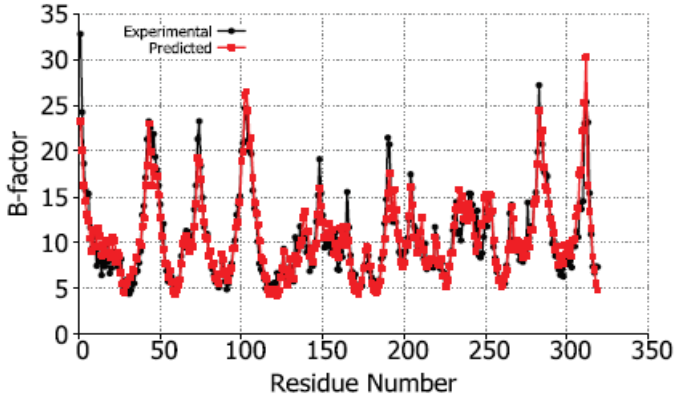
Upper parts: Lorentz kernel

Lower parts: exponential kernel







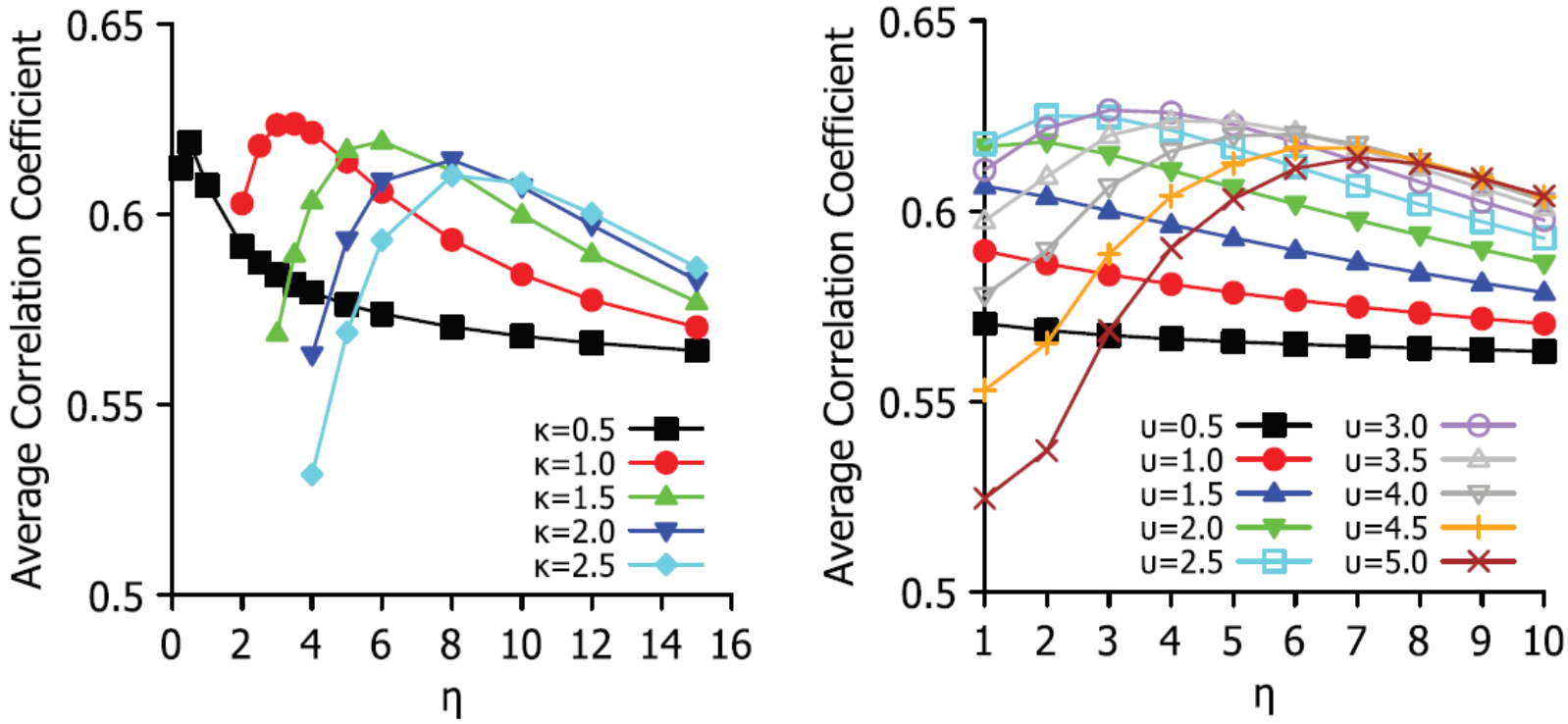


Protein ID: 1DF4 **Protein ID: 2Y7L**

Parameter testing

Correlation function	Parameter range		Average correlation coefficient
$e^{-(r/\eta)^{\kappa}}$	$1.0 \le \eta \le 10.0$	$0.5 \le \kappa \le 10.0$	0.676
$\frac{1}{1+(r/\eta)^{\upsilon}}$	$1.0 \le \eta \le 10.0$	$0.5 \le \upsilon \le 10.0$	0.673

$$C_c = \frac{\Sigma_{i=1}^{N}\left(B_i^e - \bar{B}^e\right)\left(B_i^t - \bar{B}^t\right)}{\left[\Sigma_{i=1}^{N}\left(B_i^e - \bar{B}^e\right)^2 \Sigma_{i=1}^{N}\left(B_i^t - \bar{B}^t\right)^2\right]^{1/2}}$$



Parameter testing for exponential (left chart) and Lorentz (right chart) using the dataset with 365 proteins.

Performance of our FRI

Accuracy: (10% improvement)

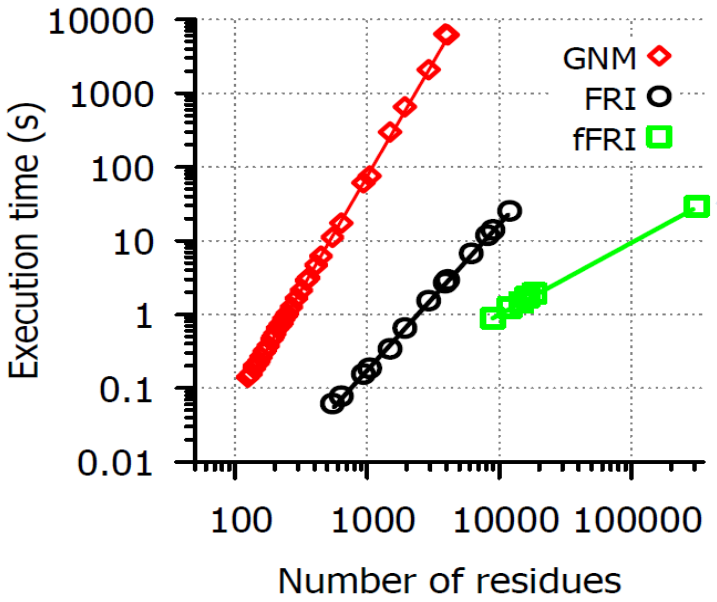
PDB set	pfFRI	GNM	NMA
Small	0.594	0.541	0.480
Medium	0.605	0.550	0.482
Large	0.591	0.529	0.494
Superset	0.626	0.565	NA

`.. atomic mean-square displacements is essentially determined by spatial variations in local packing density..”

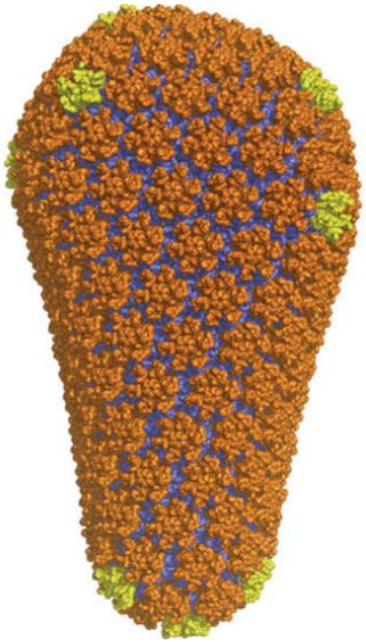
Bertil Halle, PNAS, Vol. 99, No.3, 1274-1279, 2002

Exponential parameters	Avg. CC	Lorentz parameters	Avg. CC
κ=0.5, η=0.5	0.615 (8.8%)	υ=2.5, η=2.0	0.622 (10.1%)
κ=1.0, η=3.0	0.623 (10.3%)	υ=3.0, η=3.0	0.626 (10.8%)
κ=1.5, η=6.0	0.619 (9.6%)	υ=3.5, η=4.0	0.623 (10.3%)

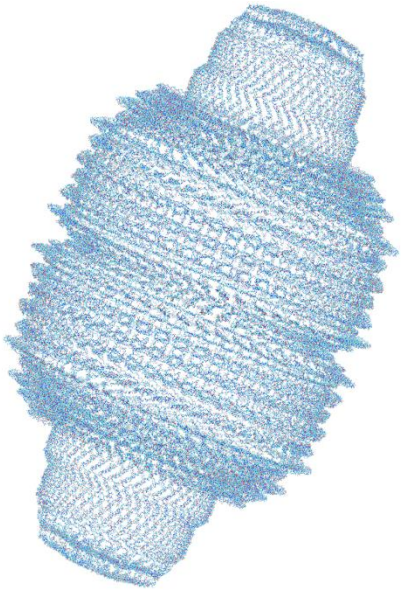
Time: fFRI O(N)

