

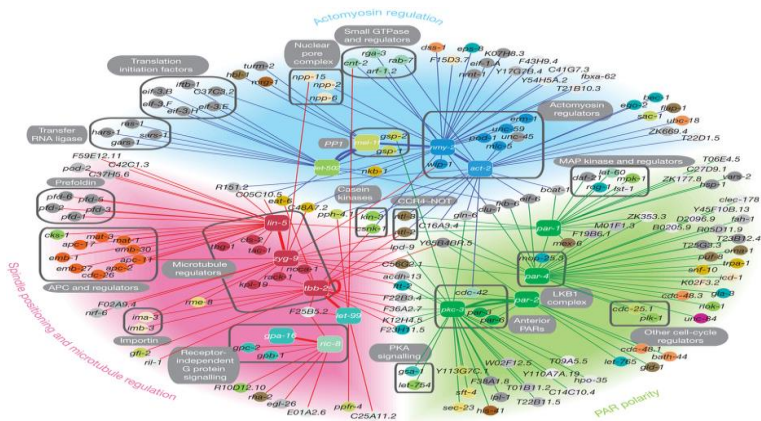
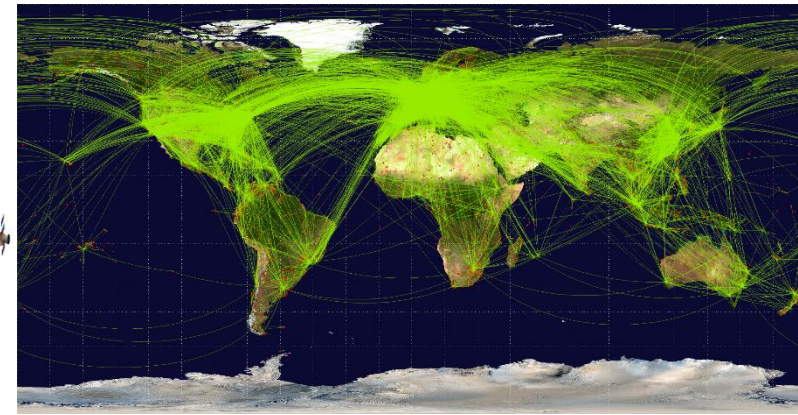
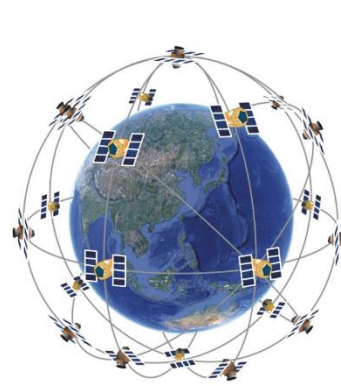
Topological data analysis (TDA) based machine learning models for biomolecular data analysis

Kelin Xia

***School of Physical and Mathematical Sciences,
School of Biological Sciences,
Nanyang Technological University***

***Winter school on topological data analysis at
Novosibirsk State University, Feb 3-8, 2020***

**Fund: ASPIRE, NTU-JSPS, MOE-Tier 1(2017,2018,2019), MOE-
Tier 2(2018)**



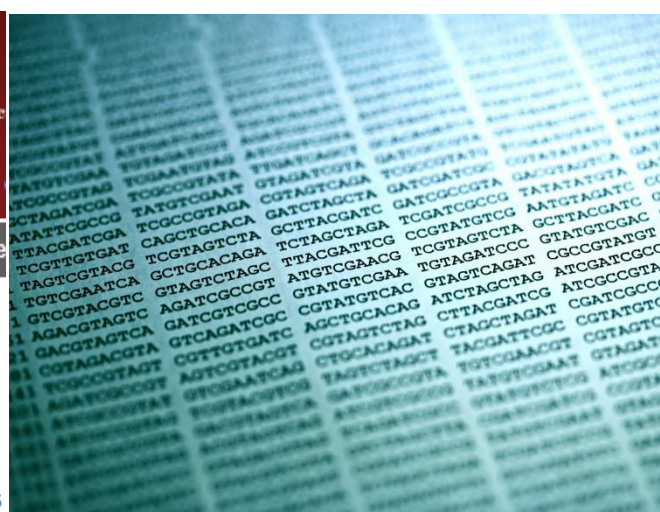
RCSB PDB An Information Portal to 133759 Biological Macromolecular Structures
PROTEIN DATA BANK

Gene Expression Omnibus

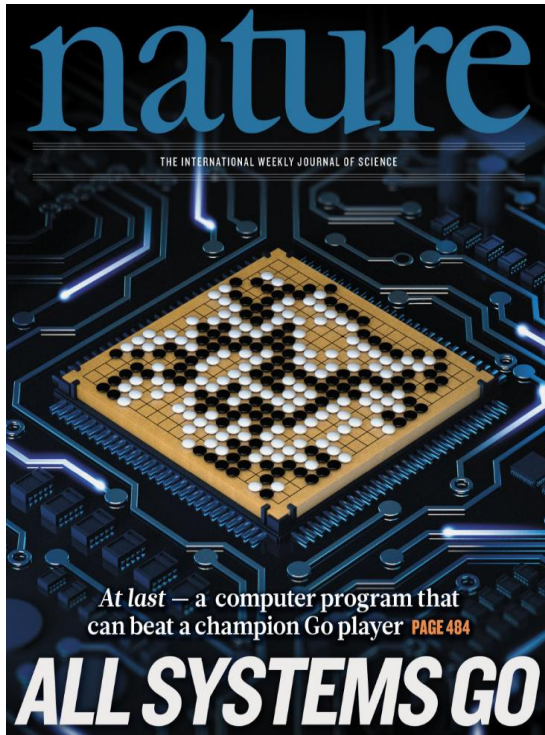
GenBank Celebrates 25 years of service

nature International weekly journal of science
Home | News & Comment | Research | Careers & Jobs |
Archive | Volume 487 | Issue 7407 | News | Article

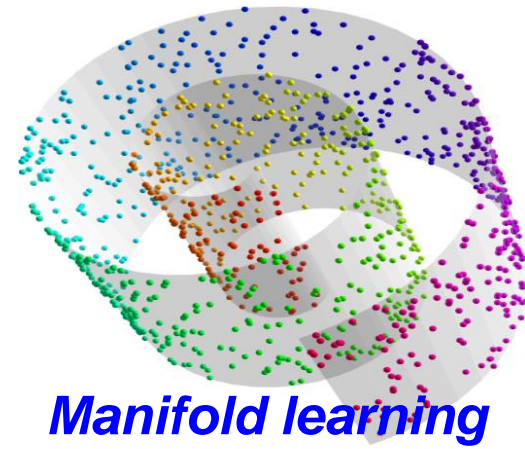
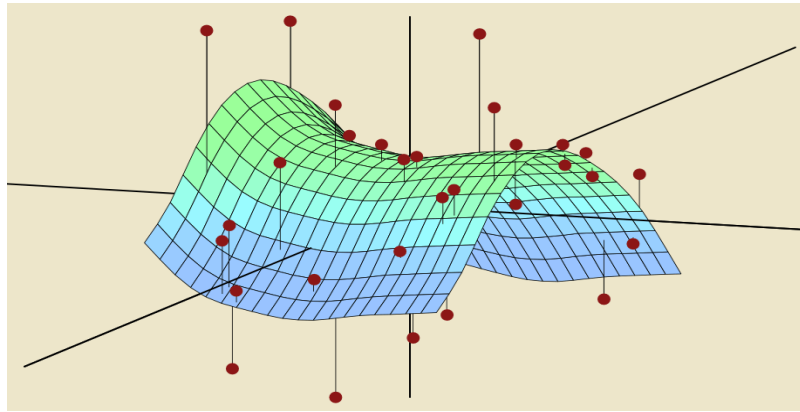
NATURE | NEWS
Gene data to hit milestone
With close to one million gene-expression data sets



Deep learning

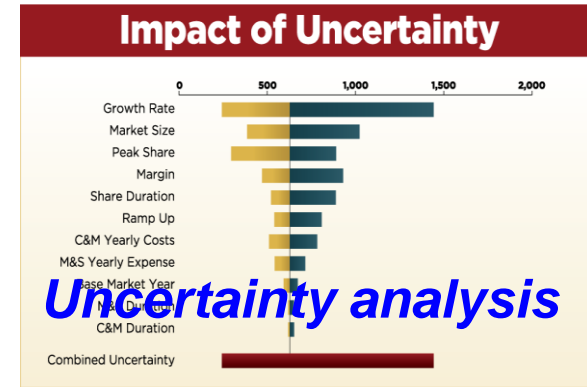


Statistic learning

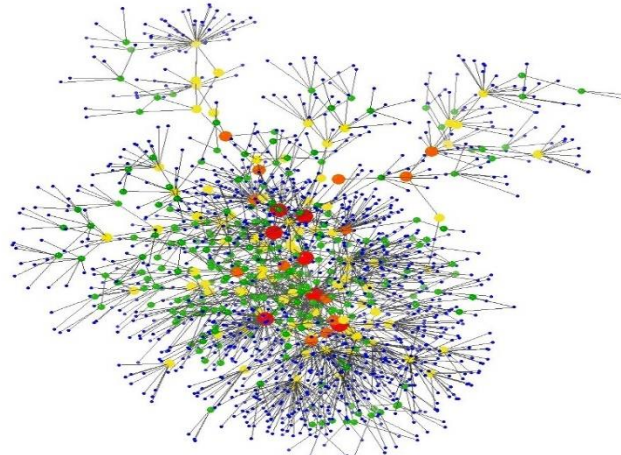
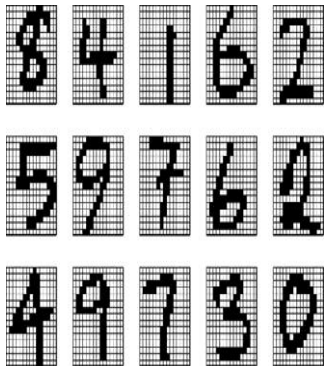


Manifold learning

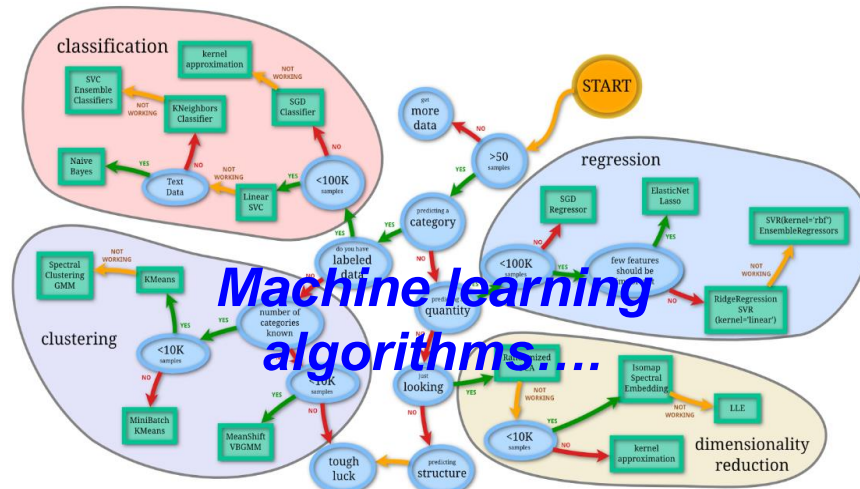
How to deal with big data?



Pattern recognition



Graph modeling and analysis



Representation and Feature Learning

“The success of machine learning algorithms generally depends on data representation...”

Y. Bengio, etc, “Representation Learning: A Review and New Perspectives”

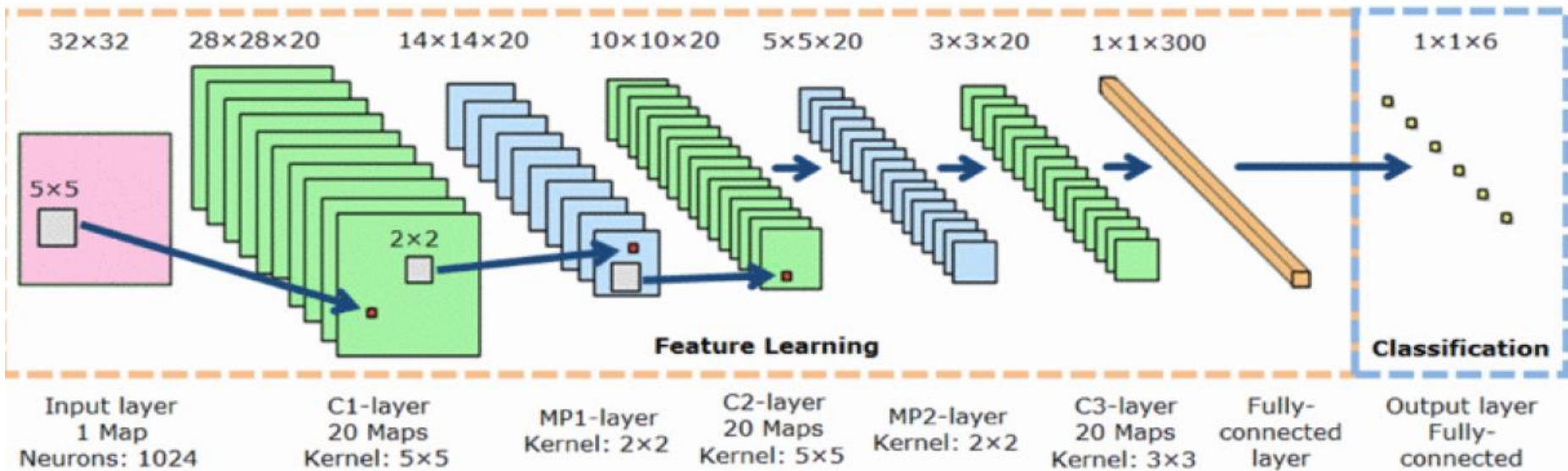
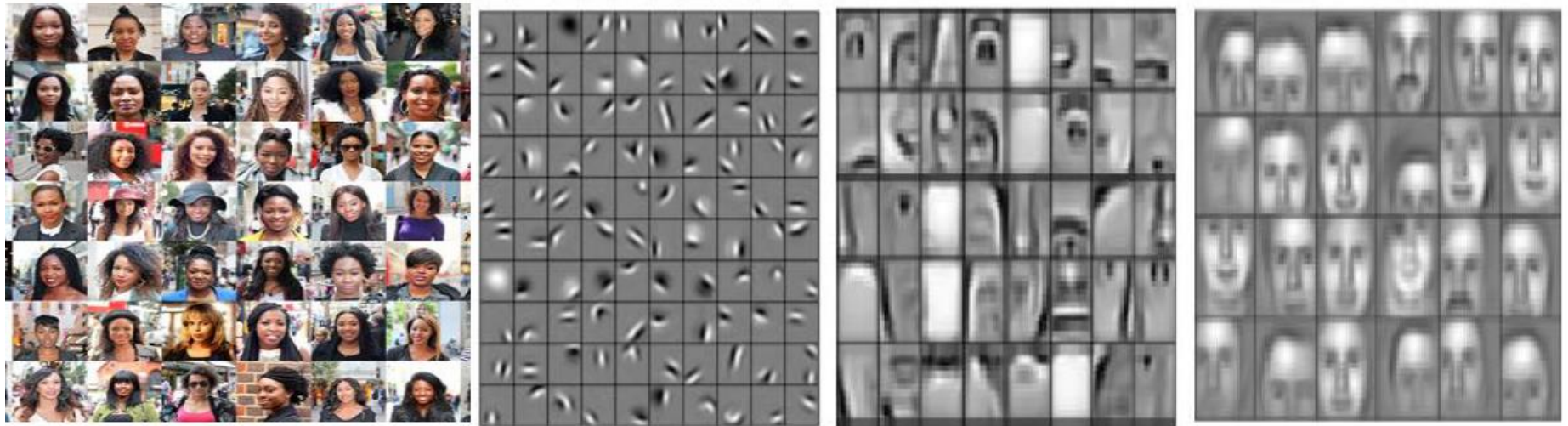


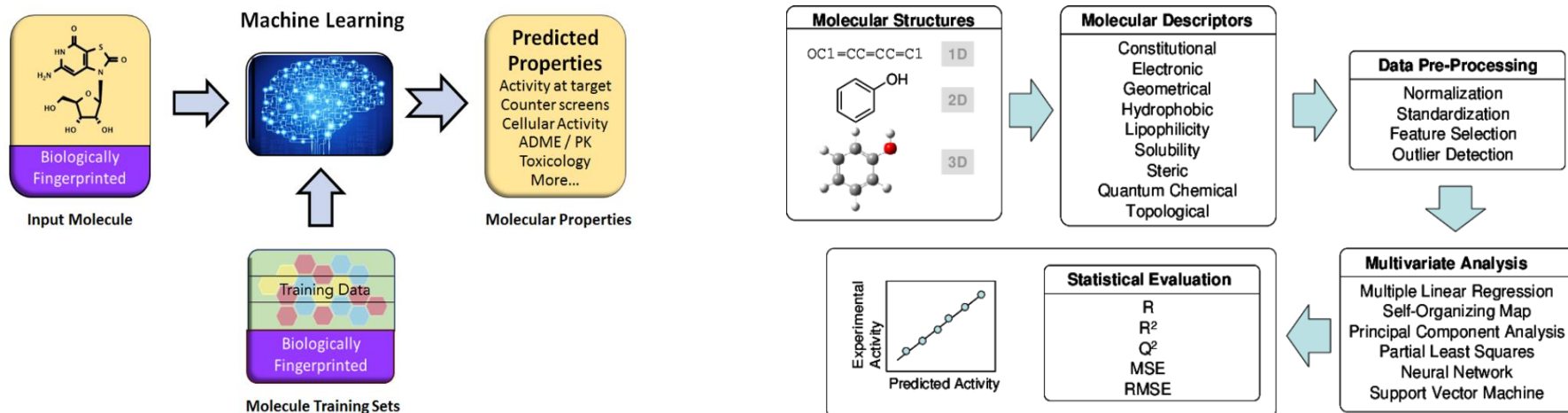
“The deep learning research aims at discovering learning algorithms that discover multiple levels of distributed representations...”

Y. Bengio, “Deep Learning of Representations: Looking Forward”

Feature learning is key to data analysis!

Fukushima (1980) – Neo-Cognitron; LeCun (1998) – Convolutional Neural Networks (CNN);...





Molecular descriptors (>5000) directly determine the performance of learning models!

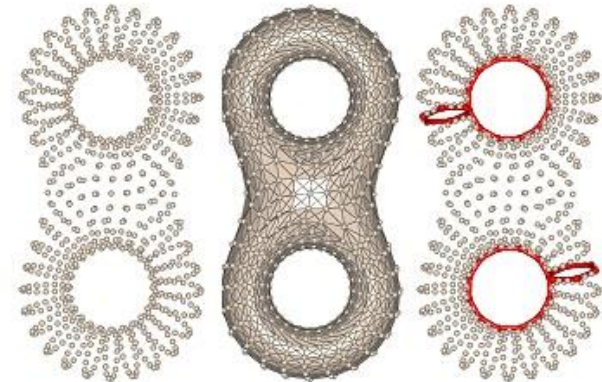
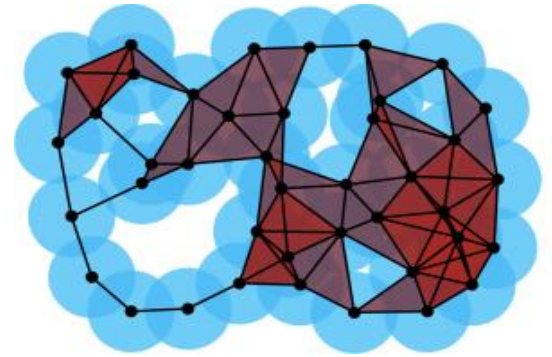
Common chemical descriptors for QSAR/QSPR analysis

Chemical descriptors	Based on	Examples
Theoretical descriptors		
0D	Molecular formula	Molecular weights, atom counts, bond counts
1D	Chemical graph	Fragment counts, functional group counts
2D	Structural topology	Weiner index, Balaban index, Randic index, BCUTS
3D	Structural geometry	WHIM, autocorrelation, 3D-MORSE, GETAWAY
4D	Chemical conformation	Volsurf, GRID, Raptor
Experimental descriptors		
Hydrophobic parameters	Hydrophobicity	Partition coefficients (logP), hydrophobic substituent constant (π)
Electronic parameters	Electronic properties	Acid dissociation constant, Hammett constant
Steric parameters	Steric properties	Taft steric constant, Charton's constant

Topological Data Analysis (TDA)

**Topological invariant;
Homology;
Homotopy;
Simplicial complex;
Morse theory;
Reeb graph;**

.....

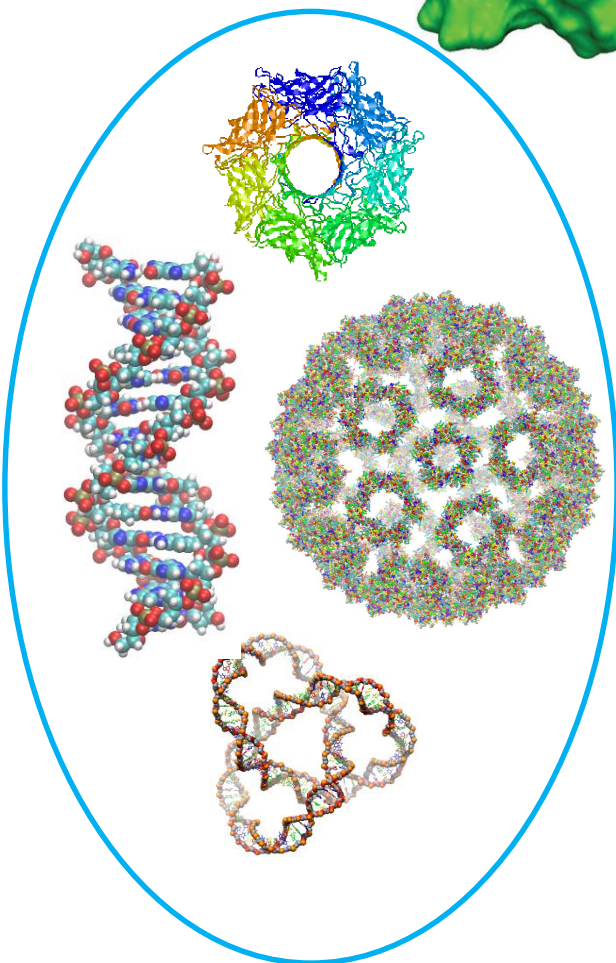
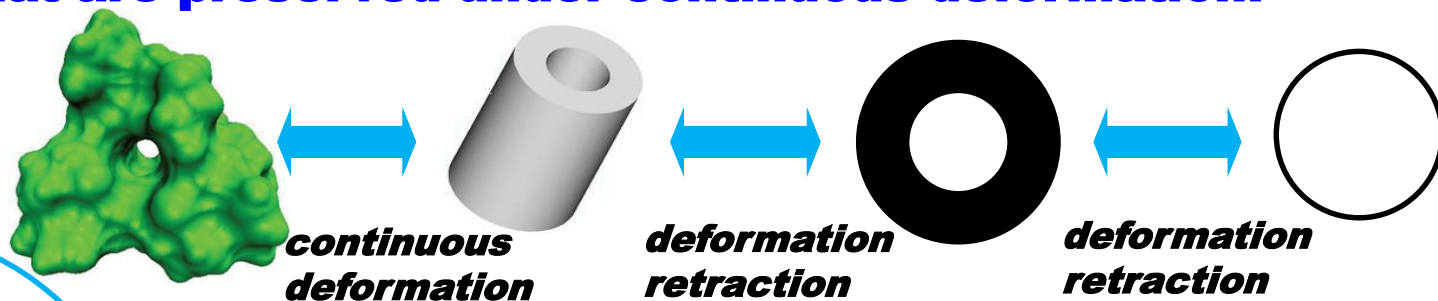


**Computational Geometry;
Computational topology;
Algebraic topology**

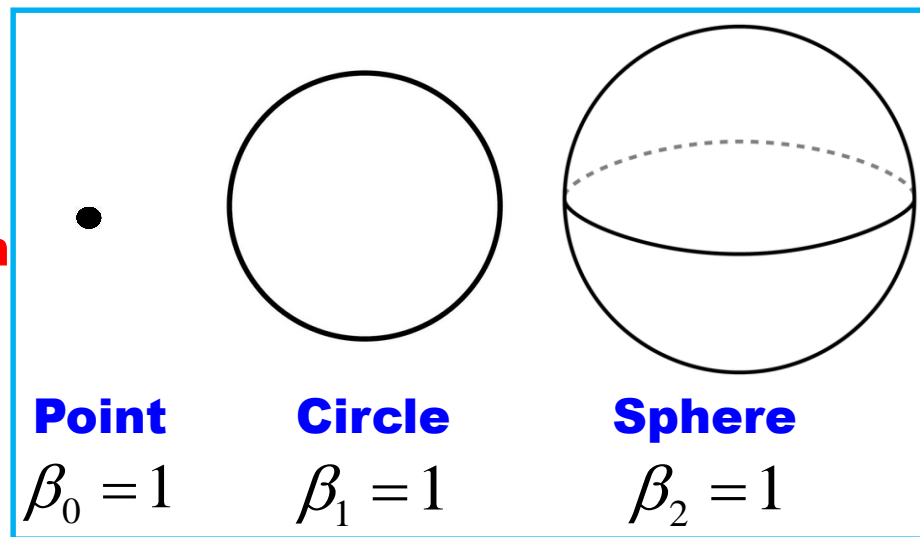
Topological invariant--Betti number

Properties that are preserved under continuous deformation!

Homotopy equivalent:



Topological simplification

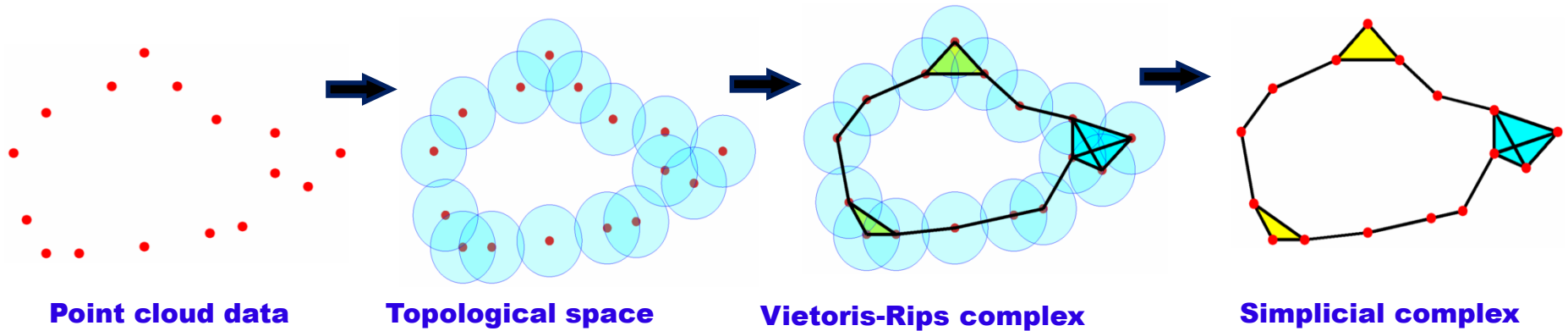


β_0 is the number of **connected components**

β_1 is the number of **tunnels or circles**

β_2 is the number of **voids or cavities**

Topological data analysis



Chain group: $C_k(K, \mathbb{Z}_2)$

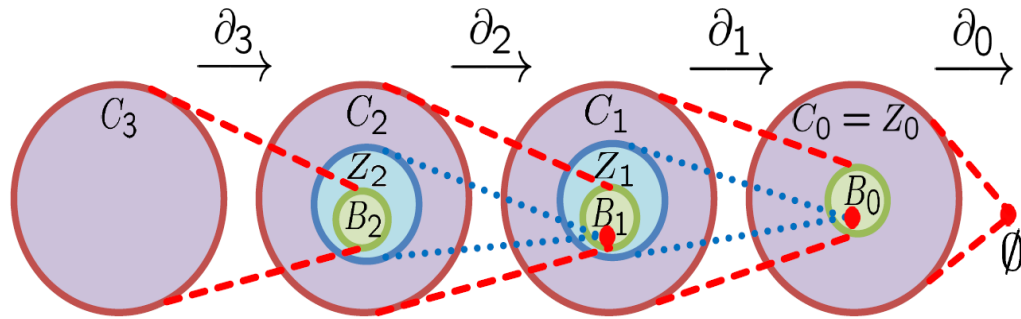
Boundary operator: $\partial_k \sigma^k = \sum_{i=0}^k (-1)^i \{v_0, v_1, \dots, \hat{v}_i, \dots, v_k\}$

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

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

Quotient group:

$$H_k = Z_k / B_k$$



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

The topological information can be calculated!!

Opportunities, **challenges** and **promises**

Opportunities from topological methods:

- ❖ **New approach for big data characterization and classification.**
- ❖ **Dramatic reduction of dimensionality and data size.**
- ❖ **Applicable to a variety of fields.**

Challenges with topological methods:

- **Geometric methods are inundated with structural details.**
- **Topology incurs too much reduction of original information.**
- **Topology is hardly used for quantitative prediction.**

Promises from persistent homology:

- ✓ **Embeds geometric information in topological invariants.**
- ✓ **Bridges the gap between geometry and topology.**

Researchers:

**Frosini (1991),
Robins (2000),
Edelsbrunner, Letscher and Zomorodian (2002),
Kaczynski, Mischaikow and Mrozek (2004),
Zomorodian and Carlsson (2005),
Ghrist (2008),
Dey and Wang(2009),**

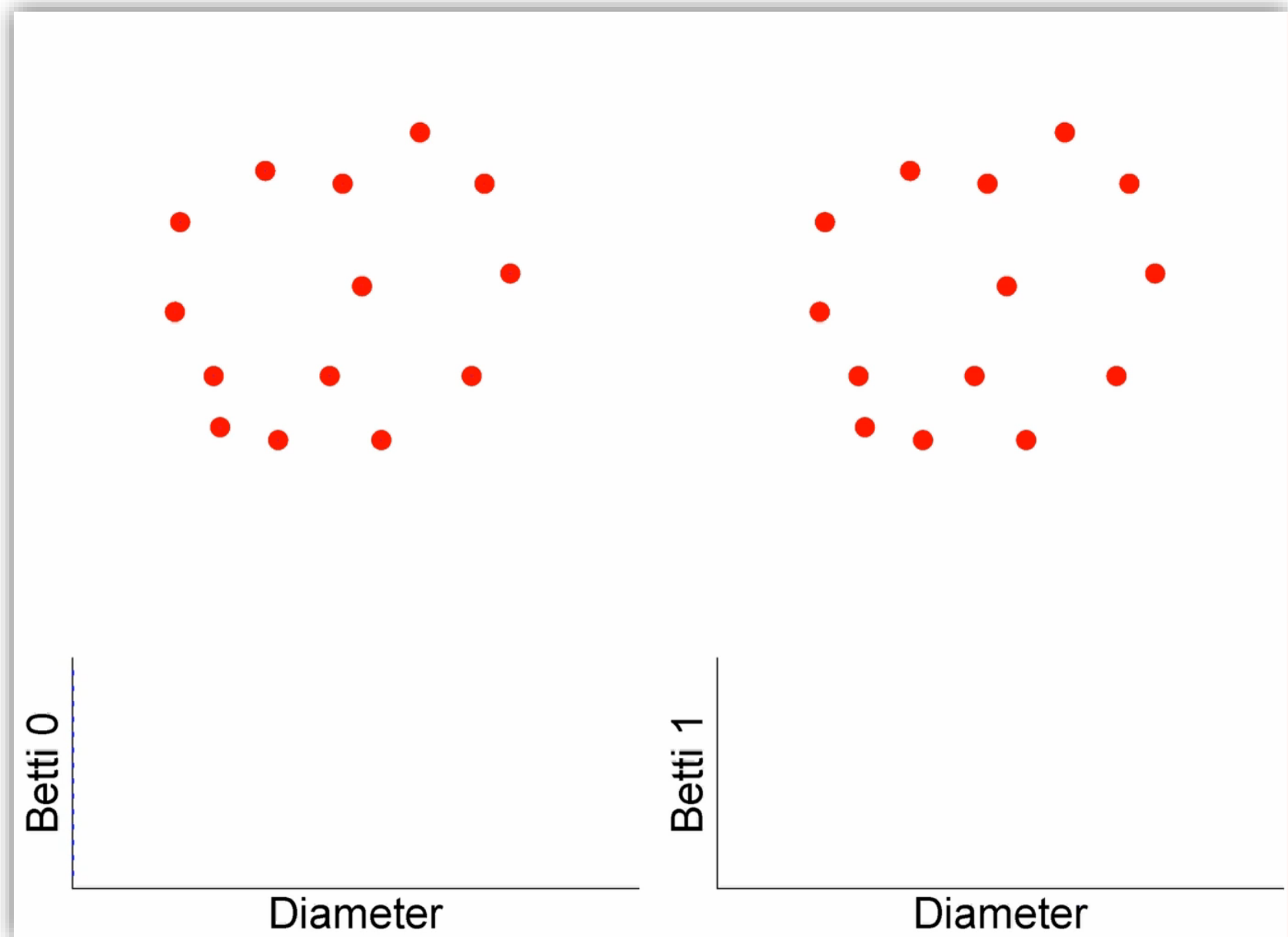
.....

Softwares:

**Javaplex,
Perseus,
Dipha,
Dionysus,**

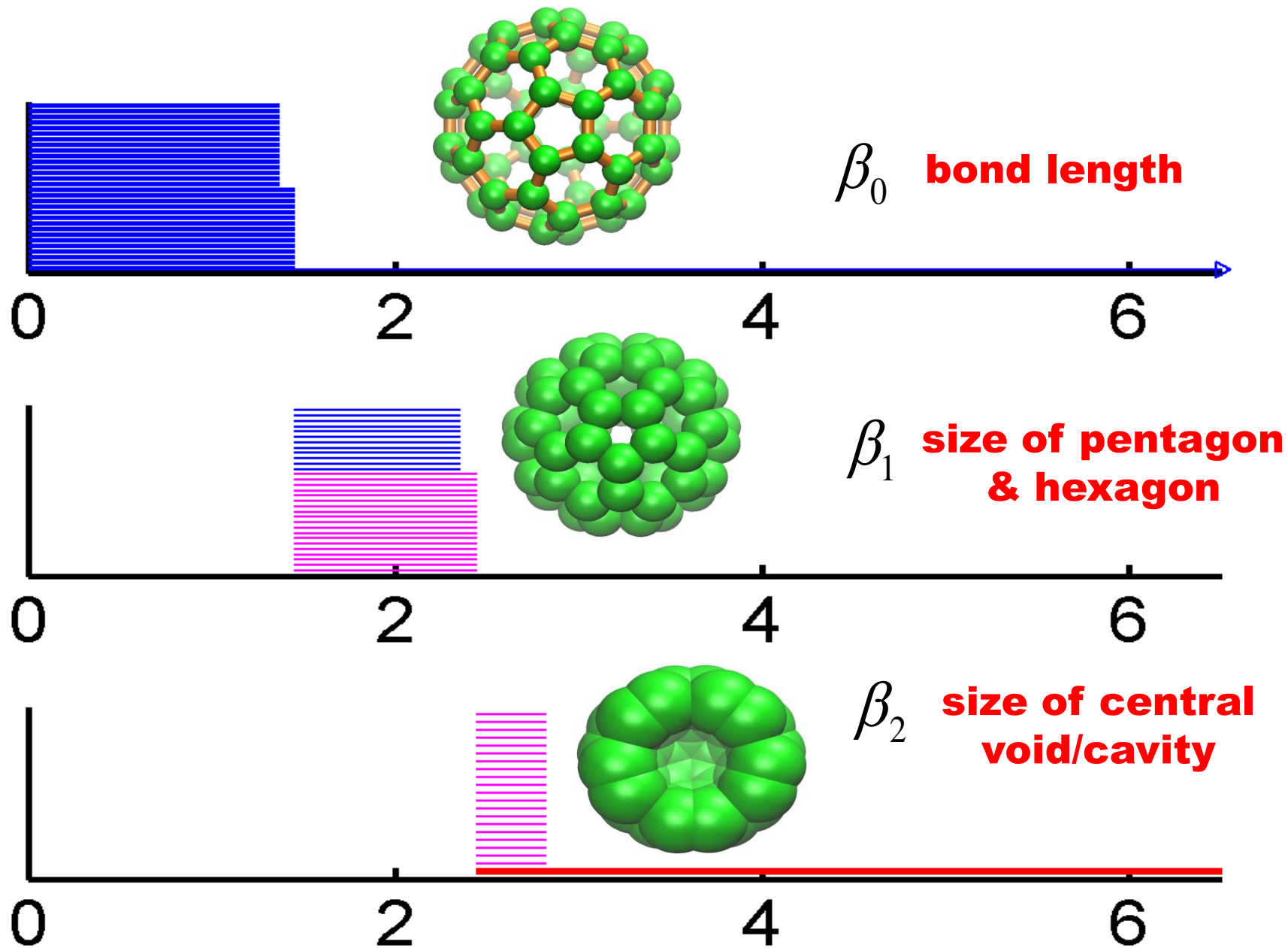
.....

Persistent homology and persistent barcodes

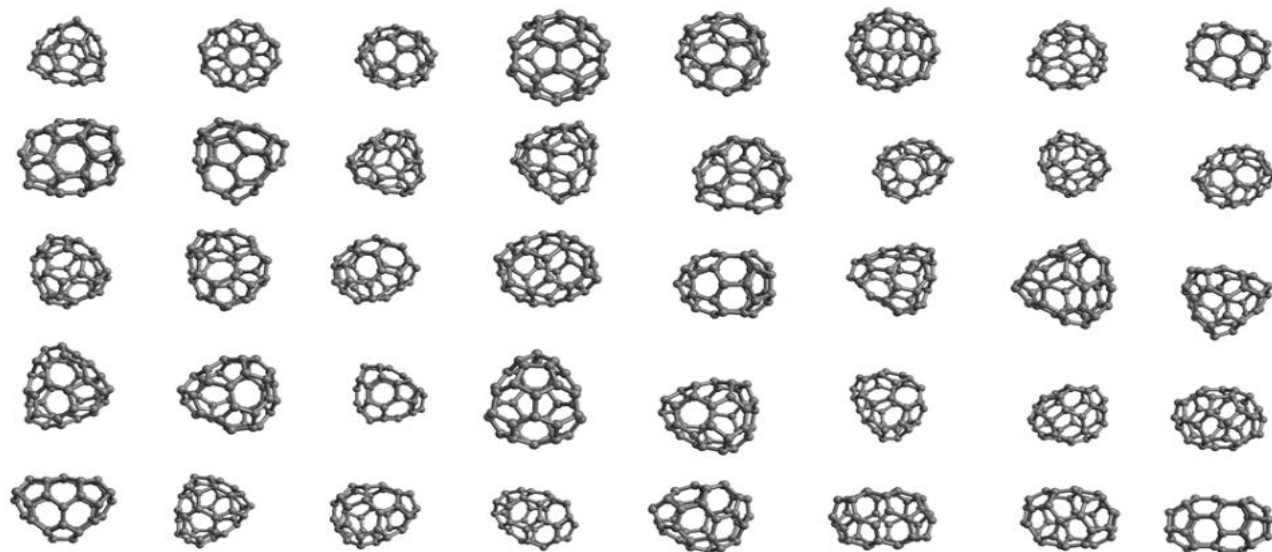


PHA of fullerene C₆₀

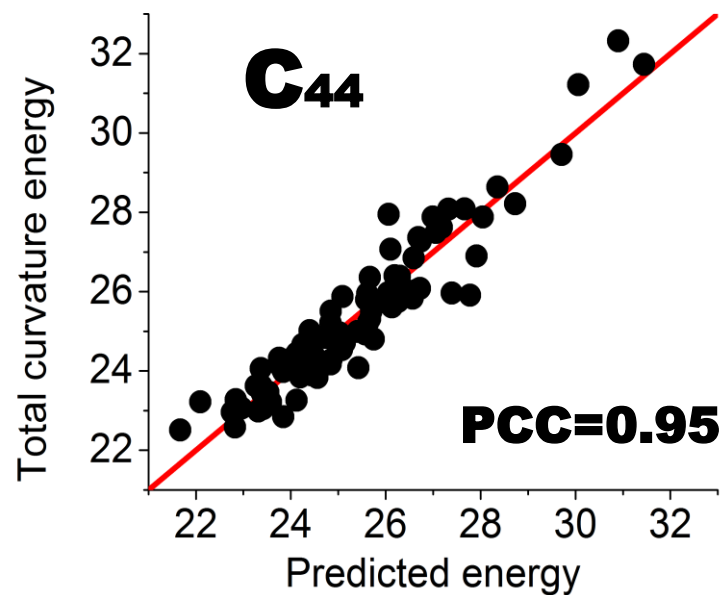
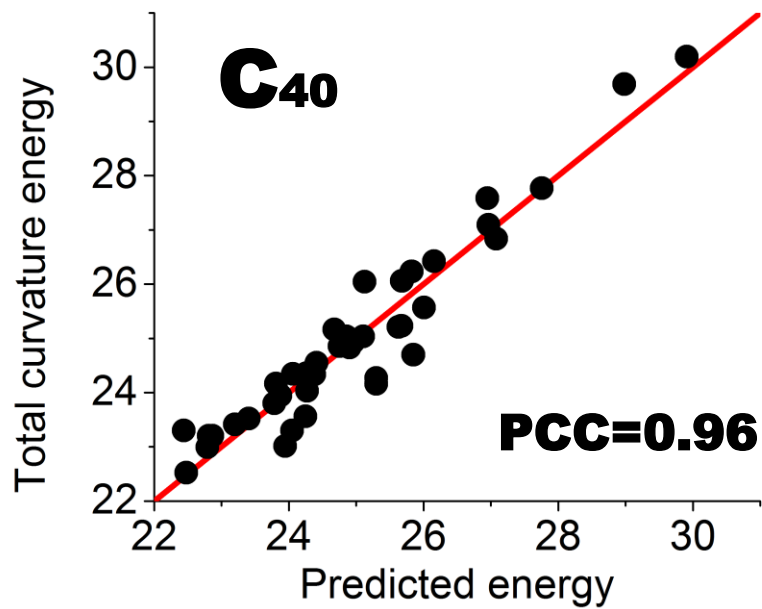
(Xia, Feng, Tong & Wei, JCC, 2015)



Fullerene isomers



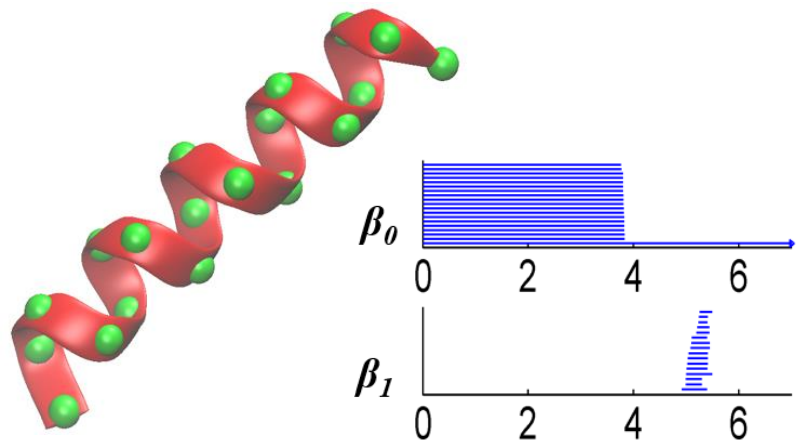
$$E \propto 1/L(\beta_2)$$



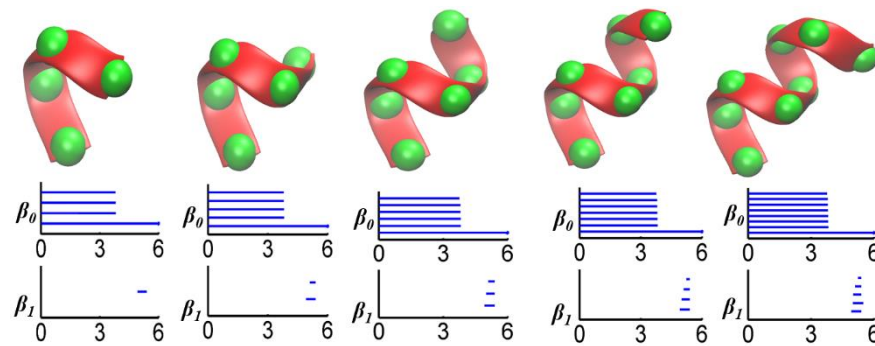
(Xia, Feng, Tong & Wei, JCC, 2015)

Biomolecular topological fingerprint (TF)

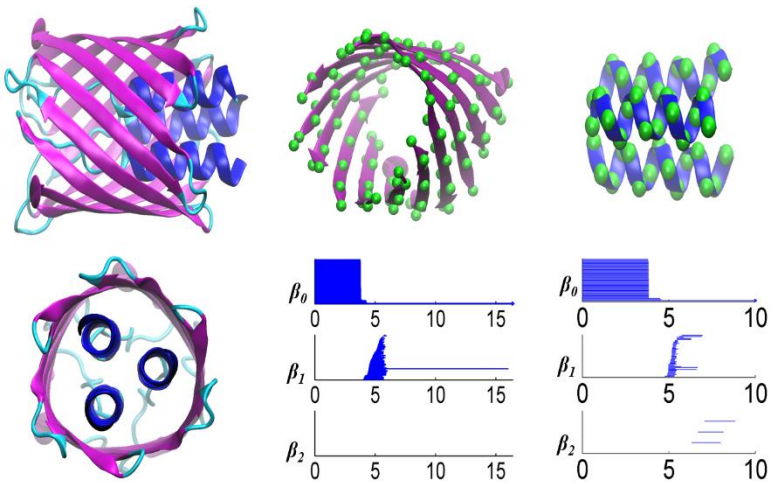
TF for alpha helix



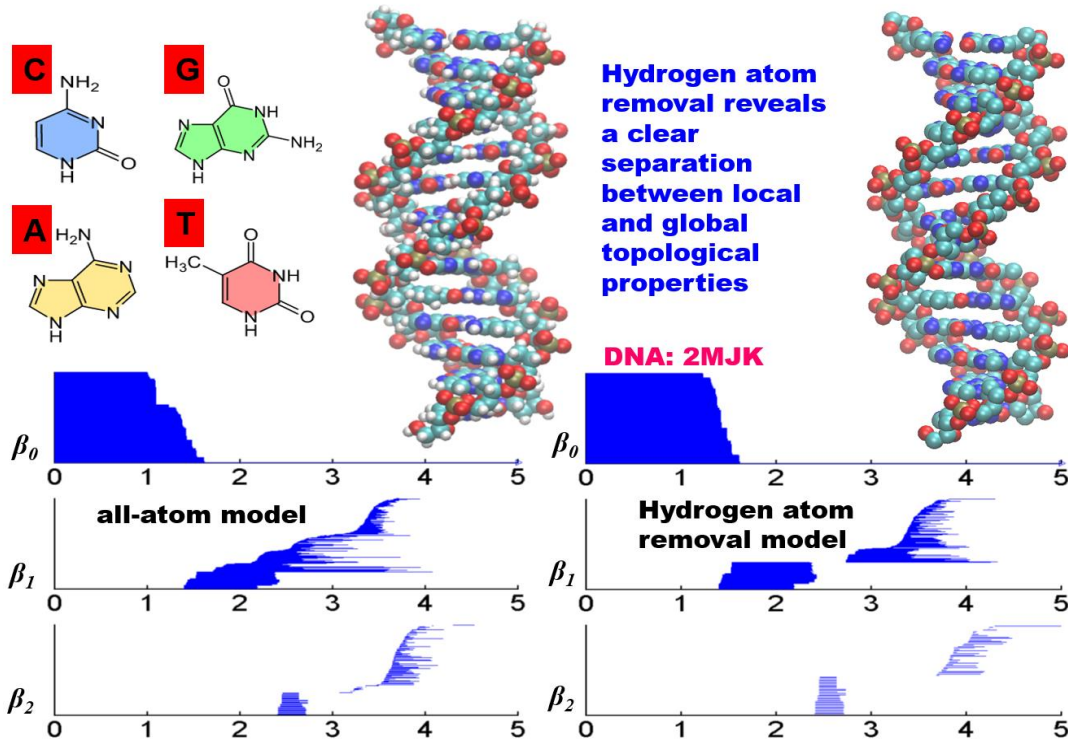
Slicing method



TF for beta barrel

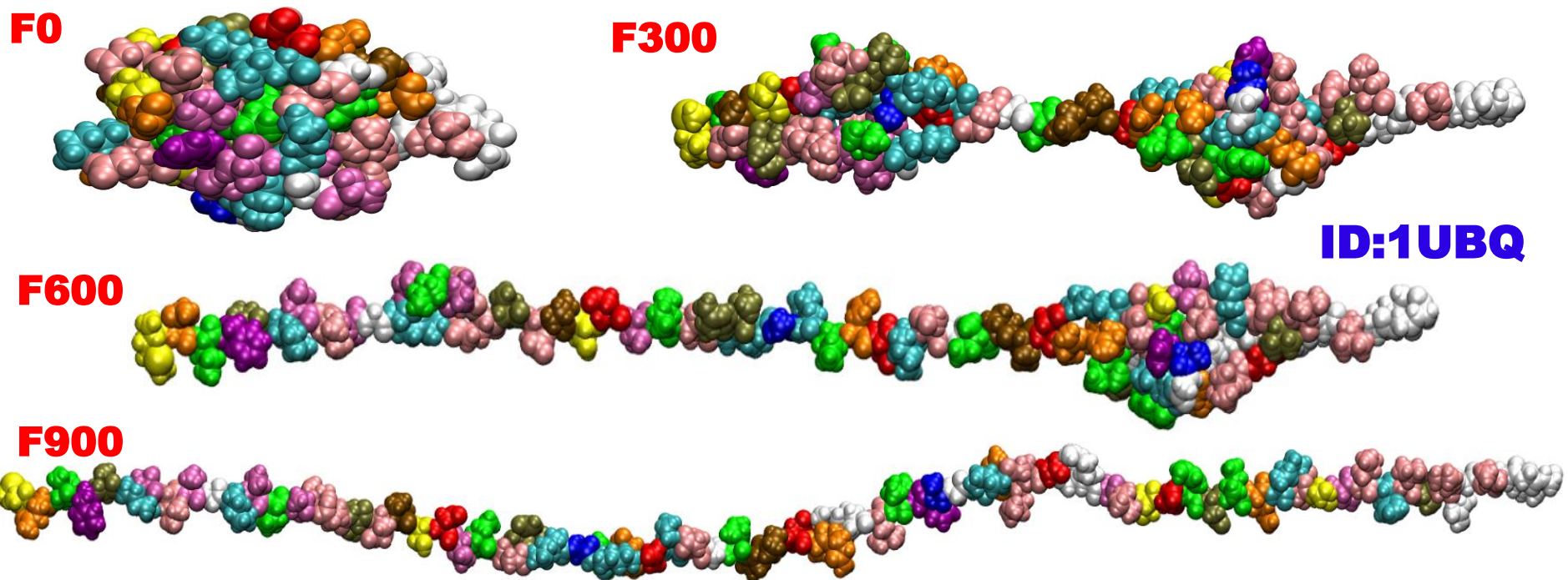
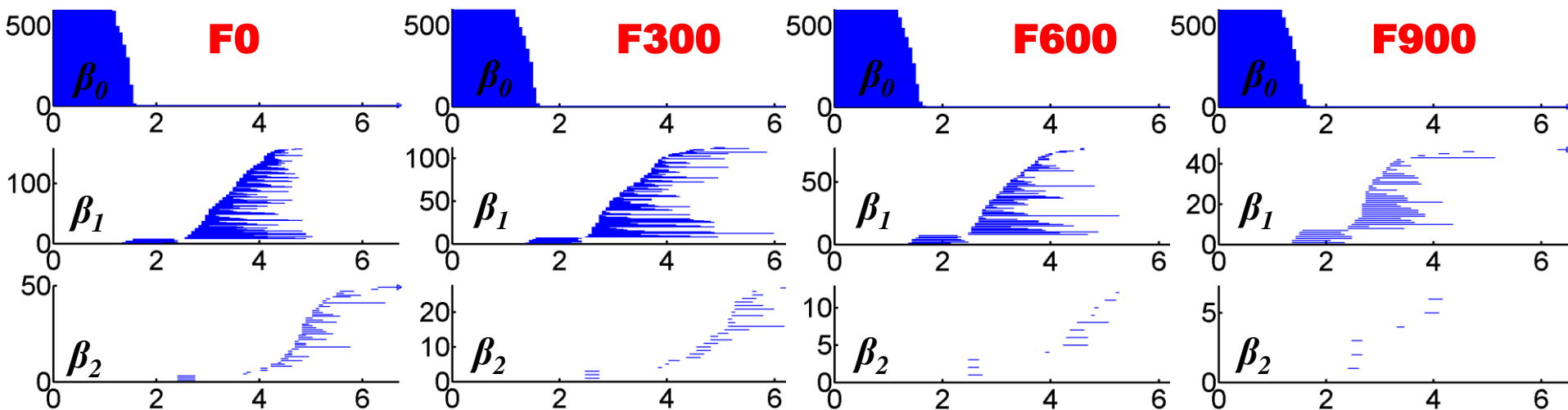


Topological fingerprints of DNA

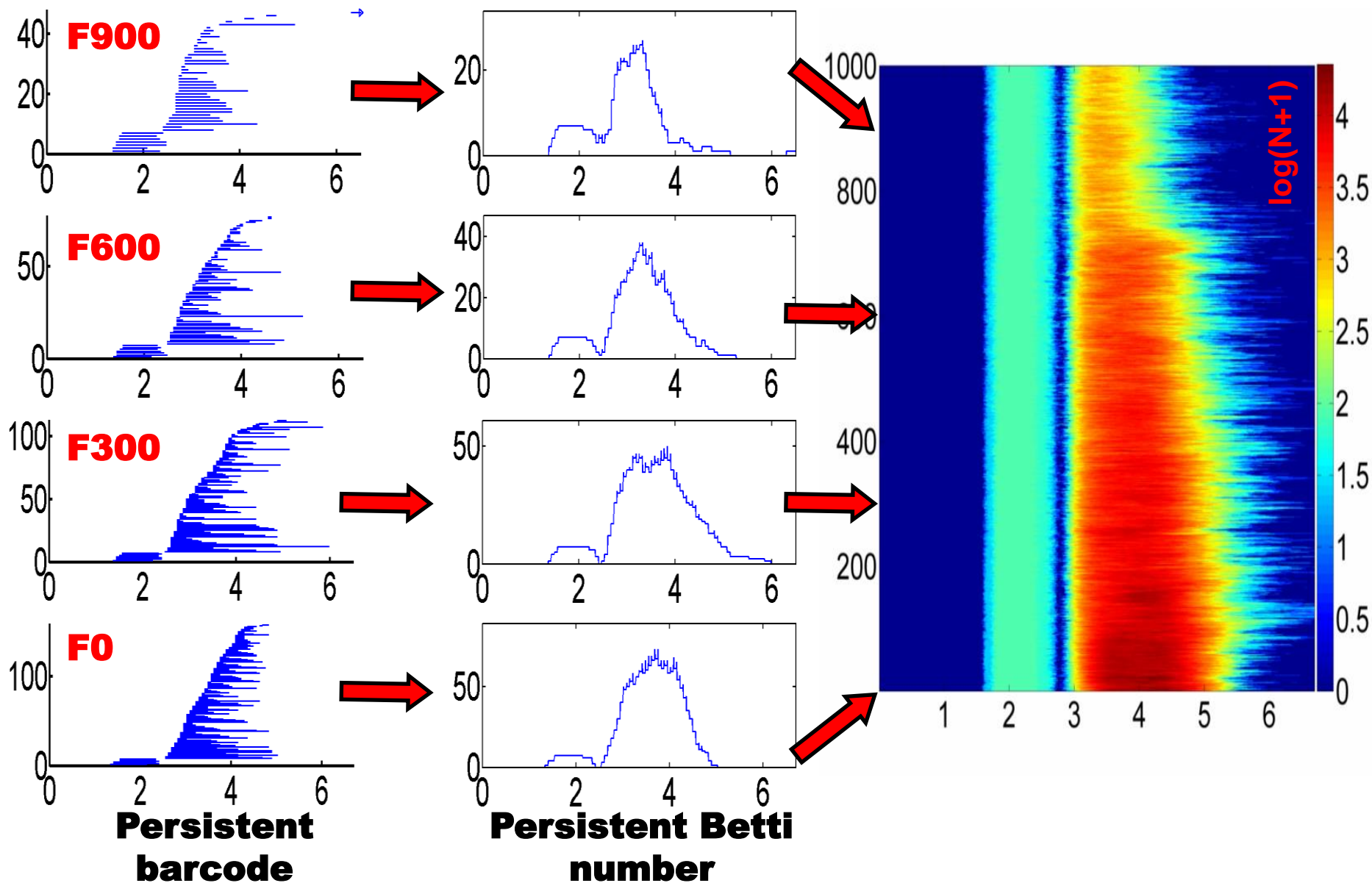


Topic--Topological analysis of protein folding

(Xia & Wei, JCC, 2015)

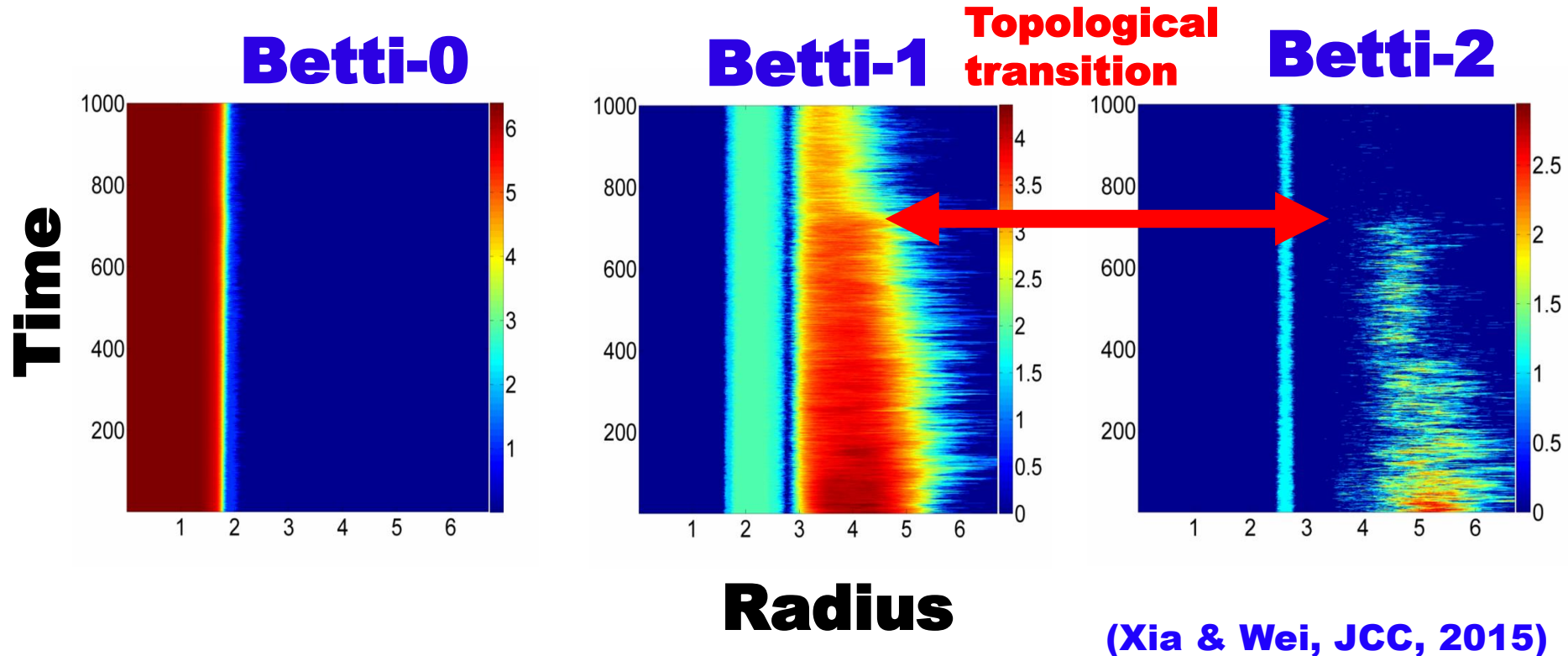
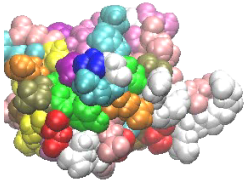


The protein folding process can be characterized by a proposed two dimensional filtration representation



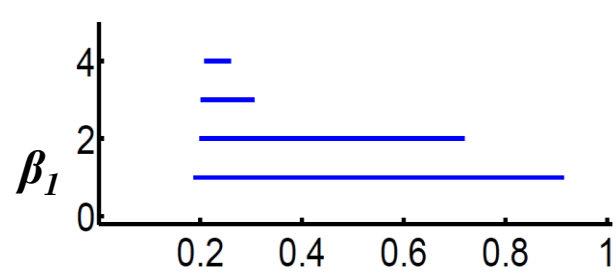
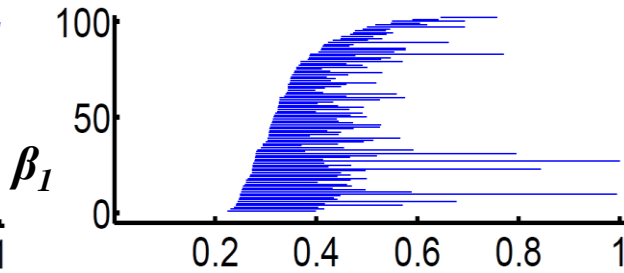
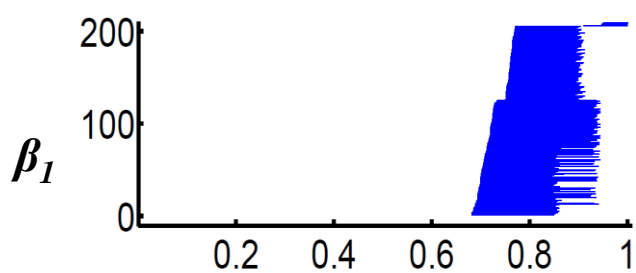
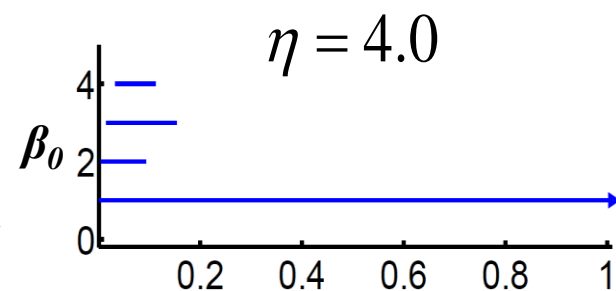
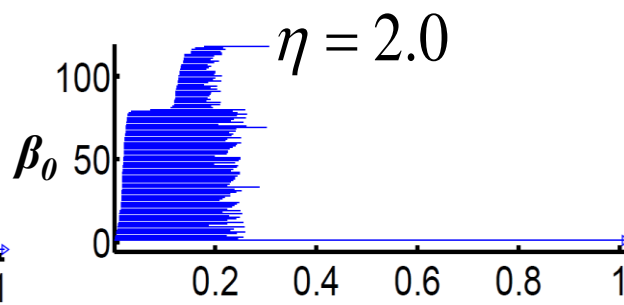
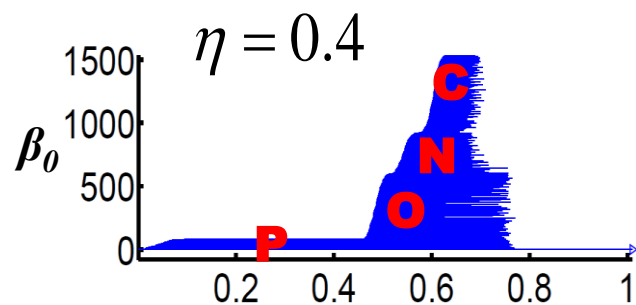
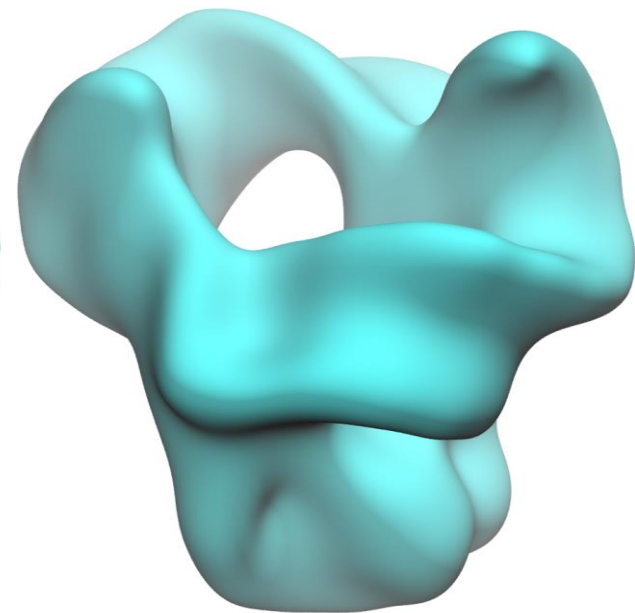
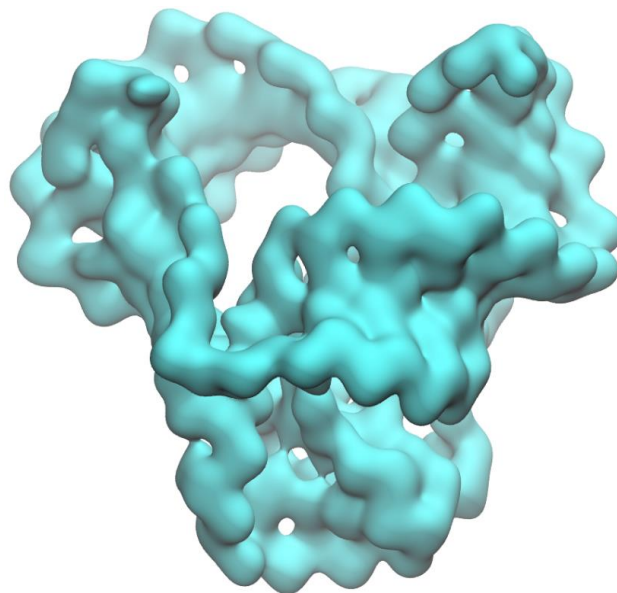
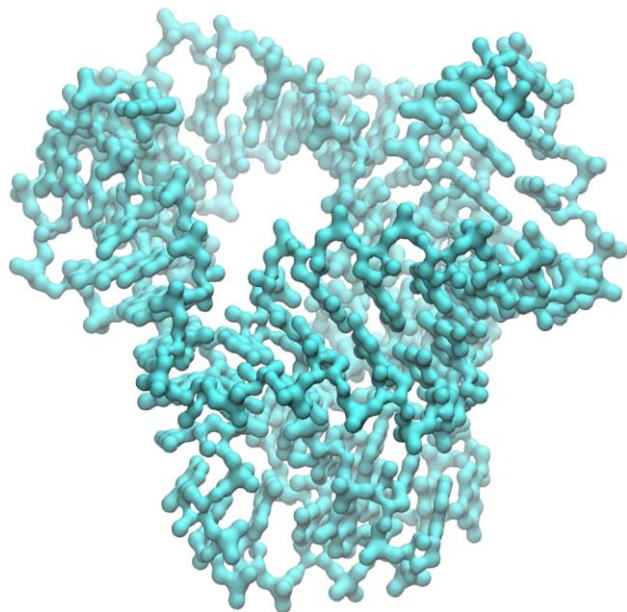
2D persistence for protein unfolding

ID:1UBQ

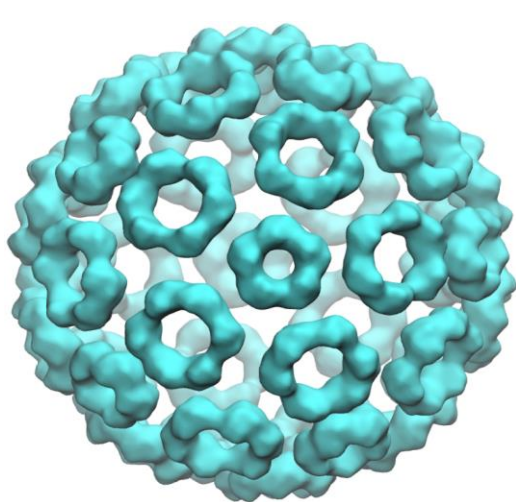


PHA for multiresolution representations

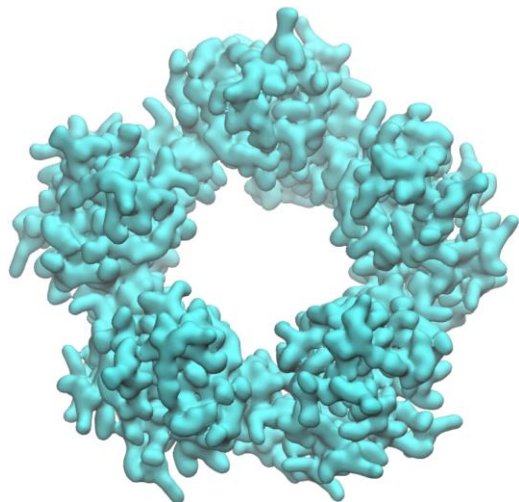
(Xia & Wei, JCB, 2015)



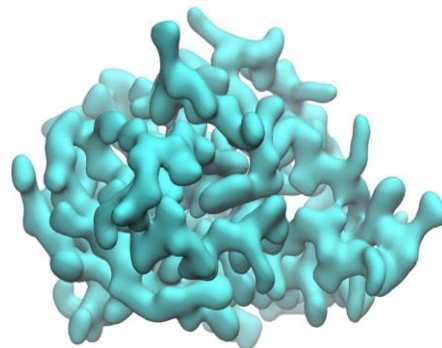
Multiresolution of the virus capsid



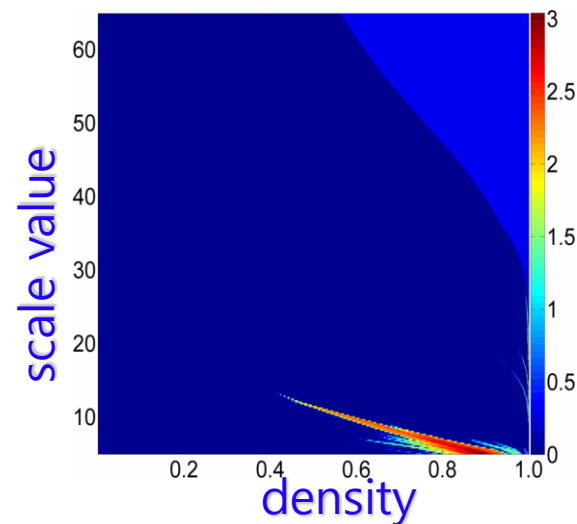
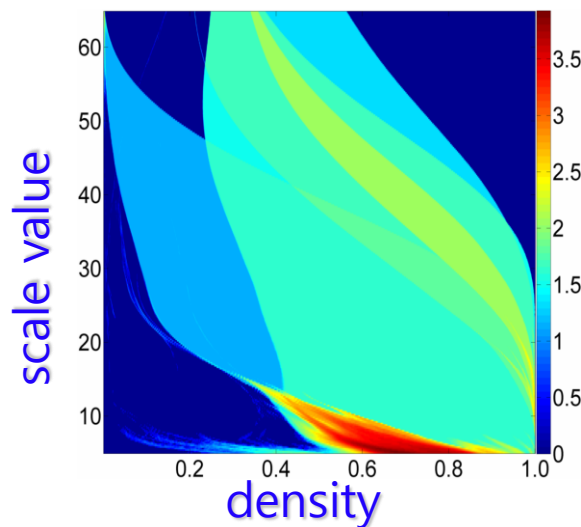
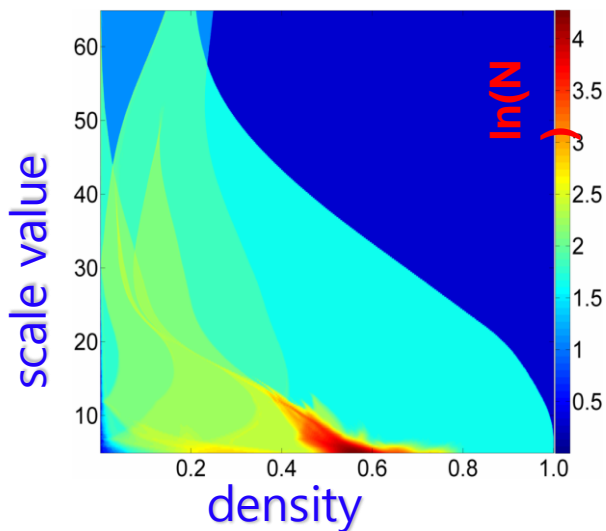
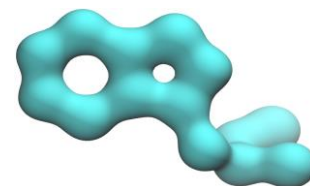
Betti-0



Betti-1



Betti-2

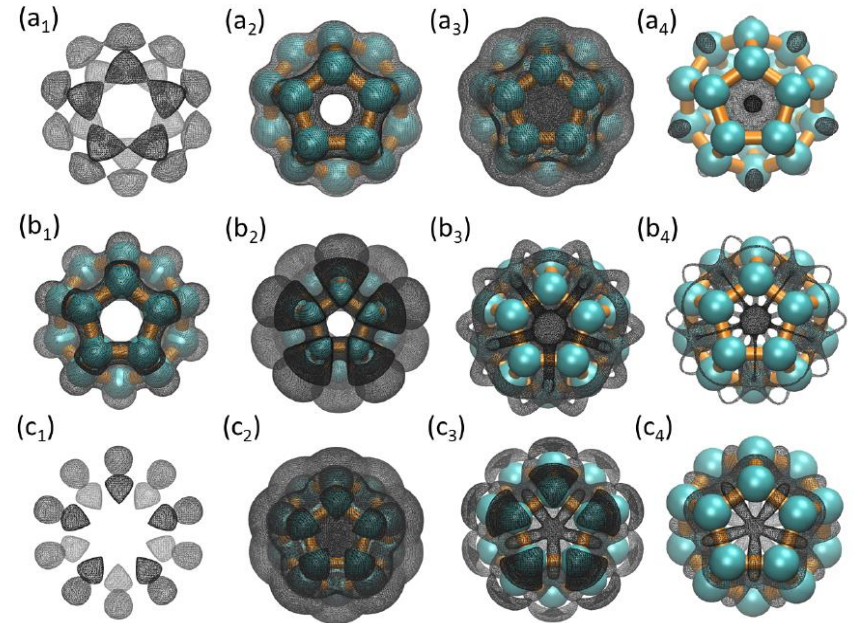
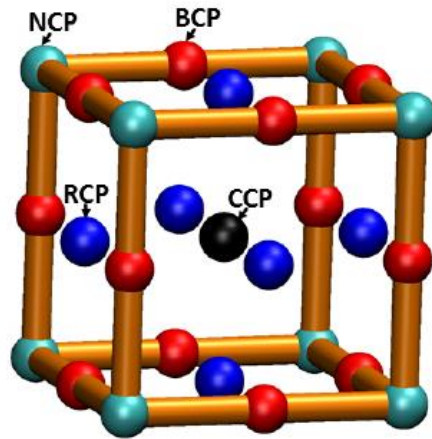
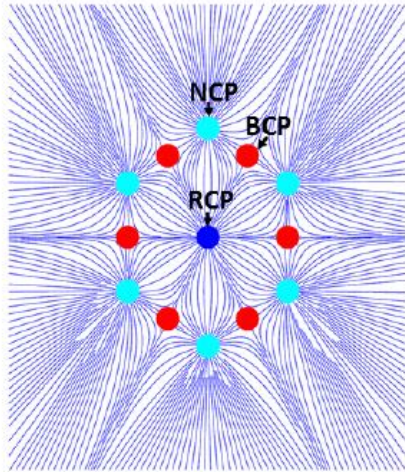
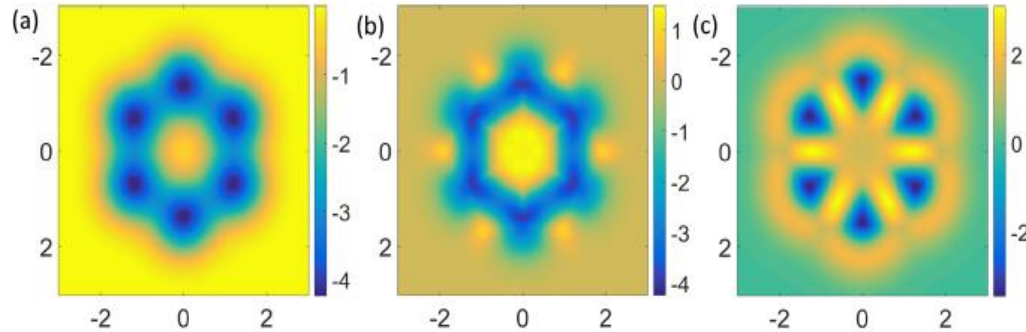


Topic—Atom in molecule

Atoms in Molecules

Richard F. W. Bader

McMaster University, Hamilton, Ontario, Canada



(Xia & Wei, arXiv, 2017)

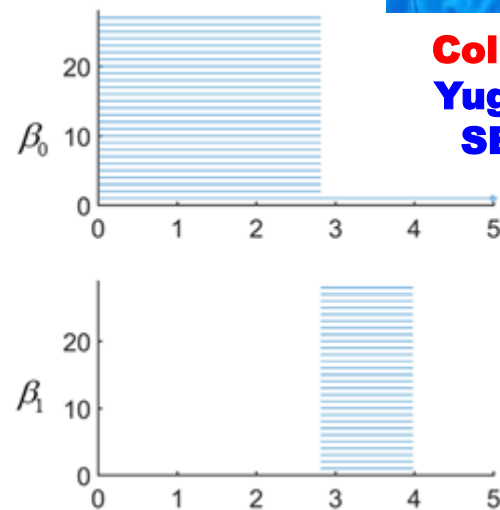
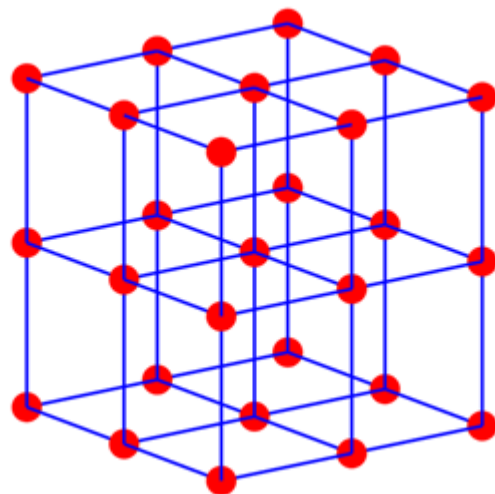
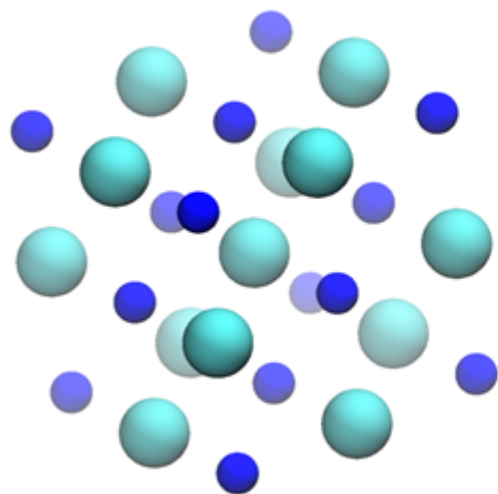
Figs: An illustration of second and third eigenvalue, mean curvature isosurfaces.

