Topological modelling and analysis for biomolecular data: 2 persistent homology

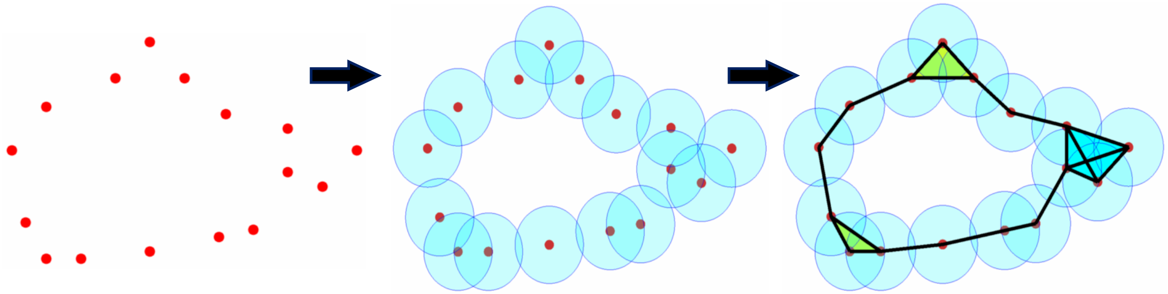
Kelin Xia

School of Physical and Mathematical Sciences, Nanyang Technological University

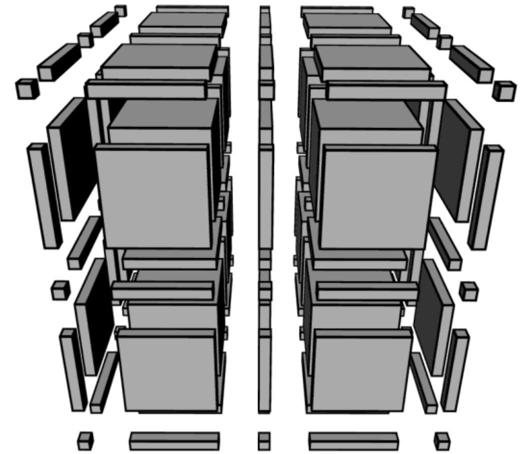
应用拓扑短期课程与报告会, 12月13-16, 2020

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

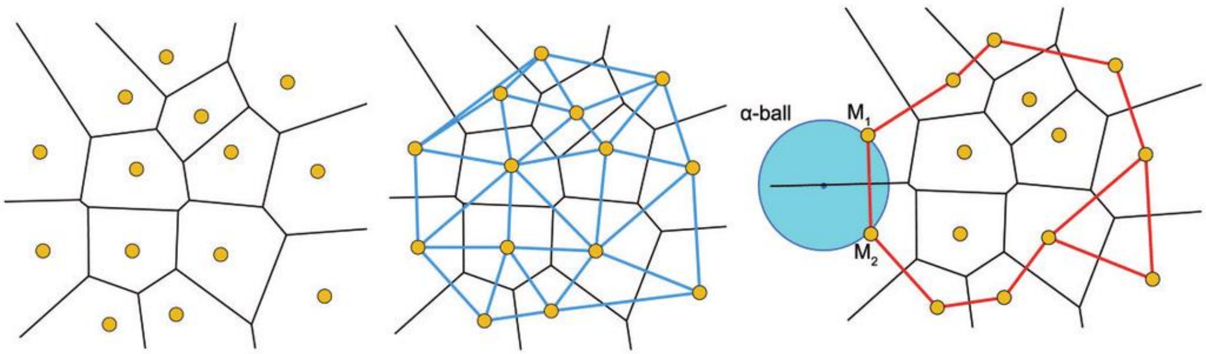
Vietoris-Rips complex



Cubical complex



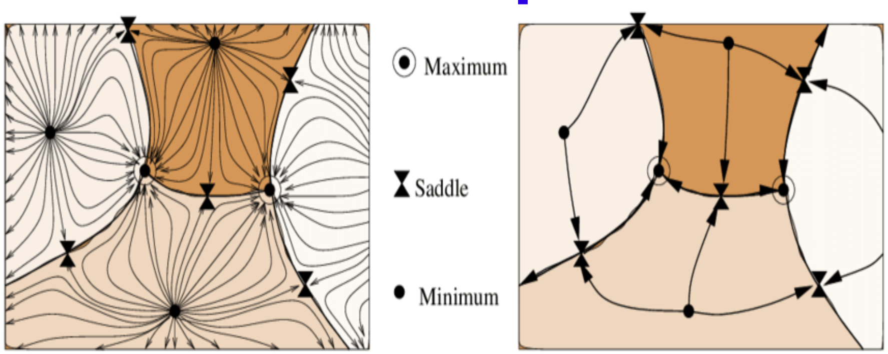
Alpha complex



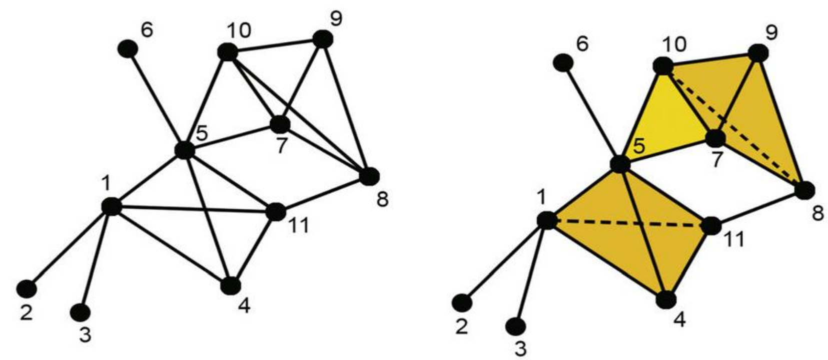
Voronoi diagram Delaunay triangulation Alpha shape

Simplicial complexes

Morse complex



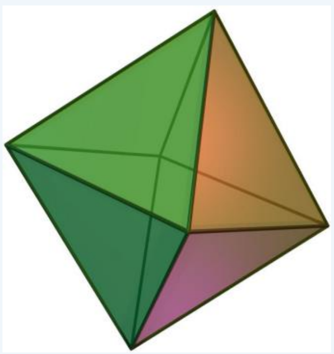
Clique complex



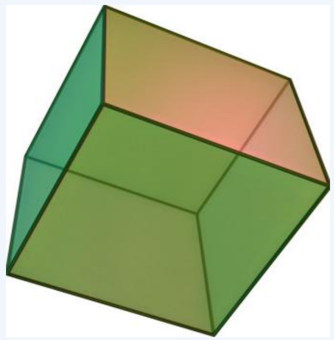
Euler Characteristic

POLYHEDRONS

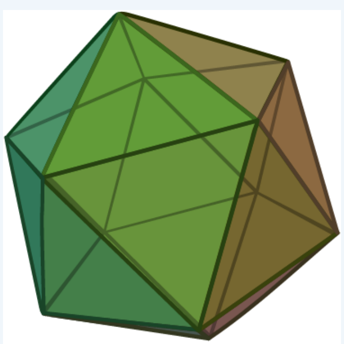
$$\chi(M) = V_{vertex} - E_{edge} + F_{face}$$



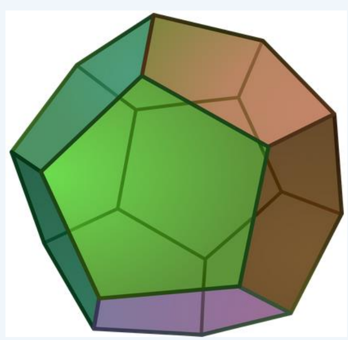
$\chi = 6 - 12 + 8 = 2$



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



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

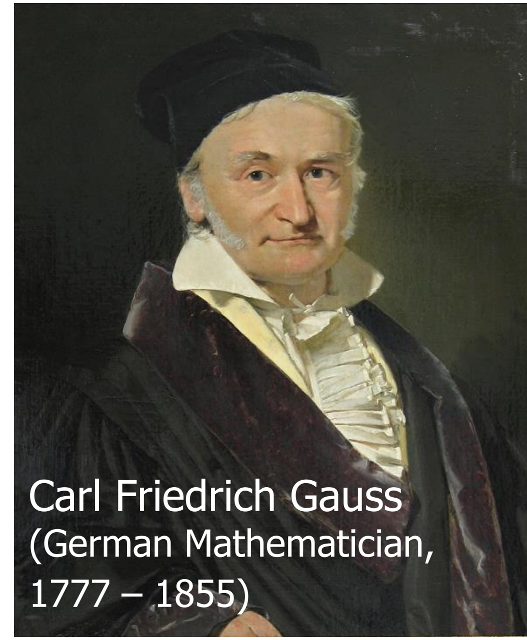


$\chi = 20 - 30 + 12 = 2$

$$\chi = 2$$

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

"Spherical cow", Wikipedia



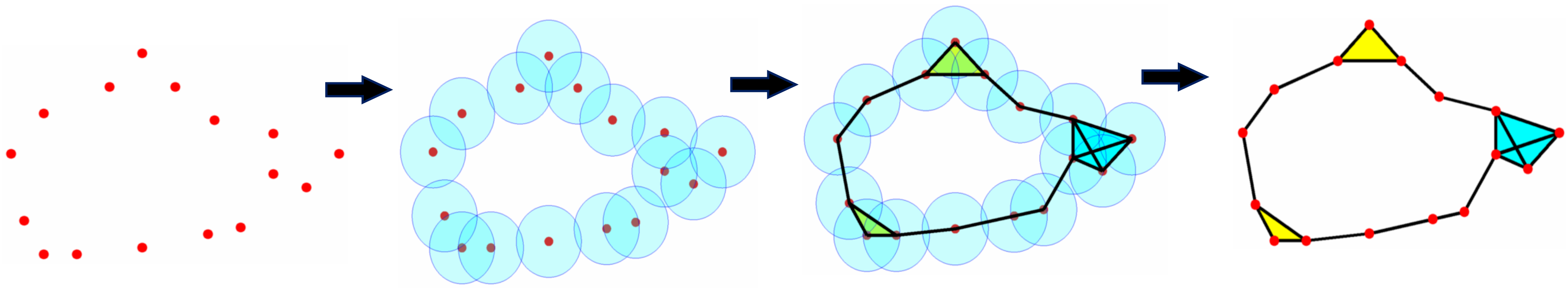
Gauss-Bonnet Theorem

Gaussian curvature

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

Connection between differential geometry and topology

Topological data analysis (TDA)



Point cloud data **Topological space** **Vietoris-Rips complex** **Simplicial complex**

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

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

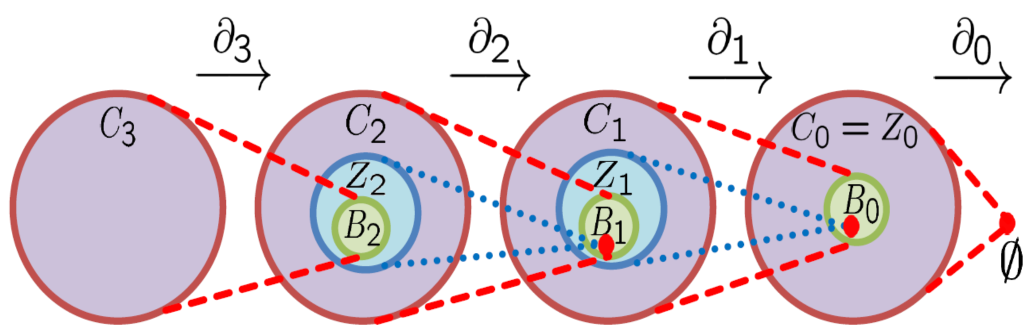
$Z_k = \operatorname{Ker} \partial_k$

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

Quotient group:

$H_k = Z_k / B_k$

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

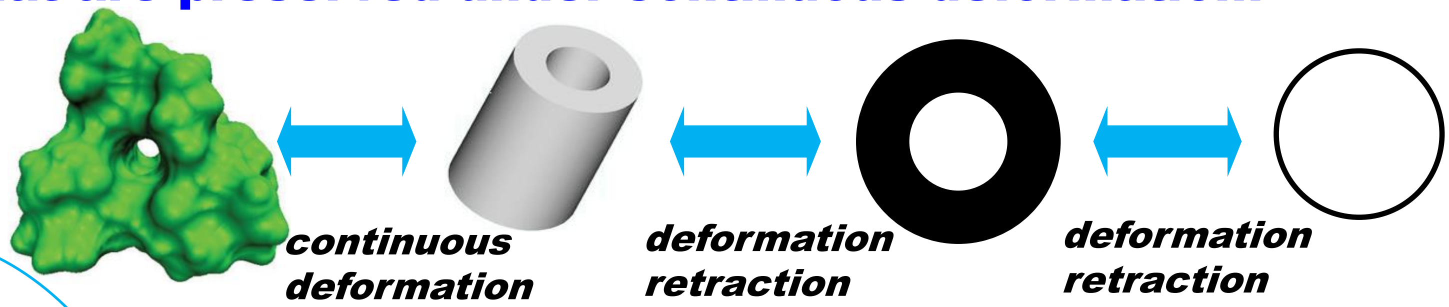


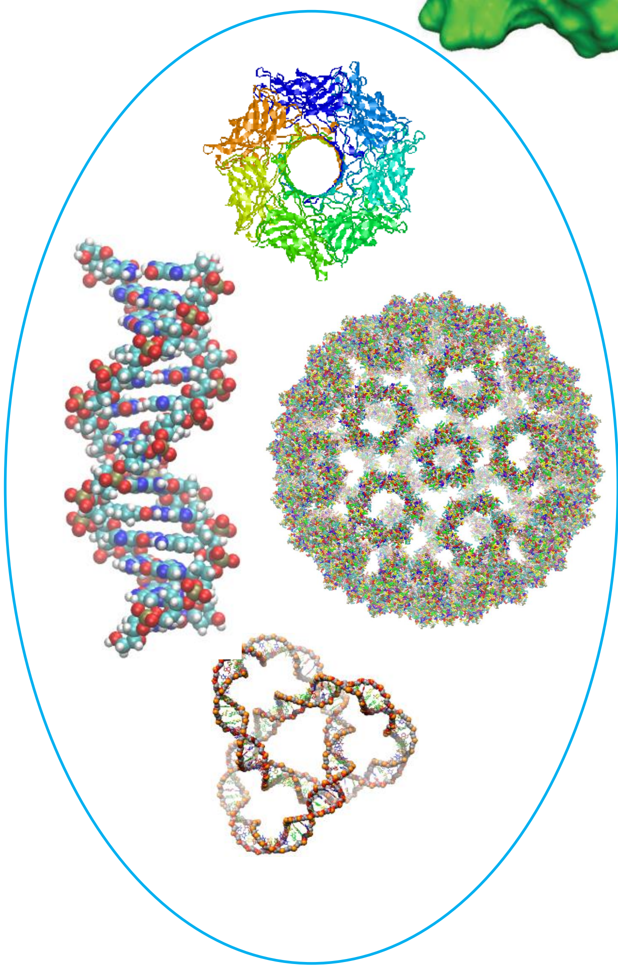
The topological information can be calculated!!

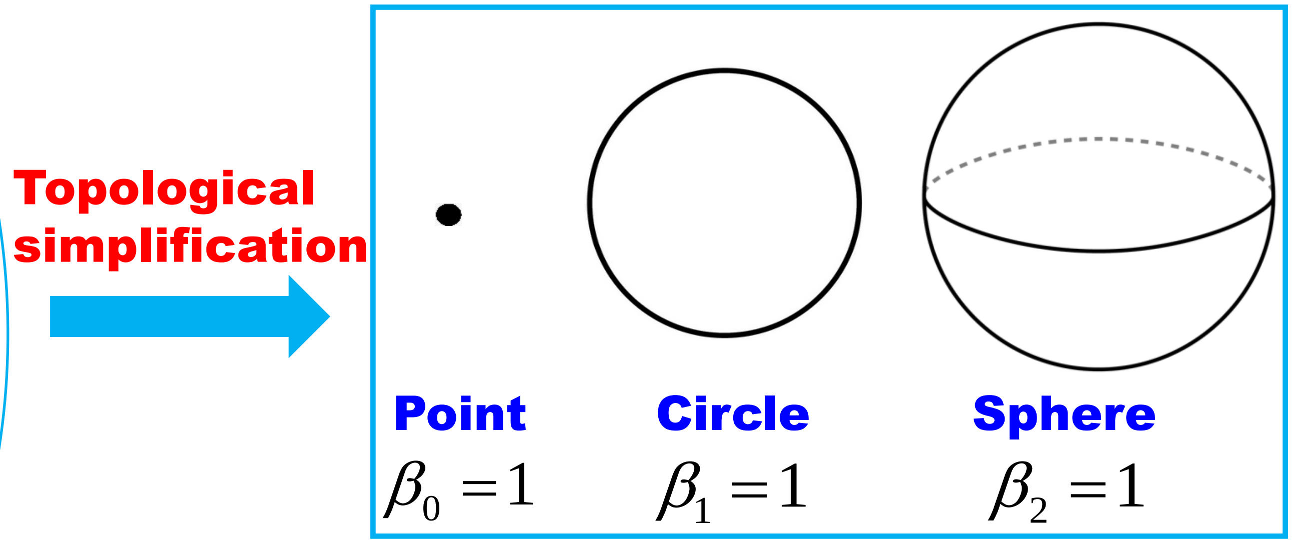
Topological invariant--Betti number

Properties that are preserved under continuous deformation!

Homotopy equivalent:







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**

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.