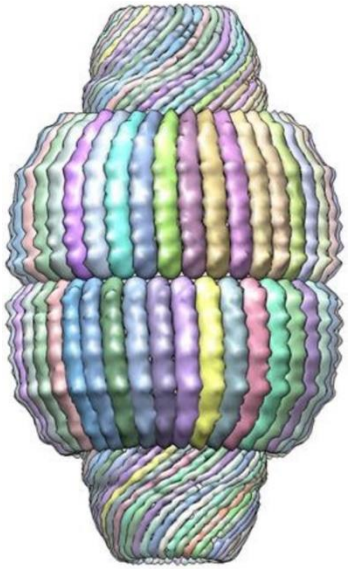


HIV Virus capsid (313 236 residues) in less than 30 seconds

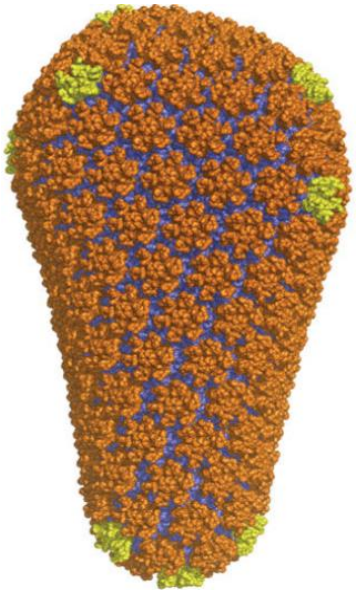


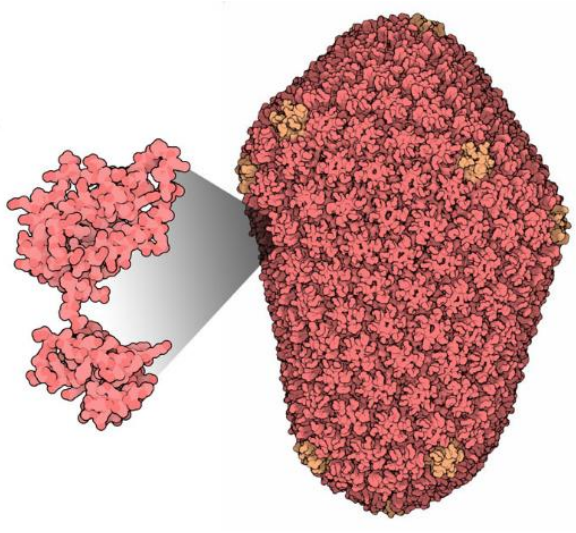
Extremely large biomolecules



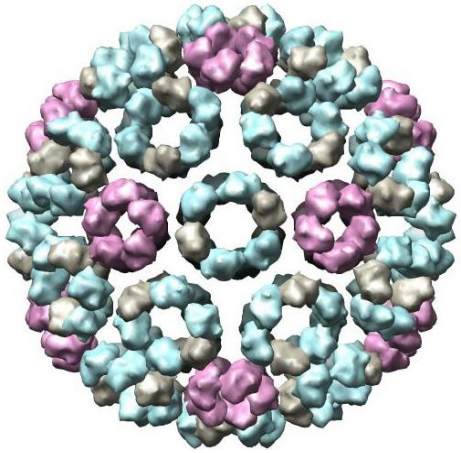


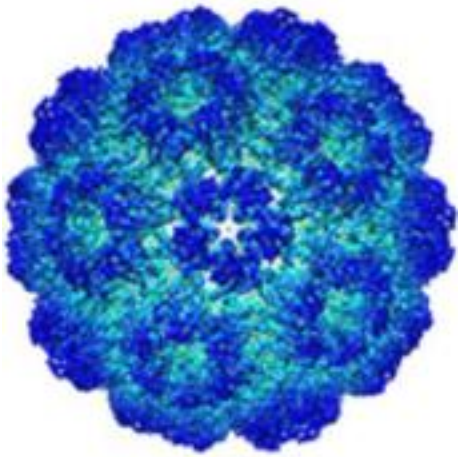
Vault particle



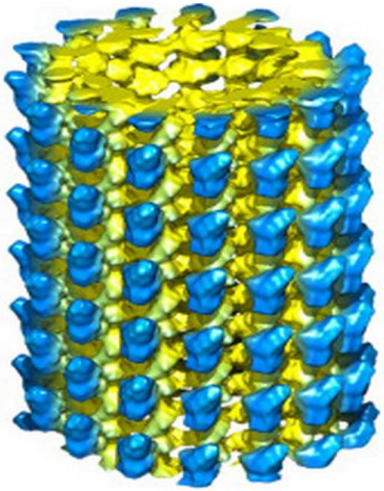


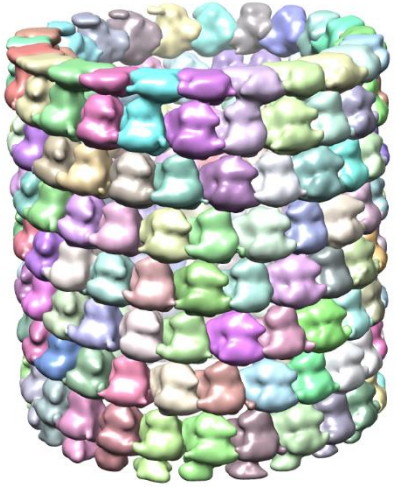
HIV virus capsid





Poliovirus capsid

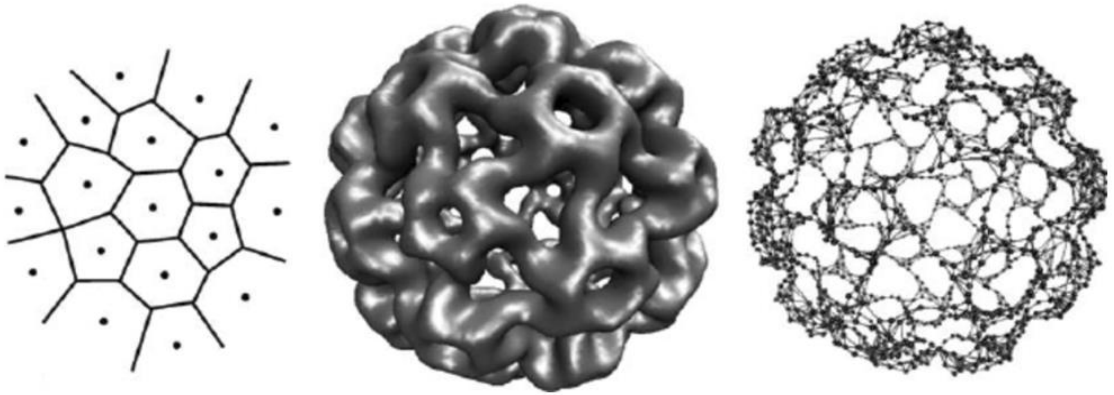




microtubule

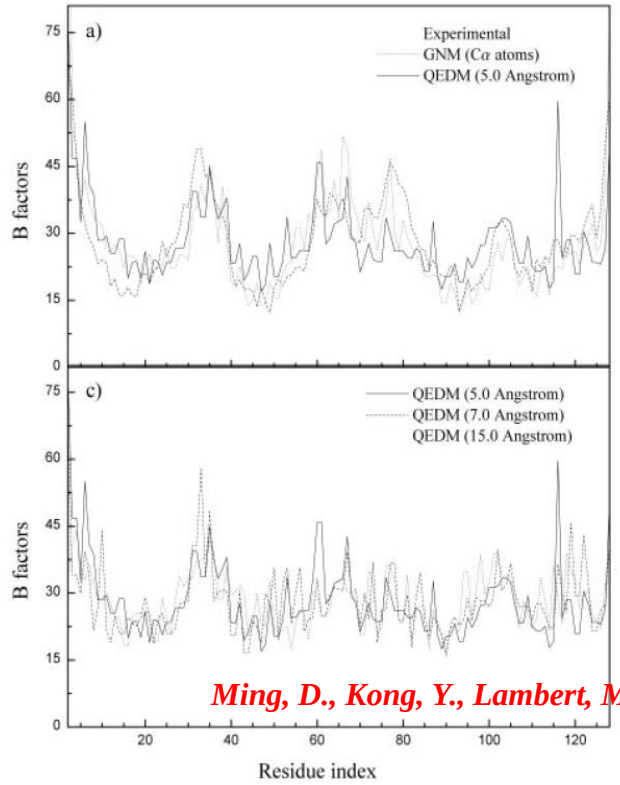
Quantized elastic deformational model (QEDM)

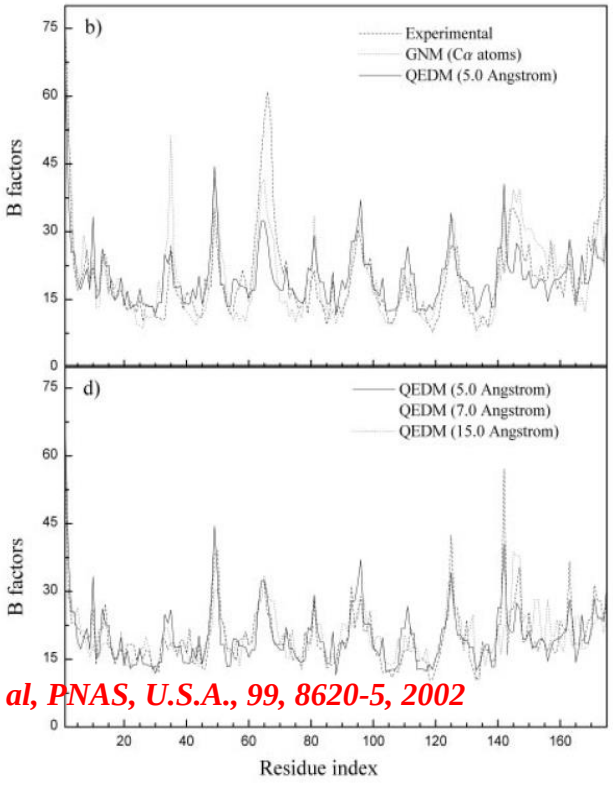
Voronoi Tessellation (Vector quantization)



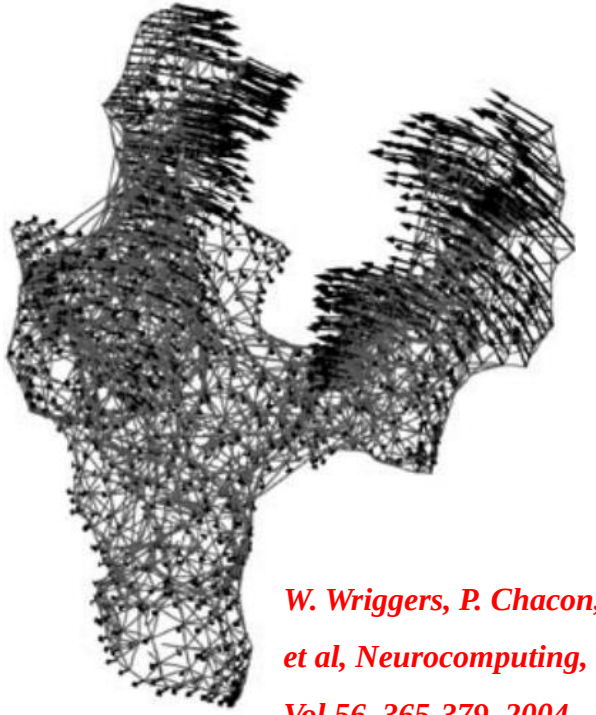
W. Wriggers, P. Chacon, et al, Neurocomputing, Vol 56, 365-379, 2004

Deformational motions are determined by GNM and ANM





Ming, D., Kong, Y., Lambert, M.A. et al, PNAS, U.S.A., 99, 8620-5, 2002

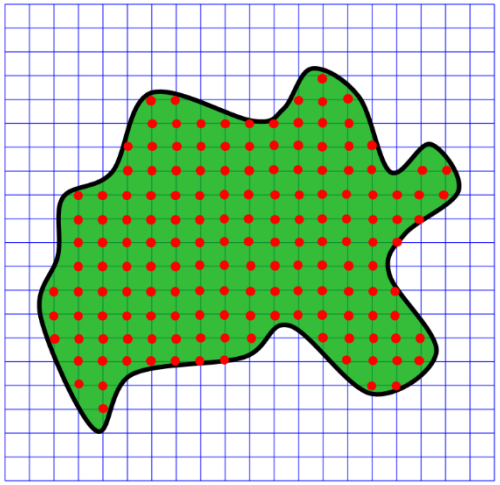


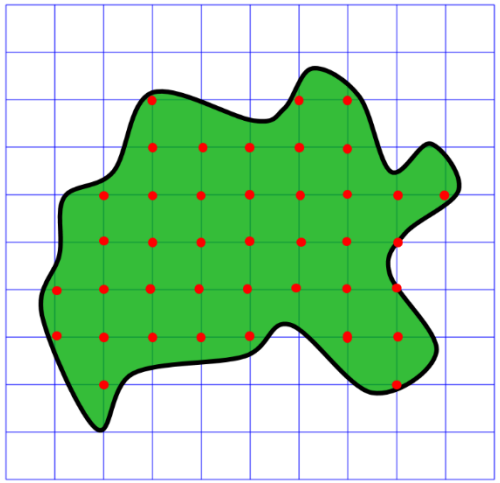
W. Wriggers, P. Chacon, et al, Neurocomputing, Vol 56, 365-379, 2004

Multiscale virtual particle model

Virtual particle generation:

Various types of meshes:
Cartesian grid;
Tetrahedral mesh;
Hexahedron; Voronio tessellation, etc.