PHA for hydrogen-bonding network

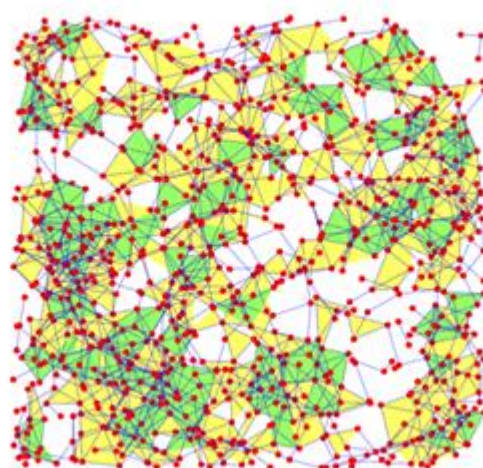
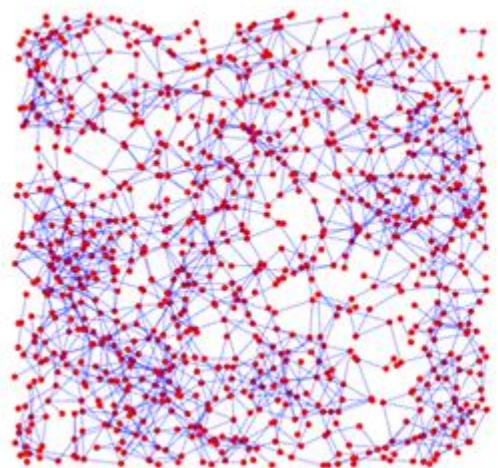


Collaborator
Yuguang Mu
SBS, NTU

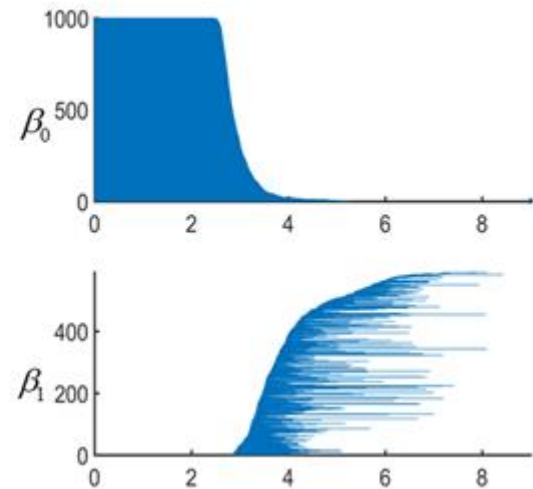
Ion crystal structure fingerprint



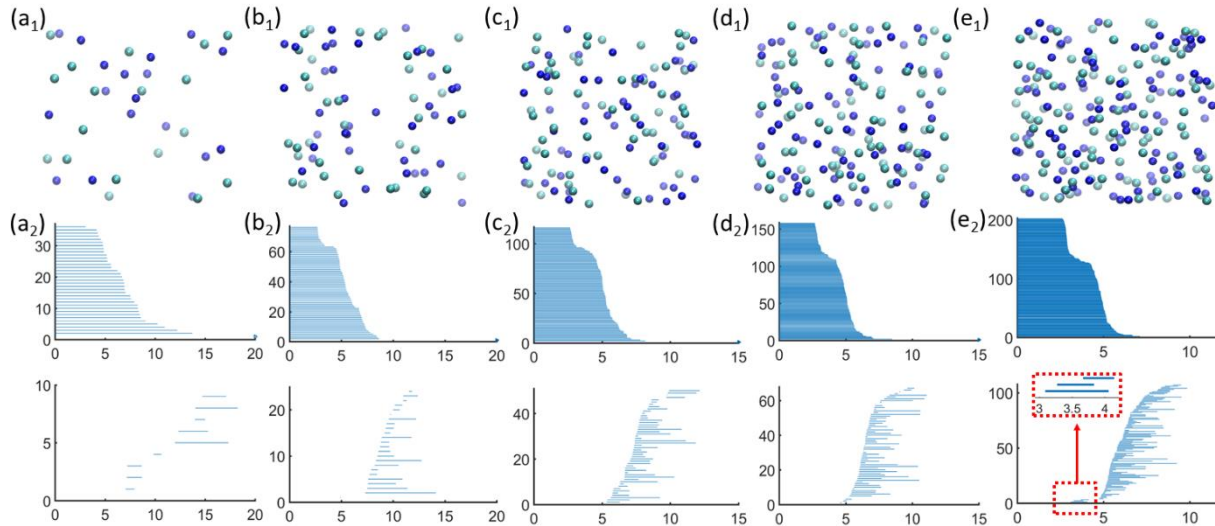
Water hydrogen-bonding fingerprint



(Xia, PCCP, 2018)

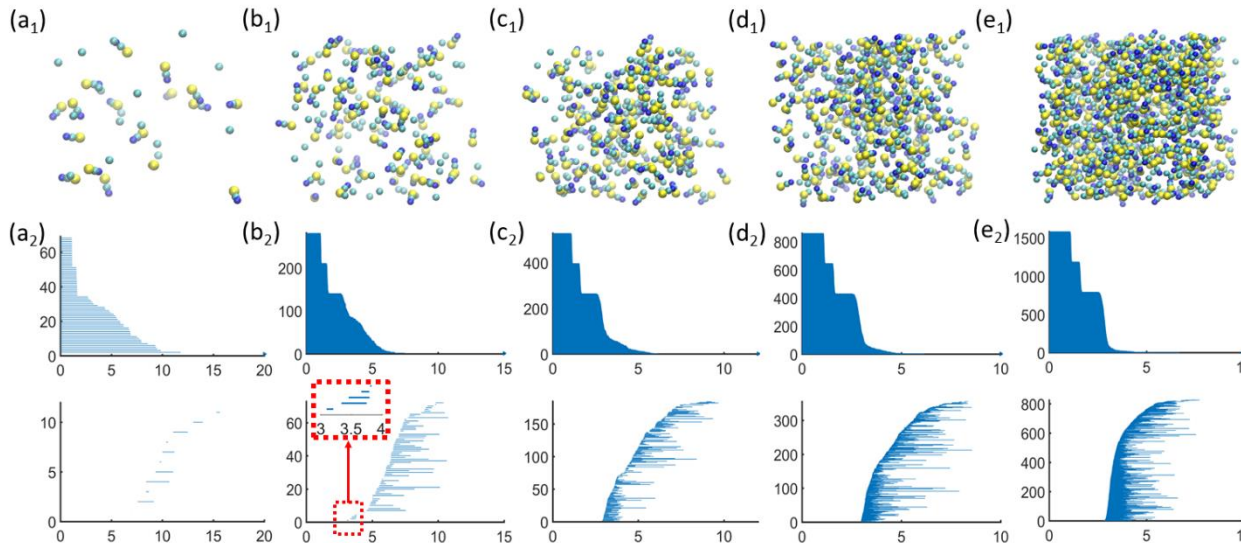


Two morphological types of ion aggregation



NaCl systems with concentrations: 1M, 2M, 3M, 4M and 5M

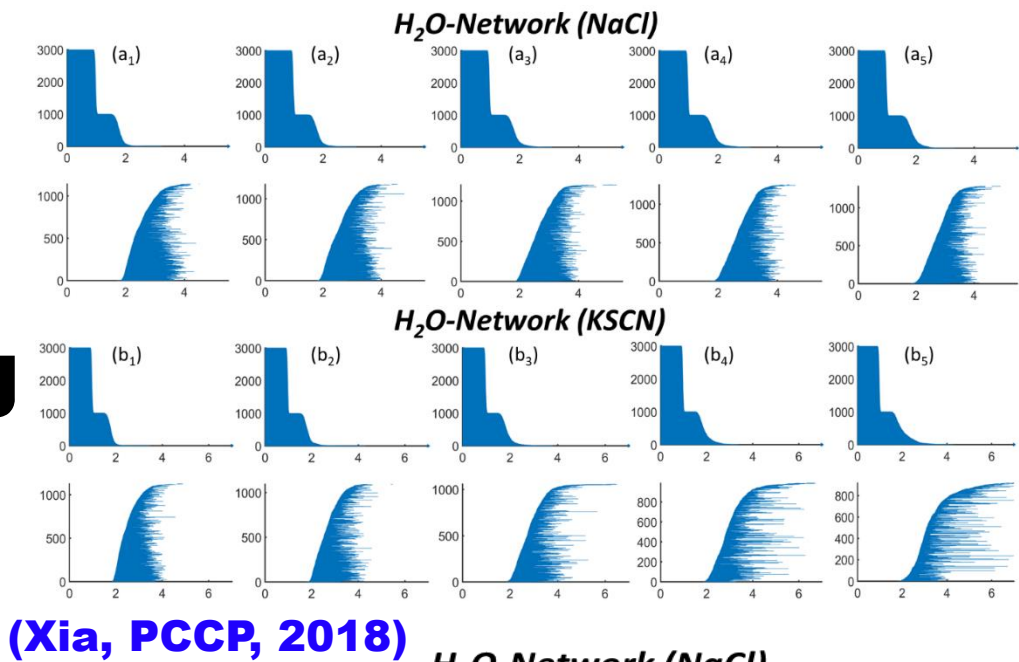
Type 1: local clusters



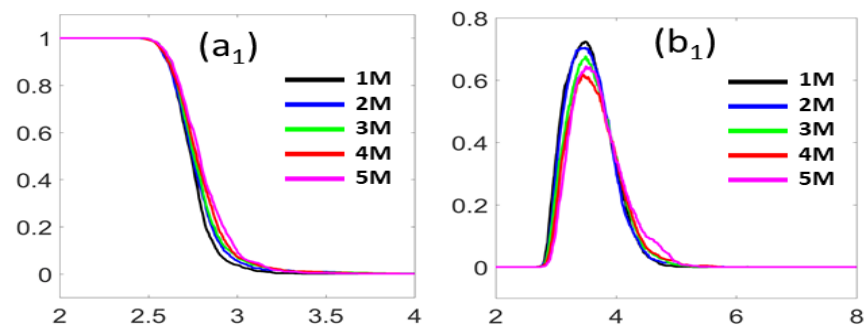
KSCN systems with concentrations: 1M, 3M, 5M, 7M and 10M

Type 2: extended ion network.

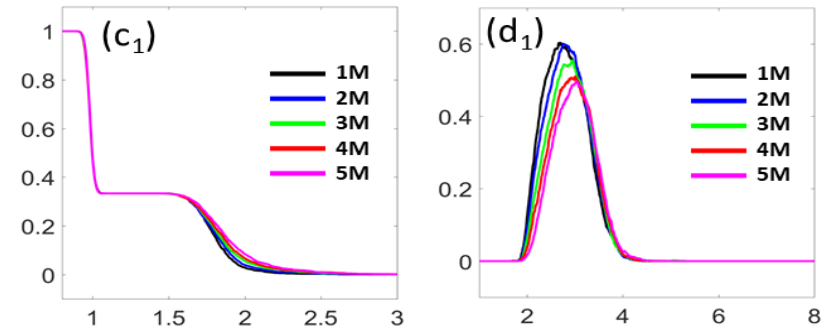
Two types of hydrogen-bonding networks



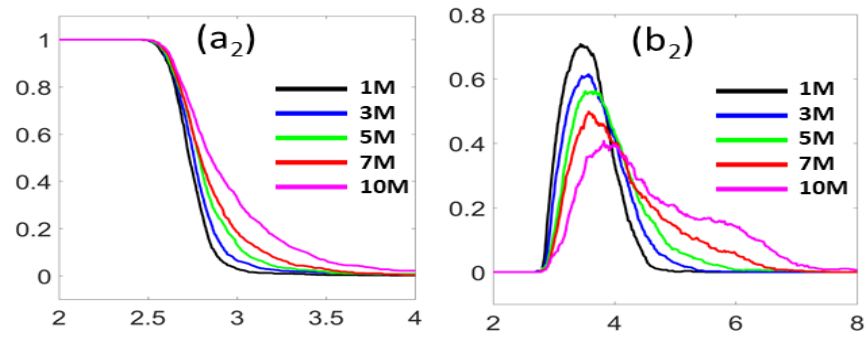
***O*-Network (NaCl)**



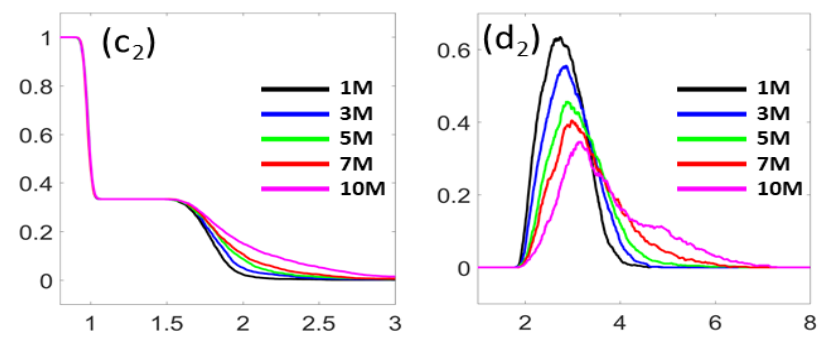
***H₂O*-Network (NaCl)**



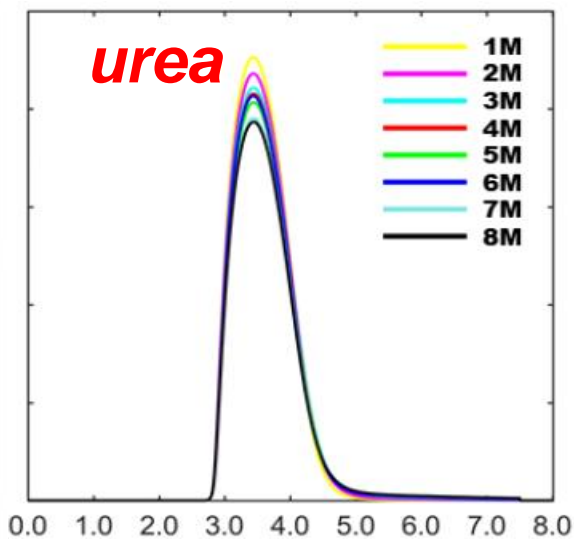
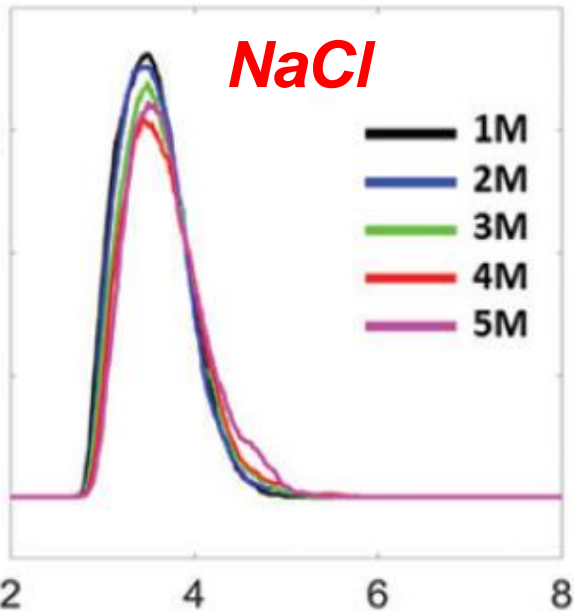
***O*-Network (KSCN)**



***H₂O*-Network (KSCN)**



Type 1: Structure “breaking”

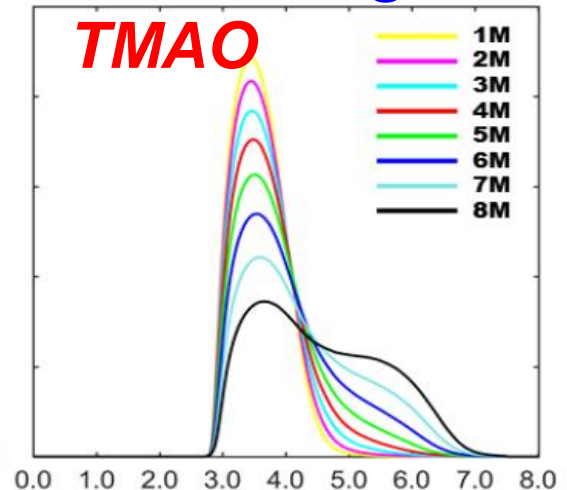
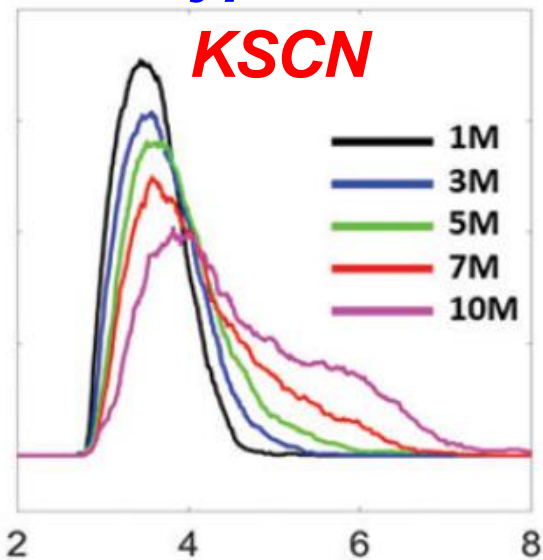


Two types of hydrogen-bonding networks from ion and osmolyte systems

Type 1:
preserve protein structure;

Type 2:
denature protein

Type 2: Structure “making”



Weighted persistent homology



Collaborator
Jie Wu
Math, NUS

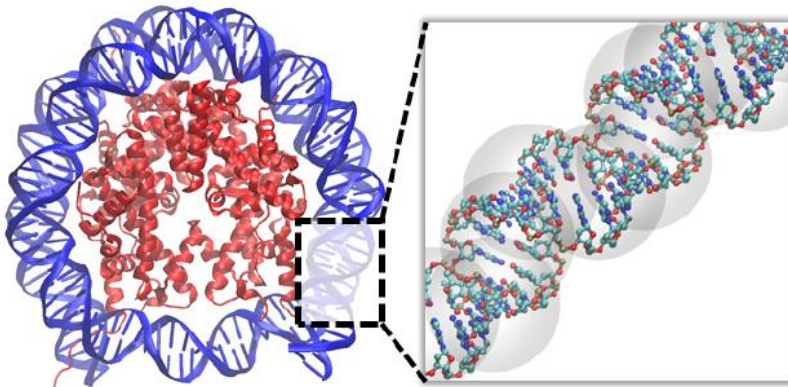
- **Weighted alpha complex;**
- **Weighted Vietoris-Rips;**
- **k-distance based models;**
- **Rigidity function based models;**
- **Weighted clique rank homology;**
- **Physics-aware models;**
- **Weighted simplicial homology;**
- **.....**

- *New filtration*
- *Weighted boundary map*

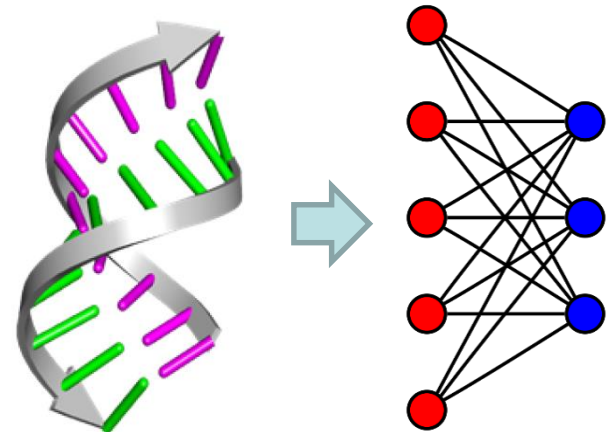
$$\partial_n(\sigma) = \sum_{i=0}^n \frac{w(\sigma)}{w(d_i(\sigma))} (-1)^i d_i(\sigma)$$

Labels: Simplex weight, Boundary operator, n-Simplex

Localized Persistent homology (LPH)



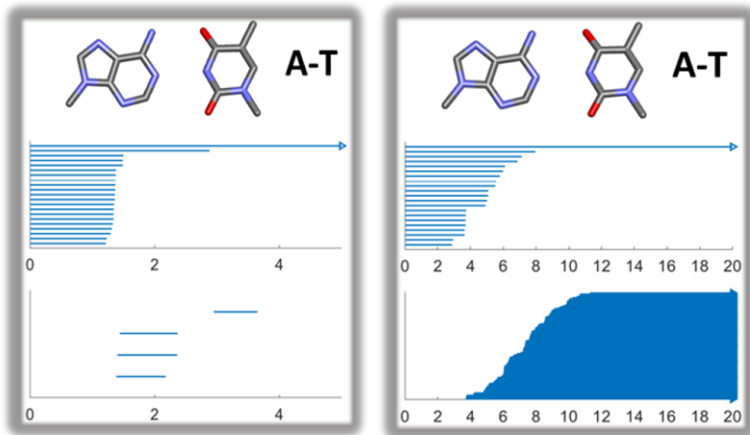
Interactive Persistent homology (IPH)



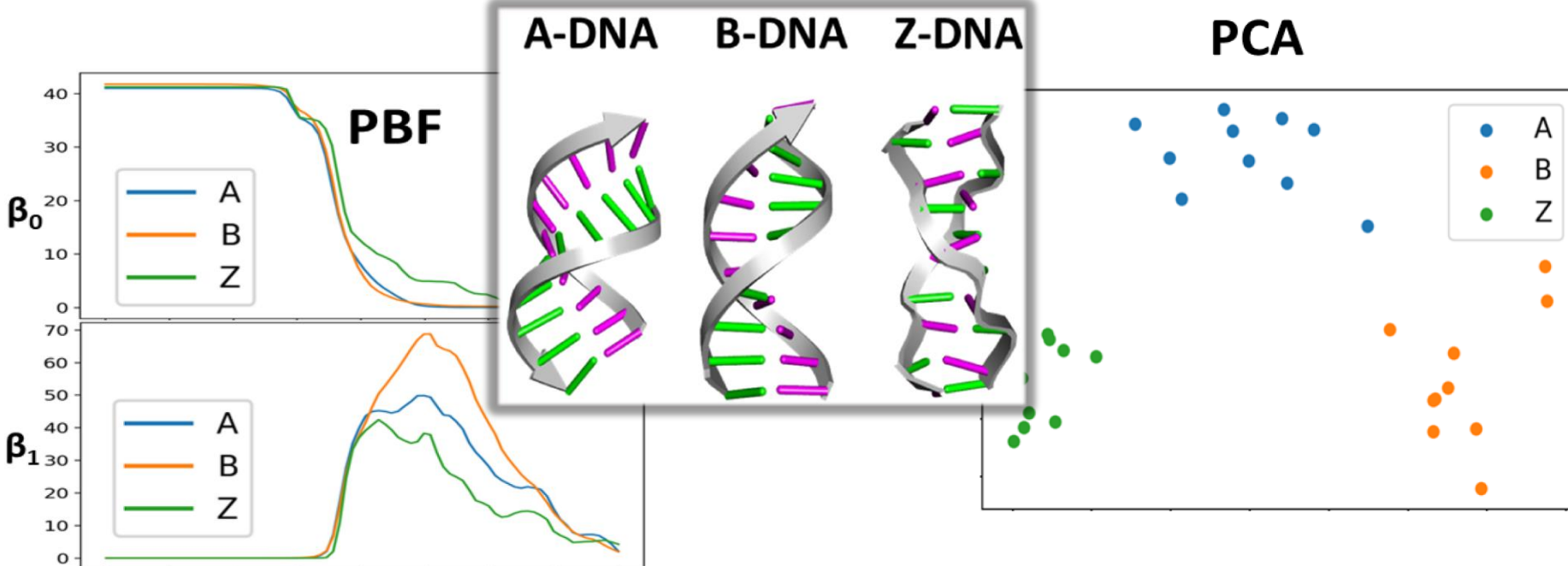
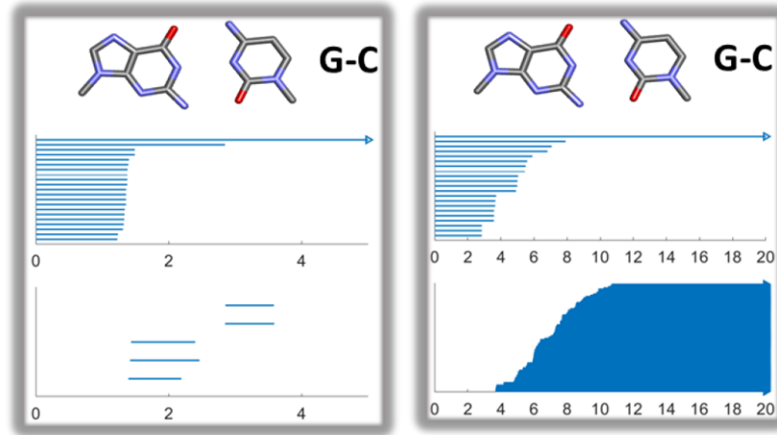
(Cang, Mu, Wei, PLOS Comp. Biol., 2018)

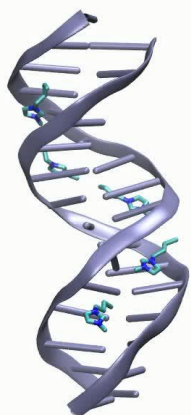
WPH for DNA classification

PH VS Interactive PH (AT)



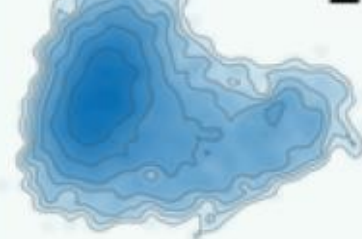
PH VS Interactive PH (GC)



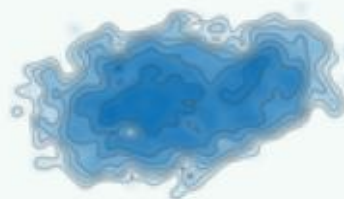


WPH for DNA clustering

(a₁) *ALL_IL*



(a₂) *IL1*



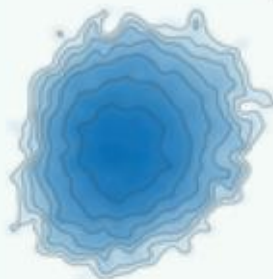
(a₃) *IL2*



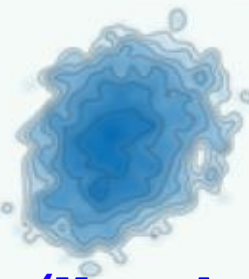
(a₄) *IL3*



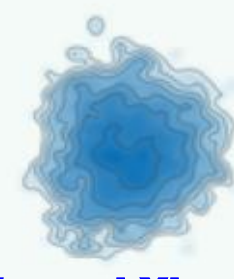
(b₁) *ALL_Wat*



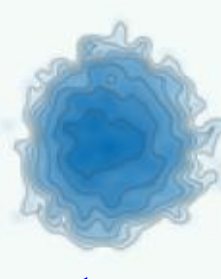
(b₂) *Wat1*



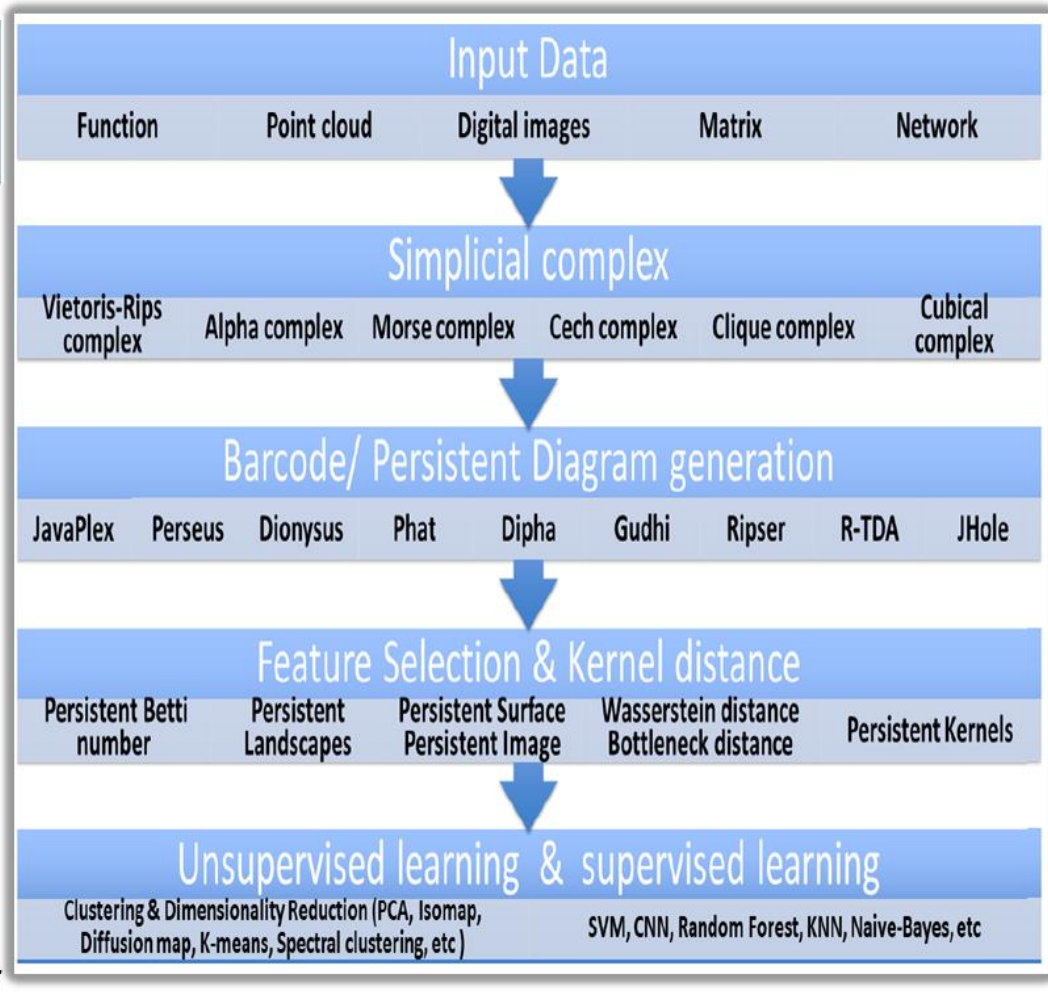
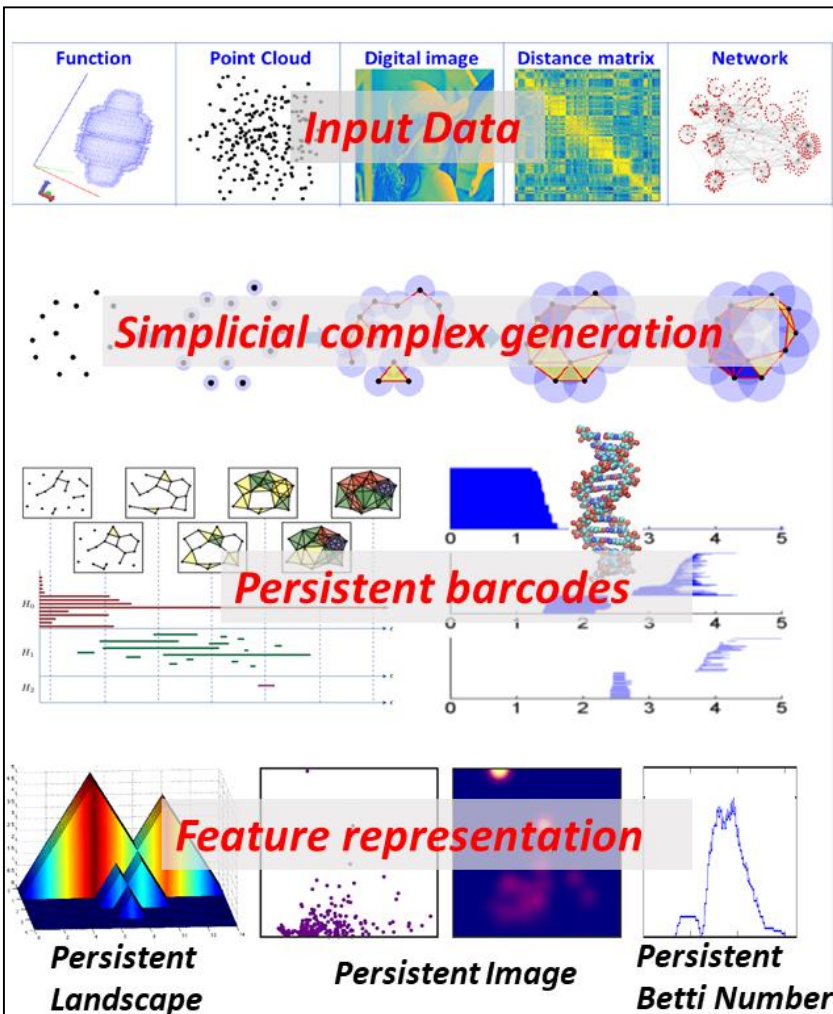
(b₃) *Wat2*



(b₄) *Wat3*

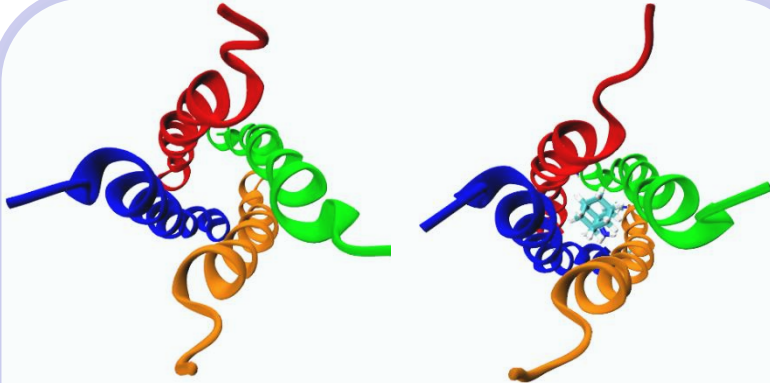


TDA based machine learning models

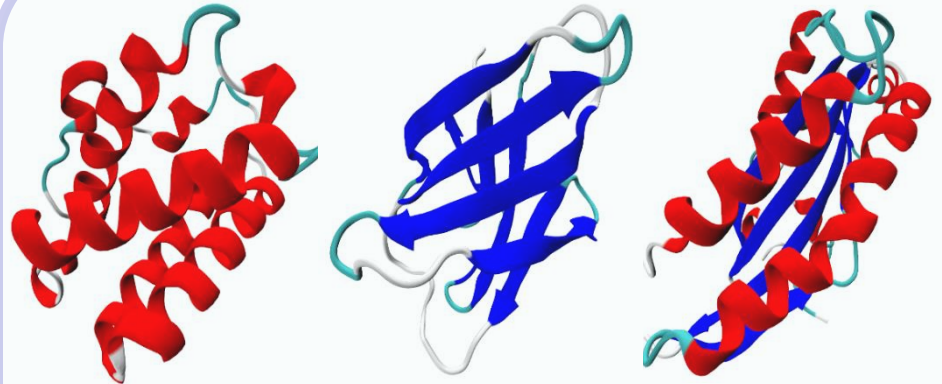


(Pun, Xia and Lee, submitted, 2020)

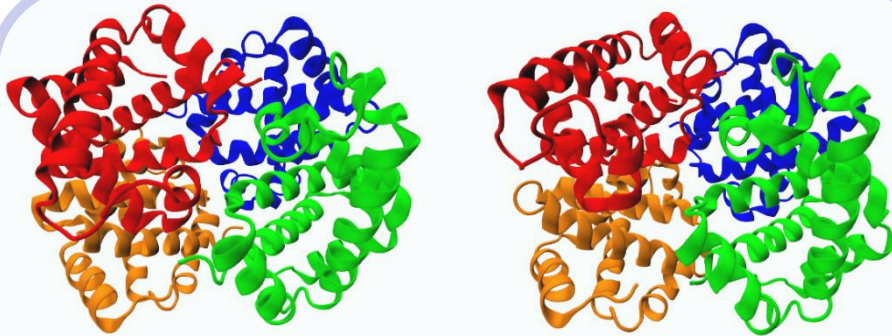
Topological fingerprint based machine learning method



Influenza A virus drug inhibition: 96% Accuracy



Protein domains: 85% Accuracy



Hemoglobins in their relaxed and taut forms: 80% accuracy

55 classification tasks of protein superfamilies over 1357 proteins from Protein Classification Benchmark Collection: 82% accuracy

Recent progress in TDA based drug design

Collaborator
Guowei Wei
MSU, USA



MORE AT SIAM **siam news**

HOME HAPPENING NOW GET INVOLVED RESEARCH

Get Involved | September 01, 2016

Mathematical Molecular Bioscience and Biophysics

A Recurring Theme at the SIAM Conference on the Life Sciences

MORE AT SIAM **siam news**

HOME HAPPENING NOW GET INVOLVED RESEARCH

Research | December 01, 2017

Persistent Homology Analysis of Biomolecular Data

Topological learning based predictions

Classification of ligands & decoys
DUD database 128,374 protein-ligand/decoy pairs

Cang and Wei, PLOS CB, 2018

Prediction RMSD of LogP (Star set)

Wu and Wei, JCC, 2018

Prediction correlations for 2648 mutations on globular proteins

Cang and Wei, PLOS CB, 2017

Prediction correlations for 223 mutations on membrane proteins

Binding affinity prediction of PDBBind v2013 core set of 195 complexes

Cang and Wei, PLOS CB, 2017

D3R Grand Challenge 2 (2016-2017)

Given: Farnesoid X receptor (FXR) and 102 ligands
Tasks: Dock 102 ligands to FXR, and predict their poses, binding free energies and energy ranking

Stage 1

- Pose Predictions (partials)
- Scoring (partials)
- Free Energy Set 1 (partials)
- Free Energy Set 2 (partials)

Stage 2

- Scoring (partials)
- Free Energy Set 1 (partials)
- Free Energy Set 2 (partials)

Grand Challenge 2 (Nguyen et al, JCAMD, 2018)

D3R Grand Challenge 3 (2017-2018)

(Nguyen et al, JCAMD, 2018)

Cathepsin Stage 1A

- Pose Predictions (partials)
- Affinity Rankings excluding Kds > 10 μM
- Cathepsin Stage 1
- Scoring (partials)
- Free Energy Set
- VEGFR2
- Scoring (partials)
- JAK2 SC3
- Scoring
- Free Energy Set
- Active / Inactive Classification
- VEGFR2
- Scoring (partials)
- JAK2 SC3
- Scoring
- Free Energy Set
- Affinity Rankings for Cocrystallized Ligands
- Cathepsin Stage 1
- Scoring (partials)
- Free Energy Set

Cathepsin Stage 1B

- Pose Prediction
- Affinity Rankings excluding Kds > 10 μM
- Cathepsin Stage 2
- Scoring (partials)
- Free Energy Set
- JAK2 SC2
- Scoring (partials)
- TIE2
- Scoring
- Free Energy Set 2
- JAK2 SC2
- Scoring (partials)
- TIE2
- Scoring (partials)
- Free Energy Set 1
- Cathepsin Stage 2
- Scoring (partials)
- Free Energy Set

p38-α Scoring (partials)

ABL1 Scoring (partials)

D3R Grand Challenge 4 (2018-2019)

Pose Predictions

- BACE Stage 1A
- Pose Predictions (Partials)

Affinity Predictions

- Cathepsin Stage 1
- Combined Ligand and Structure Based Scoring
- Ligand Based Scoring (No participation)
- Structure Based Scoring
- Free Energy Set

BACE Stage 1B

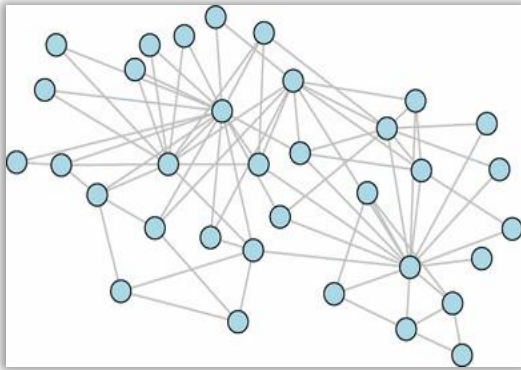
- Pose Prediction (Partials)

2/3 2/3 2/2 1/2

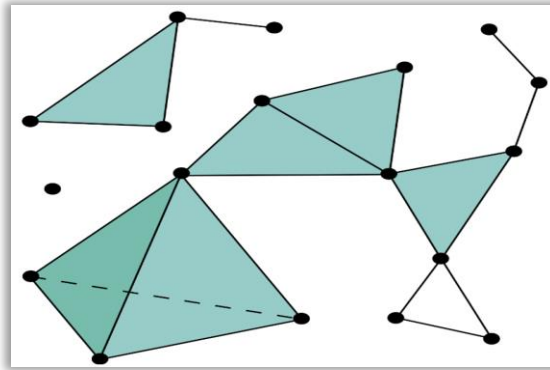
2/5 2/3 2/4 2/4 1/3 3/3 1/7 3/7 2/5

TDA is based on the multiscale simplicial complex

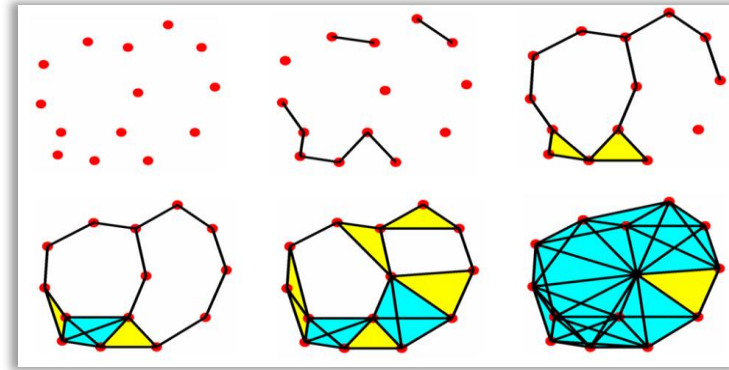
□ Graph



□ Simplicial complex



□ Multiscale simplicial complexes



❖ Graph models and measurements:

Graph Laplacian; Fiedler Eigenvalue; Fiedler eigenvector; Shortest path; Clique; Cluster coefficient; Closeness; Centrality; Betweenness; Modularity; Cheeger constant; Erdos number; Percolation...

❖ Simplicial complex models and measurements:

Combinatorial Laplacian; Hodge theory; Betti number; Euler characteristics; Homology; Cohomology; Morse theory; Knot polynomials...

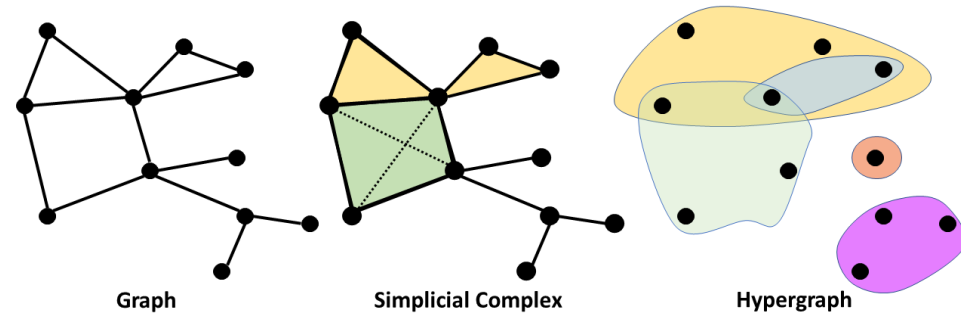
❖ Multiscale simplicial complex:

Persistent homology; Persistent cohomology...

Persistent Spectral theory (PerSpect)

Spectral models

- Spectral graph
- Spectral simplicial complex
- Spectral hypergraph



Filtration

- Nested sequence of Graphs
- Nested sequence of Simplicial Complexes
- Nested sequence of Hypergraph Laplacian

$$G^0 \subseteq G^1 \subseteq \dots \subseteq G^m$$

$$K^0 \subseteq K^1 \subseteq \dots \subseteq K^m$$

$$H^0 \subseteq H^1 \subseteq \dots \subseteq H^m$$

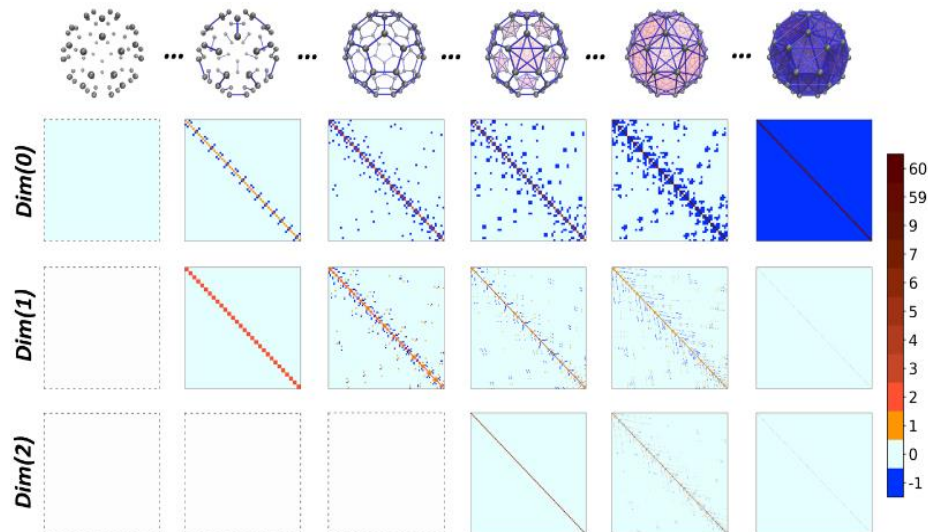
PerSpect = Spectral models + Filtration

- ❖ Persistent spectral graph
- ❖ Persistent spectral simplicial complexes
- ❖ Persistent spectral hypergraph

Persistent spectral simplicial complex

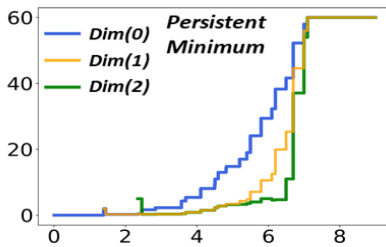
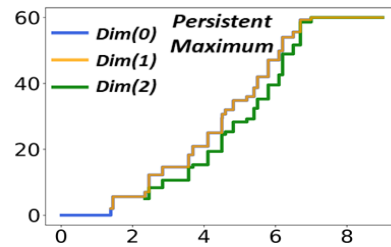
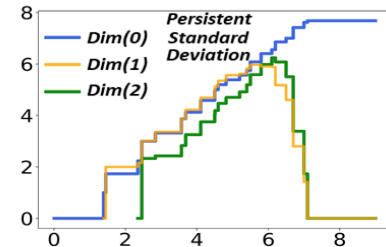
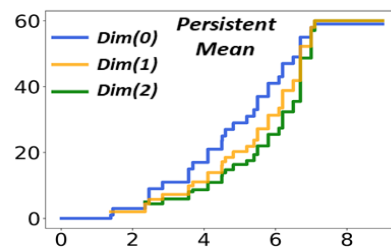
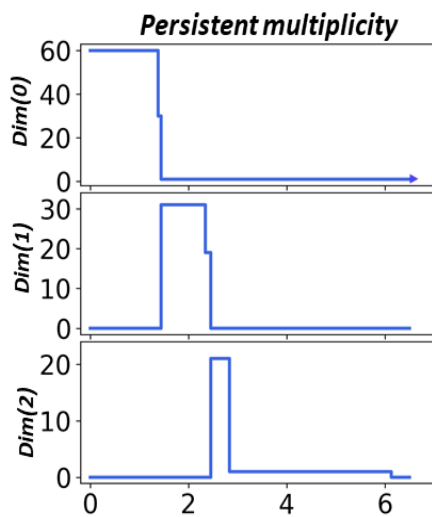
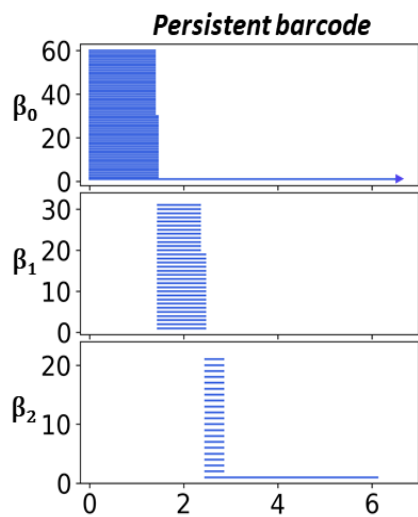
Boundary operator

$$B_k(i, j) = \begin{cases} 1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \sim \sigma_j^k \\ -1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \not\sim \sigma_j^k \\ 0, & \text{if } \sigma_i^{k-1} \not\subset \sigma_j^k. \end{cases}$$



Combinatorial Laplacian (Hodge Laplacian)

$$L_k = B_k^T B_k + B_{k+1} B_{k+1}^T.$$

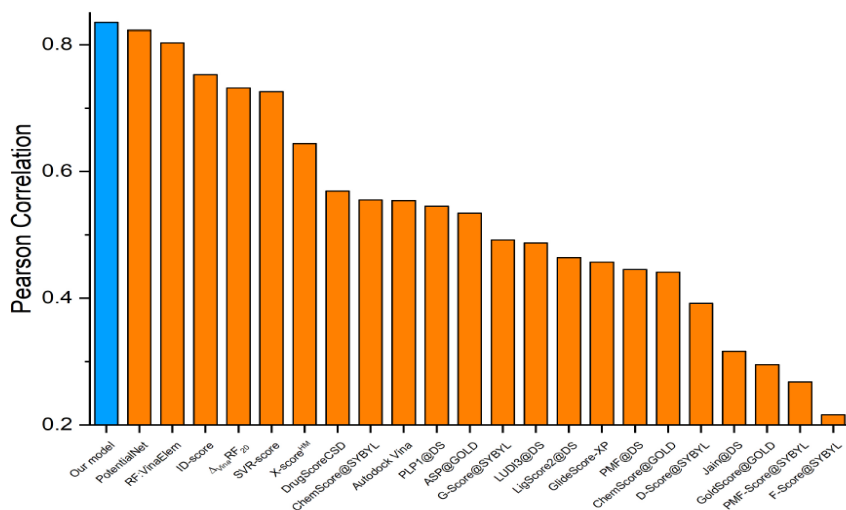


Multiplicity of zero eigenvalues (Persistent multiplicity) from PerSpect simplicial complex is equivalent to persistent Betti number.

PerSpect variables change with filtration parameter and incorporate in them related geometric information.

Ours: 0.836

Dataset 2007

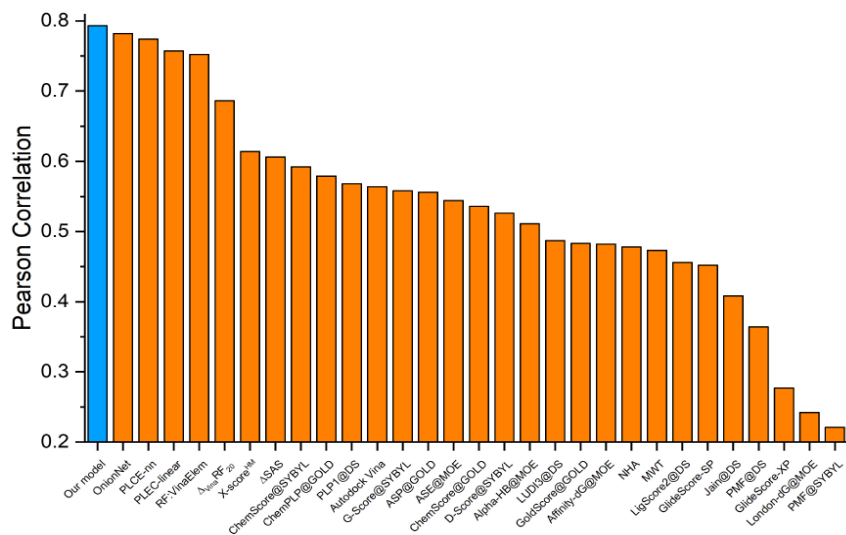


Benchmark testing with PDBbind datasets

**Model setting:
Spectral vectors
+
Random forest**

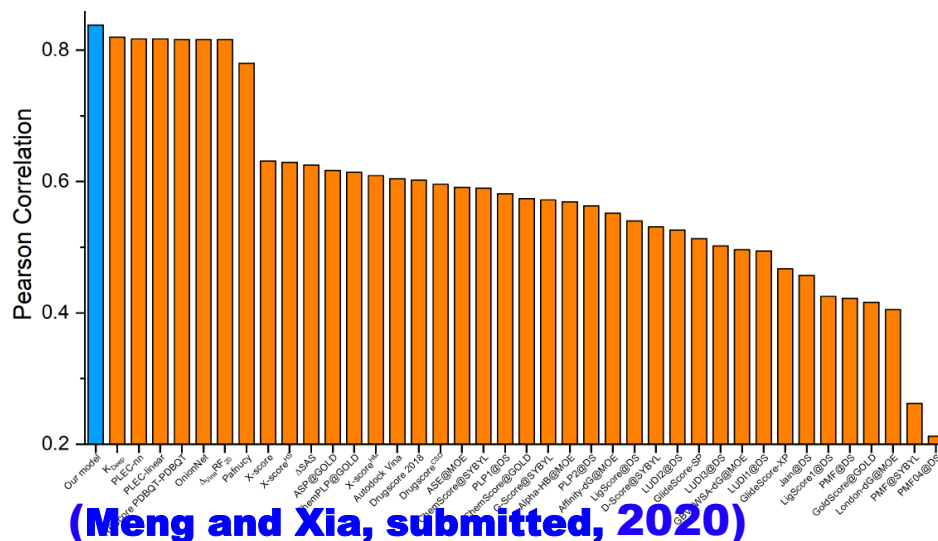
Ours: 0.793

Dataset 2013



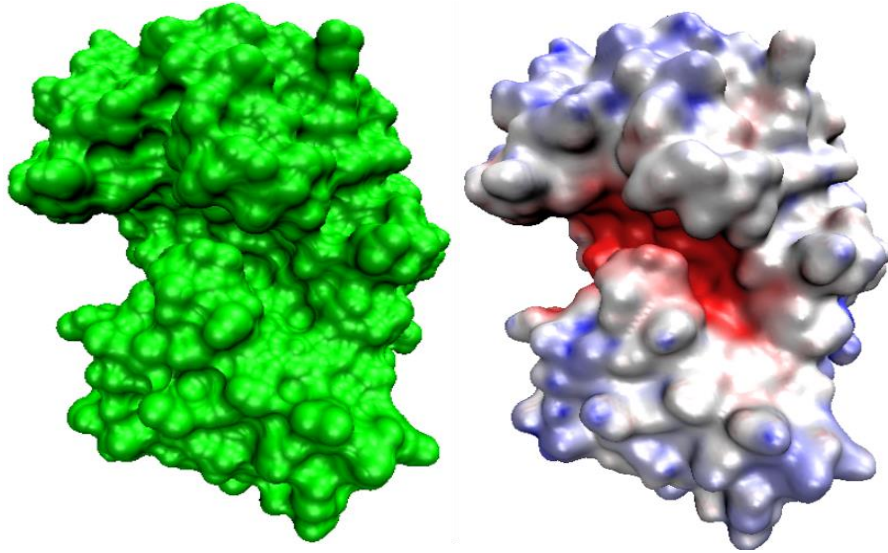
Ours: 0.840

Dataset 2016

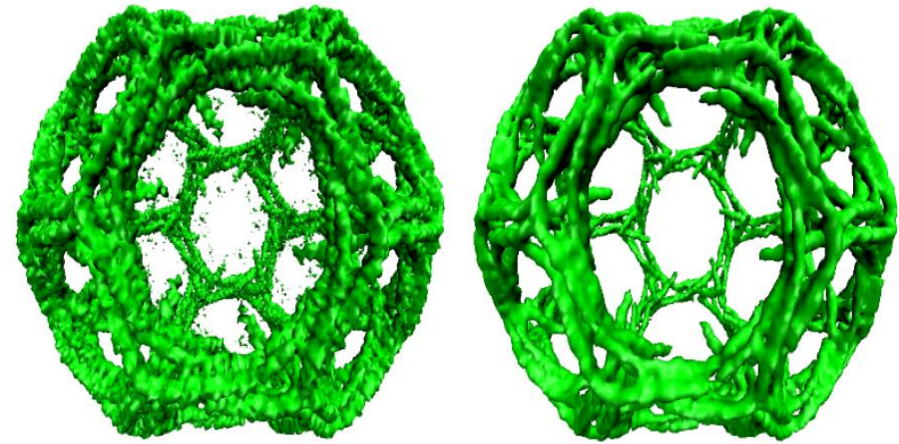


(Meng and Xia, submitted, 2020)

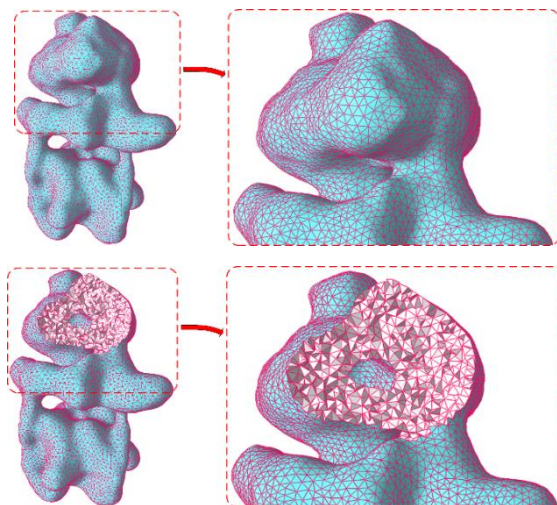
Variational multiscale modeling



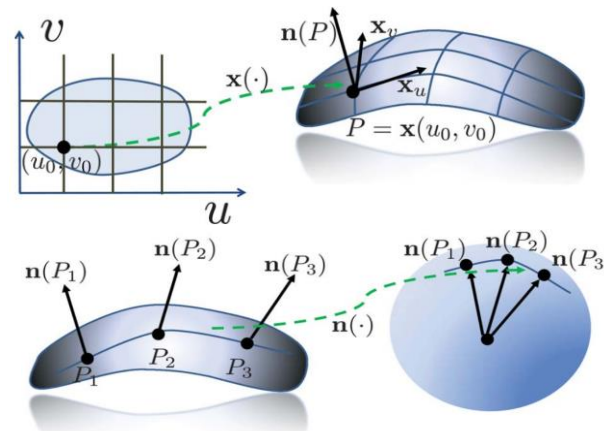
Geometric flow for noise reduction



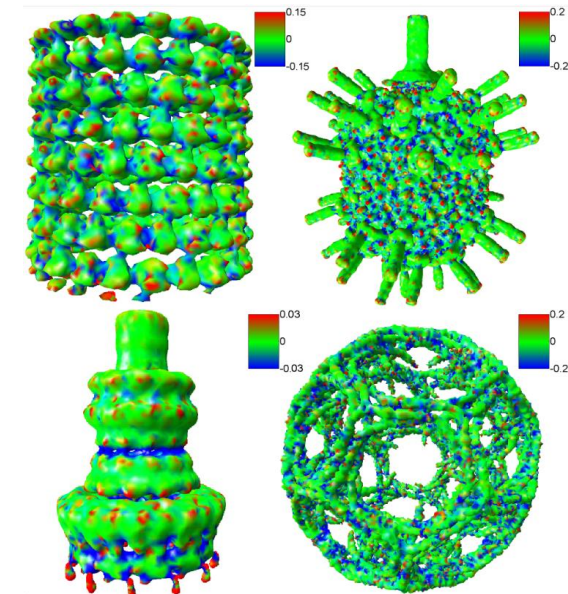
Geometric modeling of biomolecules



**Delaunay triangulation
based mesh generation**



**First and second
fundamental form**
(Feng, Xia, etc., JCC, 2013)



Gaussian curvature

Differential geometry based solvation model

The total functional:

$$G_{\text{total}}^{\text{PB}}[S, \Phi] = \int_{\Omega} \left\{ \gamma |\nabla S| + pS + S \left[-\frac{\epsilon_m}{2} |\nabla \Phi|^2 + \Phi \rho_m \right] \right. \\ \left. + (1 - S) \left[-\frac{\epsilon_s}{2} |\nabla \Phi|^2 - k_B T \sum_{\alpha} \rho_{\alpha 0} \left(e^{-\frac{q_{\alpha} \Phi + U_{\alpha} - \mu_{\alpha 0}}{k_B T}} - 1 \right) \right] \right\} d\mathbf{r}.$$

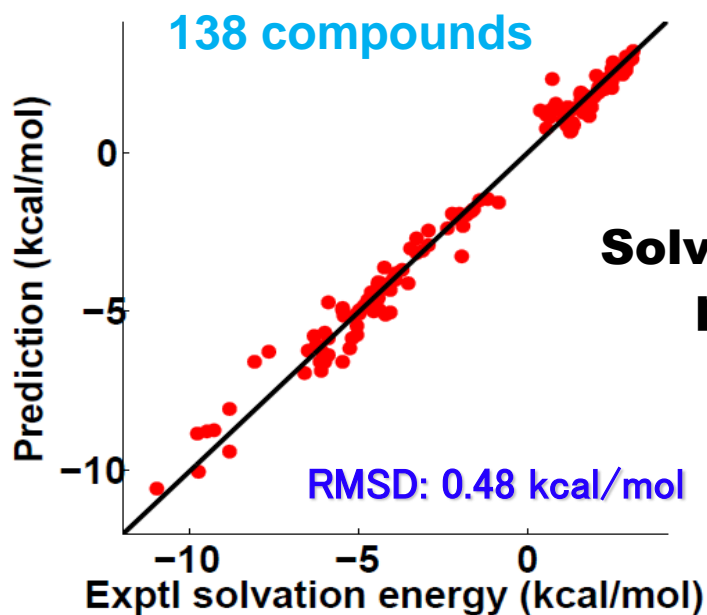
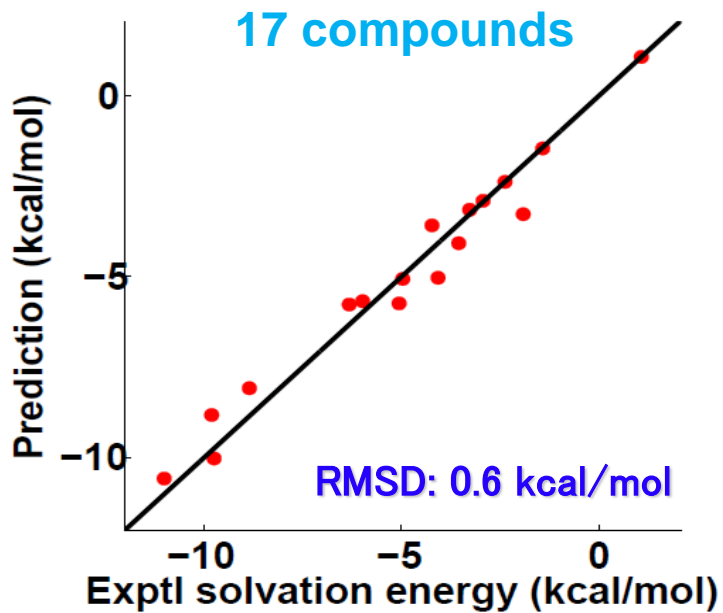
The generalized Poisson-Boltzmann equation:

$$-\nabla \cdot (\epsilon(S) \nabla \Phi) = S \rho_m + (1 - S) \sum_{\alpha} q_{\alpha} \rho_{\alpha 0} e^{-\frac{q_{\alpha} \Phi + U_{\alpha} - \mu_{\alpha 0}}{k_B T}}$$

The generalized mean curvature flow equation:

$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|} \right) + V_1 \right]$$

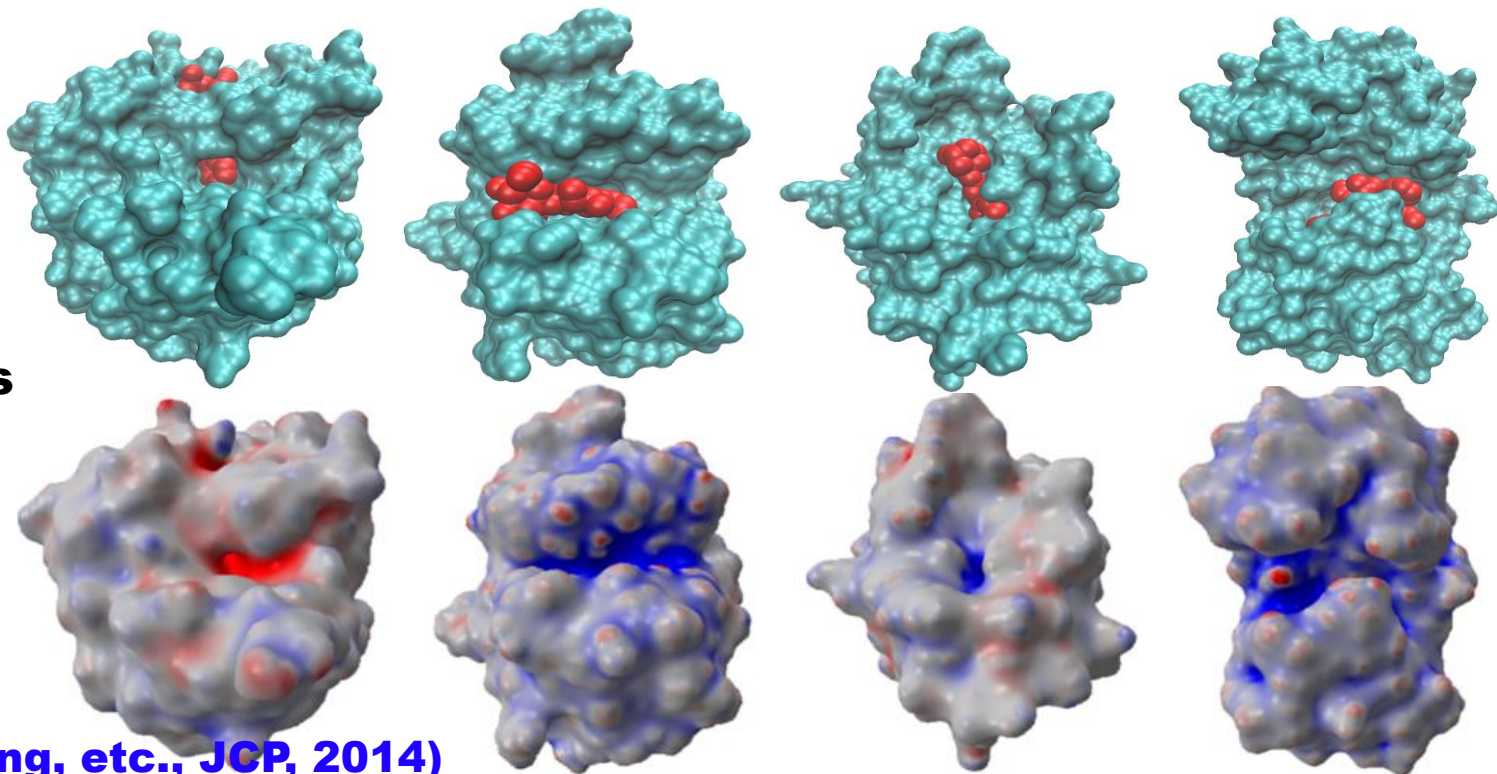
$$V_1 = -p + \frac{\epsilon_m}{2} |\nabla \Phi|^2 - \Phi \rho_m - \frac{\epsilon_s}{2} |\nabla \Phi|^2 - k_B T \sum_{\alpha} \rho_{\alpha 0} \left(e^{-\frac{q_{\alpha} \Phi + U_{\alpha} - \mu_{\alpha 0}}{k_B T}} - 1 \right)$$



Solvation energy prediction

Binding sites prediction

(Xia, Feng, etc., JCP, 2014)

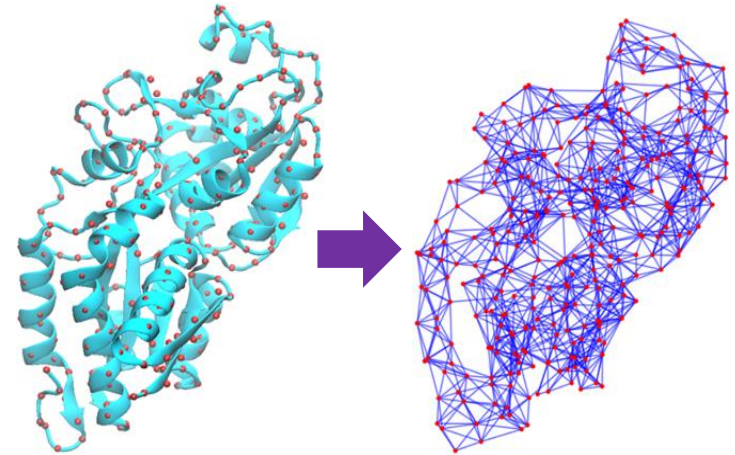
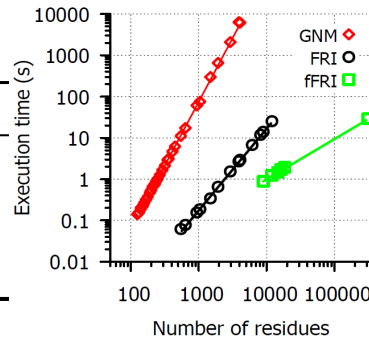


Graph modeling of biomolecules

(Opron, Xia, Wei, JCP, 2014)

Accuracy: (10%)

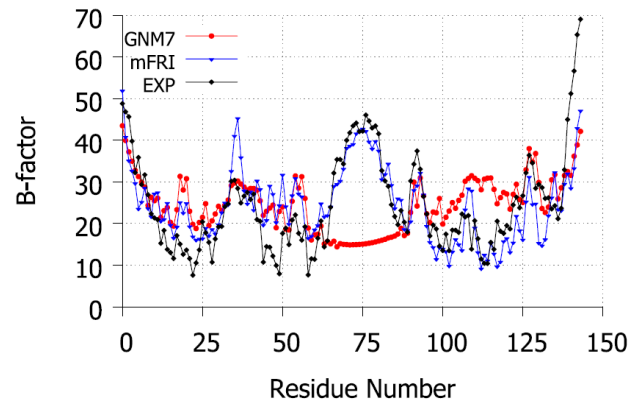
PDB set	pfFRI	GNM
Small	0.594	0.541
Medium	0.605	0.550
Large	0.591	0.529
Superset	0.626	0.565



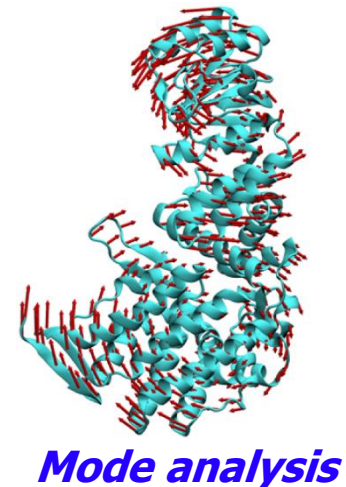
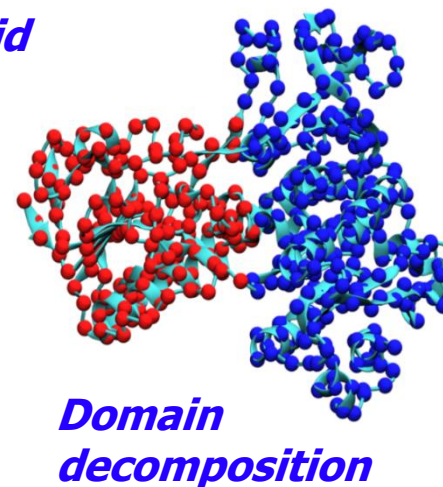
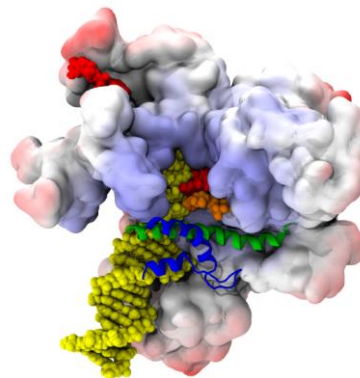
Exponential parameters	Avg. CC	Lorentz parameters	Avg. CC
$\kappa=0.5, \eta=0.5$	0.615 (8.8%)	$\nu=2.5, \eta=2.0$	0.622 (10.1%)
$\kappa=1.0, \eta=3.0$	0.623 (10.3%)	$\nu=3.0, \eta=3.0$	0.626 (10.8%)
$\kappa=1.5, \eta=6.0$	0.619 (9.6%)	$\nu=3.5, \eta=4.0$	0.623 (10.3%)

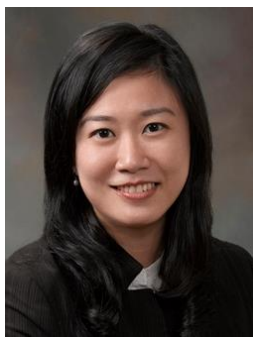
Further development: multiscale FRI, anisotropic FRI...