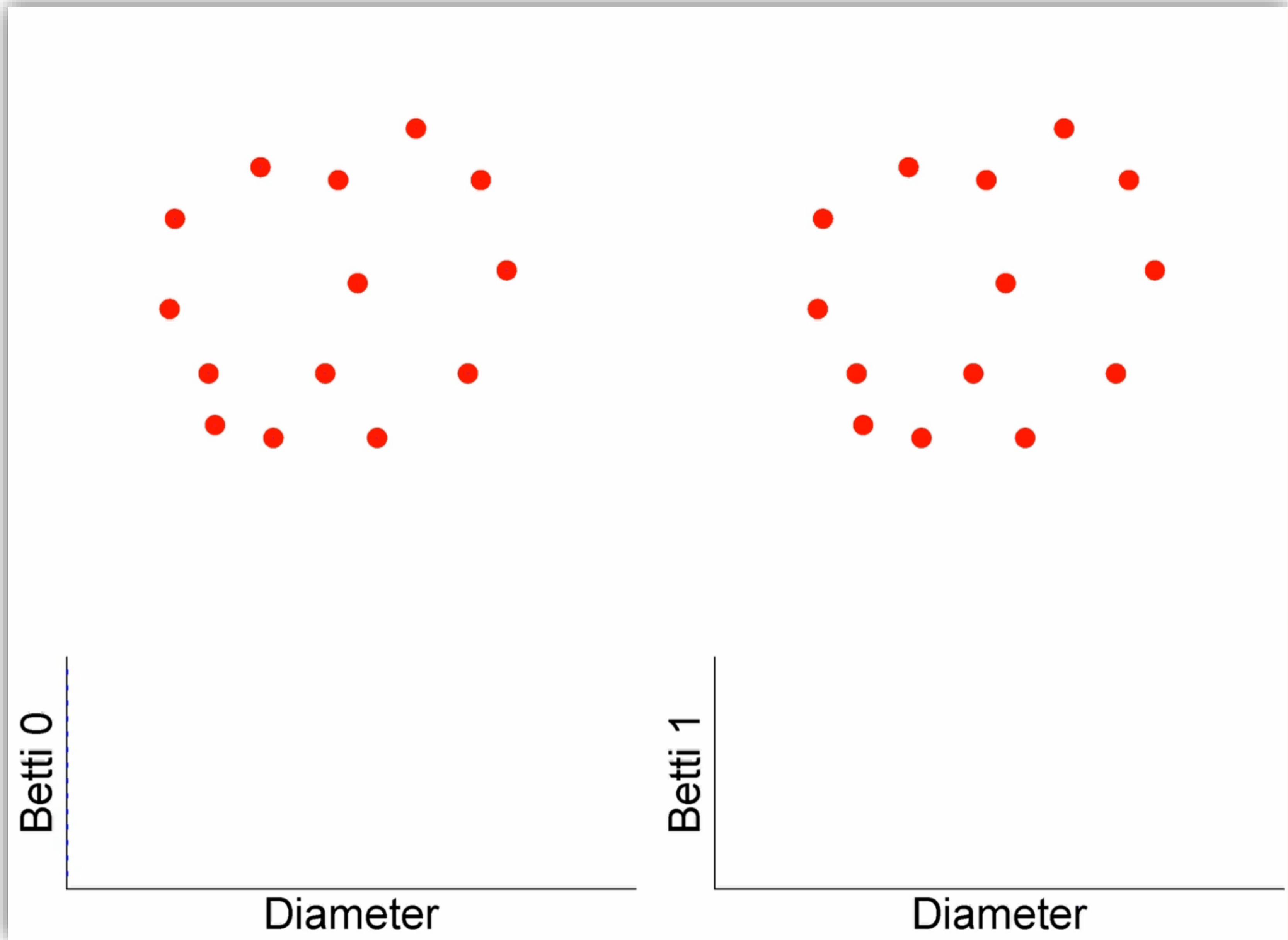
Promises from persistent homology:

- ✓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,
......

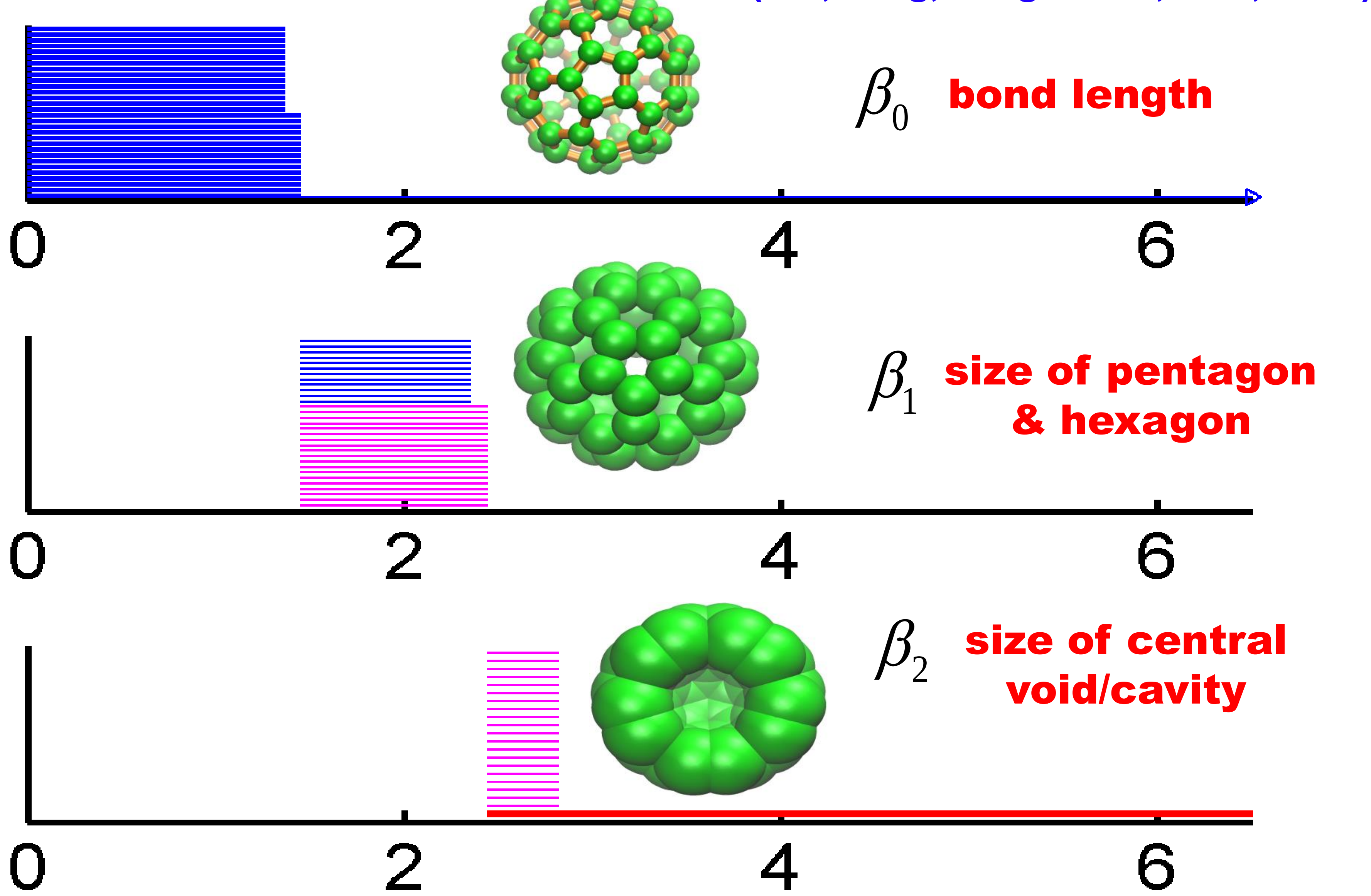
Persistent homology and persistent barcodes



Simplest molecules to start with:

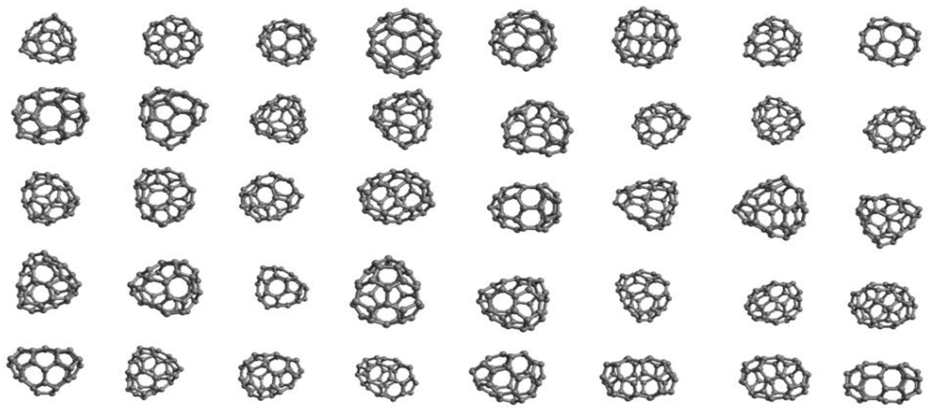
fullerene C_{60}

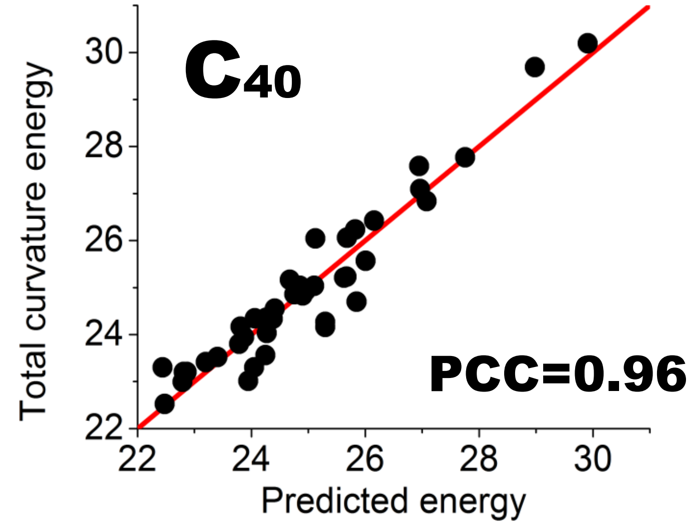
(Xia, Feng, Tong & Wei, JCC, 2015)

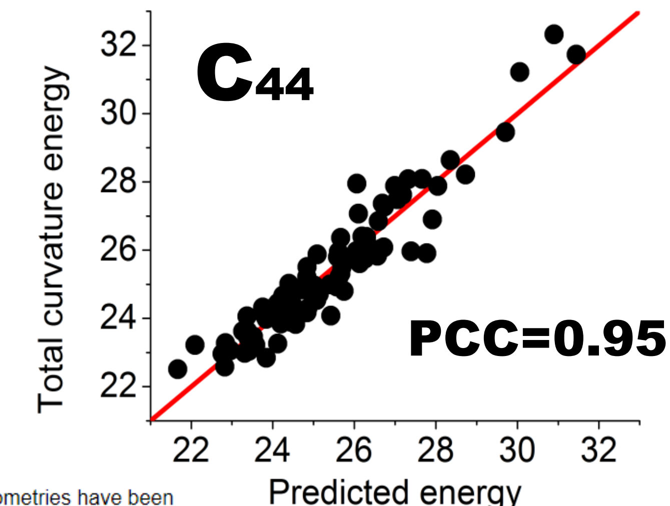


Fullerene isomers

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

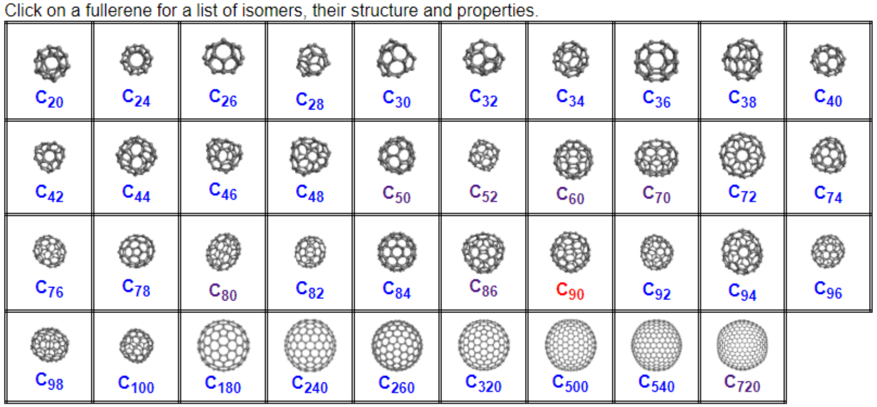






C_n Fullerenes

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



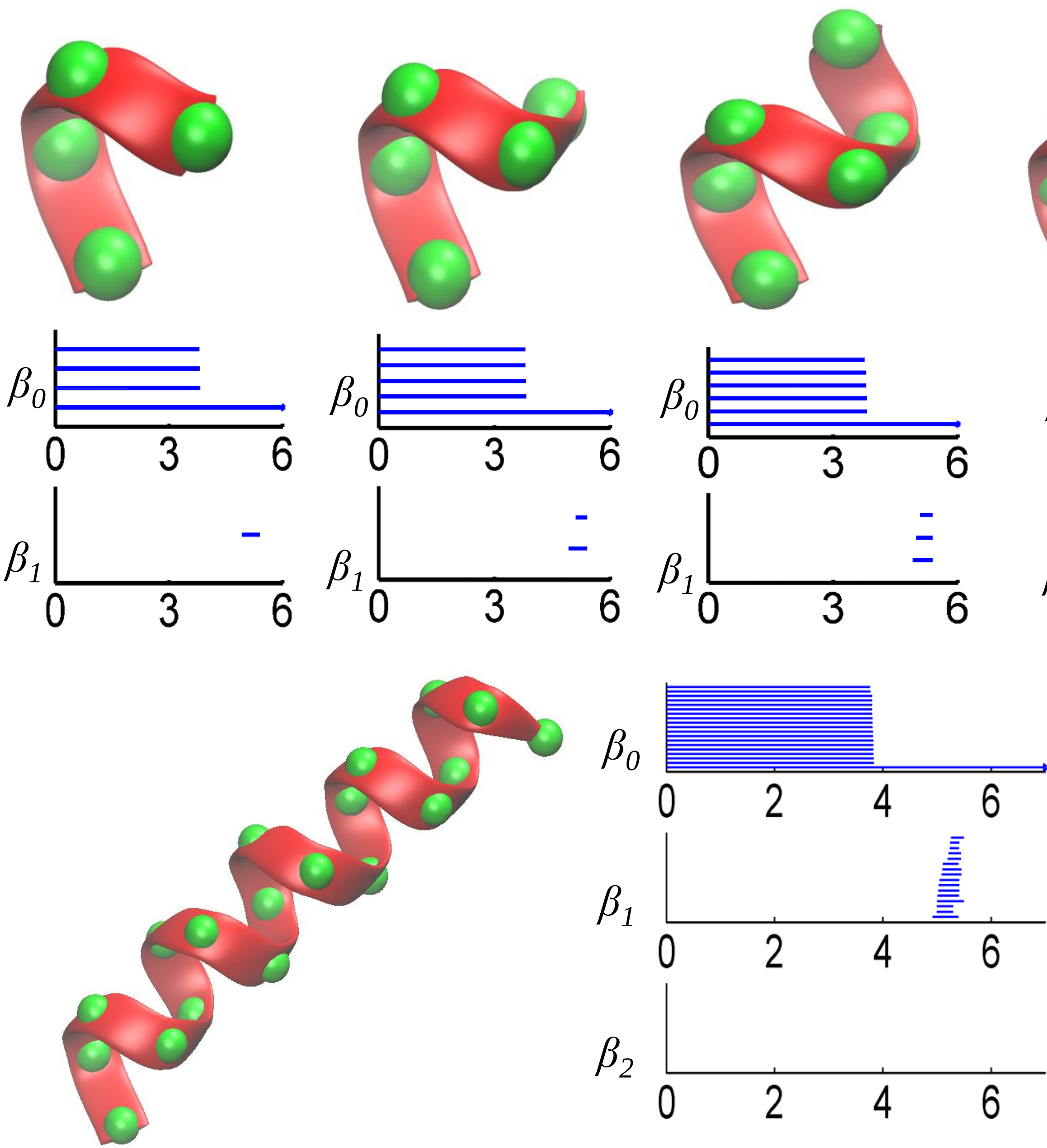
The fullerene geometries are based on structures in the Fullerene Library that has been created by M. Yoshida. The geometries have been reoptimized using a fast Dreiding-like forcefield that is built into the free Discovery Studio Visualizer. The numbering scheme of fullerene isomers seems to agree with 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.