Connection between particles:

$$\gamma(\mathbf{r}_I, \mathbf{r}_J, \Omega_I, \Omega_J, \mu^s(\mathbf{r}), \eta^{\mathrm{MVP}}) = \gamma_1(\Omega_I, \Omega_J, \mu^s(\mathbf{r})) \cdot \gamma_2(\mathbf{r}_I, \mathbf{r}_J, \eta^{\mathrm{MVP}})$$

Density contribution:

$$\gamma_1(\Omega_I, \Omega_J, \mu^s(\mathbf{r})) = \left(1 + a \int_{\Omega_I} \mu^s(\mathbf{r}) d\mathbf{r}\right)\left(1 + a \int_{\Omega_J} \mu^s(\mathbf{r}) d\mathbf{r}\right)$$

Distance contribution:

$$\gamma_2(\mathbf{r}_I, \mathbf{r}_J, \eta^{\mathrm{MVP}}) = e^{-\left(\|\mathbf{r}_I - \mathbf{r}_J\| / \eta^{\mathrm{MVP}}\right)^{\kappa}}, \quad \kappa > 0.$$

Multiscale virtual particle based Gaussian network model (MVP-GNM)

Potential function:

$$V^{\mathrm{MVP-GNM}} = \frac{1}{2}\Delta\mathbf{r}^T L^{\mathrm{MVP-GNM}} \Delta\mathbf{r}$$

$$L_{ij}^{\mathrm{MVP-GNM}} = \begin{cases} -\gamma(\mathbf{r}_I, \mathbf{r}_J, \Omega_I, \Omega_J, \mu^s(r)) & i \neq j \\ -\sum_{i\neq j}^{N} L_{ij} & i = j \end{cases}.$$

Moore–Penrose pseudoinverse:

$$\sum_{k=2}^{N} \lambda_k^{-1} \left[\mathbf{v}_k \mathbf{v}_k^T\right]_{ii}$$

eigenvector

eigenvalue

Predicted b-factor:

$$\frac{8\pi^2}{3} < \Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_i >, \ \forall i = 1, 2, \cdots, N$$

$$< \Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j >= \frac{3k_B T}{\gamma}(L^{-1})_{ij}, \ \forall i = 1, 2, \cdots, N$$

Multiscale representation of biomolecules

Kernel function:

$$\phi(\| r - r_j \|; \eta) = 1, \quad as \quad \| r - r_j \| \to 0$$

$$\phi(\| r - r_j \|; \eta) = 0, \quad as \quad \| r - r_j \| \to \infty$$

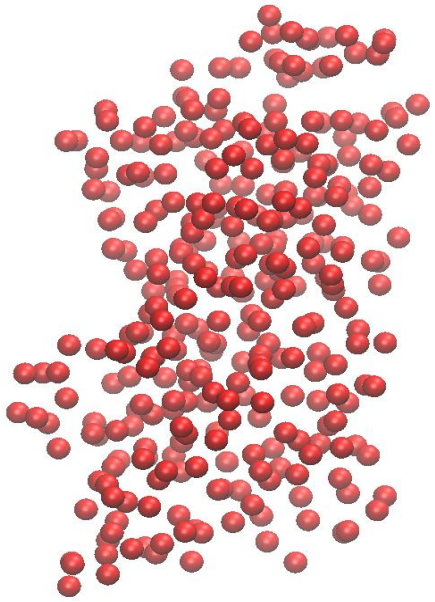
We use kernel

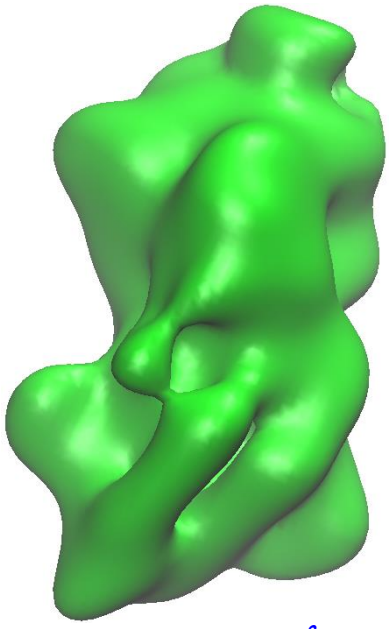
$$\phi(\| r - r_j \|; \eta) = e^{-(\|r - r_j\|/\eta)^2}$$

Rigidity function:

$$\mu(r) = \sum_{j}^{N} w_j \phi(\| r - r_j \|; \eta)$$

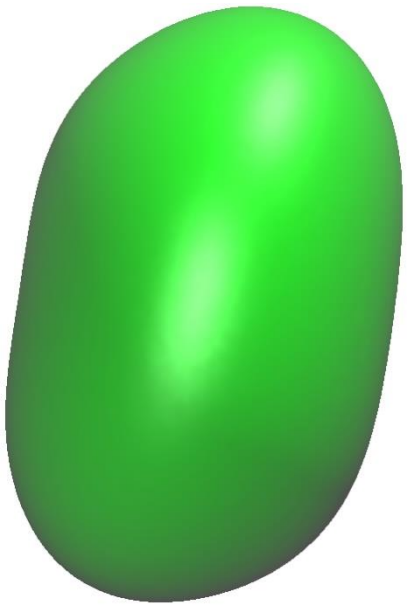
Protein ID: 2ABH





Resolution: 5Å

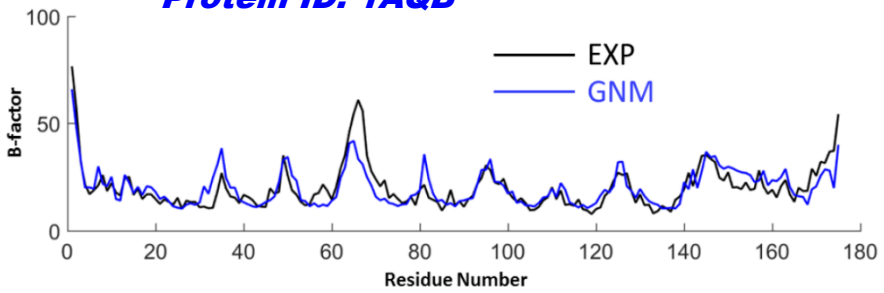
Resolution: 10Å

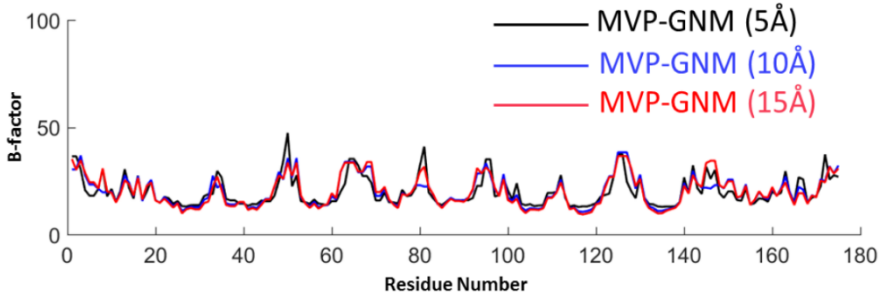


Resolution: 15Å

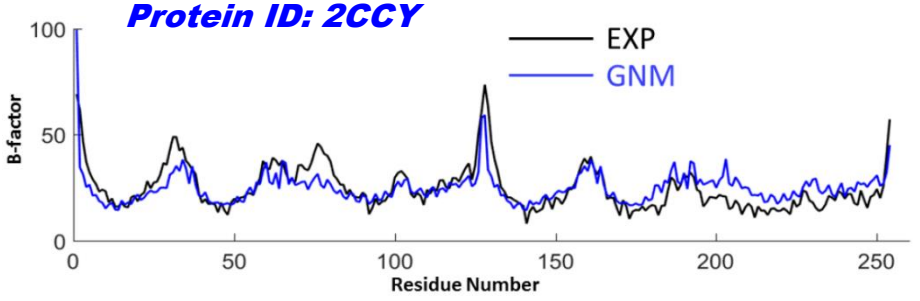
Validation of MVP-GNM

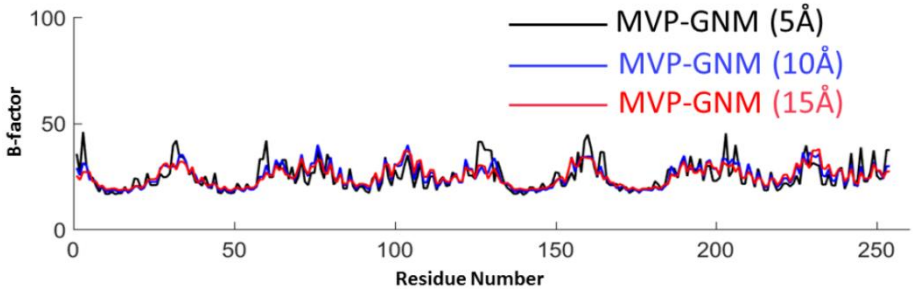
Protein ID: 1AQB



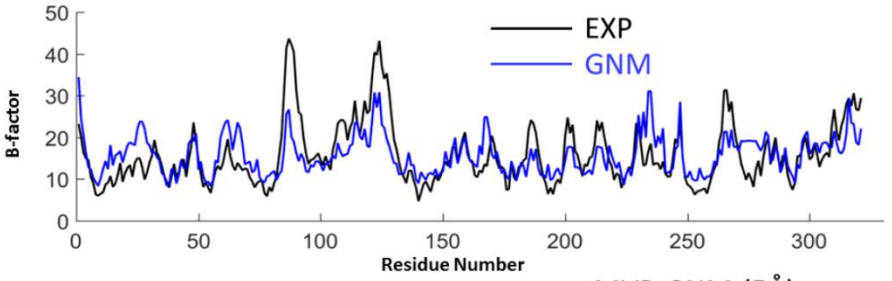


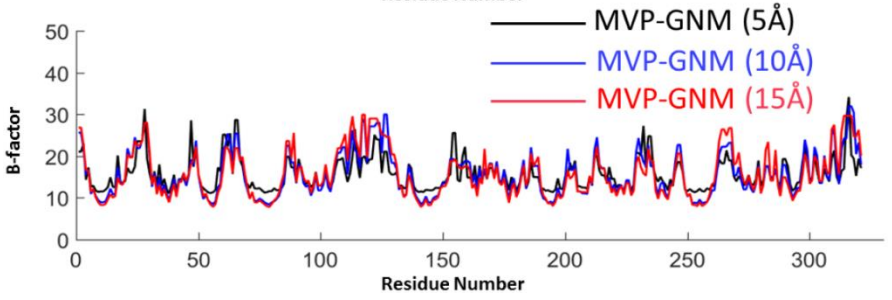
Protein ID: 2CCY





Protein ID: 2ABH





Cartesian grid with grid spacing 4Å!!