Multiscale FRI



Protein-Nucleic Acid Flexibility



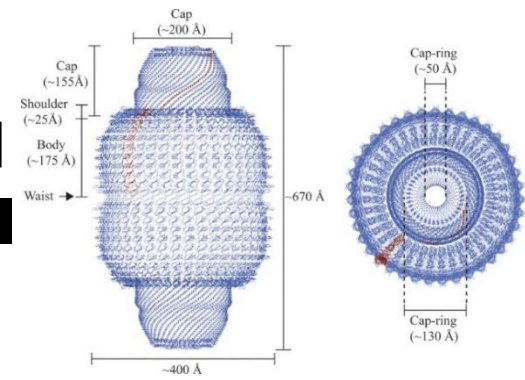


Collaborator
Sierin Lim
SCBE, NTU

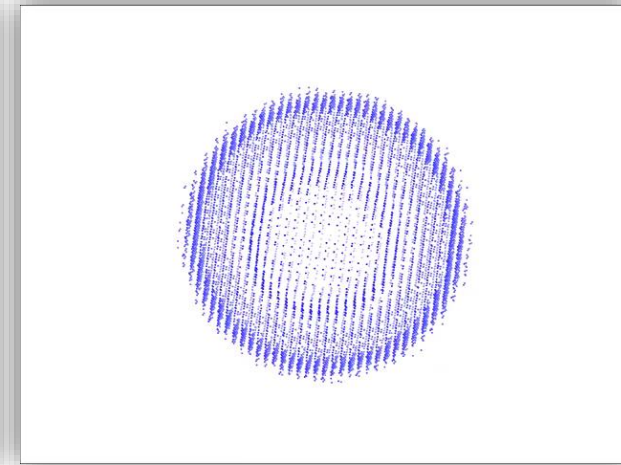
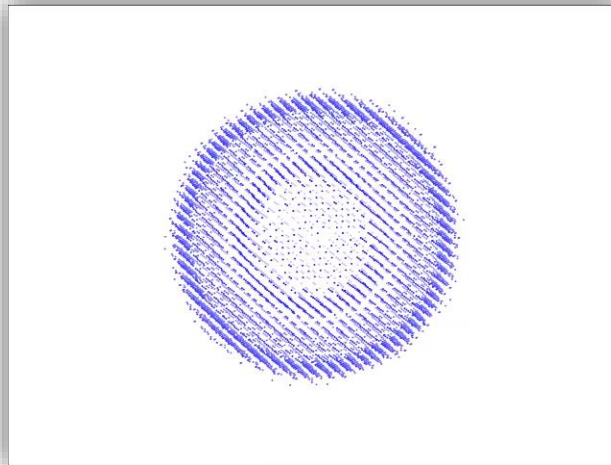
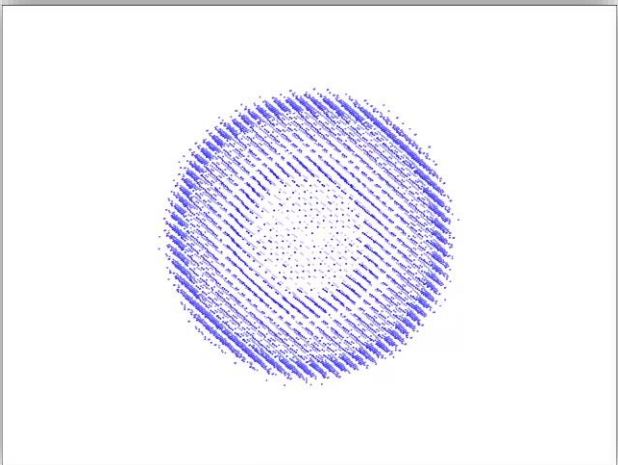
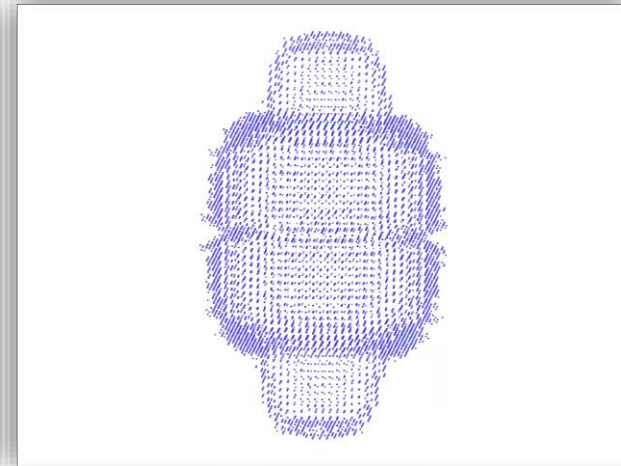
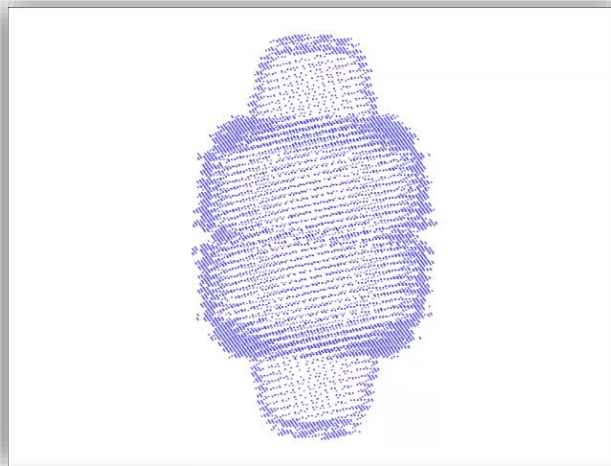
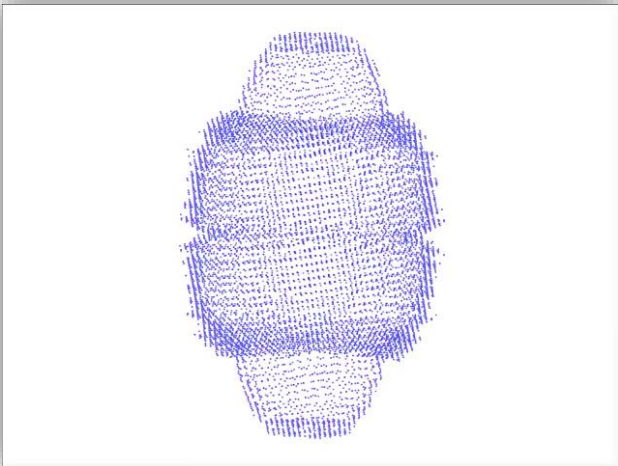


Collaborator
Takafumi Ueno
Tokyo Tech

Multiscale Virtual particle based elastic network model of Vault (on-going)

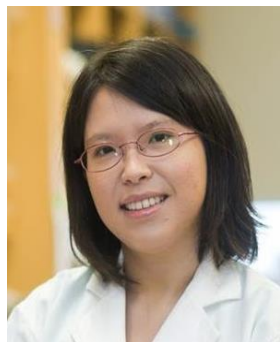


(Xia, PCCP, 2018)



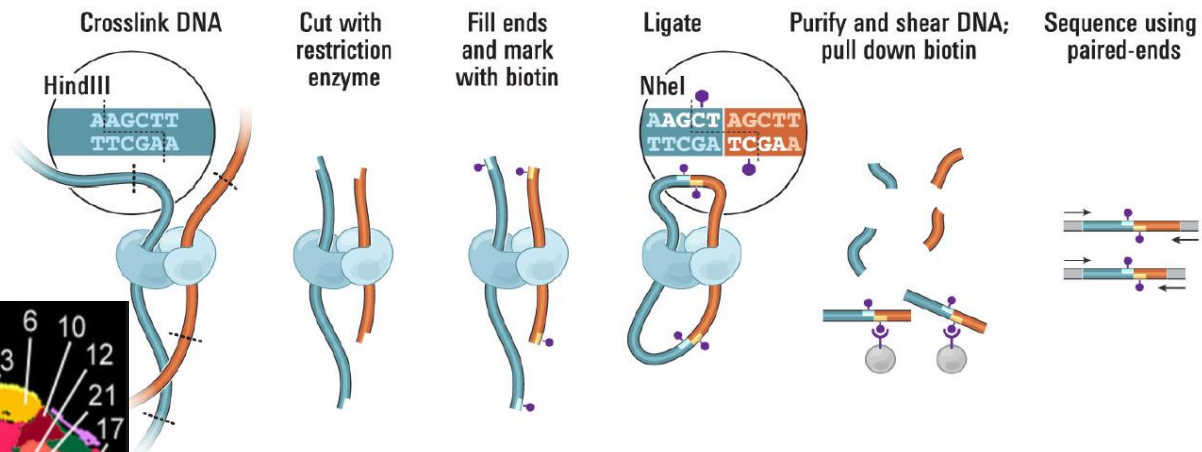


Collaborator
Jiajie Peng
CS, NWPU

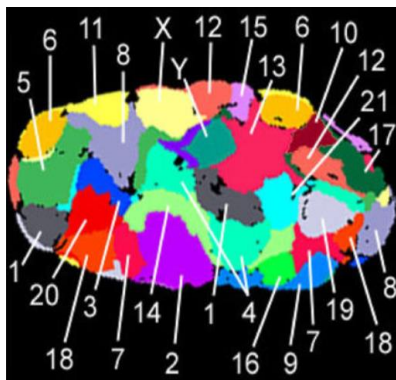
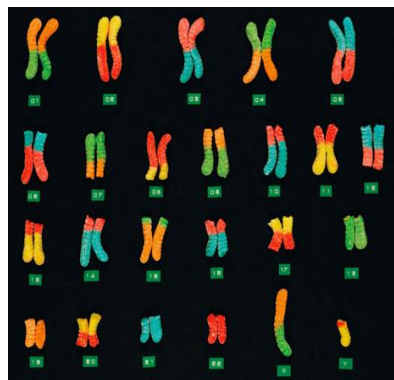


Collaborator
Melissa Fullwood
SBS, NTU

Hi-C Data analysis (on-going)

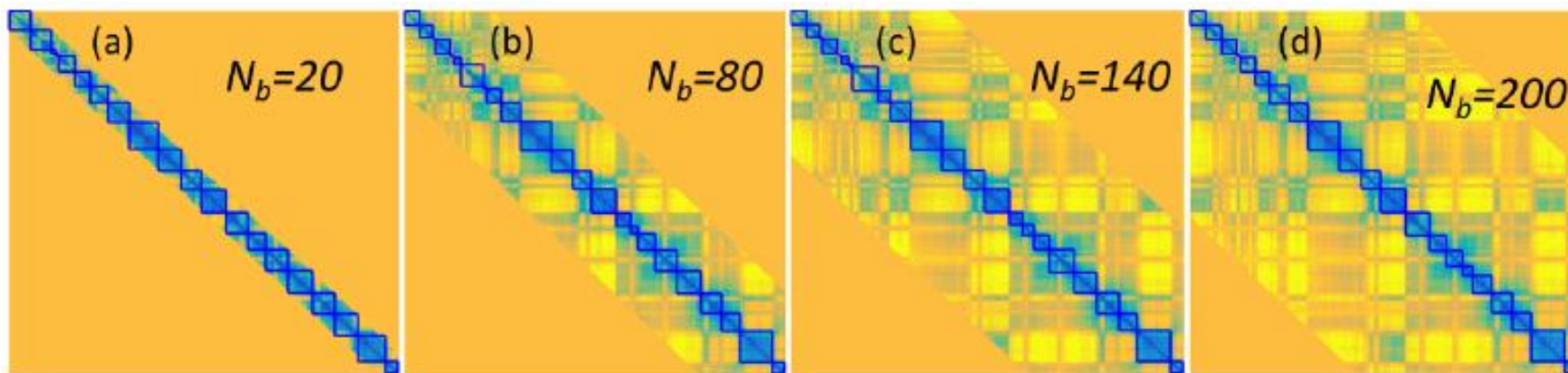


Chromosome Conformation Capture



(Xia, Plos one, 2018)

A multiscale spectral graph model for Hi-C data analysis



(Peng, Yang, Xia, Bioinformatics, revised 2019)

Data analysis--Machine learning and data mining:

Statistic learning;
Machine learning;
Deep learning;
Nonlinear dimensionality reduction
Data mining...

***Combine together
to provide features
for data analysis***

Geometric representation and modeling;

Topological representation and modeling;

Combinatorial representation and modeling...

Physics and

Biophysics models:

Fokker-Planck equation, Brownian dynamics, Langevin dynamics, molecular dynamics, master equation, Poisson-Nernst-Planck equations, Kohn-Sham equation, Navier-Stokes equation, Laplace-Beltrami equation, mean curvature flow, Poisson-Boltzmann equation, Maxwell's equations, anisotropic diffusion equation...

Part 1: Biomolecular topology

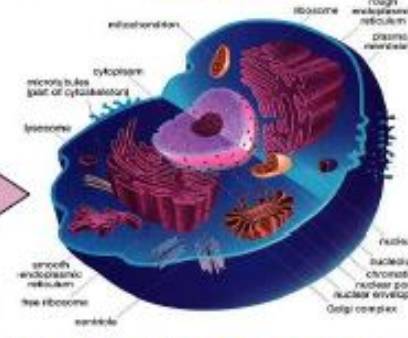
Species



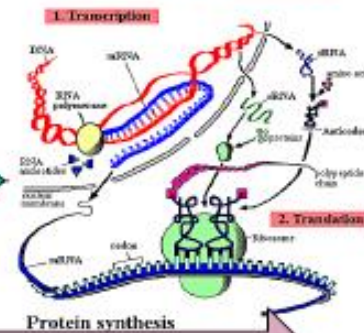
Organs



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

A Brief Summary of Modern Biological Science

1960

2000

2019

Organismal biology
(i.e., nonliving organisms, living organisms, developmental biology, morphology, anatomy, physiology, and medicine)

Molecular organismal biology, organomics, connectomics, foodomics, physiomics, pharmacogenomics, ...

Ecology

Molecular ecology

Evolution (i.e., life, and evolutionary biology)

Molecular evolution

Molecular and cellular biology
(i.e., cell biology, biochemistry, molecular biology, and genetics)

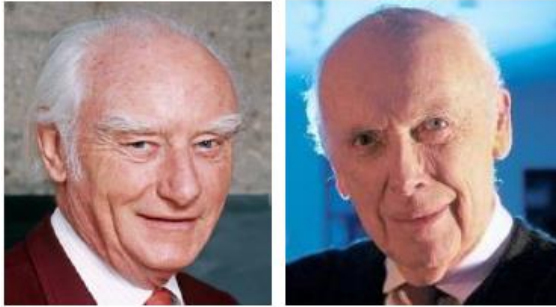
Omics (e.g., genomics, proteomics, metabolomics, metagenomics, lipidomics, glycomics, transcriptomics, epigenomics, ...)

Macroscopic

Mesososcopic

Microscopic

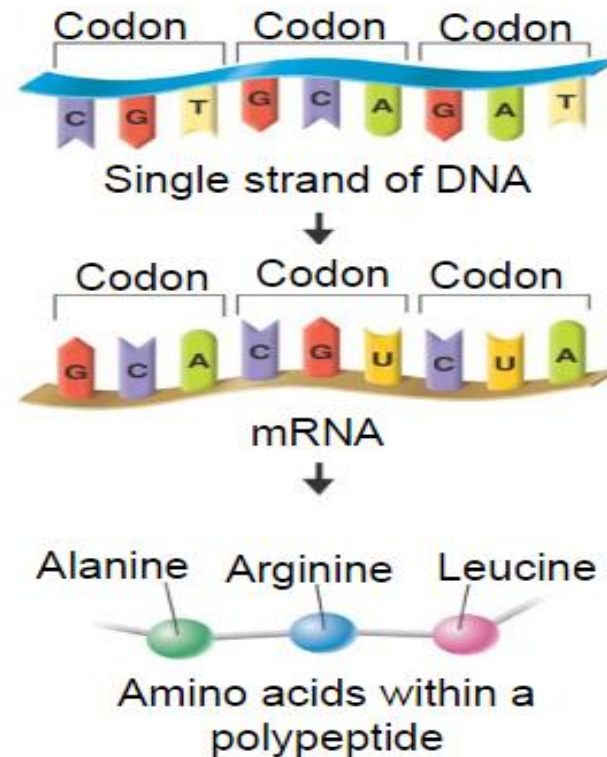
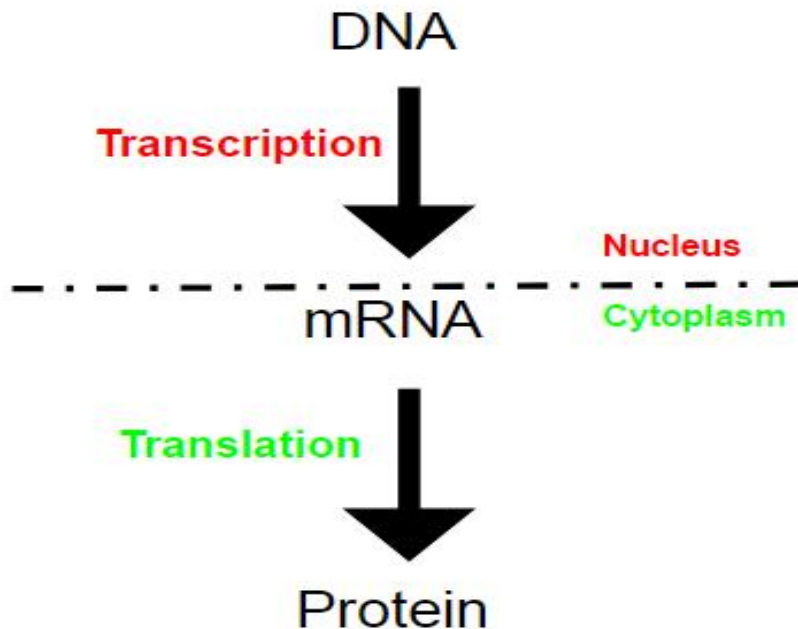
A Brief History of Biological Science



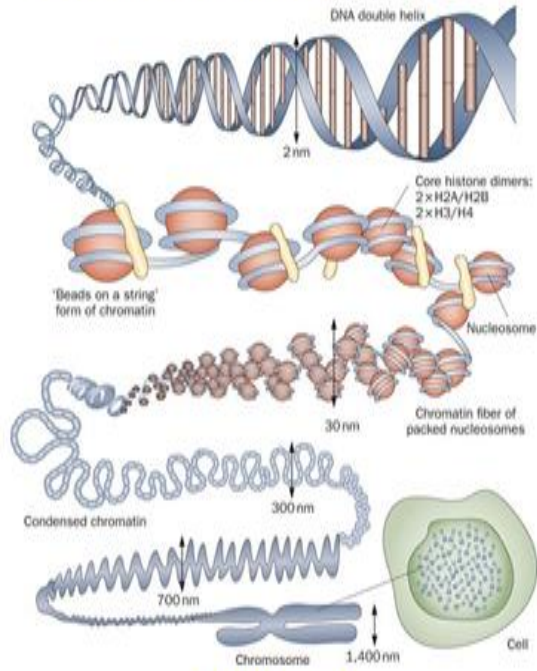
Francis Crick (1916 - 2004) and James Watson (1928 -) Nobel Prize in Physiology or Medicine in 1962 "for discoveries concerning the molecular structure of

nucleic acids and its significance for information transfer in living material", i.e., the **Central Dogma** of molecular biology:

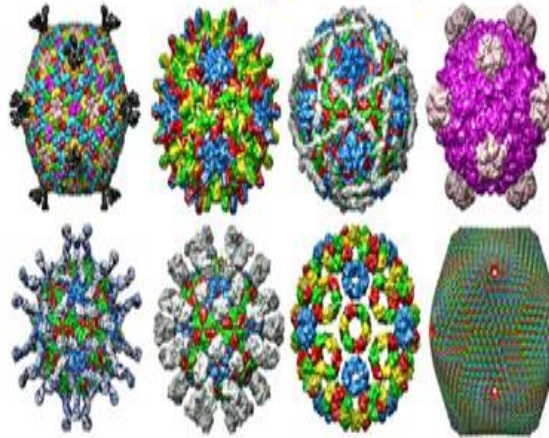
Central Dogma in a nutshell



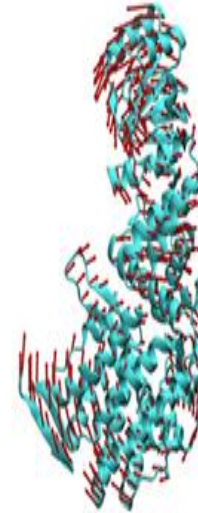
Chromosome structure



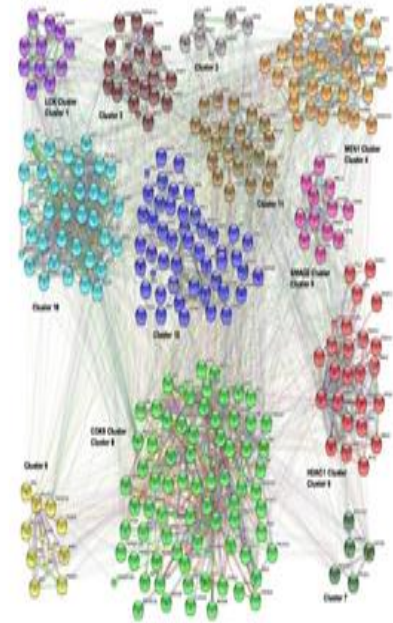
Virus self-assembly



Molecular dynamics

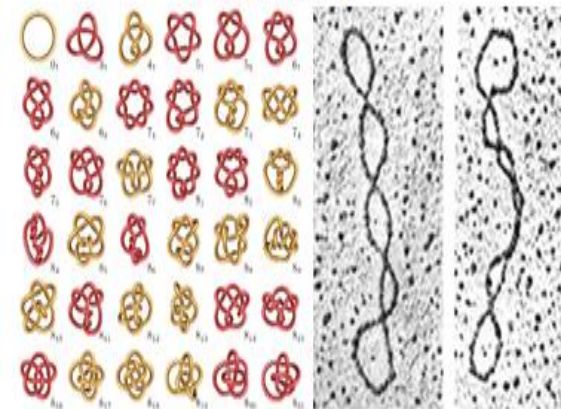


Protein-protein Interactions

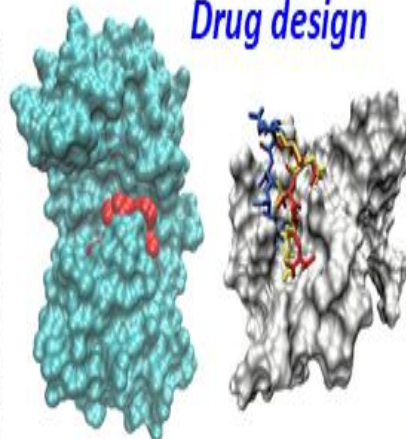


Biomolecular Topology

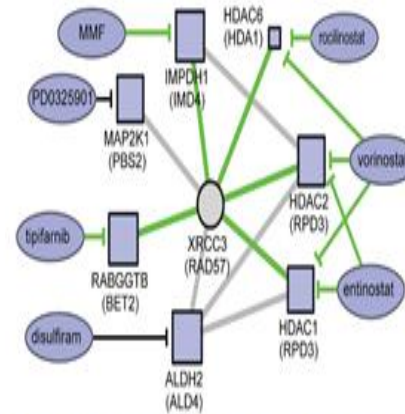
DNA knots



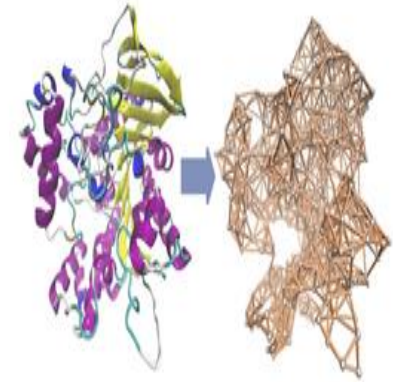
Drug design



Biomolecular regulatory network



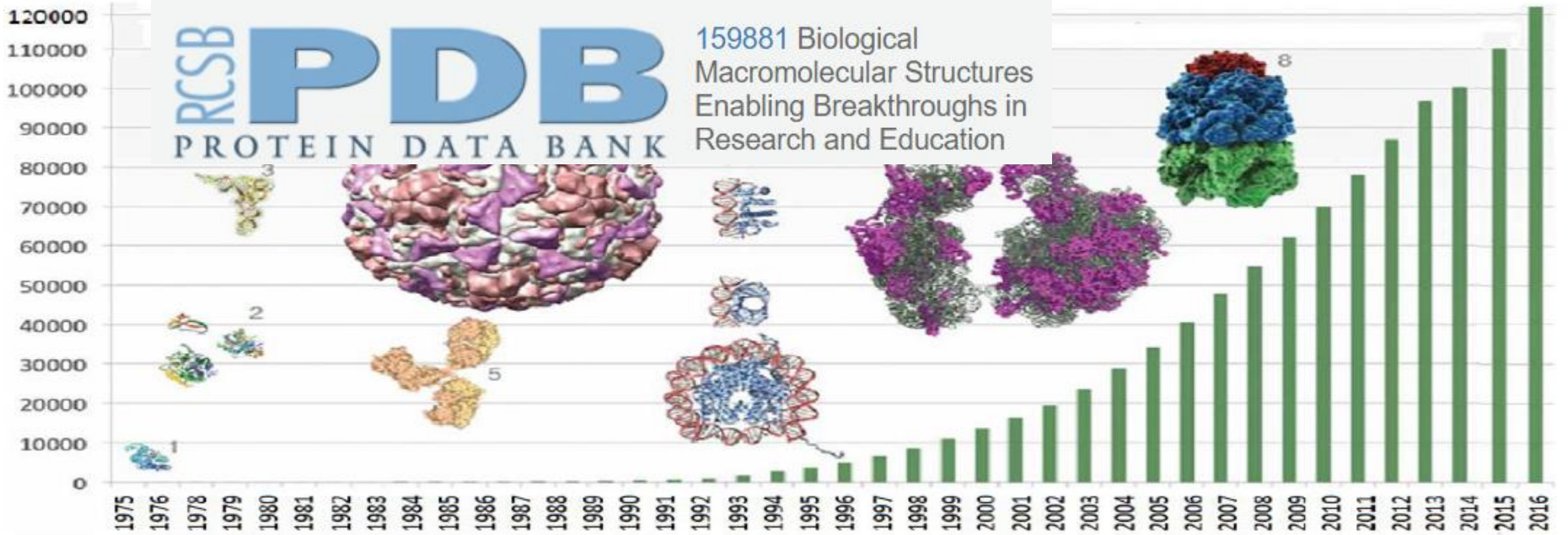
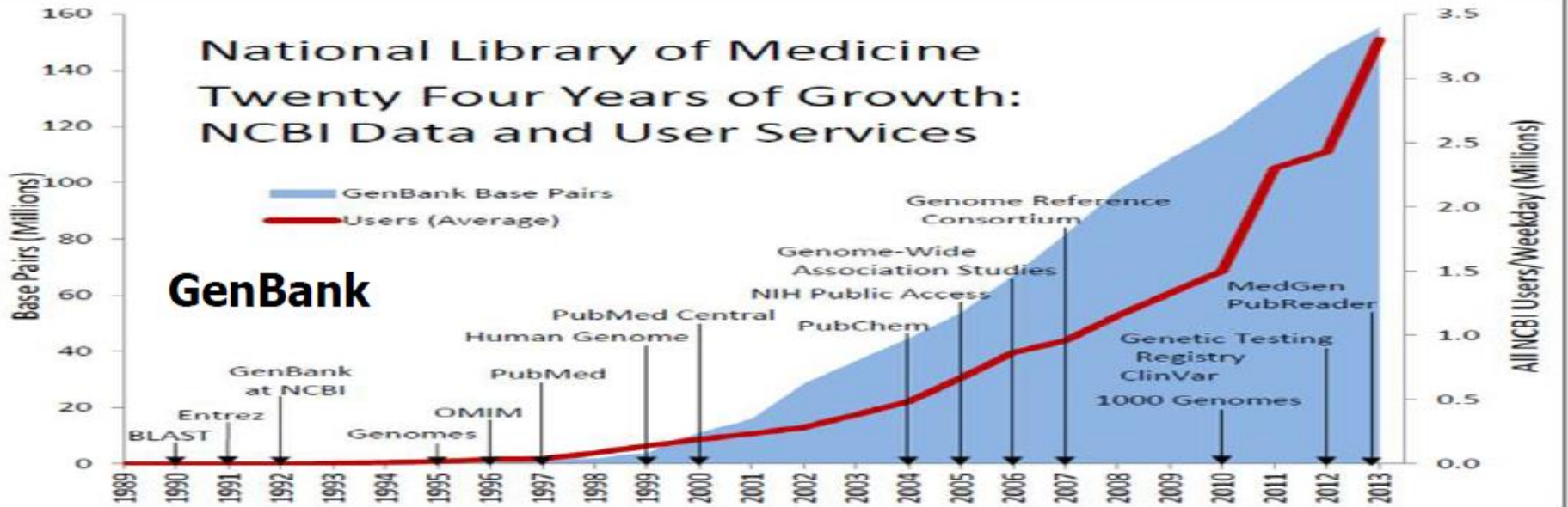
Protein structure network



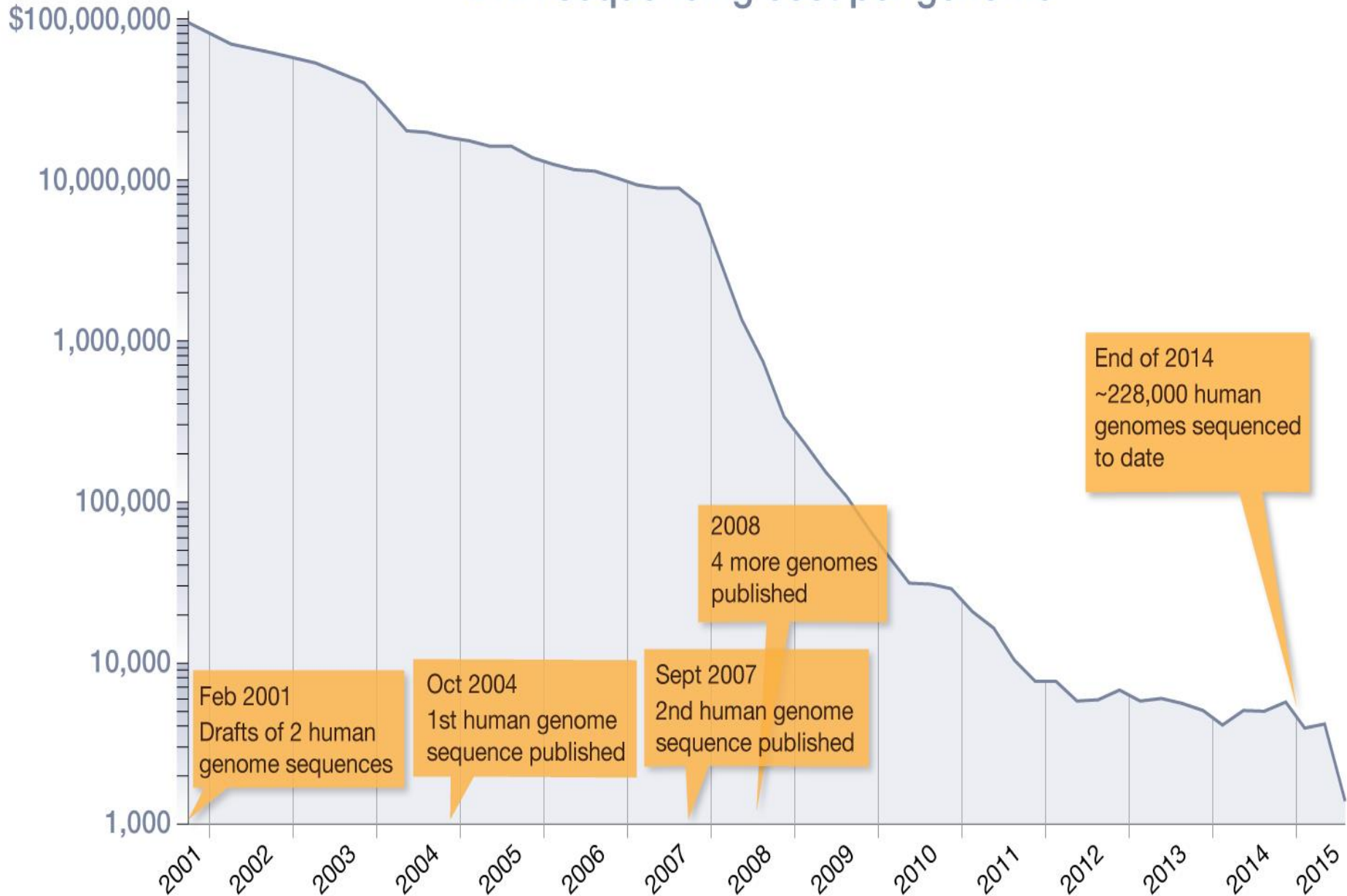
Biological data

National Library of Medicine Twenty Four Years of Growth: NCBI Data and User Services

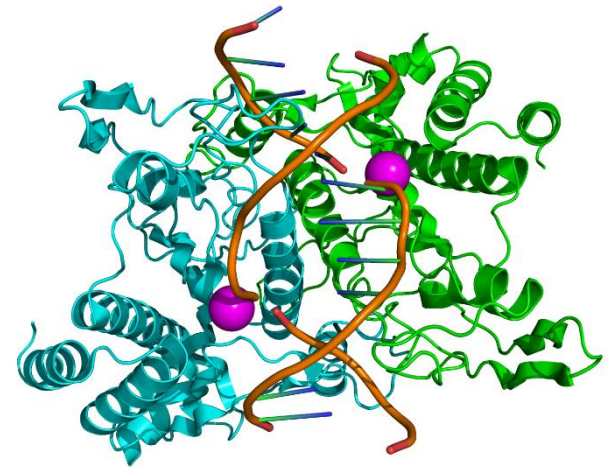
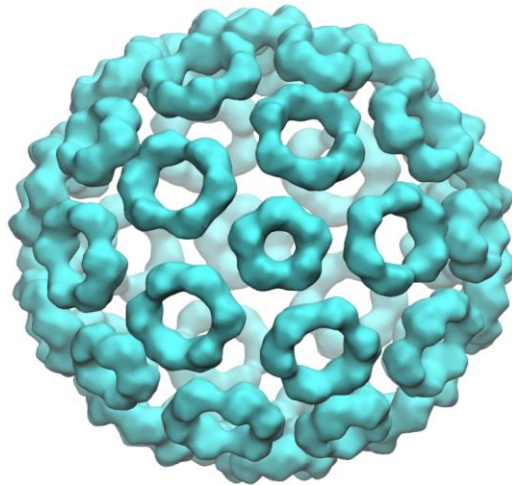
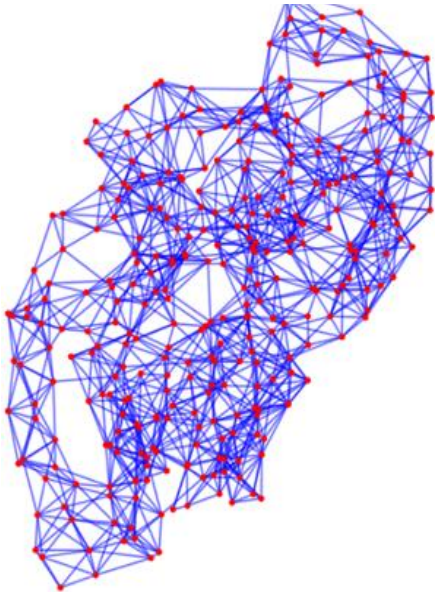
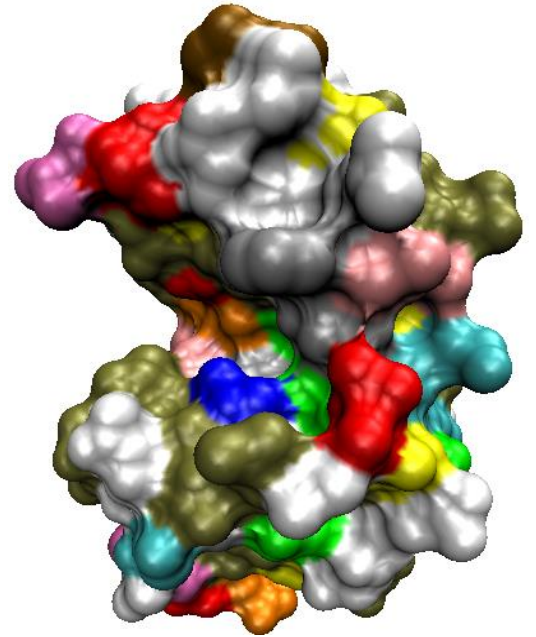
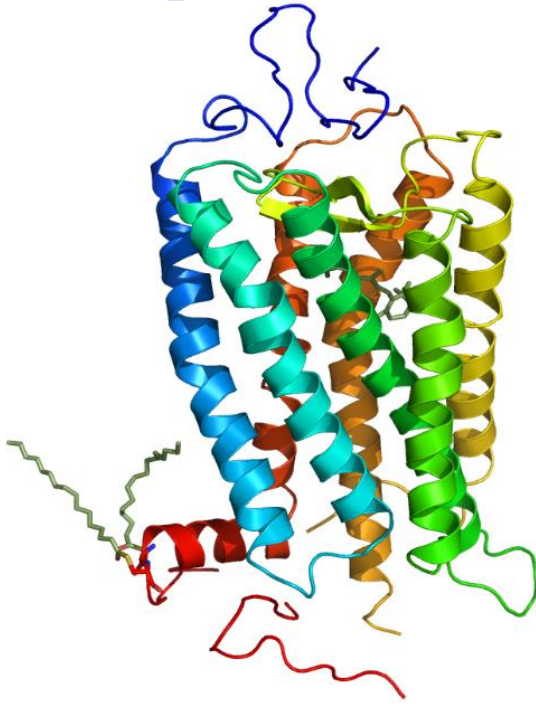
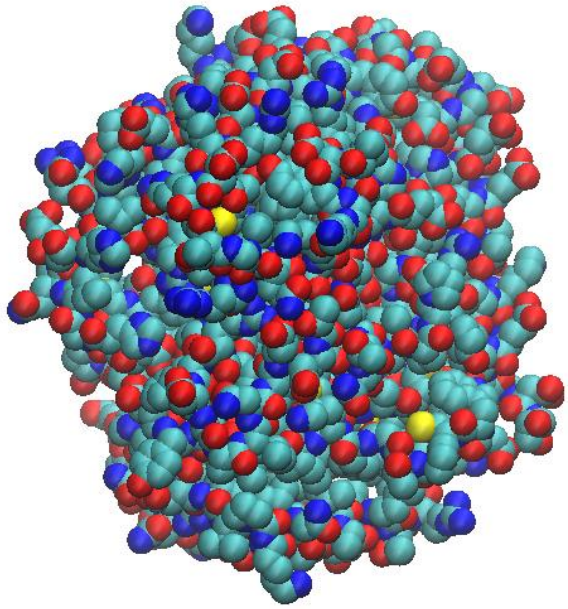
GenBank



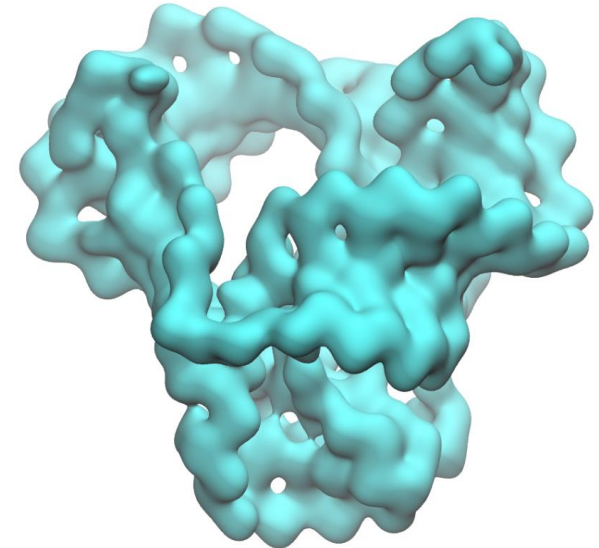
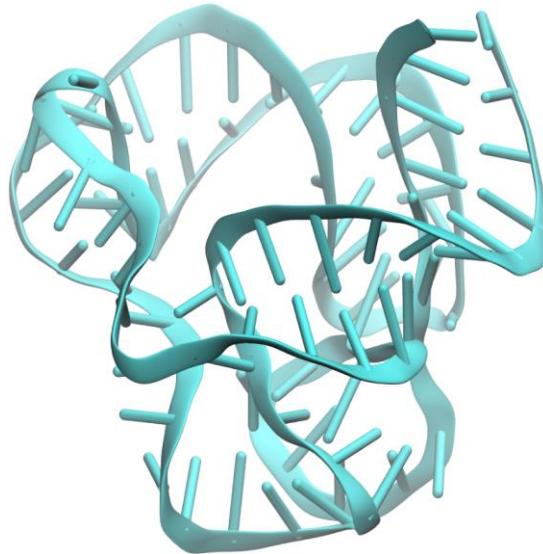
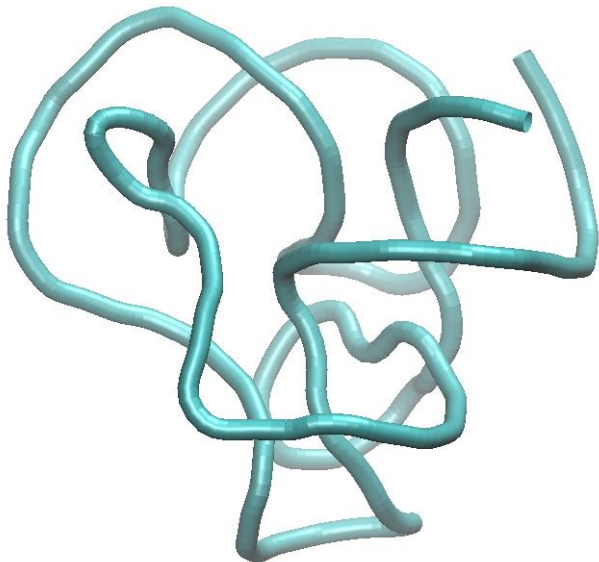
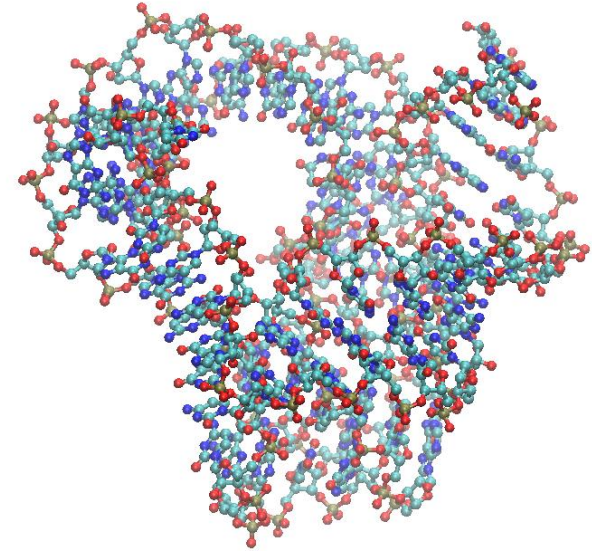
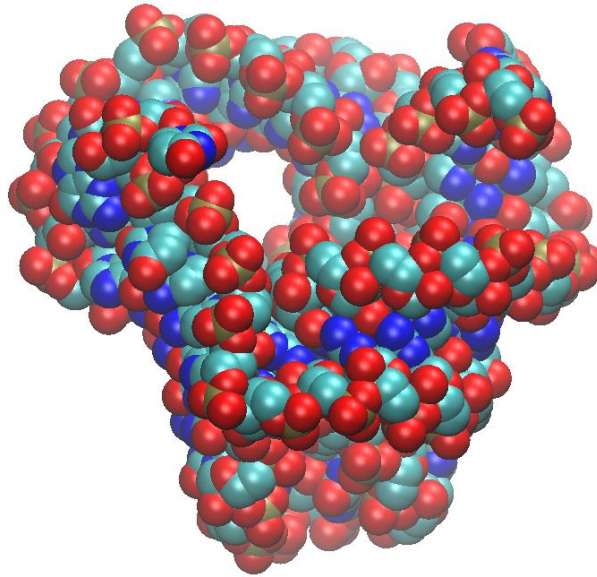
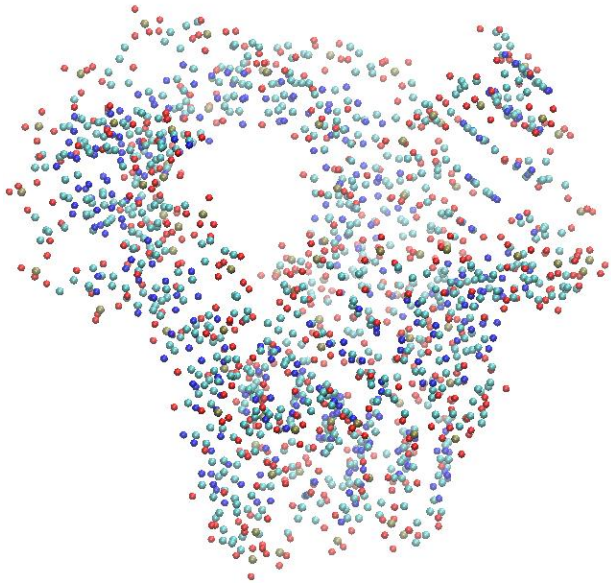
DNA sequencing cost per genome



All proteins?



Same biomolecule?

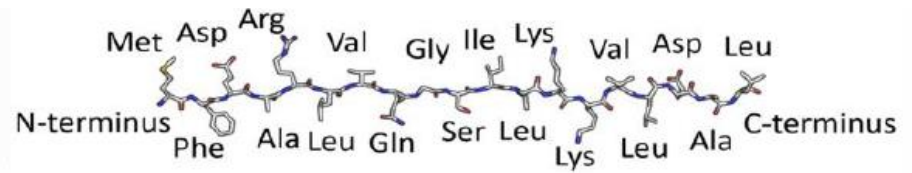


Protein Structure

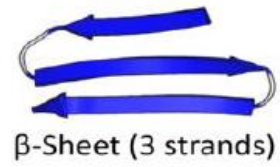
Polypeptides formed from sequences of 20 amino acids.

Primary structure
Secondary structure
Tertiary structure
Quaternary structure

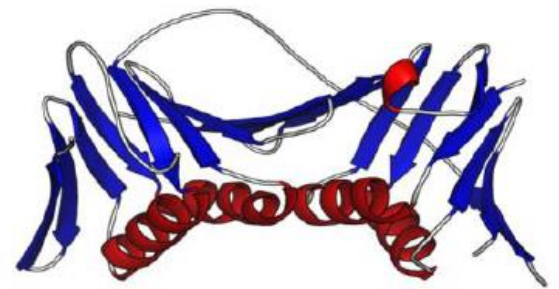
Primary



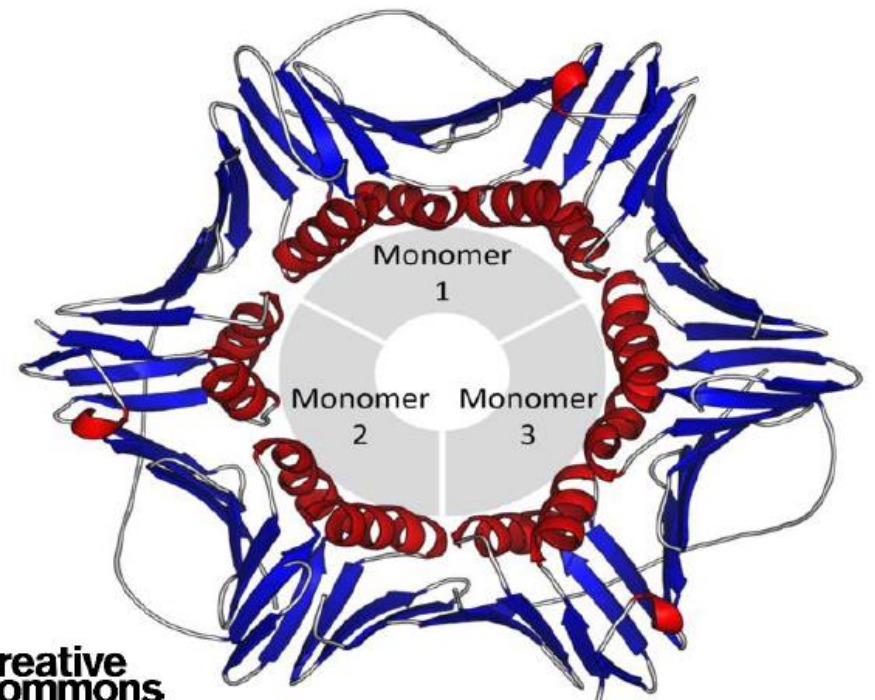
Secondary



Tertiary

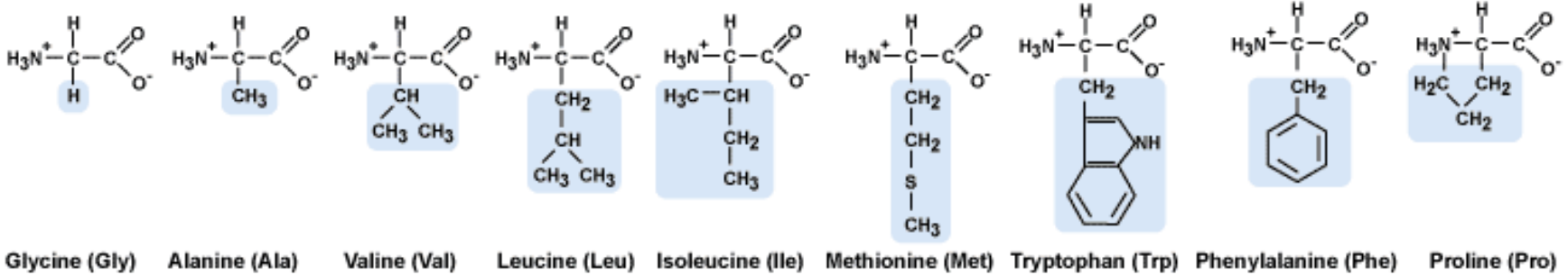


Quaternary

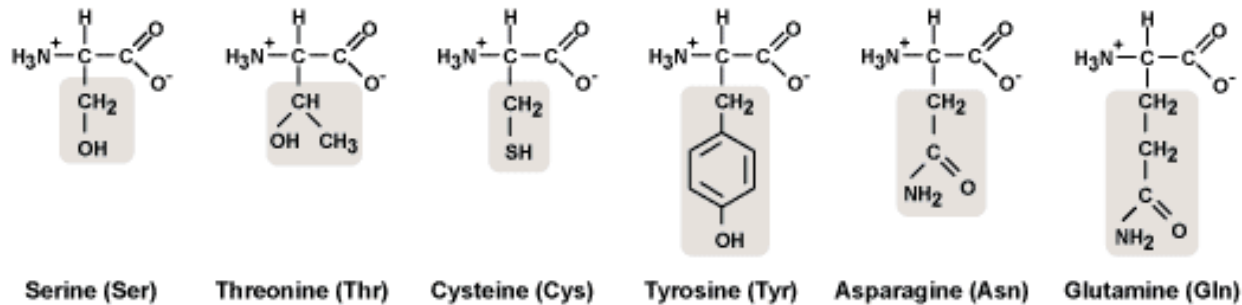


20 common Amino Acids

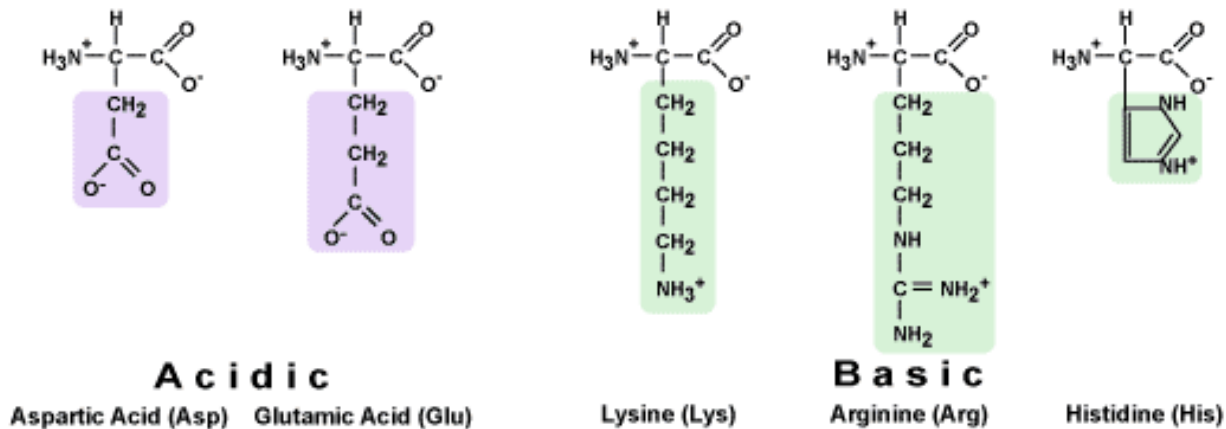
NONPOLAR



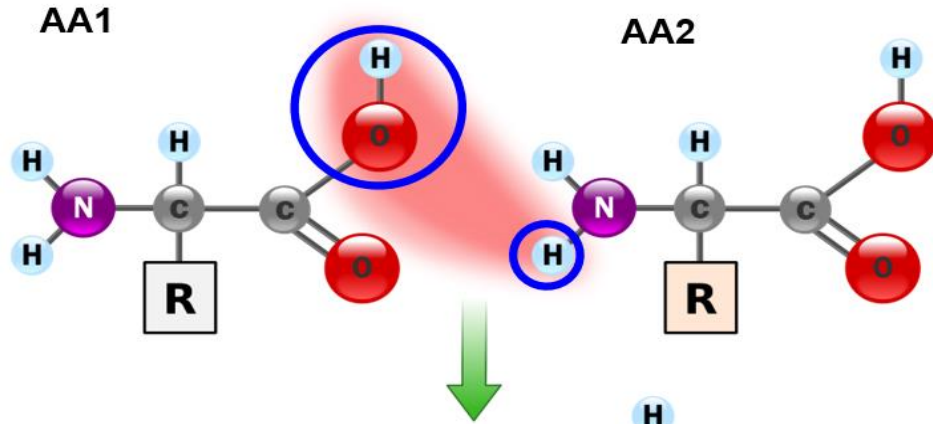
POLAR



Electrically Charged

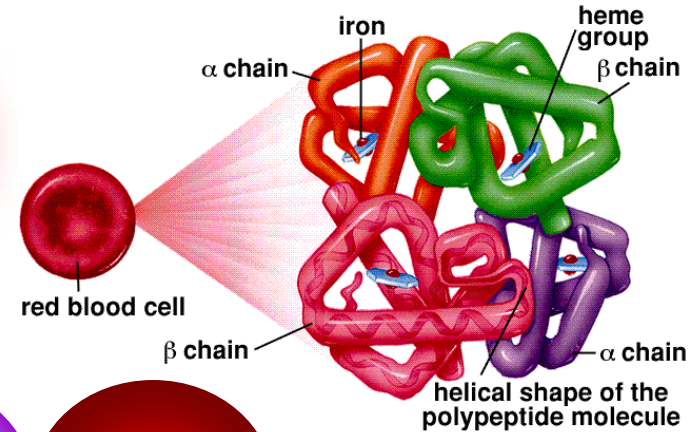
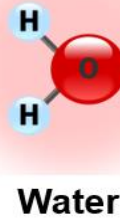
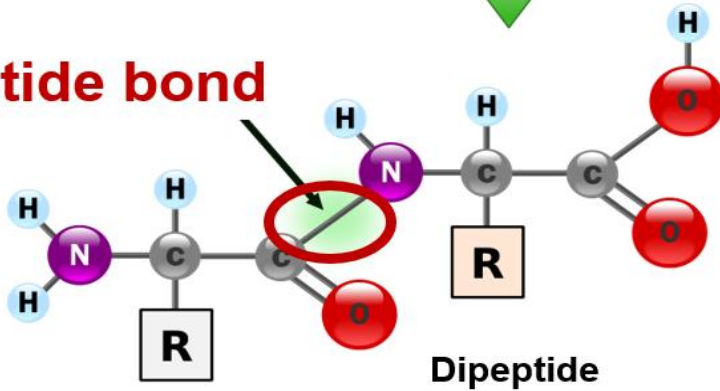


Peptide chain



Peptide bond:
C-N ~100kcal/mol

Peptide bond



Valine

Histadine

Leucine

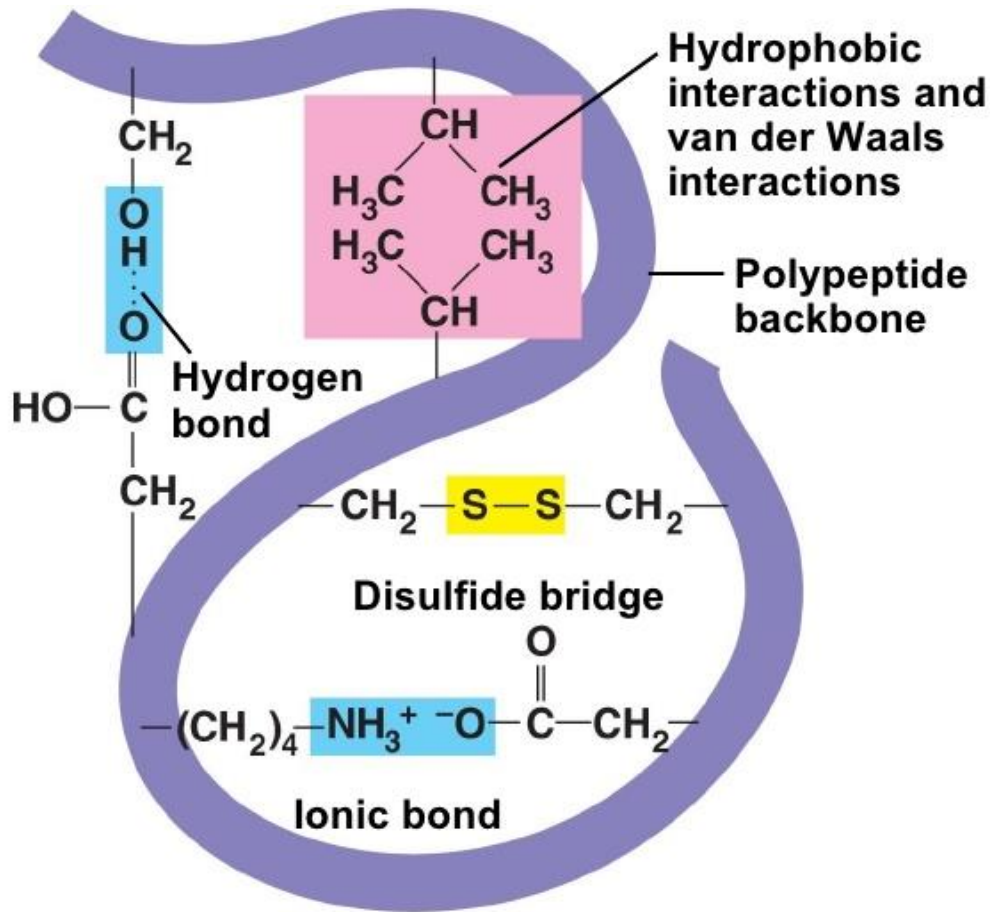
Threonine

Proline

Glutamic
Acid

hemoglobin

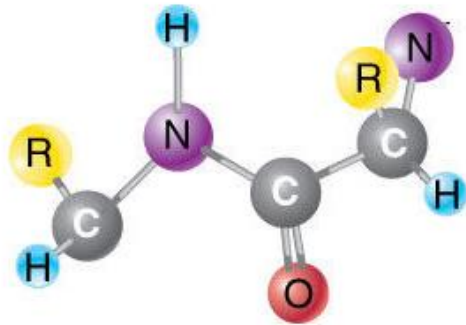
Protein Bond Types



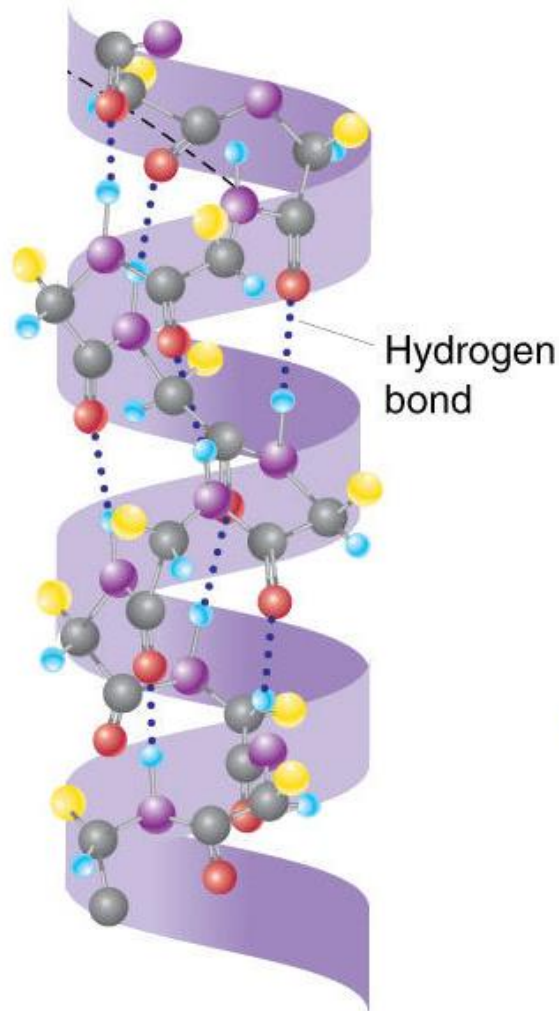
- ❖ **Disulfide bond**
~60kcal/mol
- ❖ **Salt (Ionic) bond**
~20kcal/mol
- ❖ **Hydrogen bond**
~10kcal/mol
- ❖ **Hydrophobic interaction and van der Waals**
~1kcal/mol

Covalent bond: C-C ~100kcal/mol

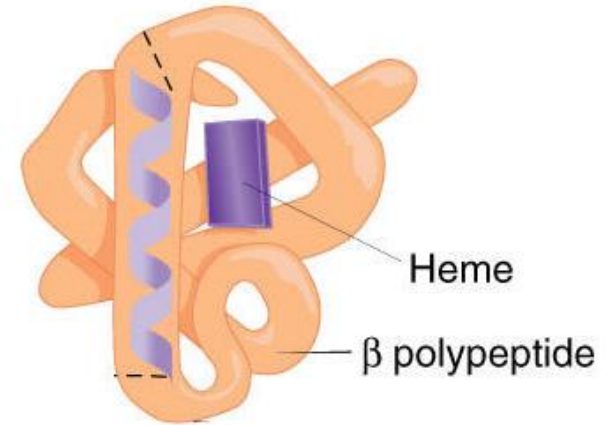
Protein Structures



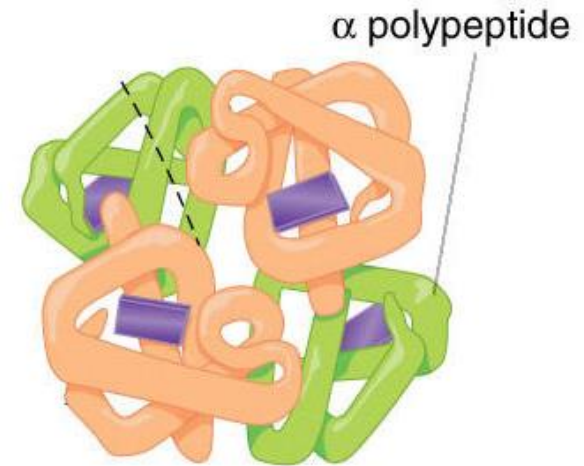
(a) Primary structure



(b) Secondary structure



(c) Tertiary structure



(d) Quaternary structure

Nuclei Acid Structure

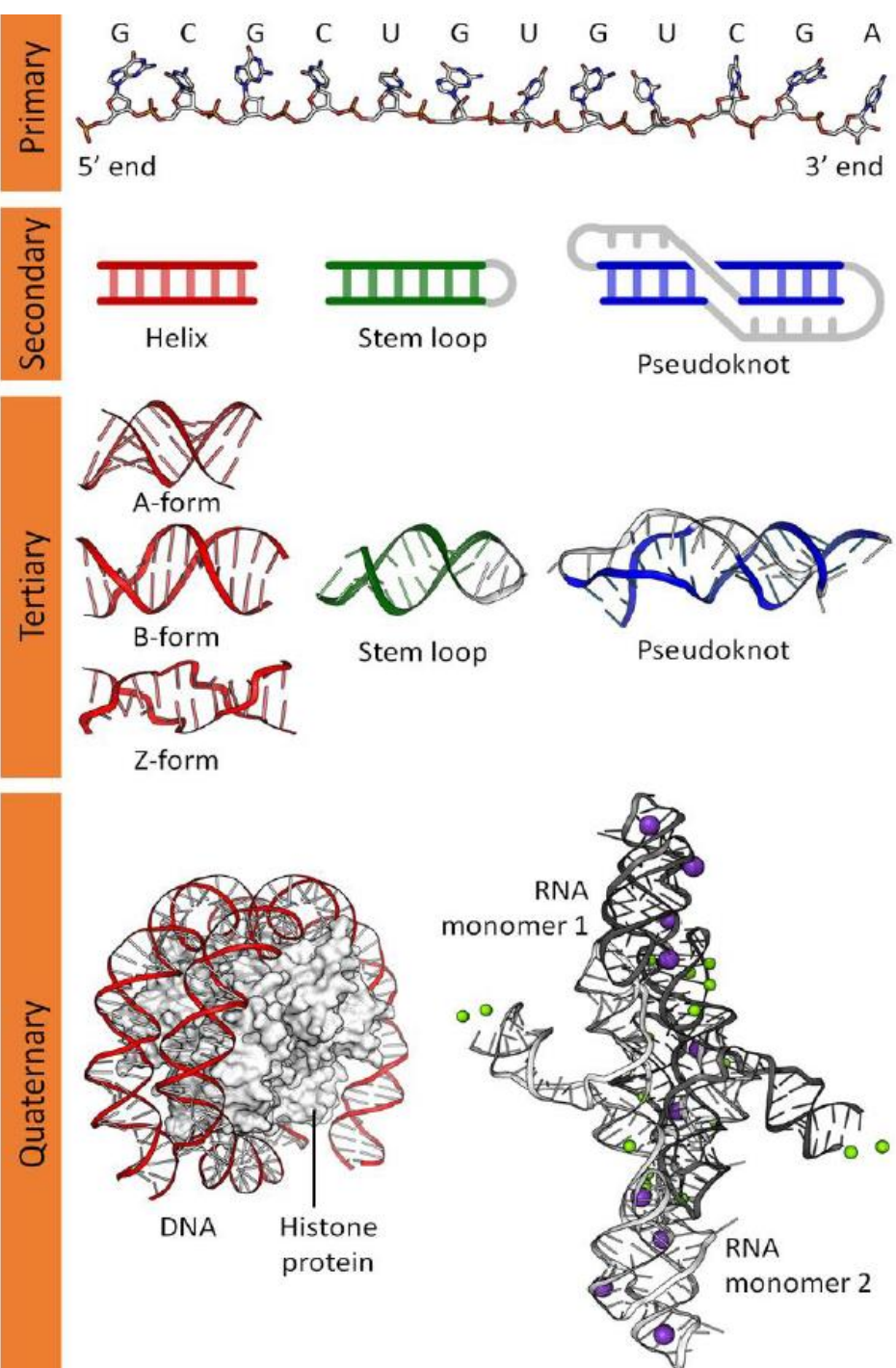
Nitrogenous base (Adenine, Guanine, Cytosine, Thymine (in DNA), Uracil (in RNA))

5-carbon sugar called deoxyribose (found in DNA) and ribose (found in RNA).

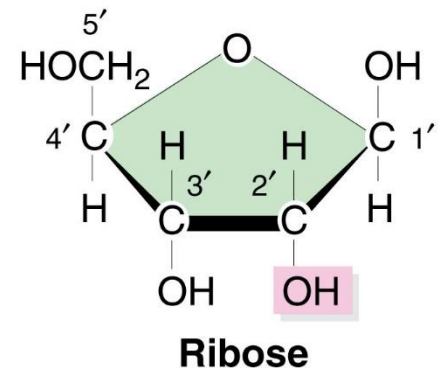
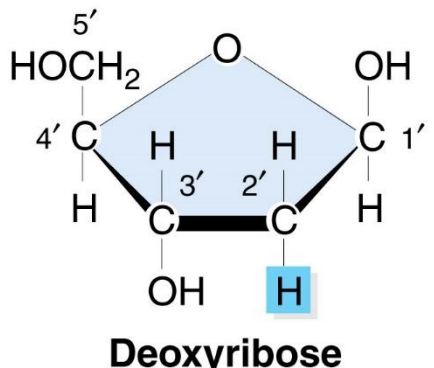
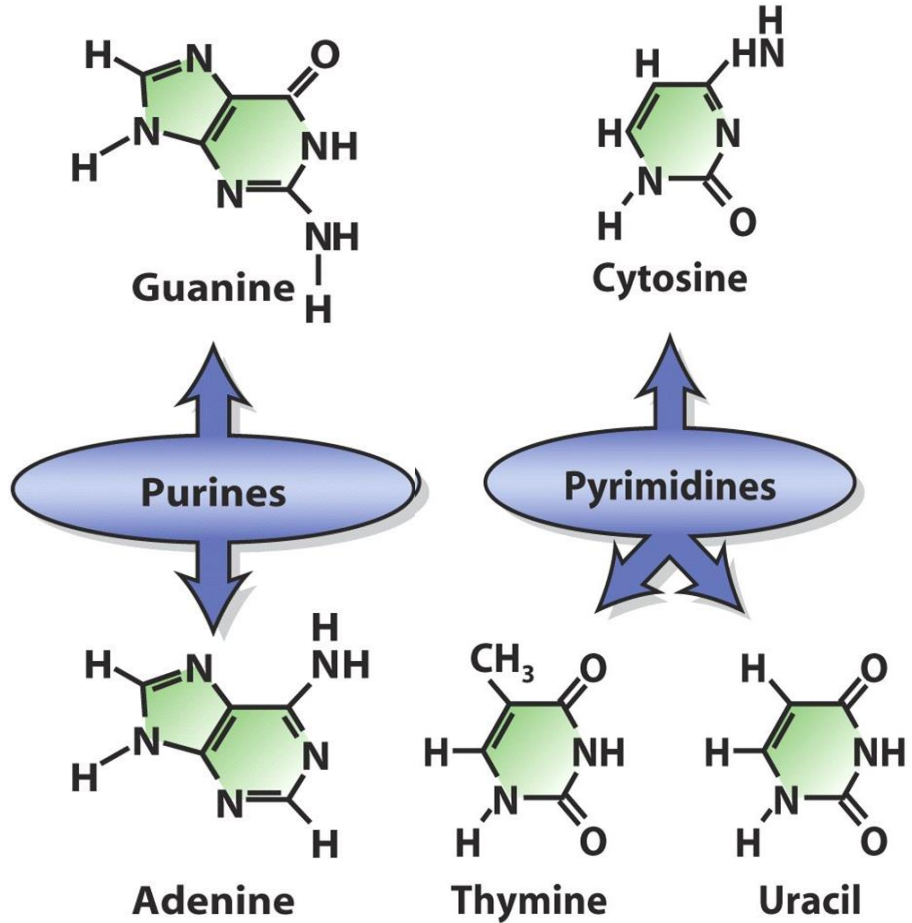
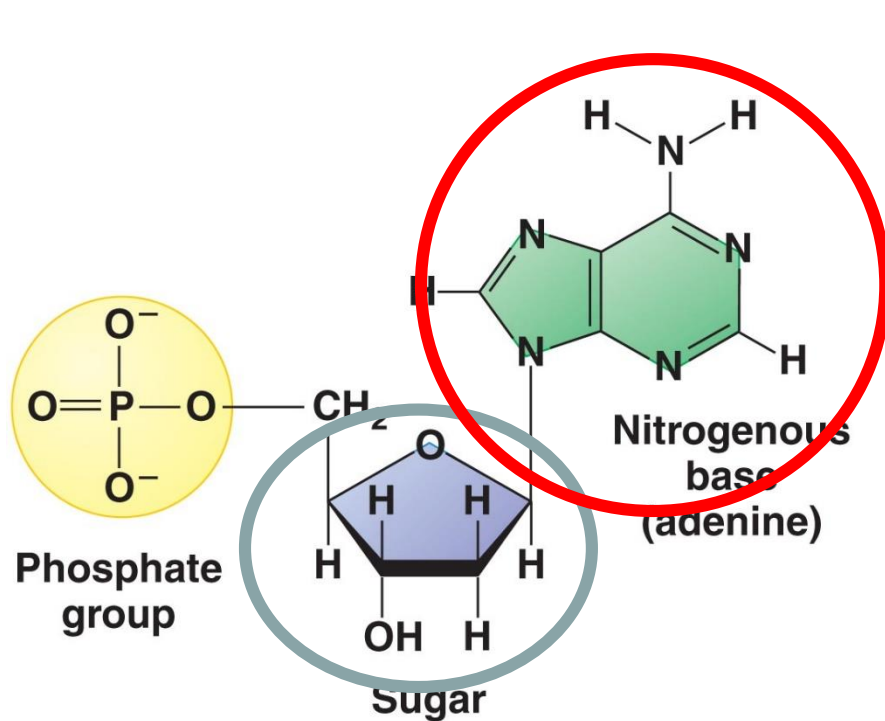
One or more phosphate groups.

Primary structure
Secondary structure
Tertiary structure
Quaternary structure

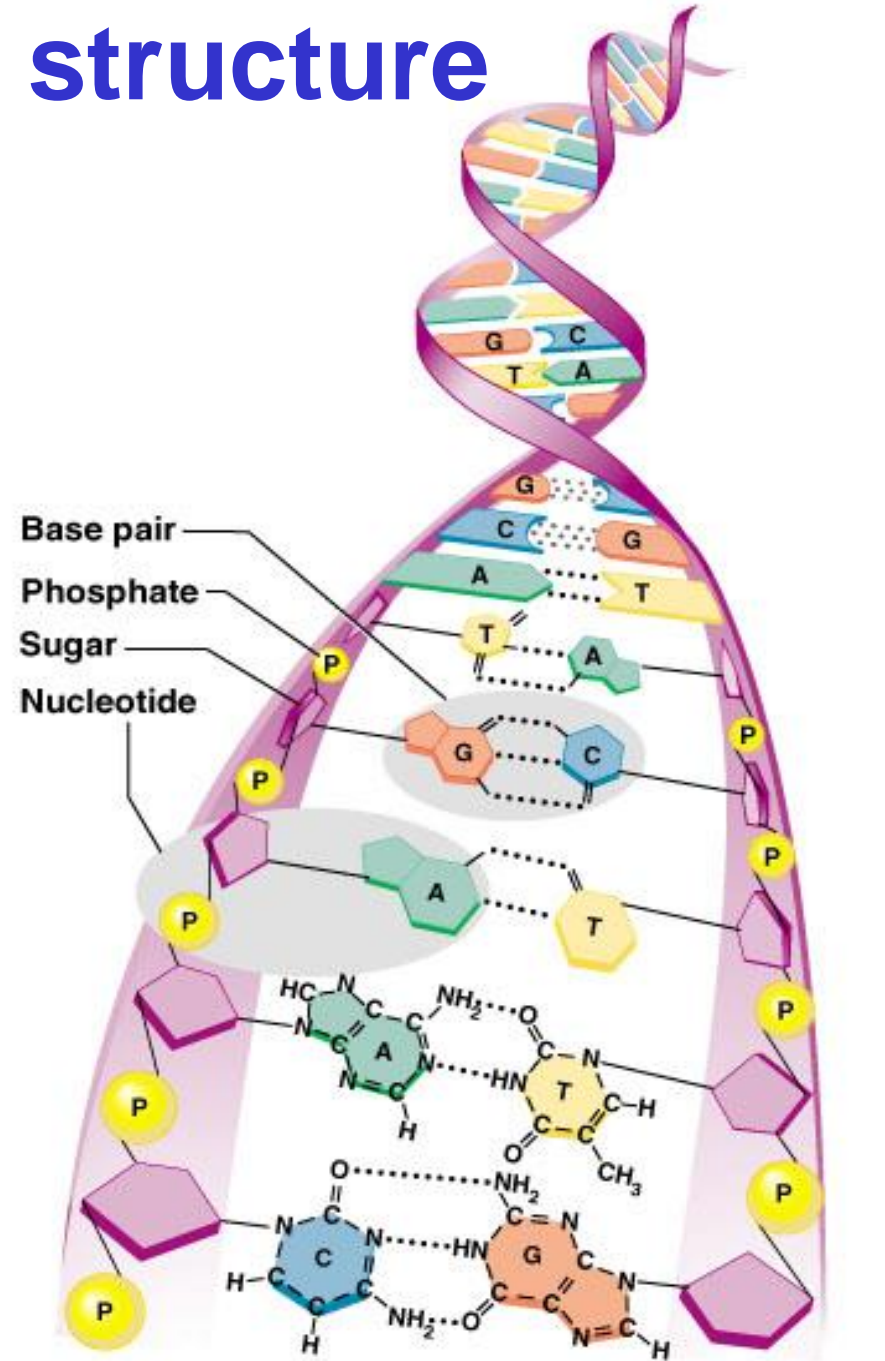
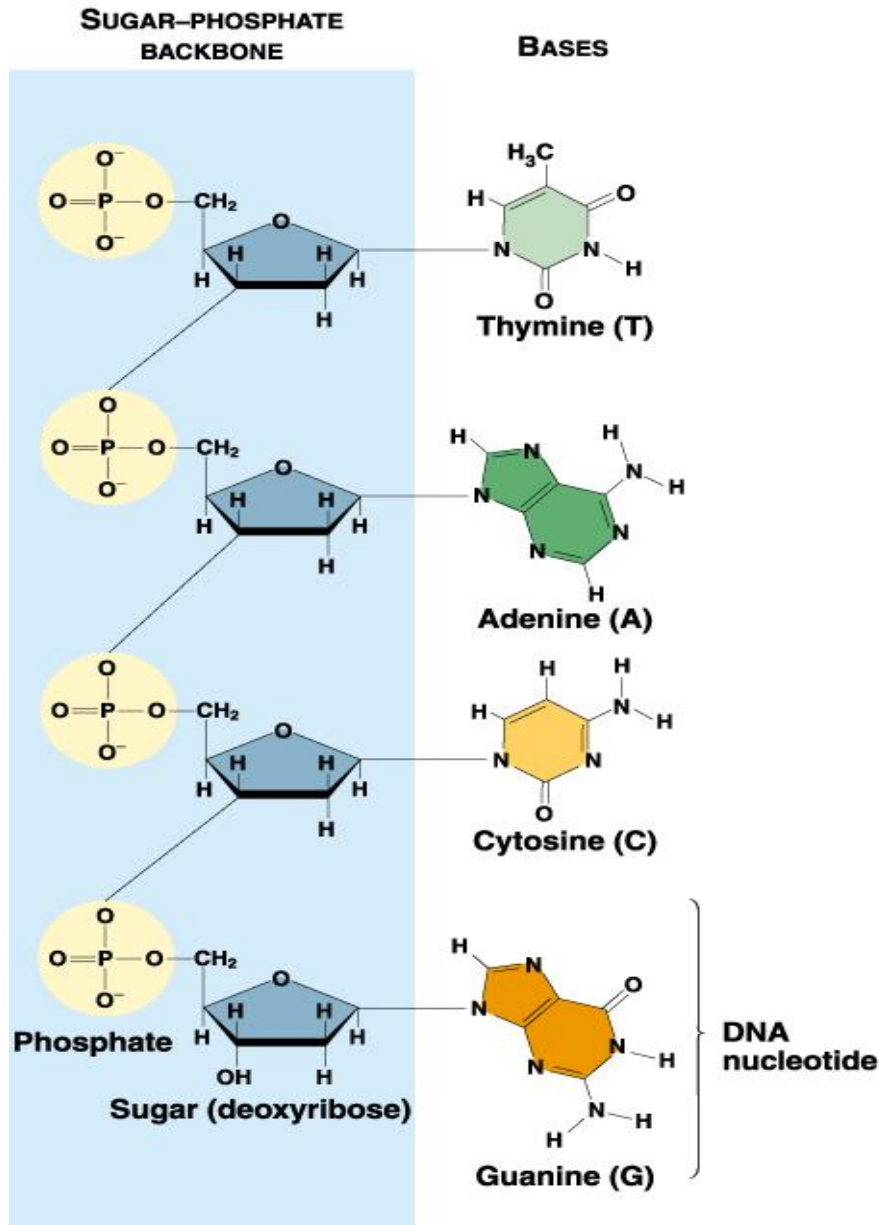
[Image credit: Thomas Shafee](#)



Nuclei acid structure



Nuclei acid structure



Experimental tools and data

- **Experimental tools**
 - X-Ray Crystallography
 - NMR Spectroscopy
 - Cryogenic electron microscopy
- **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

index	name	resname	chain	resid	x	y	z			segname	
ATOM	22	N	ALA	B	3	-4.073	-7.587	-2.708	1.00	0.00	BH
ATOM	23	HN	ALA	B	3	-3.813	-6.675	-3.125	1.00	0.00	BH
ATOM	24	CA	ALA	B	3	-4.615	-7.557	-1.309	1.00	0.00	BH
ATOM	25	HA	ALA	B	3	-4.323	-8.453	-0.704	1.00	0.00	BH
ATOM	26	CB	ALA	B	3	-4.137	-6.277	-0.676	1.00	0.00	BH
ATOM	27	HB1	ALA	B	3	-3.128	-5.950	-0.907	1.00	0.00	BH
ATOM	28	HB2	ALA	B	3	-4.724	-5.439	-1.015	1.00	0.00	BH
ATOM	29	HB3	ALA	B	3	-4.360	-6.338	0.393	1.00	0.00	BH
ATOM	30	C	ALA	B	3	-6.187	-7.538	-1.357	1.00	0.00	BH
ATOM	31	O	ALA	B	3	-6.854	-6.553	-1.264	1.00	0.00	BH
ATOM	32	N	ALA	B	4	-6.697	-8.715	-1.643	1.00	0.00	BH
ATOM	33	HN	ALA	B	4	-6.023	-9.463	-1.751	1.00	0.00	BH
ATOM	34	CA	ALA	B	4	-8.105	-9.096	-1.934	1.00	0.00	BH
ATOM	35	HA	ALA	B	4	-8.287	-8.878	-3.003	1.00	0.00	BH
ATOM	36	CB	ALA	B	4	-8.214	-10.604	-1.704	1.00	0.00	BH
ATOM	37	HB1	ALA	B	4	-7.493	-11.205	-2.379	1.00	0.00	BH
ATOM	38	HB2	ALA	B	4	-8.016	-10.861	-0.665	1.00	0.00	BH
ATOM	39	HB3	ALA	B	4	-9.245	-10.914	-1.986	1.00	0.00	BH
ATOM	40	C	ALA	B	4	-9.226	-8.438	-1.091	1.00	0.00	BH
ATOM	41	O	ALA	B	4	-10.207	-7.958	-1.667	1.00	0.00	BH

000

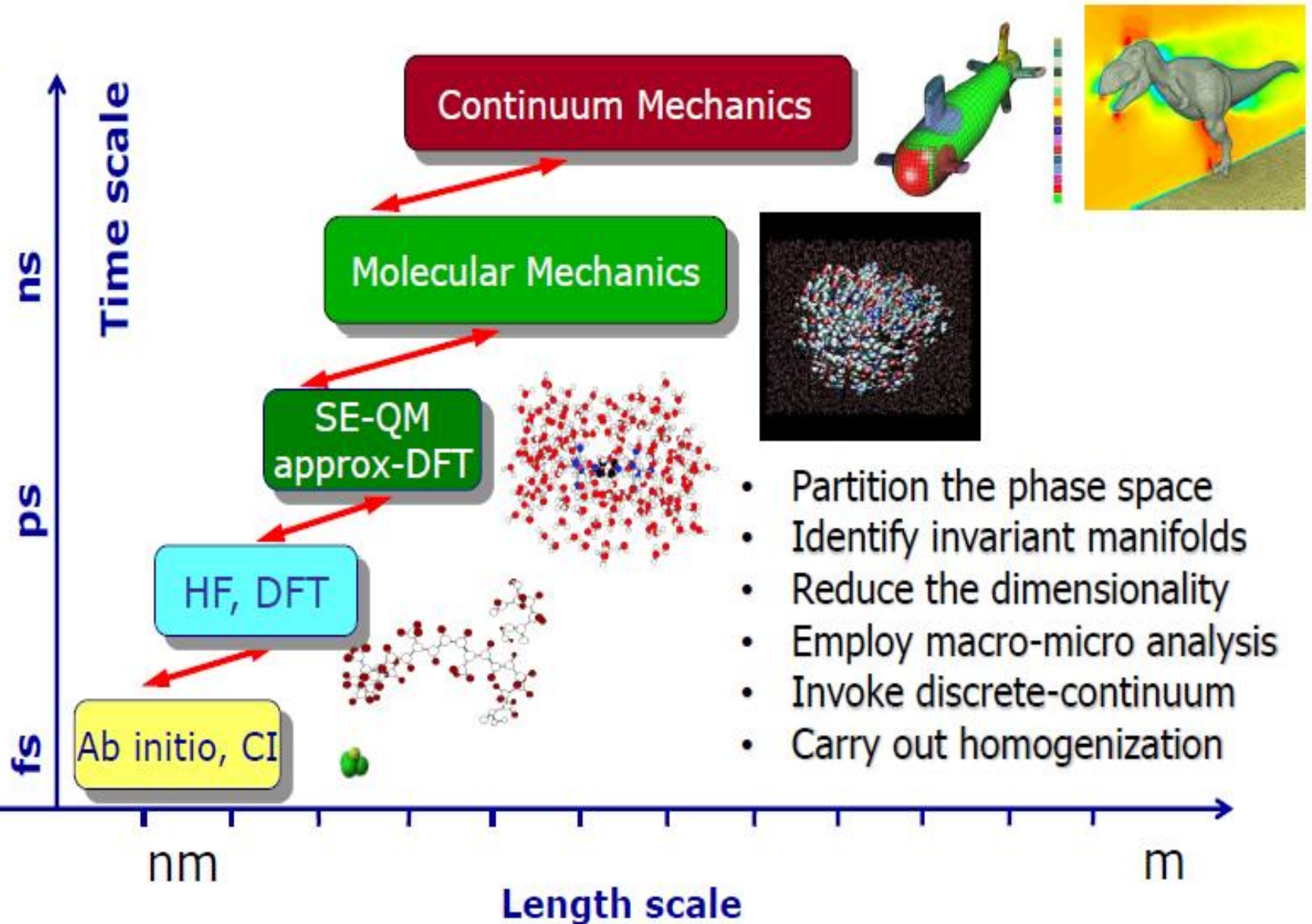
10 20 30 40 50 60 70

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

Multiscale Coupling



Molecular Mechanics (MM)

Using classical particle assumption.