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

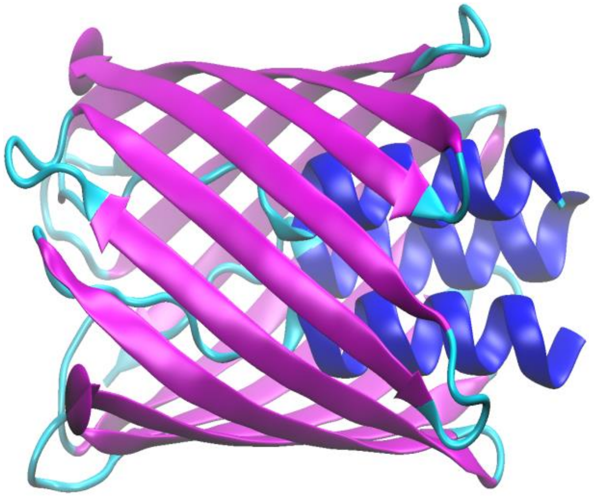
Topological fingerprints of an alpha helix

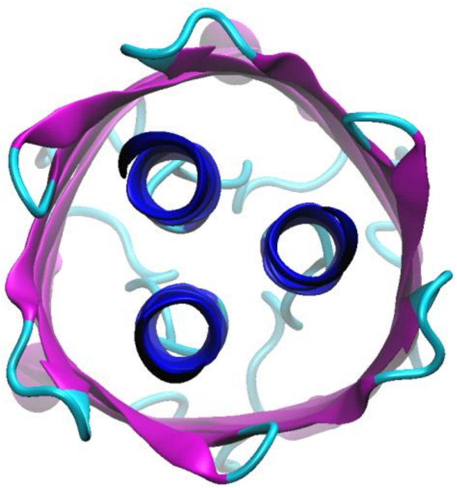


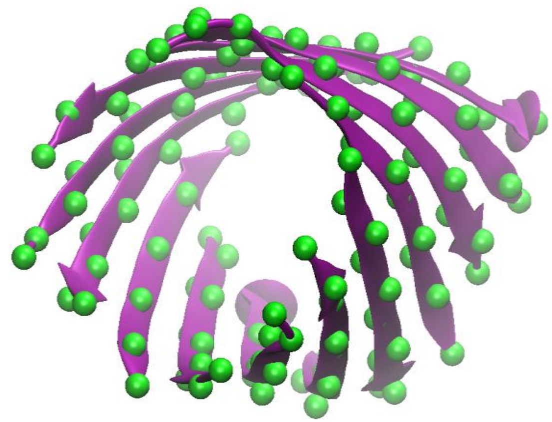
(Xia & Wei, IJNMBE, 2014)

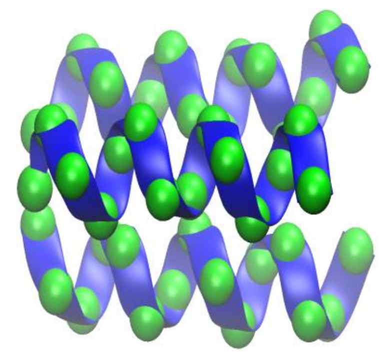
Topological fingerprints of beta barrel

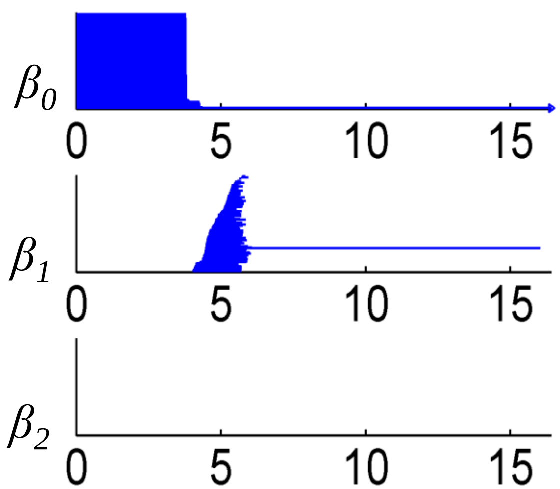
Protein:2GR8

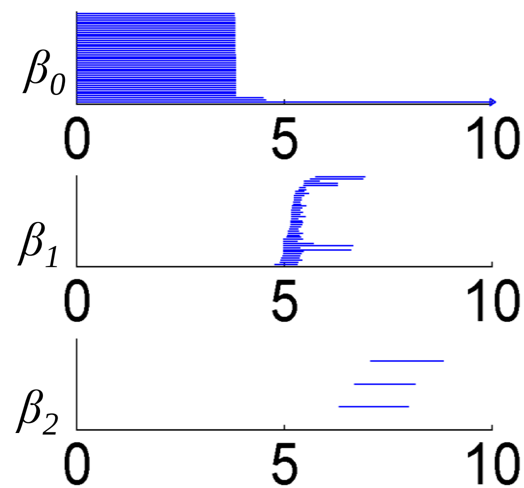




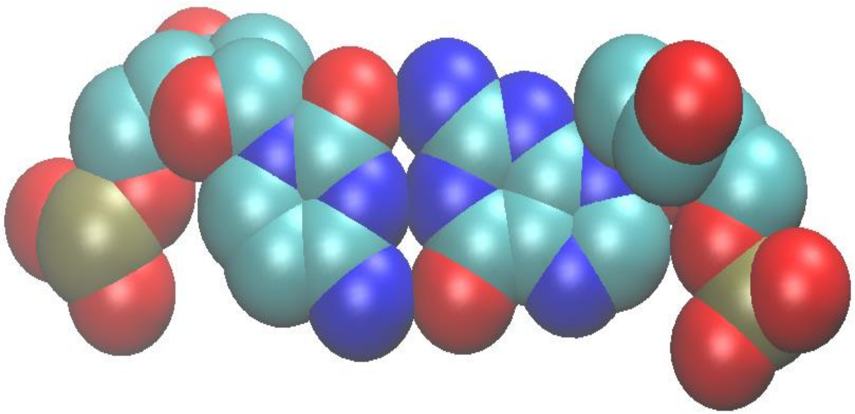


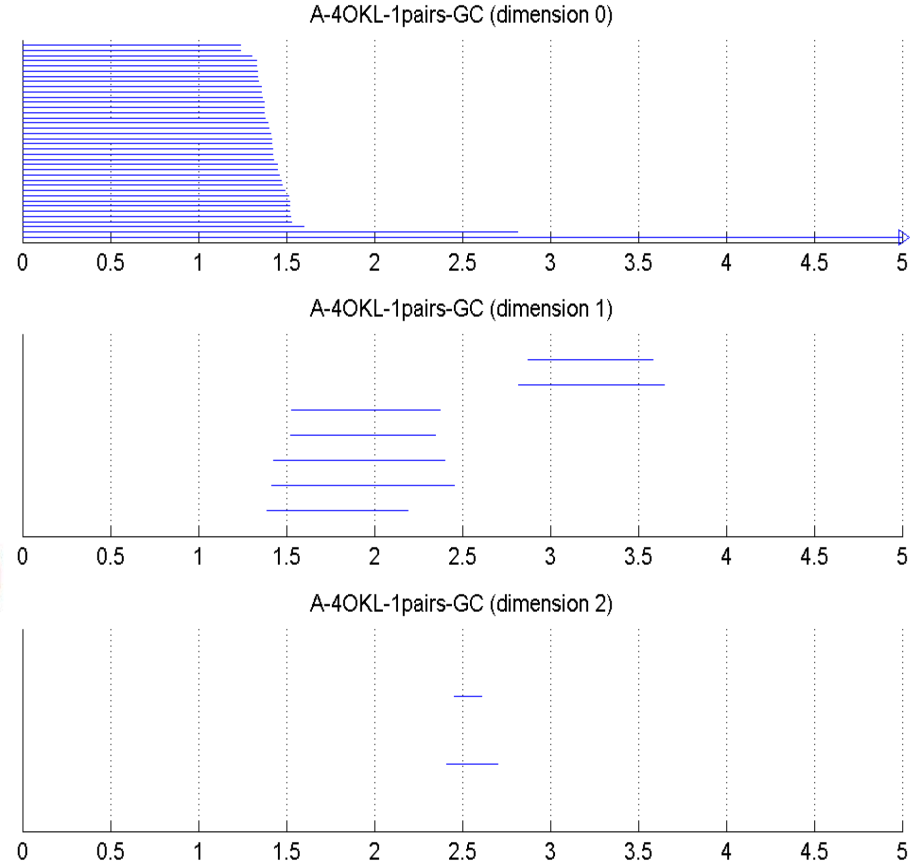


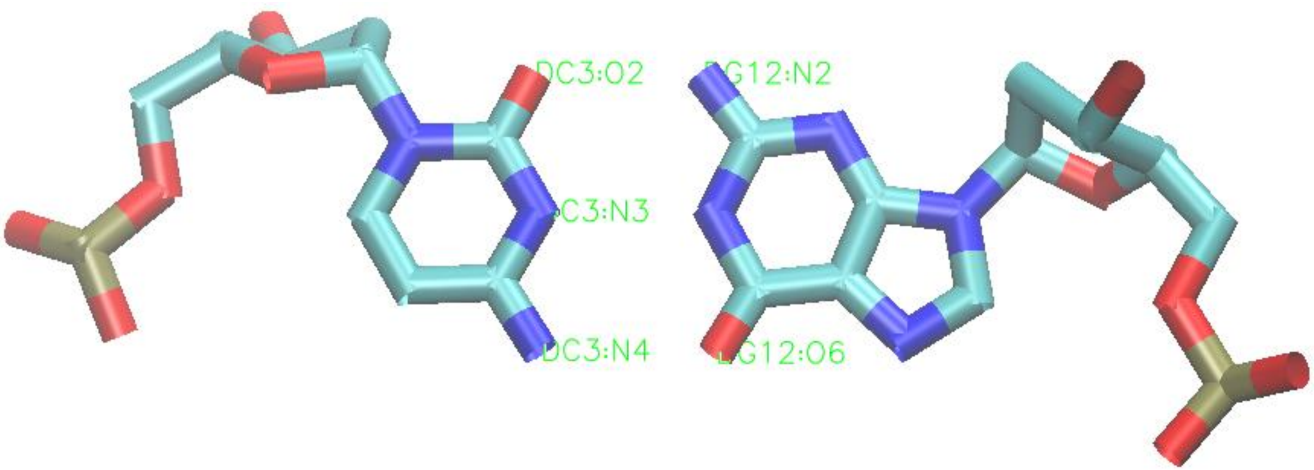




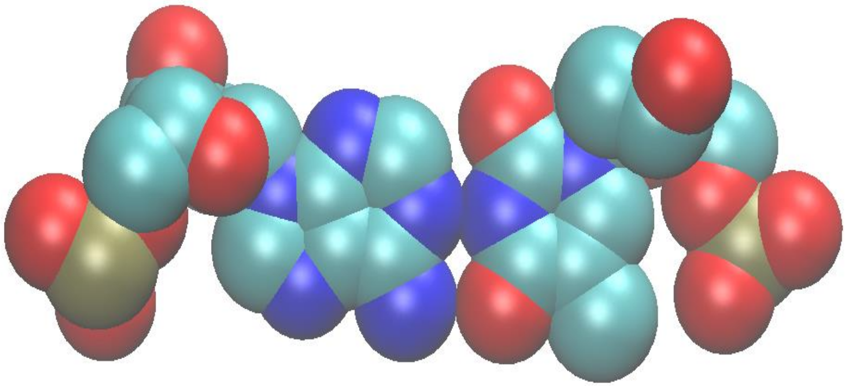
DNA:G-C pair

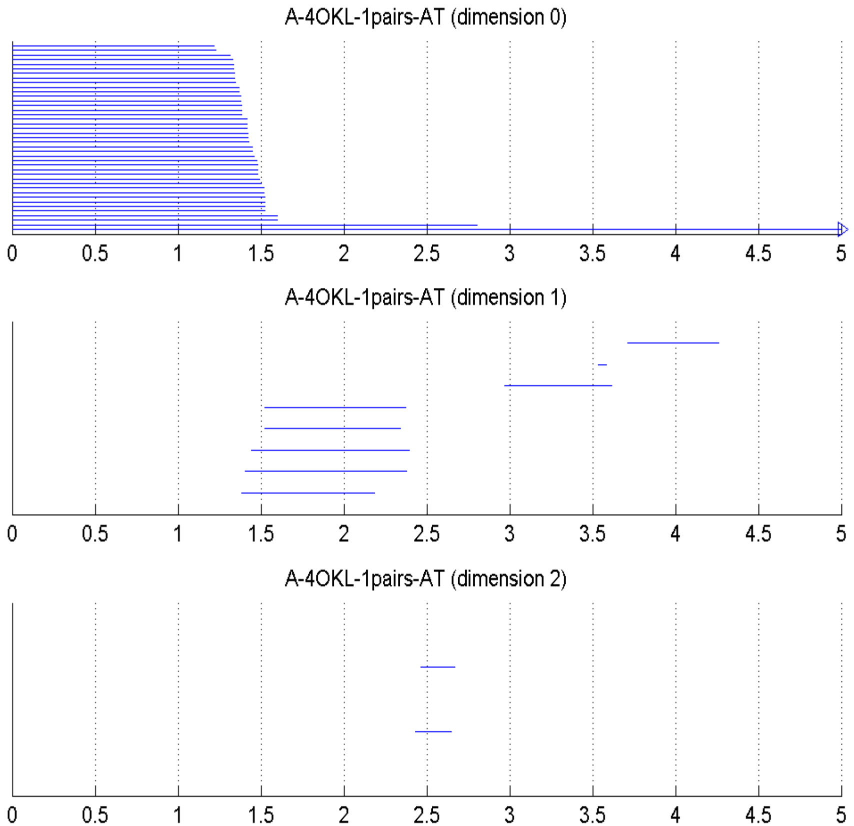


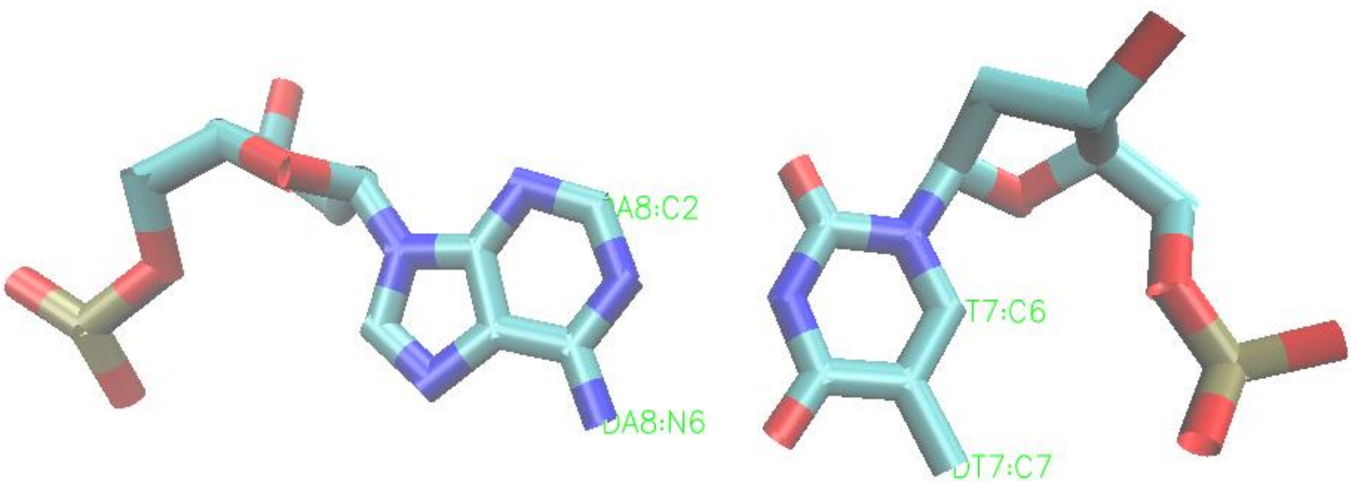

A-4OKL-1pairs-GC (dimension 0)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
A-4OKL-1pairs-GC (dimension 1)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
A-4OKL-1pairs-GC (dimension 2)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5


DC3:O2
G12:N2
C3:N3
DC3:N4
G12:O6

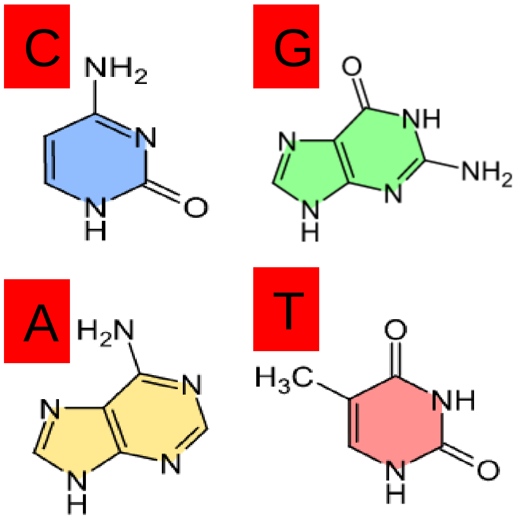
DNA: A-T pair

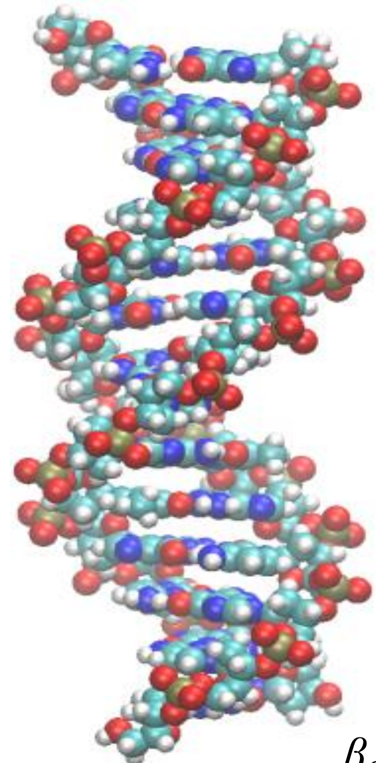



A-4OKL-1pairs-AT (dimension 0)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
A-4OKL-1pairs-AT (dimension 1)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
A-4OKL-1pairs-AT (dimension 2)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

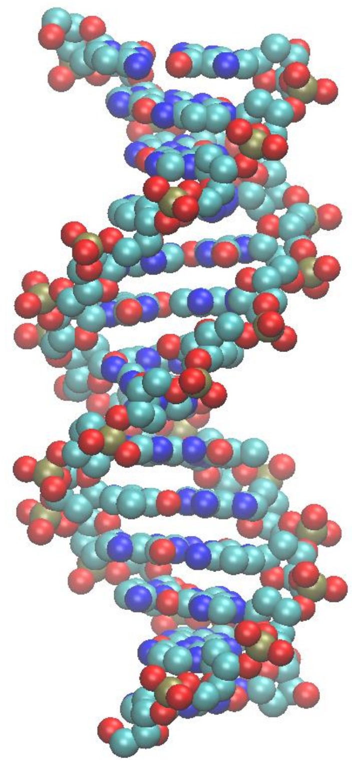

DA8:C2
DA8:N6
DT7:C6
DT7:C7

DNA topological fingerprints:

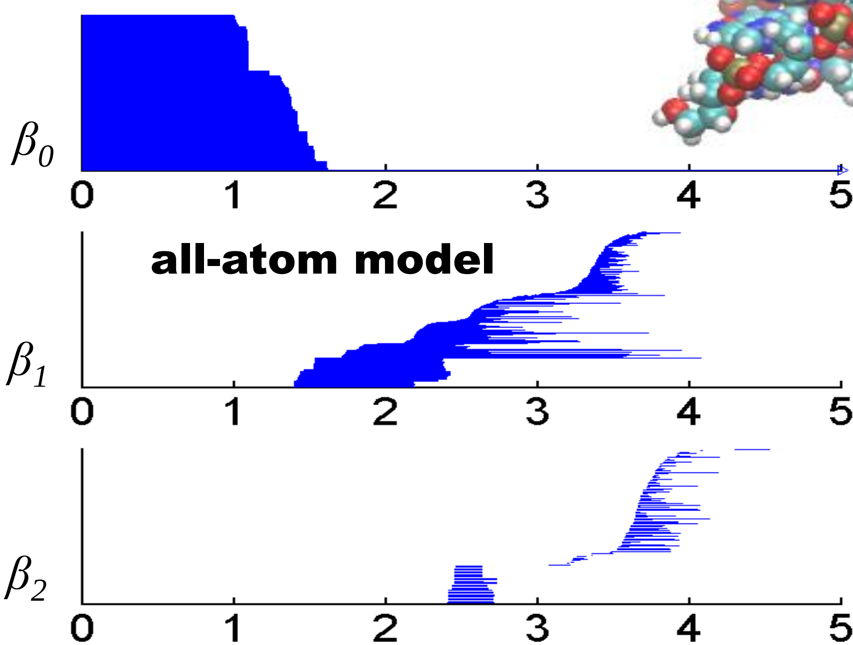


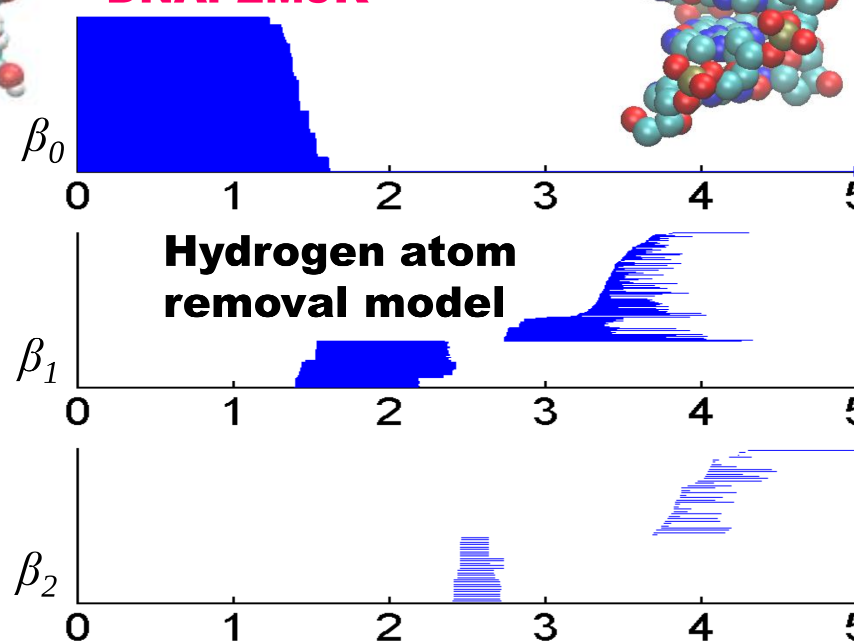


Hydrogen atom removal reveals a clear separation between local and global topological properties