	GNM	MVP-GNM (5Å)	MVP-GNM (10Å)	MVP-GNM (15Å)
1AQB	0.822	0.666	0.657	0.699
2CCY	0.739	0.623	0.507	0.439
2ABH	0.647	0.550	0.731	0.775

Multiscale virtual particle based anisotropic network model (MVP-ANM)

Potential function:

$$V^{\mathrm{MVP-ANM}} = \frac{1}{2}\Delta\mathbf{R}^T H^{\mathrm{MVP-ANM}} \Delta\mathbf{R}$$

$$H_{IJ}^{\mathrm{MVP-ANM}} = -\frac{\gamma_{IJ}}{r_{ij}^2}\begin{bmatrix} (x_J-x_I)(x_J-x_I) & (x_J-x_I)(y_J-y_I) & (x_J-x_I)(z_J-z_I) \\ (y_J-y_I)(x_J-x_I) & (y_J-y_I)(y_J-y_I) & (y_J-y_I)(z_J-z_I) \\ (z_J-z_I)(x_J-x_I) & (z_J-z_I)(y_J-y_I) & (z_J-z_I)(z_J-z_I) \end{bmatrix} \quad I \neq J.$$

$$H_{II}^{\mathrm{MVP-ANM}} = -\sum_{I\neq J} H_{IJ}^{\mathrm{MVP-ANM}}, \ \forall i = 1, 2, \cdots, N.$$

Moore–Penrose pseudoinverse:

eigenvector

$$\sum_{k=7}^{3N} \lambda_k^{-1} \left[\mathbf{v}_k \mathbf{v}_k^T\right]_{ii}$$

eigenvalue

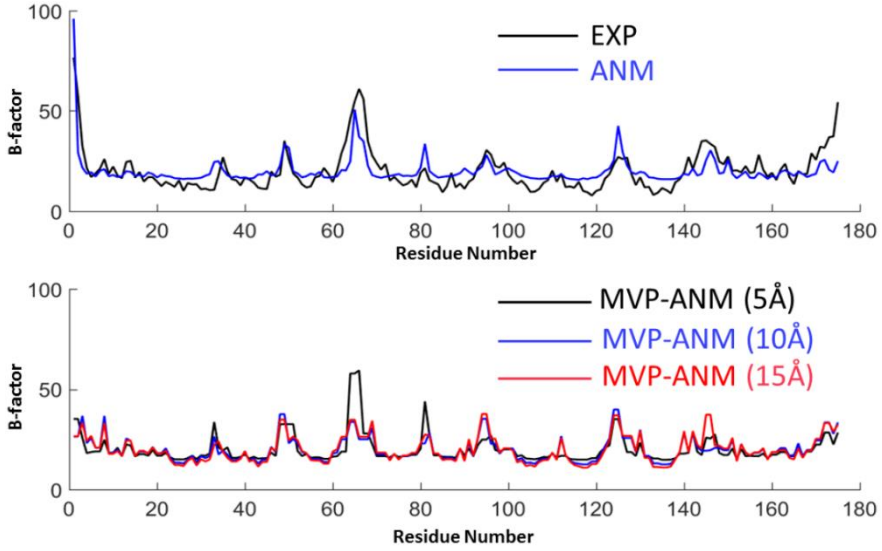
Predicted b-factor:

$$\frac{8\pi^2}{3} \sum_{j=3i-2}^{3i} < \Delta\mathbf{R}_j \cdot \Delta\mathbf{R}_j >, \ \forall i = 1, 2, \cdots, N.$$

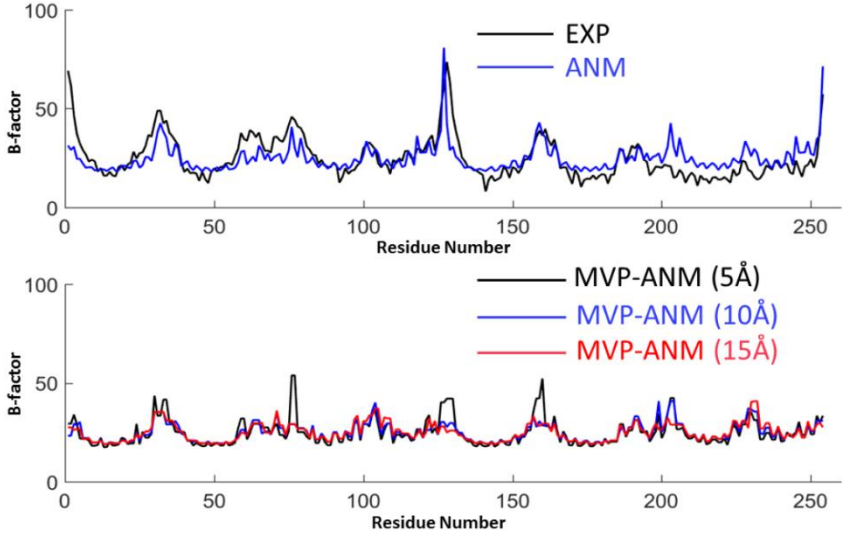
$$< \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j >= \frac{3k_B T}{\gamma}(H^{-1})_{ij}, \ \forall i, j = 1, 2, \cdots, 3N.$$