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

$$\begin{aligned} \text{Approximation: } U = & \sum_{\text{bonds}} K_d (d - d_0)^2 + \sum_{\text{angle}} K_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{dihedrals}} K_\chi (1 + \cos(n\chi - \delta)) \\ & + \sum_{\text{nonbond}} \left\{ \epsilon_{ij} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - \left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon r_{ij}} \right\} \end{aligned}$$

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

Foundations of Biophysics

Continuum Mechanics (CM)

Hydrodynamics (HD)

Electrodynamics (ED)

Thermodynamics (TD)

Molecular Mechanics (MM)

Kinetic Theory (KT)

Statistical Mechanics (SM)

Quantum Mechanics (QM)

↑
Computability

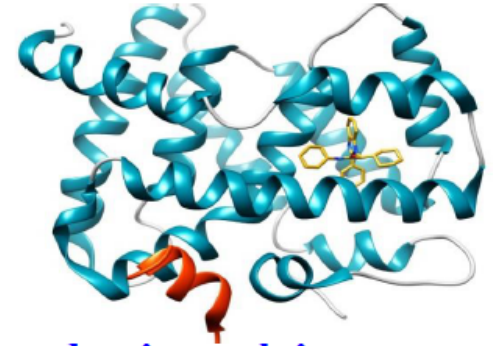
↓
Reliability

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
NPDSL DLVCS



- 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.

ALPHA FOLD

Protein Sequence

SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLG

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

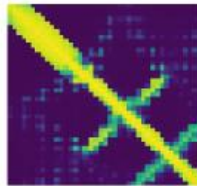


Neural Network

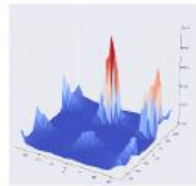


Databases

Distance Predictions



Angle Predictions



Score



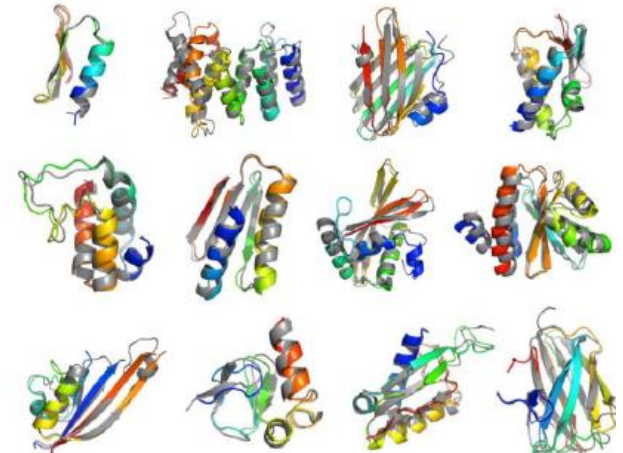
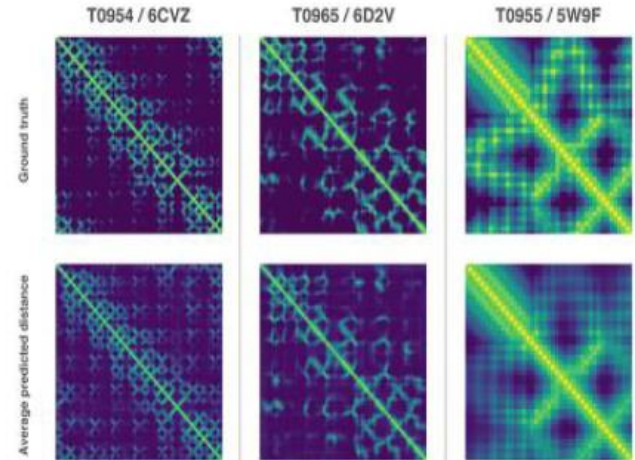
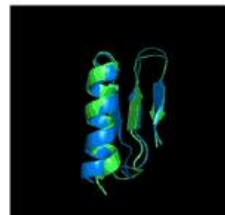
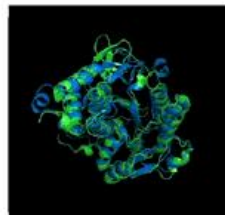
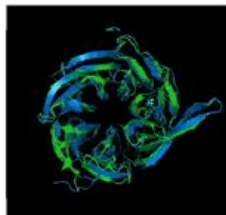
Gradient Descent

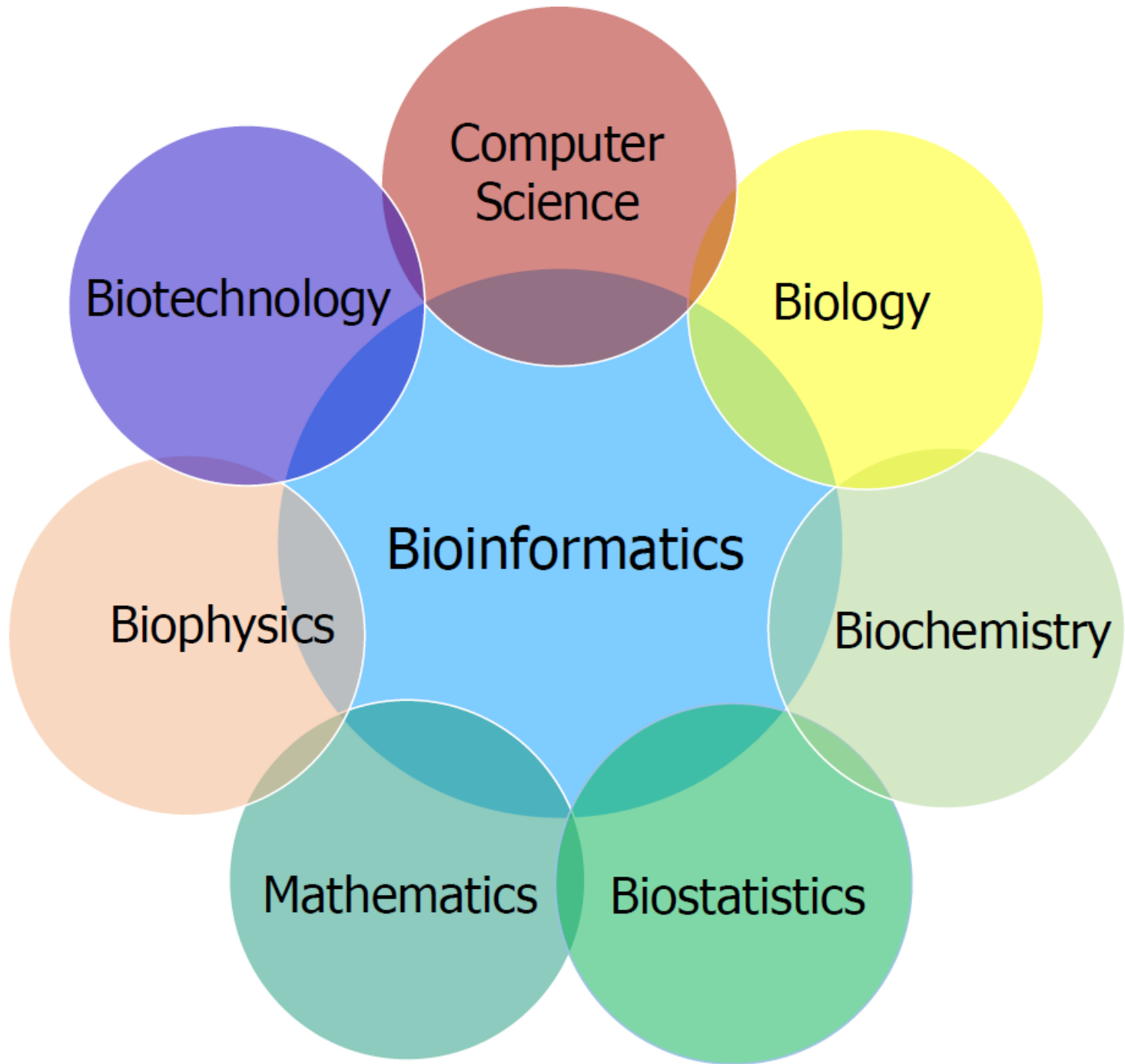
T0954 / 6CVZ

T0965 / 6D2V

T0955 / 5W9F

Structures:
Ground truth (green)
Predicted (blue)

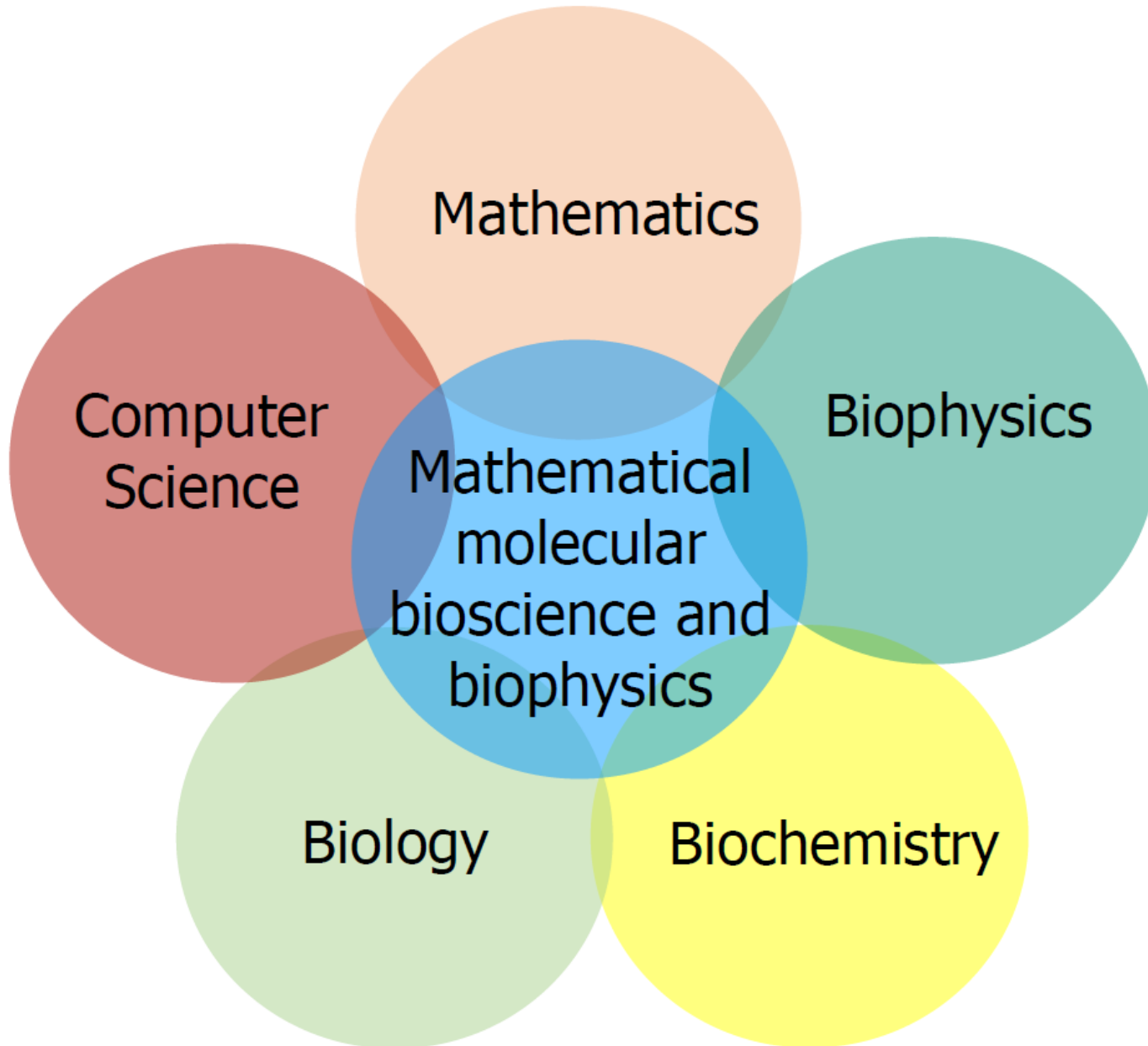




New Trends in Biological Science

- 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

D.W. Sumners, Isabel .K. Darcy , Mariel Vazquez, Dorothy Buck, Tamar Schlick, Erica Flapan, Christian Reidys, Yusu Wang, Peter Rogen, Jack Quine,

Topology

Algebra

Christine Heitsch, David Murrugarra, Reidun Twarock
Natasha Jonoska, R Brijder, HJ Hooeboom,
Julie Mitchell

Group theory

Combinatorics

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, ...

Part 2-1: Topological modeling of biomolecules

Graph theory for molecular bioscience

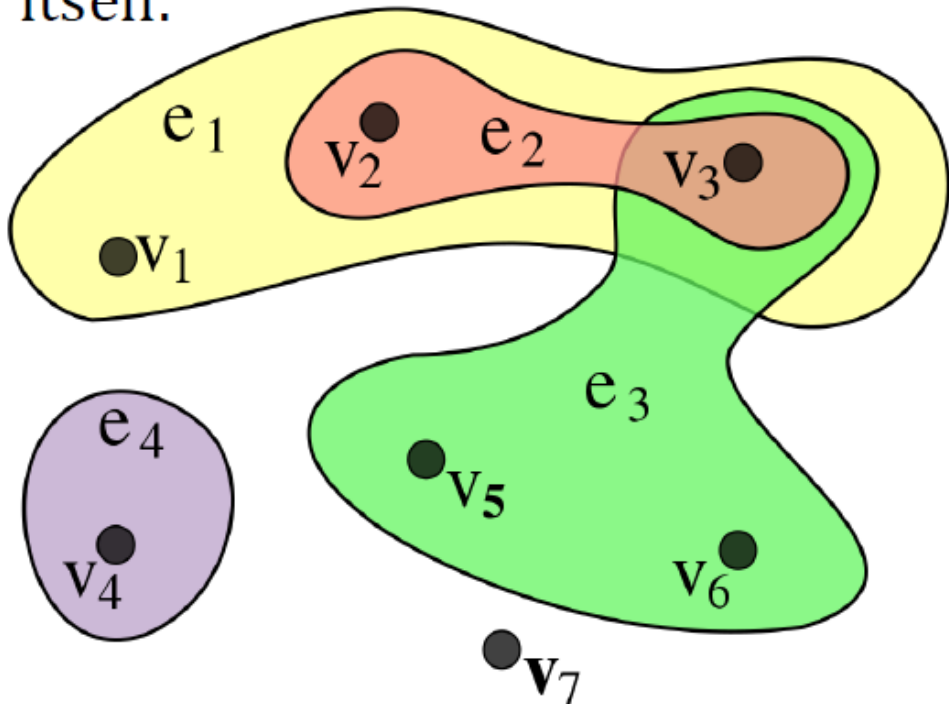
- Structural stability and flexibility analysis
- Surface modeling
- Visualization
- Biomolecular domain analysis and hinge detection
- Entropy estimation
- Modeling of a wide range of biomolecular interactions
- Prediction of a wide variety of chemical and biological properties, including binding affinity,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

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.



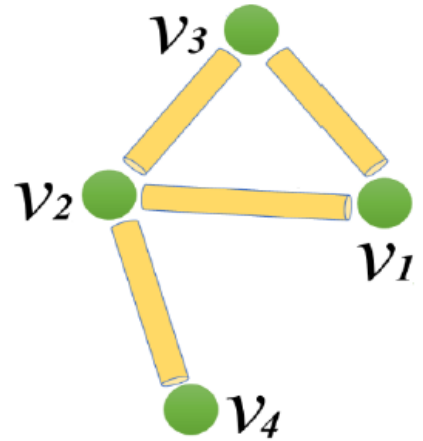
Algebraic graph Theory

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

Example: A simple (undirected) graph:

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

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



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

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

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

$$L = D - A$$

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

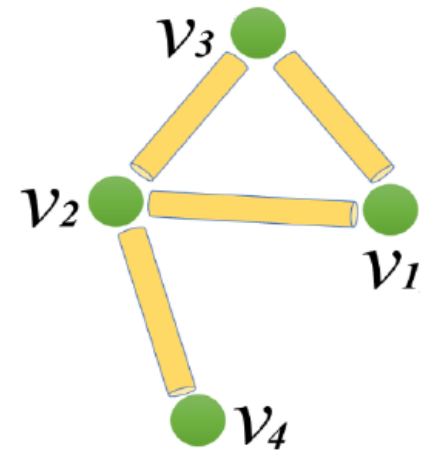
Degree of node i

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:

- $0 \leq \lambda_1^L \leq \lambda_2^L \dots 0 \leq \lambda_{\text{Max}}^L$
- $\lambda_1^L = 0$ 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 $\lambda_1^L \leq \lambda_2^L \dots$
 $0 \leq \lambda_{\text{Max}}^L$. Then

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

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

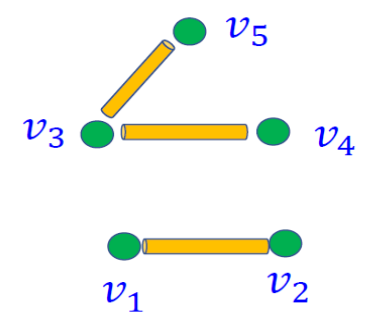
Graph Theory

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:

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


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

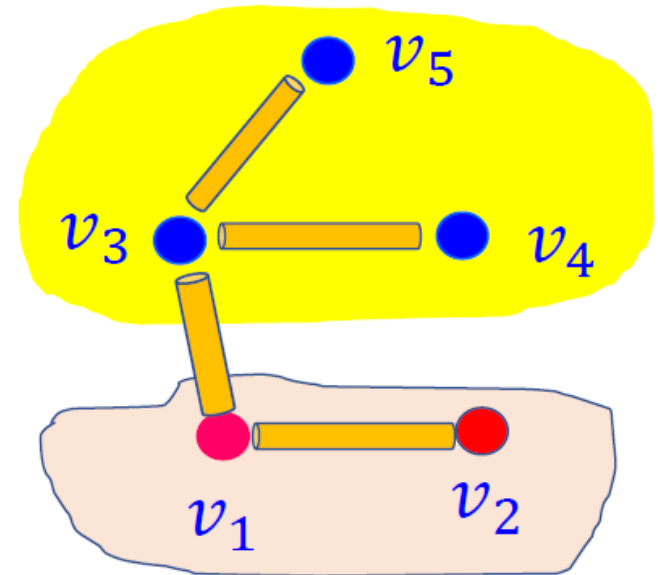
Eig = [0, 0, 1, 2, 3]

Two disconnected components (harmonic part due to the topology)

Spectral bisection

Example II:

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



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

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

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:

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

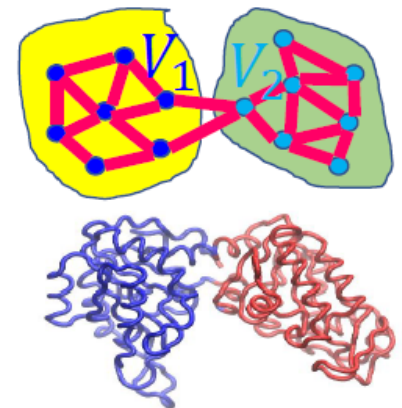
where s_i is a membership variable,

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

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.



Gaussian Network Model (GNM)—Laplacian model

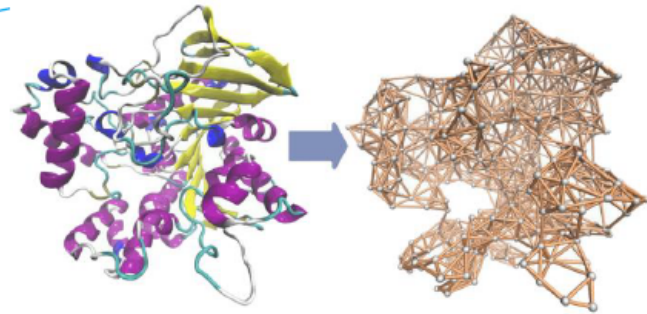
(Bahar, Atilgan and Erman, FD, 1997)

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

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

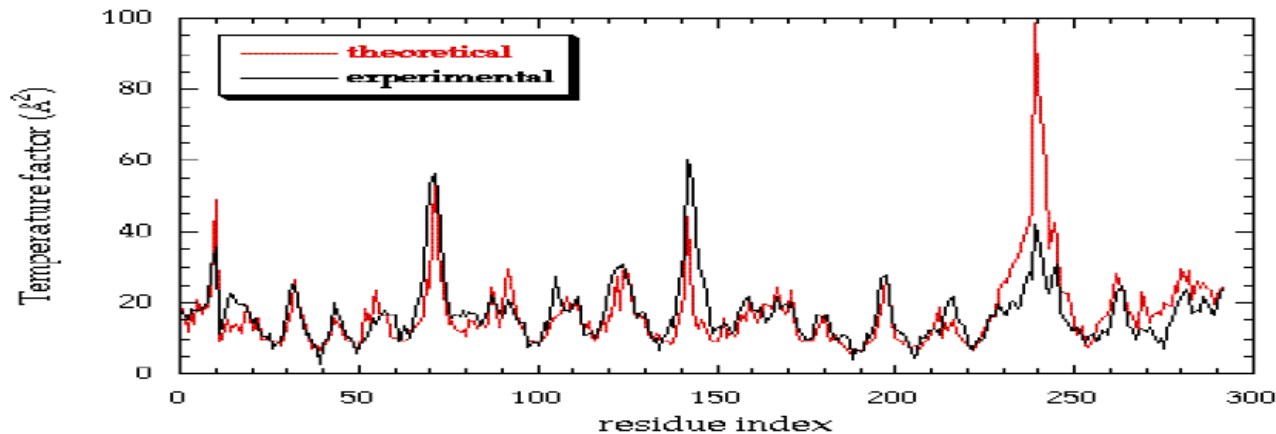
$$(L^{-1})_{jj} = \sum_{k=2}^N \frac{1}{\lambda_k} [u_k u_k^T]_{jj}$$

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

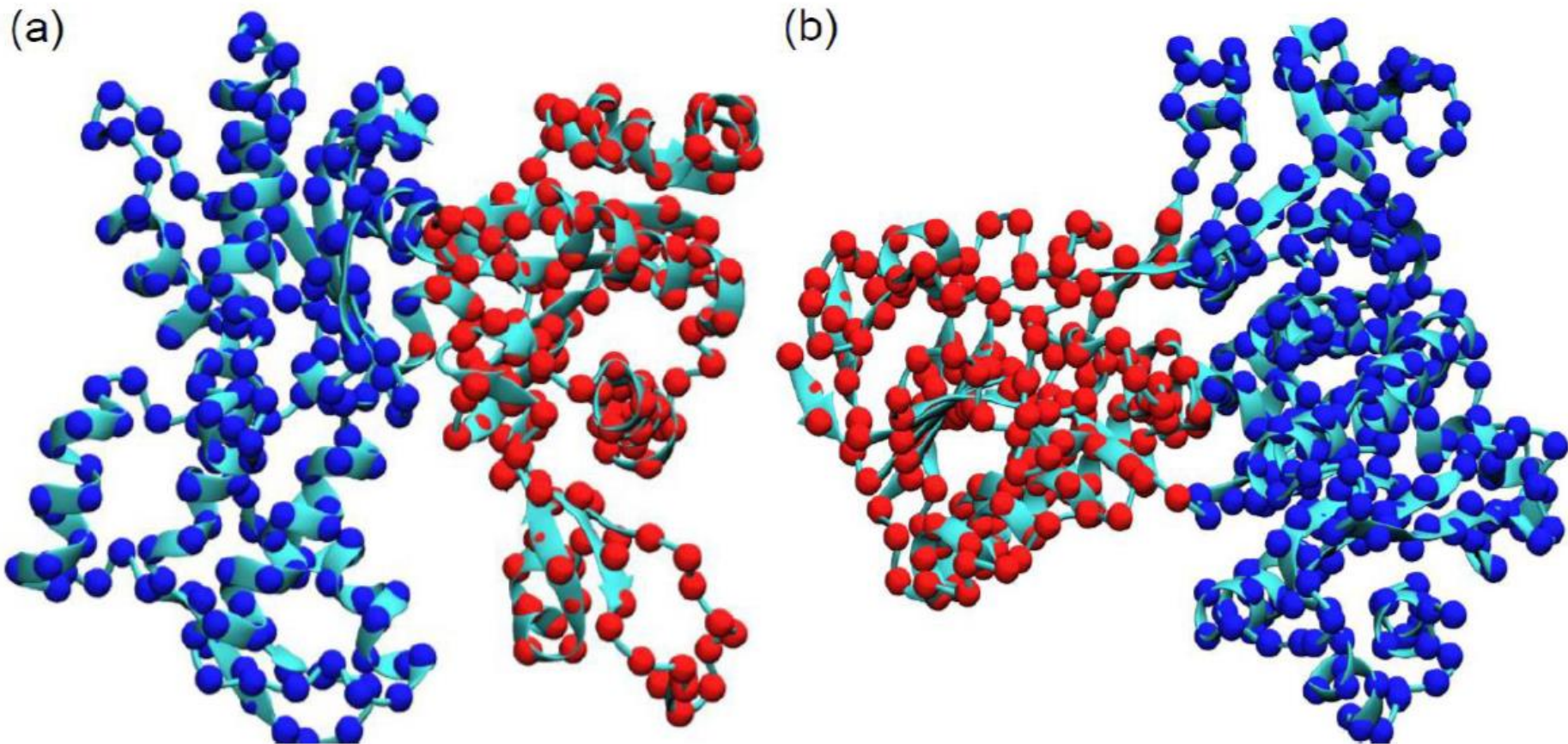


Moore–Penrose pseudoinverse

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

$$r_{ij}^d = |\mathbf{r}_{ij}^d|$$

$$r_{ij} = |\mathbf{r}_{ij}|$$

$$\Delta \mathbf{R} = \{\Delta x_1, \Delta y_1, \Delta z_1, \dots, \Delta x_N, \Delta y_N, \Delta z_N\}$$

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

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

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

Moore–Penrose pseudoinverse:

eigenvector

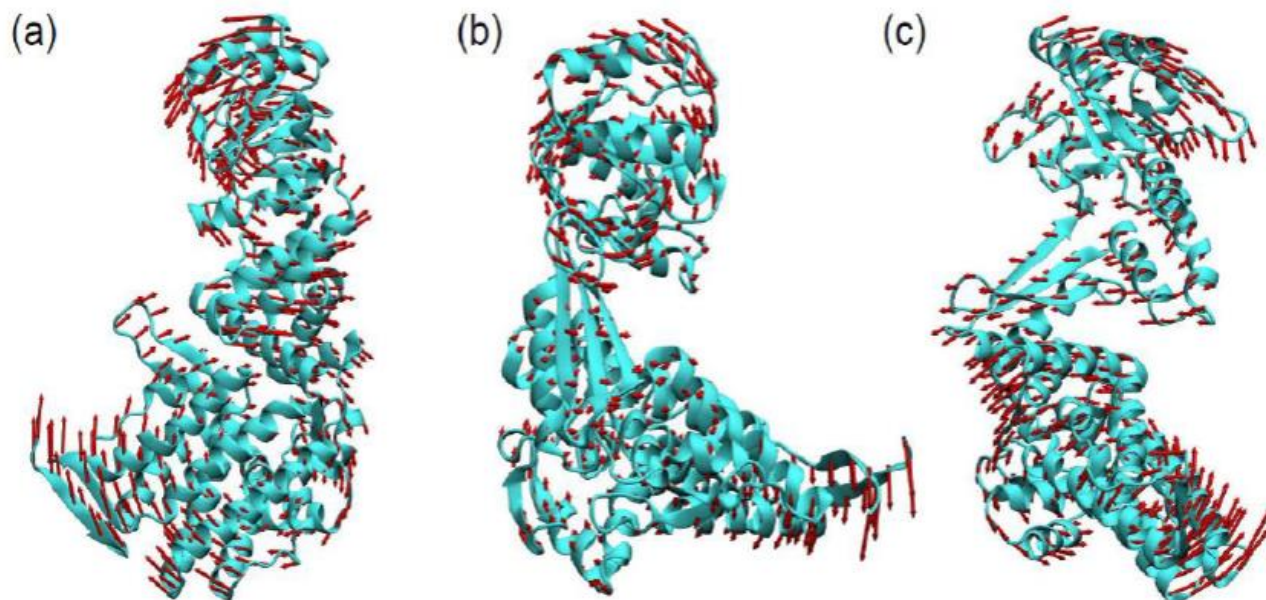
$$(H^{-1})_{ii} = \sum_{k=7}^{3N} \lambda_k^{-1} [\mathbf{v}_k \mathbf{v}_k^T]_{ii}$$

eigenvalue

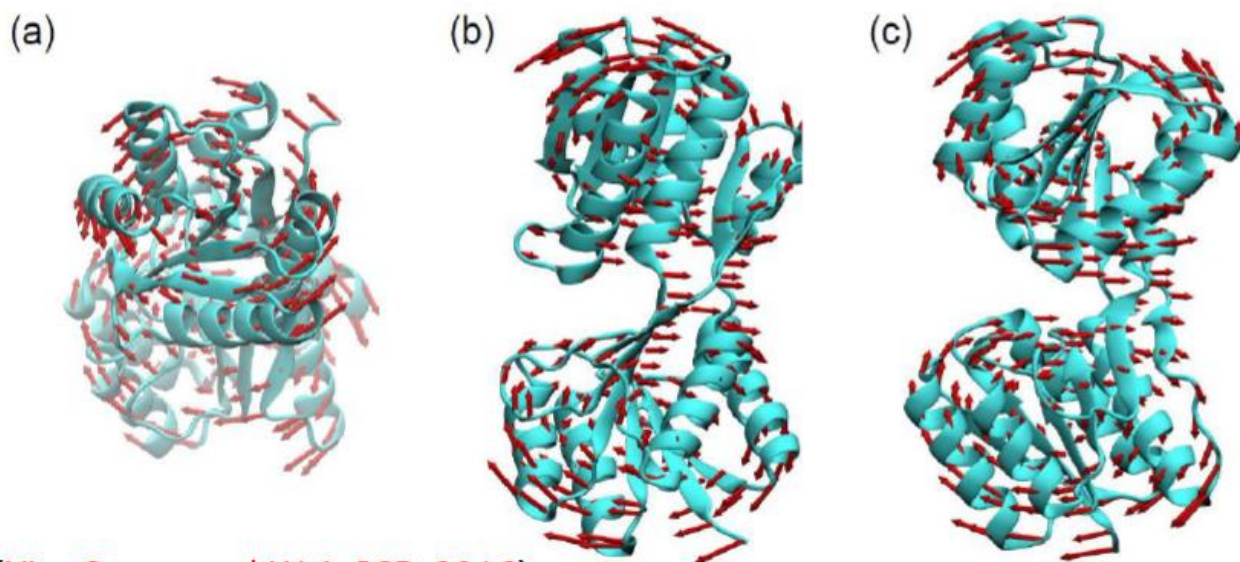
Predicted b-factor:

$$B_i^{\text{ANM}} = \frac{8\pi^2}{3} \sum_{j=3i-2}^{3i} \langle \Delta \mathbf{R}_j \cdot \Delta \mathbf{R}_j \rangle, \quad \forall i = 1, 2, \dots, N.$$

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

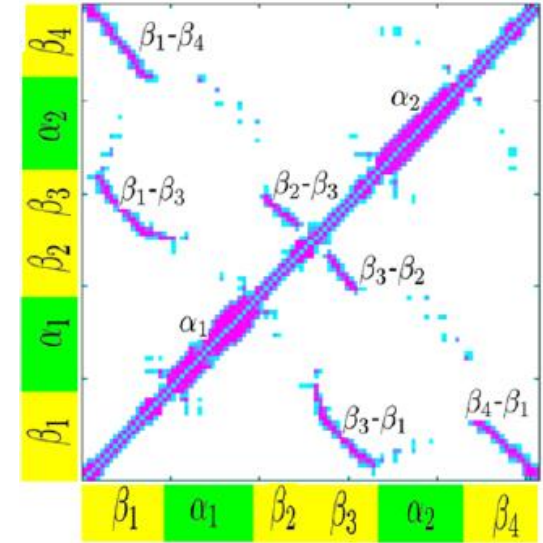
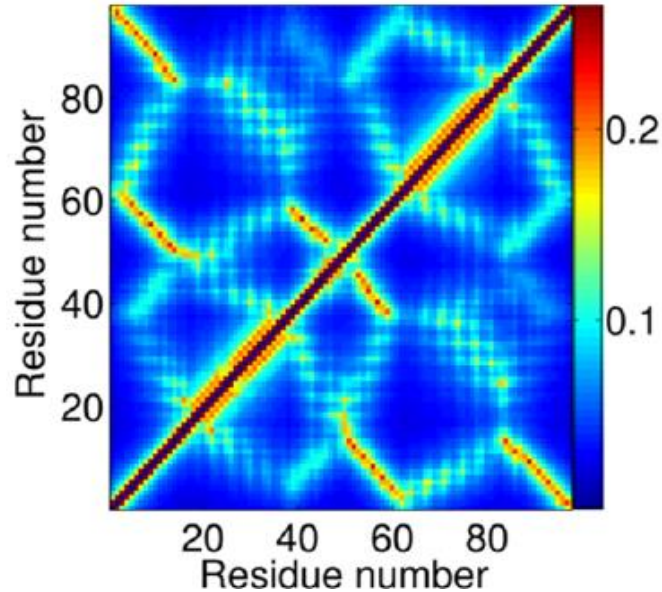
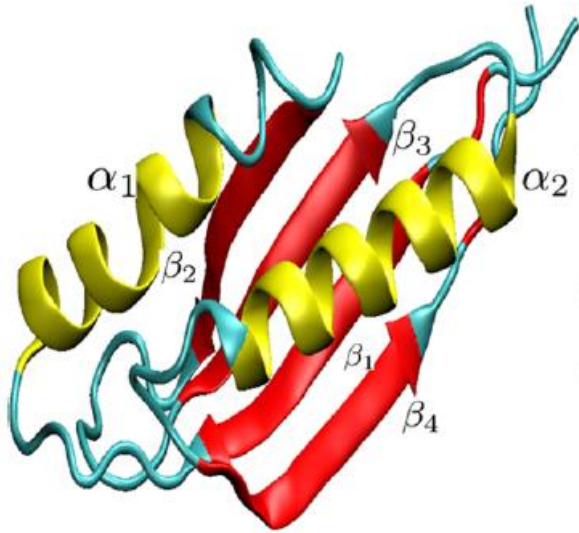


The motions of 1GRU (chain A). The 7th, 8th, and 9th nANM modes are demonstrated in (a)–(c), respectively



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

Geometry to topology mapping



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:

$$\phi(\|r_i - r_j\|; \eta) = 1, \text{ as } \|r_i - r_j\| \rightarrow 0$$

$$\phi(\|r_i - r_j\|; \eta) = 0, \text{ as } \|r_i - r_j\| \rightarrow \infty$$

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

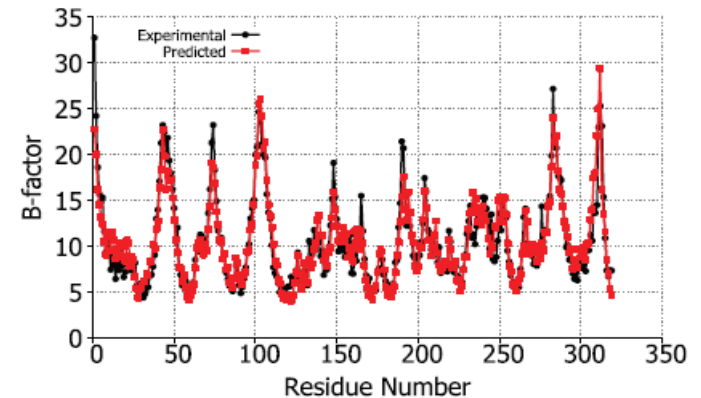
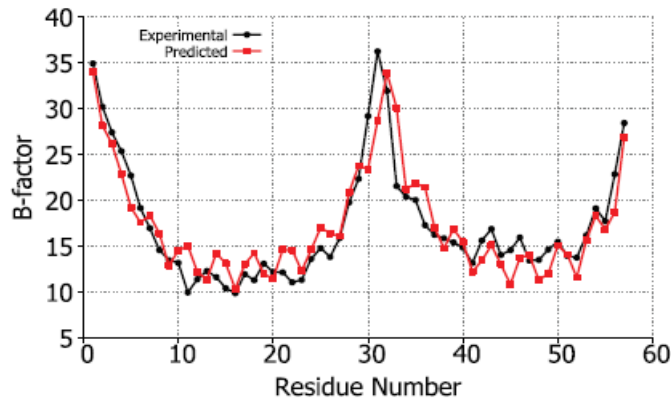
Generalized Lorentz: $\phi(\|r_i - r_j\|; \eta) = \frac{1}{1 + (\|r_i - r_j\|/\eta)^\nu}$

Flexibility rigidity index (FRI)

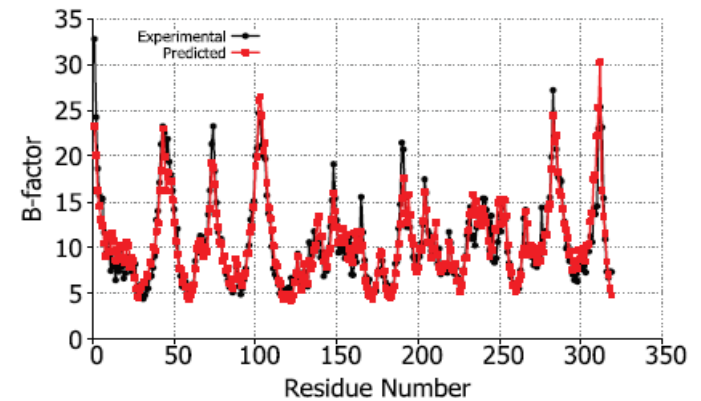
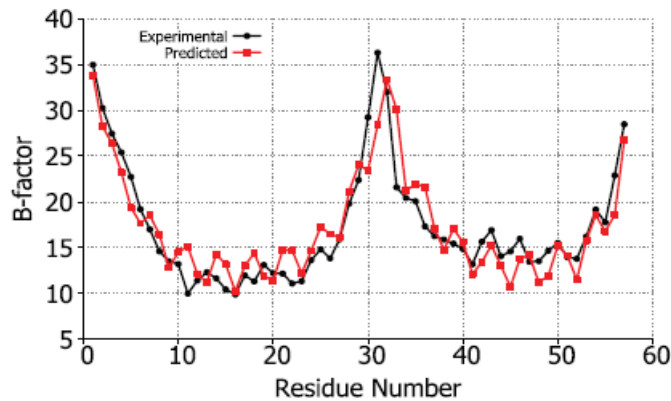
Rigidity index:
$$\mu_i = \sum_{j \neq i} w_j \phi(\|r_i - r_j\|; \eta)$$

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**



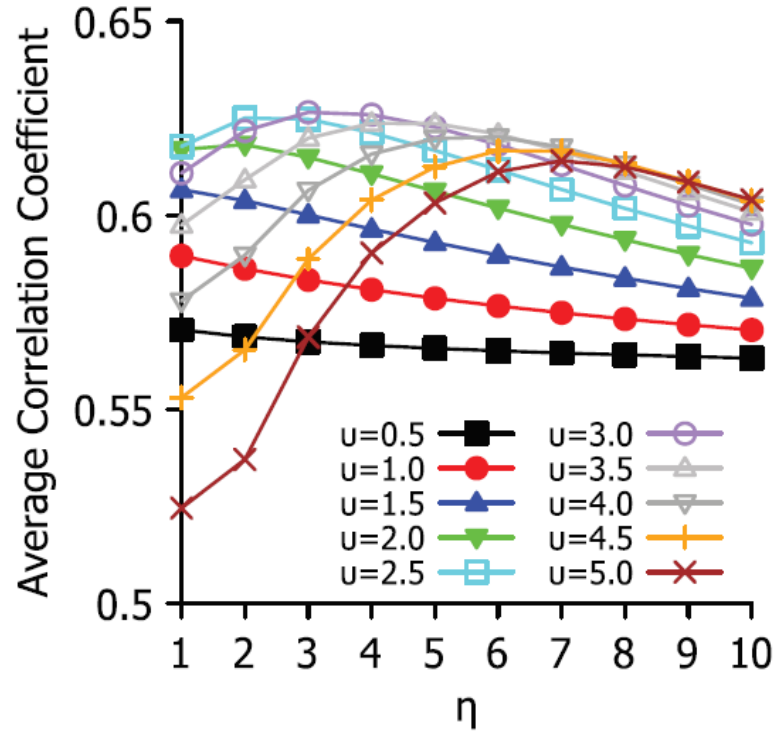
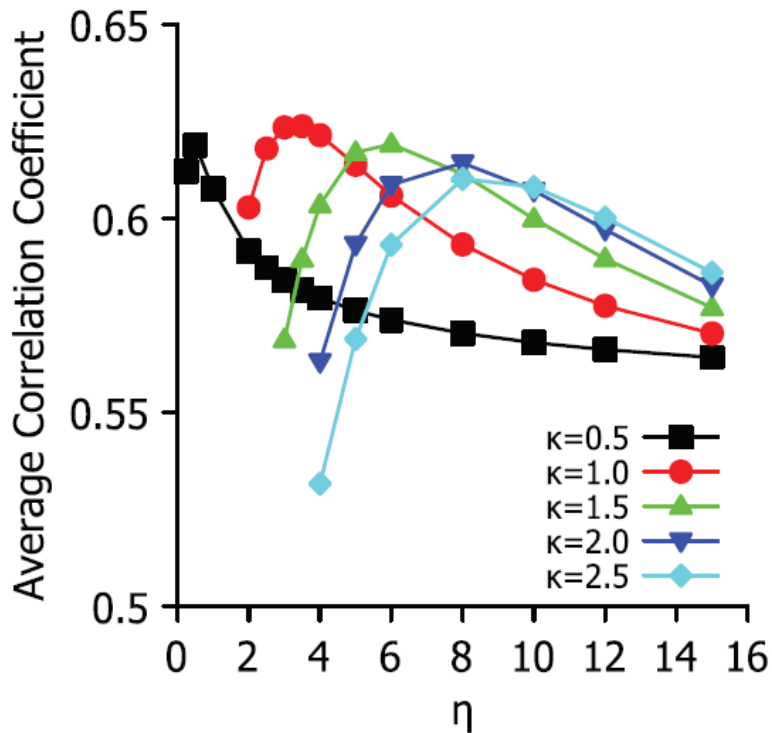
Protein ID: 1DF4

Protein ID: 2Y7L

Parameter testing

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

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



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

Performance of our FRI

Accuracy: (10% improvement)

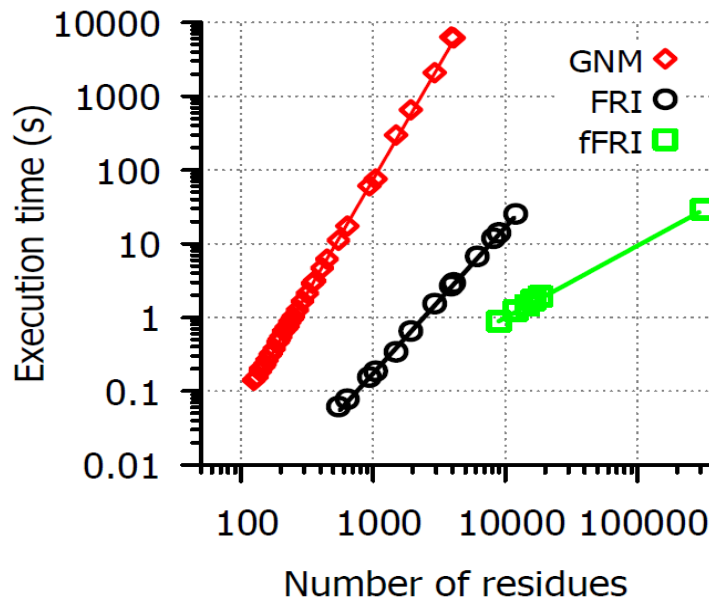
PDB set	pfFRI	GNM	NMA
Small	0.594	0.541	0.480
Medium	0.605	0.550	0.482
Large	0.591	0.529	0.494
Superset	0.626	0.565	NA

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

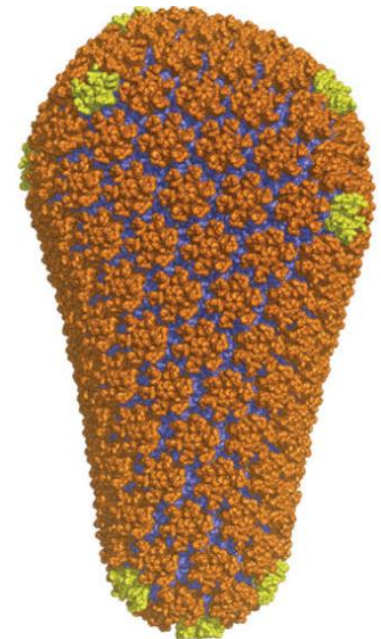
Bertil Halle, PNAS, Vol. 99, No.3, 1274-1279, 2002

Exponential parameters	Avg. CC	Lorentz parameters	Avg. CC
$\kappa=0.5, \eta=0.5$	0.615 (8.8%)	$\nu=2.5, \eta=2.0$	0.622 (10.1%)
$\kappa=1.0, \eta=3.0$	0.623 (10.3%)	$\nu=3.0, \eta=3.0$	0.626 (10.8%)
$\kappa=1.5, \eta=6.0$	0.619 (9.6%)	$\nu=3.5, \eta=4.0$	0.623 (10.3%)

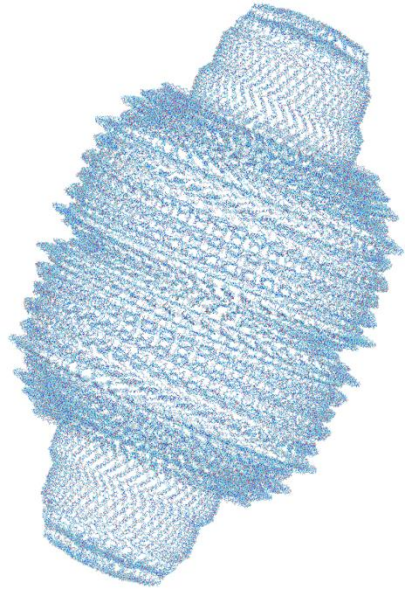
Time: fFRI $O(N)$



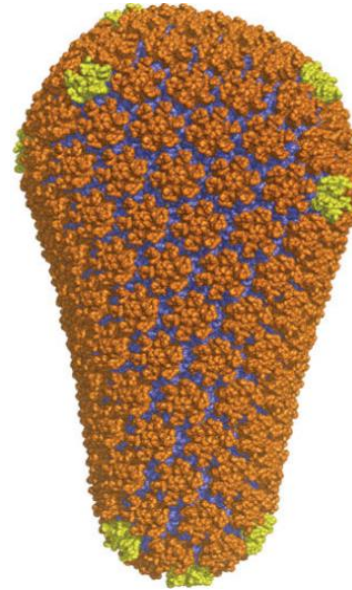
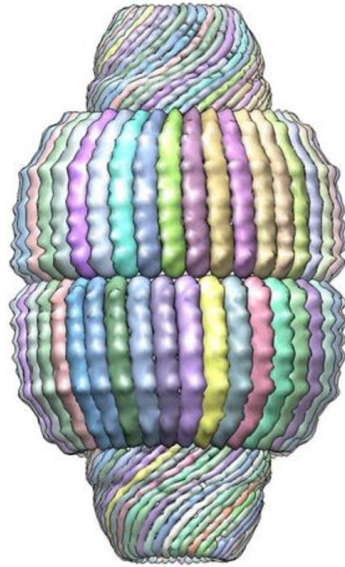
HIV Virus capsid (313 236 residues) in less than 30 seconds



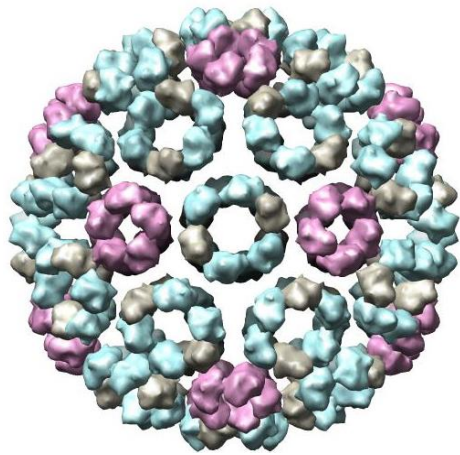
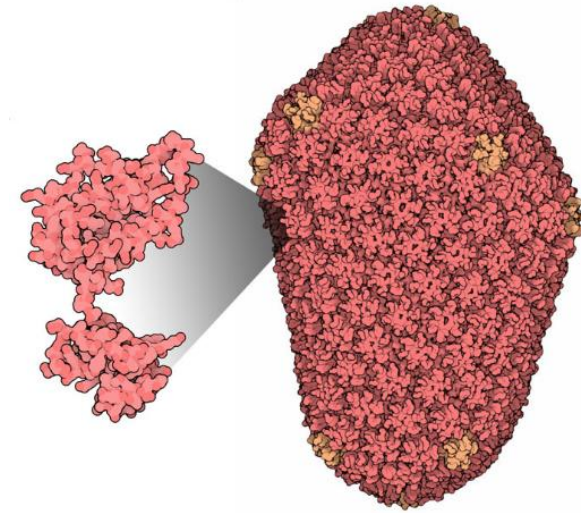
Extremely large biomolecules



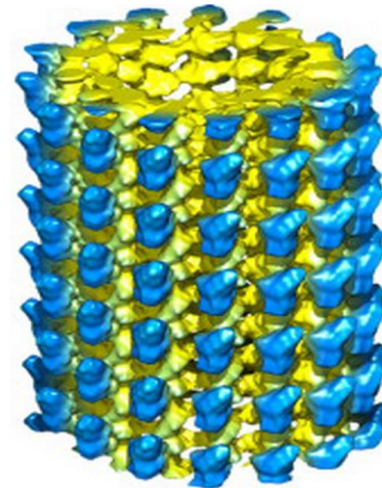
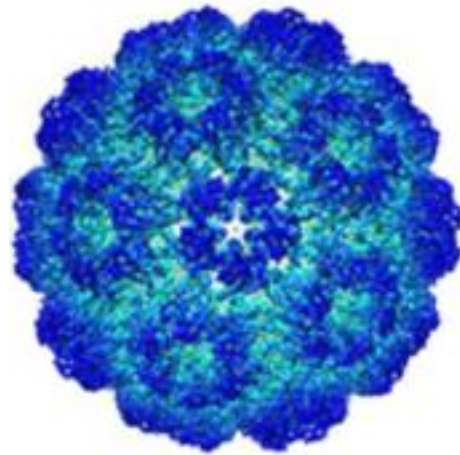
Vault particle



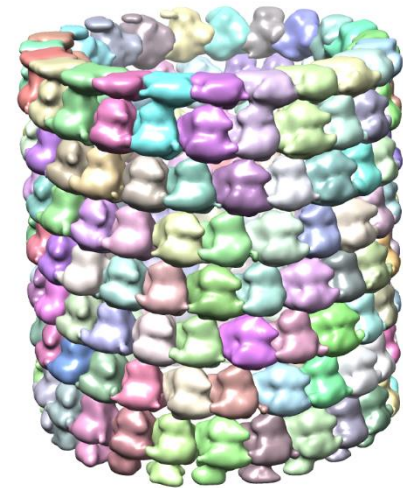
HIV virus capsid



Poliovirus capsid

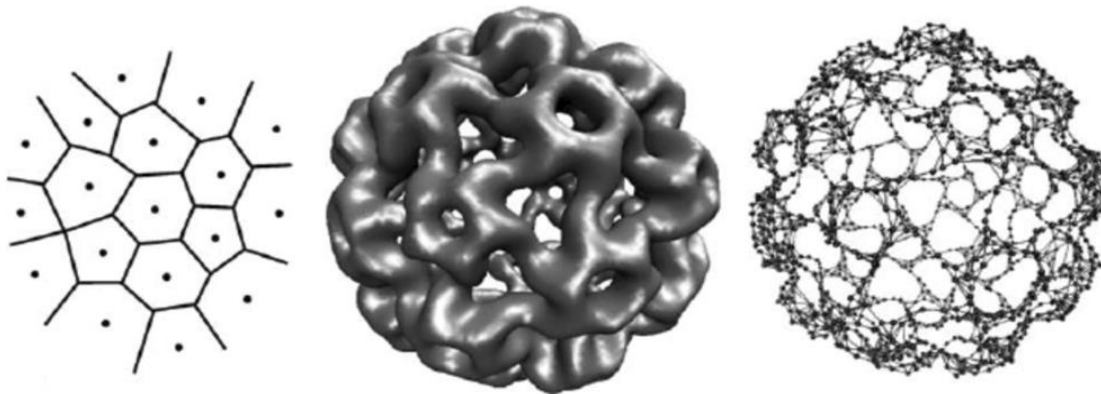


microtubule



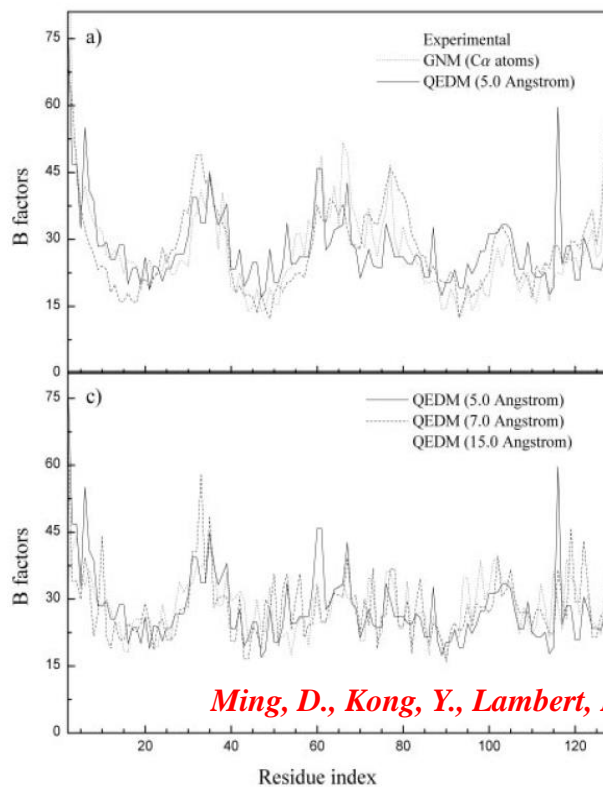
Quantized elastic deformational model (QEDM)

Voronoi Tessellation (Vector quantization)

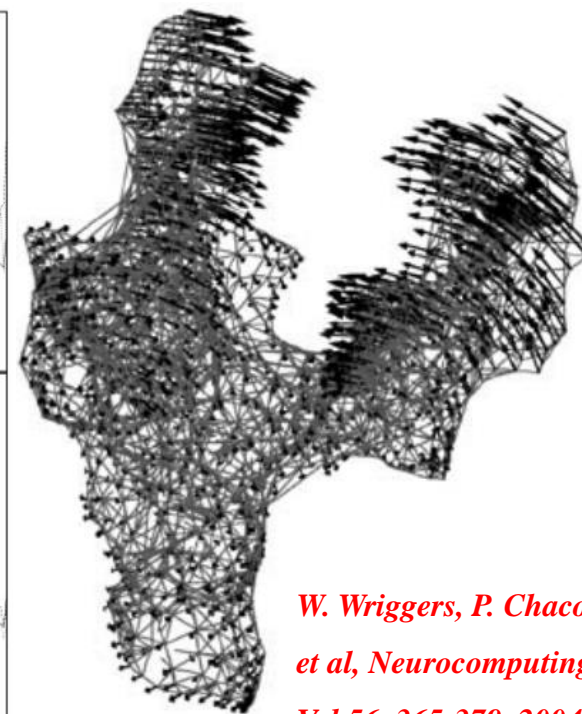
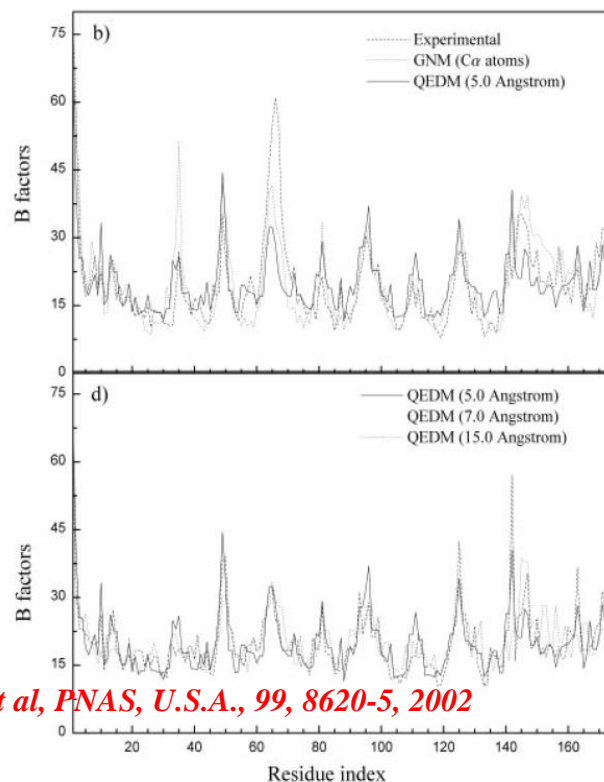


W. Wriggers, P. Chacon, et al, Neurocomputing, Vol 56, 365-379, 2004

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

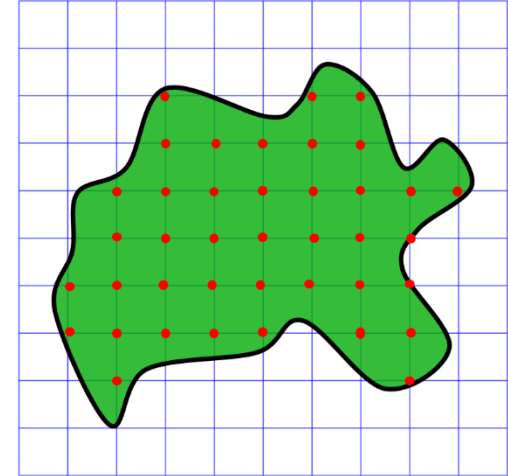
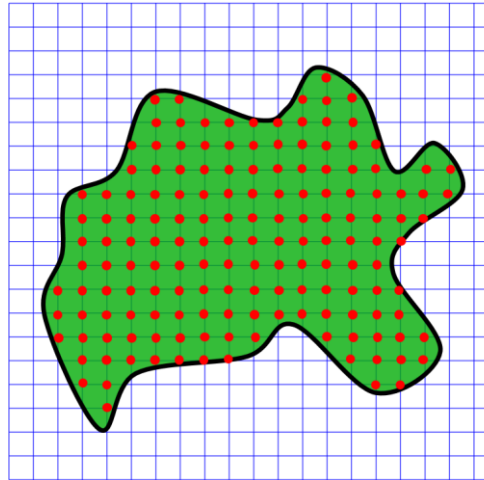
Virtual particle generation:

Various types of meshes:

Cartesian grid;

Tetrahedral mesh;

Hexahedron; Voronoi tessellation, etc.



Connection between particles:

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

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:

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

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

Potential function:

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

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

Moore-Penrose pseudoinverse:

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

eigenvector

Predicted b-factor:

eigenvalue

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

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

Multiscale representation of biomolecules

Kernel function:

$$\phi(\|r - r_j\|; \eta) = 1, \text{ as } \|r - r_j\| \rightarrow 0$$

$$\phi(\|r - r_j\|; \eta) = 0, \text{ as } \|r - r_j\| \rightarrow \infty$$

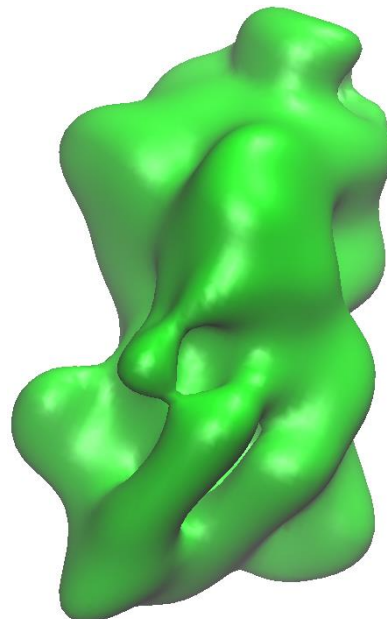
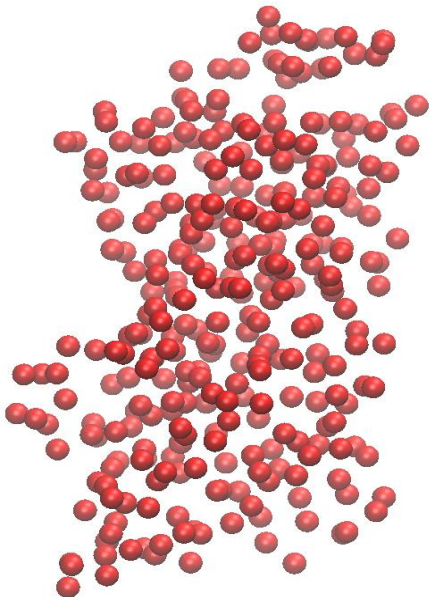
We use kernel

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

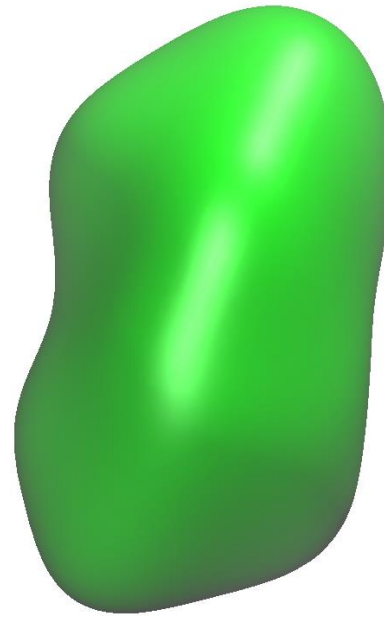
Rigidity function:

$$\mu(r) = \sum_j^N w_j \phi(\|r - r_j\|; \eta)$$

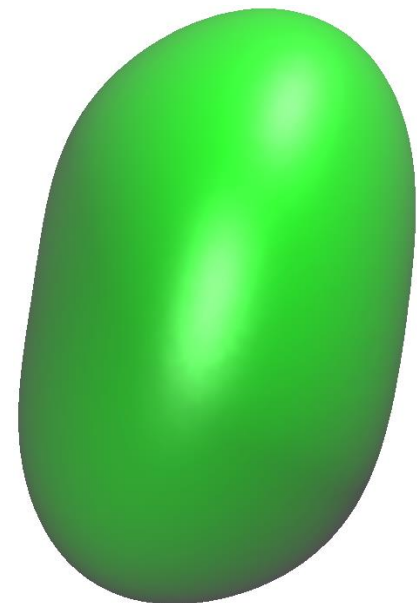
Protein ID: 2ABH



Resolution: 5Å



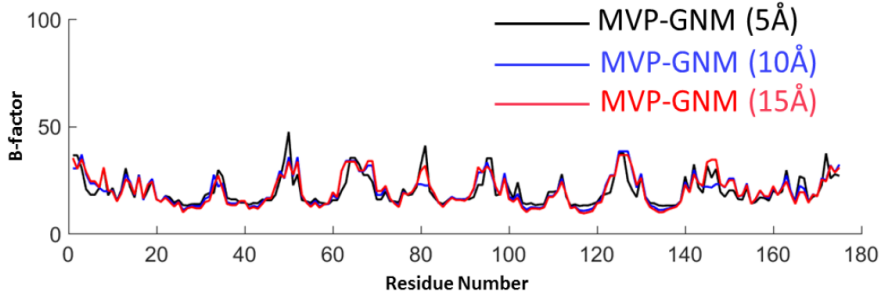
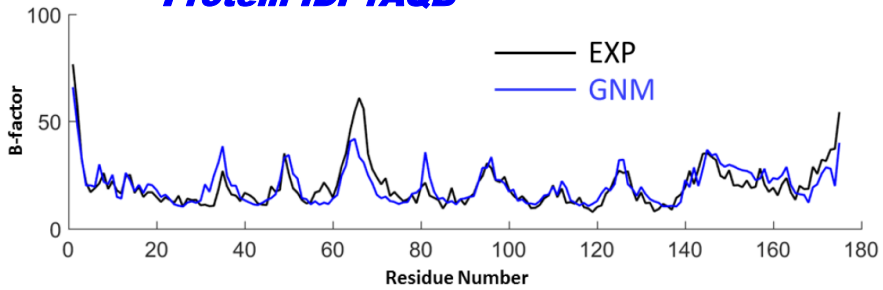
Resolution: 10Å



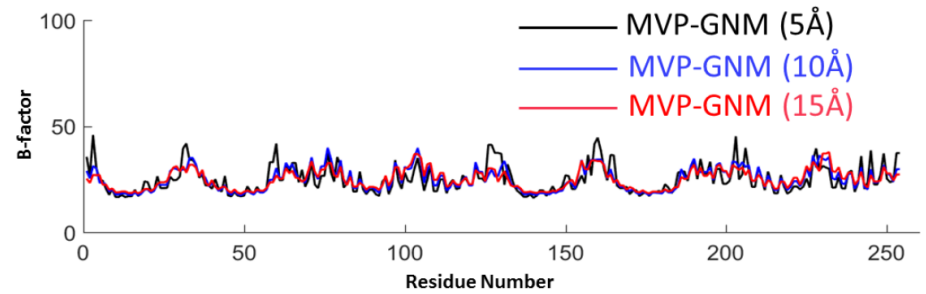
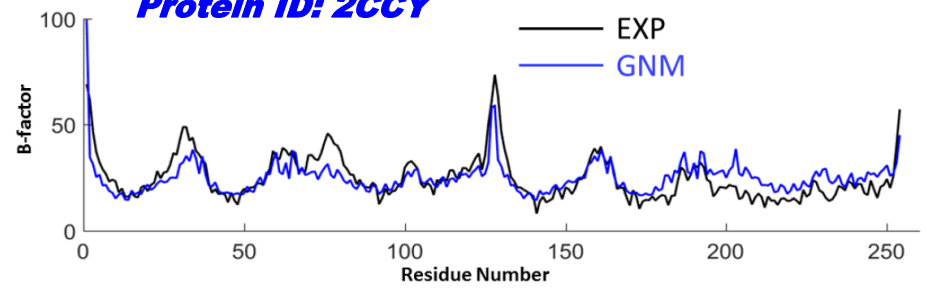
Resolution: 15Å

Validation of MVP-GNM

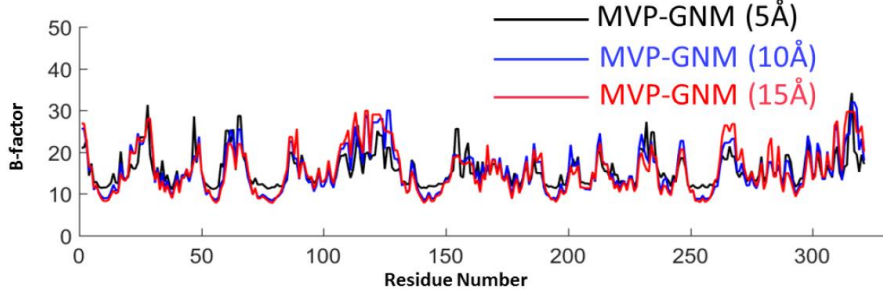
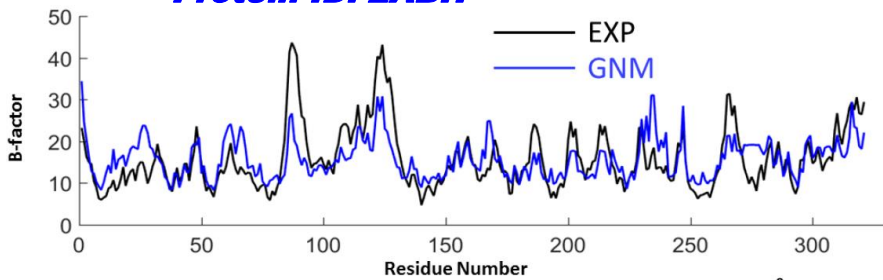
Protein ID: 1AQB



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:

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

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

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

Moore-Penrose pseudoinverse:

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

eigenvector

eigenvalue

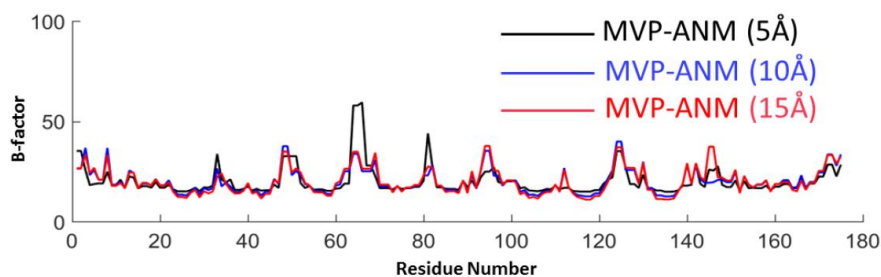
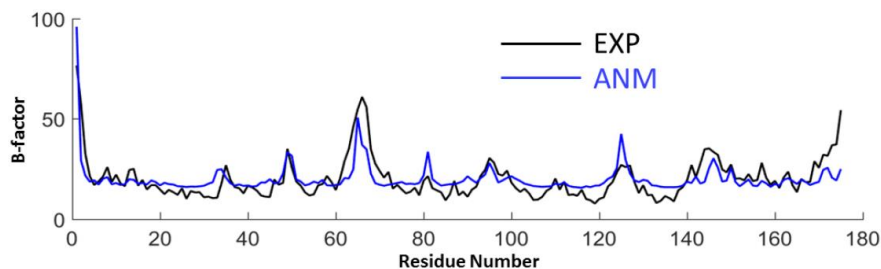
Predicted b-factor:

$$\frac{8\pi^2}{3} \sum_{j=3i-2}^{3i} \langle \Delta \mathbf{R}_j \cdot \Delta \mathbf{R}_j \rangle, \quad \forall i = 1, 2, \dots, N.$$

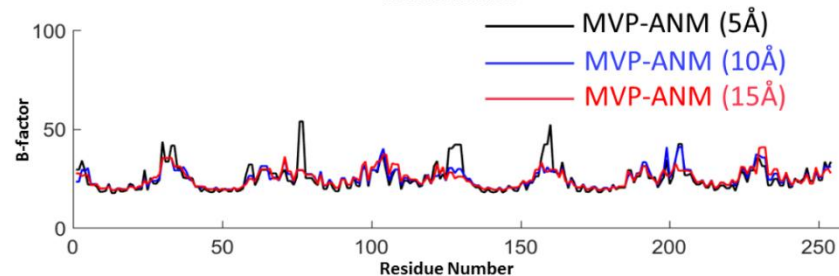
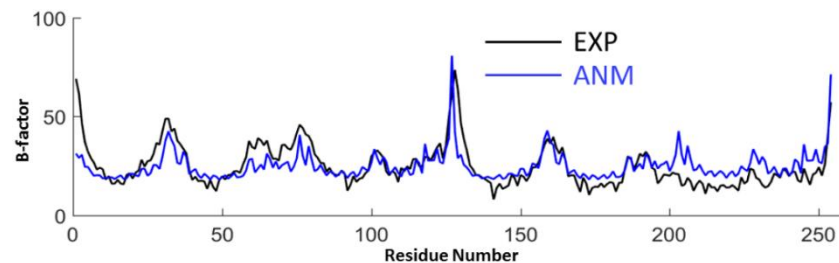
$$\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle = \frac{3k_B T}{\gamma} (H^{-1})_{ij}, \quad \forall i, j = 1, 2, \dots, 3N.$$

Validation of MVP-ANM

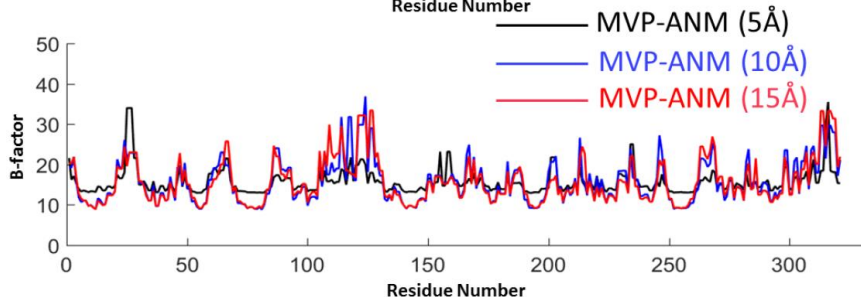
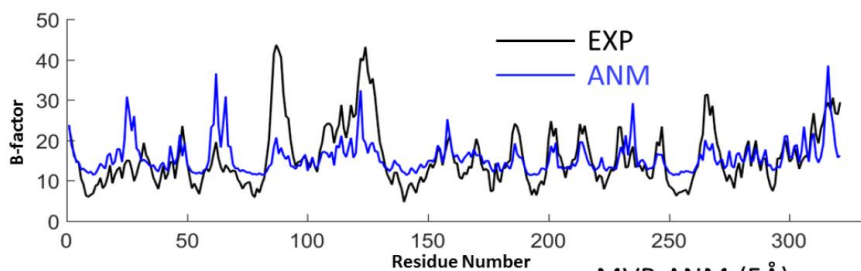
Protein ID: 1AQB



Protein ID: 2CCY



Protein ID: 2ABH



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

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

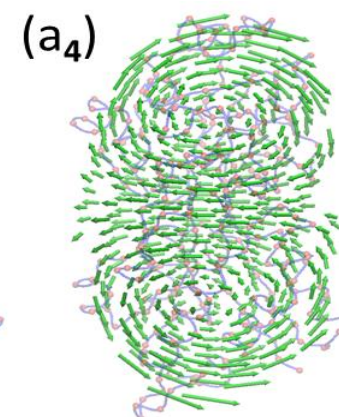
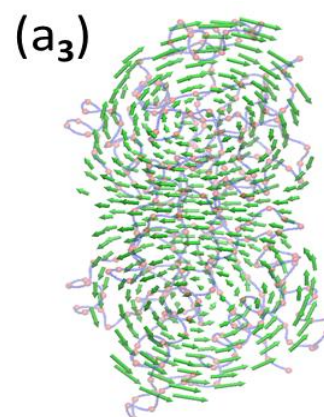
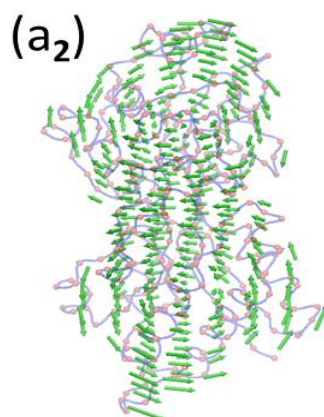
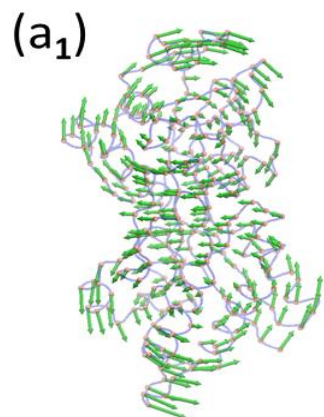
MVP-ANM