DNA: 2MJK

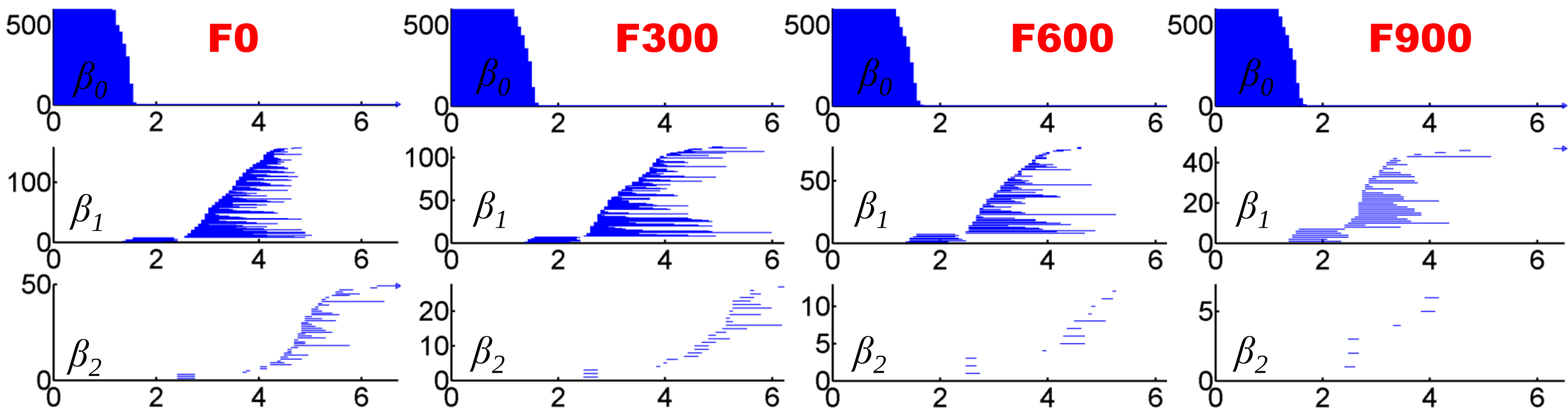


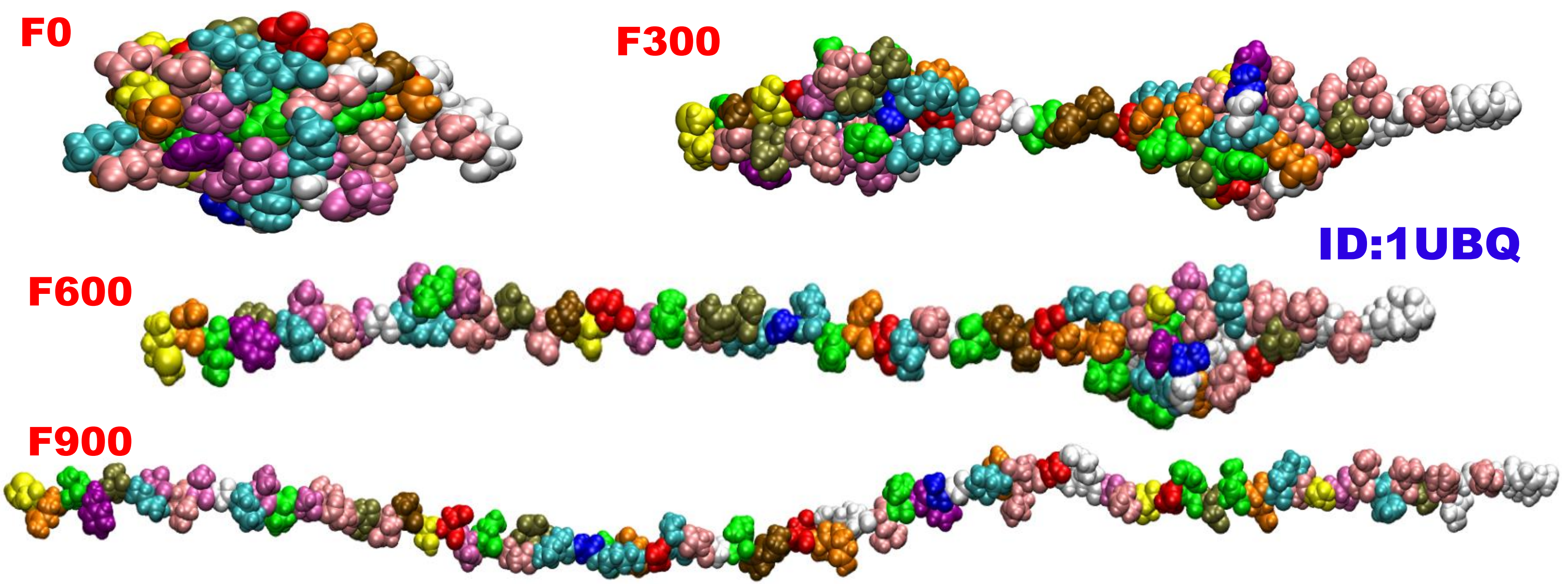


Part 1: Point cloud data based analysis

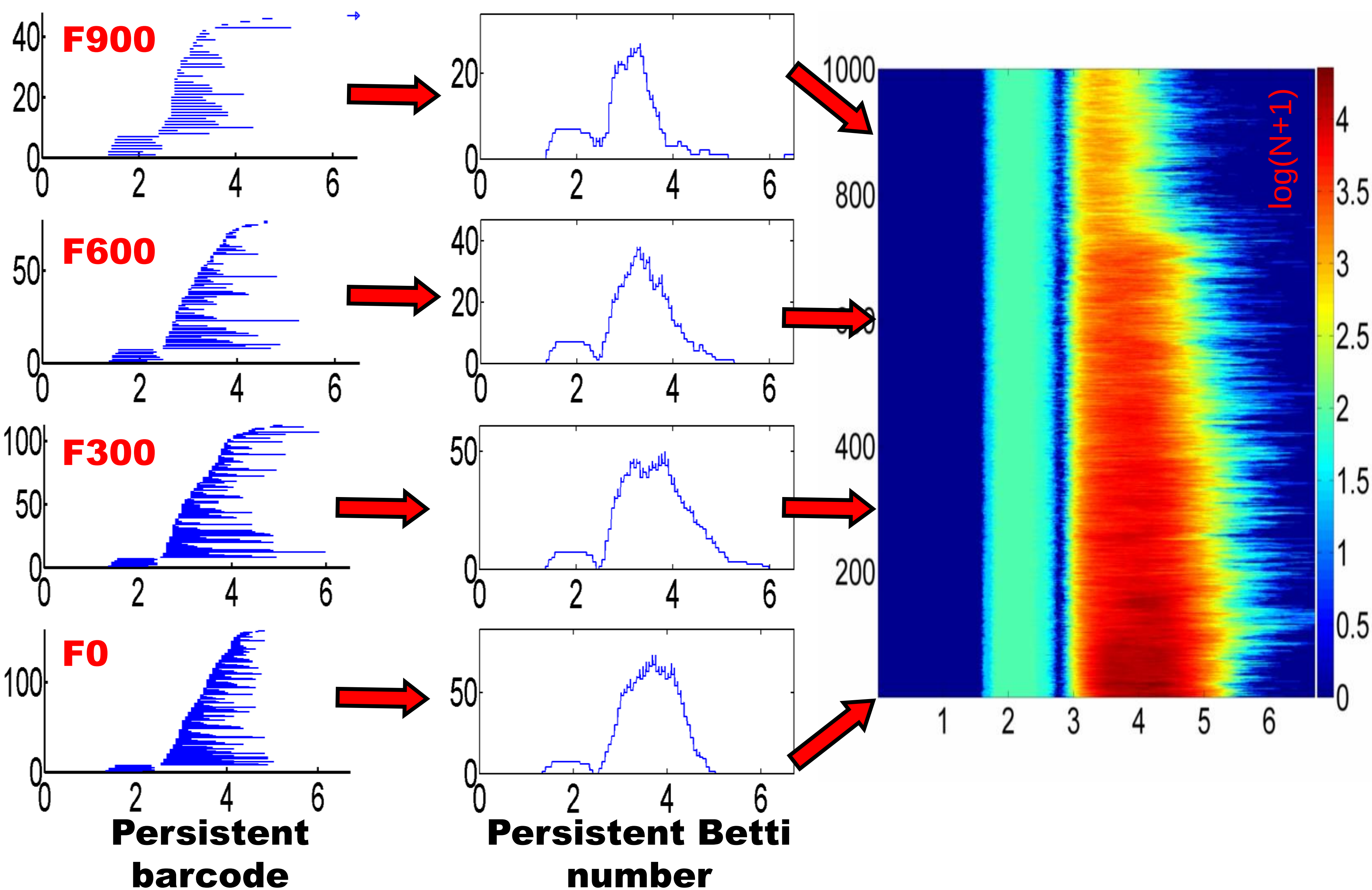
Topic 1--Topological analysis of protein folding

(Xia & Wei, JCC, 2015)

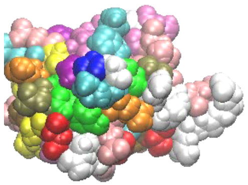




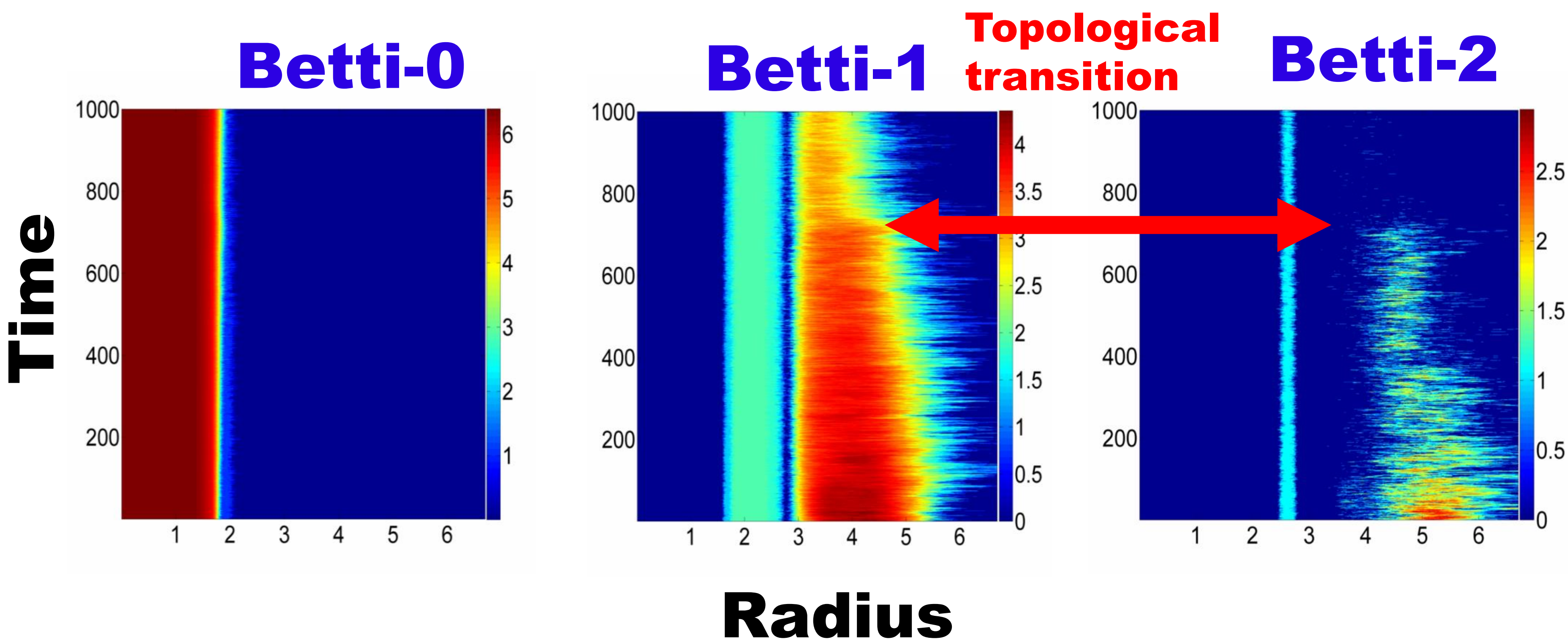
The protein folding process can be characterized by a proposed two dimensional filtration representation



2D persistence in protein unfolding



ID:1UBQ



Topic 2-- 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*

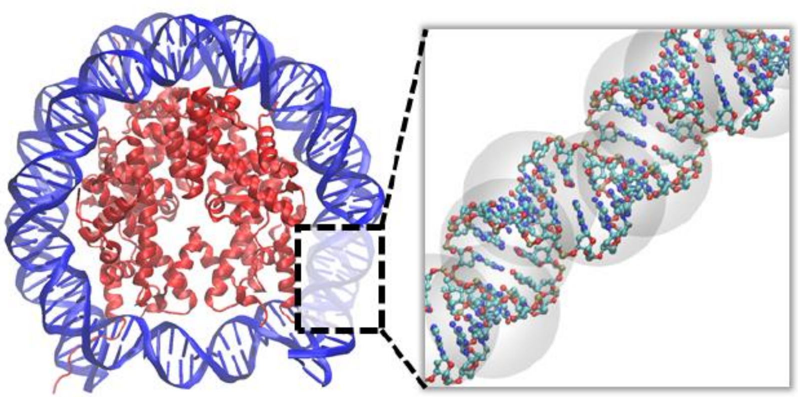
Simplex weight

Boundary operator

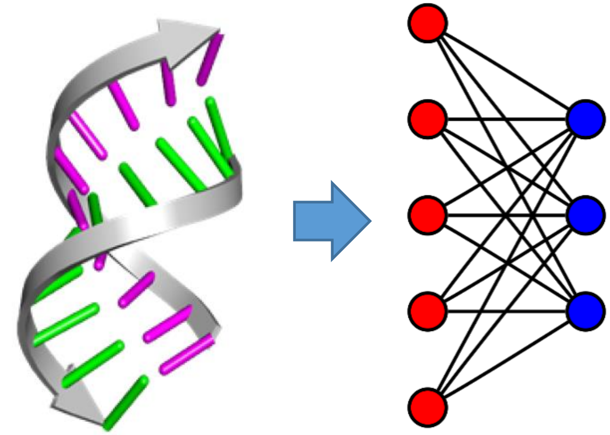
$$\partial_n(\sigma) = \sum_{i=0}^{n} \frac{w(\sigma)}{w(d_i(\sigma))}(-1)^i d_i(\sigma)$$

n-Simplex

Localized Persistent homology (LPH)



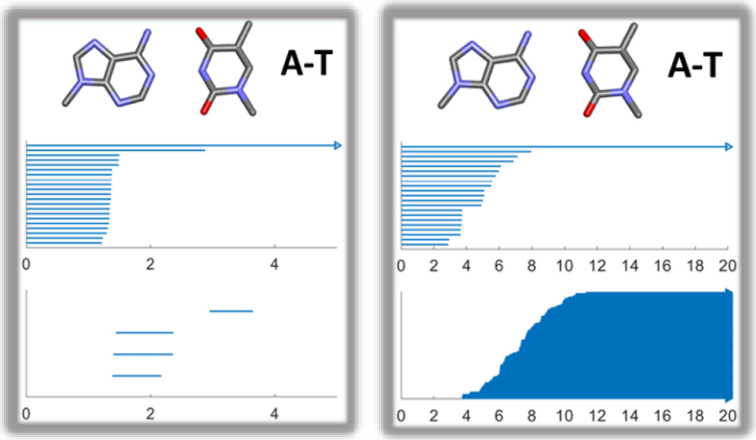
Interactive Persistent homology (IPH)



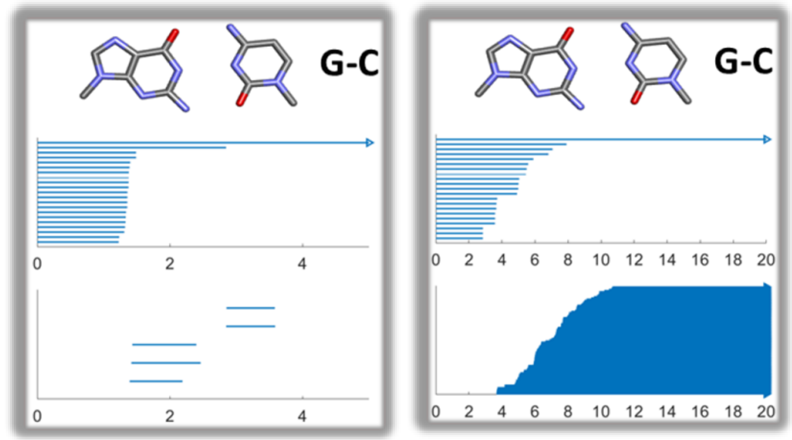
(Cang, Mu, Wei, PLOS Comp. Biol., 2018)