Validation of MVP-ANM

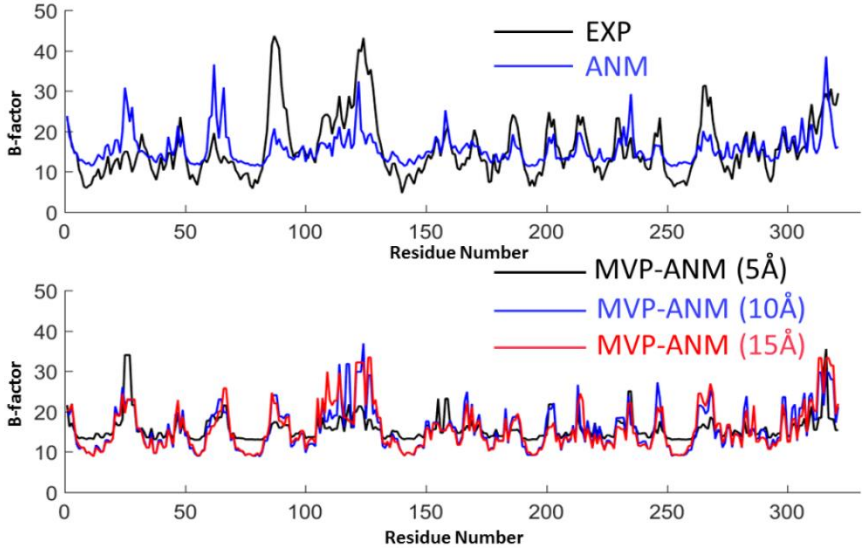
Protein ID: 1AQB



Protein ID: 2CCY



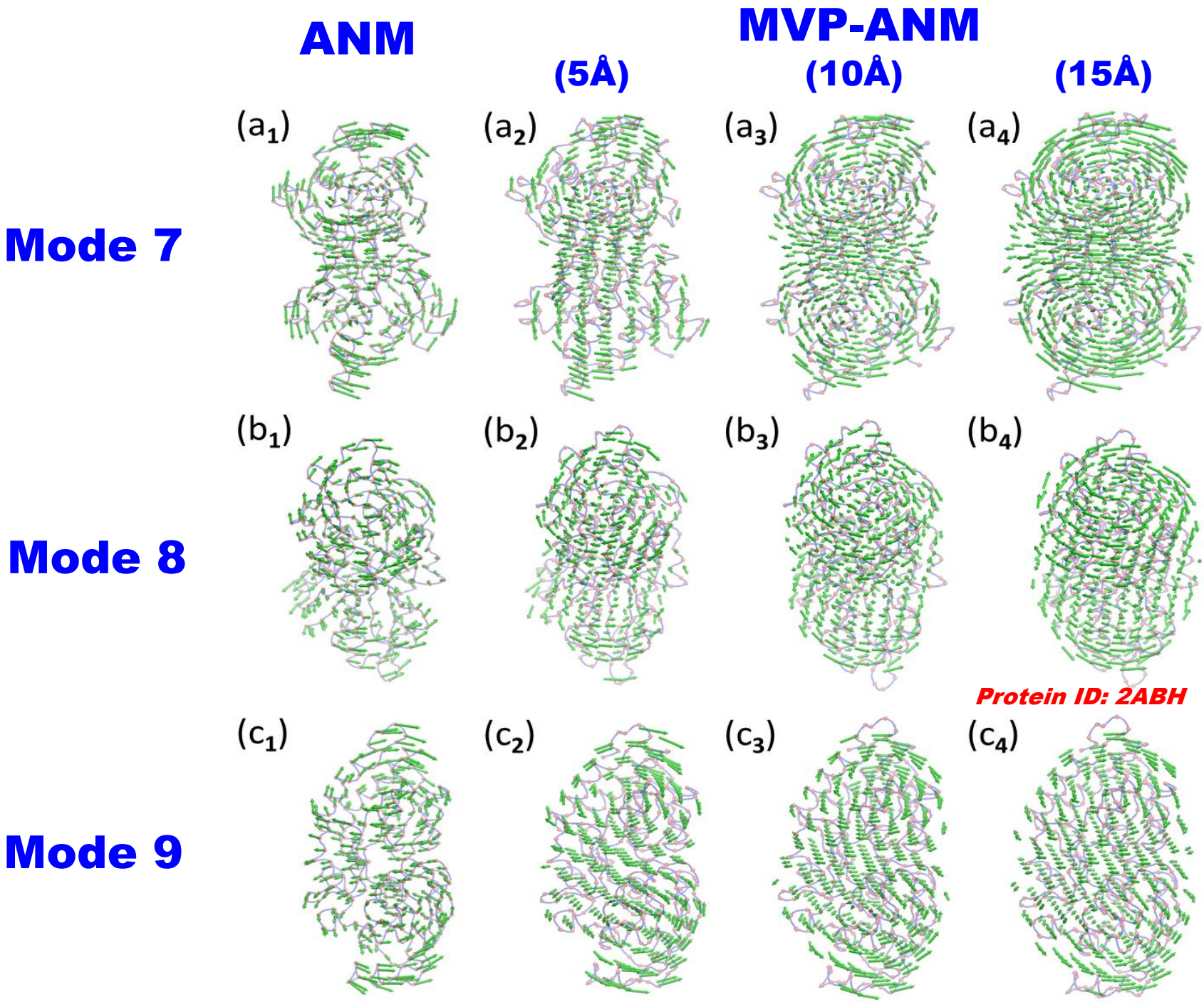
Protein ID: 2ABH



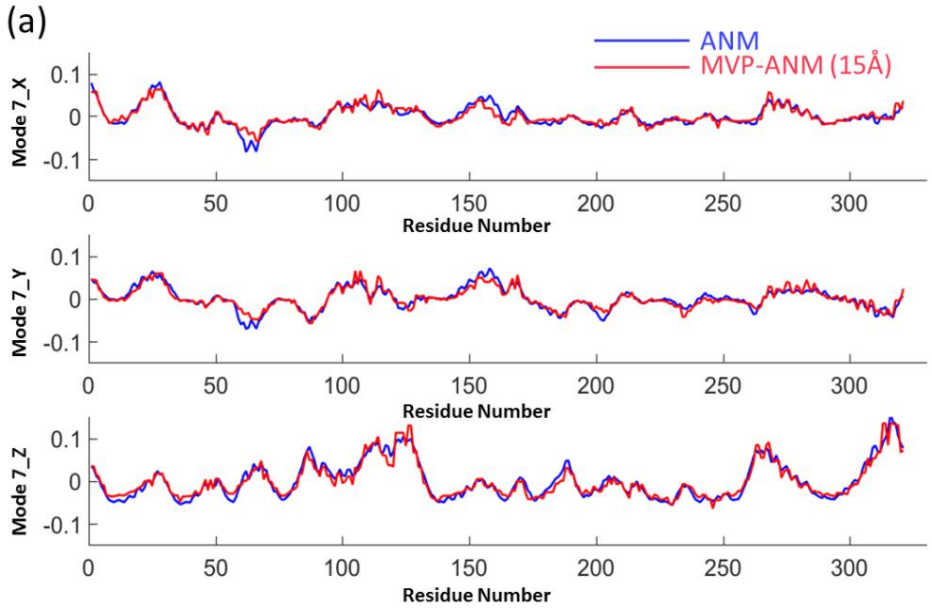
We use kernel $\phi(\| r - r_j \|; \eta) = e^{-(\|r - r_j\|/\eta)^2}$

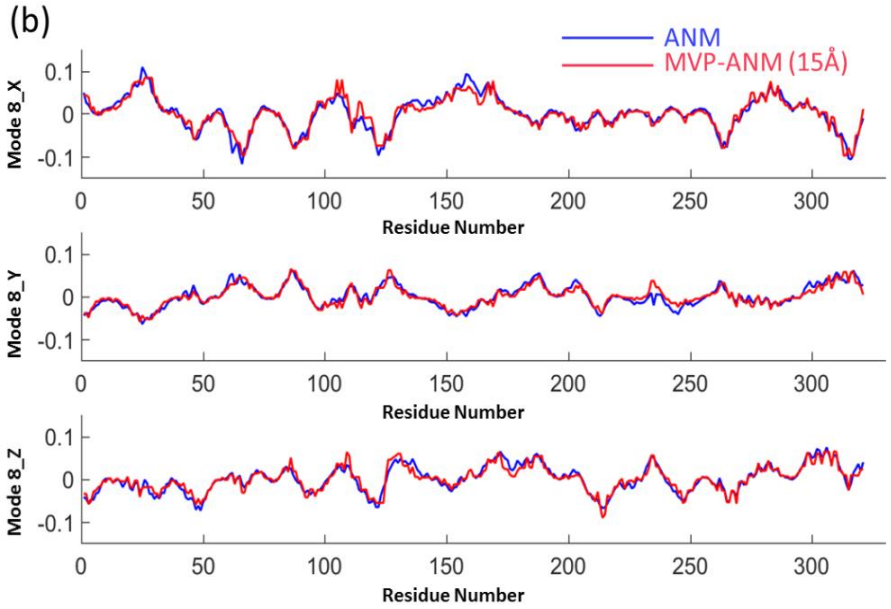
Cartesian grid with grid spacing 5Å

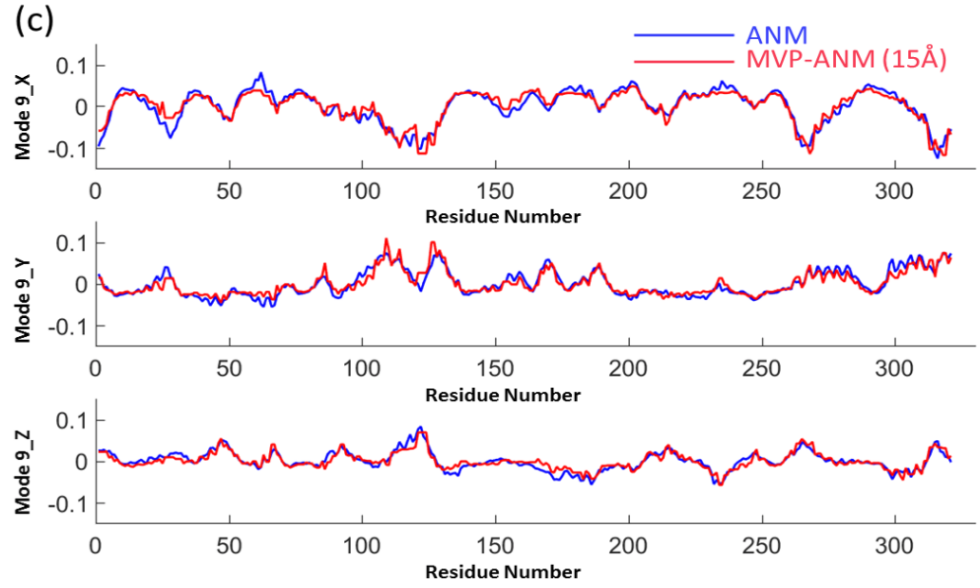
	GNM	MVP-ANM (5Å)	MVP-ANM (10Å)	MVP-ANM (15Å)
1AQB	0.725	0.696	0.593	0.646
2CCY	0.664	0.627	0.450	0.435
2ABH	0.548	0.442	0.743	0.760


ANM
MVP-ANM
(5Å)
(10Å)
(15Å)
Mode 7
Mode 8
Mode 9
(a1)
(a2)
(a3)
(a4)
(b1)
(b2)
(b3)
(b4)
(c1)
(c2)
(c3)
(c4)
Protein ID: 2ABH

Normal modes for protein 2ABH



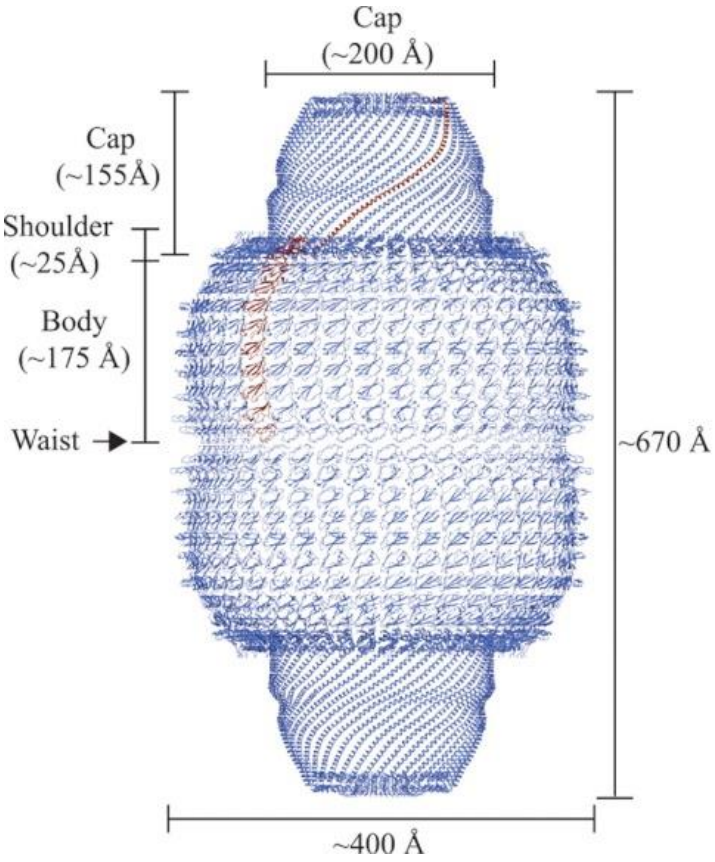


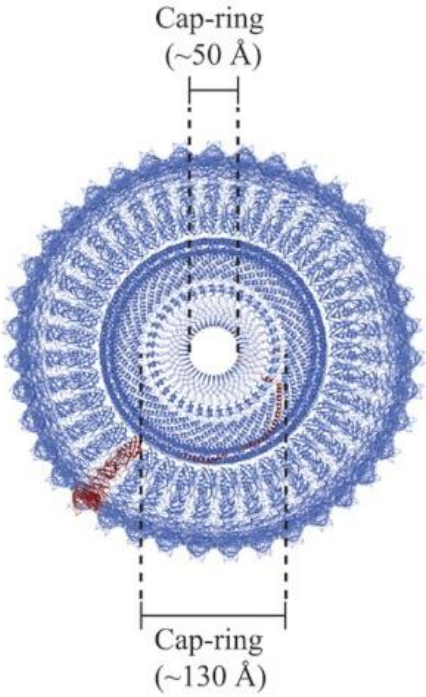


Resolution parameter is 15Å

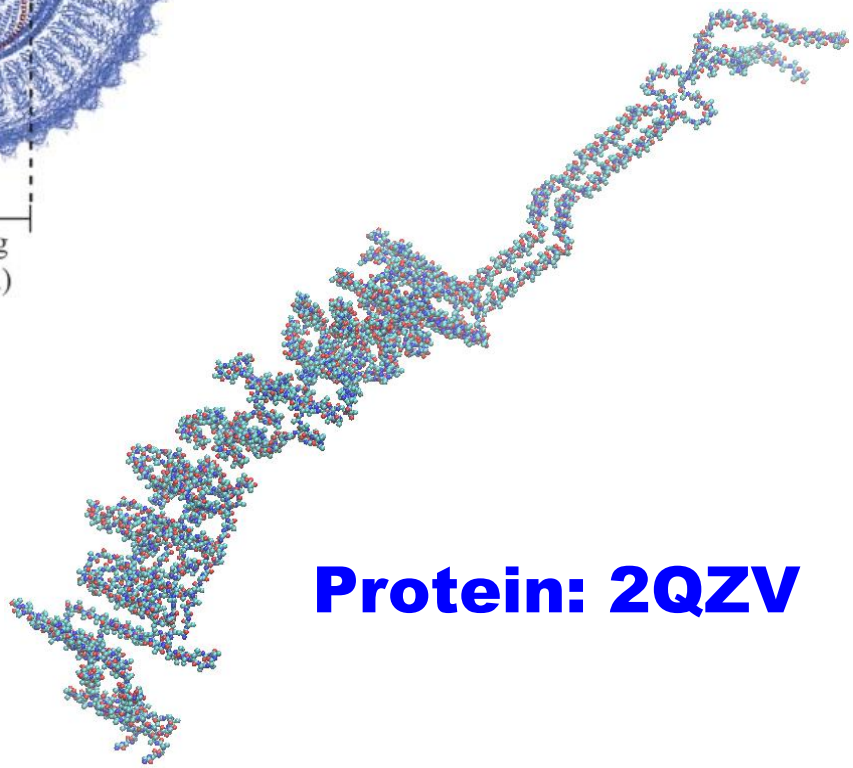
	MVP-ANM (X)	MVP-ANM (Y)	MVP-ANM (Z)
Mode 7	0.914	0.929	0.960
Mode 8	0.948	0.930	0.943
Mode 9	0.936	0.921	0.906