(5Å)

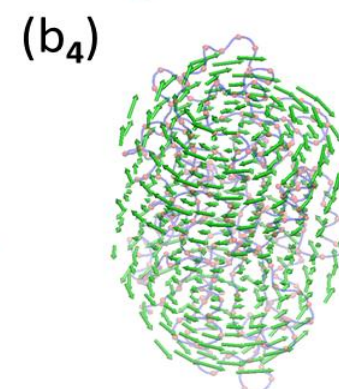
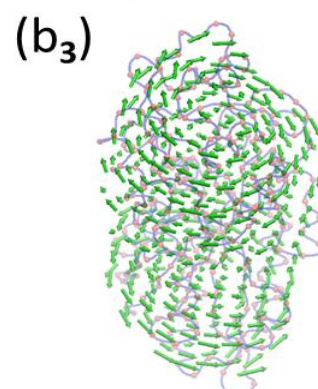
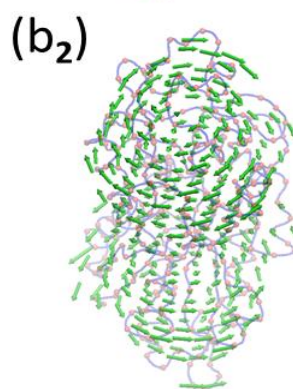
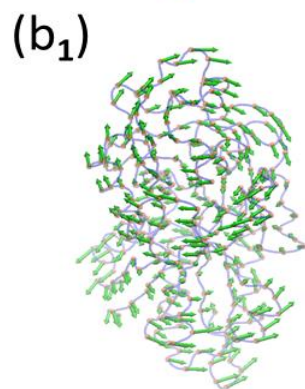
(10Å)

(15Å)

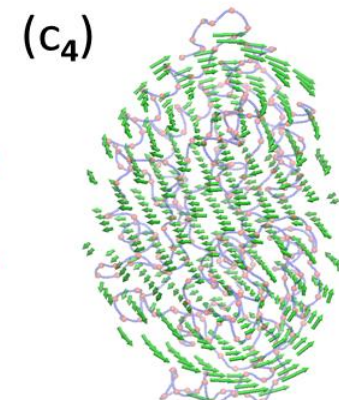
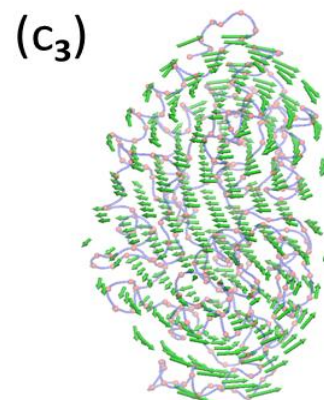
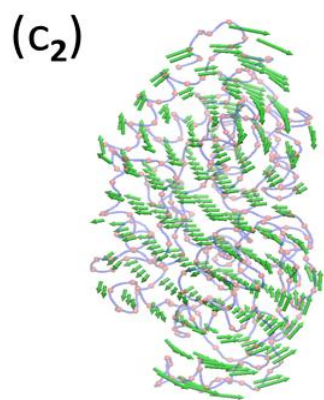
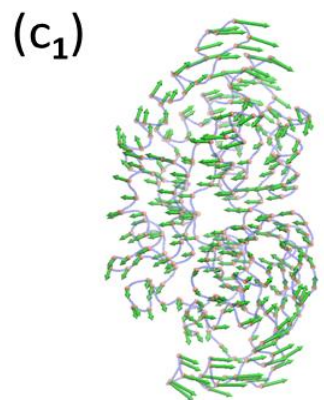
Mode 7



Mode 8

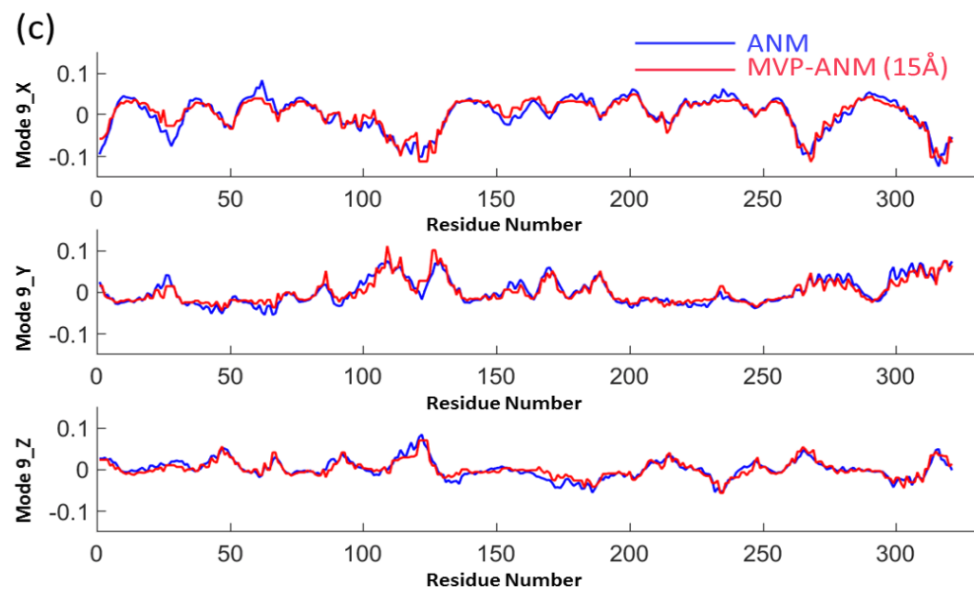
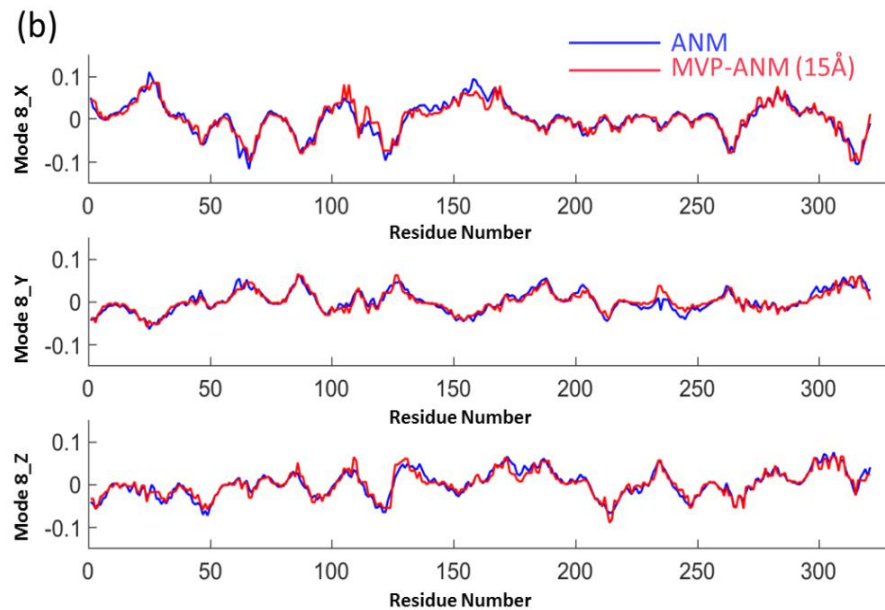
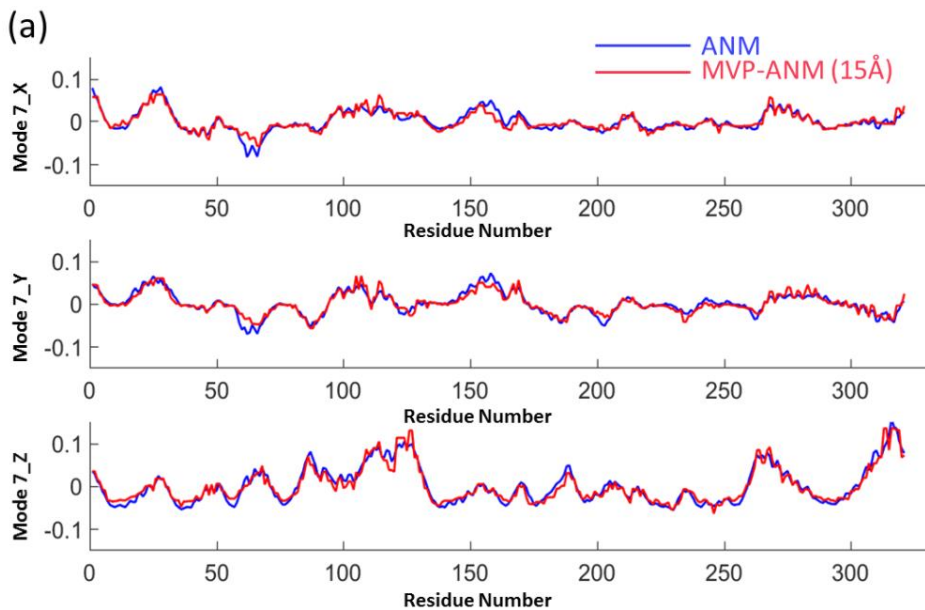


Mode 9



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

ANM

MVP-ANM

(5Å)

(10Å)

(15Å)

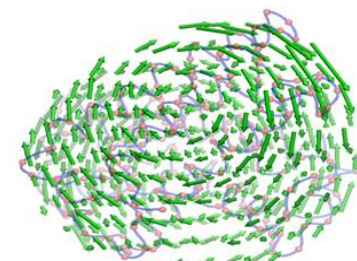
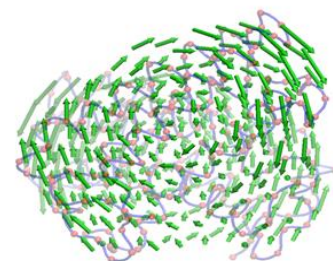
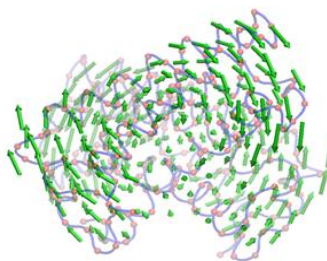
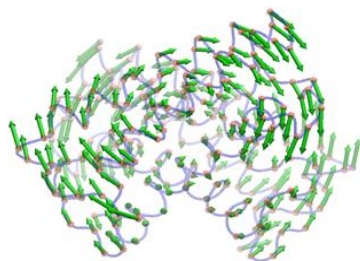
(a₁)

(a₂)

(a₃)

(a₄)

Mode 7



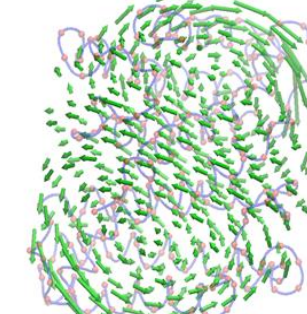
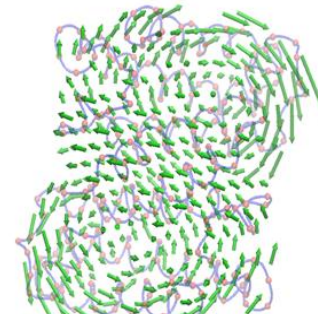
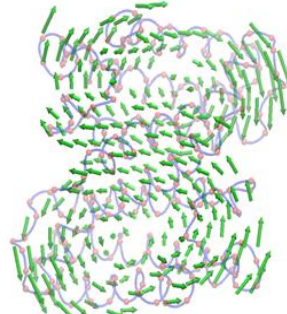
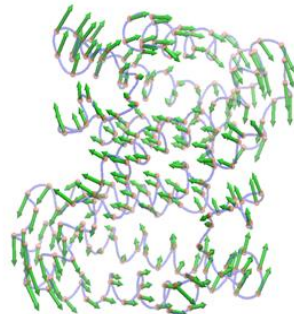
(b₁)

(b₂)

(b₃)

(b₄)

Mode 8



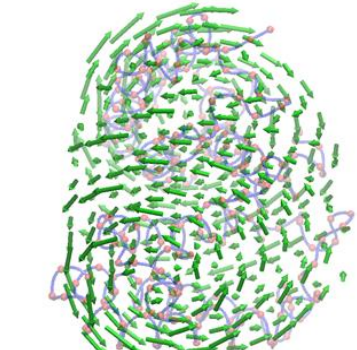
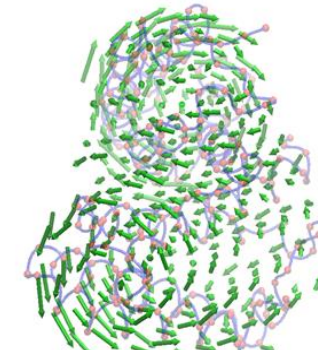
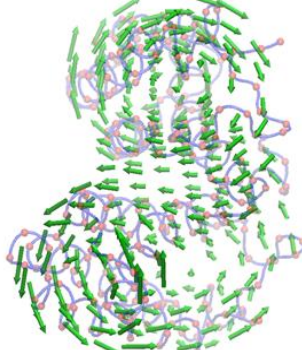
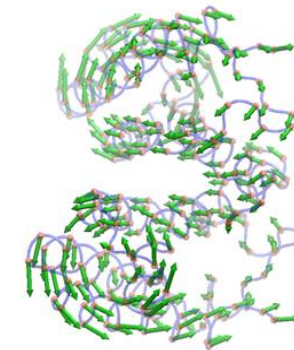
(c₁)

(c₂)

(c₃)

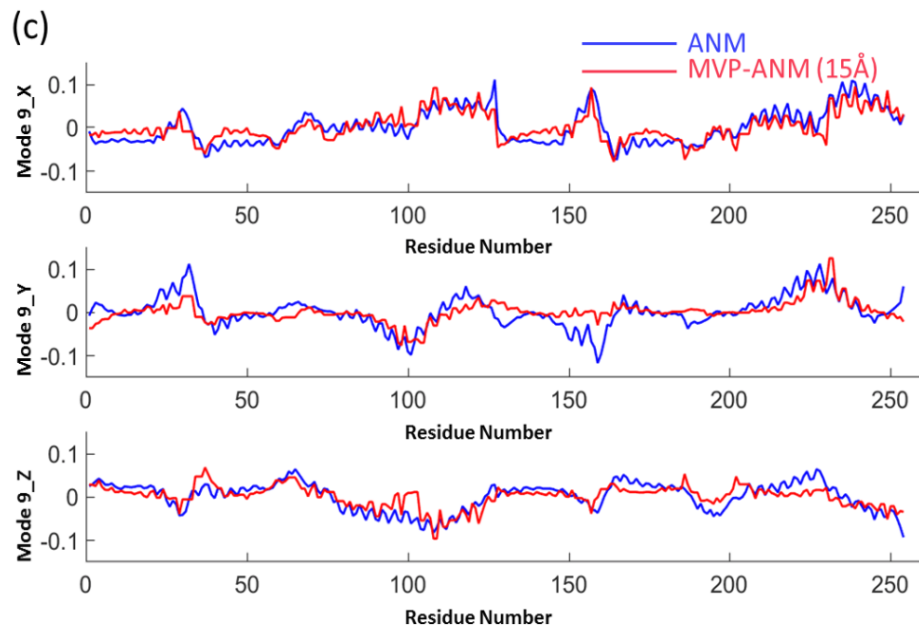
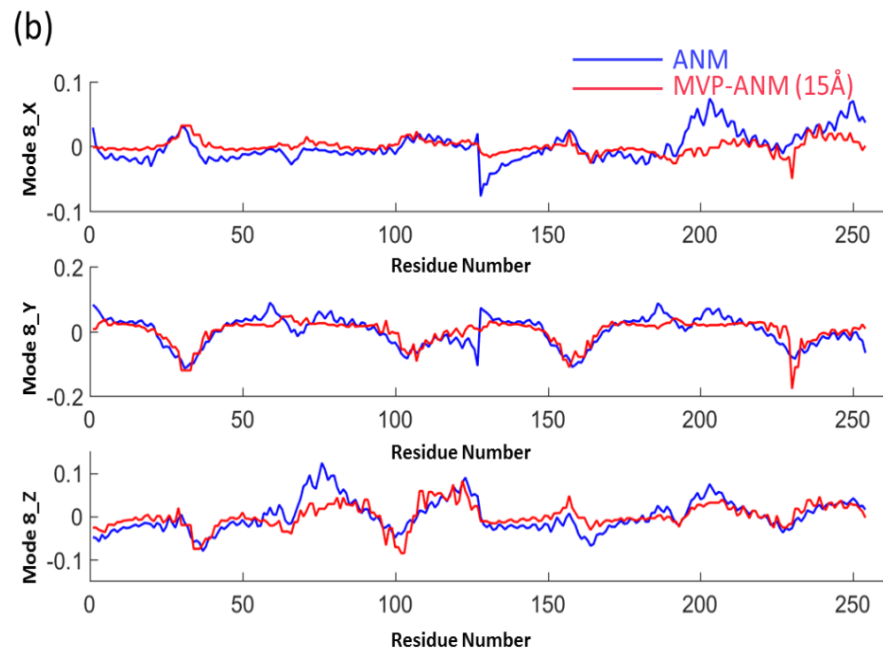
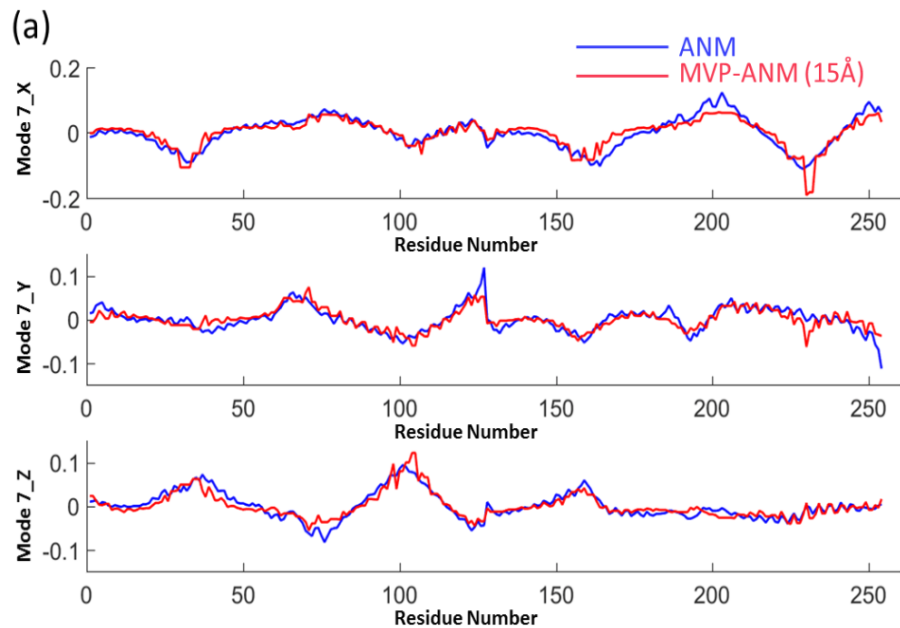
(c₄)

Mode 9



Protein ID: 2CCY

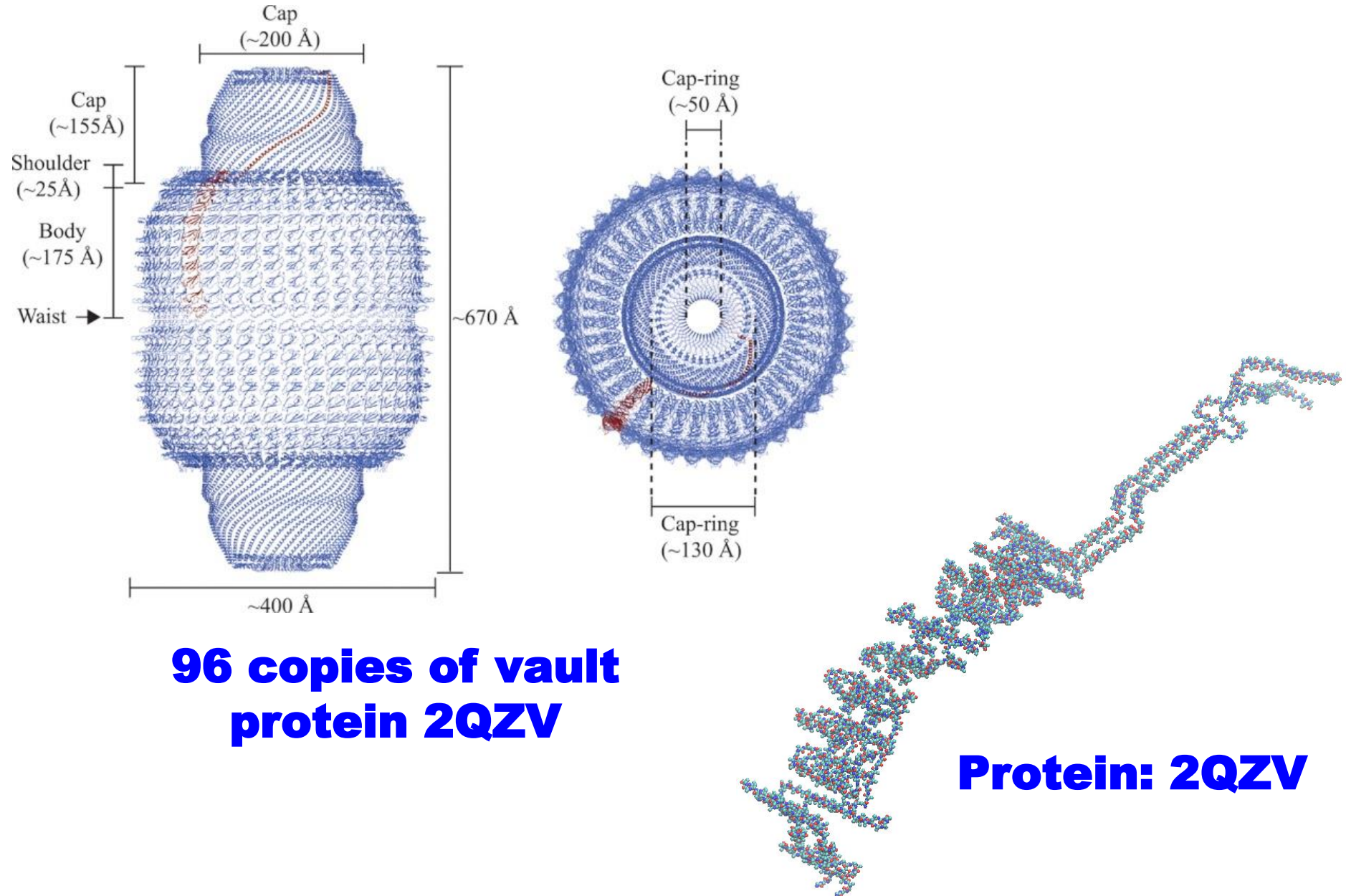
Normal modes for protein 2CCY



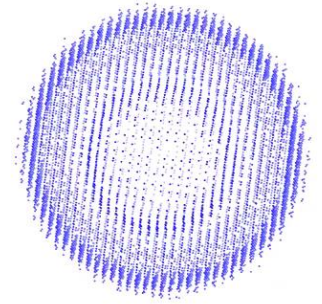
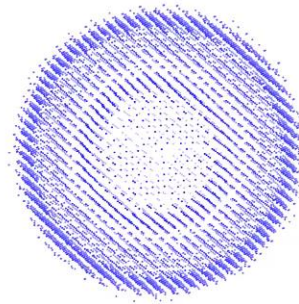
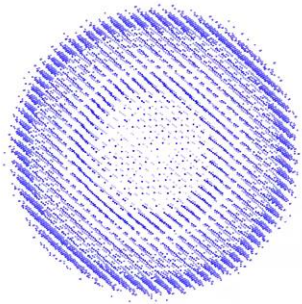
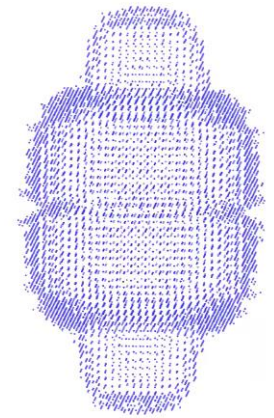
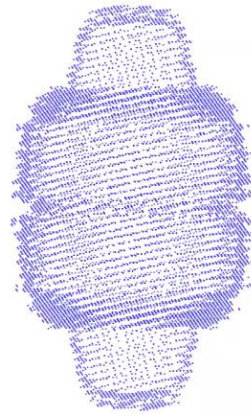
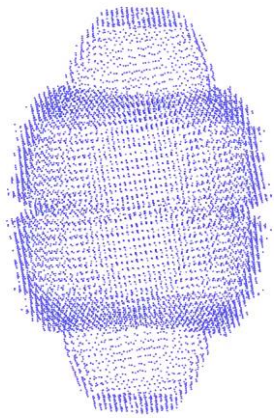
Resolution parameter is 15Å

	MVP-ANM (X)	MVP-ANM (Y)	MVP-ANM (Z)
Mode 7	0.896	0.834	0.910
Mode 8	0.461	0.799	0.746
Mode 9	0.827	0.681	0.759

The dynamics of Vault Shell



Multiscale virtual particle based elastic network model of Vault



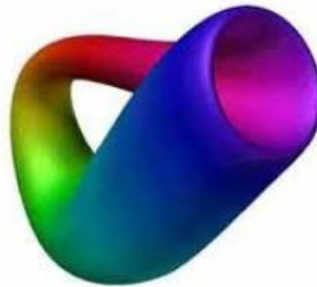
Part 2-2: Topological modeling of biomolecules

Classical Topology

Möbius Strips (1858)



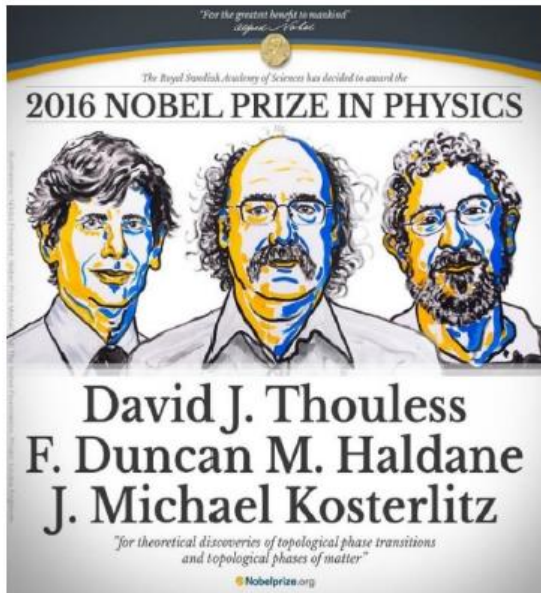
Klein Bottle (1882)



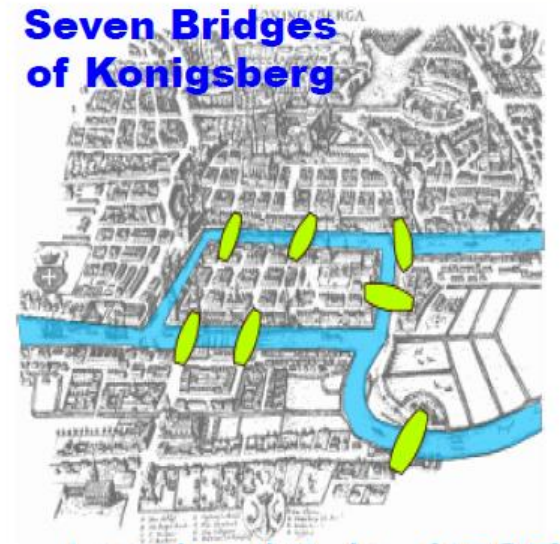
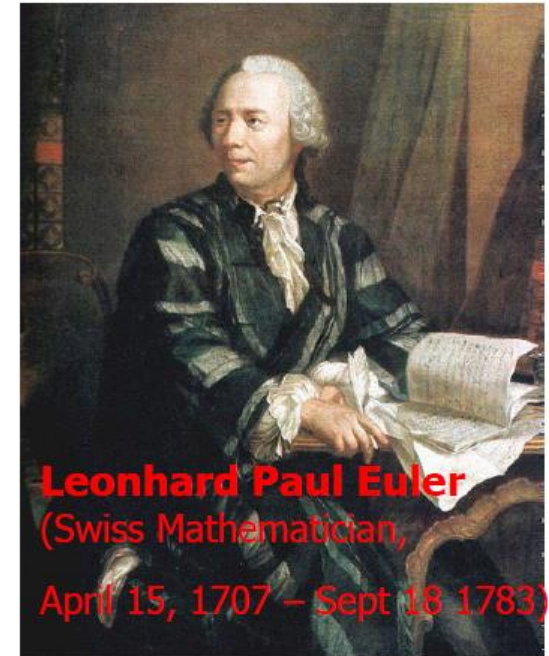
Torus



Double Torus



Augustin-Louis Cauchy,
Ludwig Schläfli,
Johann Benedict Listing,
Bernhard Riemann, and
Enrico Betti

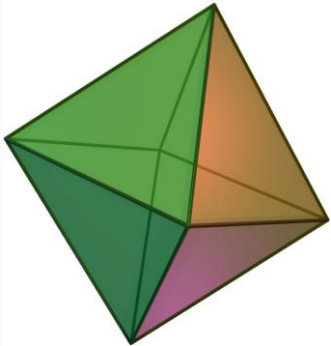


Leonhard Euler (1735)

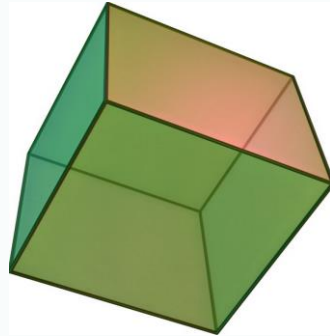
Euler Characteristic

POLYHEDRONS

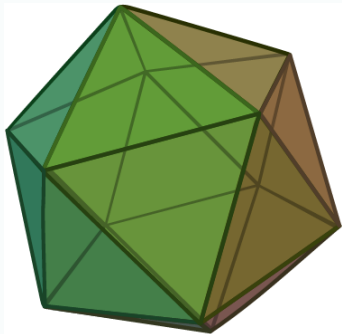
$$\chi(M) = V_{\text{vertex}} - E_{\text{edge}} + F_{\text{face}}$$



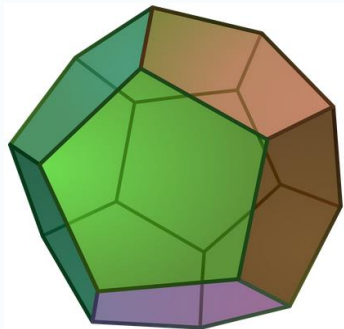
$$\chi = 6 - 12 + 8 = 2$$



$$\chi = 8 - 12 + 6 = 2$$



$$\chi = 12 - 30 + 20 = 2$$



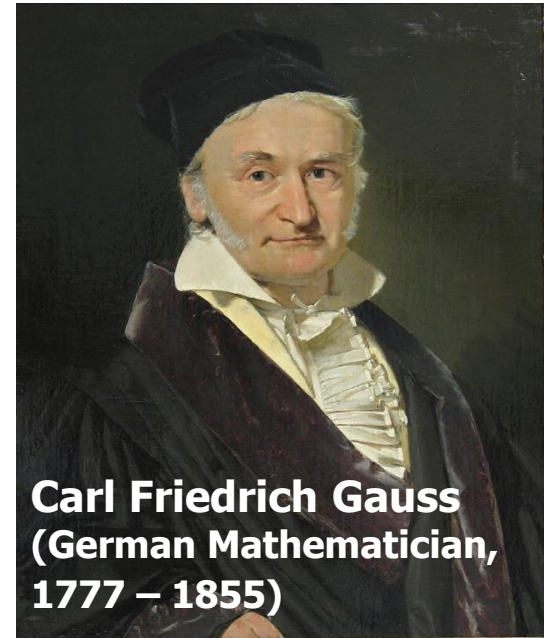
$$\chi = 20 - 30 + 12 = 2$$

$$\chi = 2$$

$$K = \frac{1}{R^2}$$



"Spherical cow", Wikipedia



Carl Friedrich Gauss
(German Mathematician,
1777 – 1855)

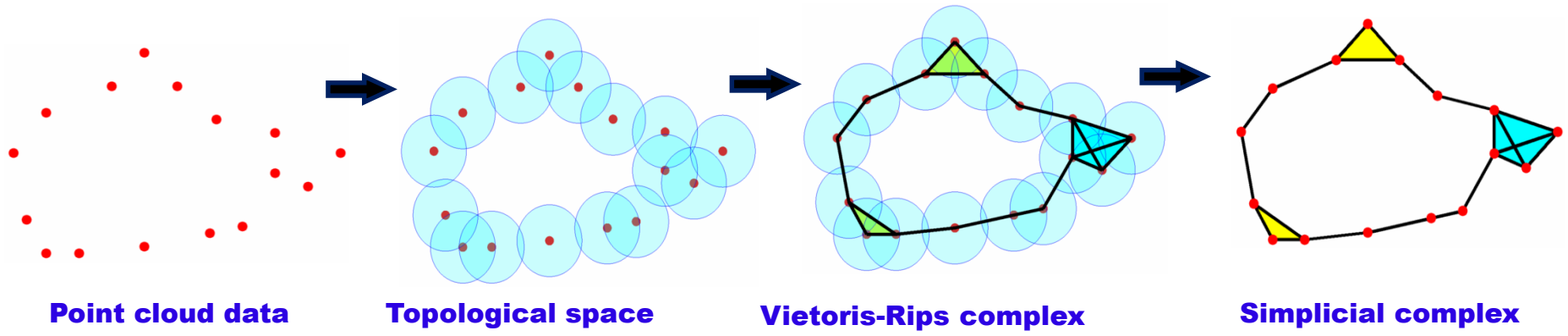
Gauss-Bonnet Theorem

Gaussian curvature

$$\int_M K dA = 2\pi\chi(M)$$

**Connection between
differential geometry
and topology**

Topological data analysis



Chain group: $C_k(K, \mathbb{Z}_2)$

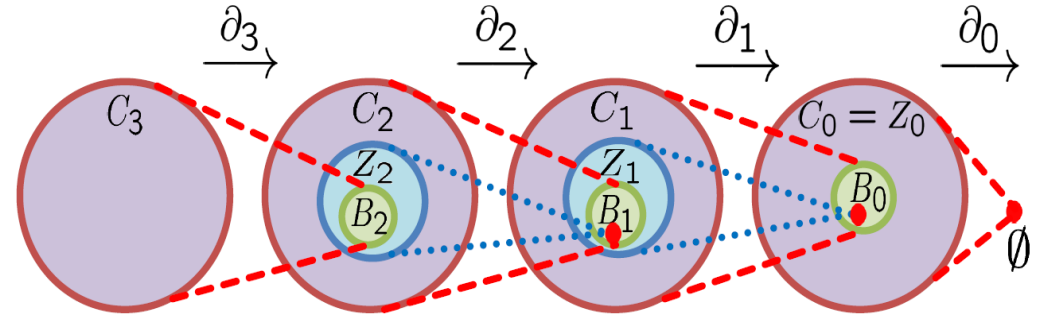
Boundary operator: $\partial_k \sigma^k = \sum_{i=0}^k (-1)^i \{v_0, v_1, \dots, \hat{v}_i, \dots, v_k\}$

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

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

Quotient group:

$$H_k = Z_k / B_k$$



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

The topological information can be calculated!!

Opportunities, **challenges** and **promises**

Opportunities from topological methods:

- ❖ **New approach for big data characterization and classification.**
- ❖ **Dramatic reduction of dimensionality and data size.**
- ❖ **Applicable to a variety of fields.**

Challenges with topological methods:

- **Geometric methods are inundated with structural details.**
- **Topology incurs too much reduction of original information.**
- **Topology is hardly used for quantitative prediction.**

Promises from persistent homology:

- ✓ **Embeds geometric information in topological invariants.**
- ✓ **Bridges the gap between geometry and topology.**

Researchers:

**Frosini (1991),
Robins (2000),
Edelsbrunner, Letscher and Zomorodian (2002),
Kaczynski, Mischaikow and Mrozek (2004),
Zomorodian and Carlsson (2005),
Ghrist (2008),
Dey and Wang(2009),**

.....

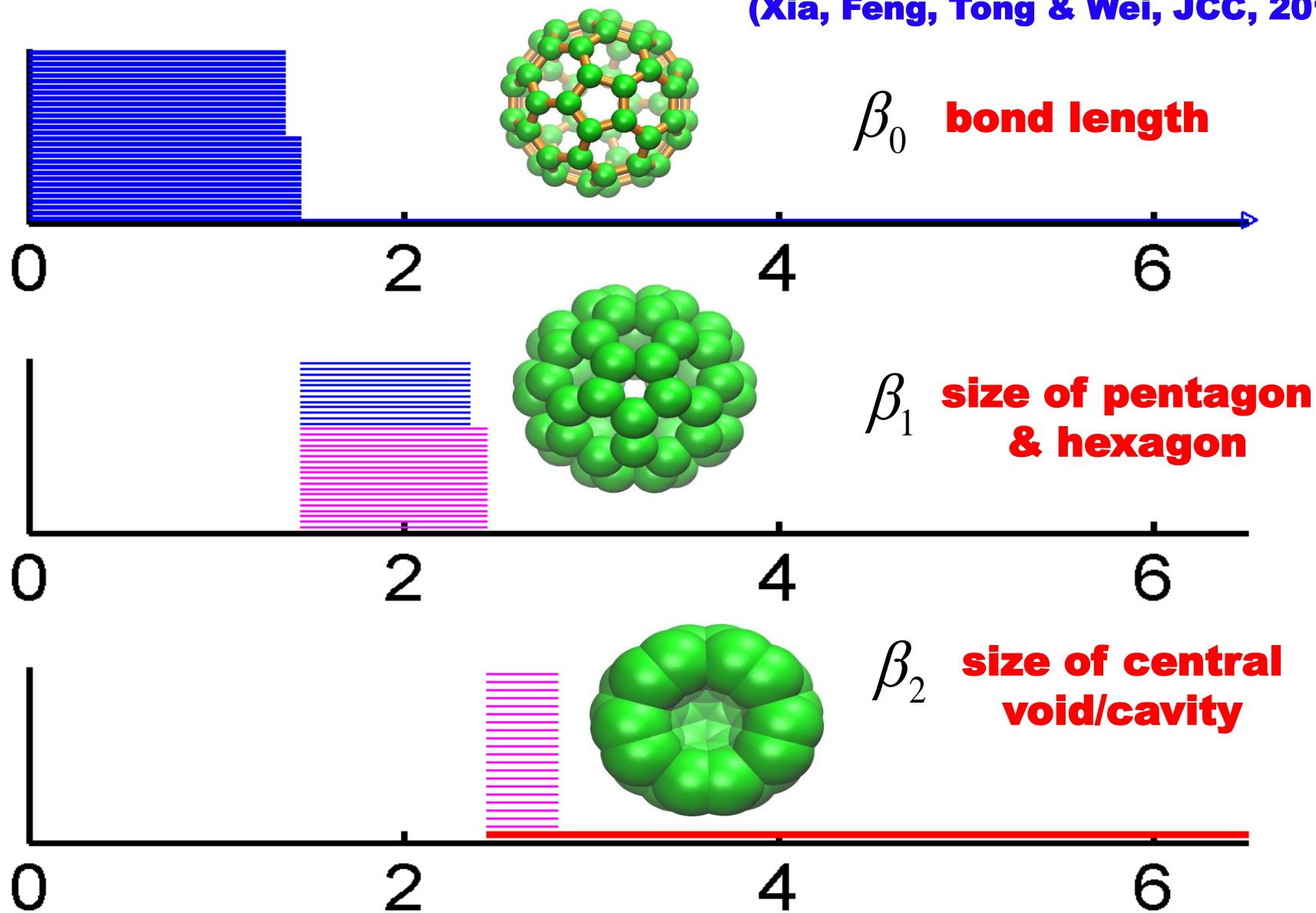
Softwares:

**Javaplex,
Perseus,
Dipha,
Dionysus,**

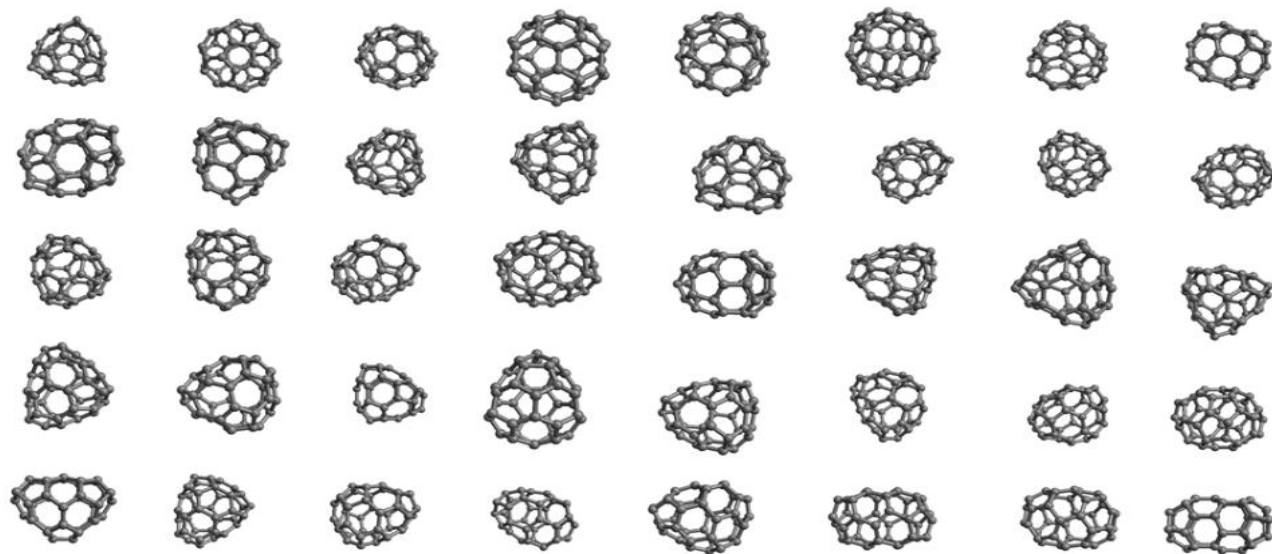
.....

Simplest molecules to start with: fullerene C₆₀

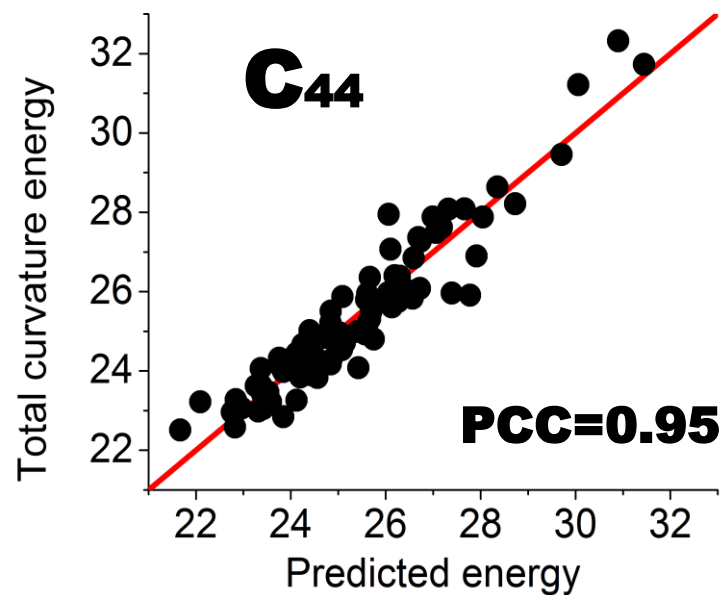
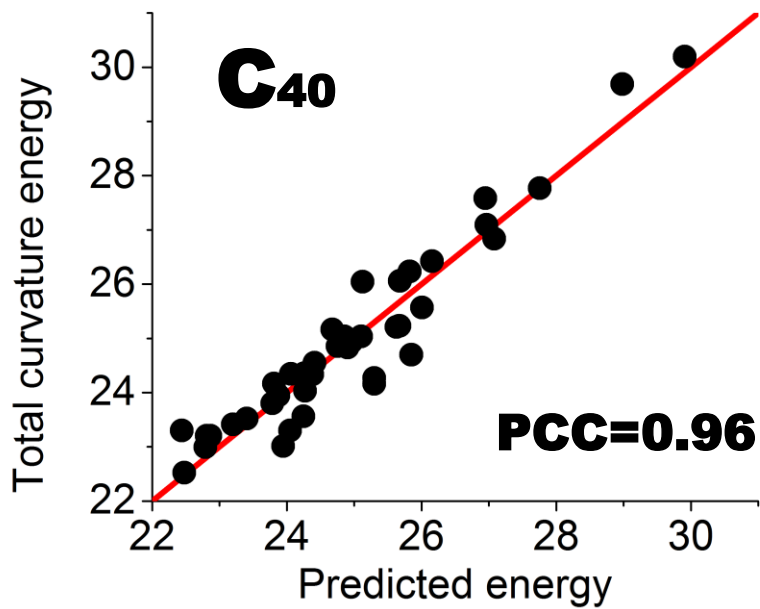
(Xia, Feng, Tong & Wei, JCC, 2015)



Fullerene isomers



$$E \propto 1/L(\beta_2)$$










































(Xia, Feng, Tong & Wei, JCC, 2015)

Fullerene isomers

<http://www.nanotube.msu.edu/fullerene/fullerene-isomers.html>

C_n Fullerenes

Click on a fullerene for a list of isomers, their structure and properties.

 C ₂₀	 C ₂₄	 C ₂₆	 C ₂₈	 C ₃₀	 C ₃₂	 C ₃₄	 C ₃₆	 C ₃₈	 C ₄₀
 C ₄₂	 C ₄₄	 C ₄₆	 C ₄₈	 C ₅₀	 C ₅₂	 C ₆₀	 C ₇₀	 C ₇₂	 C ₇₄
 C ₇₆	 C ₇₈	 C ₈₀	 C ₈₂	 C ₈₄	 C ₈₆	 C ₉₀	 C ₉₂	 C ₉₄	 C ₉₆
 C ₉₈	 C ₁₀₀	 C ₁₈₀	 C ₂₄₀	 C ₂₆₀	 C ₃₂₀	 C ₅₀₀	 C ₅₄₀	 C ₇₂₀	

The fullerene geometries are based on structures in the Fullerene Library that has been created by M. Yoshida. The geometries have been reoptimized using a fast Dreiding-like forcefield that is built into the free [Discovery Studio Visualizer](#). The numbering scheme of fullerene isomers seems to agree with that used in the monograph "An atlas of fullerenes" by P. W. Fowler and D. E. Manolopoulos.

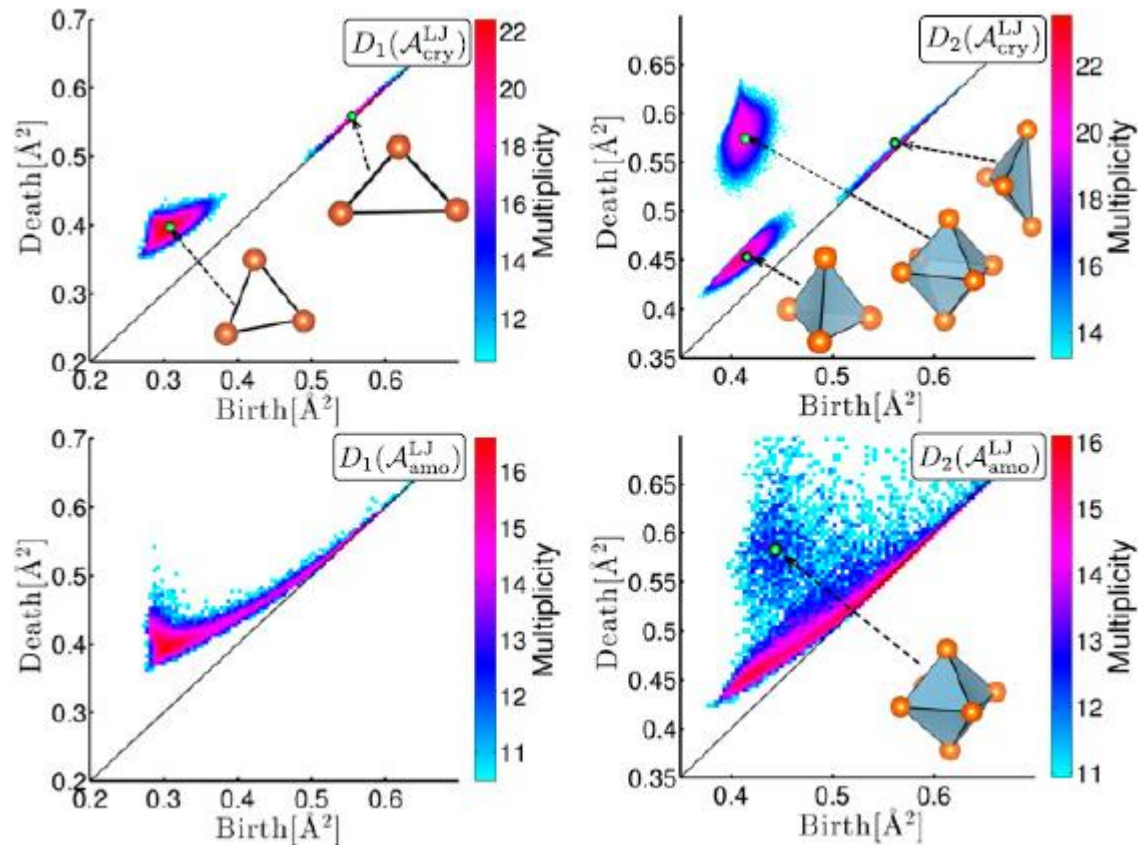
The curvature energy is an estimate of the formation energy of the particular isomer with respect to graphite. This estimate, provided by Jie Guan, is based on local curvature, as defined in the publication by Jie Guan, Zhongqi Jin, Zhen Zhu, Chern Chuang, Bih-Yaw Jin, and David Tománek, entitled *Local Curvature and Stability of Two-Dimensional Systems*, *Phys. Rev. B* **90**, 245403 (2014).

The web resource at <http://www.nanotube.msu.edu/fullerene/fullerene-isomers.html> has been provided by [David Tomanek](#) and Nick Frederick at the [Michigan State University Computational Nanotechnology Lab](#). It is linked to the Supplementary Information provided with the monograph [Guide through the Nanocarbon Jungle: Buckyballs, Nanotubes, Graphene, and Beyond](#).

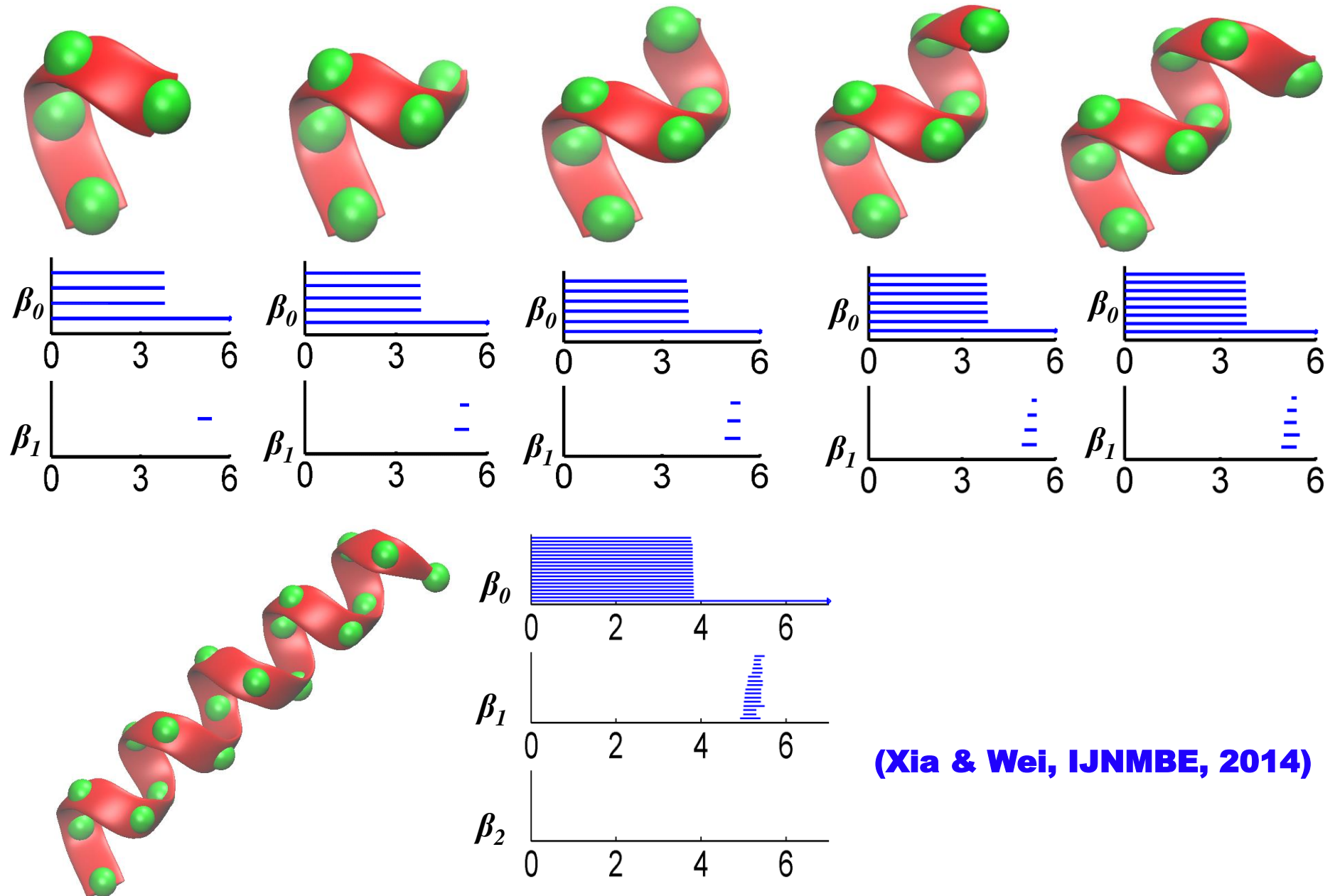
(Nano) Material

Hierarchical structures of amorphous solids characterized by persistent homology

Yasuaki Hiraoka^{a,1,2}, Takenobu Nakamura^{a,1}, Akihiko Hirata^a, Emerson G. Escolar^a, Kaname Matsue^b, and Yasumasa Nishiura^a

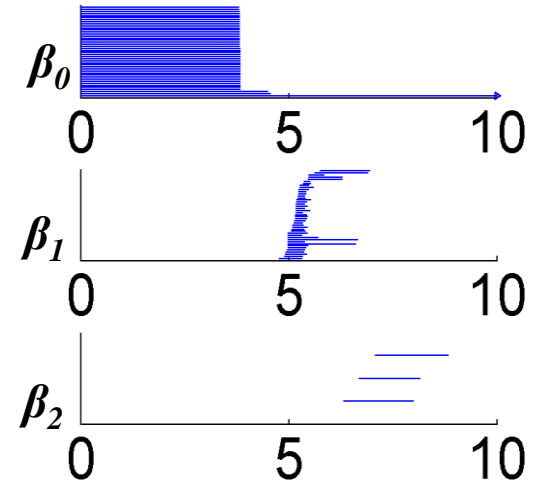
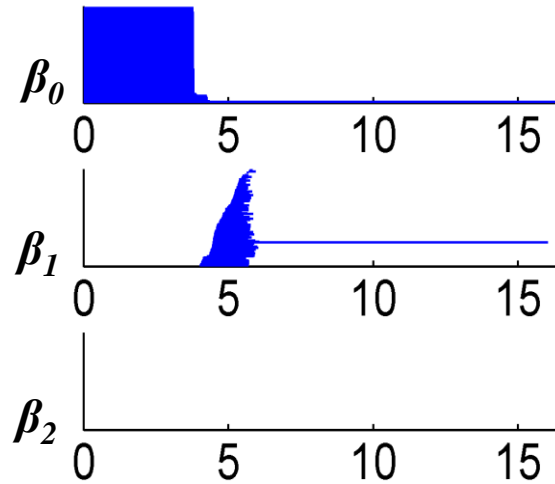
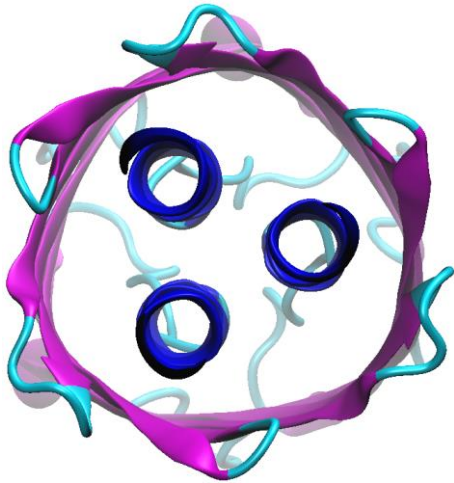
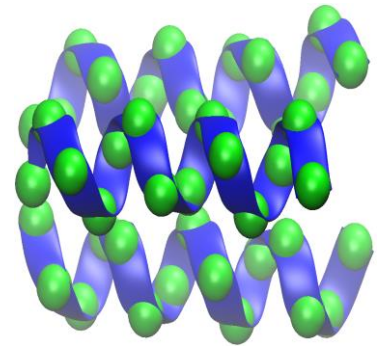
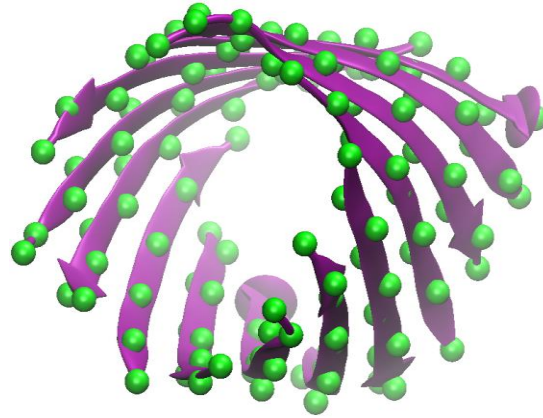
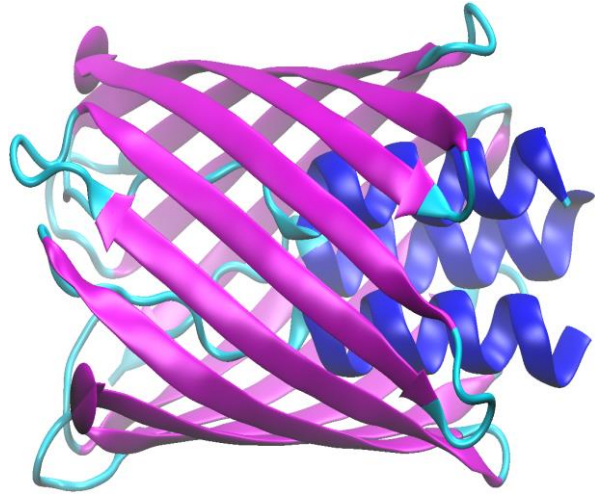


Topological fingerprints of an alpha helix

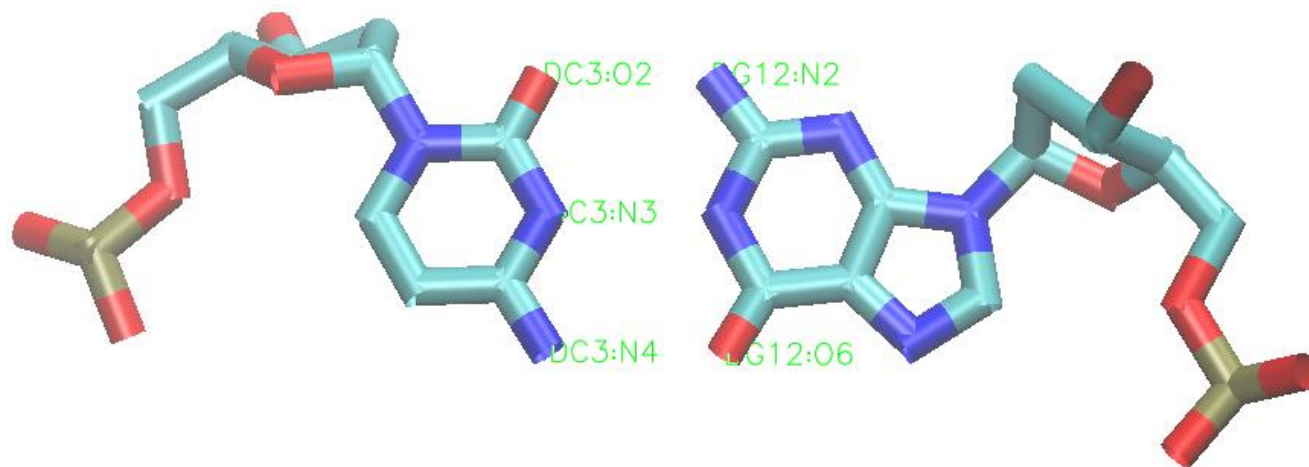
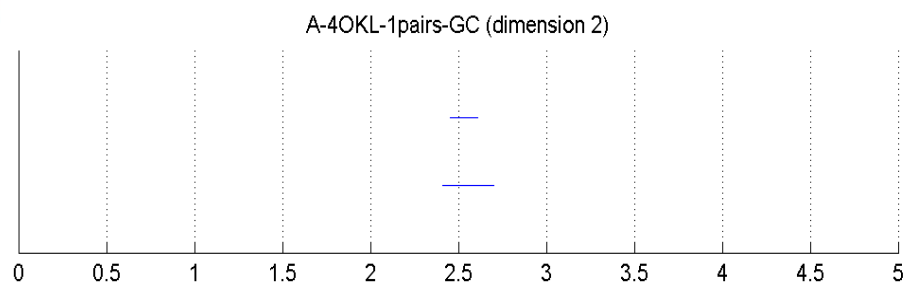
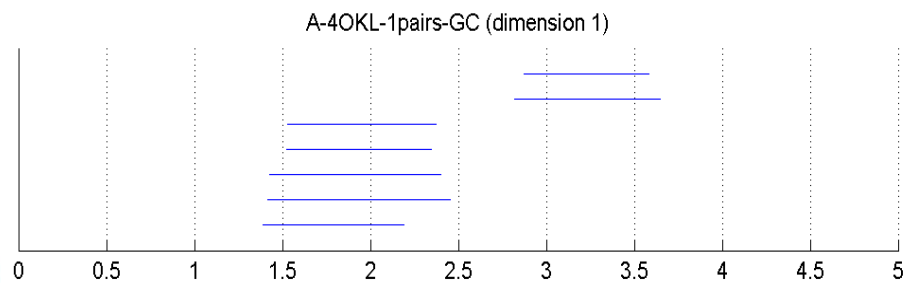
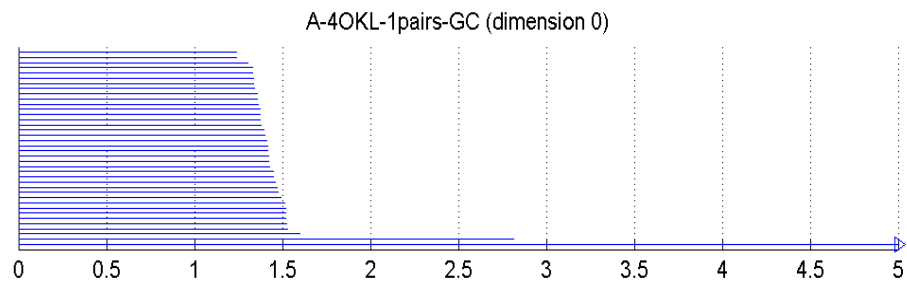
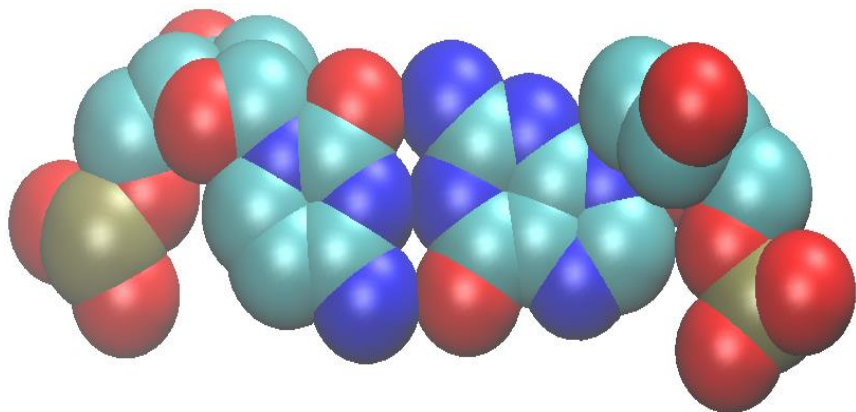


Topological fingerprints of beta barrel

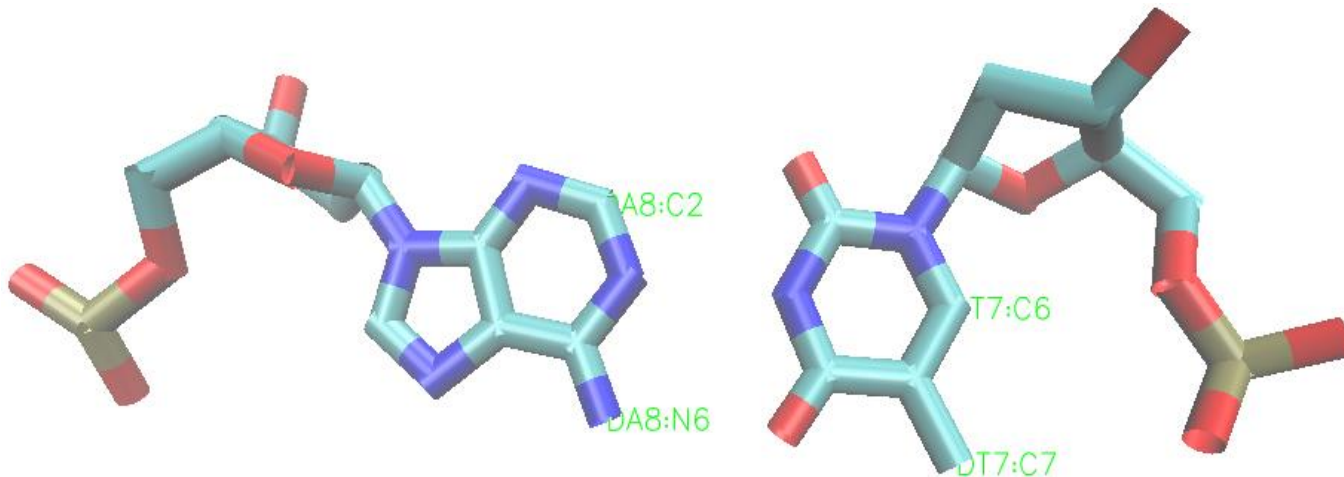
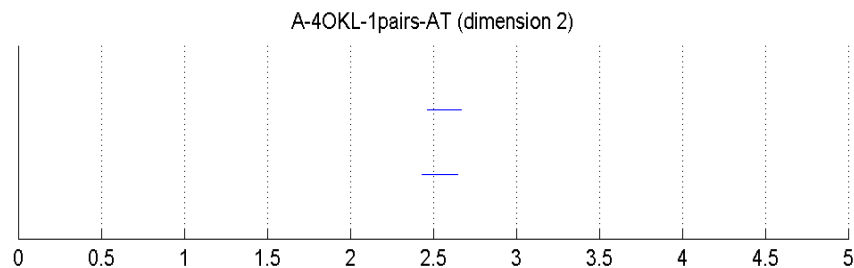
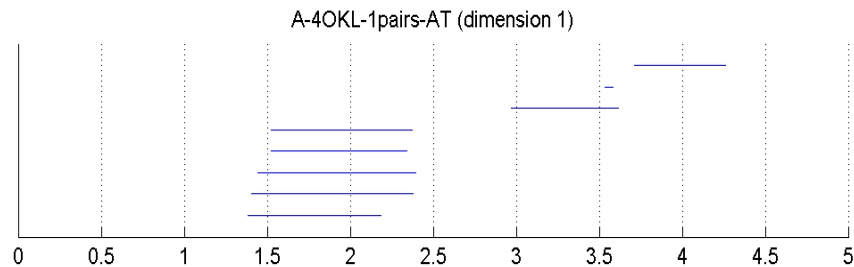
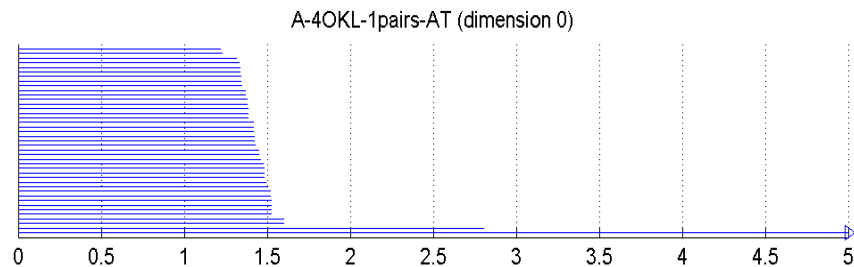
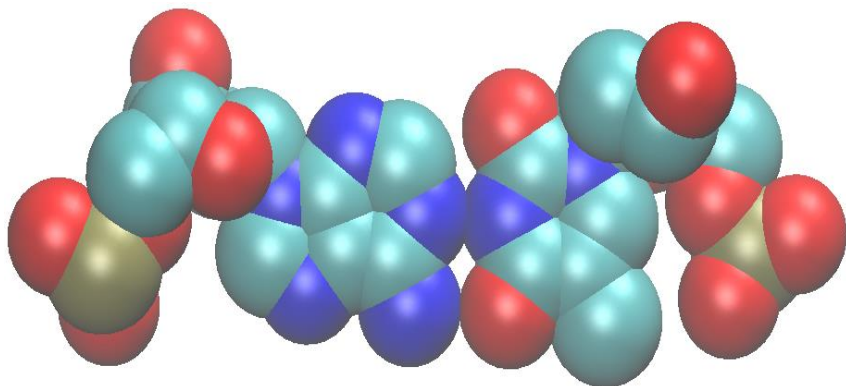
Protein:2GR8

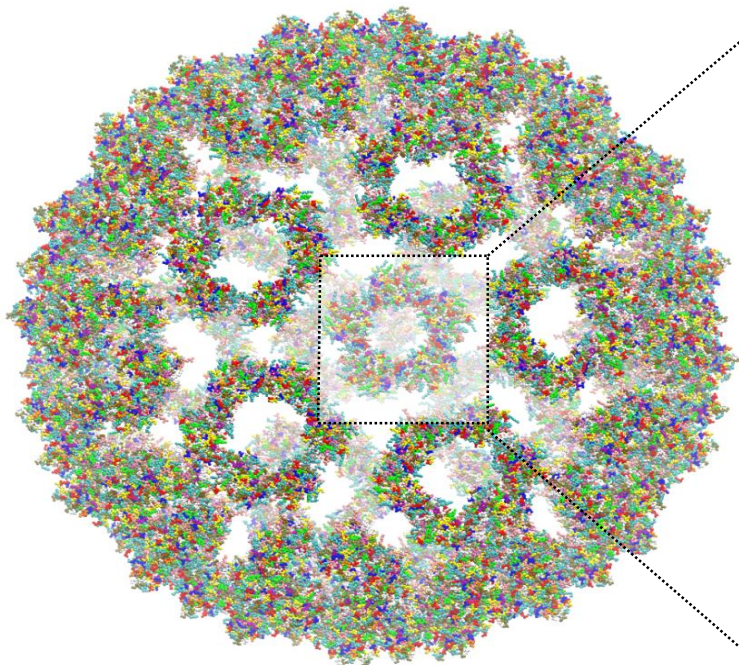


DNA:G-C pair

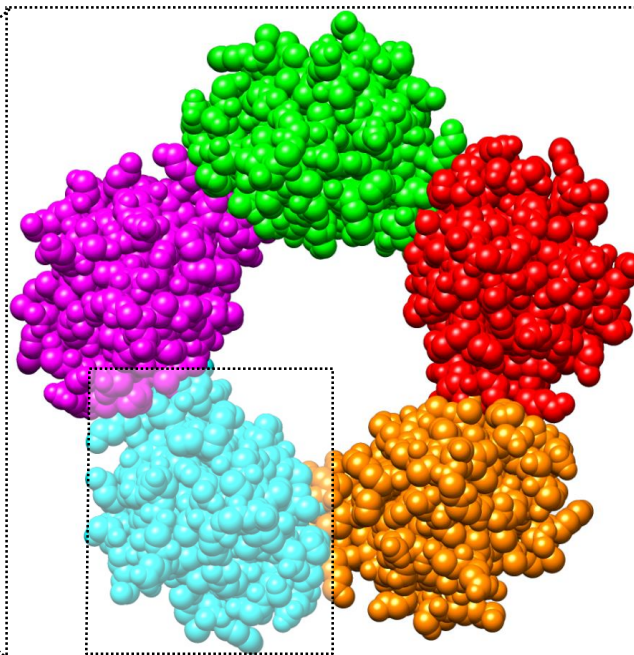


DNA: A-T pair



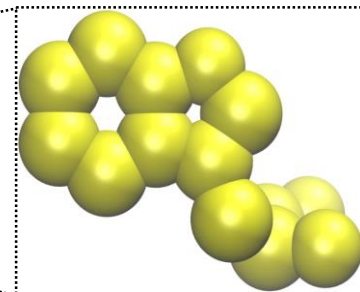
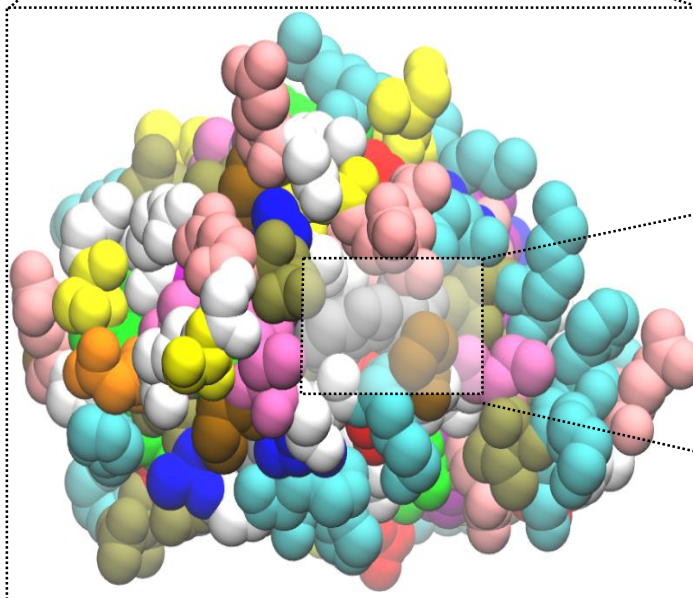


Protein ID:1DYL



(Xia & Wei, JCP, 2015)

**Topic--
Multiresolution
PHA of excessively
large biomolecular
data**



Flexibility-rigidity analysis

Kernel function:

$$\phi(\|r - r_j\|; \eta) = 1, \text{ as } \|r - r_j\| \rightarrow 0$$

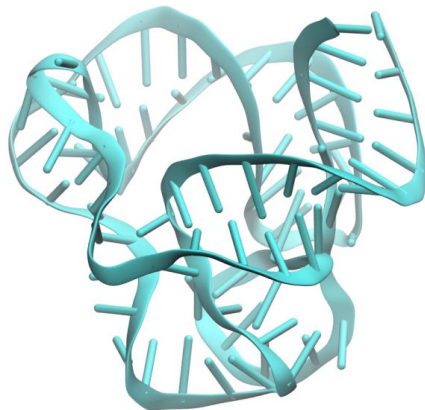
$$\phi(\|r - r_j\|; \eta) = 0, \text{ as } \|r - r_j\| \rightarrow \infty$$

For example:

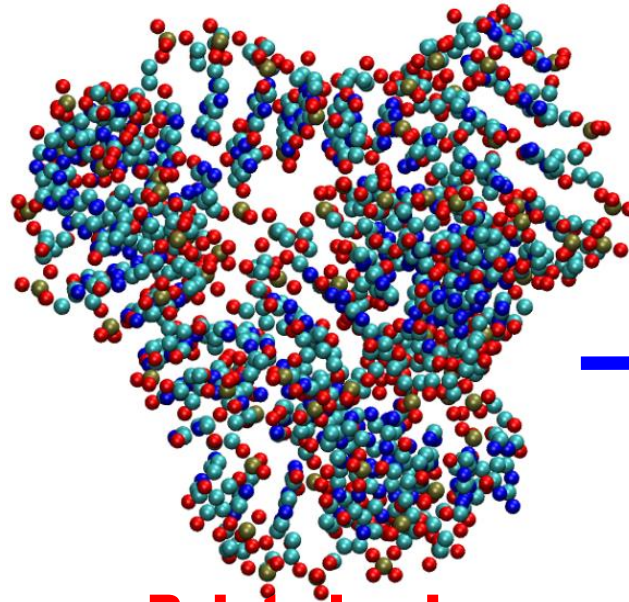
$$\phi(\|r - r_j\|; \eta) = e^{-(\|r - r_j\|/\eta)^\kappa}, \kappa > 0$$

$$\mu(r) = \sum_j^N w_j \phi(\|r - r_j\|; \eta)$$

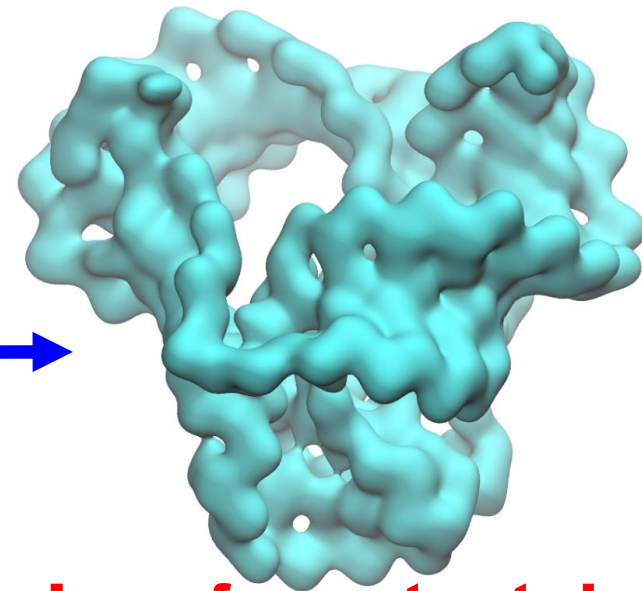
Rigidity function:



RNA: 4QG3



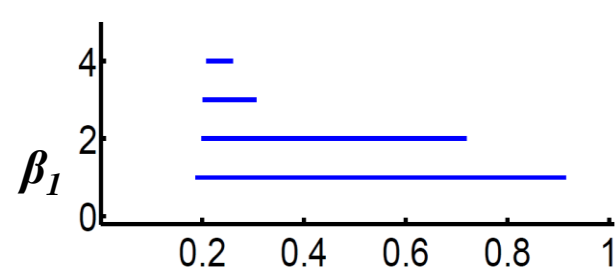
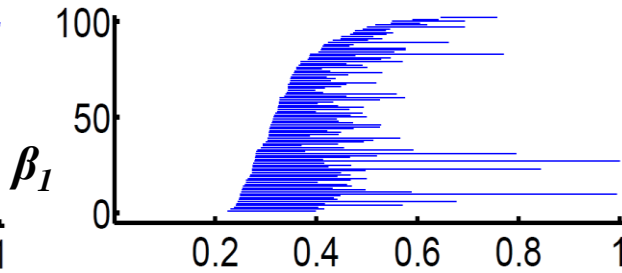
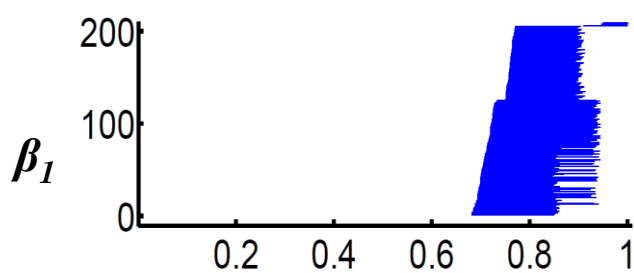
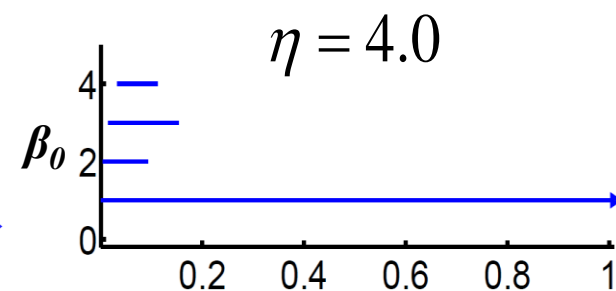
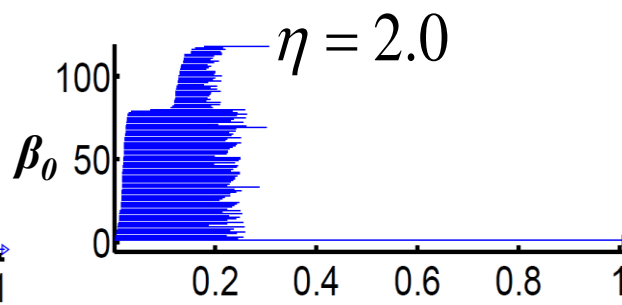
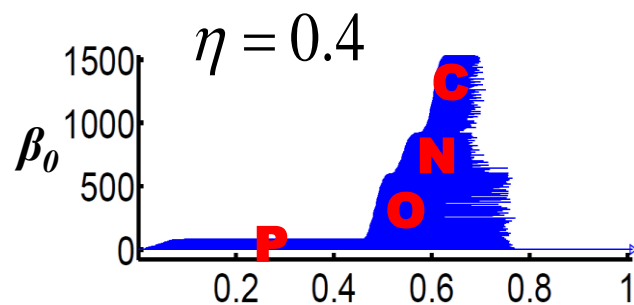
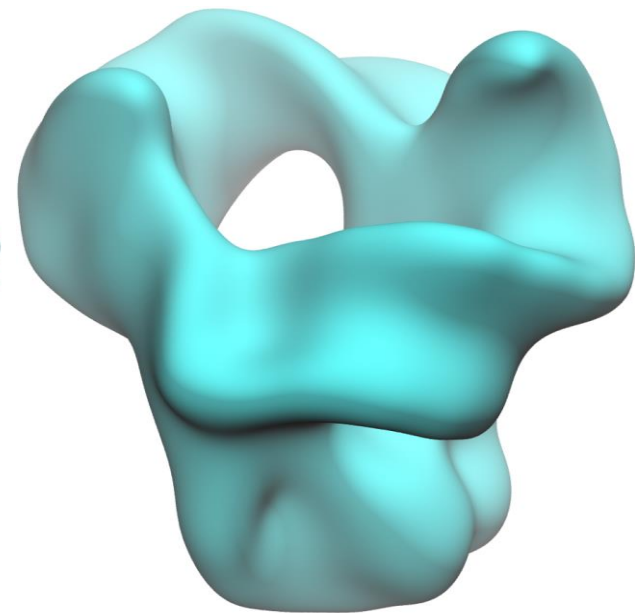
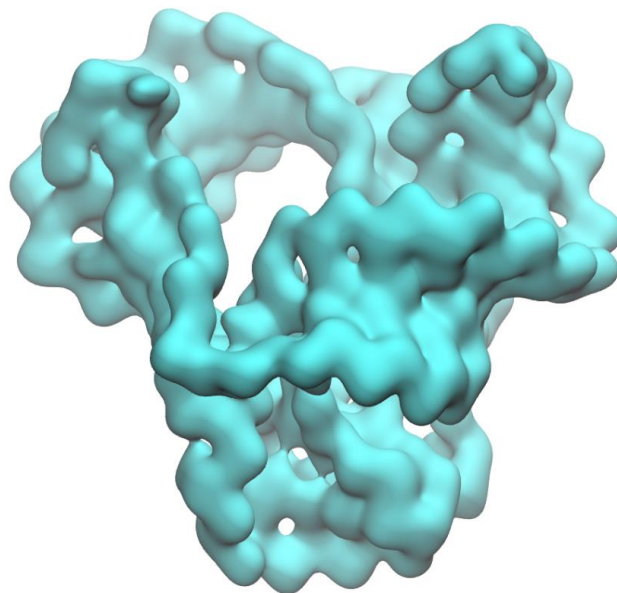
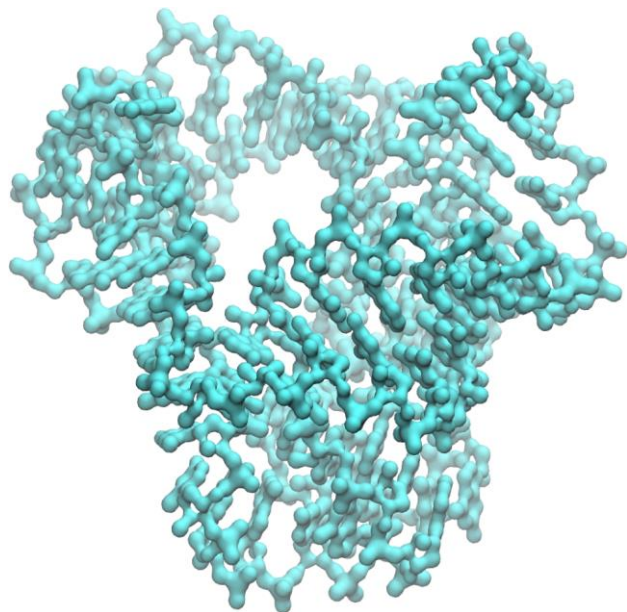
Point cloud representation



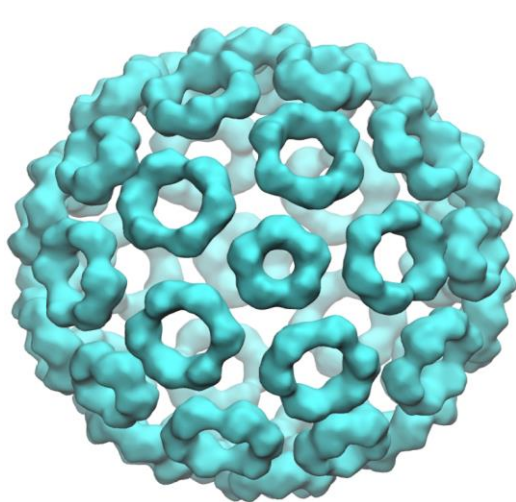
Isosurface extracted from rigidity function

PHA for multiresolution representations

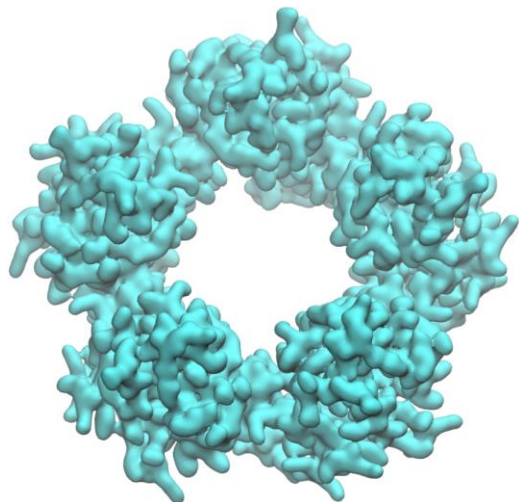
(Xia & Wei, JCB, 2015)



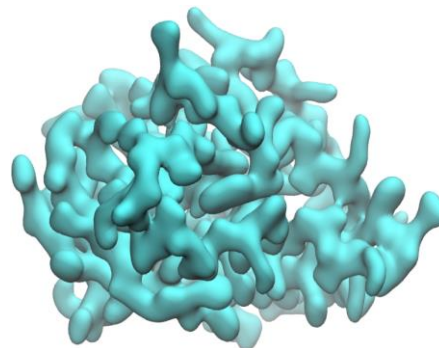
Multiresolution of the virus capsid



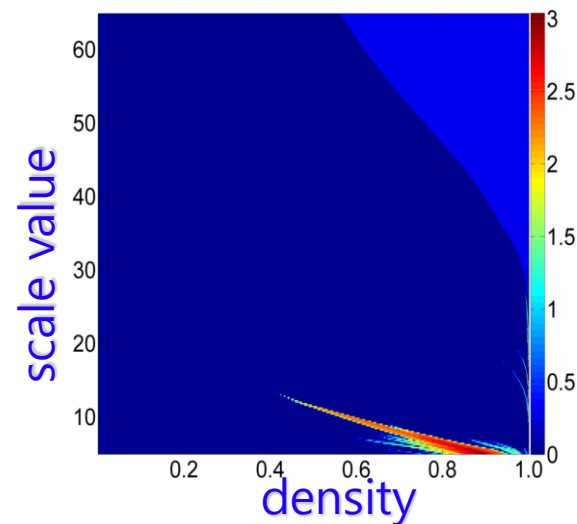
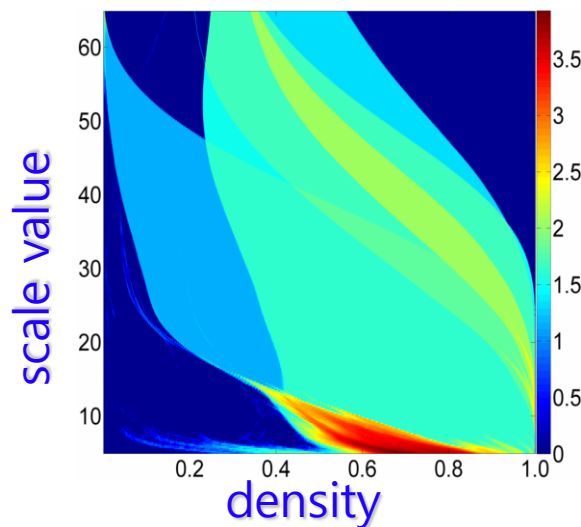
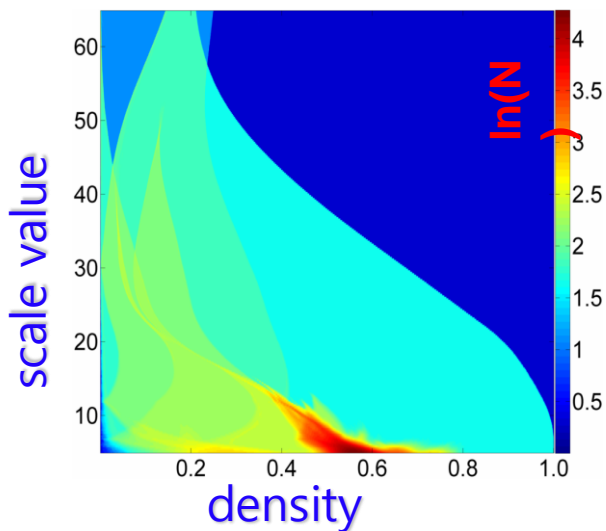
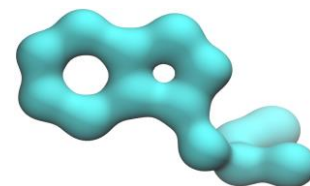
Betti-0



Betti-1



Betti-2



Topic--PHA for ill-posed inverse problems

**Microtubule
(EMD1129)**

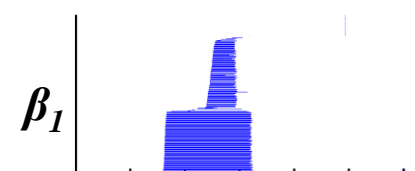
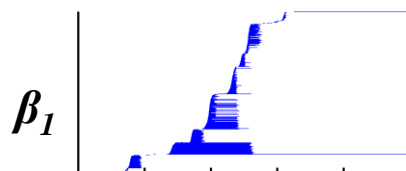
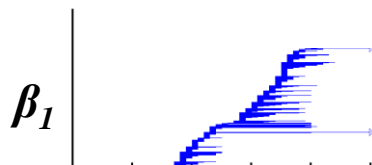
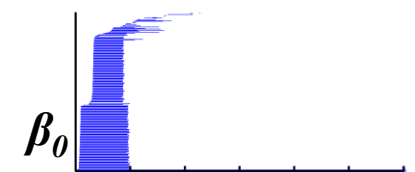
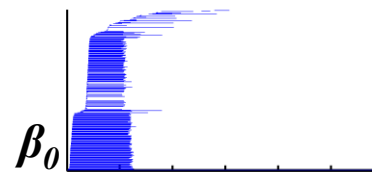
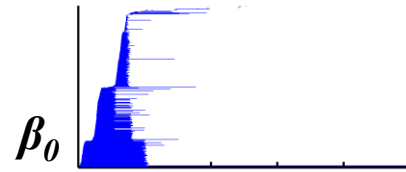
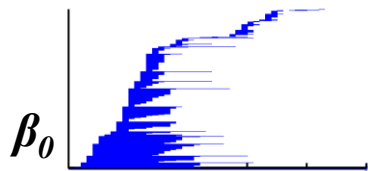
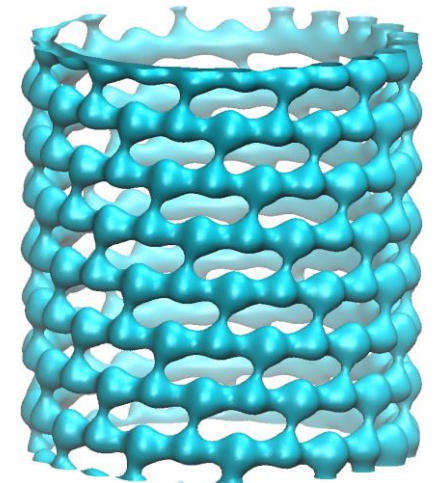
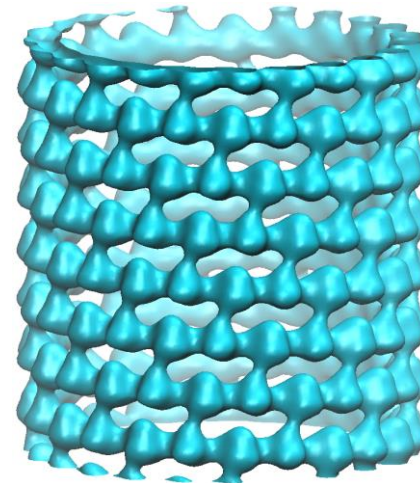
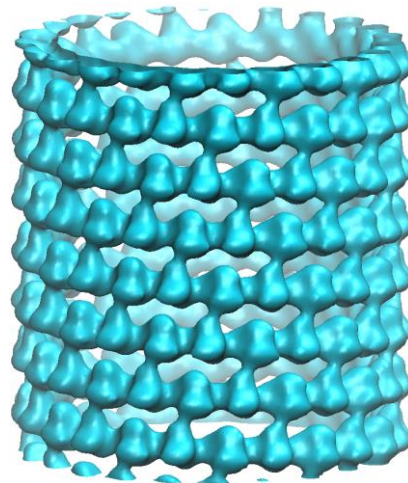
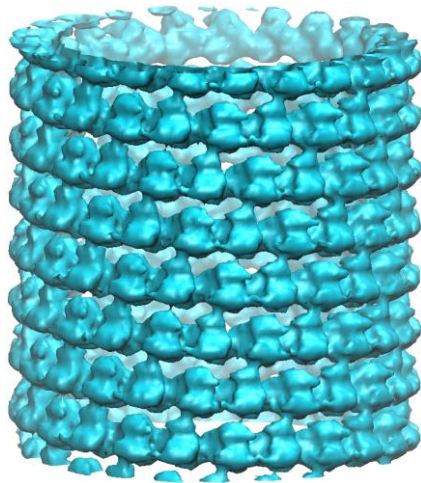
(Xia & Wei, IJNMBE, 2015)

**Original
data**

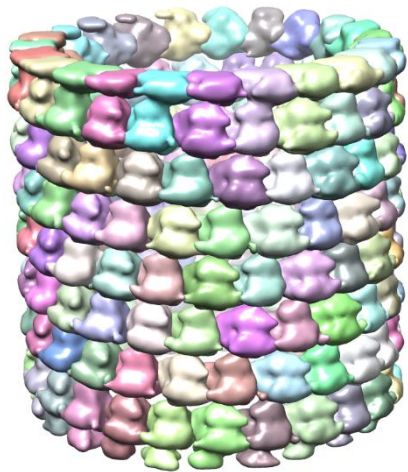
**Ten
iterations**

**Twenty
iterations**

**Forty
iterations**

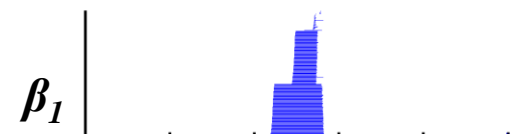
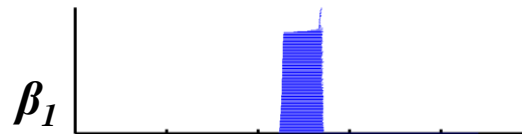
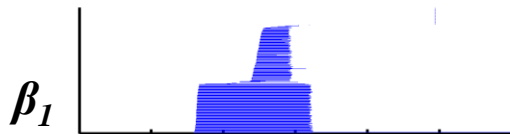
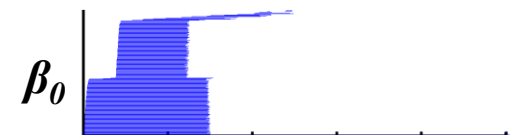
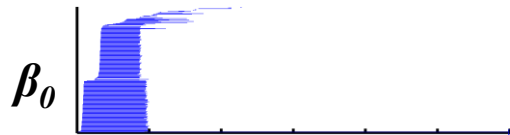
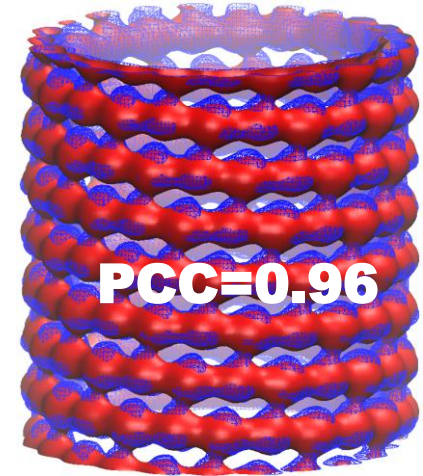
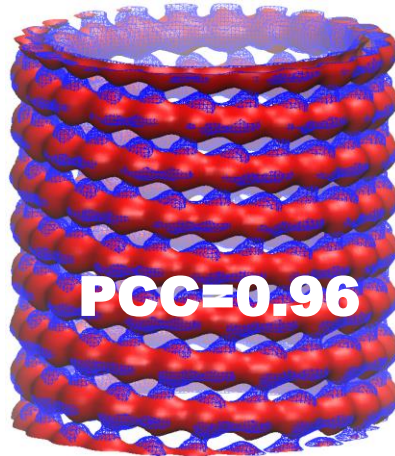
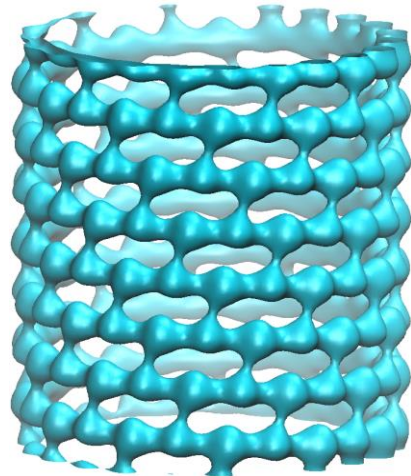


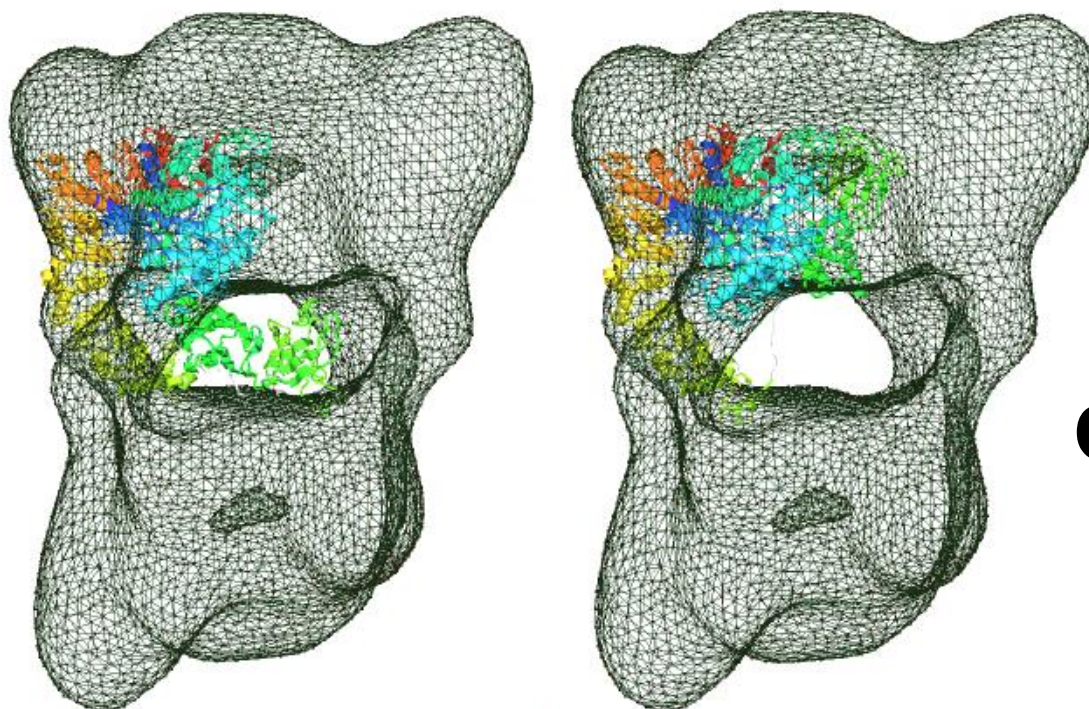
PHA for ill-posed inverse problems



Fitted with one-type of tubulins

Fitted with two-types of tubulins





Cry-EM docking

Figure 1. Manual fitting of predicted models of beta subunit into phosphorylase kinase using MVP-Fit. Left panel shows the result with rigid-body fitting while the right shows that with further local flexible fitting.

Molecular Dynamics Flexible Fitting

[Main](#) | [Method](#) | [Software](#) | [Docu](#)

Biophysical *Journal*
Article



Flexible Fitting of Atomic Models into Cryo-EM Density Maps Guided by Helix Correspondences

Hang Dou,^{1,*} Derek W. Burrows,¹ Matthew L. Baker,² and Tao Ju¹

¹Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri and ²Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas

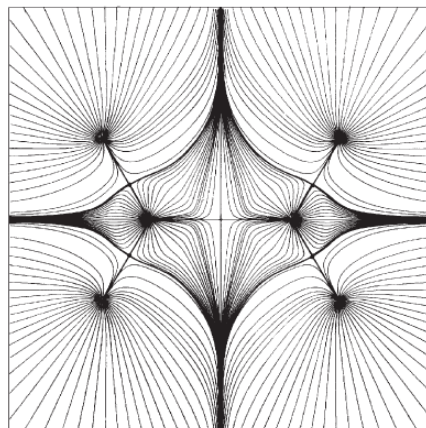
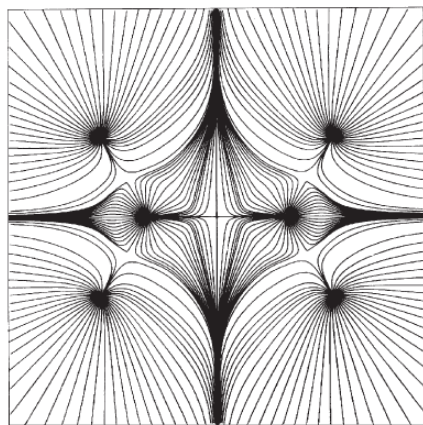
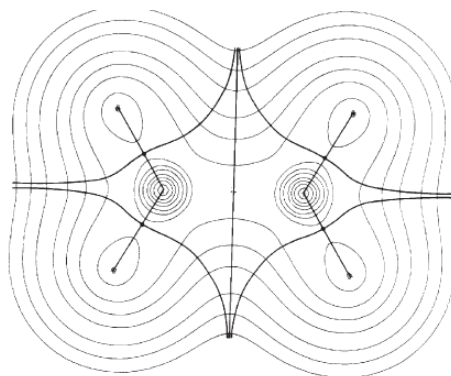
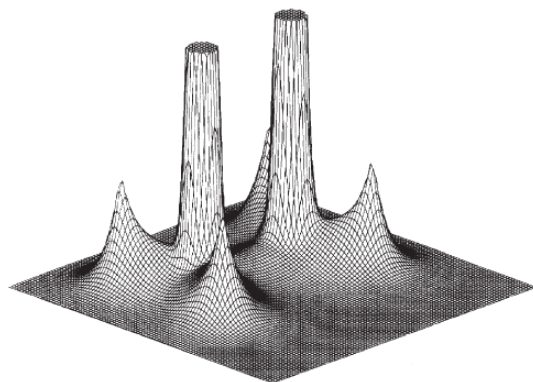
MDFF for cryo-EM

The molecular dynamics flexible fitting (MDFF) **method** can be used to fit the density map into a molecular dynamics (MD) simulation of the atomic

xMDFF for X-ray Crystallography

xMDFF is an MDFF-based approach for determining structures from low-iteratively updating electron density map. It addresses significant large-s

Use the menu above to navigate the MDFF website. For examples of MC



Atoms in Molecules

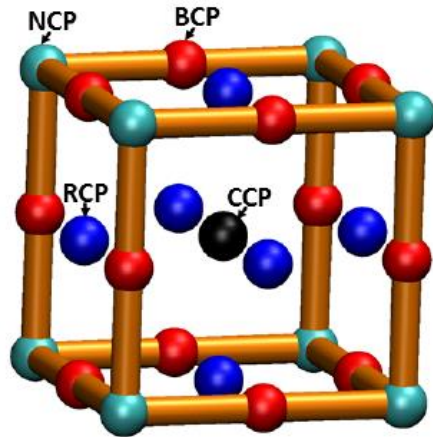
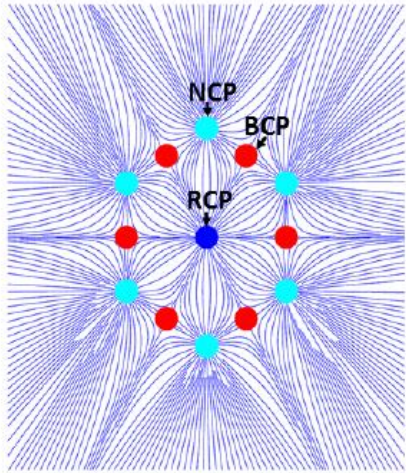
Richard F. W. Bader

McMaster University, Hamilton, Ontario, Canada

Table 1: The classification of critical points into four basic types, including nucleic critical point (NCP), bond critical point (BCP), ring critical point (RCP) and cage critical point (CCP), as demonstrated in Fig. 17.

	Rank	Signature	Poincaré index	Simplex	Property
NCP	3	-3	1	0-simplex	local maxima
BCP	3	-1	-1	1-simplex	saddle
RCP	3	1	1	2-simplex	saddle
CCP	3	3	-1	3-simplex	local minima

Atoms in molecule (On-going)

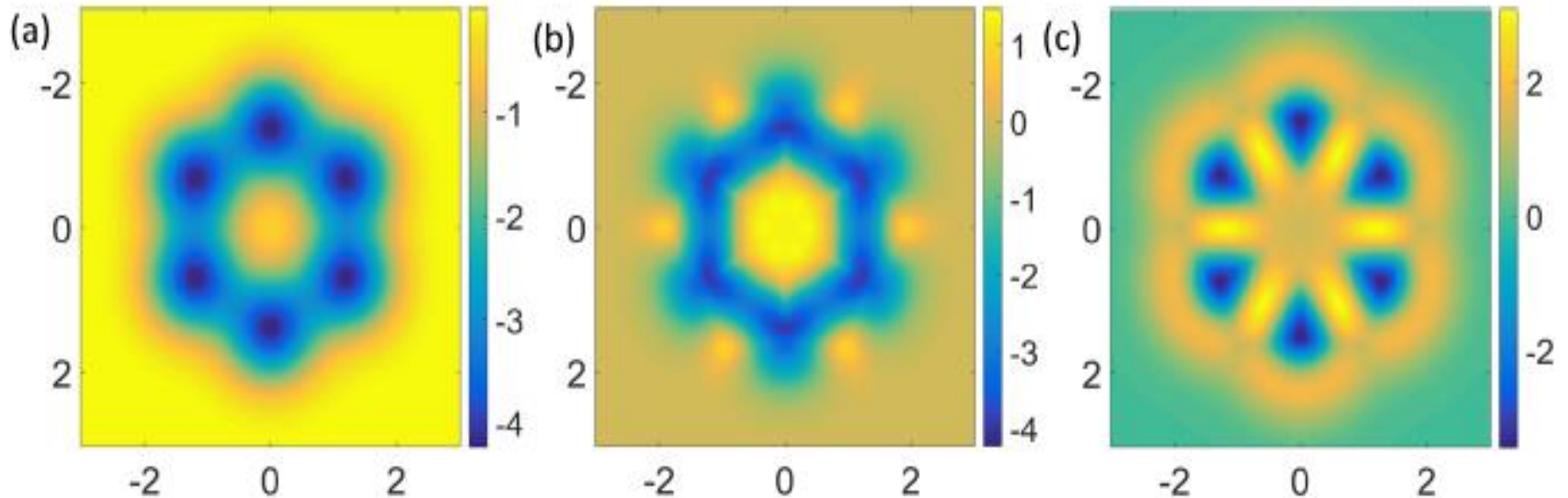


Atoms in Molecules

Richard F. W. Bader

McMaster University, Hamilton, Ontario, Canada

(Xia & Wei, arXiv, 2017)



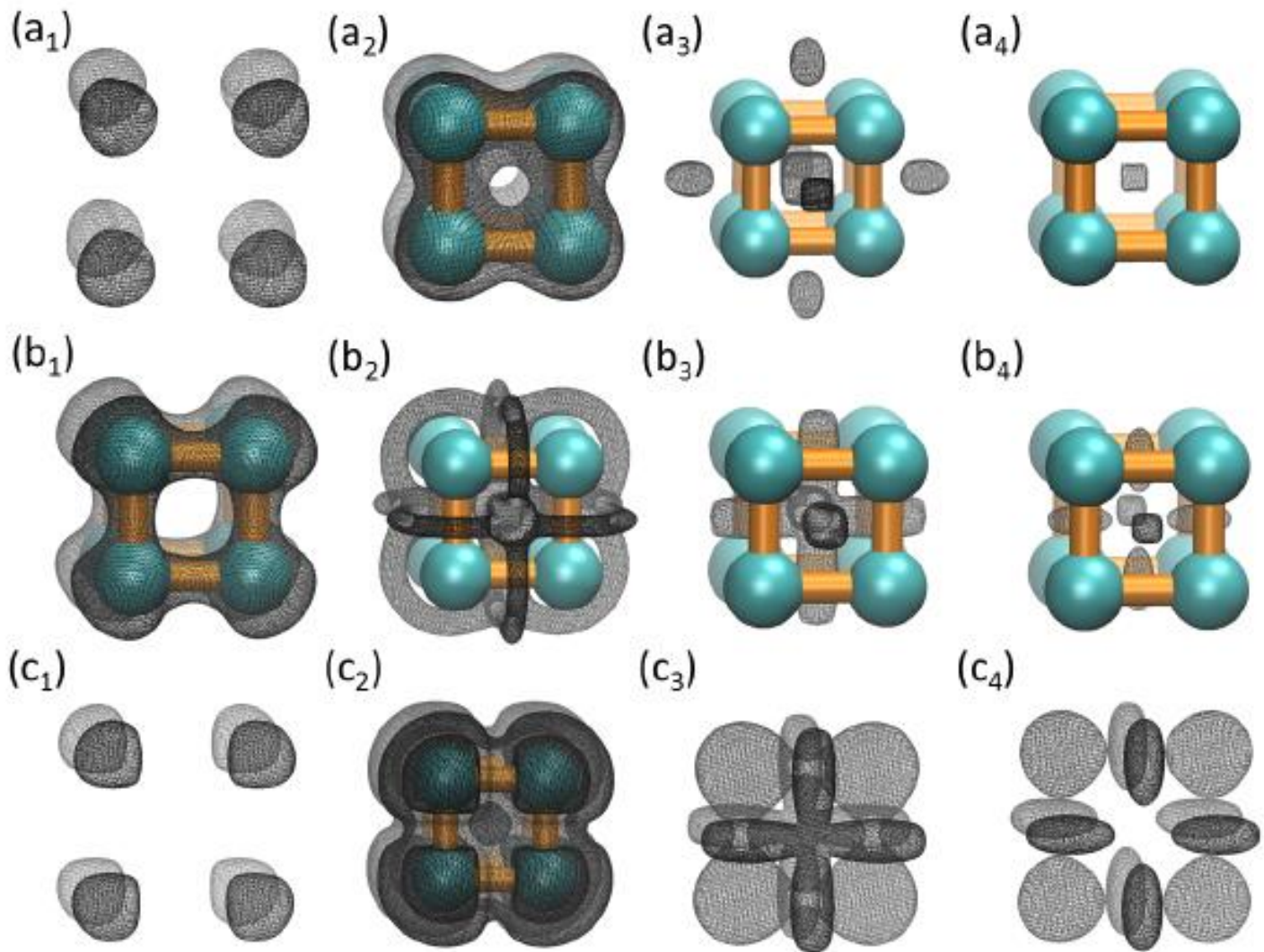
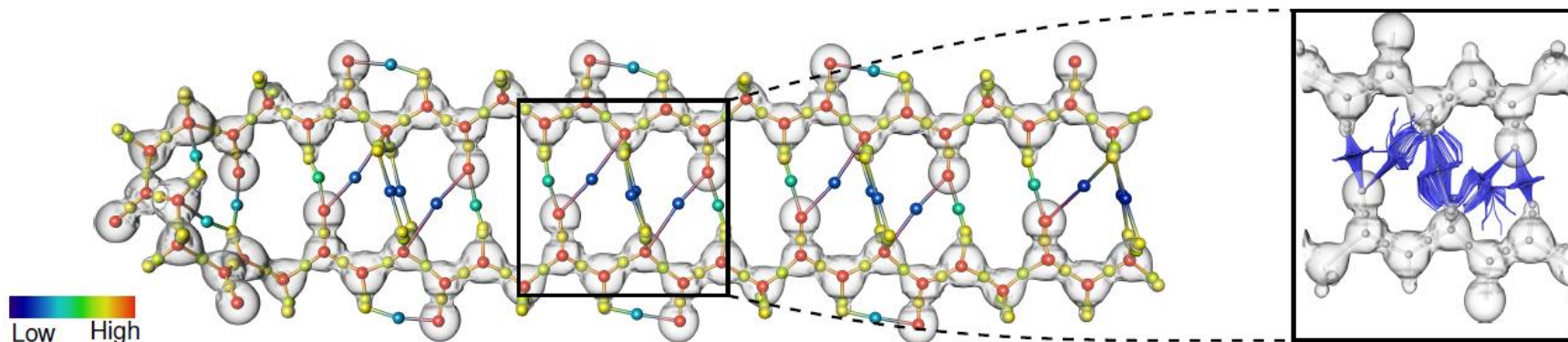





Figure 20: Eigenvalue maps obtained from different isovalues (or level-set values) for a cubic structure. (a) The isosurfaces for the first eigenvalue. The isovalues from (a_1) to (a_4) are -3.0, -1.5, 0.1 and 0.9. (b) The isosurfaces for the second eigenvalue. The isovalues from (b_1) to (b_4) are -1.0, 0.5, 1.0 and 1.5. (c) The isosurfaces for the third eigenvalue. The isovalues from (c_1) to (c_4) are -1.0, 1.5, 2.0 and 2.5.

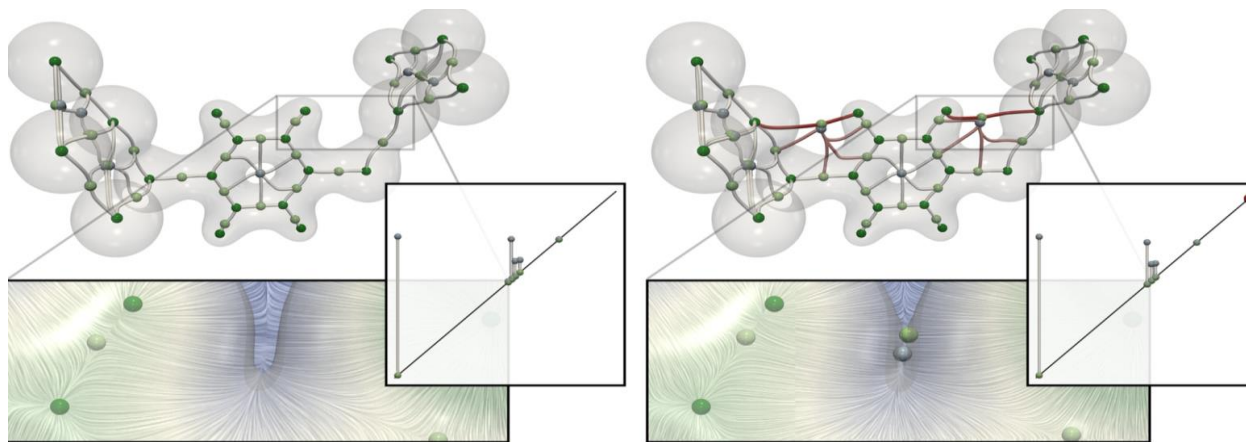
Characterizing Molecular Interactions in Chemical Systems

David Günther, Roberto A. Boto, Julia Contreras-Garcia, Jean-Philip Piquemal, Julien Tierny

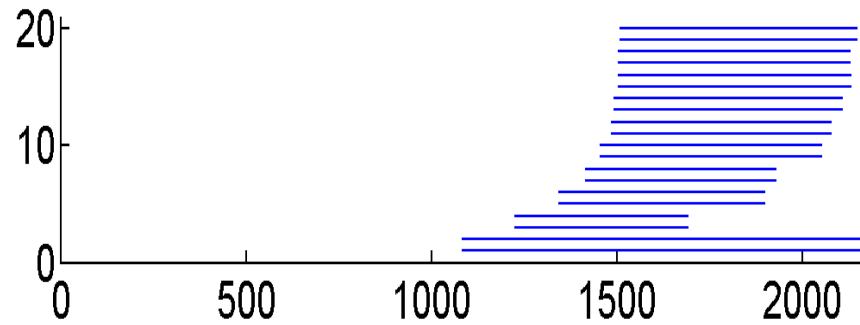
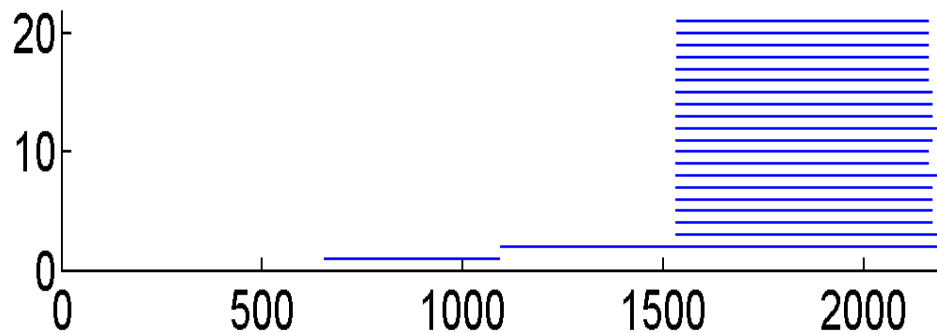
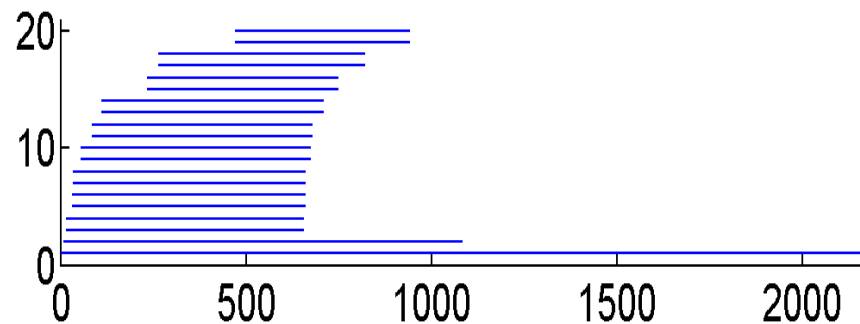
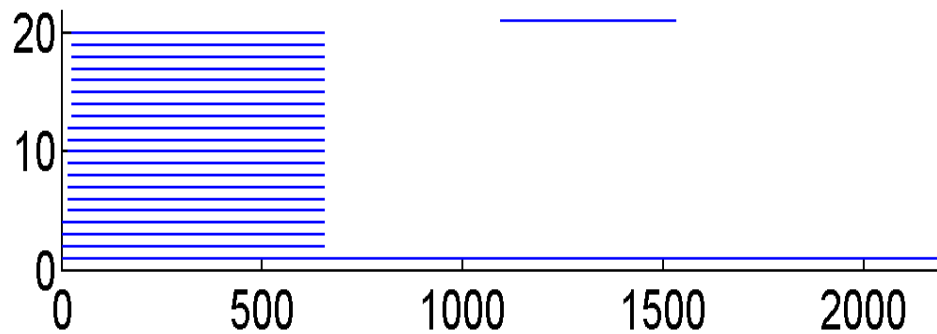
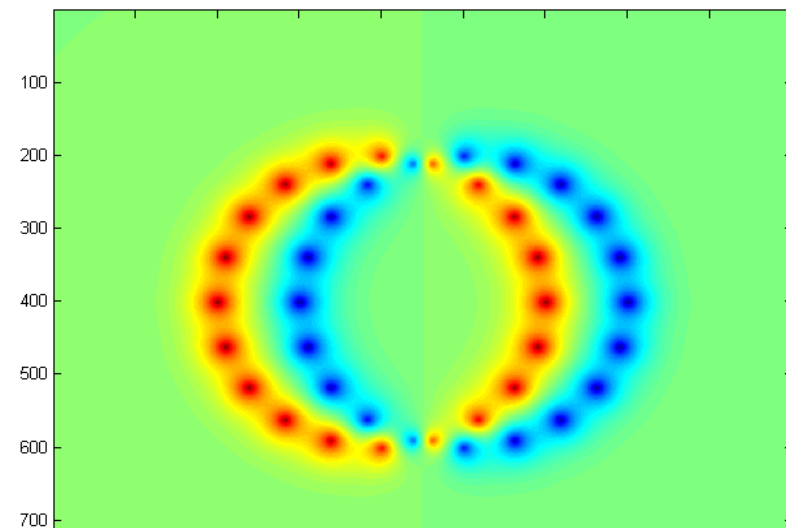
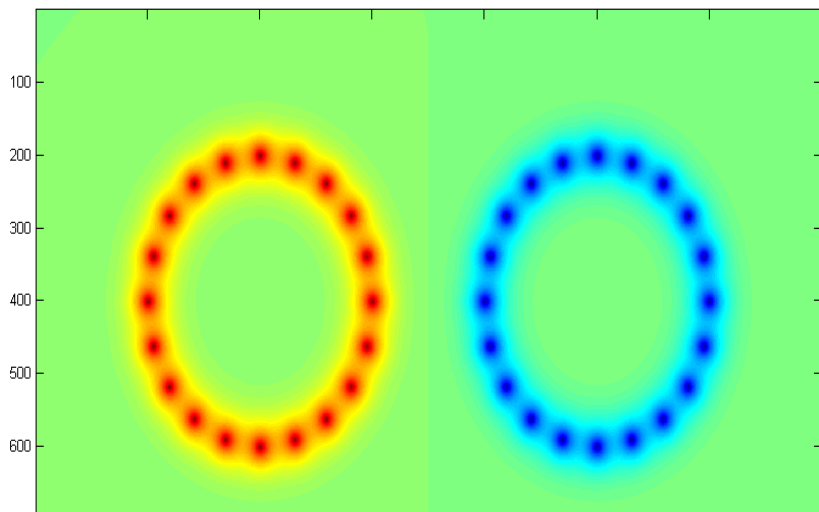


A Topological Data Analysis perspective on noncovalent interactions in relativistic calculations

Małgorzata Olejniczak¹  | André Severo Pereira Gomes²  | Julien Tierny³ 

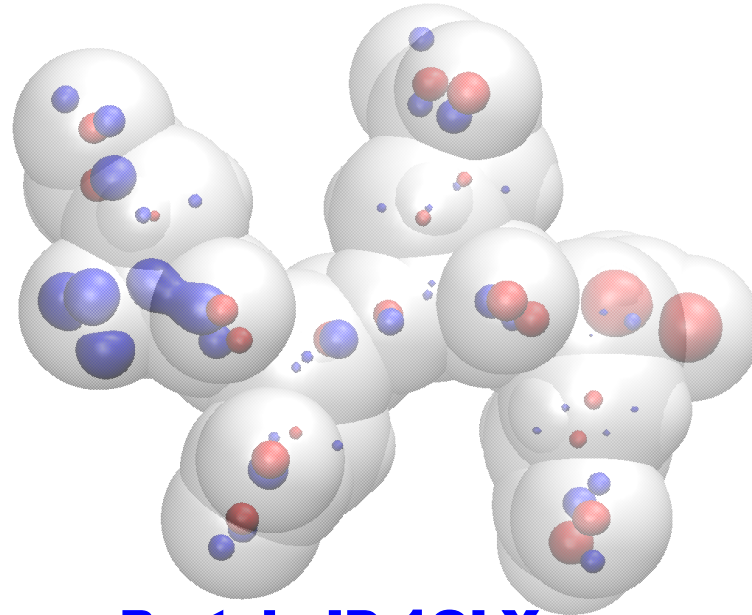


$$potential = \sum_i e^{-\left(\frac{|r-r_i|}{\sigma}\right)^2} - \sum_i e^{-\left(\frac{|r-r_i|}{\sigma}\right)^2}$$



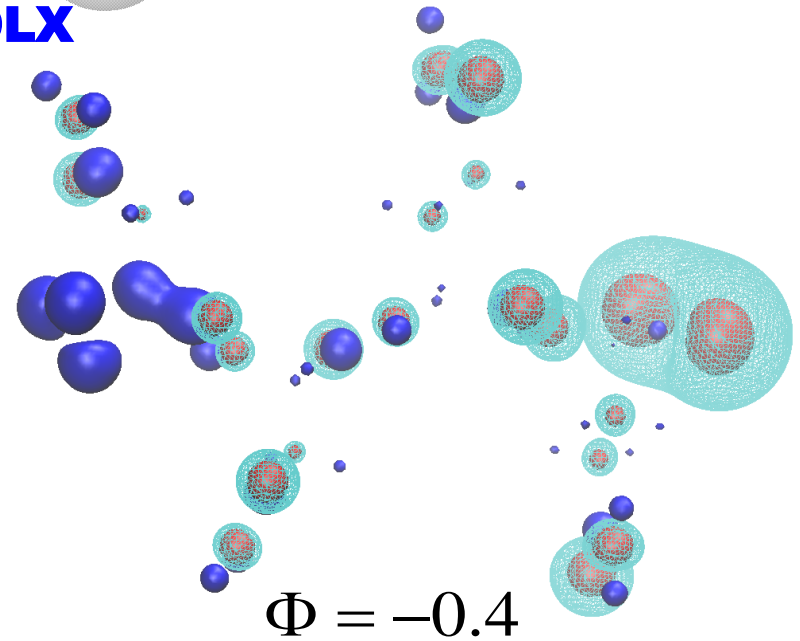
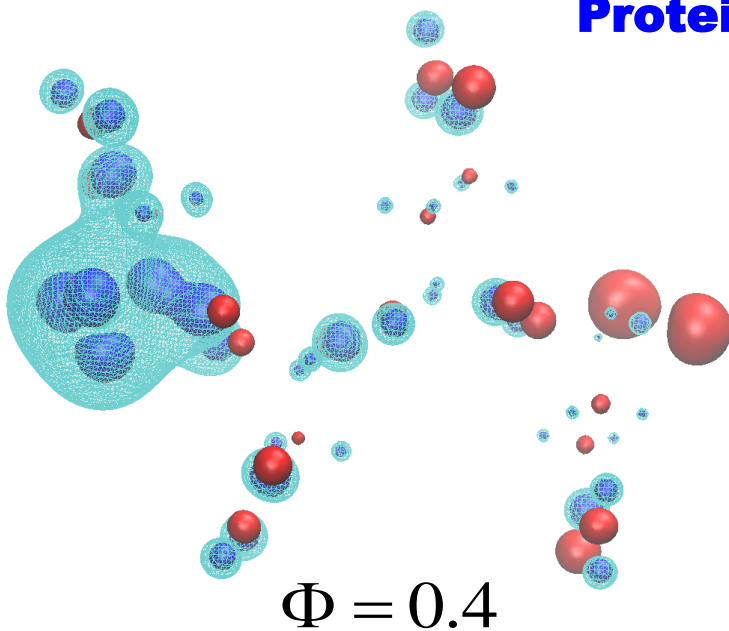
Protein Electrostatic Potential

$$\Phi = \sum_j \frac{q_j}{\epsilon_0 |r - r_j|}$$

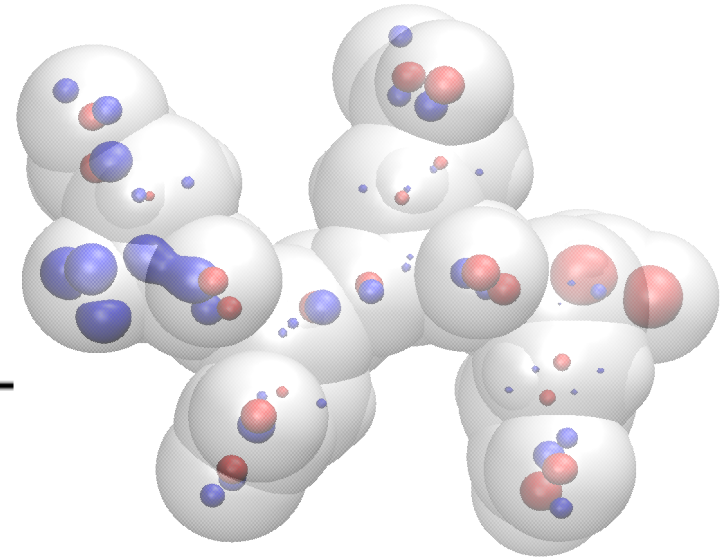
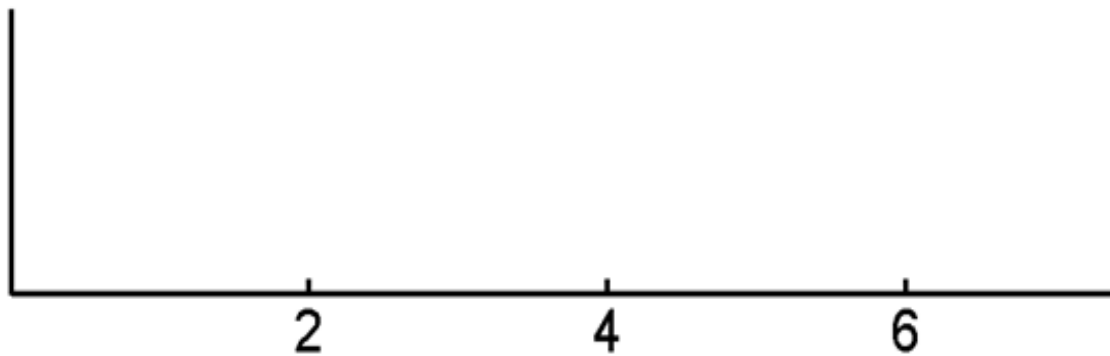
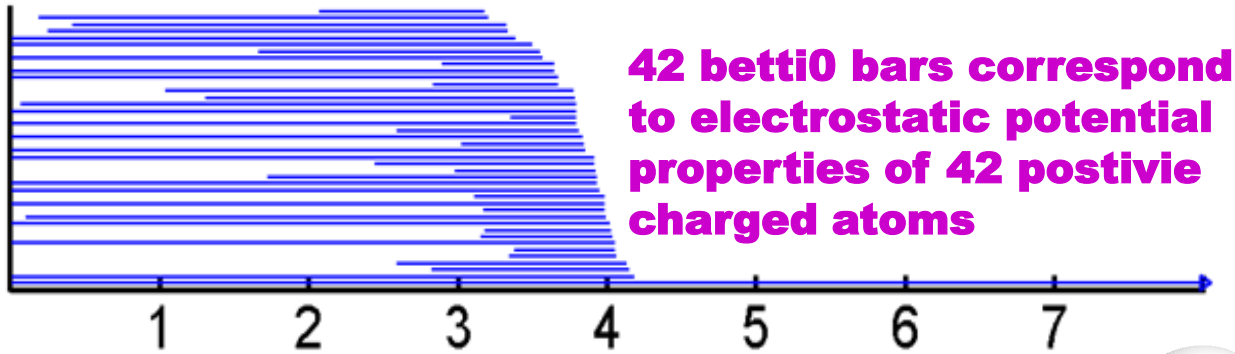


The **positive** charges are represented by **blue color**
The **negative** charges are marked by **red color**

Protein ID:1OLX



Persistent homology analysis for electrostatic potential



Topic-- Weighted persistent homology

Collaborator
Jie Wu
Math, NUS



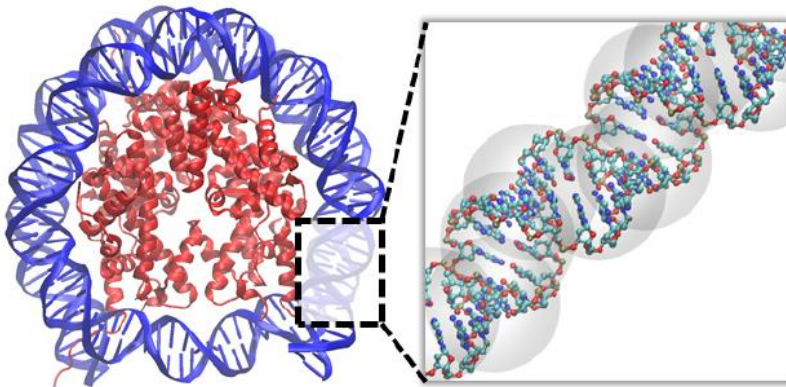
- **Weighted alpha complex;**
- **Weighted Vietoris-Rips;**
- **k-distance based models;**
- **Rigidity function based models;**
- **Weighted clique rank homology;**
- **Physics-aware models;**
- **Weighted simplicial homology;**
- **.....**

- *New filtration*
- *Weighted boundary map*

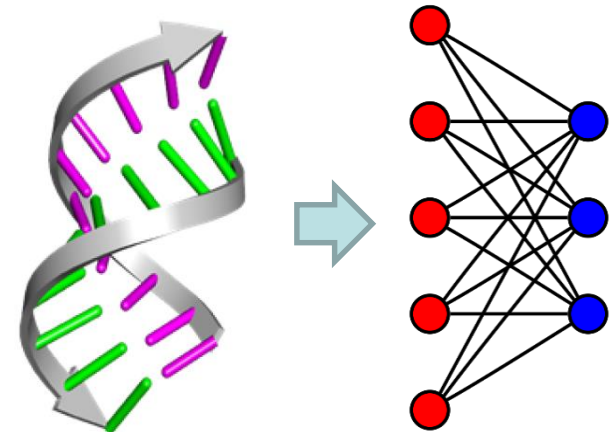
$$\partial_n(\sigma) = \sum_{i=0}^n \frac{w(\sigma)}{w(d_i(\sigma))} (-1)^i d_i(\sigma)$$

Labels: **Simplex weight** (pointing to $w(\sigma)$), **Boundary operator** (pointing to $d_i(\sigma)$), **n-Simplex** (pointing to the summation index n)

Localized Persistent homology (LPH)

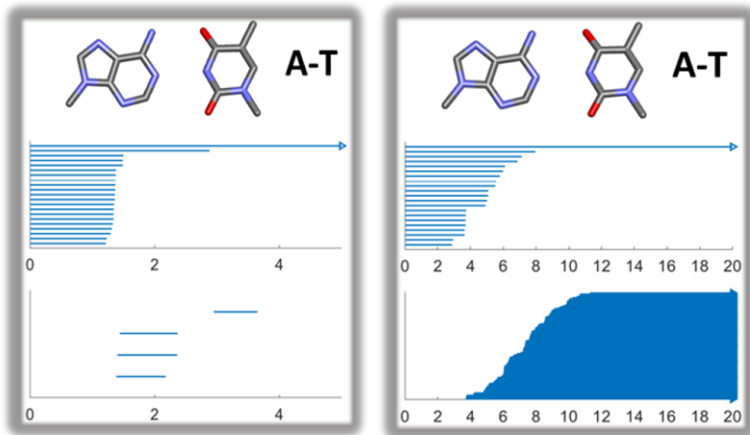


Interactive Persistent homology (IPH)

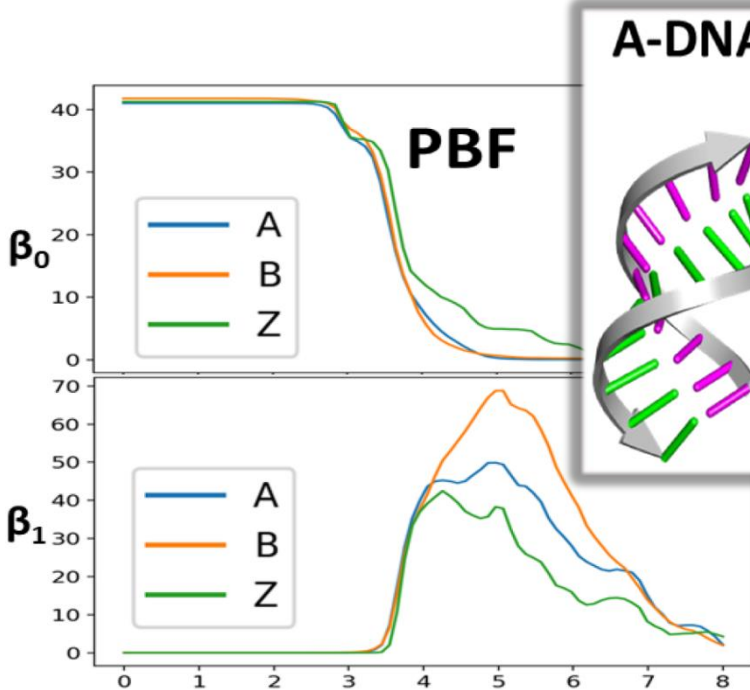
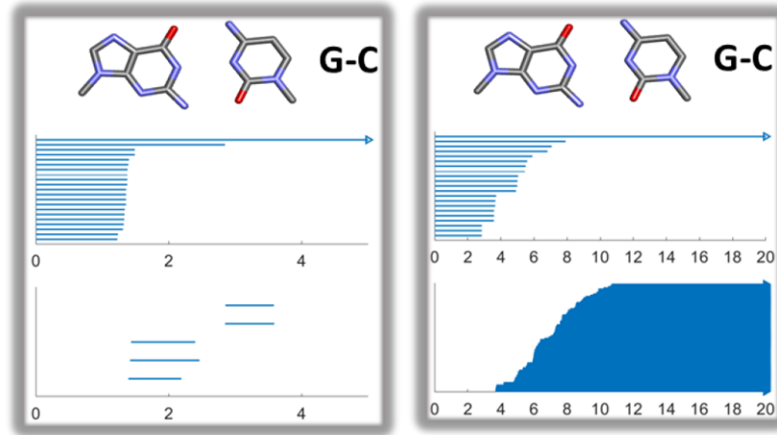


WPH for DNA classification

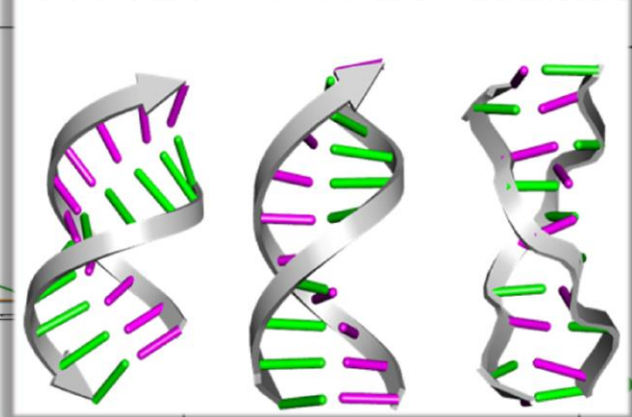
PH VS Interactive PH (AT)



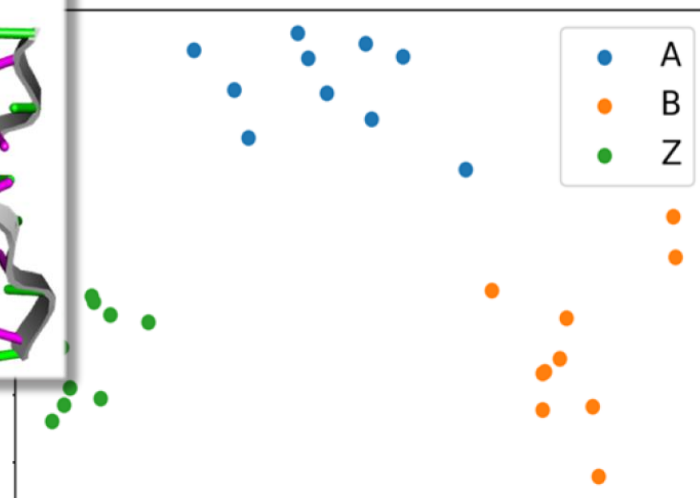
PH VS Interactive PH (GC)



A-DNA B-DNA Z-DNA



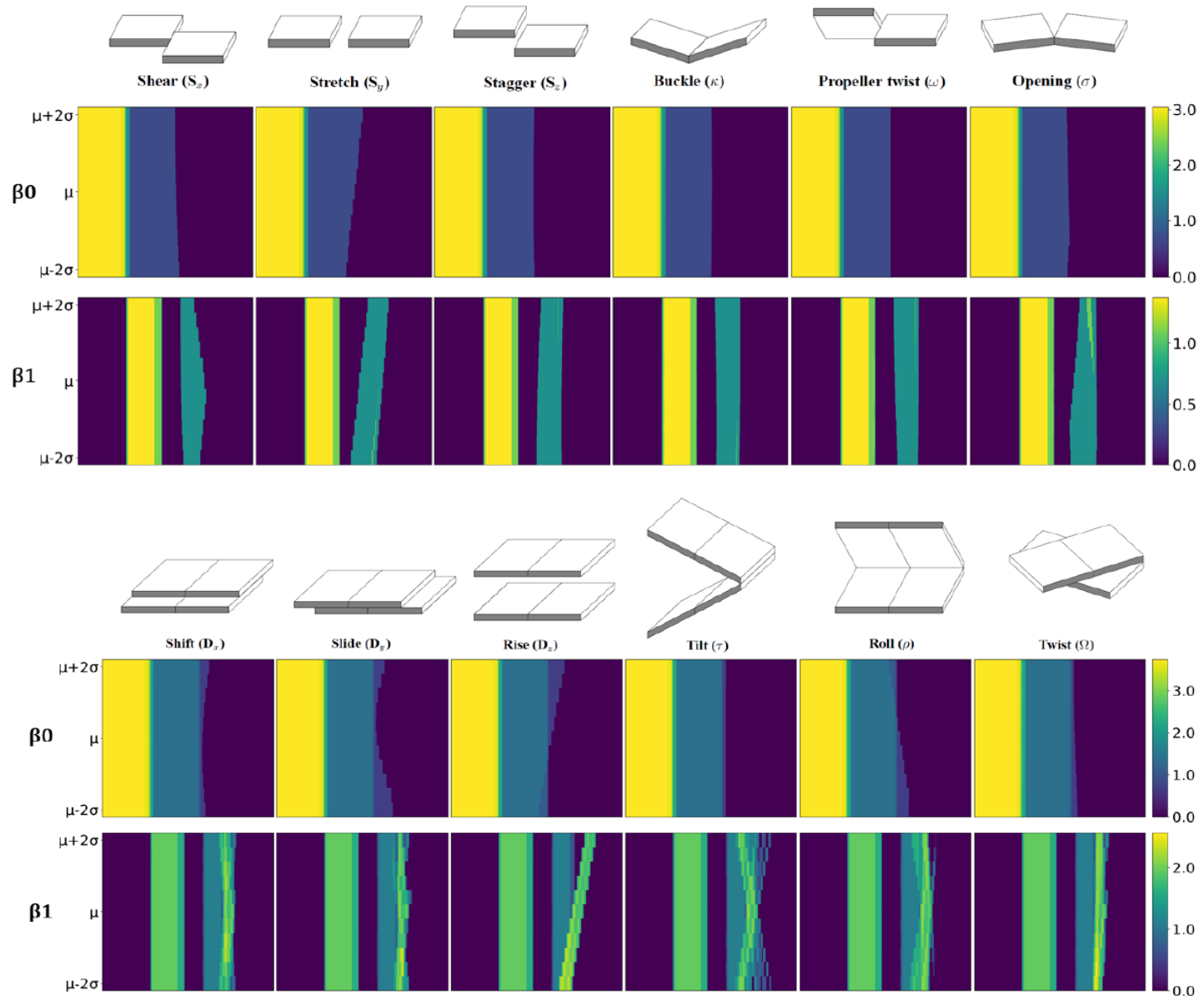
PCA

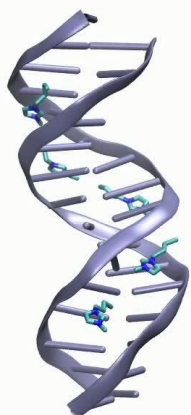


(Meng, Anand, Lu, Wu, and Xia, Sci. Rep., 2020)

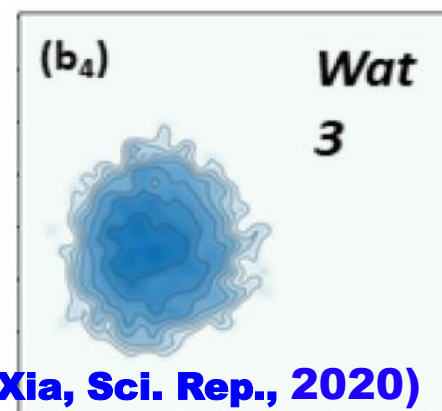
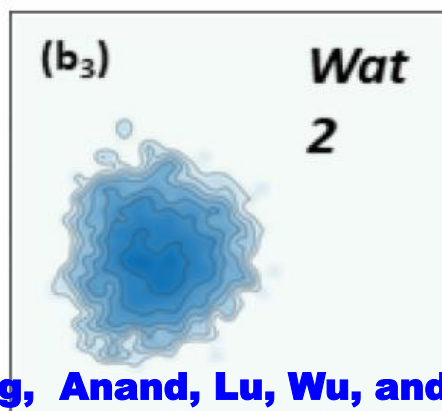
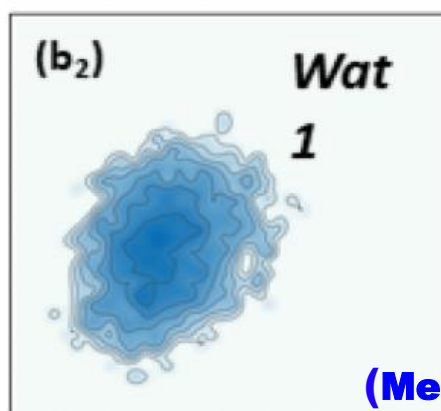
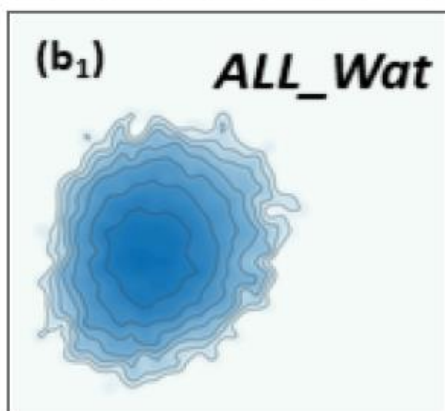
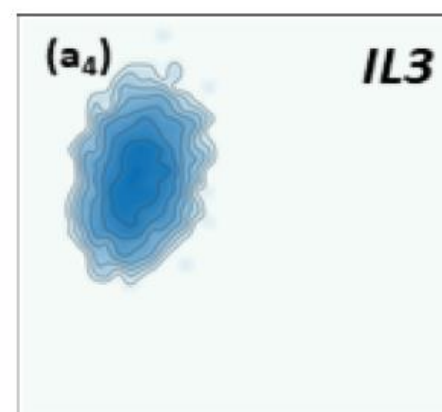
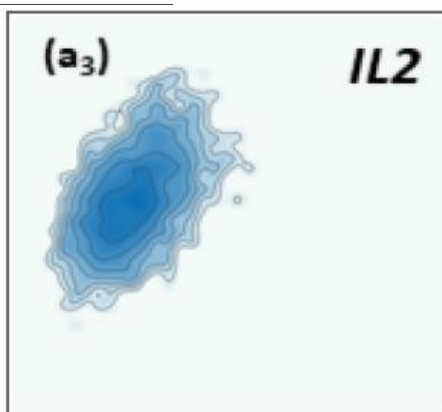
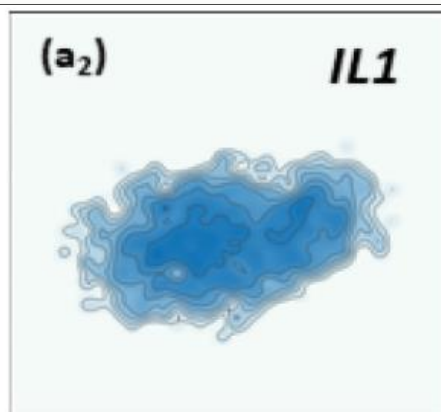
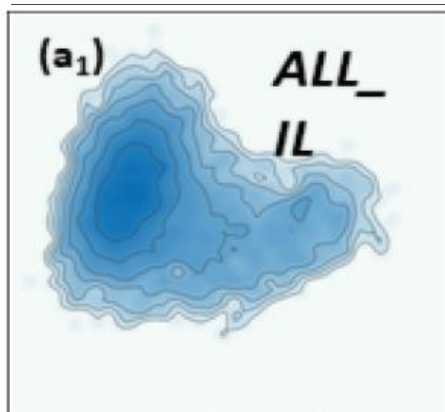
Cambridge University Engineering Department

Helix computation Scheme (CEHS)

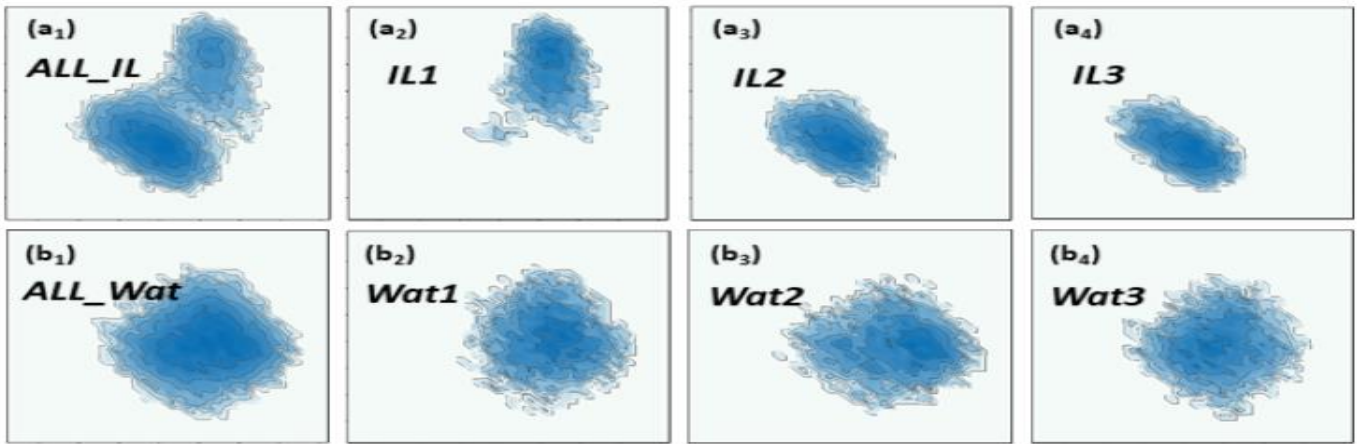




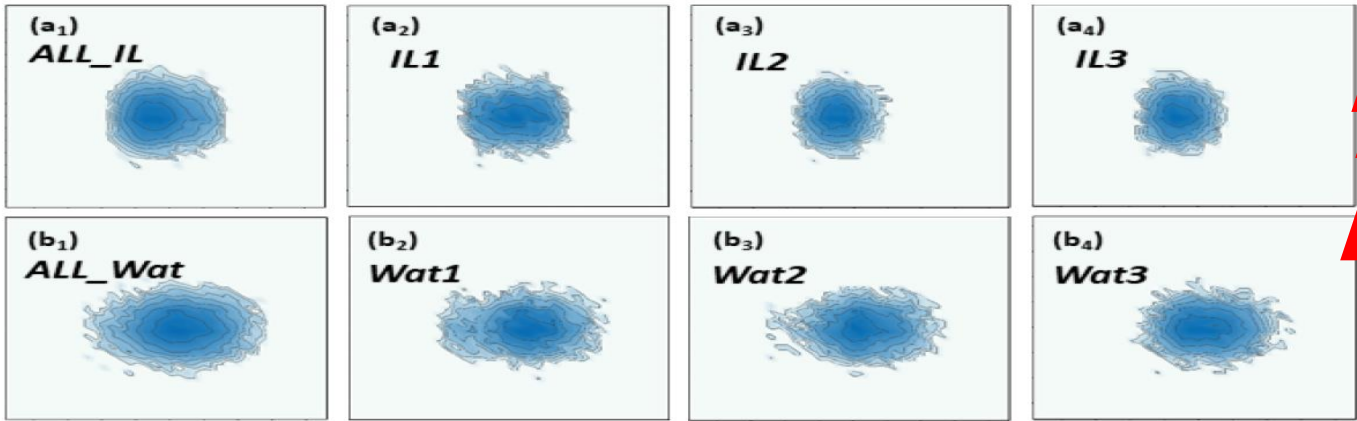
WPH for DNA clustering



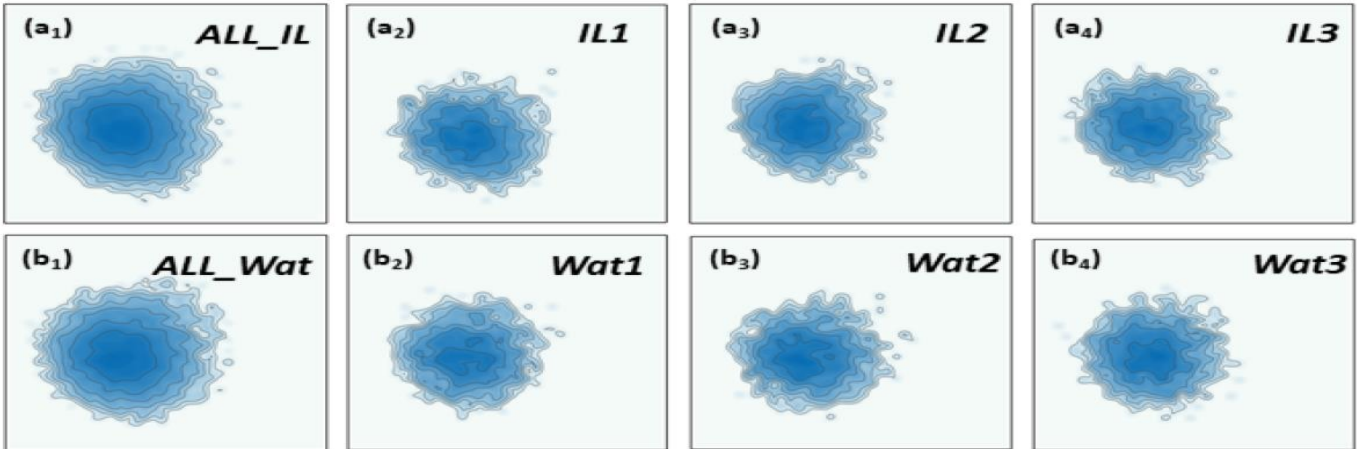
(Meng, Anand, Lu, Wu, and Xia, Sci. Rep., 2020)