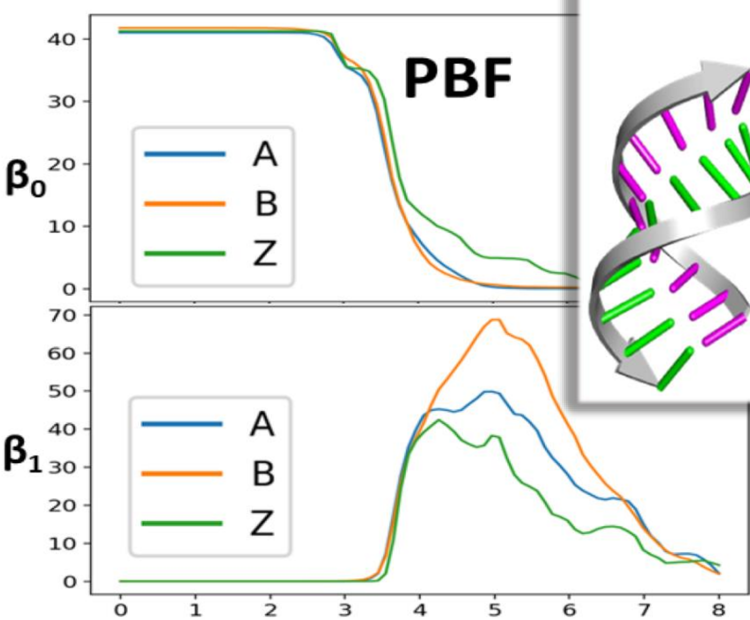
WPH for DNA classification

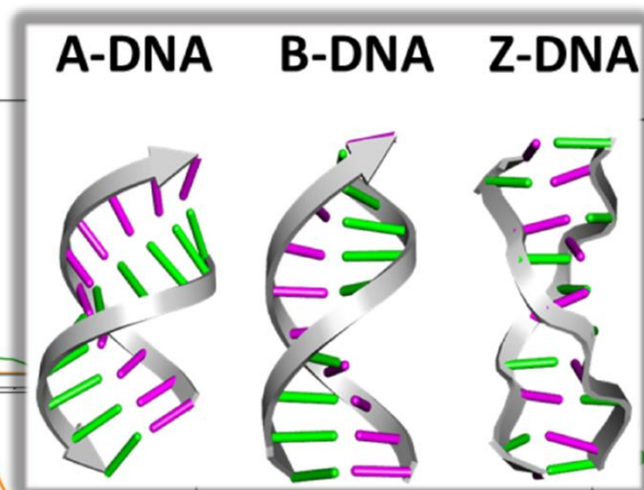
PH *VS* Interactive PH (AT)

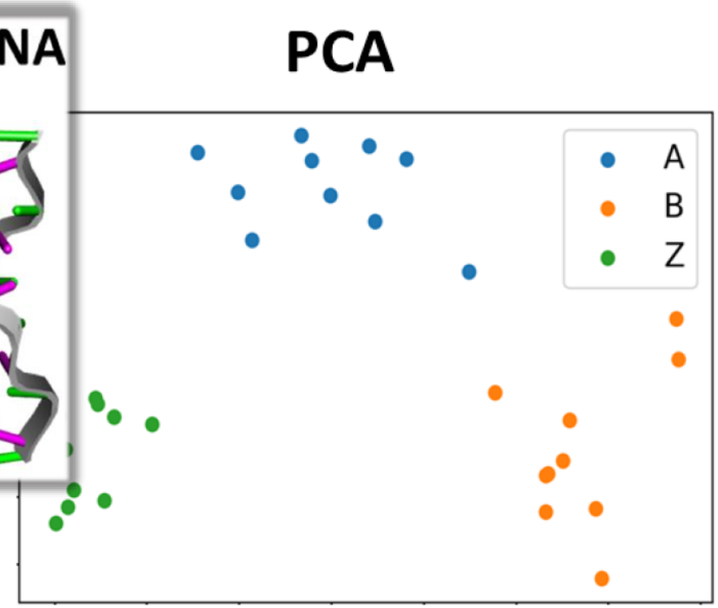


PH *VS* Interactive PH (GC)



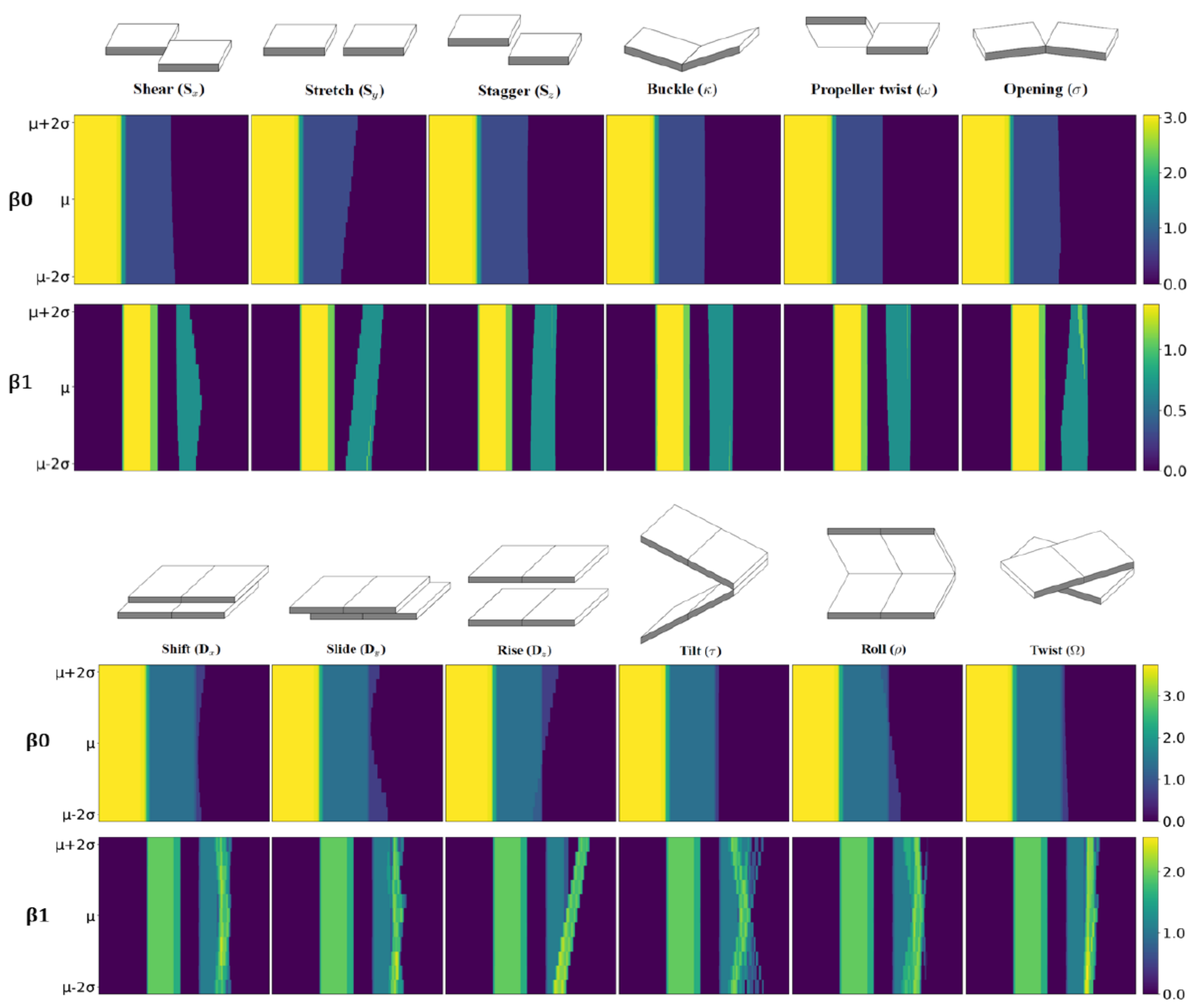


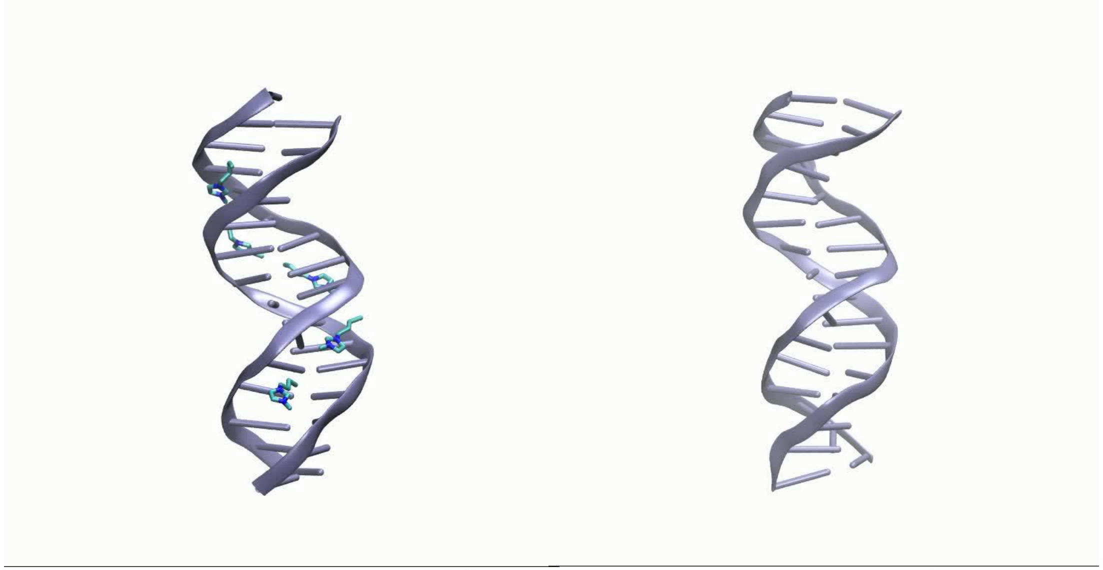




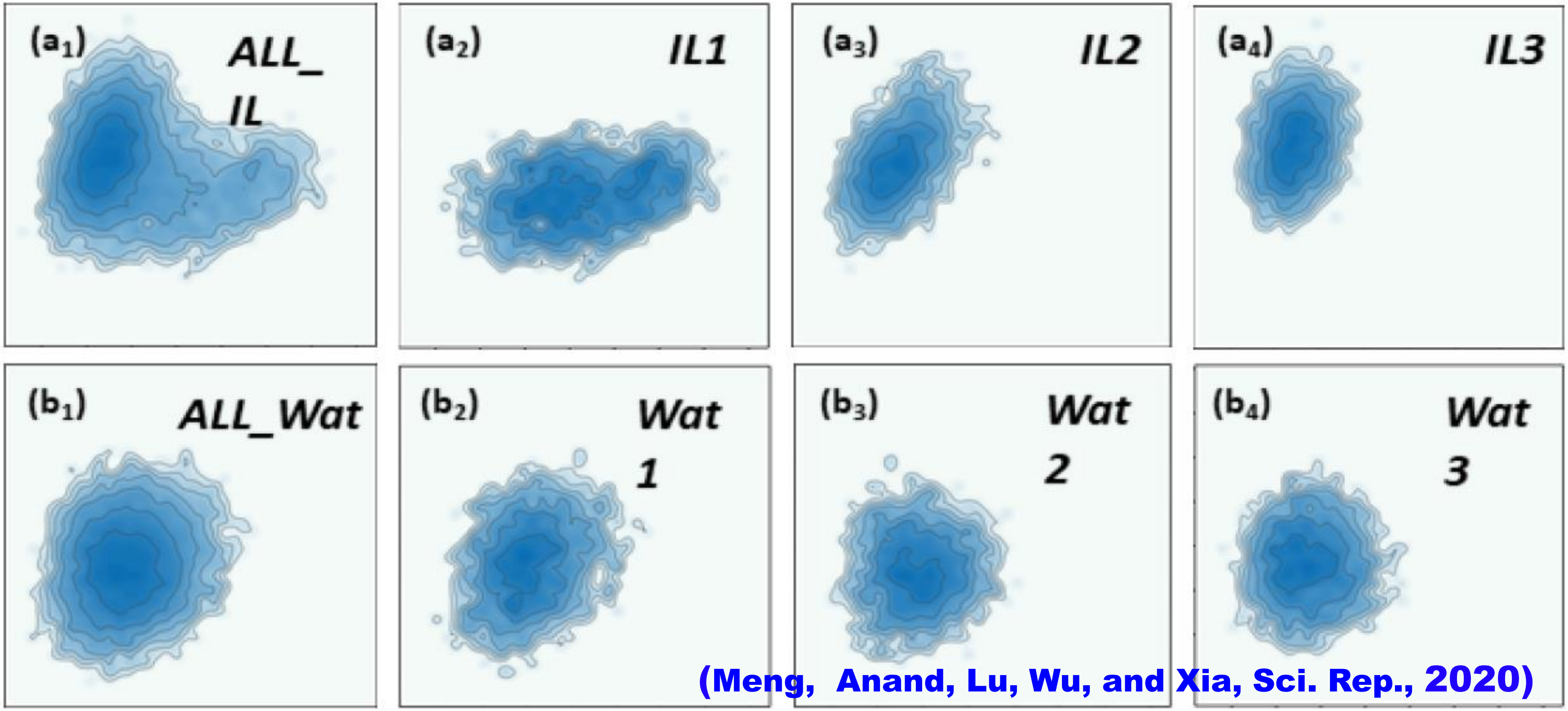
(Meng, Anand, Lu, Wu, and Xia, Sci. Rep., 2020)

Cambridge University Engineering Department Helix computation Scheme (CEHS)

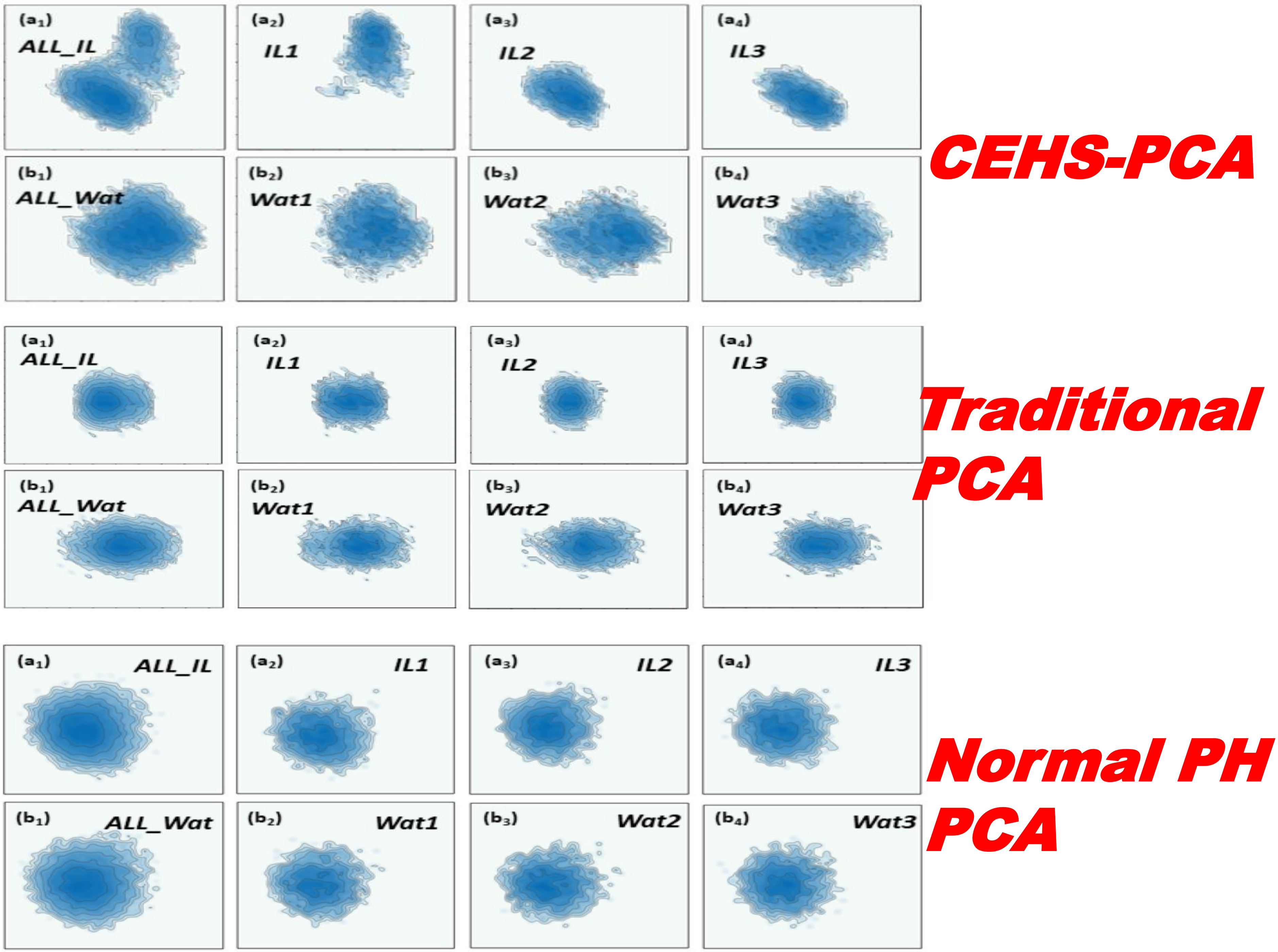




WPH for DNA clustering



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

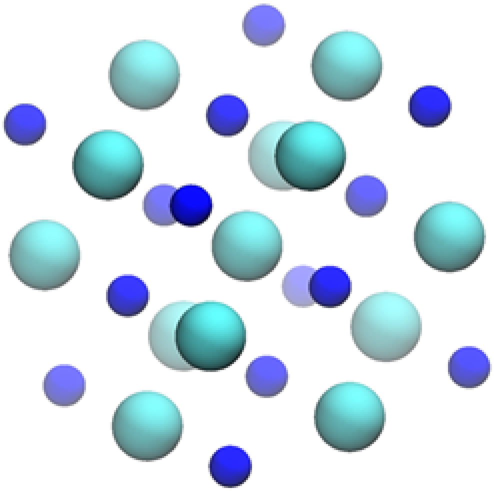


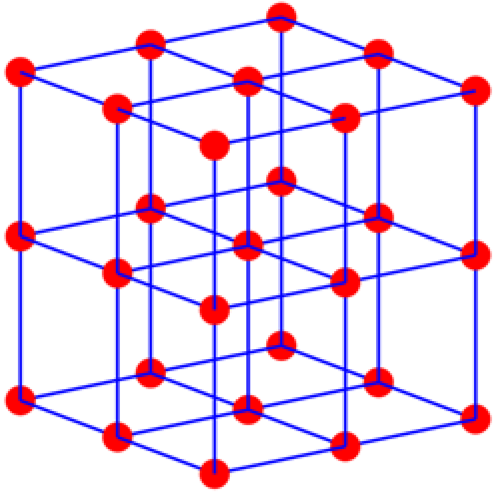
(a1) ALL_IL
(a2) IL1
(a3) IL2
(a4) IL3
(b1) ALL_Wat
(b2) Wat1
(b3) Wat2
(b4) Wat3
CEHS-PCA
(a1) ALL_IL
(a2) IL1
(a3) IL2
(a4) IL3
(b1) ALL_Wat
(b2) Wat1
(b3) Wat2
(b4) Wat3
Traditional PCA
(a1) ALL_IL
(a2) IL1
(a3) IL2
(a4) IL3
(b1) ALL_Wat
(b2) Wat1
(b3) Wat2
(b4) Wat3
Normal PH PCA