The dynamics of Vault Shell



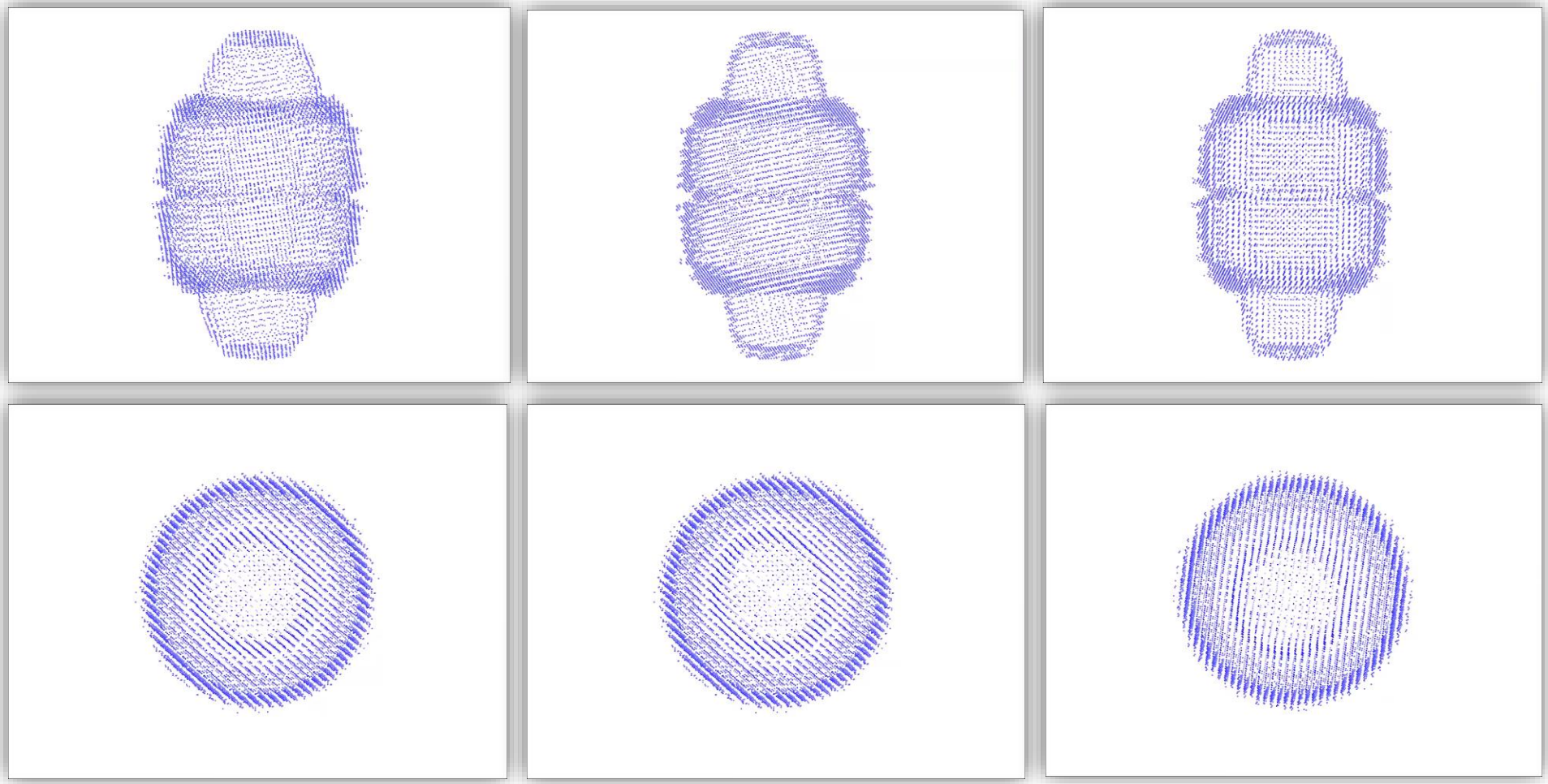


96 copies of vault protein 2QZV

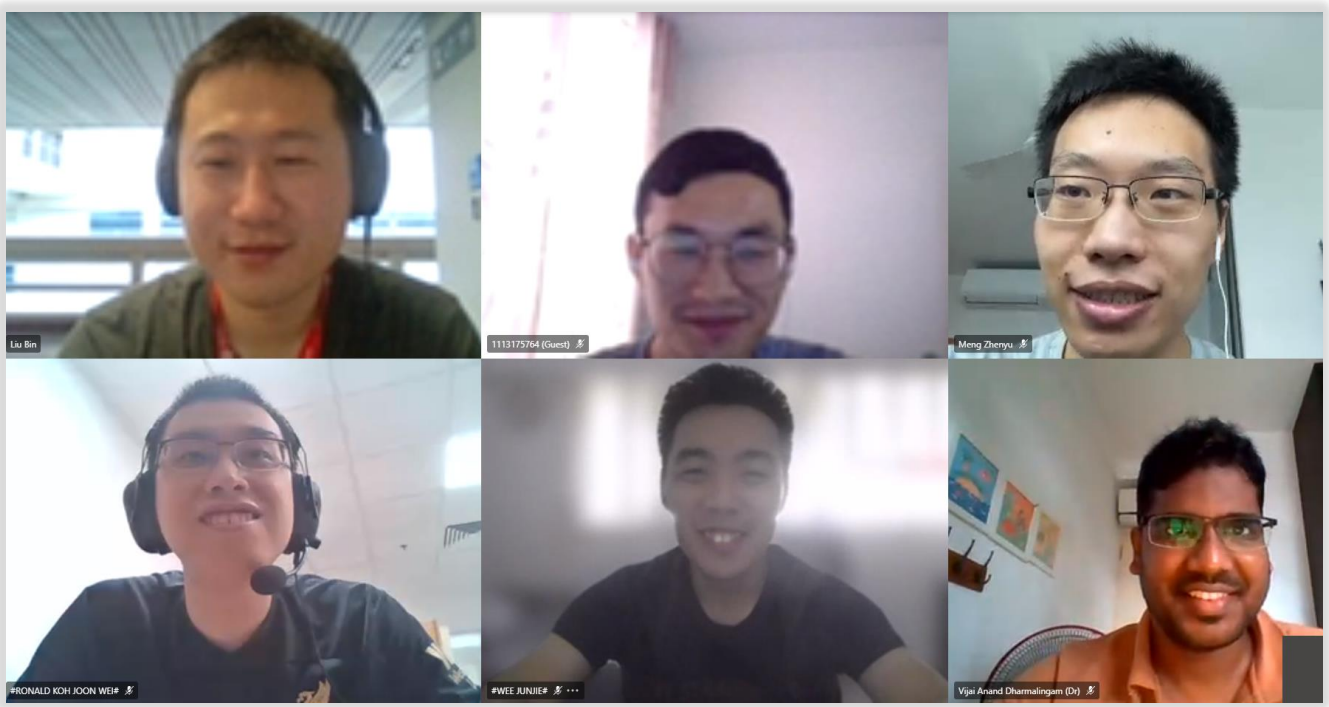


Protein: 2QZV

Multiscale virtual particle based elastic network model of Vault



Group members



Grant support

NTU-JSPS (2019-2022)
Alibaba-NTU (2020-2021)
Merlion (2020-2022)
MOE-Tier 1 (2018-2021, 2019-2022)
MOE-Tier 2 (2018-2021, 2021-2024)

Biophysics

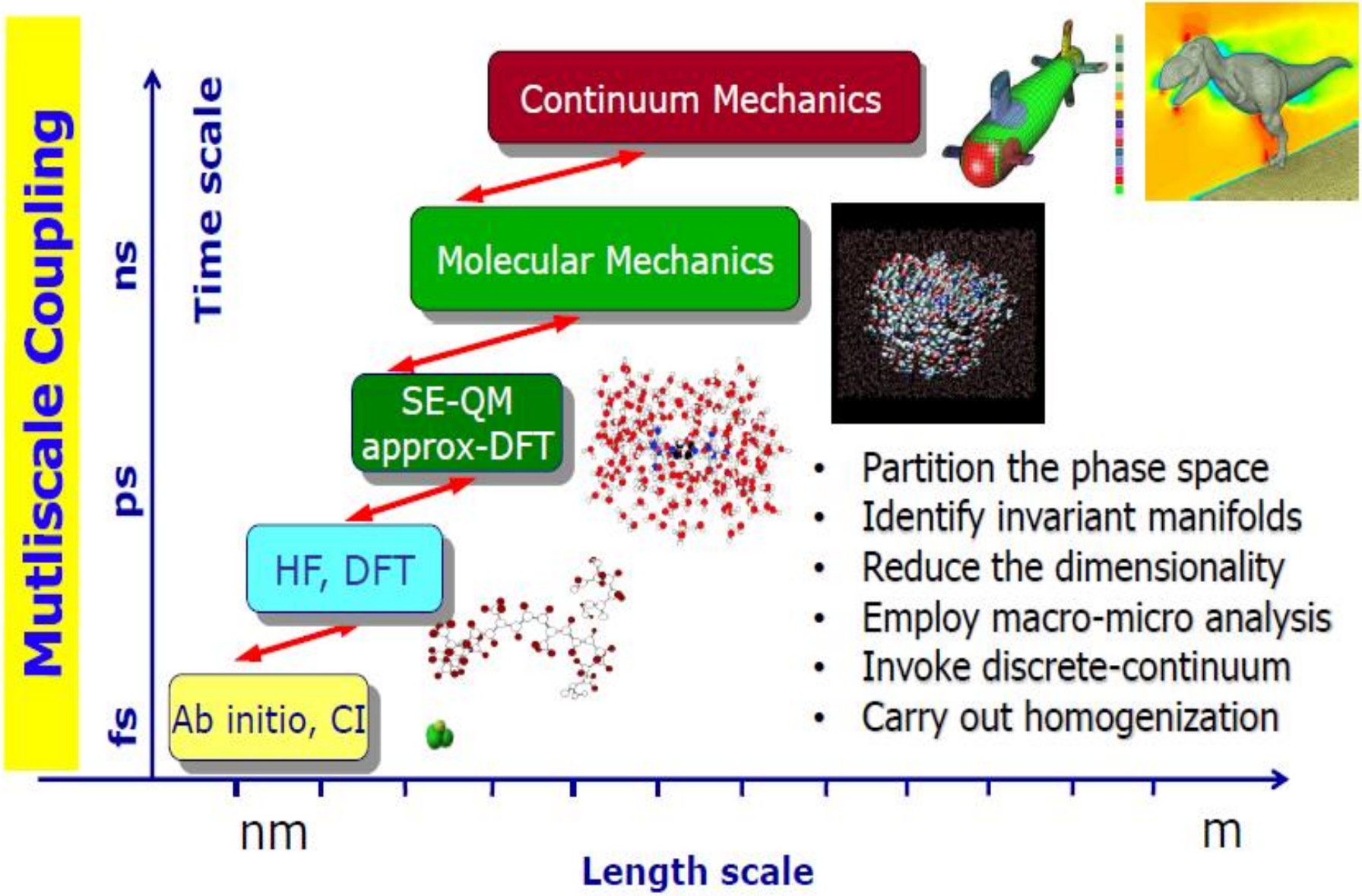
1) Biophysics is an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena.