CEHS-PCA

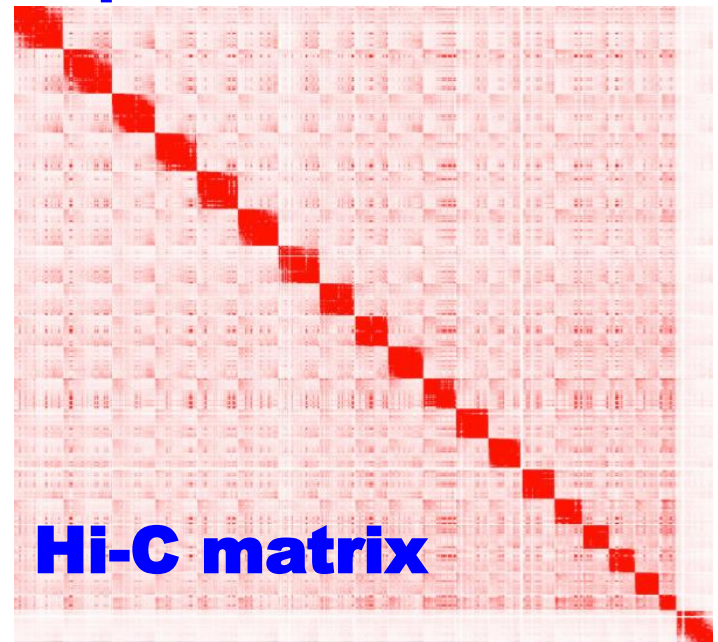
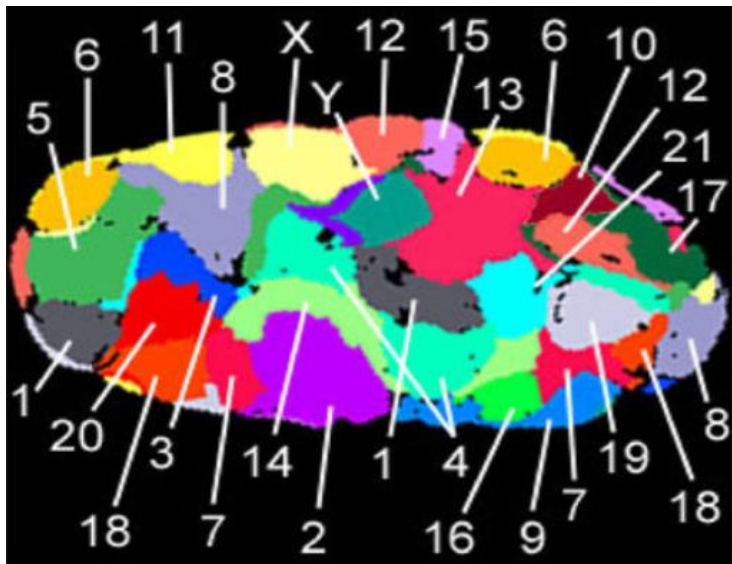
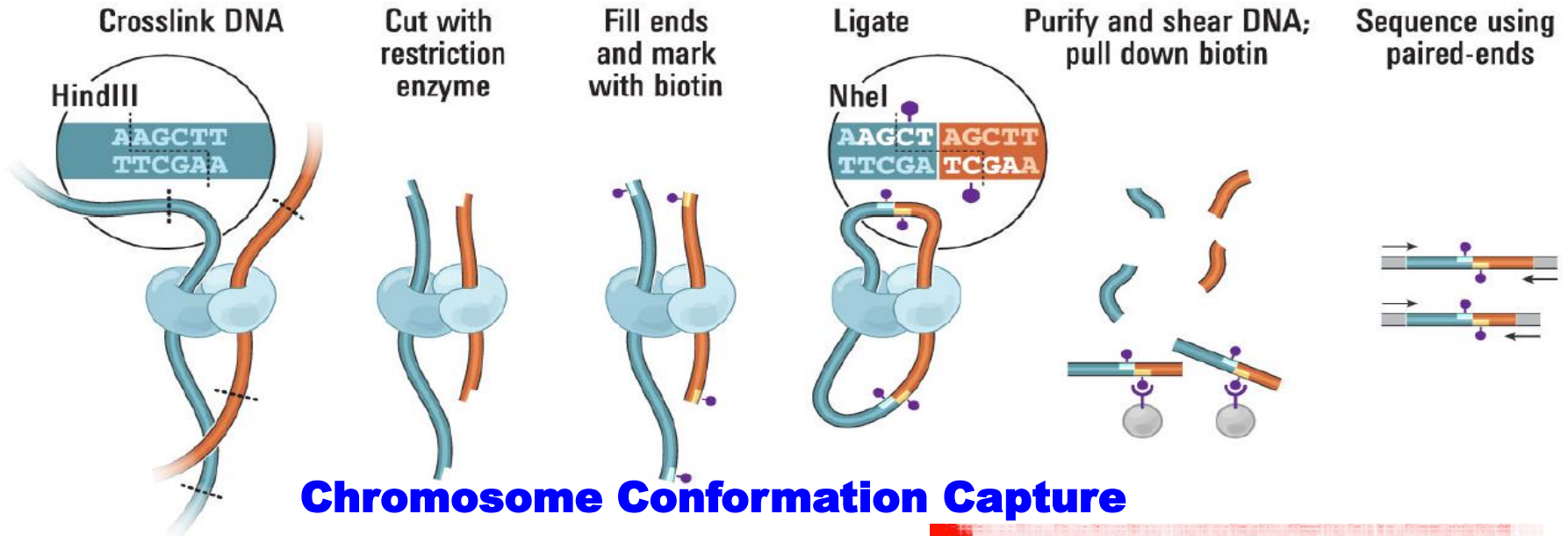


**Traditional
PCA**

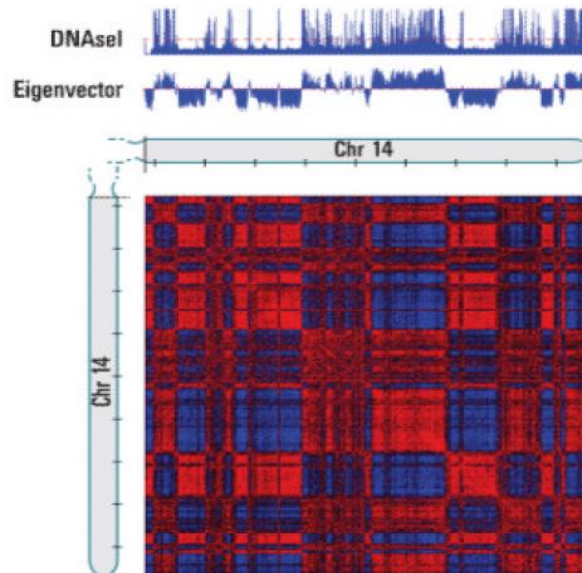
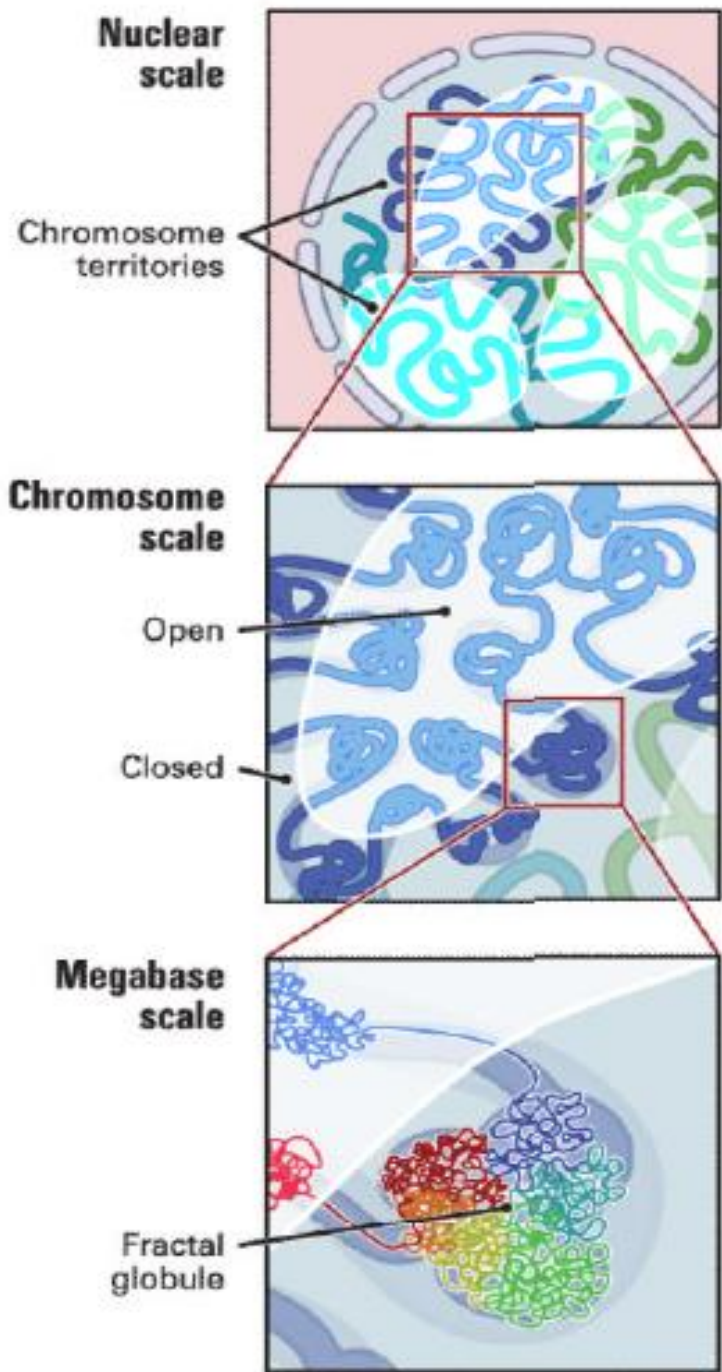


**Normal PH
PCA**

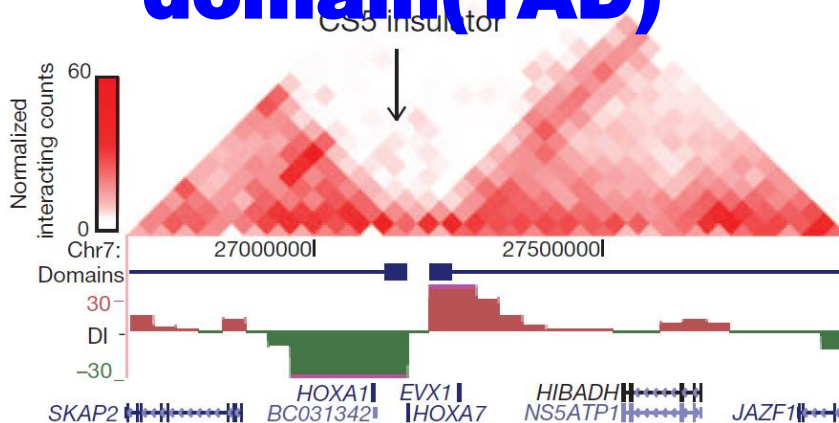
Hi-C data analysis



Genomic compartment



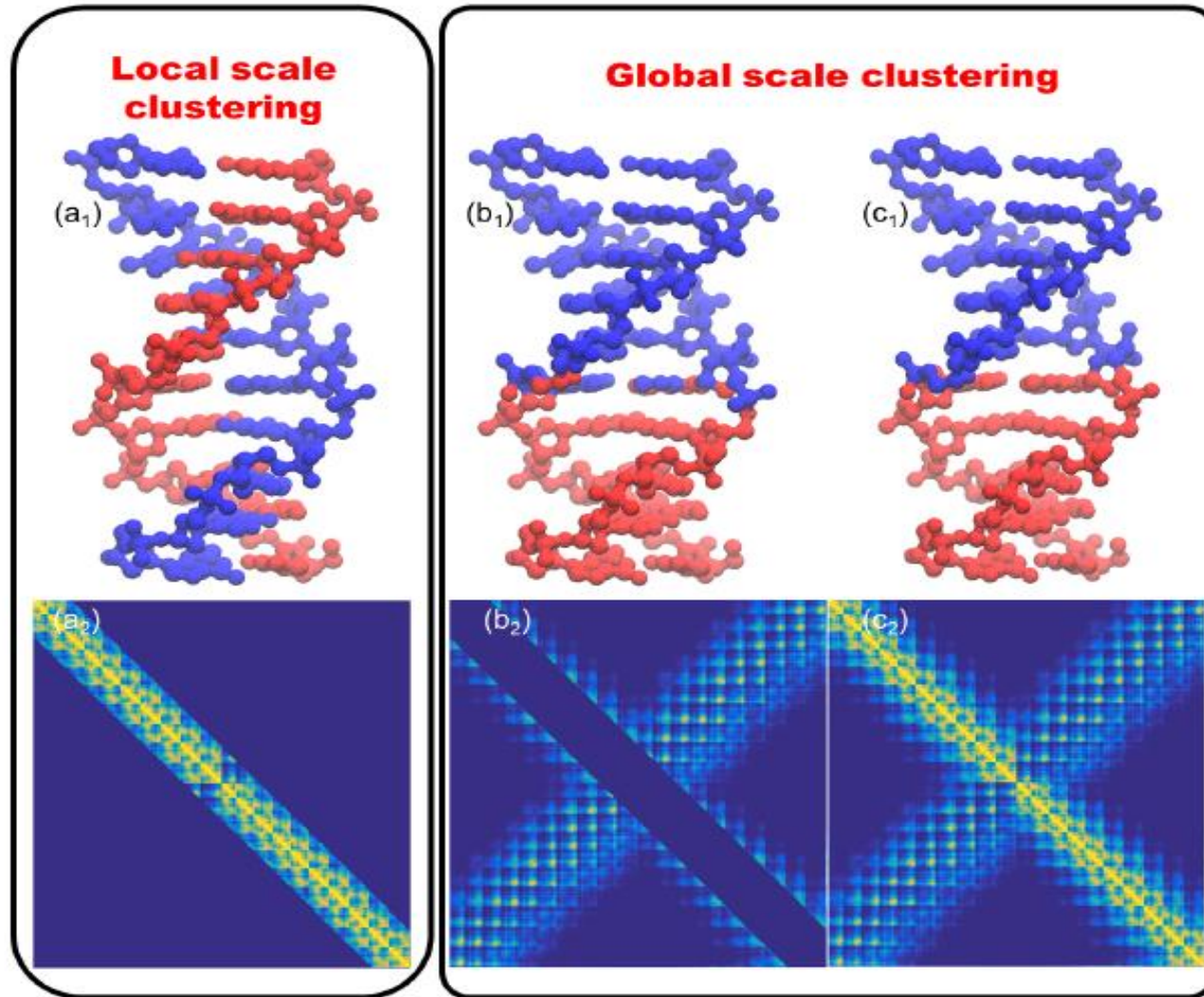
Topological associated domain (TAD)



Megabase-sized local chromatin domain

Sequence-based multiscale models

Kelin Xia, PLOS ONE, 2018



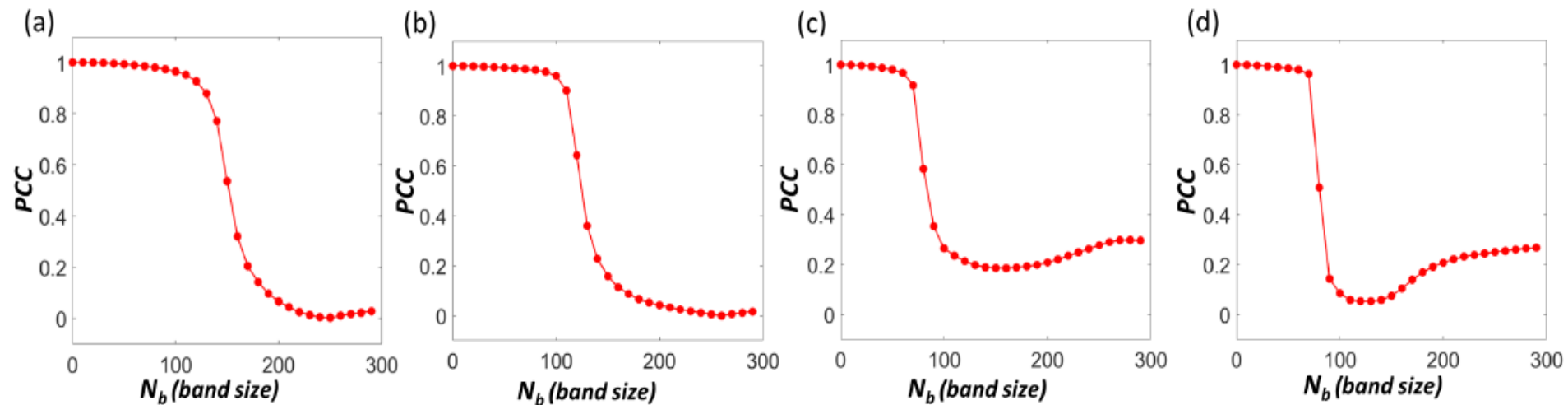
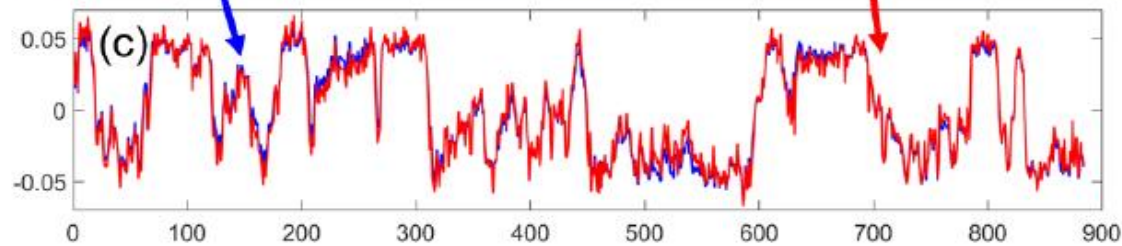
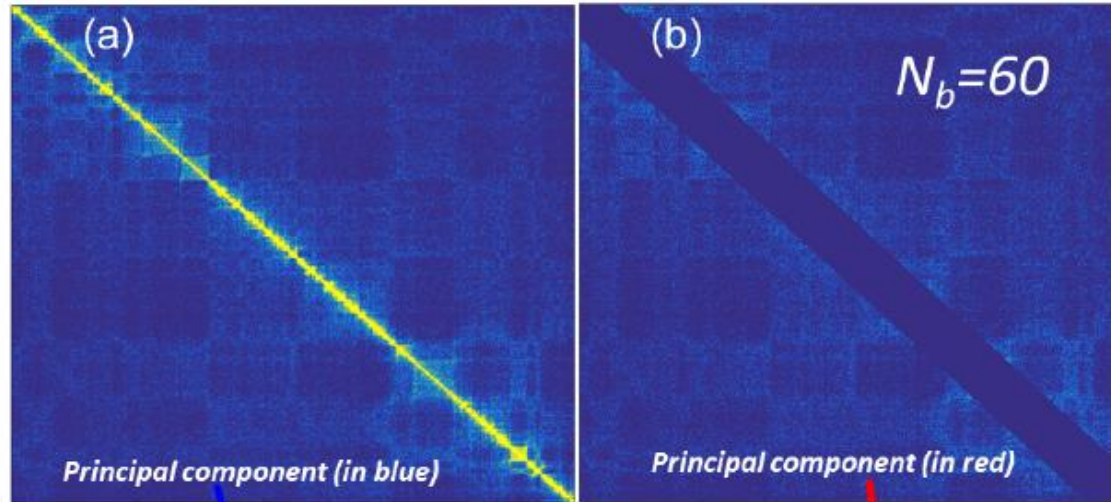
**Optimize
sequence
information**

**Optimize
spatial
information**

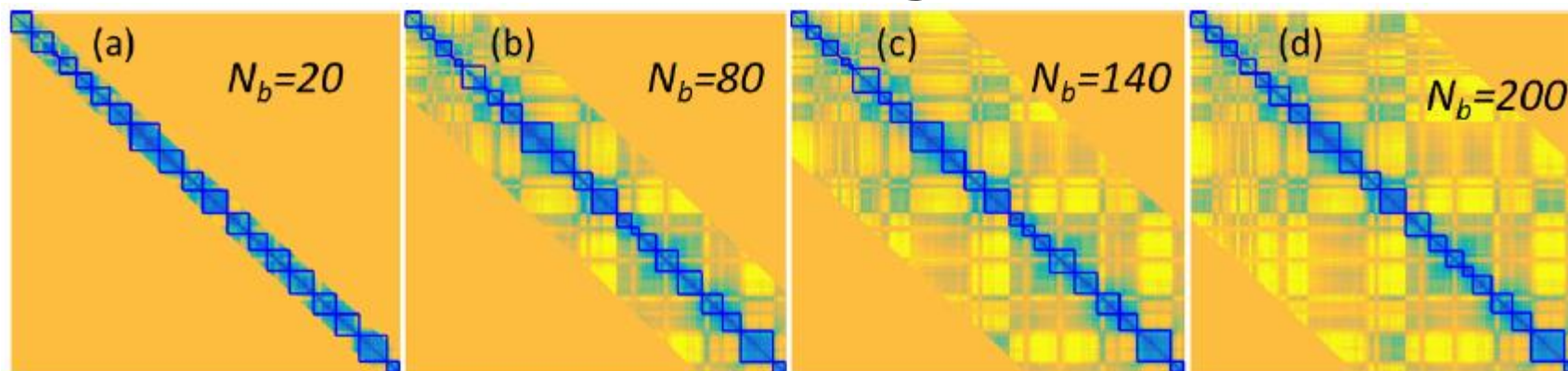
TAD

Genomic compartment

Genomic compartment analysis

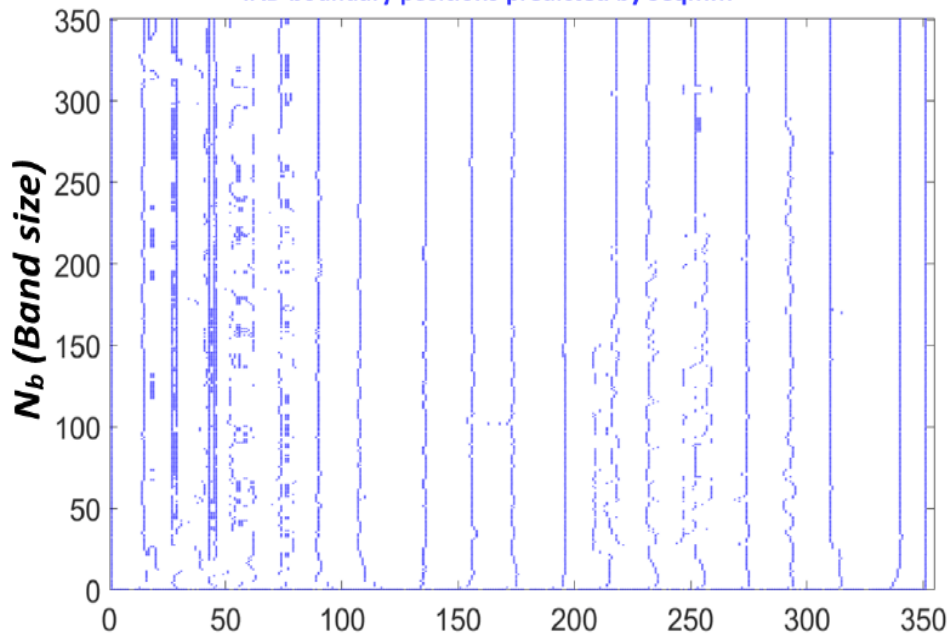


TAD analysis



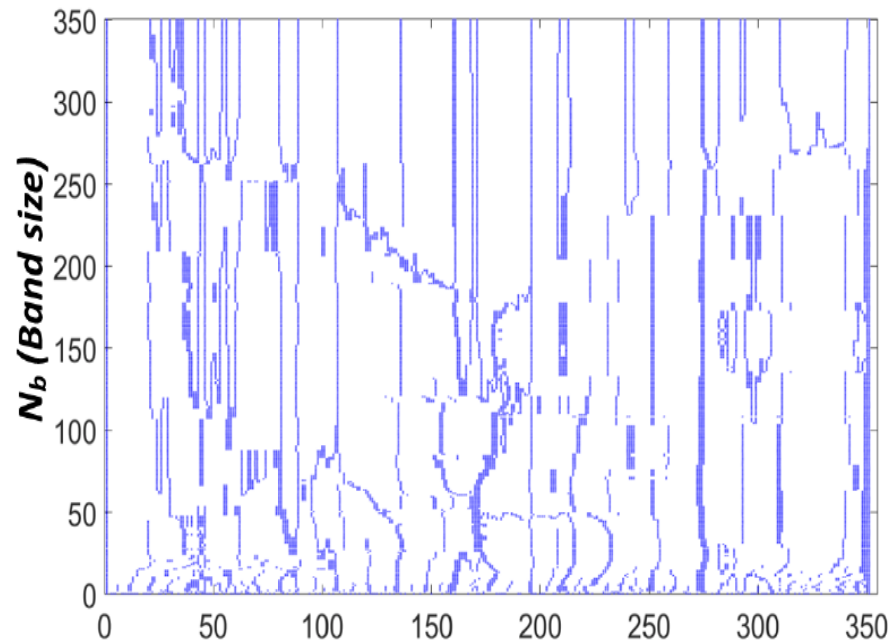
IMR90, cell chromosome 22, Resolution:100kb

TAD boundary positions predicted by SeqMM



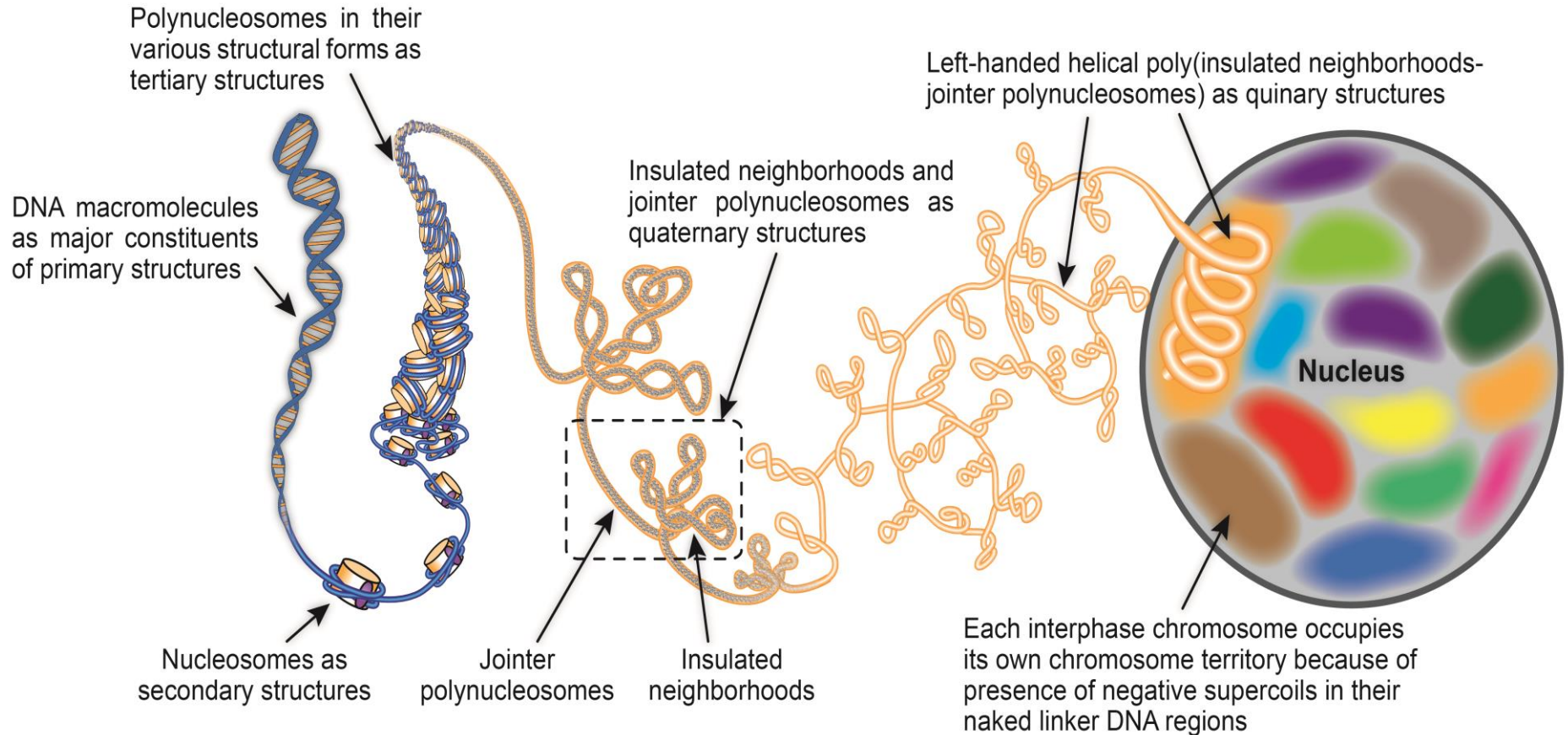
Our SeqMM

TAD boundary positions predicted by Spectral method



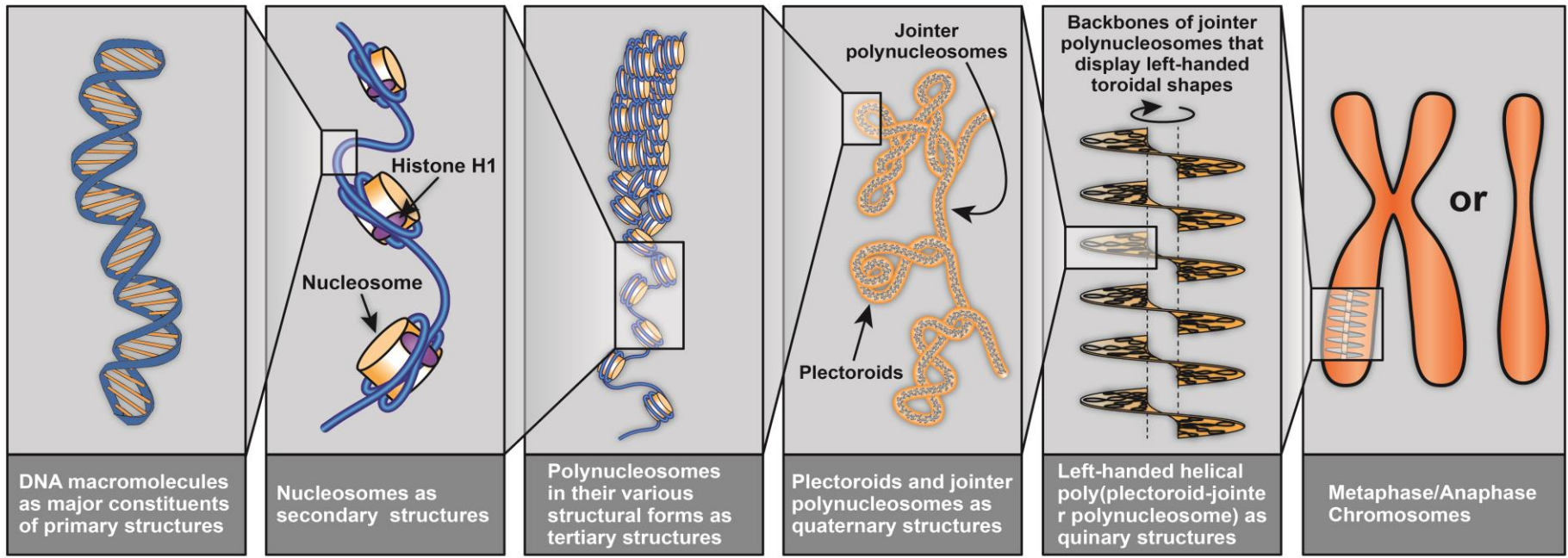
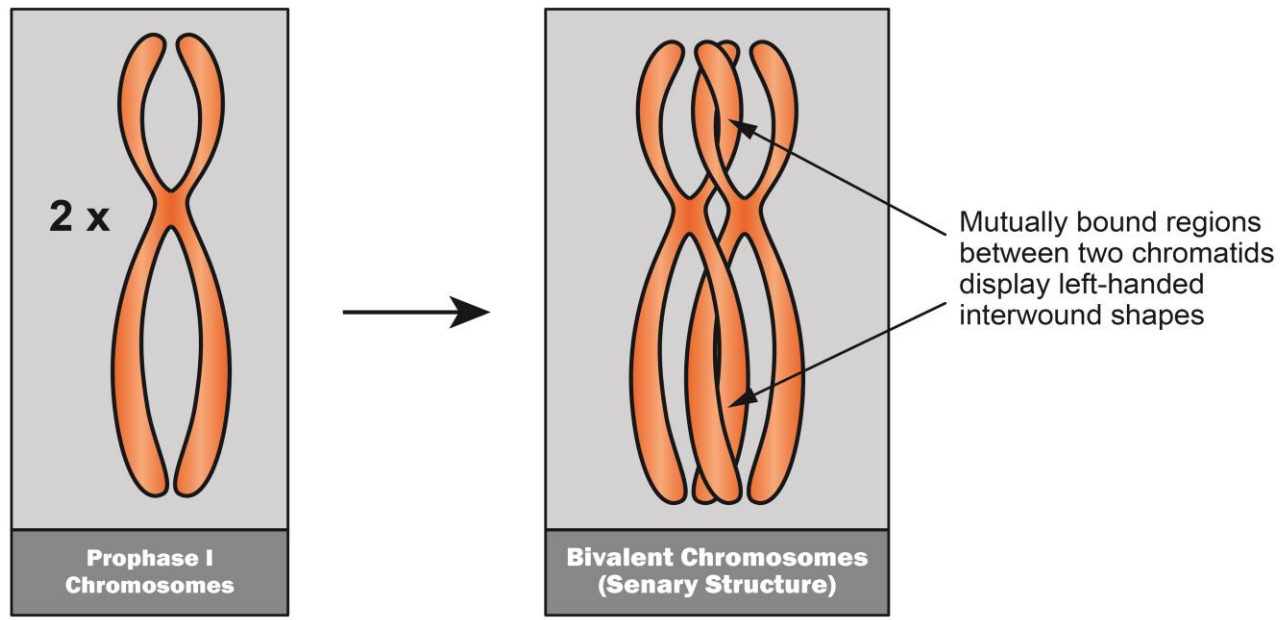
Chen's spectral method

Multiscale Knots



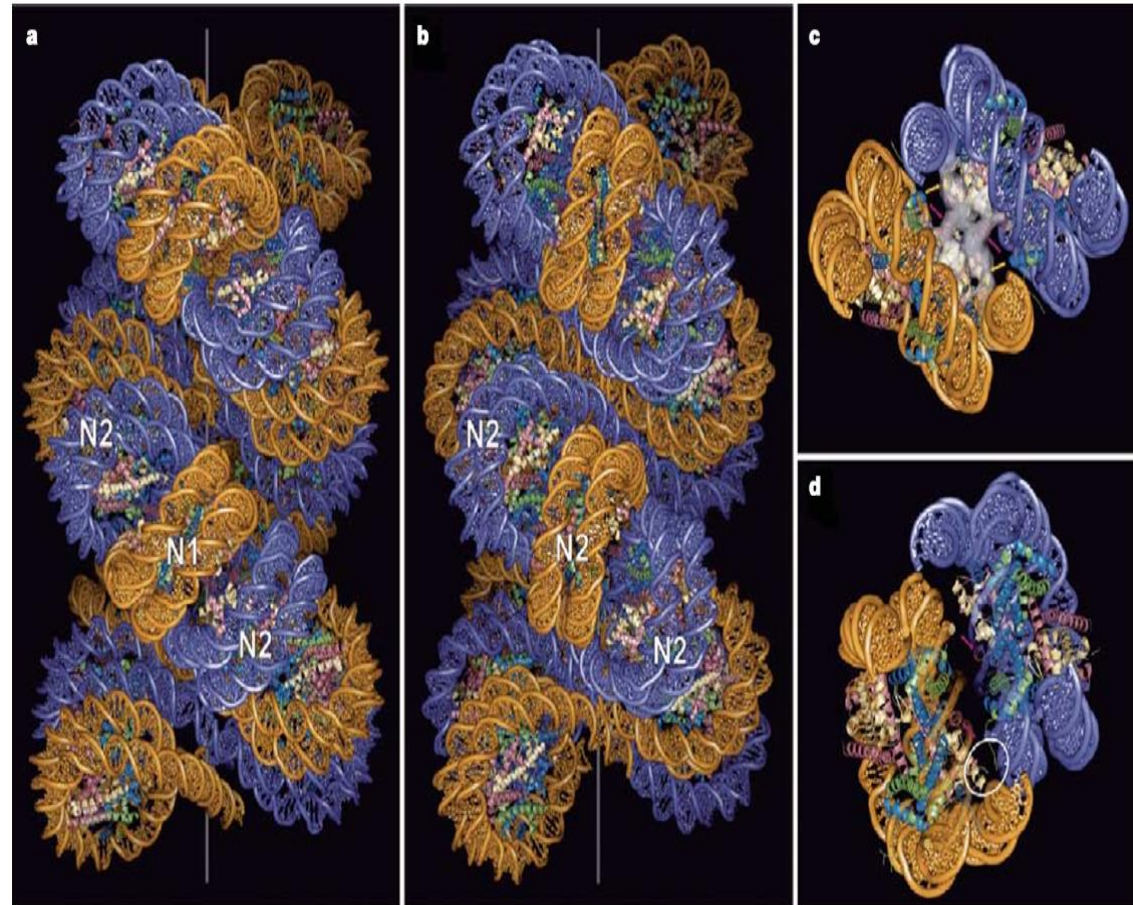
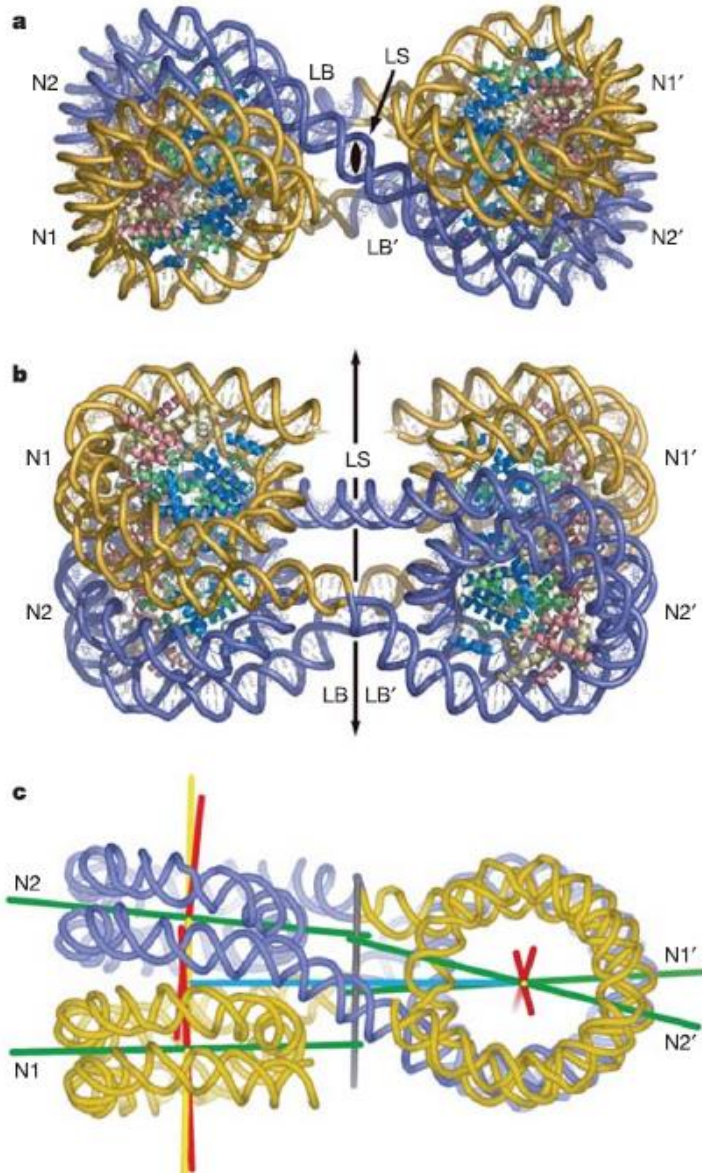
Supercoiling Theory and Model of Chromosomal Structures in Eukaryotic Cells

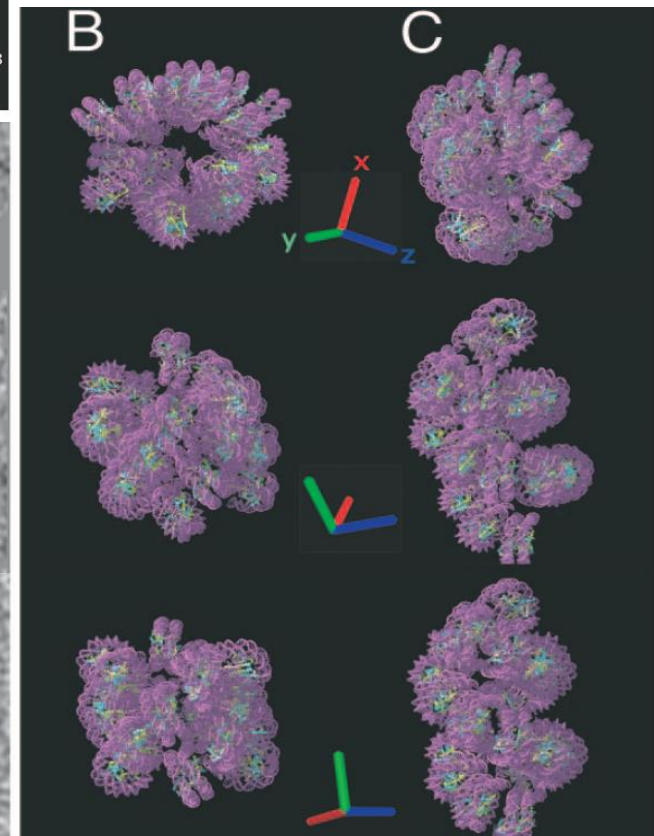
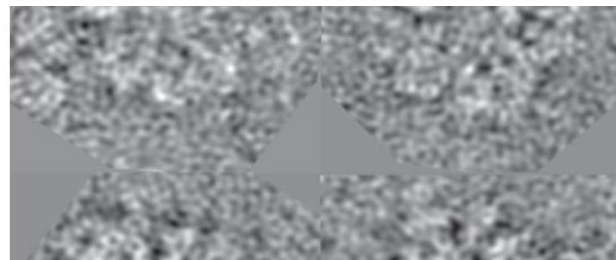
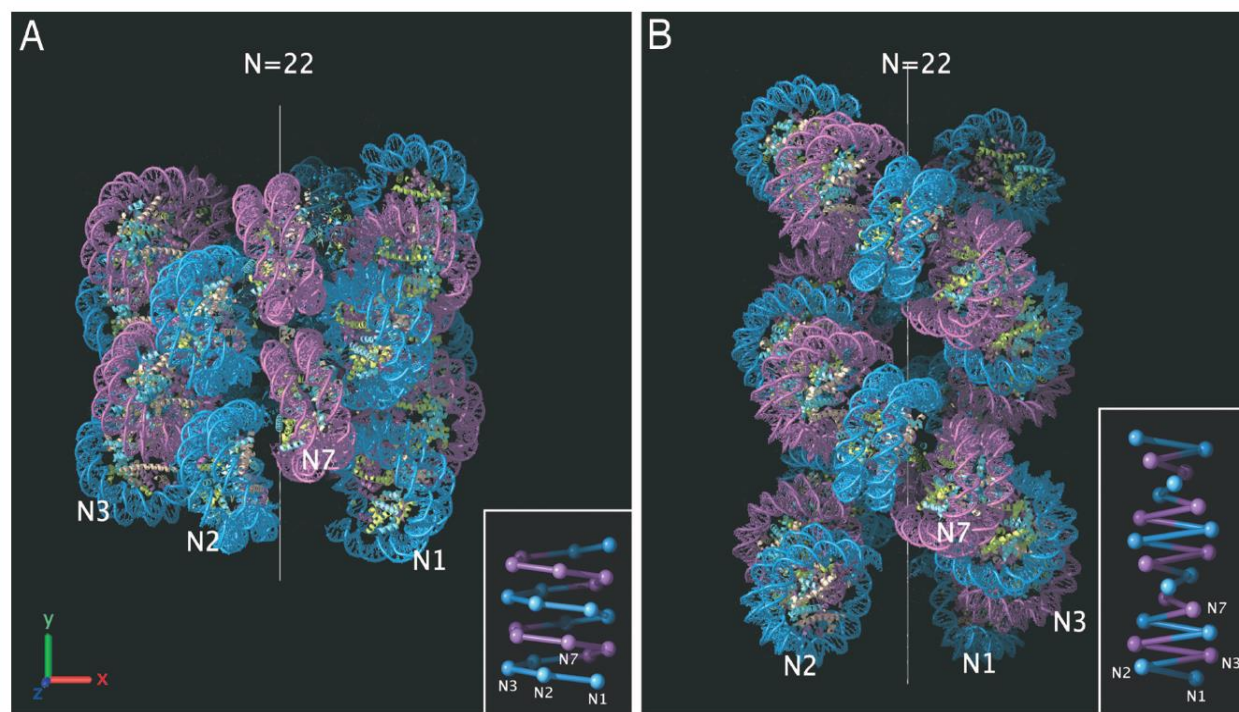
Hao Zhang, Tianhu Li*

A.**B.**

X-ray structure of a tetranucleosome and its implications for the chromatin fibre

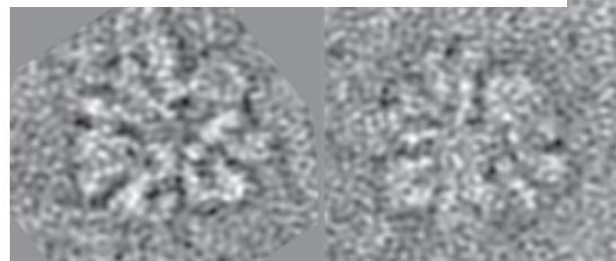
Thomas Schalch¹, Sylwia Duda¹, David F. Sargent¹ & Timothy J. Richmond¹

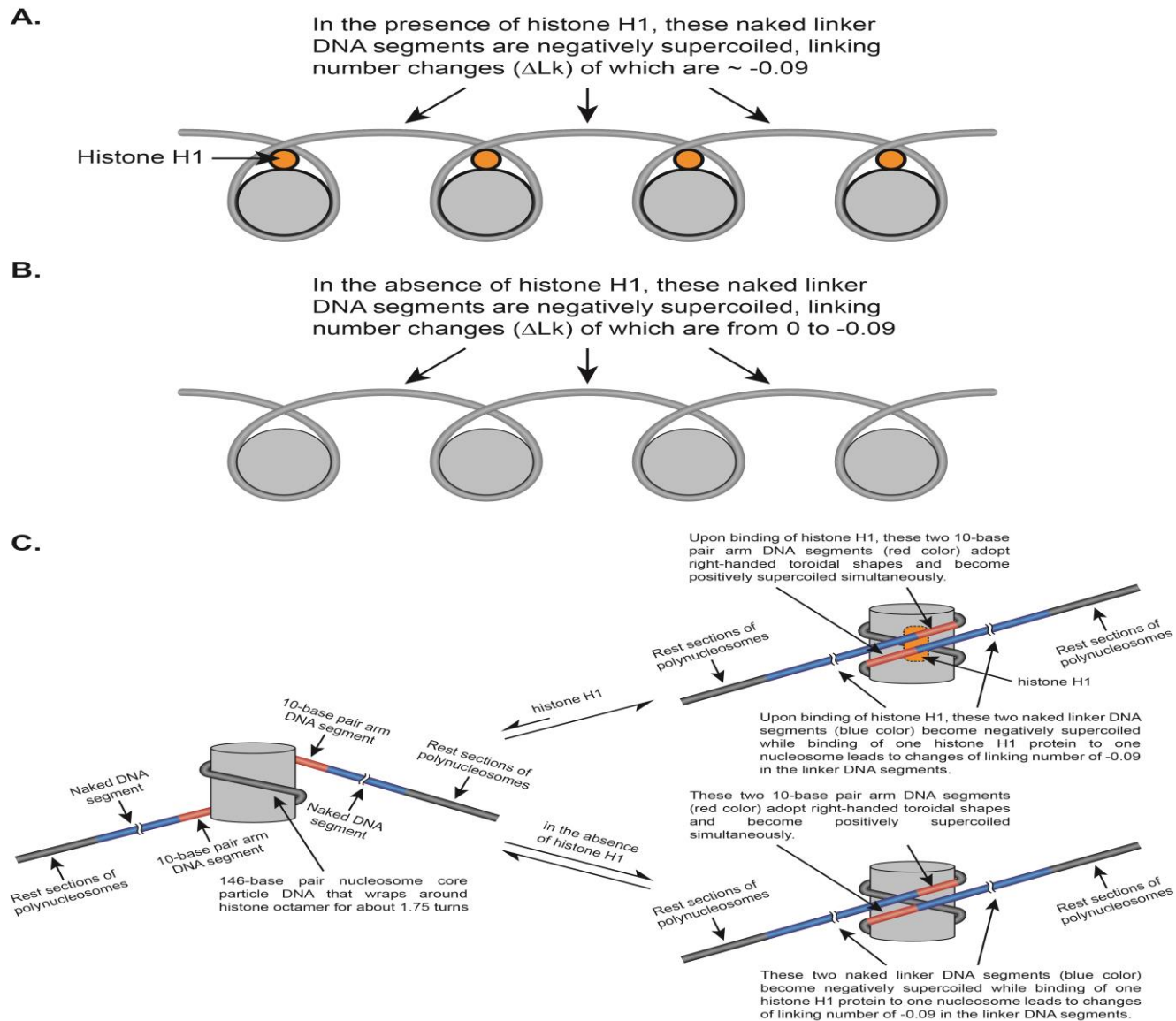




EM measurements define the dimensions of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure

Philip J. J. Robinson, Louise Fairall, Van A. T. Huynh*, and Daniela Rhodes†






Supercoiling Theory and Model of Chromosomal Structures in Eukaryotic Cells

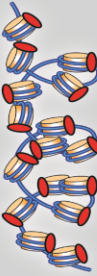
Hao Zhang, Tianhu Li*

A. Type 1: Naked linker DNA segments in this structure display left-handed toroidal shapes on the whole.




Histone H1-bound densely packed polynucleosome

B. Type 2: Naked linker DNA segments in this structure display left-handed toroidal shapes on the whole.



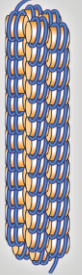
Histone H1-bound loosely packed polynucleosome

C. Type 3: Naked linker DNA segments in this structure display (1) right-handed plectonemic shapes, and/or (2) left-handed toroidal shapes.




Histone H1-bound slacked polynucleosome

D. Type 4: Naked linker DNA segments in this structure display left-handed toroidal shapes on the whole.




Histone H1-free densely packed polynucleosome

E. Type 5: Naked linker DNA segments in this structure display left-handed toroidal shapes on the whole.



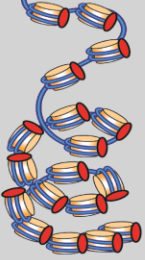
Histone H1-free loosely packed polynucleosome

F. Type 6: Naked linker DNA segments in this structure display (1) right-handed plectonemic shapes, and/or (2) left-handed toroidal shapes.



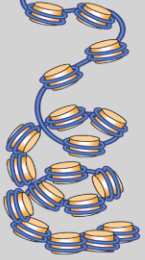
Histone H1-free slacked polynucleosome

G. Type 7: Naked linker DNA segments in this structure display (1) right-handed plectonemic shapes, and/or (2) left-handed toroidal shapes.

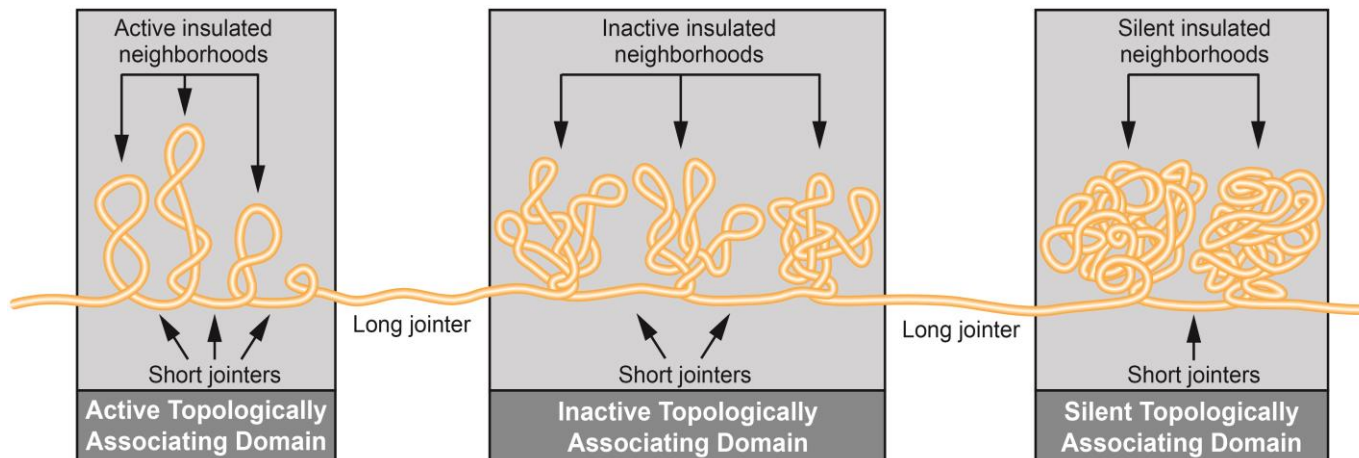
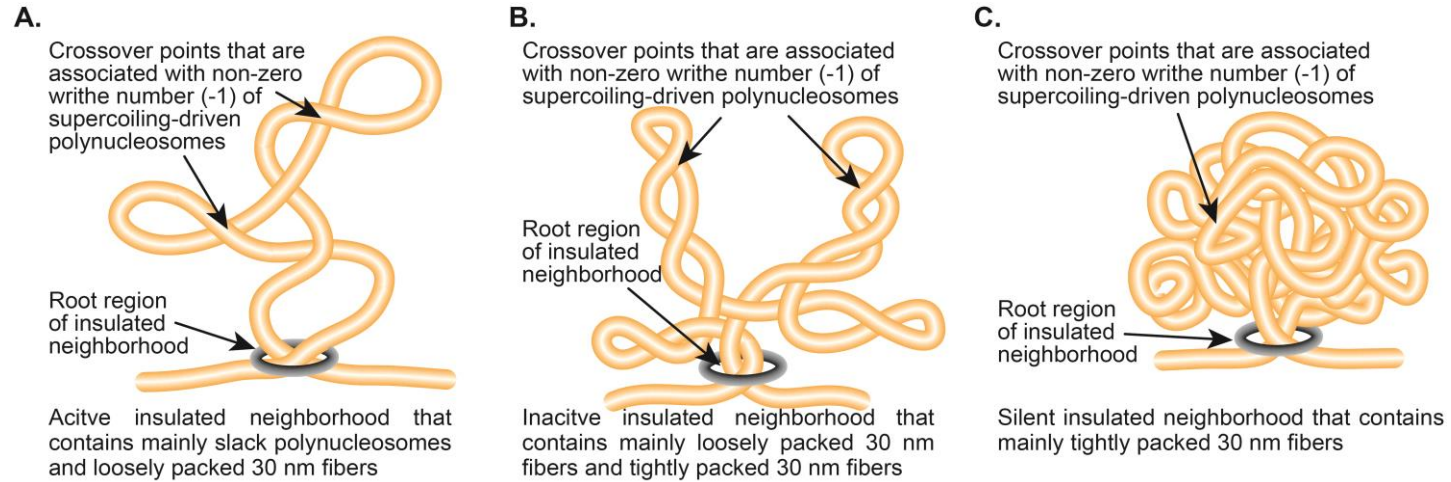
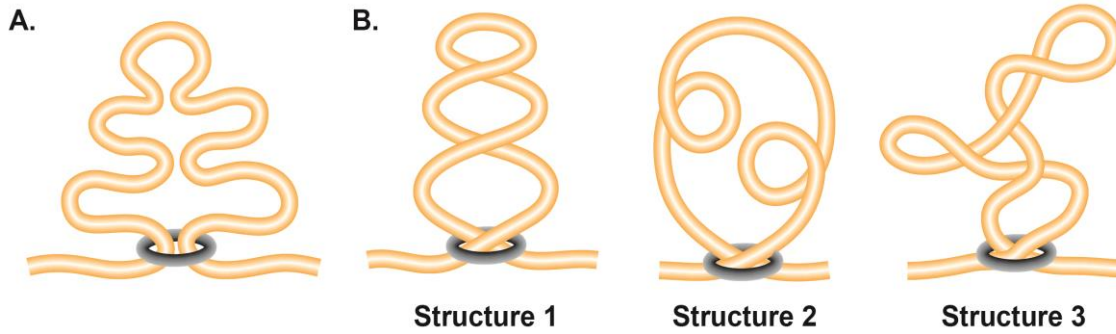


Histone H1-bound polynucleosomes with mixed lengths of linker DNA segments

H. Type 8: Naked linker DNA segments in this structure display (1) right-handed plectonemic shapes, and/or (2) left-handed toroidal shapes.



Histone H1-free polynucleosomes with mixed lengths of linker DNA segments



Part 3: TDA based machine learning for drug design

Drug design and discovery

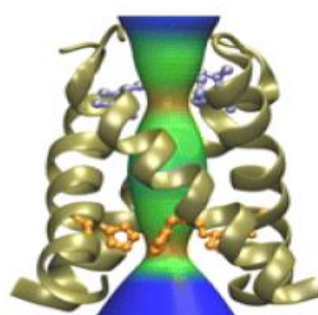


- 1) **Disease identification (physiology)**
- 2) **Target hypothesis (biochem./mole. biol.)**
- 3) **Virtual screening: drug pose, binding affinity, solubility, partition coefficient, toxicity, and side-effects (biophysics/bioinformatics)**
- 4) **Drug structural optimization in the target binding site (biochemistry/biophysics/synthetic chem.)**
- 5) **Preclinical *in vitro* and *in vivo* test**
- 6) **Clinical trials**
- 7) **Optimize drug's efficacy, pharmacokinetics, and pharmacodynamics properties (quantitative systems pharmacology)**

Influenza -- flu virus



M2 channel



Amantadine



M2-A complex



Drug Discovery Process (simplified)

Clinical Trials

Target Discovery

- Target identification
- Microarray profiling
- Target validation
- Assay development
- Biochemistry
- Clinical/Animal disease models

Lead Discovery

- High-throughput Screening (HTS)
- Fragment-based screening
- Focused libraries
- Screening collection

Lead Optimization

- Medicinal Chemistry
- Structure-based drug design
- Selectivity screens
- ADMET screens
- Cellular/Animal disease models
- Pharmacokinetics

•Preclinical Development

- Toxicology
- In vivo safety pharmacology
- Formulation
- Dose prediction

Phase 1

PK tolerability

Phase 2

Efficacy

Phase 3

Safety & Efficacy

Launch

Indication Discovery & expansion

Discovery

Development

Use

Med. Chem. ML,

Clinical Candidates

Drugs

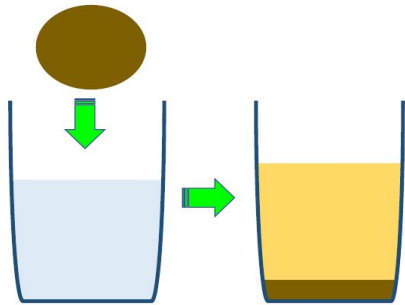
>450,000 distinct compounds
~25,000 distinct lead series

~12,000 candidates

~1,200 drugs

Solubility and partition coefficient

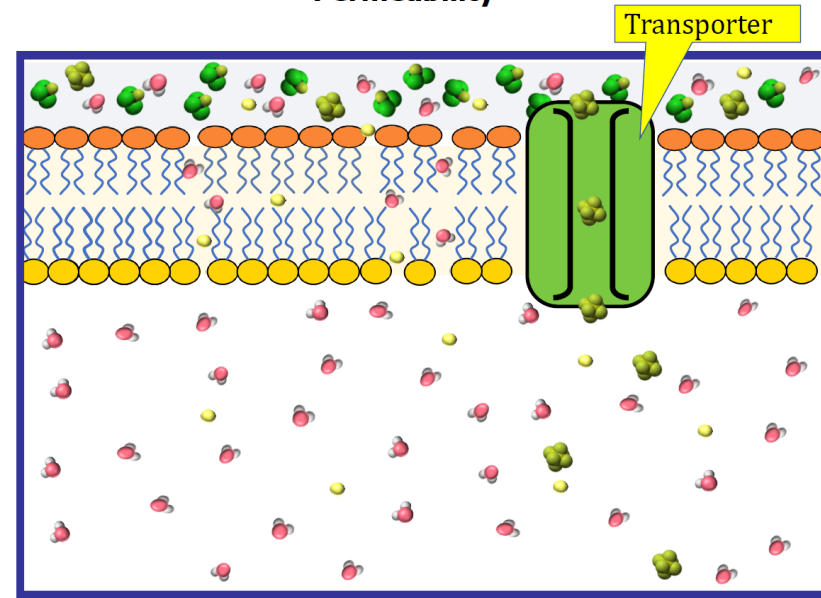
Solubility is commonly expressed as a concentration; for example, as *g* of solute per *kg* of solvent.



Partition coefficient is defined as a particular ratio of the concentrations of a solute between the two solvents



Permeability



Toxicity

Toxicity: The degree to which a substance (a toxin or poison) can harm humans or animals. **Drug toxicity** occurs when a person has accumulated too much of a drug in his bloodstream, leading to adverse effects on the body.

Bioassays:

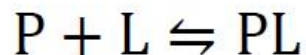
LD50 is defined as the **lethal dose** at which 50% of the population is killed in a given period of time.

LC50 is the **lethal concentration** required to kill 50% of the population. The LC50 is a measure, *e.g.* in mg/l, of the concentration of the toxin whereas a dose is a more general term.



Protein-ligand (-protein) binding

Protein (P) and ligand (L) form a protein-ligand complex (PL):



The association and dissociation constants are

$$K_a = \frac{[PL]}{[P][L]}, \quad K_d = \frac{[P][L]}{[PL]}$$

Binding affinity: $\Delta G = RT \ln K_d$

Database:

ChEMBL (as 02/28/2019):

Targets: 12,091

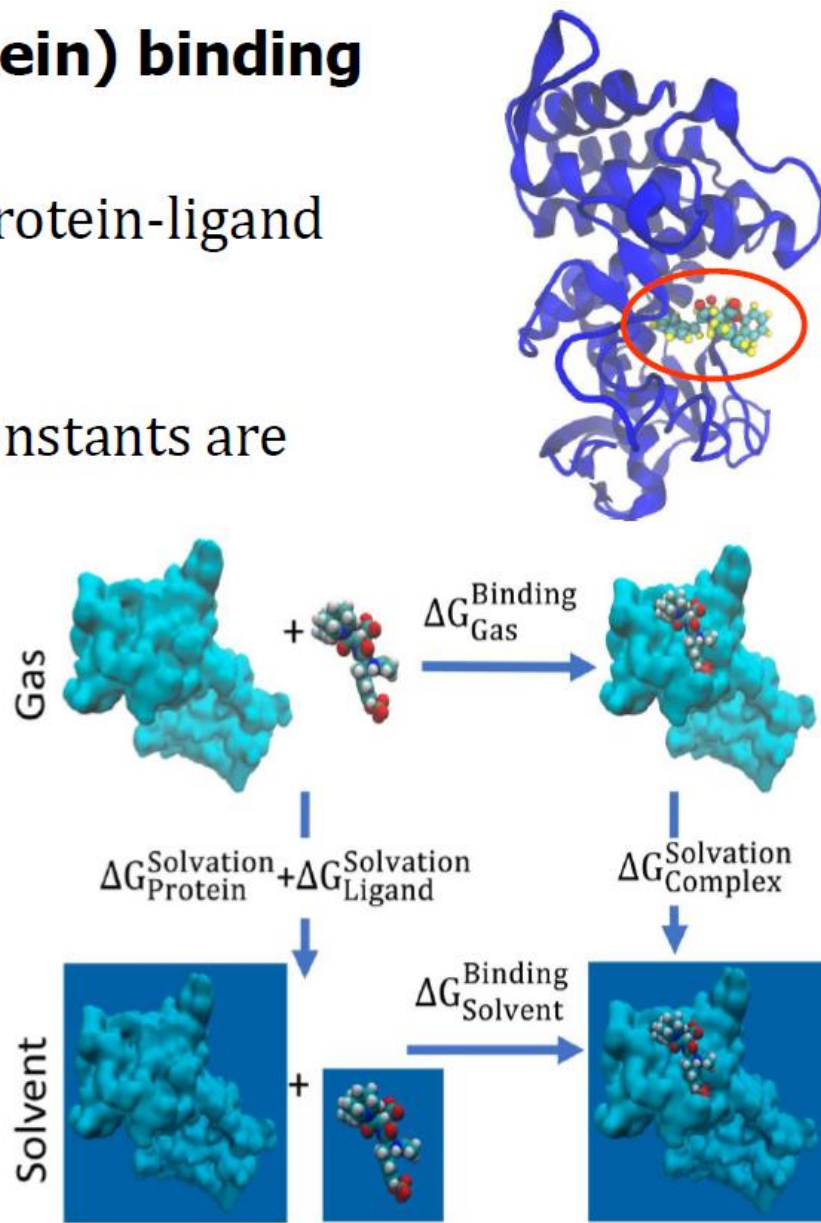
Compound records: 2,275,906

Distinct compounds: 1,828,820

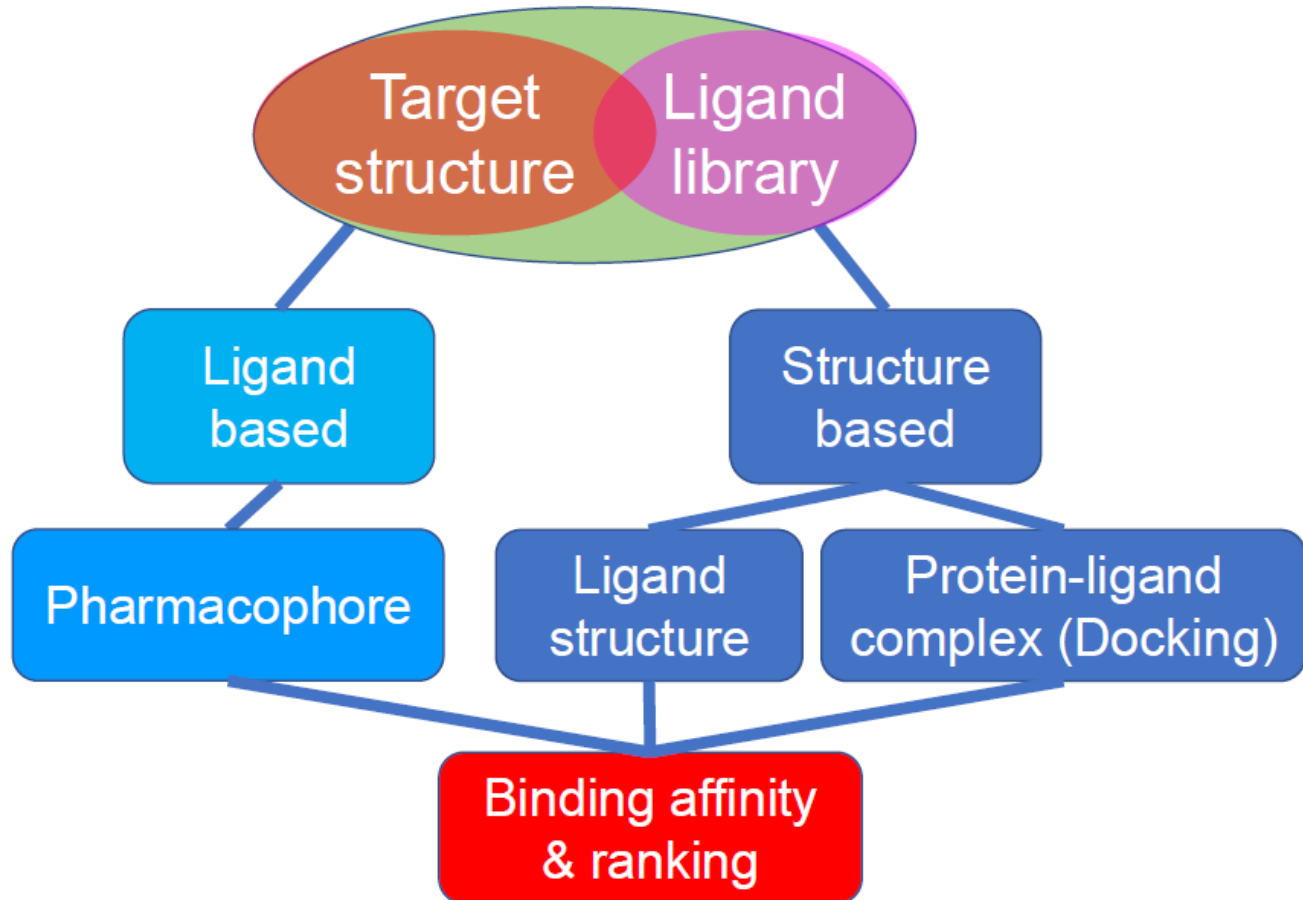
Activities: 15,207,914

Publications: 69,861

Binding DB, PubChem, PDBbind, K_i Database.



Virtual Screening



Force field

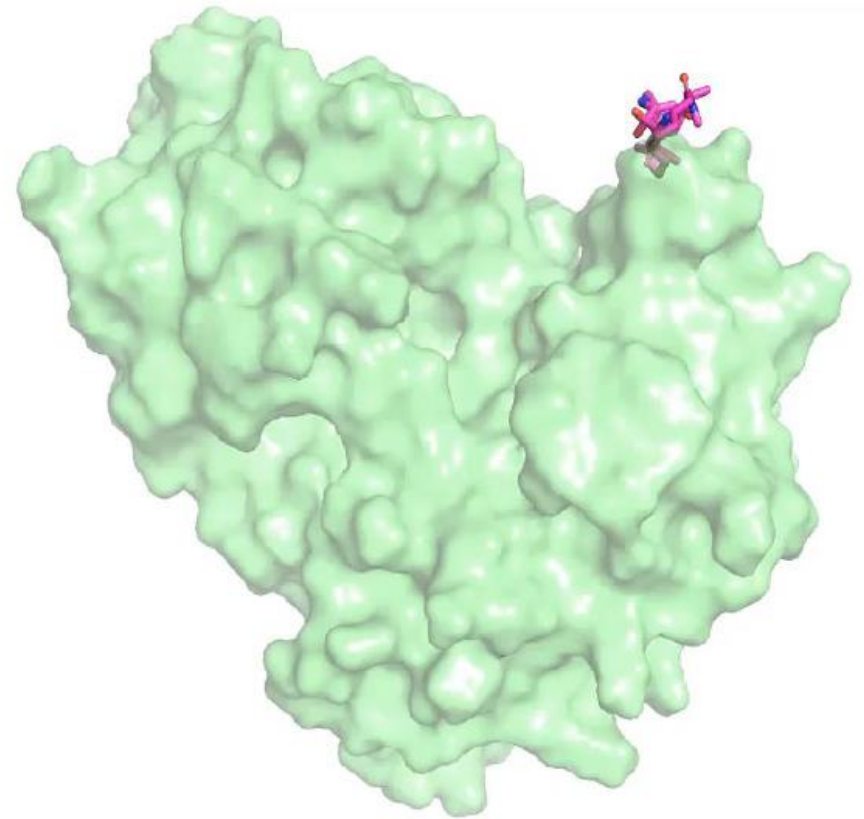
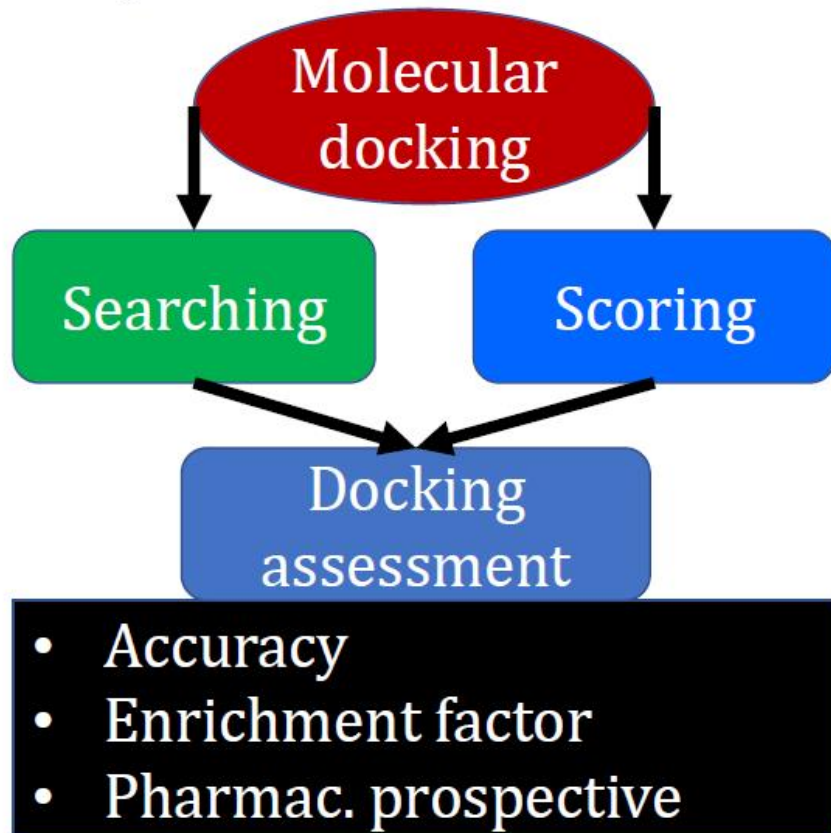
Empirical

Knowledge-based

Machine learning

Molecular docking

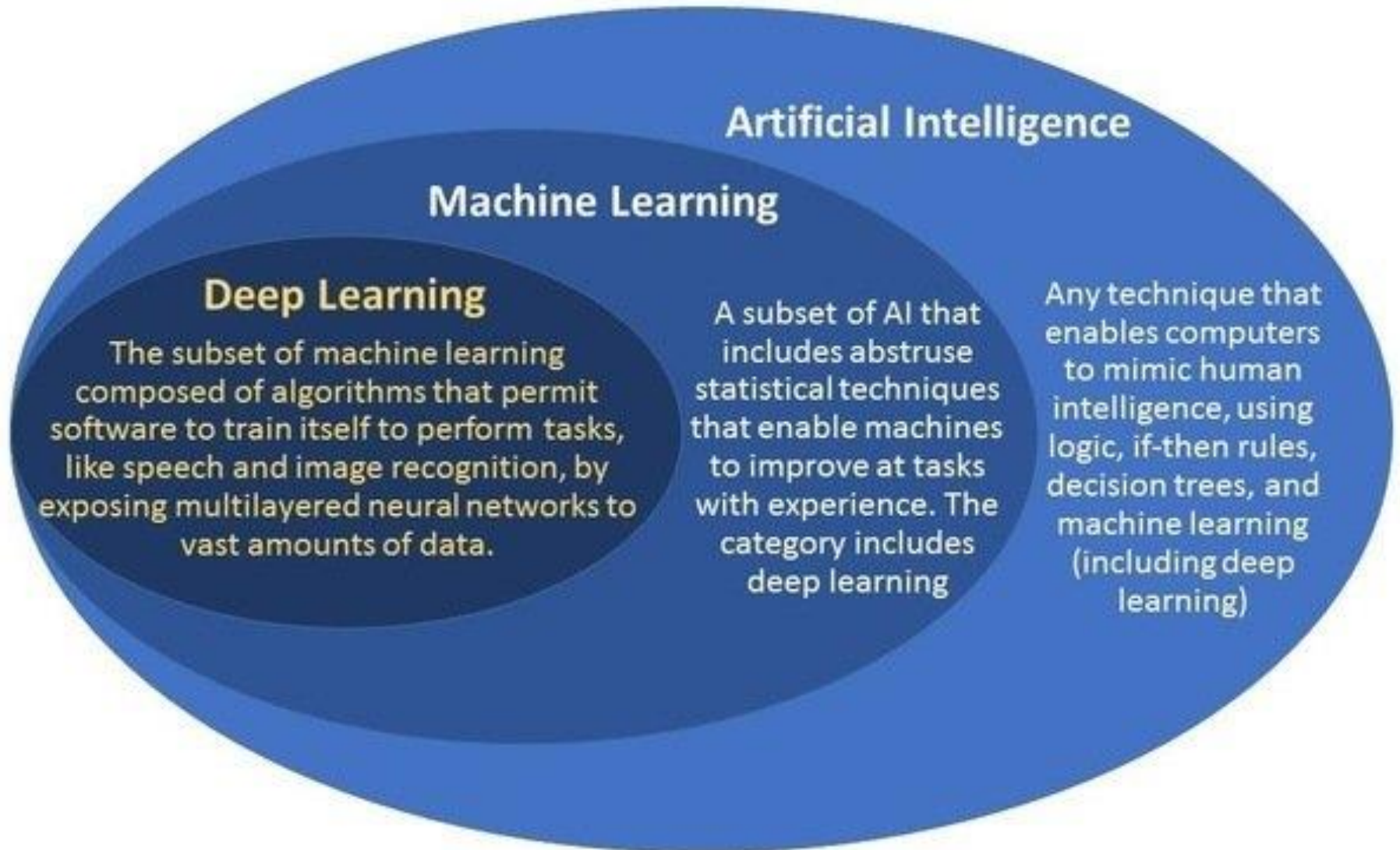
Docking is a process for searching the preferred position and orientation of one molecule to another one to form a stable complex.