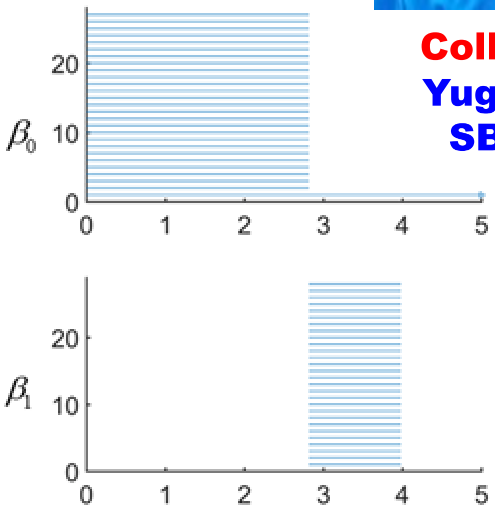
Topic 3--PHA for hydrogen-bonding network

Collaborator
Yuguang Mu
SBS, NTU

Ion crystal structure fingerprint

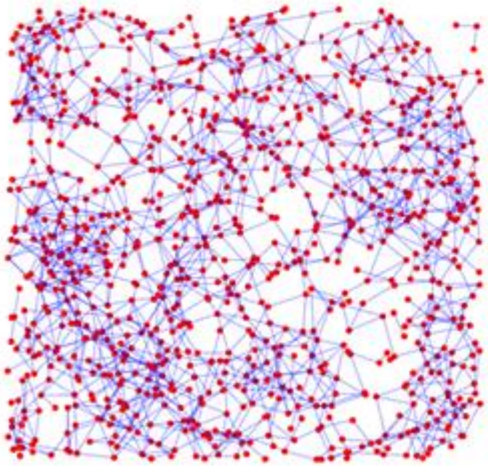


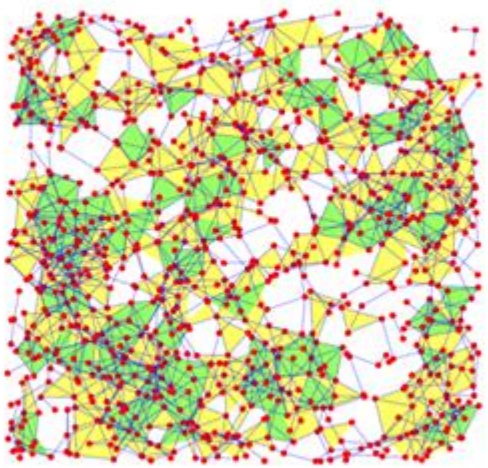


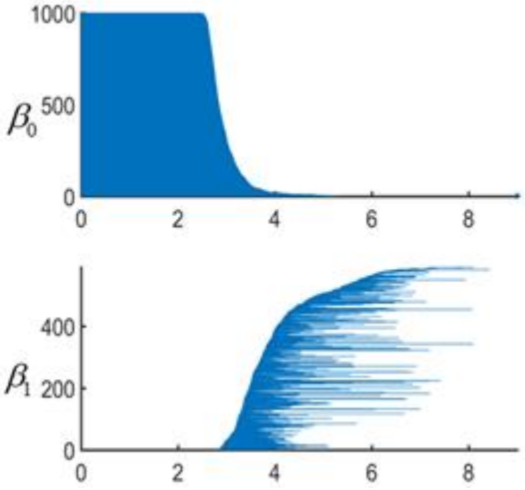


(Xia, PCCP, 2018)

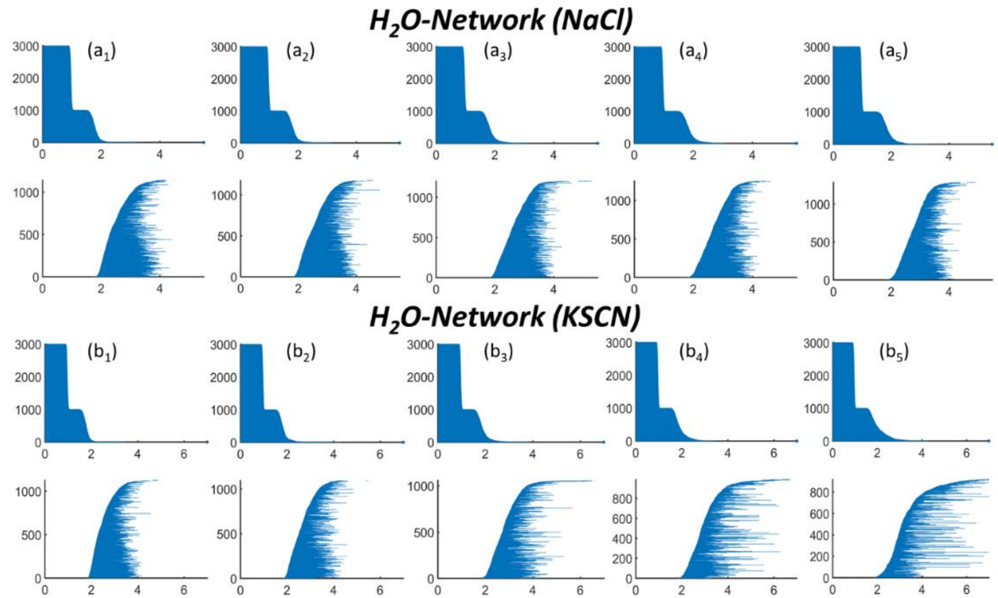
Water hydrogen-bonding fingerprint



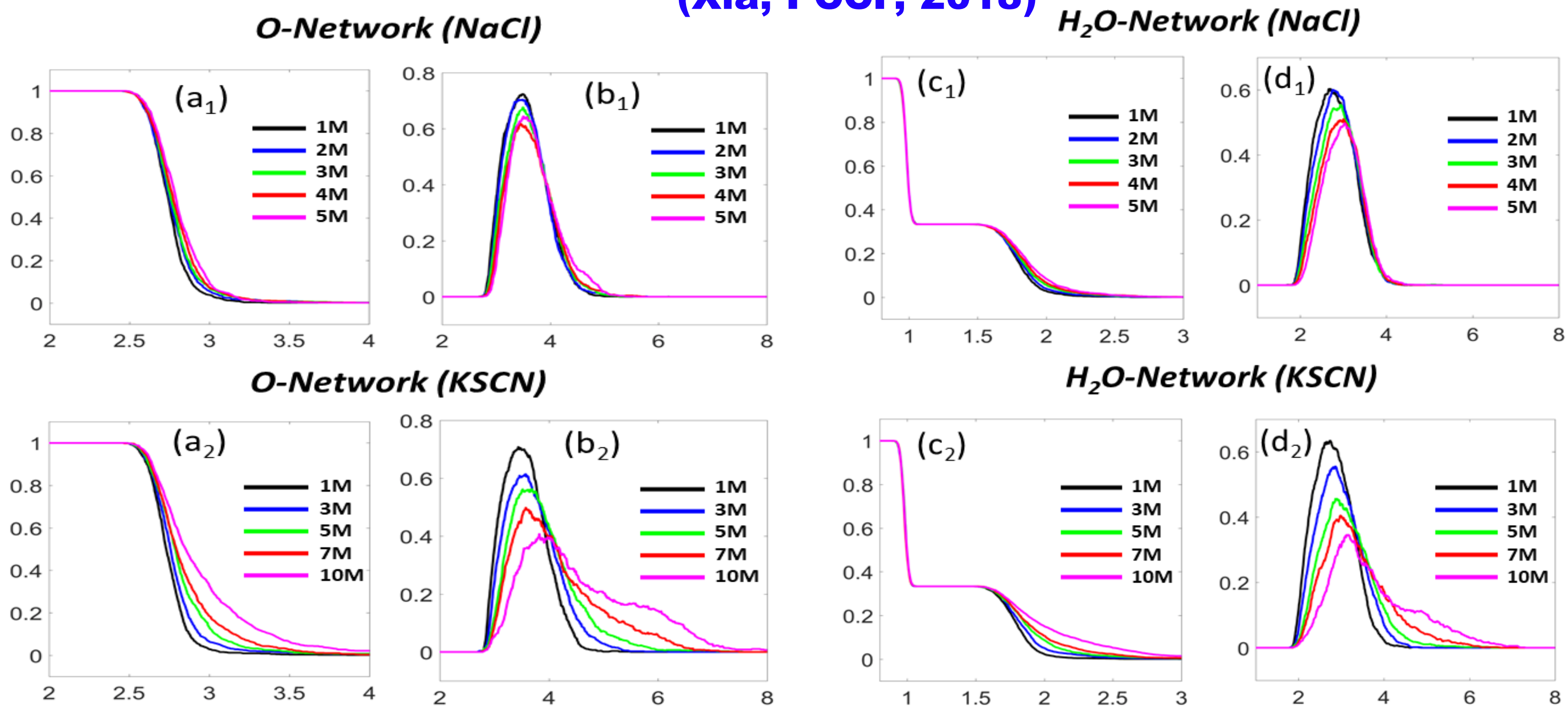




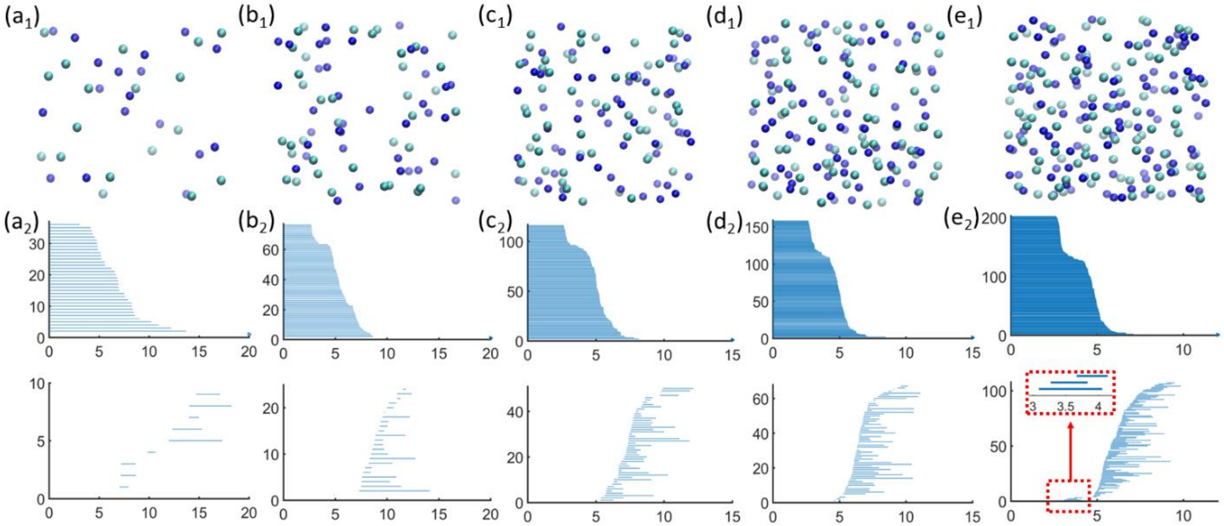
Two types of hydrogen-bonding networks



(Xia, PCCP, 2018)

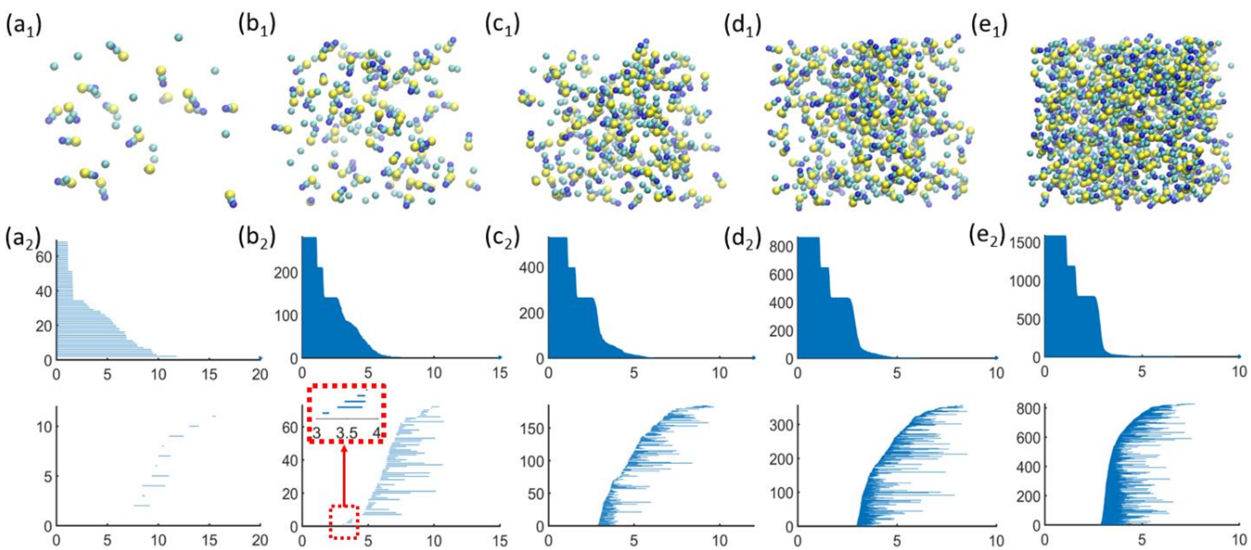


Two morphological types of aggregation



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

Type 1: local clusters

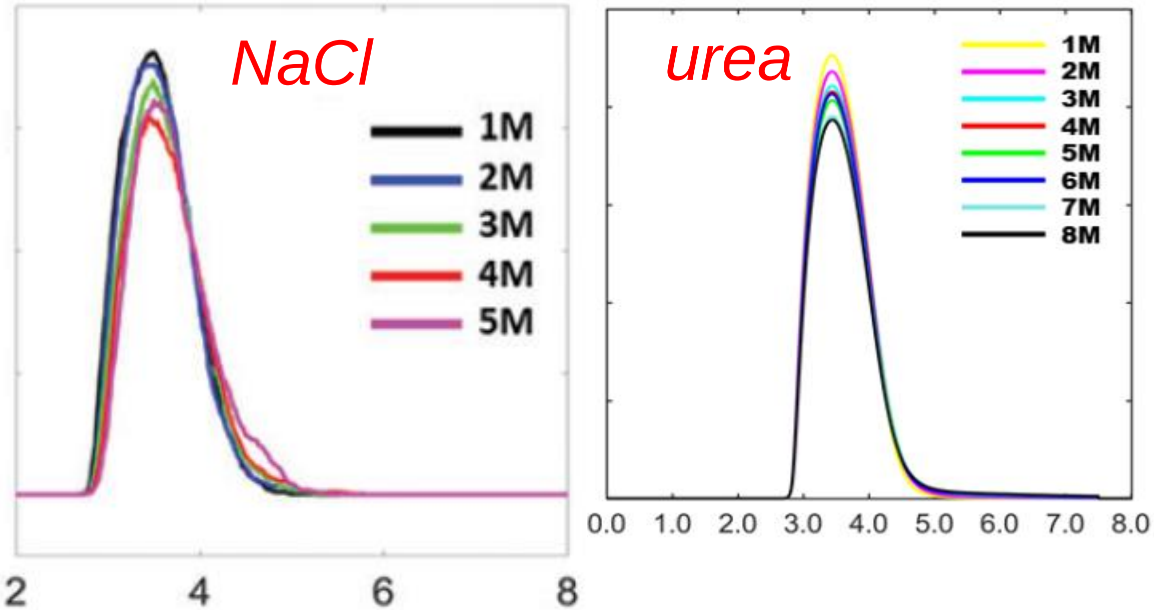


KSCN systems with concentrations: 1M, 3M, 5M, 7M and 10M

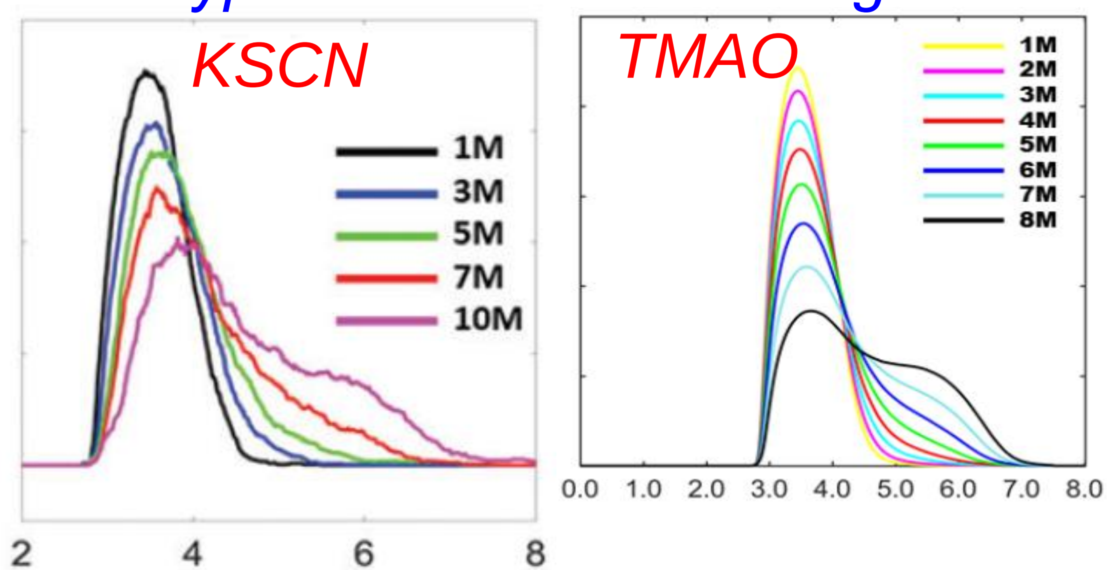
Type 2: extended ion network.