2) Biophysics covers all scales of biological organization, from molecular to organismic and populations

3) Molecular biophysics applies physical approach to model biomolecular systems and understand their interactions and structure-function relationship.

4) Unlike data-driven bioinformatics and knowledge-based systems biology, biophysics is mechanistic.

Hierarchy of Methods



Molecular Mechanics (MM)

Using classical particle assumption.

Newton's second law: $m_i \frac{d^2}{dt} \boldsymbol{r}_i = \boldsymbol{F} = -\nabla U(\boldsymbol{r}_1(t), \boldsymbol{r}_2(t), \ldots, \boldsymbol{r}_n(t))$

Aproximation: $U = \sum_{\text{bonds}} K_d \, (d - d_0)^2 + \sum_{\text{angle}} K_\theta \, (\theta - \theta_0)^2$

$$+ \sum_{\text{dihedrals}} K_\chi \, (1 + \cos(n\chi - \delta))$$

$$+ \sum_{\text{nonbond}} \left\{ \epsilon_{ij} \left[\left(\frac{R_{ij}^{\min}}{\boldsymbol{r}_{ij}} \right)^{12} - \left(\frac{R_{ij}^{\min}}{\boldsymbol{r}_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon \boldsymbol{r}_{ij}} \right\}$$

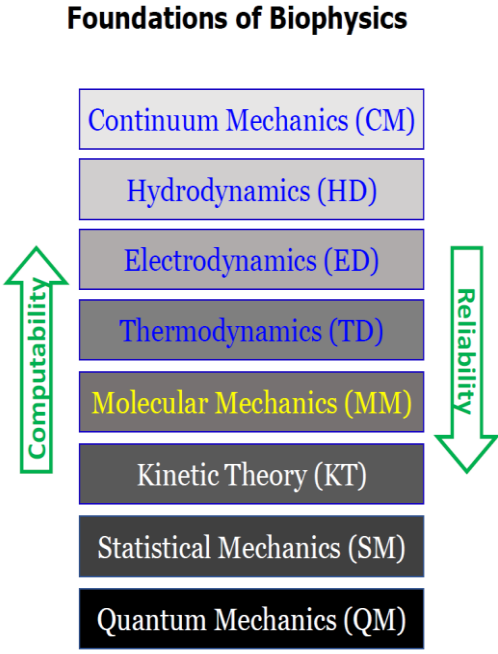
Langevin equation: $m\ddot{\boldsymbol{r}} = -\nabla V(\boldsymbol{r}) - \gamma \dot{\boldsymbol{r}} + \sqrt{2\gamma k_\beta T} R(t)$,

where, $\langle R(t) \rangle = 0$, $\langle R(t) R(t') \rangle = \delta(t - t')$

Explicit solvent/Implicit solvent/Coarse Grained

MM Software: AMBER/CHARMM/NAMD/TINKER/GROMOS

Research issues: high-order force fields, coarse grained methods, implicit MD, etc.

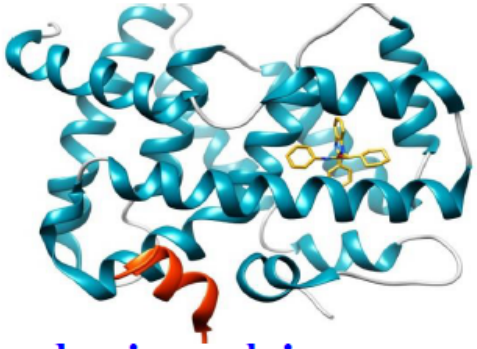


Bioinformatics

1) Sequence analysis (DNA sequencing, sequence assembly, genome annotation, comparative genomics, pan genomics, computational evolution, genetics, cancer mutation, etc.).
2) Gene and protein expression (Gene expression analysis, protein expression analysis, gene regulation, genotype–phenotype map, etc.).
3) Systems biology (Pan networks, integrated systems analysis).
4) Cellular organization (Microscopy and image analysis, protein localization, membrane mechanics, chromatin analysis, etc.).
5) Structural bioinformatics (Biomolecular structure and interaction, structure-function relationship, protein folding, protein design, etc..
6) Database and software.
7) Unlike biophysics, bioinformatics is data-driven.

Protein Structure Prediction

SVYDAAAQLTADVKKDLRDSW
KVIGSDKKGNGVALMTTLFAD
NQETIGYFKRLGNVSQGMAND
KLRGHSITLMYALQNFIDQLD
NPDSLDLVCS .....



1) Understand protein structure-function relationship
2) Design protein with desired function
3) Drug development
4) Methods (knowledge-based):

Template-based modeling (homology modeling) is used when there is one or more similar known structures in PDB.

Ab initio structure prediction (e.g., Rosetta) is used when one cannot find any similar structure.

Deep learning (e.g., AlphaFold, CNN, RNN)

5) Evaluation:

Critical Assessment of protein Structure Prediction (CASP)
Knowledge-based methods win; QM/MM do not work well.

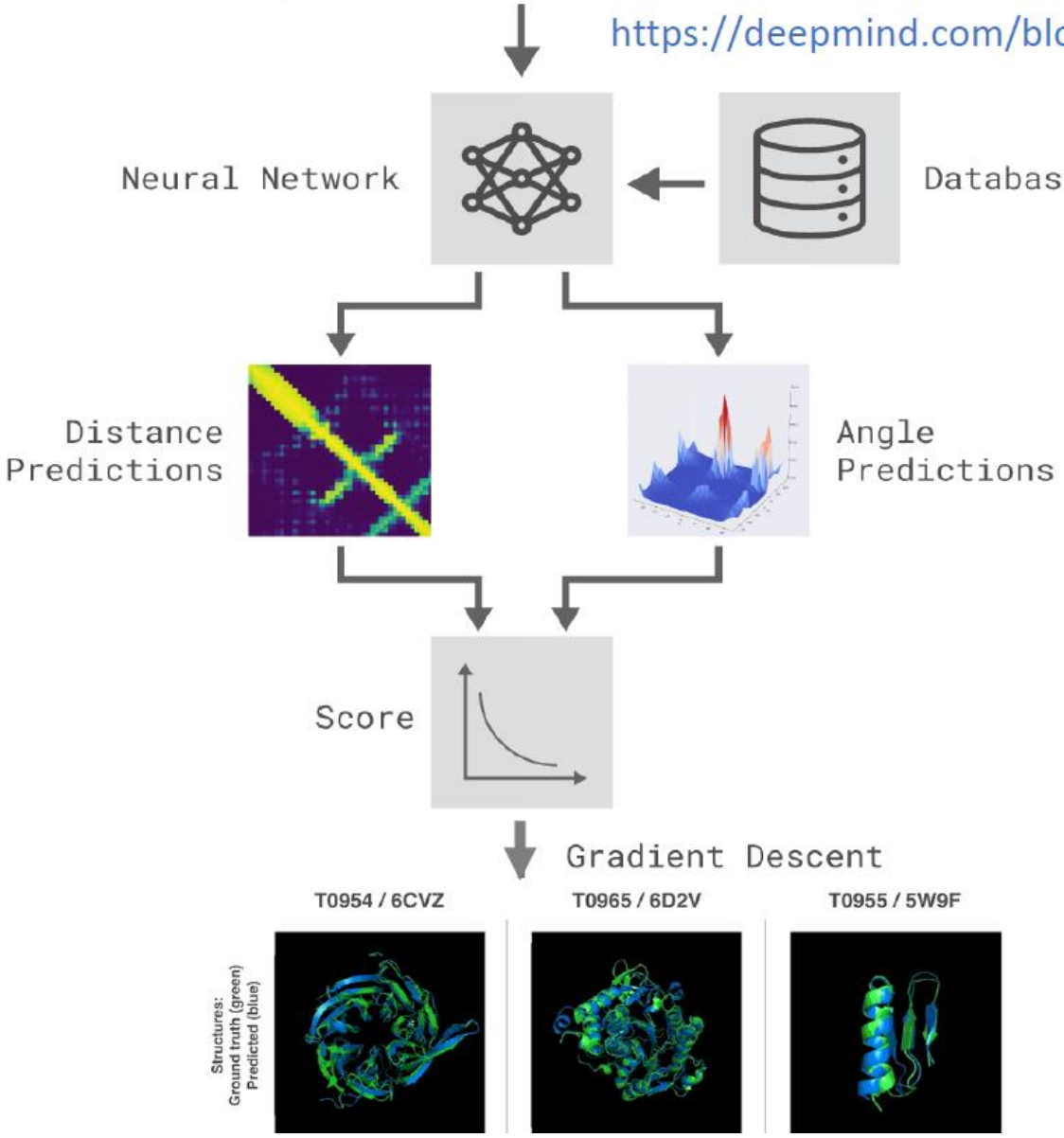
ALPHAFOLD

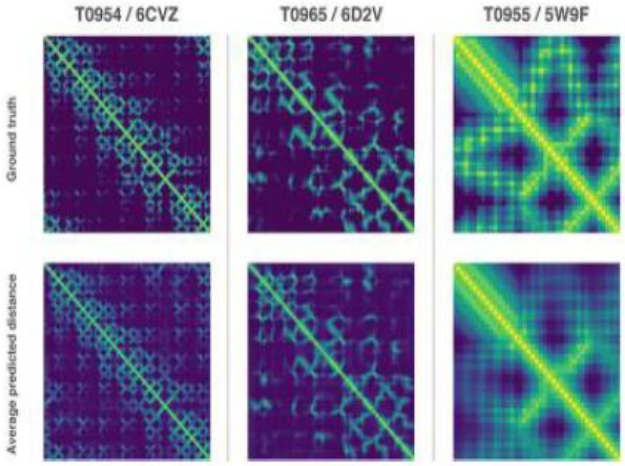
Protein Sequence

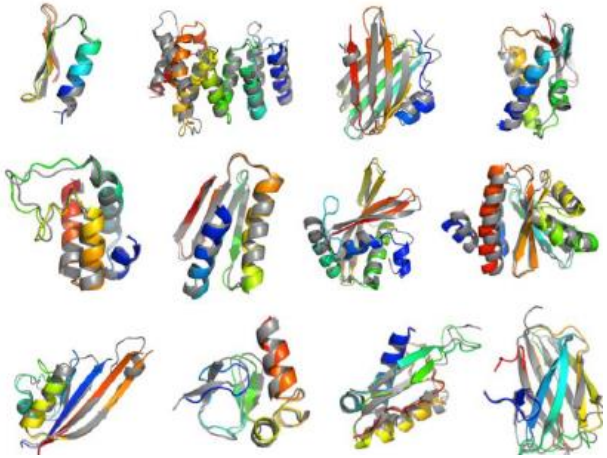
SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC

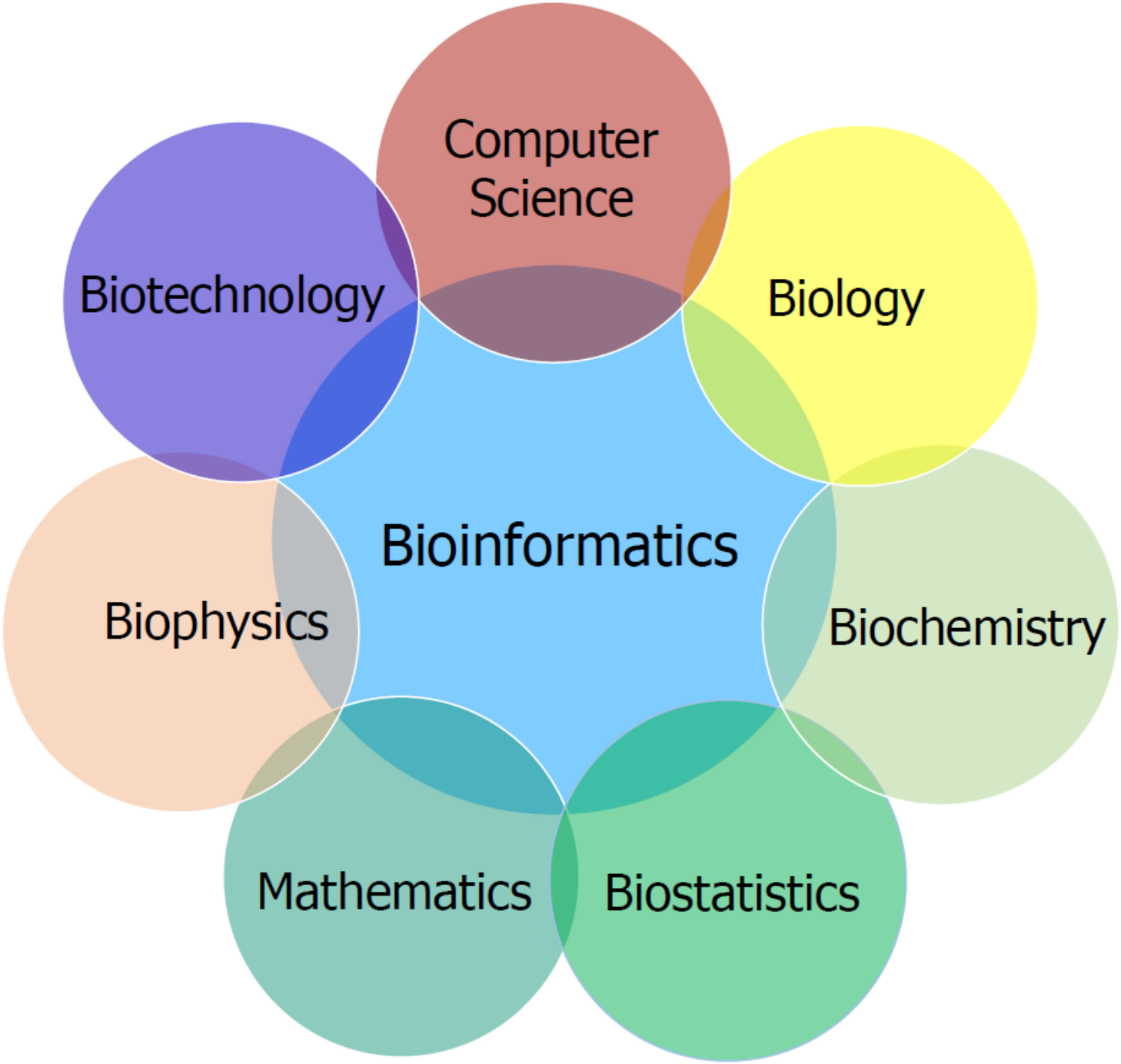
https://deepmind.com/blog/alphafold/

Google DeepMind





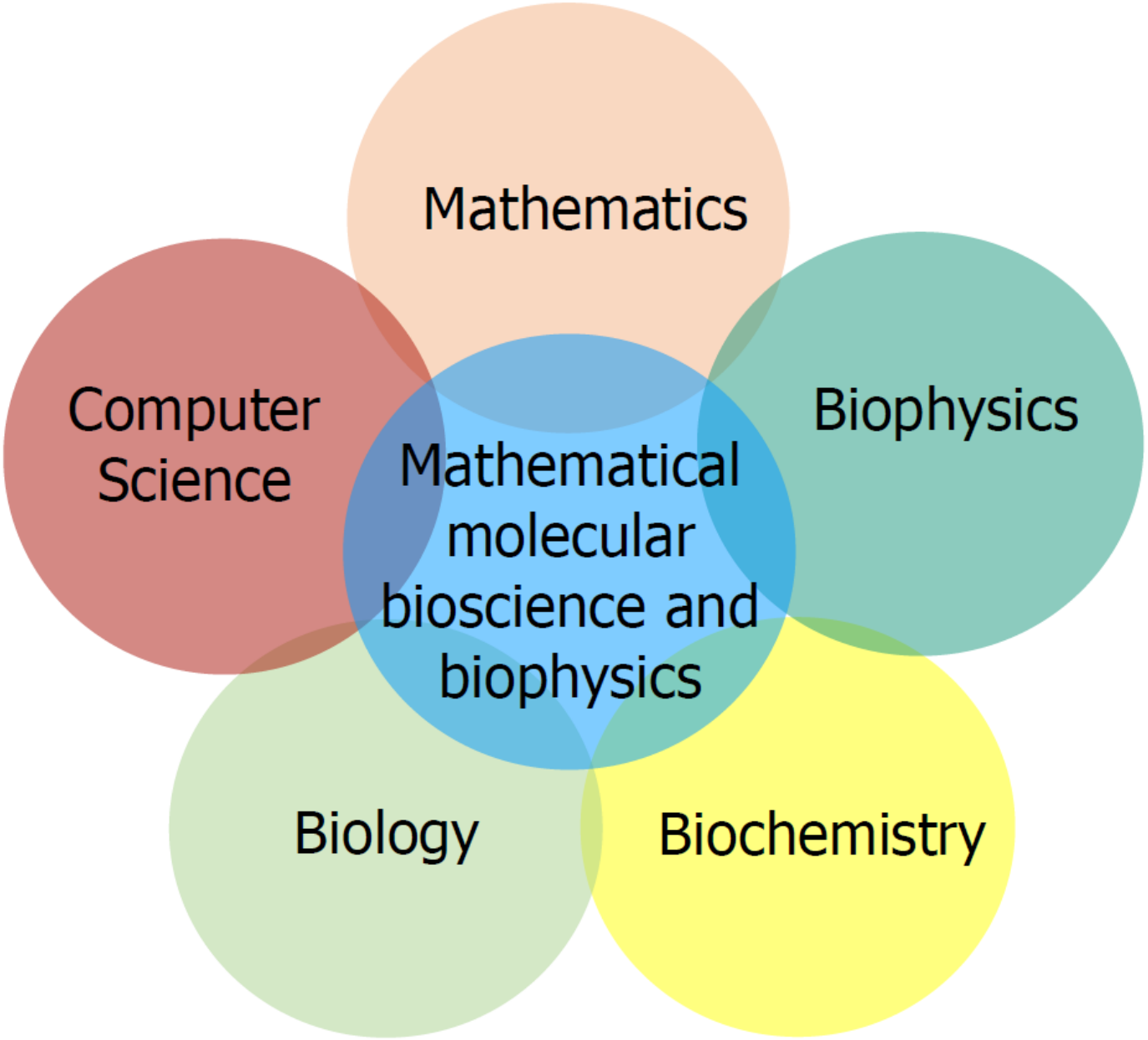



Computer Science
Biotechnology
Biology
Bioinformatics
Biophysics
Biochemistry
Mathematics
Biostatistics

New Trends in Biological Science

1) New bioscience is based on molecules and/or omics.
2) Integrative biology (from molecules, organism to environment).
3) Integration of biophysics, systems biology, and bioinformatics.
4) Integration of mathematics, data science, theoretical biology and experimental biology.
5) Mathematical molecular bioscience and biophysics.
6) Quantitative systems pharmacology (from systems biology, biomechanics, systems physiology to systems pharmacology).
7) Personalized medicine (precision medicine).

Mathematical Molecular Bioscience and Biophysics



Mathematical Molecular Bioscience and Biophysics

1) It concerns the mathematical foundation of biological science.
2) It is based on molecular bioscience and omics in contrast to macroscopic biosciences.
3) It overlaps with molecular biophysics, systems biology and bioinformatics but is distinguished from any mathematical biology that is macroscopic and phenomenological.
4) It exploits existing mathematics for describing biological observations and dynamics.
5) It makes use of computational algorithms and methods from mathematics, machine learning and statistics.
6) It has applications to a wide range of biological problems, including protein design, drug discovery, precision medicine, to mention only a few.
7) It generates new mathematics from biological challenges.

Mathematics Commonly Used in Molecular Bioscience

Geometry Topology	D.W. Sumners, Isabel .K. Darcy , Mariel Vazquez, Dorothy Buck, Tamar Schlick, Erica Flapan, Christian Reidys, Yusu Wang, Peter Rogen, Jack Quine,
Algebra Group theory Combinatorics	Christine Heitsch, David Murrugarra, Reidun Twarock Natasha Jonoska, R Brijder, HJ Hoogeboom, Julie Mitchell
Analysis Calculus/Variation ODE and PDE Numerical analysis	M Karplus, M Levitt, A Warshel, B Honig, E Alxov, A Onufriev,… B.S. Eisenberg, Chun Liu, Weishi Liu, Yun Kyong Hyon, TC Lin, JL Liu, TL Horng, YN Young, HX Huang, Lei Zhang, Tom Chou J.A. McCammon, Michael Holst, Jingfang Huang, Benzhuo Lu, Nathan Baker, Bo Li, LT Cheng, MX Chen, Shenggao Zhou, Keith Promislow, Shibin Dai, Nir Gavish, Robert Krasny, DX Xie, LR Scott, Wei Cai, ZL Xu, Amit Singer, D. Kozakov, R Rizzo, D. Green, R Ryham, LJ Cowen, …