Docking a ligand to BACE

(Kaifu Gao, Duc Nguyen, and Wei, 2019)

Machine learning based data analysis



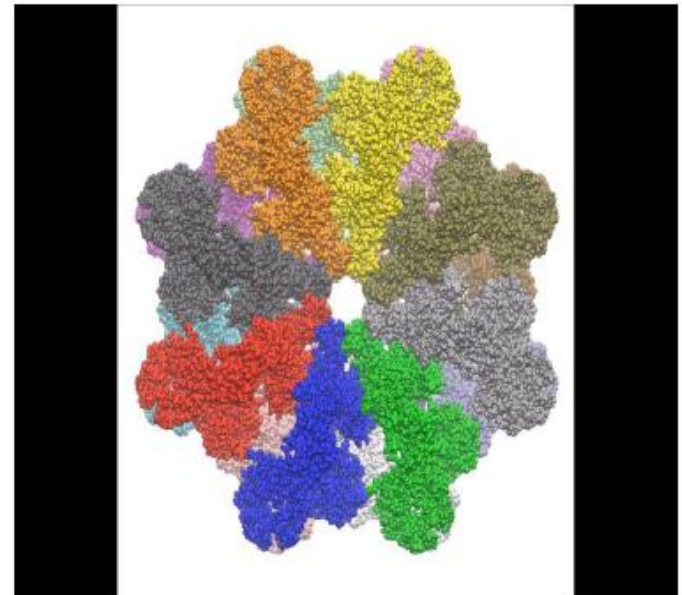
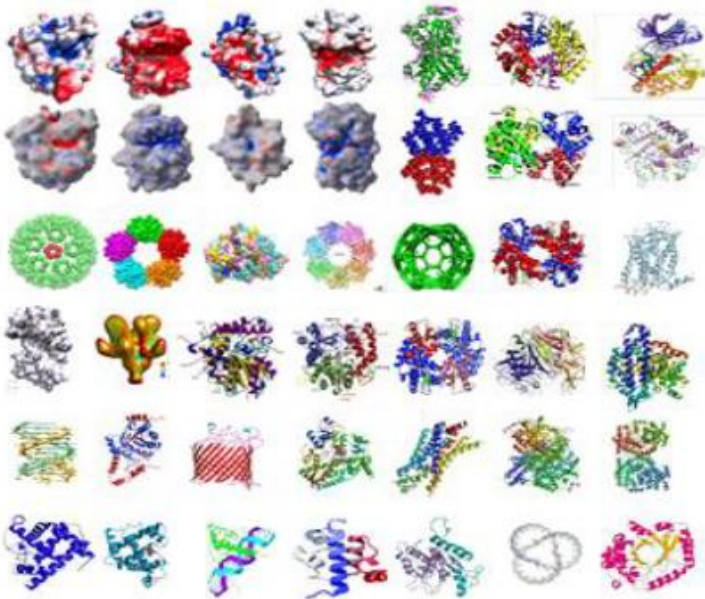
How to do deep learning for 3D biomolecular data?

Obstacles for deep learning of 3D biomolecules:

- **Geometric dimensionality:** \mathbb{R}^{3N} , where $N \sim 5000$ for a protein.
- **Machine learning dimensionality:** $> 1024^3 m$, where m is the number of atom types in a protein.
- **Molecules have different sizes --- non-scalable.**
- **Complexity:** intermolecular & intramolecular interactions

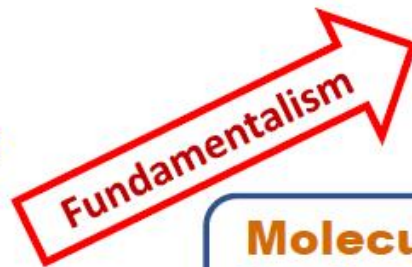
Solution:

- **Geometric simplification, dimension reduction & scale unification**



Two schools of thinking

Given a protein with N atom and an average of n electrons in each atom



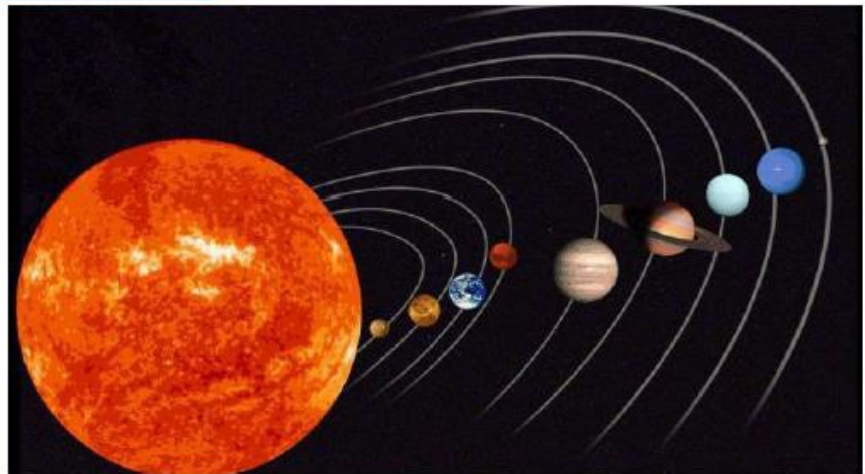
Quantum Mechanics
 \mathbb{R}^{3Nn+3N}

QM/MM \mathbb{R}^K
 $3N < K < 3N(n+1)$

Molecular Mechanics
 \mathbb{R}^{3N}

Multiscale Coarse-grain
 \mathbb{R}^M ($3 < M < 3N$)

Poisson-Boltzmann, PNP, etc. \mathbb{R}^3



Differentiable Manifold
 \mathbb{R}^2

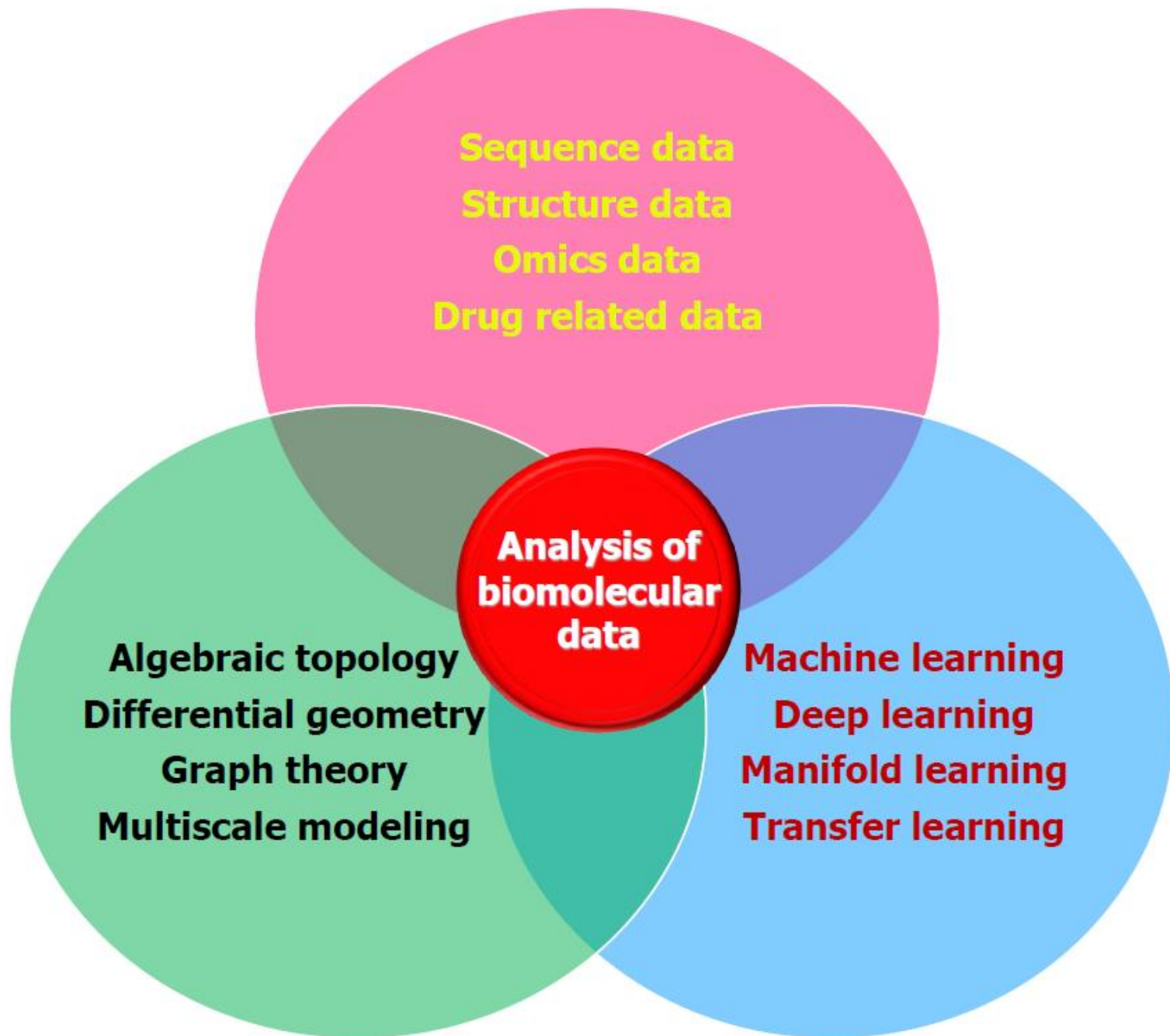
Algebraic Topology
 \mathbb{R}^1

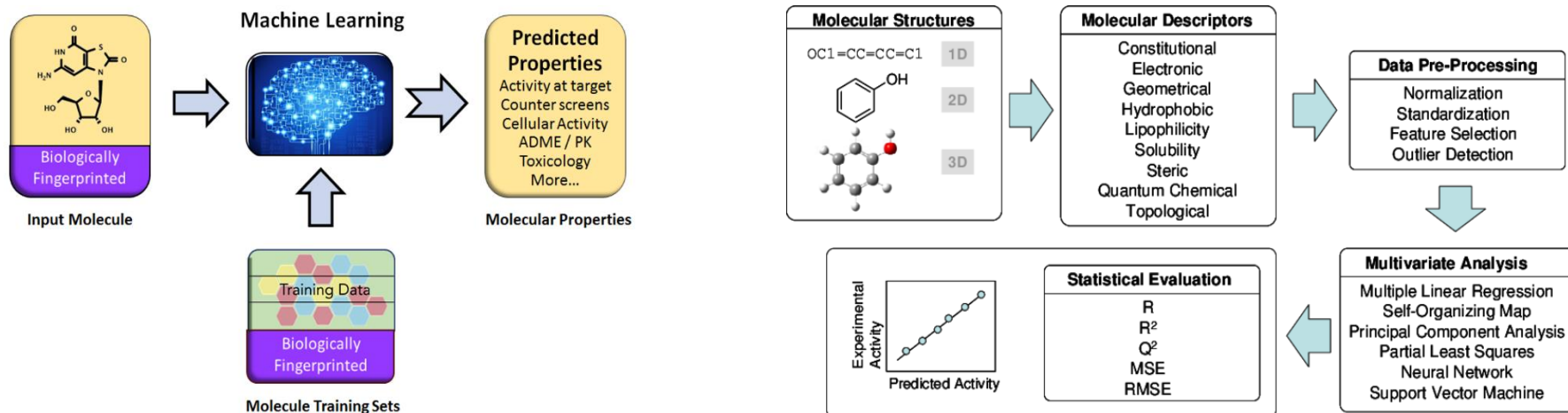
Basic hypothesis:
Intrinsic physics lies on low-dimensional manifolds in a high dimensional space



Graph Theory
 \mathbb{R}^0

Geo-Top Indices
 \mathbb{R}^0



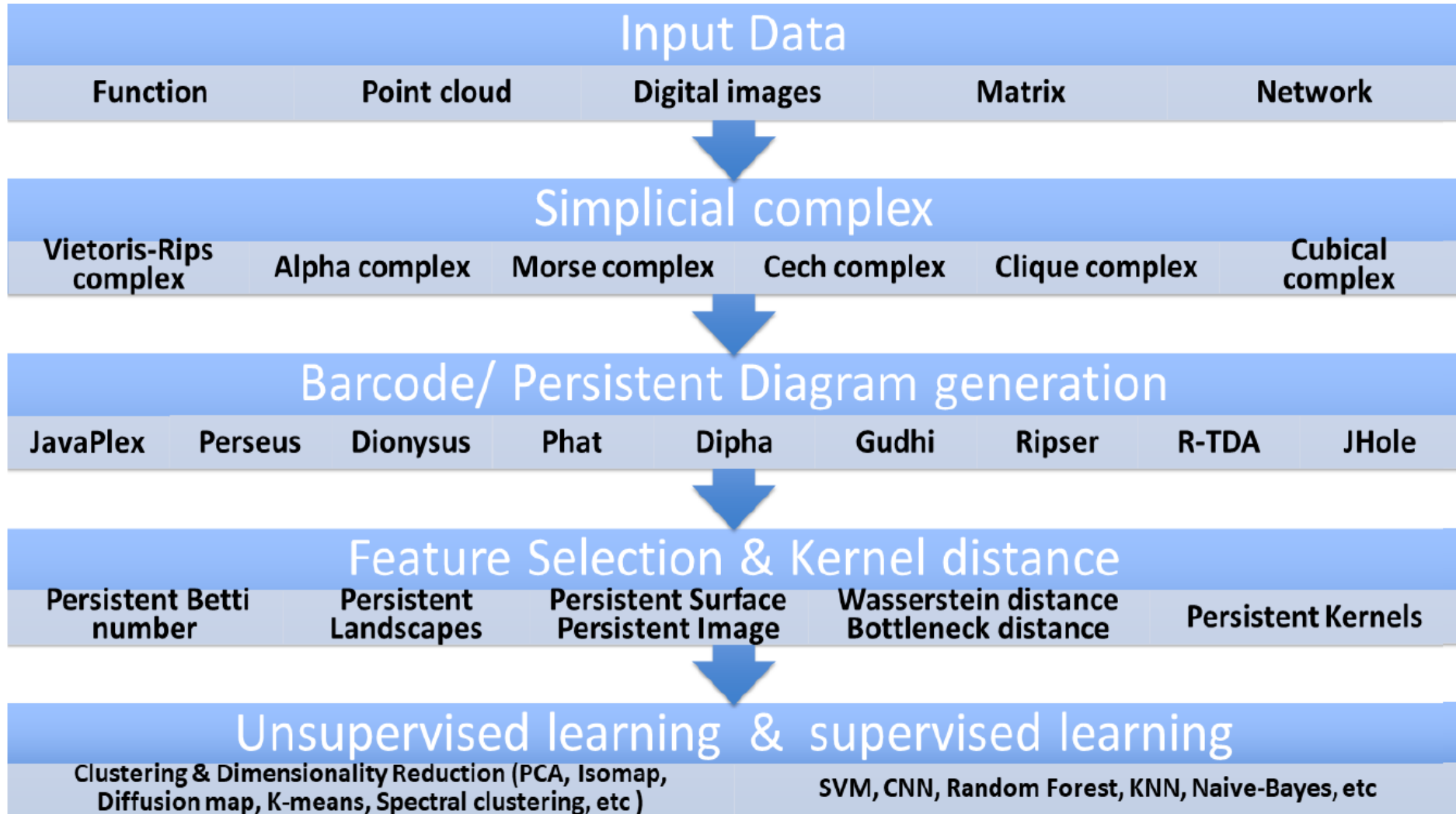


Molecular descriptor directly determines the performance of the learning models!

Common chemical descriptors for QSAR/QSPR analysis

Chemical descriptors	Based on	Examples
Theoretical descriptors		
0D	Molecular formula	Molecular weights, atom counts, bond counts
1D	Chemical graph	Fragment counts, functional group counts
2D	Structural topology	Weiner index, Balaban index, Randic index, BCUTS
3D	Structural geometry	WHIM, autocorrelation, 3D-MORSE, GETAWAY
4D	Chemical conformation	Volsurf, GRID, Raptor
Experimental descriptors		
Hydrophobic parameters	Hydrophobicity	Partition coefficients (logP), hydrophobic substituent constant (π)
Electronic parameters	Electronic properties	Acid dissociation constant, Hammett constant
Steric parameters	Steric properties	Taft steric constant, Charton's constant

TDA based machine learning models



(Pun, Xia and Lee, submitted, 2018)

Guowei Wei group's works

SIAM NEWS DECEMBER 2017



Research | December 01, 2017

Persistent Homology Analysis of Biomolecular Data

By Guo-Wei Wei

SIAM NEWS SEPTEMBER 2016



Get Involved | September 01, 2016

Mathematical Molecular Bioscience and Biophysics

A Recurring Theme at the SIAM Conference on the Life Sciences

By Guo-Wei Wei

Professor
Mathematics,
Electrical & Computer Engineering,
Biochemistry & Molecular Biology,
Michigan State University , USA

Software packages:

- [MIBPB](#): Online server for electrostatic analysis using the second-order accurate Poisson-Boltzmann solver.
- [ESES](#): Open-source online server for the generation of Eulerian solvent excluded surface.
- [PPD](#): Online server for Protein Pocket Detection.
- [FRI](#): Online server for the flexibility analysis of biomolecules based on flexibility and rigidity index.
- [RI-Score](#): Online server for geometric graph theory or rigidity index (RI) based scoring function for protein ligand binding affinity prediction.
- [TML-BP](#): Online server for topological learning for protein-ligand binding affinity prediction.
- [TML-MP](#): Online server for topology based machine learning for the prediction of protein folding stability change upon mutation.
- [TDL-BP](#): Online server for topological deep learning for protein-ligand binding affinity prediction.
- [TDL-MP](#): Online server for topological deep learning for the prediction of protein folding stability change upon mutation.
- [TopP-S](#): Online server for topological learning of partition coefficient (LogP) and aqueous solubility (LogS).
- [TopTox](#): Online server for computing element-specific topological descriptors (ESTDs) for toxicity endpoint predictions.

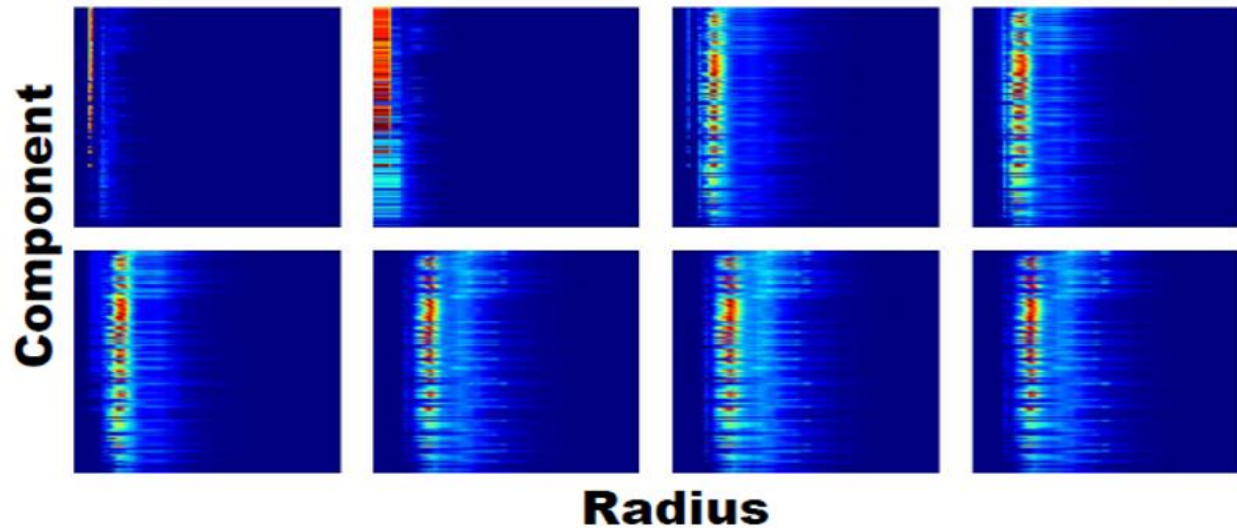
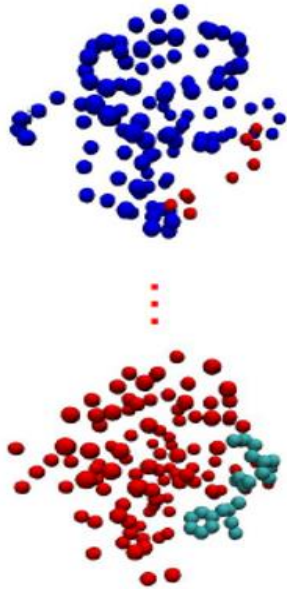
Recent progress in topology based drug design **(By Guowei Wei's group)**

Element specific persistent homology (ESPH) method

Proteins: (C, N, O, S)

Ligands: (C, N, O, S, P, F, Cl, Br, I)

*Cross protein-ligand ESTFs: one type from protein and the other from the ligand. **Totally 36 sets of ESTFs** in each topological dimension*

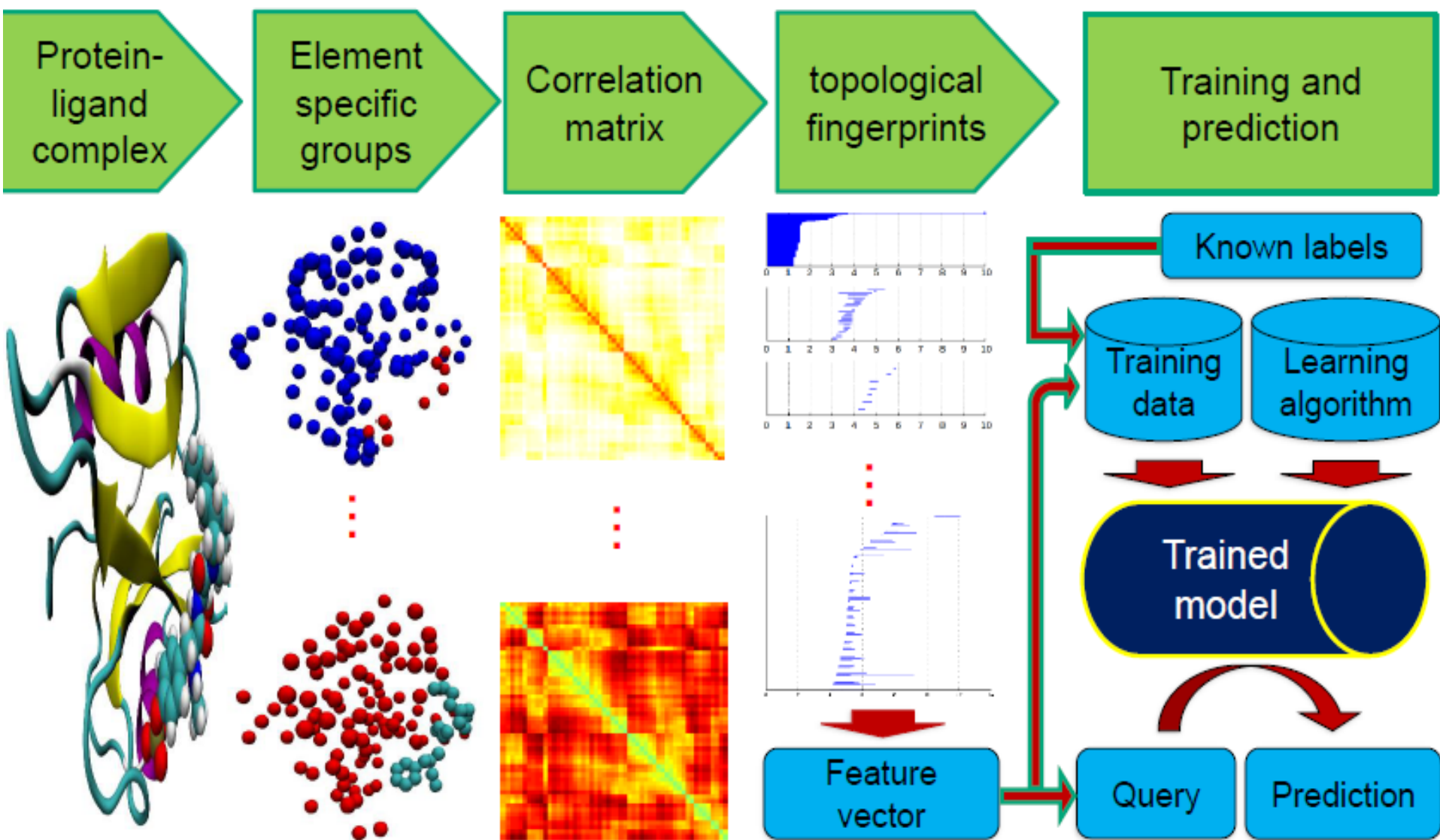


Components are generated from element specific persistent homology. Eight channels are constructed from births, deaths and persistences at Betti-0, Betti-1 and Betti-2.

(Cang & Wei, IJNMBE, 2017)

Topology based learning architecture

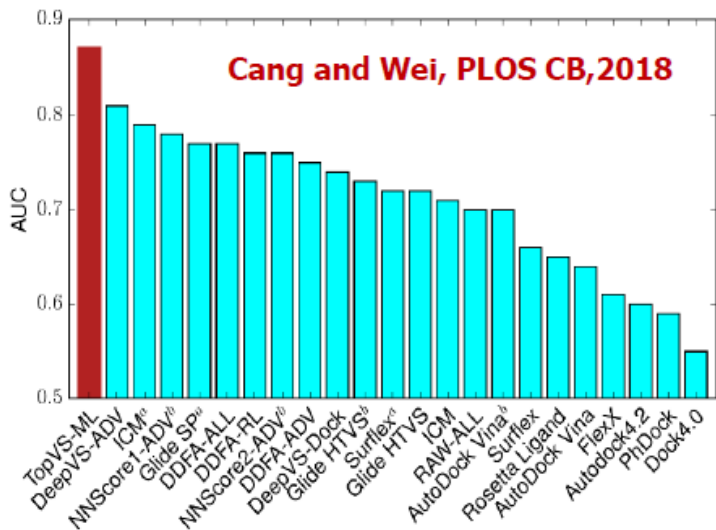
(Cang & Wei, IJNMBE, 2017)



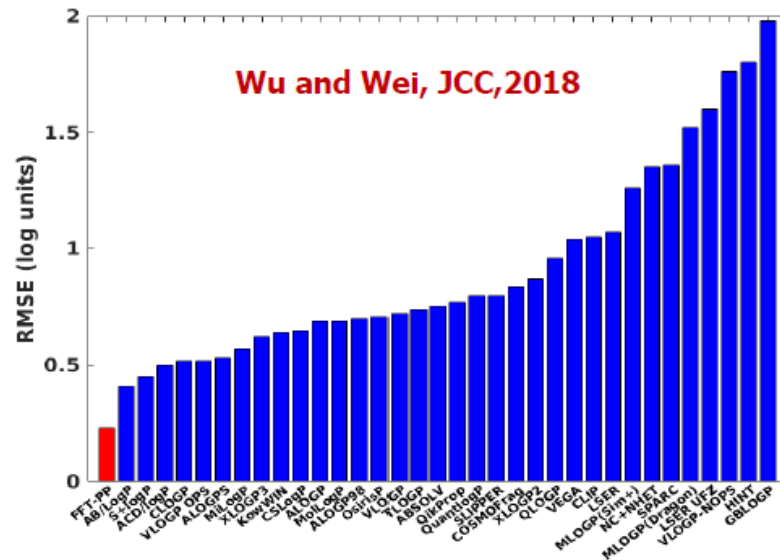
Topological learning based predictions

Classification of ligands & decoys

DUD database 128,374 protein-ligand/decoy pairs

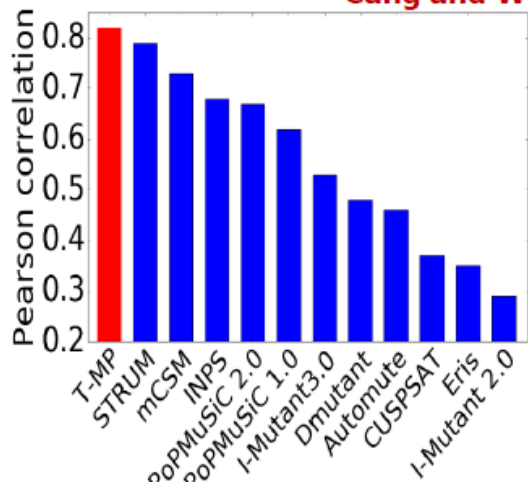


Prediction RMSD of LogP (Star set)

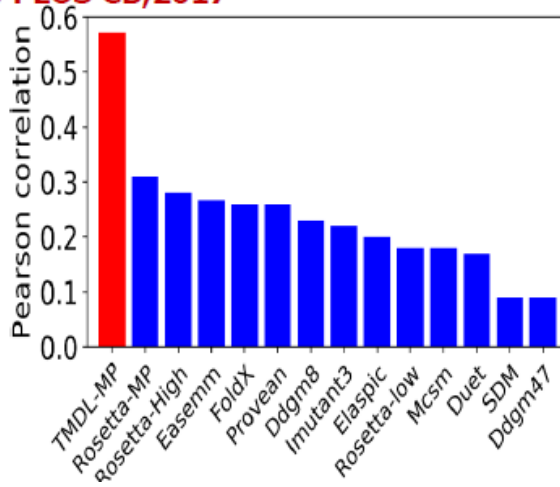


Prediction correlations for 2648 mutations on globular proteins

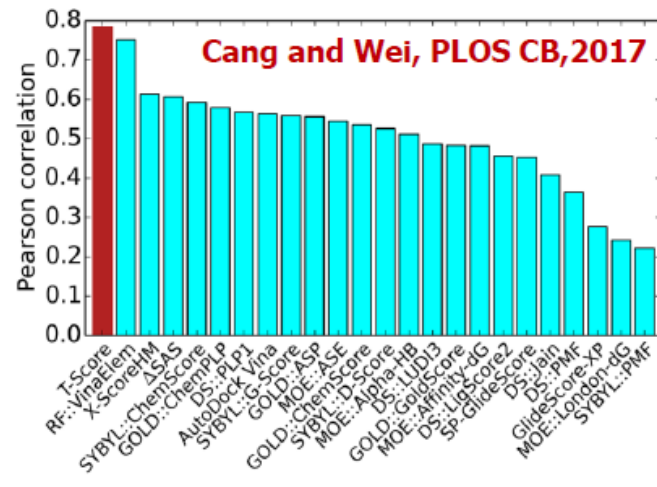
Cang and Wei, PLOS CB, 2017



Prediction correlations for 223 mutations on membrane proteins



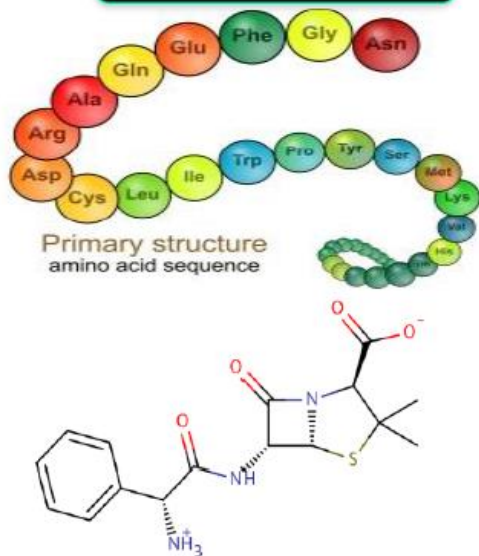
Binding affinity prediction of PDBBind v2013 core set of 195 complexes



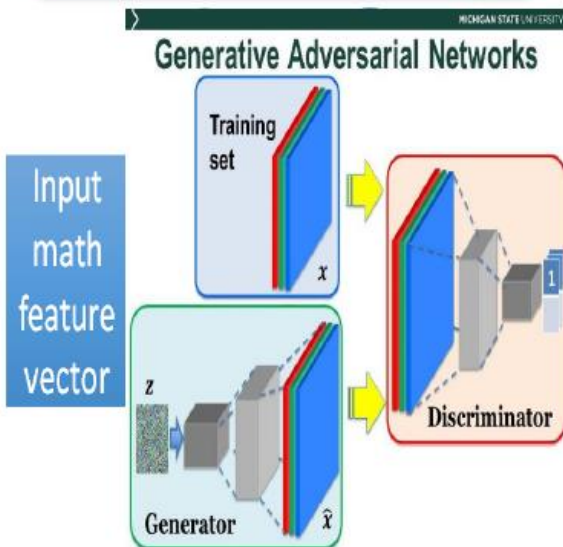
Drug Design Data Resource (D3R) Grand Challenge



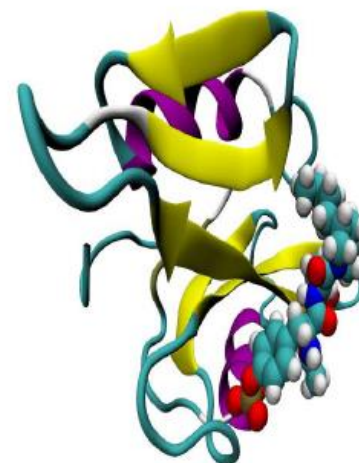
Given data



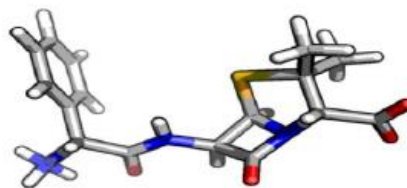
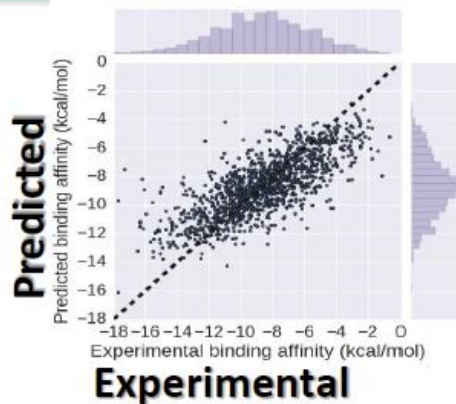
Math based GAN



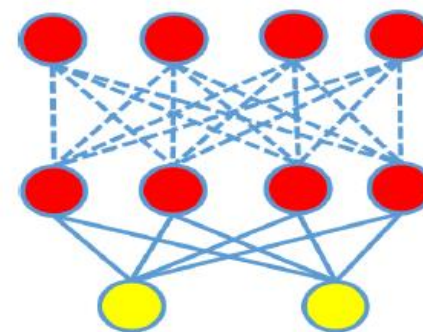
Predicted complex



Final predictions to be compared with experiments



Drug pose



(Nguyen et al, JCAMD, 2018)

D3R Grand Challenge 2 (2016-2017)



Given: Farnesoid X receptor (FXR) and 102 ligands

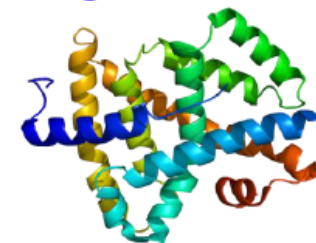
Tasks: Dock 102 ligands to FXR, and predict their poses, binding free energies and energy ranking

Stage 1

- [Pose Predictions \(partials\)](#)
- [Scoring \(partials\)](#)
- [Free Energy Set 1 \(partials\)](#)
- [Free Energy Set 2 \(partials\)](#)

Stage 2

- [Scoring \(partials\)](#)
- [Free Energy Set 1 \(partials\)](#)
- [Free Energy Set 2 \(partials\)](#)

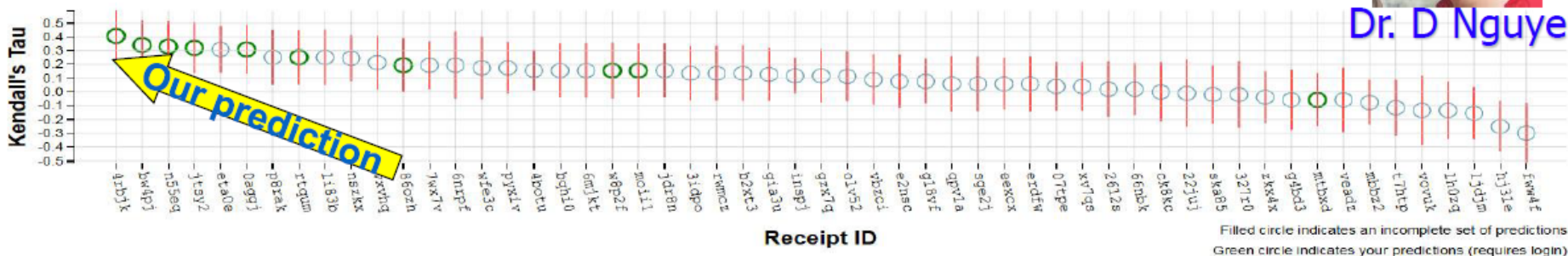


Dr. D Nguyen

Grand Challenge 2

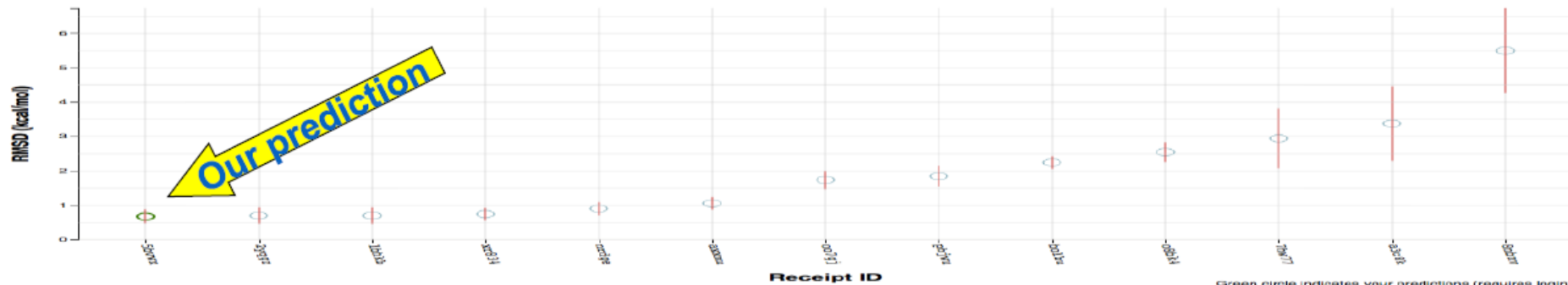
(Nguyen et al, JCAMD, 2018)

Free Energy Set 1 (Stage 2) - Kendall's Tau



Grand Challenge 2

Free Energy Set 1 (Stage 1) - RMSD



D3R Grand Challenge 3 (2017-2018)

(Nguyen et al, JCAMD, 2018)



Pose Prediction

Cathepsin Stage 1A

[Pose Predictions \(partials\)](#)

Affinity Rankings excluding Kds > 10 μM

Cathepsin Stage 1

[Scoring \(partials\)](#)

[Free Energy Set](#)

VEGFR2

[Scoring \(partials\)](#)

JAK2 SC3

[Scoring](#)

[Free Energy Set](#)



Active / Inactive Classification

VEGFR2

[Scoring \(partials\)](#)

JAK2 SC3

[Scoring](#)

[Free Energy Set](#)



Affinity Rankings for Cocrystalized Ligands

Cathepsin Stage 1

[Scoring \(partials\)](#)

[Free Energy Set](#)



Cathepsin Stage 1B

[Pose Prediction](#)

Cathepsin Stage 2

[Scoring \(partials\)](#)

[Free Energy Set](#)

JAK2 SC2

[Scoring \(partials\)](#)

TIE2

[Scoring](#)

[Free Energy Set 2](#)



JAK2 SC2

[Scoring \(partials\)](#)

TIE2

[Scoring \(partials\)](#)

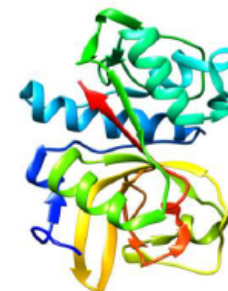
[Free Energy Set 1](#)



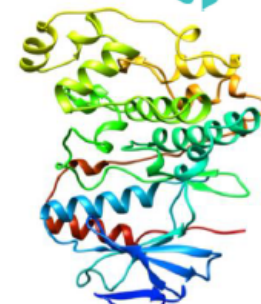
Cathepsin Stage 2

[Scoring \(partials\)](#)

[Free Energy Set](#)



Cathepsin S



Kinase: p38-α

p38-α

[Scoring](#)

ABL1

[Scoring \(partials\)](#)



p38-α

[Scoring \(partials\)](#)

ABL1

[Scoring \(partials\)](#)



Zixuan Cang



Dr. D Nguyen

D3R Grand Challenge 4 (2018-2019)



Pose Predictions

BACE Stage 1A

Pose Predictions (Partials)



BACE Stage 1B

Pose Prediction (Partials)



Affinity Predictions

Cathepsin Stage 1

Combined Ligand and Structure Based Scoring



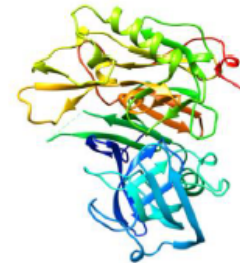
Dr. Kaifu Gao Dr. D Nguyen

Ligand Based Scoring (No participation)

Structure Based Scoring



Free Energy Set



BACE Stage 1

Combined Ligand and Structure (No participation)

Ligand Based Scoring (Partials) (No participation)

Structure Based Scoring (Partials) (No participation)

Free Energy Set (No participation)

BACE Stage 2

Combined Ligand and Structure

Ligand Based Scoring (No participation)

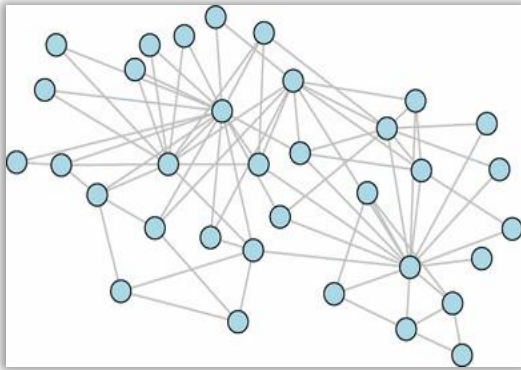
Structure Based Scoring (Partials)

Free Energy Set

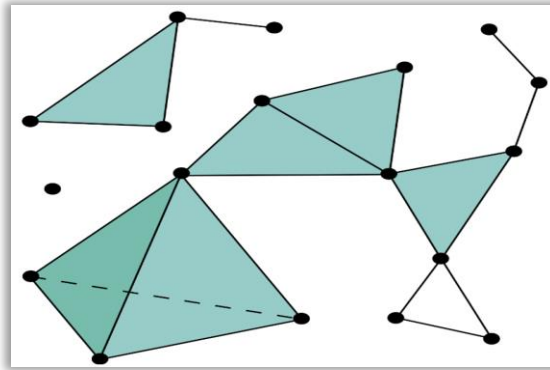


TDA is based on the multiscale simplicial complex

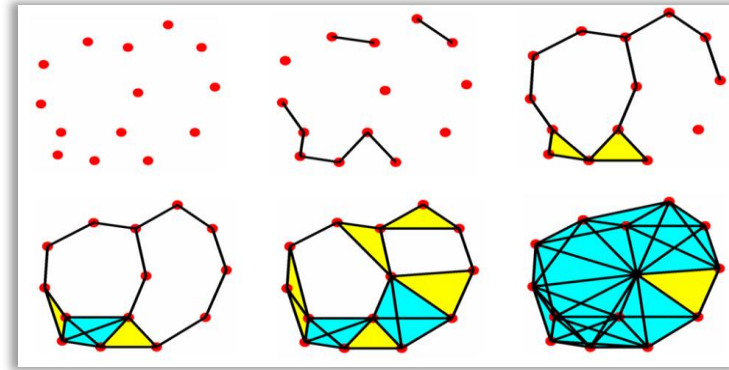
□ Graph



□ Simplicial complex



□ Multiscale simplicial complexes



❖ Graph models and measurements:

Graph Laplacian; Fiedler Eigenvalue; Fiedler eigenvector; Shortest path; Clique; Cluster coefficient; Closeness; Centrality; Betweenness; Modularity; Cheeger constant; Erdos number; Percolation...

❖ Simplicial complex models and measurements:

Combinatorial Laplacian; Hodge theory; Betti number; Euler characteristics; Homology; Cohomology; Morse theory; Knot polynomials...

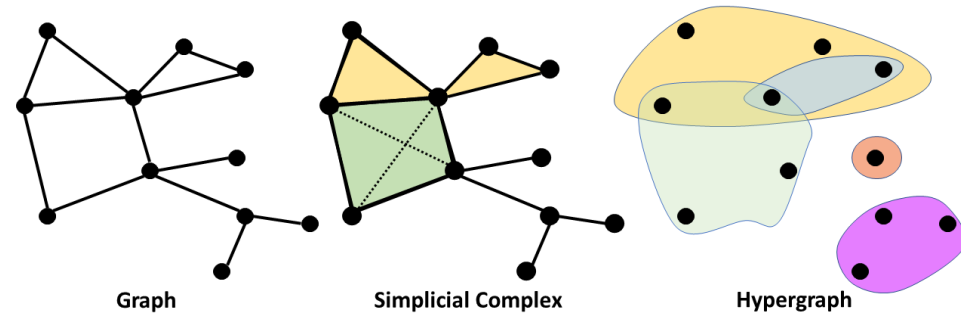
❖ Multiscale simplicial complex:

Persistent homology; Persistent cohomology...

Persistent Spectral theory (PerSpect)

Spectral models

- Spectral graph
- Spectral simplicial complex
- Spectral hypergraph



Filtration

- Nested sequence of Graphs
- Nested sequence of Simplicial Complexes
- Nested sequence of Hypergraph Laplacian

$$G^0 \subseteq G^1 \subseteq \dots \subseteq G^m$$
$$K^0 \subseteq K^1 \subseteq \dots \subseteq K^m$$
$$H^0 \subseteq H^1 \subseteq \dots \subseteq H^m$$

PerSpect = Spectral models + Filtration

- ❖ Persistent spectral graph
- ❖ Persistent spectral simplicial complexes
- ❖ Persistent spectral hypergraph

Hodge Laplacian matrix

The *adjoint* $\delta_n^* : C^{n+1}(K; \mathbb{F}) \rightarrow C^n(K; \mathbb{F})$ of the coboundary operator δ_n is defined by

$$(\delta_n f, g)_{C^{n+1}} = (f, \delta_n^* g)_{C^n},$$

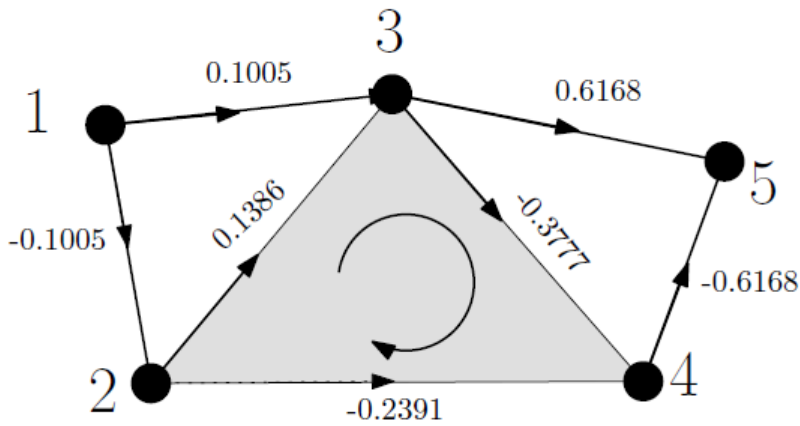
for all $f \in C^n(K; \mathbb{F})$ and $g \in C^{n+1}(K; \mathbb{F})$.

Let $\mathbb{F} = \mathbb{R}$ or \mathbb{C} . We define the *n-dimensional Laplace operator*

$\Delta_n : C^n(K; \mathbb{F}) \rightarrow C^n(K; \mathbb{F})$ by

$$\Delta_n = \delta_{n-1} \delta_{n-1}^* + \delta_n^* \delta_n,$$

where δ_n is the coboundary operator.



$$\mathbf{L}_k = \mathbf{B}_k^T \mathbf{B}_k + \mathbf{B}_{k+1} \mathbf{B}_{k+1}^T.$$

$$\mathbf{B}_1 = \begin{array}{ccccccc|c} [12] & [13] & [23] & [24] & [34] & [35] & [45] & \\ \hline -1 & -1 & 0 & 0 & 0 & 0 & 0 & [1] \\ 1 & 0 & -1 & -1 & 0 & 0 & 0 & [2] \\ 0 & 1 & 1 & 0 & -1 & -1 & 0 & [3] \\ 0 & 0 & 0 & 1 & 1 & 0 & -1 & [4] \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & [5] \end{array} \quad \mathbf{B}_2 = \begin{array}{c|c} [2, 3, 4] & \\ \hline 0 & [1, 2] \\ 0 & [1, 3] \\ 1 & [2, 3] \\ -1 & [2, 4] \\ 1 & [3, 4] \\ 0 & [3, 5] \\ 0 & [4, 5] \end{array}$$

The Laplacians are computed as

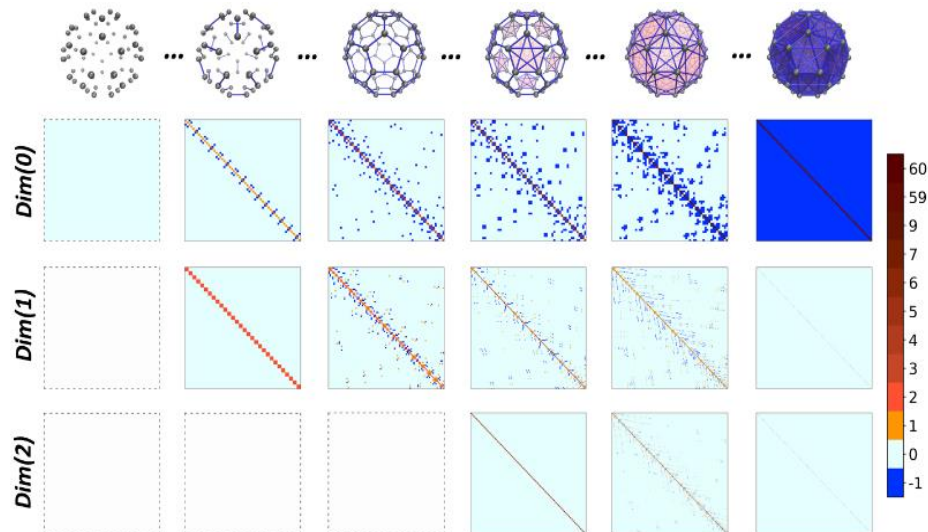
$$\mathcal{L}_0 = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix}, \quad \mathcal{L}_1 = \begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & -1 & -1 & 0 \\ -1 & 1 & 3 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 3 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 3 & 1 & -1 \\ 0 & -1 & -1 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & -1 & -1 & 1 & 2 \end{pmatrix}, \quad \mathcal{L}_2 = 3.$$

Control Using Higher Order Laplacians in Network Topologies

Persistent spectral simplicial complex

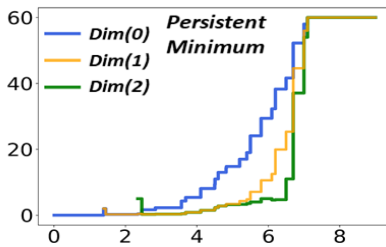
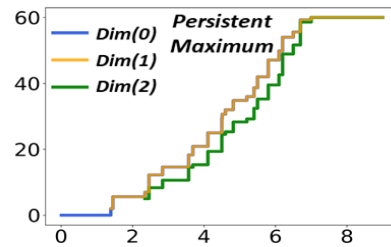
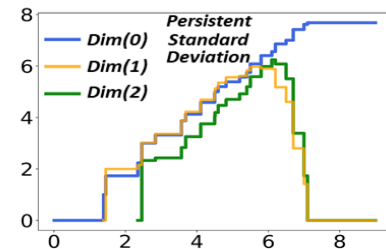
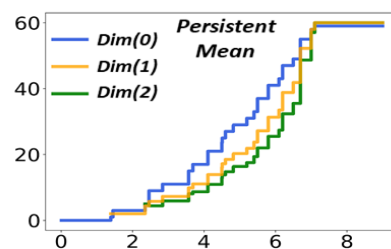
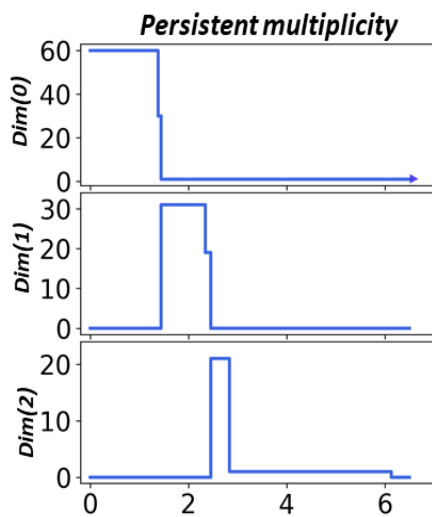
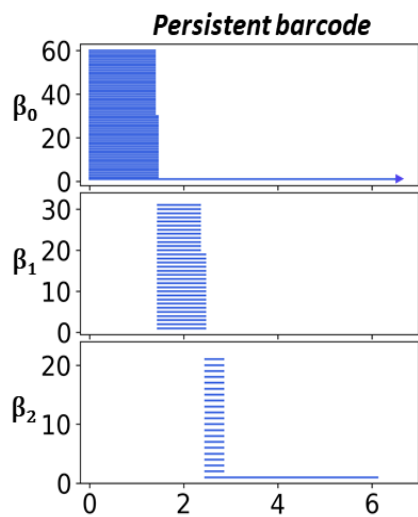
Boundary operator

$$B_k(i, j) = \begin{cases} 1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \sim \sigma_j^k \\ -1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \not\sim \sigma_j^k \\ 0, & \text{if } \sigma_i^{k-1} \not\subset \sigma_j^k. \end{cases}$$



Combinatorial Laplacian (Hodge Laplacian)

$$L_k = B_k^T B_k + B_{k+1} B_{k+1}^T.$$

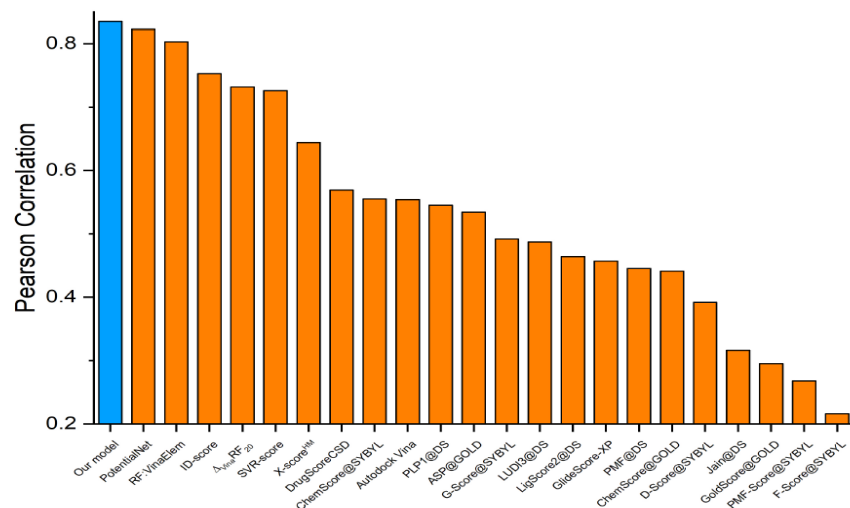


Multiplicity of zero eigenvalues (Persistent multiplicity) from PerSpect simplicial complex is equivalent to persistent Betti number.

PerSpect variables change with filtration parameter and incorporate in them related geometric information.

Ours: 0.836

Dataset 2007

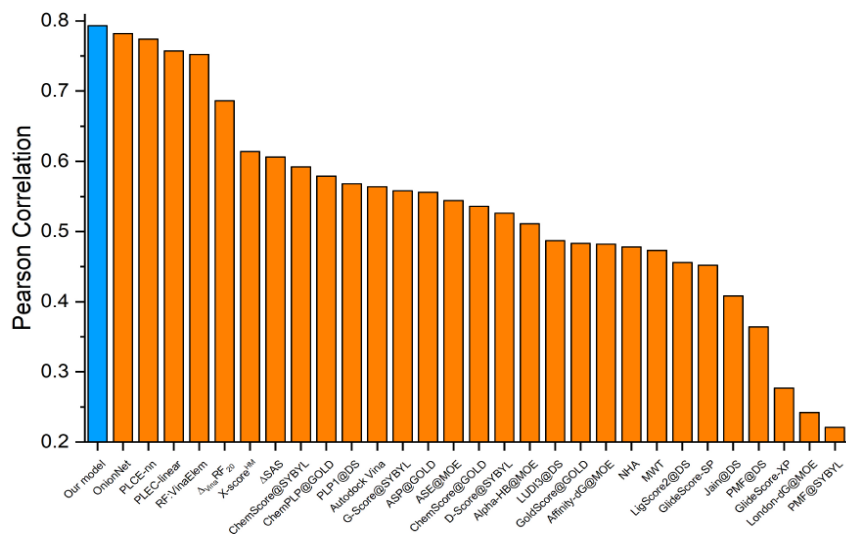


Benchmark testing with PDBbind datasets

Model setting:
Spectral vectors
+
Random forest

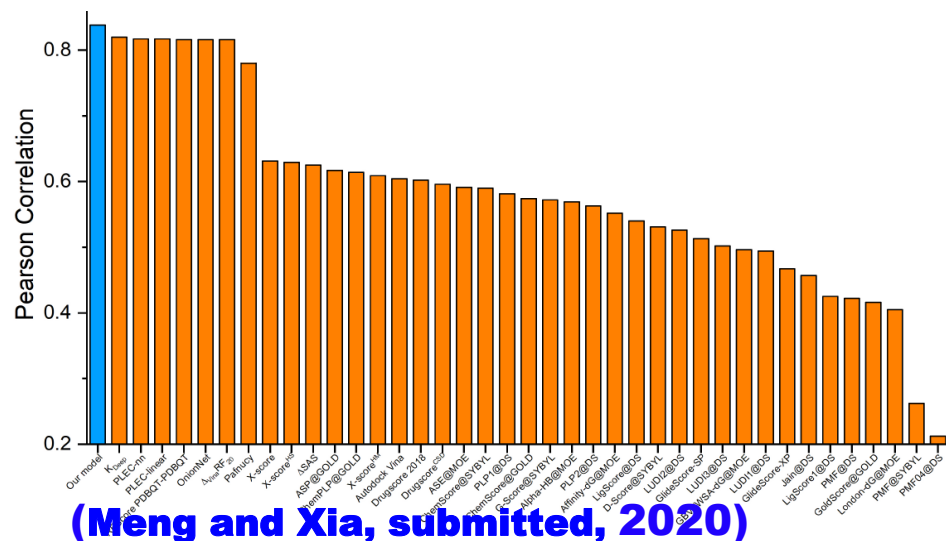
Ours: 0.793

Dataset 2013



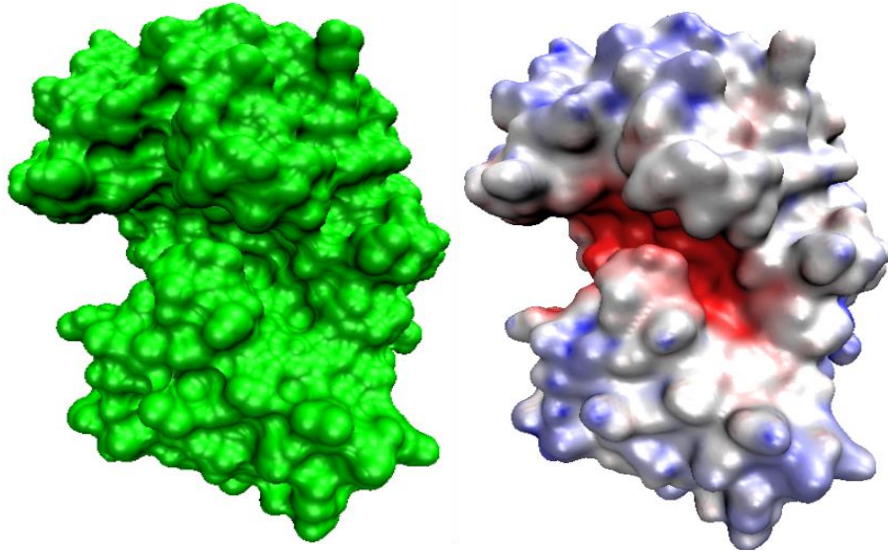
Ours: 0.840

Dataset 2016

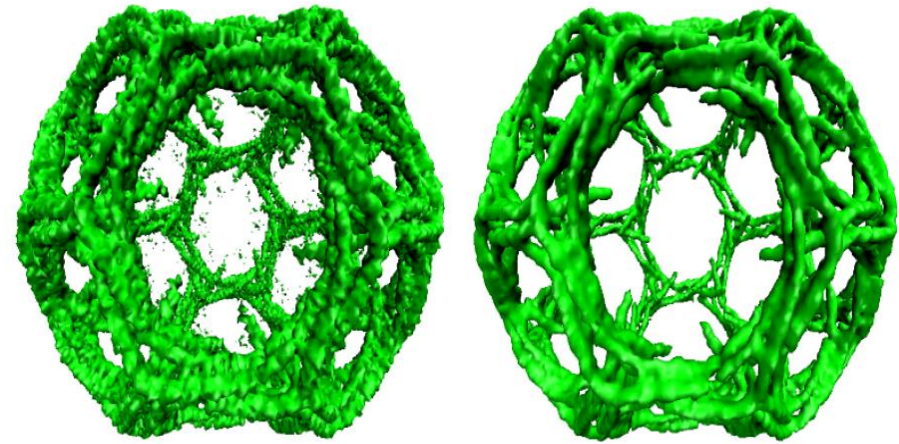


(Meng and Xia, submitted, 2020)

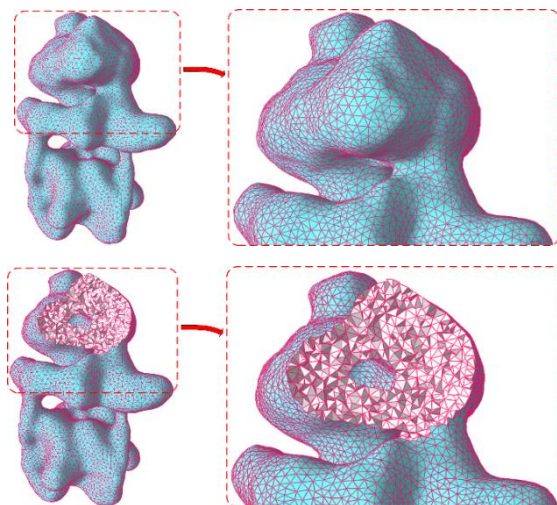
Variational multiscale modeling



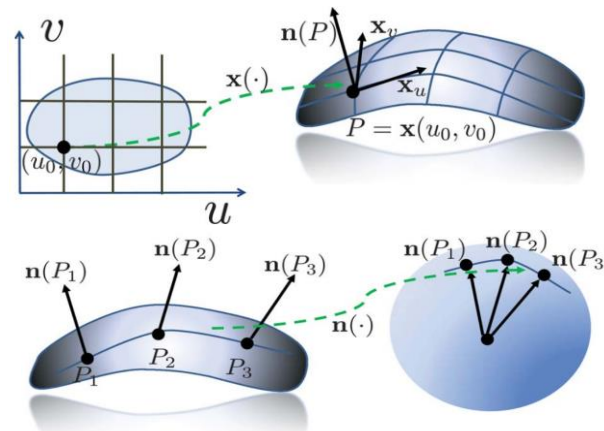
Geometric flow for noise reduction



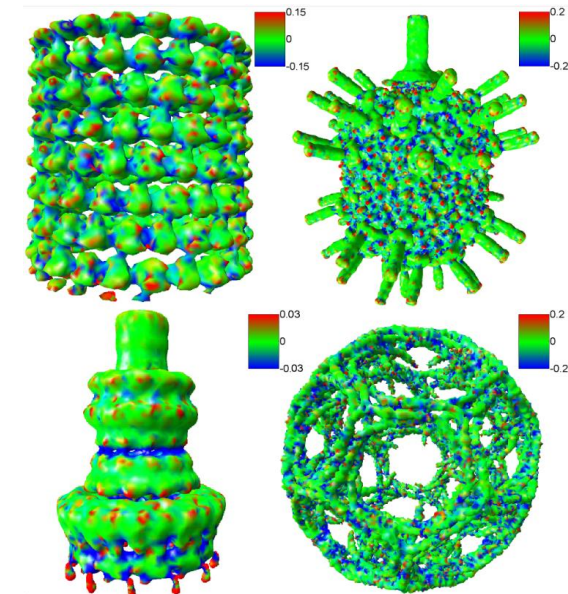
Geometric modeling of biomolecules



**Delaunay triangulation
based mesh generation**



**First and second
fundamental form**
(Feng, Xia, etc., JCC, 2013)

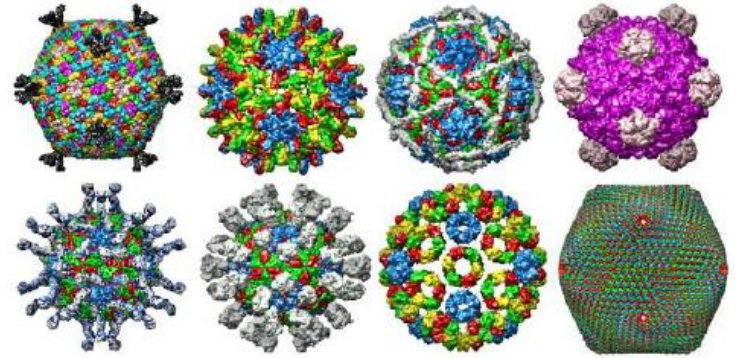


Gaussian curvature

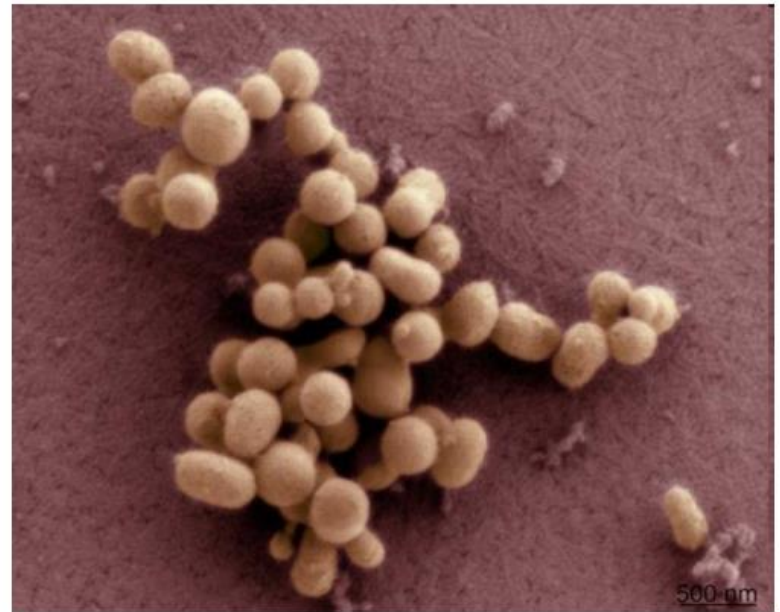


Minimal Surfaces

A way to minimize energy and maximize stability



Viral morphology



The first man-made life, Bacterium, *M. mycoides*, based on information from a computer



Free energy functional of a surface model

$$G = \gamma(\text{Area}) = \int_U \gamma \sqrt{g} du_1 du_2$$

where **gamma** is the surface tension and **g** is the Gram determinant: $g = 1 + S_1^2 + S_2^2$ of matrix $(g_{ij}) = \begin{pmatrix} 1 + S_1^2 & S_1 S_2 \\ S_2 S_1 & 1 + S_2^2 \end{pmatrix}$

Minimizing the surface free energy with the **Euler-Lagrange equation**, we arrive at the **generalized mean curvature equation**:

$$\Delta_{\Xi} S = \frac{1}{\sqrt{g}} \sum_{ij} \frac{\partial}{\partial x_i} \left(\gamma \sqrt{g} g^{ij} \frac{\partial}{\partial x_j} S \right) = \frac{1}{\sqrt{g}} \nabla \cdot \left(\frac{\gamma}{\sqrt{g}} \nabla S \right) = 0$$

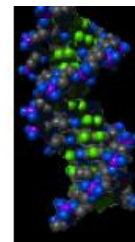
where $\Delta_{\Xi} S$ is the **Laplace-Beltrami operator**. We solve the **Laplace-Beltrami equation** below to generate **minimal molecular surfaces**:

$$\frac{\partial S}{\partial t} = \sqrt{g} \left[\nabla \cdot \left(\frac{\gamma \nabla S}{\sqrt{g}} \right) \right]$$

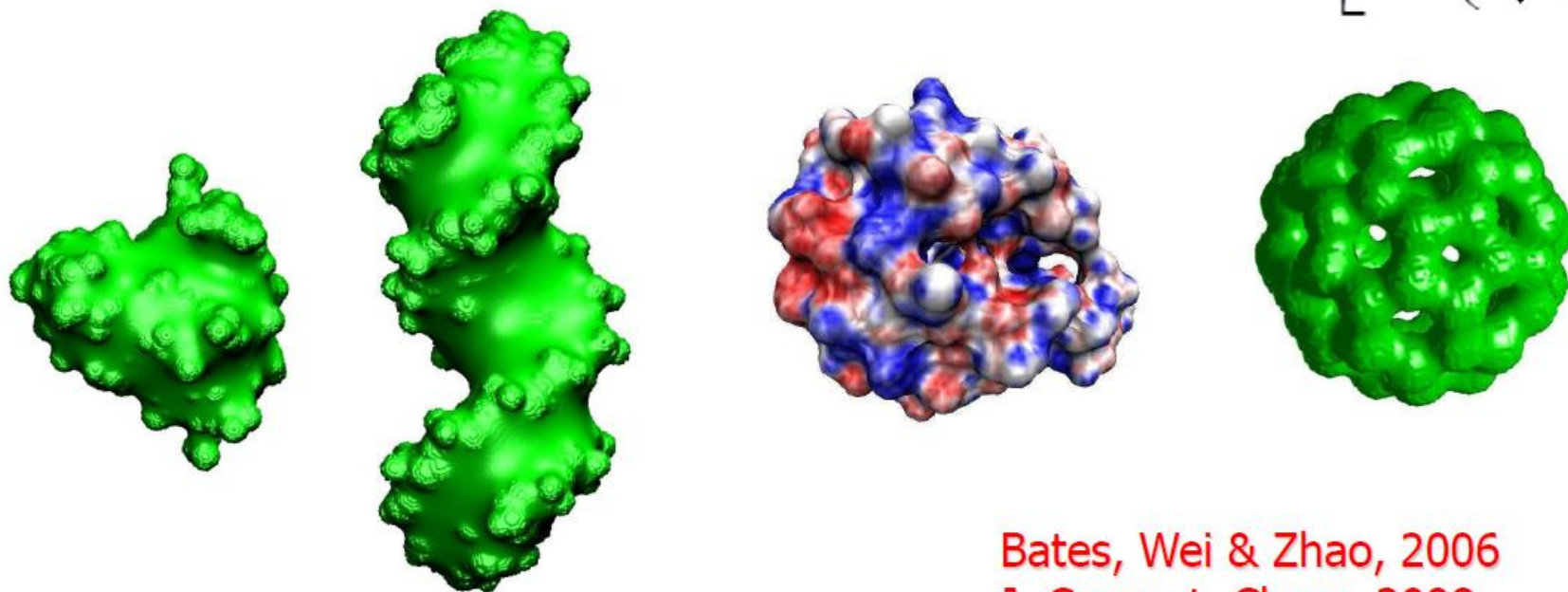
(Bates, Wei & Zhao, 2006)

Minimal Molecular surface

The first biomolecular surface constructed
with the variational principle



Generalized Laplace-Beltrami flow:
$$\frac{\partial S}{\partial t} = \sqrt{g} \left[\nabla \cdot \left(\frac{\gamma \nabla S}{\sqrt{g}} \right) \right]$$



Bates, Wei & Zhao, 2006
J. Comput. Chem. 2008

Differential geometry based **nonpolar** solvation model

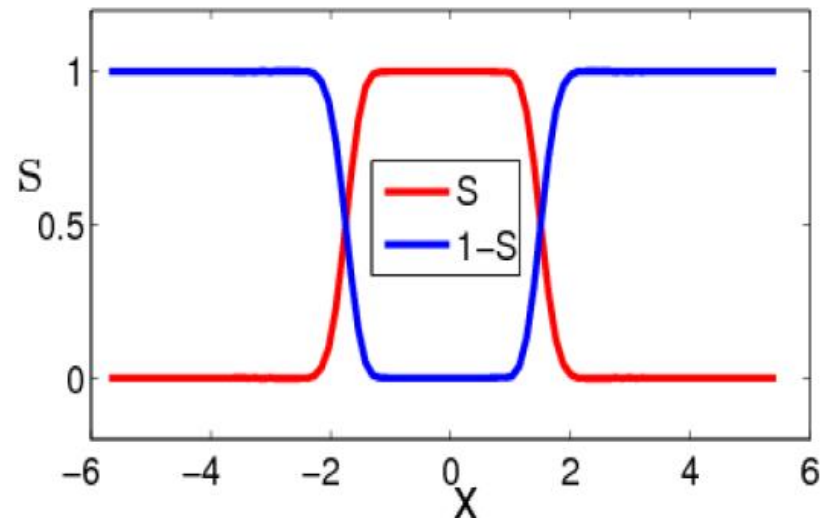
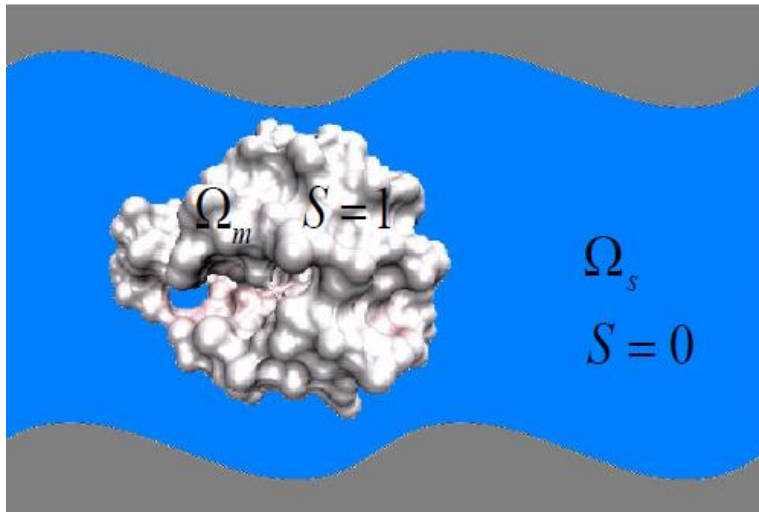
$$G = \int_{\Omega} [\gamma |\nabla S| + Sp + (1 - S)U] dr$$

(Wei, BMB, 2010;
Chen, Zhao, Baker, Bates, Wei, 2011)

area + volume + van der Waals

$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \cdot \frac{\gamma \nabla S}{|\nabla S|} - p + U \right]$$

Laplace-Beltrami equation



Differential geometry based solvation model

$$G = \int_{\Omega} [Nonpolar + Electro] dr$$

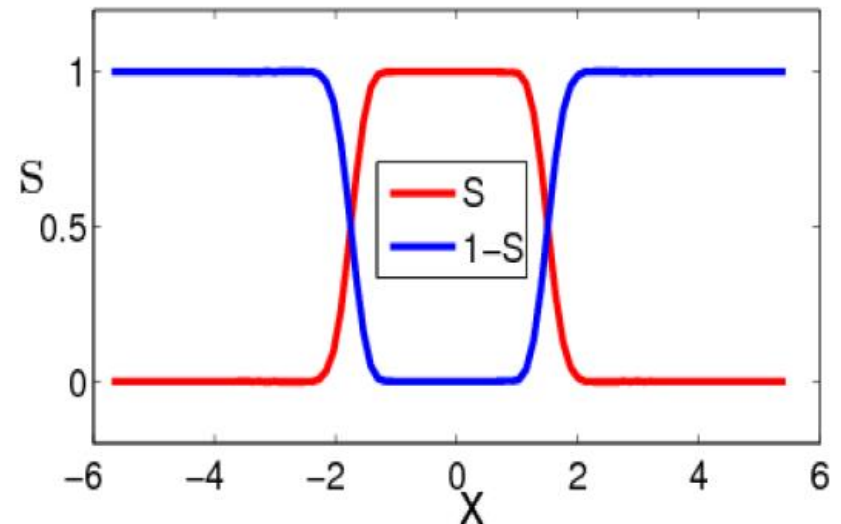
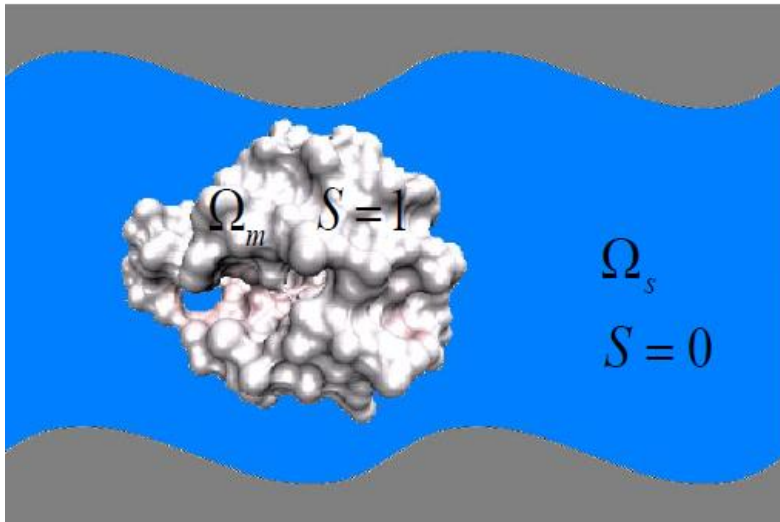
(Wei, BMB, 2010;
Chen, Baker, Wei, JCP, 2010)

Geometric = area + volume + van der Waals:

$$Nonpolar = \gamma |\nabla S| + Sp + (1 - S)U$$

Electro = electric field + solute charges + solvent ions:

$$Electro = S \left(\frac{\epsilon_m}{2} |\nabla \phi|^2 - \phi n \right) + (1 - S) \left[\frac{\epsilon_s}{2} |\nabla \phi|^2 + kT \sum_i c_i (e^{-q_i \phi / kT} - 1) \right]$$



Variation of the total free energy functional

$$\frac{\partial S}{\partial t} = \nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|} \right) - p + U - \frac{\epsilon_m - \epsilon_s}{2} |\nabla \phi|^2 + kT \sum_i c_i \left(e^{-q_i \phi / kT} - 1 \right) - \phi n$$

Generalized Laplace Beltrami equation

$$-\nabla \cdot (\epsilon(S) \nabla \phi) = (1 - S) \sum_i q_i c_i e^{-q_i \phi / kT} + S n$$

Generalized Poisson-Boltzmann equation

- Electrostatic binding and solvation energies
- pKa, pH values
- Electrostatic forces, ionic distributions
- Electrostatic matching between proteins and ligands
- Stability of protein folding
- Molecular dynamics
- A tool for rational drug design (interactions of receptor-inhibitor, protein-ligand, protein-protein, signal, enzyme, regulator, etc.)



THANK YOU

Dziękuję

Terima kasih

Gracias

감사합니다

Köszönöm

謝謝

Děkuji

Multumesc

Спасибо

Merci

Danke

Gracias

Obrigado

ありがとう

謝謝

Kiitos

Merci

Terima kasih

Teşekkür ederim

Terima kasih

Salamat

Dank U

Merci

Děkuji

Tack

謝謝

Merci