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

Type 1: Structure "breaking"



Type 2: Structure "making"

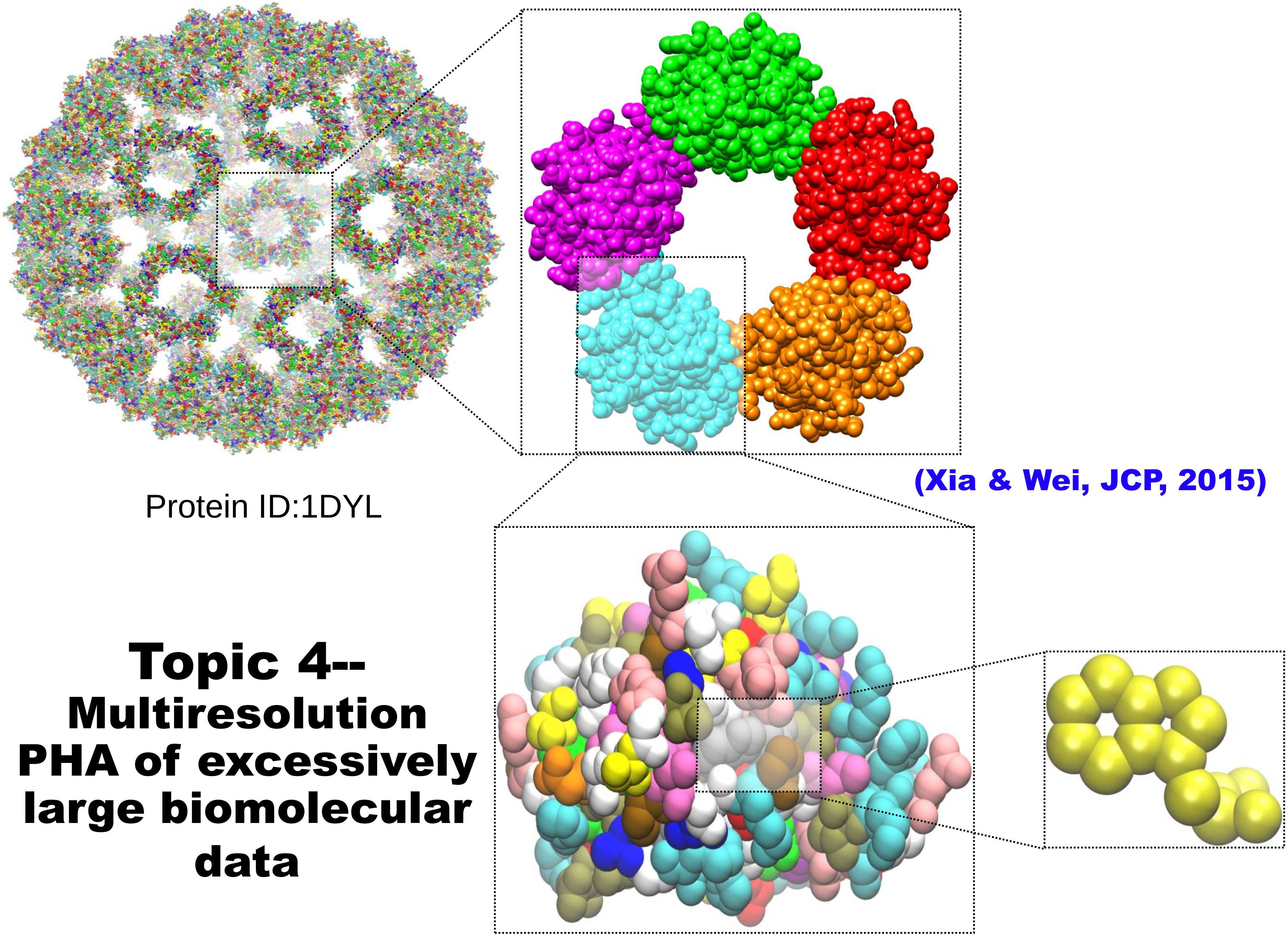


Type 1: denature protein

Type 2: preserve protein structure

(Xia, Anand, Saxena, Mu, PCCP, 2019)

Part 2: density function data based analysis



(Xia & Wei, JCP, 2015)

Topic 4-- Multiresolution PHA of excessively large biomolecular data

Morse theory

Definition 5.9 (Morse function) A smooth map $h : \mathbb{M} \to \mathbb{R}$ is a *Morse function* if all its critical points are nondegenerate.

$$H(p) = \begin{bmatrix} \frac{\partial^2 h}{\partial x^2}(p) & \frac{\partial^2 h}{\partial y \partial x}(p) \\ \frac{\partial^2 h}{\partial x \partial y}(p) & \frac{\partial^2 h}{\partial y^2}(p) \end{bmatrix}$$

Definition 5.11 (minimum, saddle, maximum) A critical point of index 0, 1, or 2, is called a *minimum*, *saddle*, or *maximum*, respectively.

Definition 5.13 (integral line) An *integral line* $\gamma : \mathbb{R} \to \mathbb{M}$ is a maximal path whose tangent vectors agree with the gradient, that is, $\frac{\partial}{\partial s} p(s) = \nabla h(p(s))$ for all $s \in \mathbb{R}$. We call $\operatorname{org} p = \lim_{s \to -\infty} p(s)$ the *origin* and $\operatorname{dest} p = \lim_{s \to +\infty} p(s)$ the *destination* of the path p.

Theorem 5.2 *Integral lines have the following properties:*

(a) *Two integral lines are either disjoint or the same.*
(b) *The integral lines cover all of* $\mathbb{M}$.
(c) *And the limits* $\operatorname{org} p$ *and* $\operatorname{dest} p$ *are critical points of* h.

The properties follow from standard differential calculus.

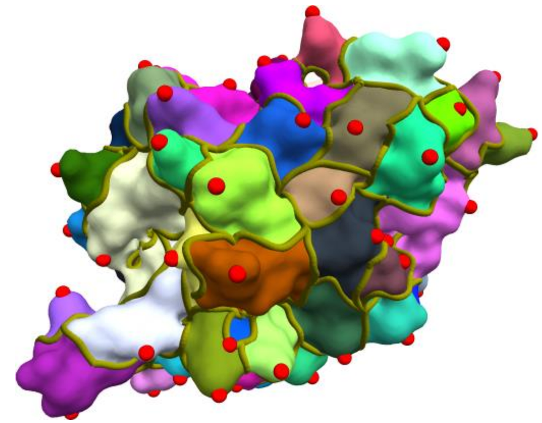
Definition 5.14 (stable and unstable manifolds) The *stable manifold* $S(p)$ and the *unstable manifold* $U(p)$ of a critical point p are defined as

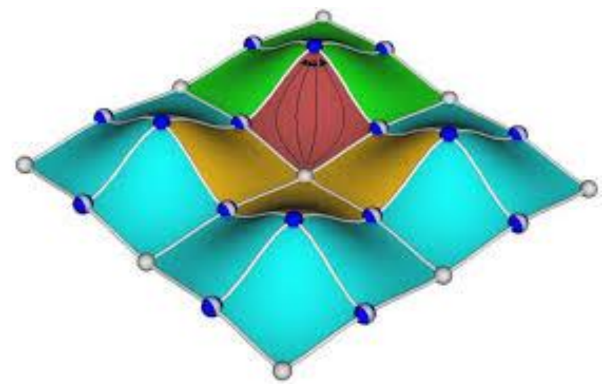
$$S(p) = \{p\} \cup \{y \in \mathbb{M} \mid y \in \operatorname{im}\gamma, \operatorname{dest}\gamma = p\}, \tag{5.4}$$
$$U(p) = \{p\} \cup \{y \in \mathbb{M} \mid y \in \operatorname{im}\gamma, \operatorname{org}\gamma = p\}, \tag{5.5}$$

where $\operatorname{im}\gamma$ is the image of the path $\gamma \in \mathbb{M}$.

Both sets of manifolds decompose $\mathbb{M}$ into open cells.

Definition 5.17 (Morse-Smale complex) Connected components of sets $U(p) \cap S(q)$ for all critical points $p, q \in \mathbb{M}$ are *Morse-Smale cells*. We refer to the cells of dimension 0, 1, and 2 as *vertices*, *arcs*, and *regions*, respectively. The collection of Morse-Smale cells form a complex, the *Morse-Smale complex*.





Topology for Computing

Afra J. Zomorodian

Density representation

Kernel function:

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

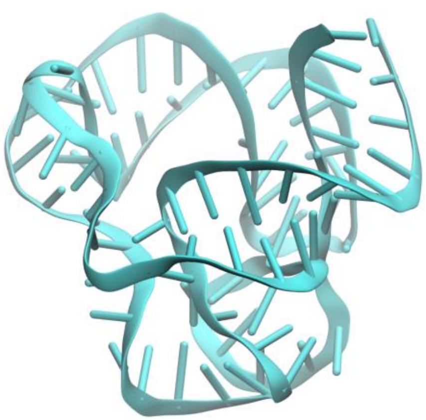
$$\phi(\| r - r_j \|; \eta) = 0, \quad as \quad \| r - r_j \| \to \infty$$

For example:

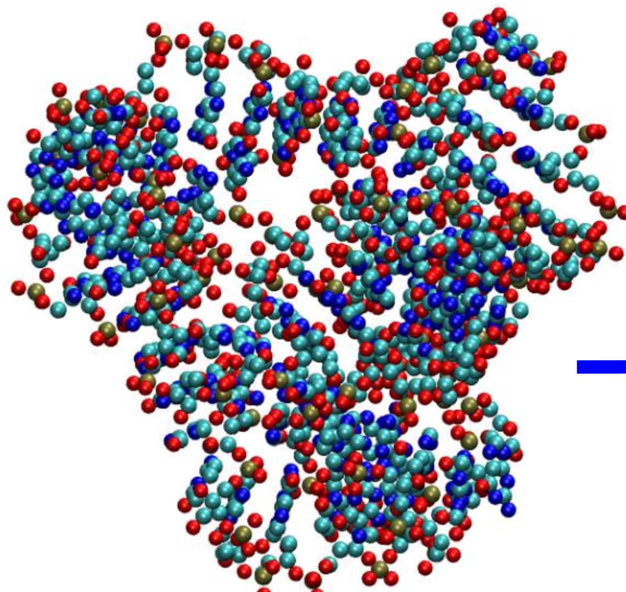
$$\phi(\| r - r_j \|; \eta) = e^{-(\| r - r_j \| / \eta)^{\kappa}}, \kappa > 0$$

Rigidity function:

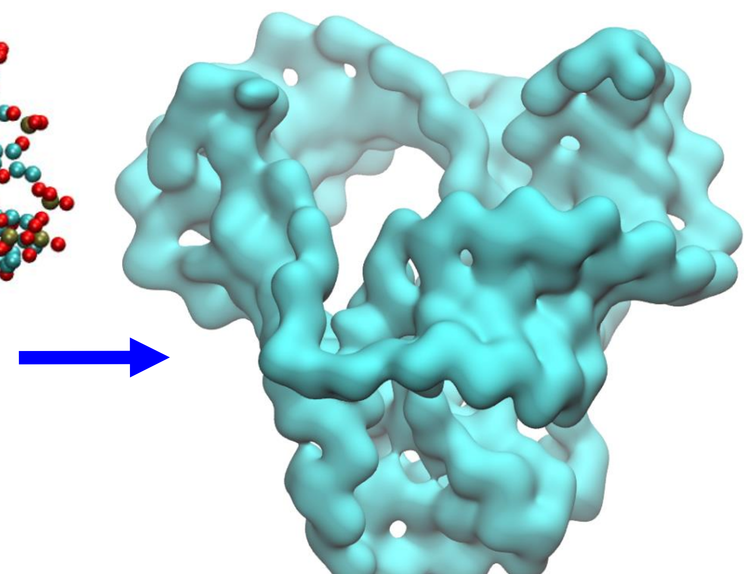
$$\boxed{\mu(r) = \sum_{j}^{N} w_j \phi(\| r - r_j \|; \eta)}$$



RNA: 4QG3



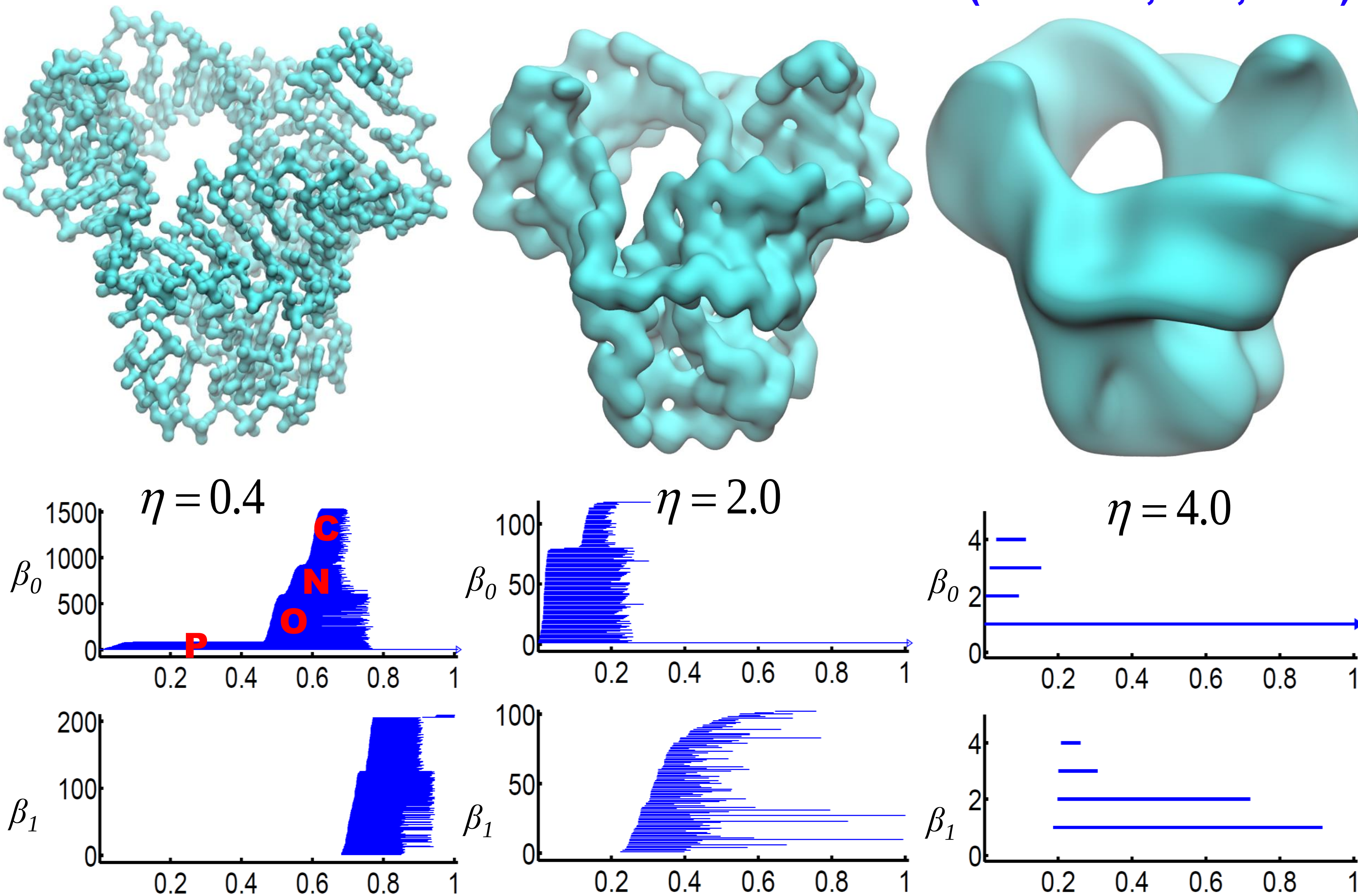
Point cloud representation



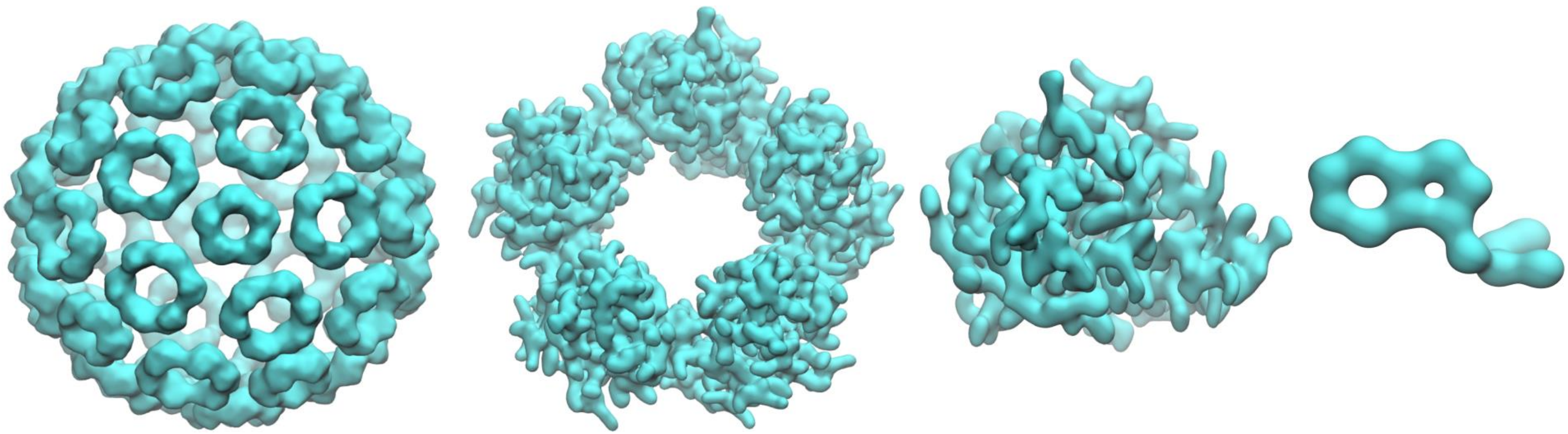
Isosurface extracted from rigidity function

PHA for multiresolution representations

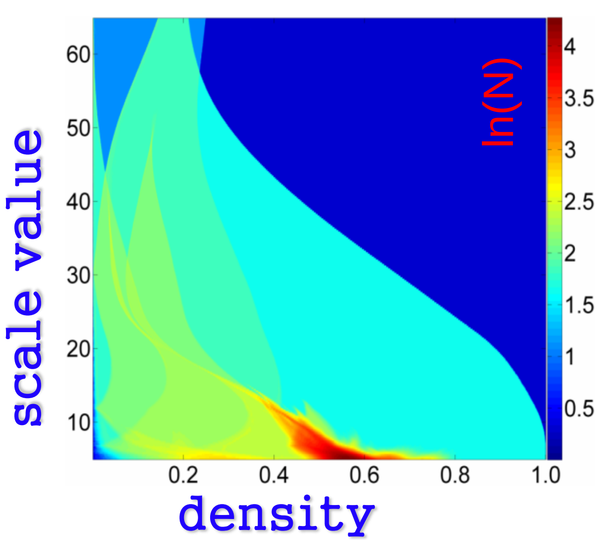
(Xia & Wei, JCB, 2015)



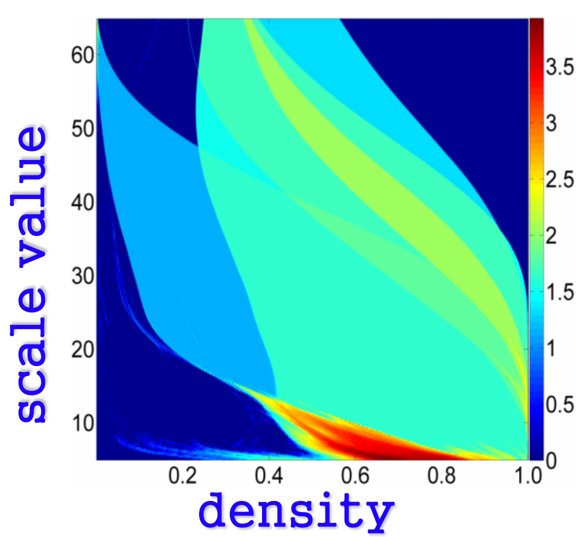
Multiresolution of the virus capsid



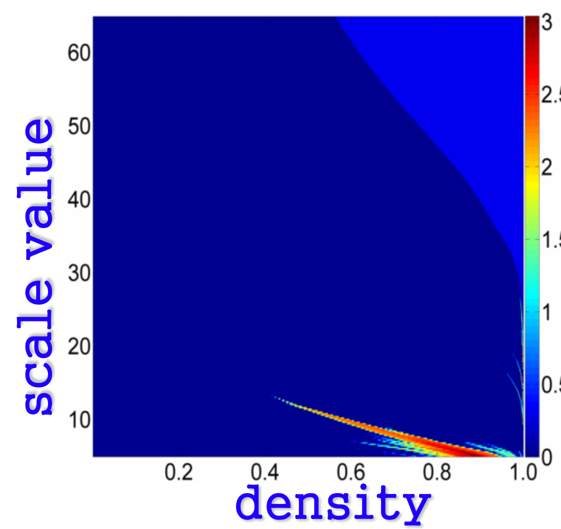
Betti-0



Betti-1

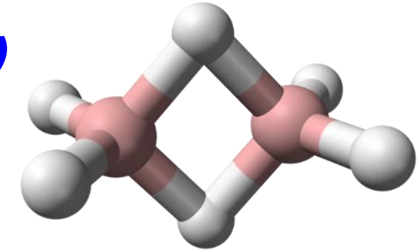


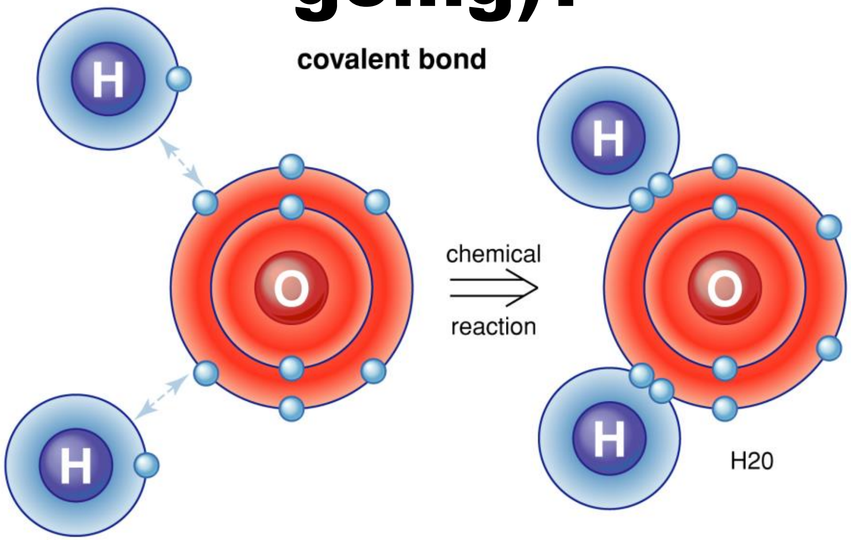
Betti-2

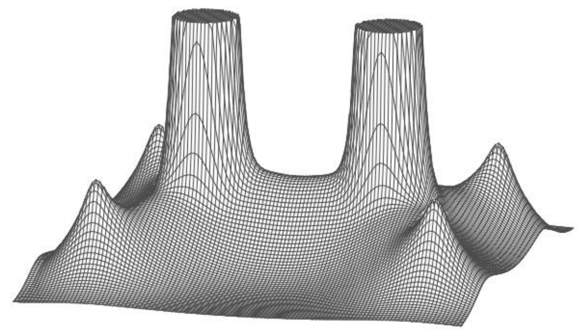


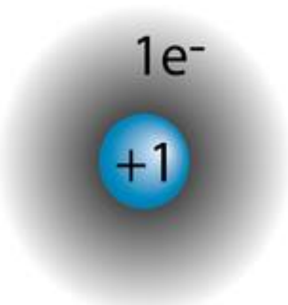
Topic 5--What is chemical bond (on-going)?

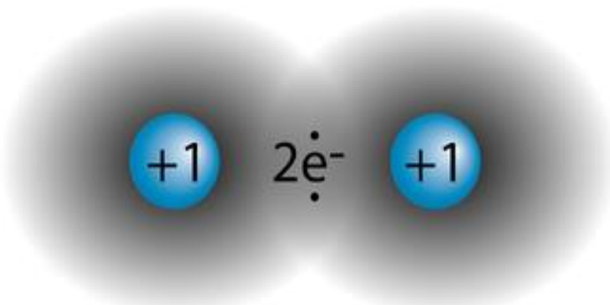
Diborane (B2H6)

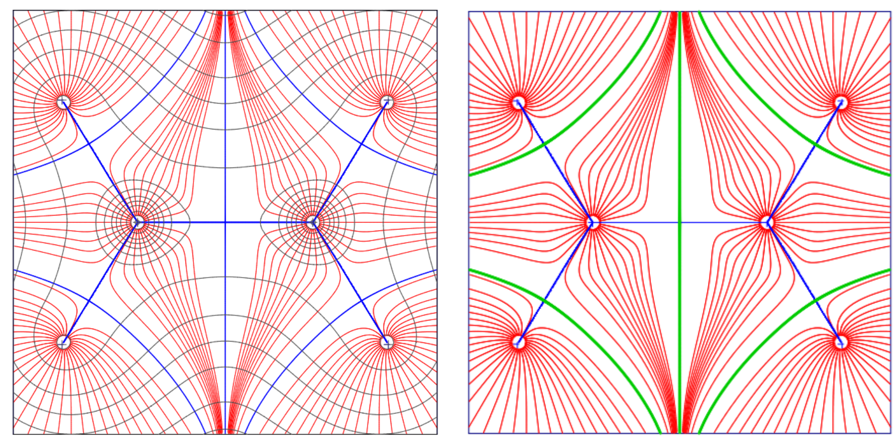












·H

H:H

H — H

Atoms in Molecules

Richard F. W. Bader

McMaster University, Hamilton, Ontario, Canada

Classification of the Critical Points

Atoms in molecule

CP's are labeled: (rank, signature)

Only 4 possible *signature* values for critical points of *rank = 3* :

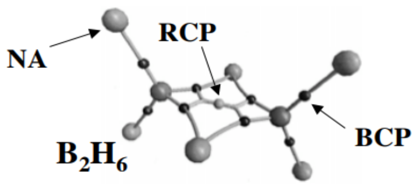
(3, -3) : Nuclear Attractor (NA)

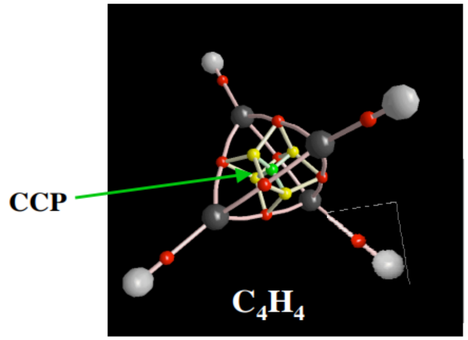
(3, -1) : Bond Critical Point (BCP)

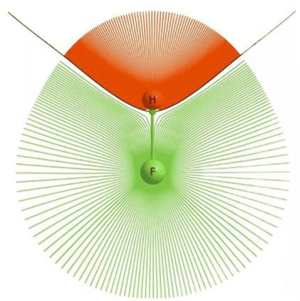
(3, +1) : Ring Critical Point (RCP)

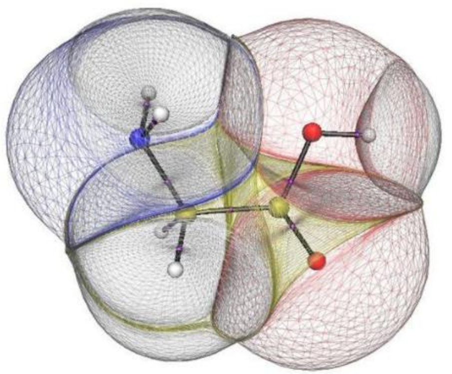
(3, +3) : Cage Critical Point (CCP)

$$A(r_c) = \begin{pmatrix} \frac{\partial^2 \rho}{\partial x^2} & \frac{\partial^2 \rho}{\partial x \partial y} & \frac{\partial^2 \rho}{\partial x \partial z} \\ \frac{\partial^2 \rho}{\partial y \partial x} & \frac{\partial^2 \rho}{\partial y^2} & \frac{\partial^2 \rho}{\partial y \partial z} \\ \frac{\partial^2 \rho}{\partial z \partial x} & \frac{\partial^2 \rho}{\partial z \partial y} & \frac{\partial^2 \rho}{\partial z^2} \end{pmatrix}_{r = r_c} \qquad \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$









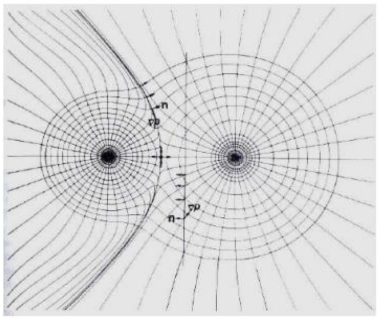
Zero-flux Surfaces

- An atom can be defined as a region of real space bounded by surfaces through which there is *no flux in the gradient vector field of* ρ, meaning that the surface is not crossed by any trajectories of $\nabla\rho(r_s)$
- An *interatomic surface* (IAS) satisfies the "zero-flux" *boundary condition*:

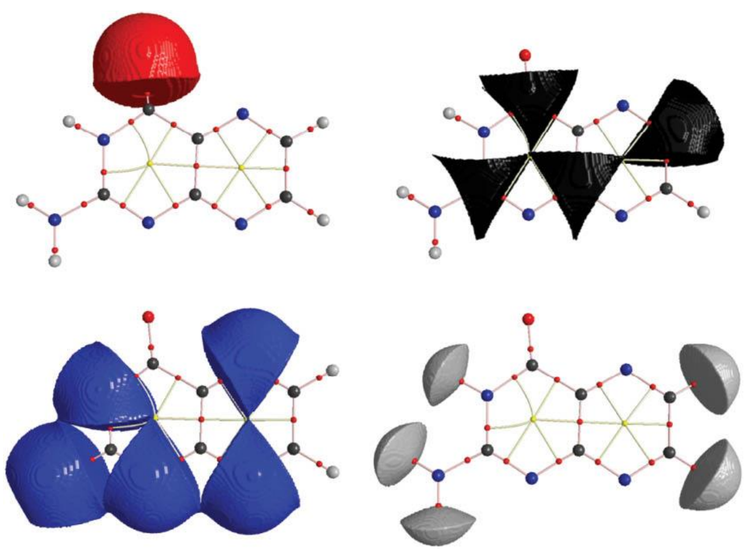
$$\nabla\rho(r_s) \bullet n(r_s) = 0 \qquad \text{for every point } r_s \text{ on the surface } S(r_s)$$

where $n(r_s)$ is the unit vector normal to the surface at r_s

- At a point on a dividing surface the gradient of the electron density has *no component* normal to the surface.



Contour map of NaCl overlaid with trajectories of $\nabla\rho$



R.F.W. Bader, *Atoms in Molecules - A quantum theory*, Oxford University Press, New York, 1990.

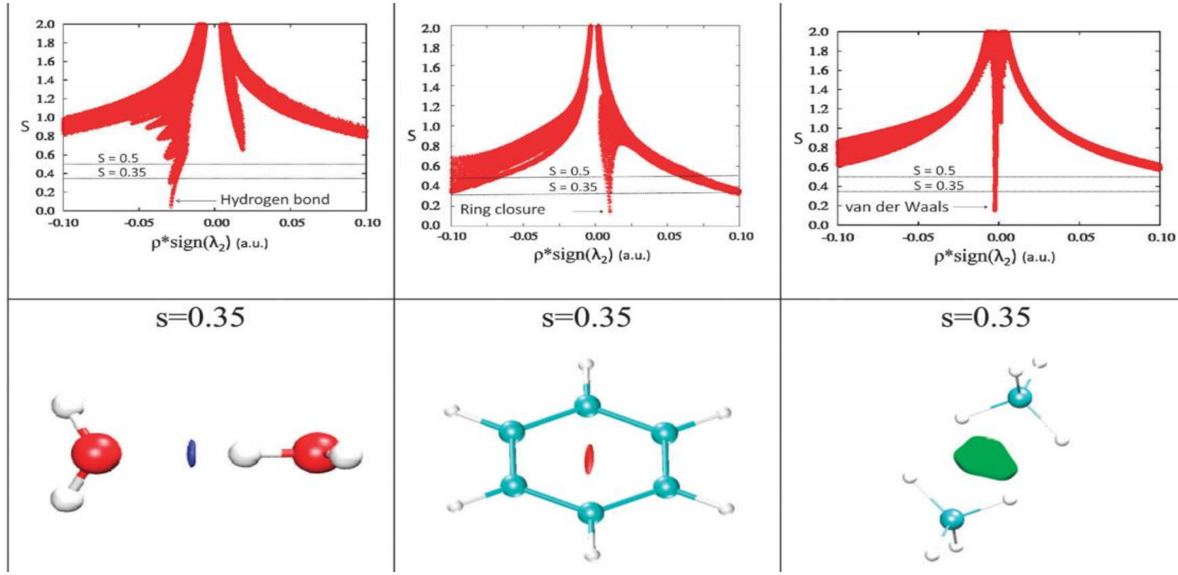
Non-covalent bond situation

$$\tilde{\rho}(\mathbf{r}) = \mathrm{sign}(\gamma_2(\mathbf{r}))\rho(\mathbf{r})$$

$$s(\mathbf{r}) = \frac{1}{2(3\pi^2)^{\frac{1}{3}}}\frac{|\nabla\rho(\mathbf{r})|}{\rho(\mathbf{r})^{\frac{4}{3}}}$$

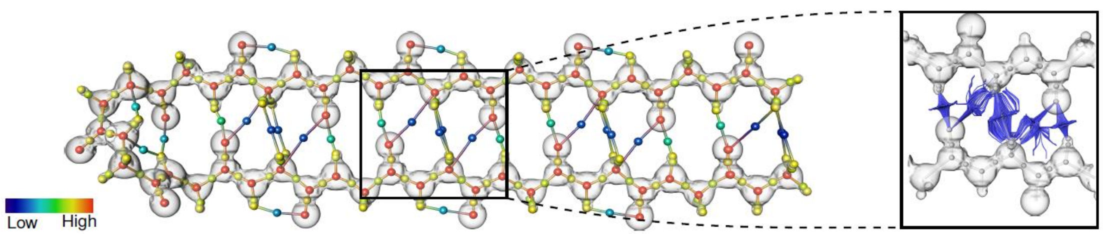
Revealing Noncovalent Interactions

Erin R. Johnson,[§] Shahar Keinan,[§] Paula Mori-Sánchez,[§] Julia Contreras-García,[§] Aron J. Cohen,[#] and Weitao Yang*[,§]



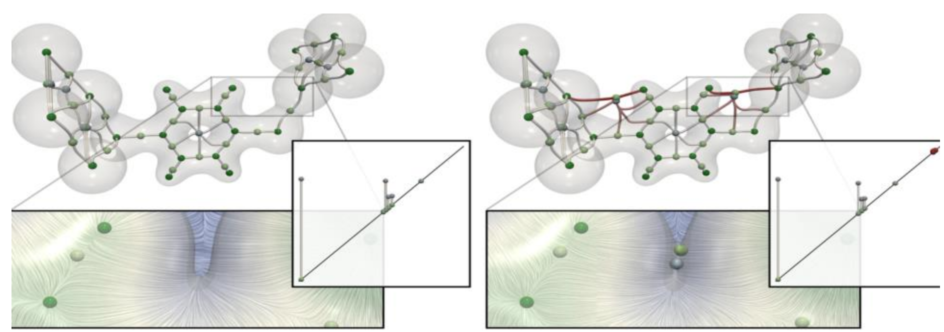
Characterizing Molecular Interactions in Chemical Systems

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

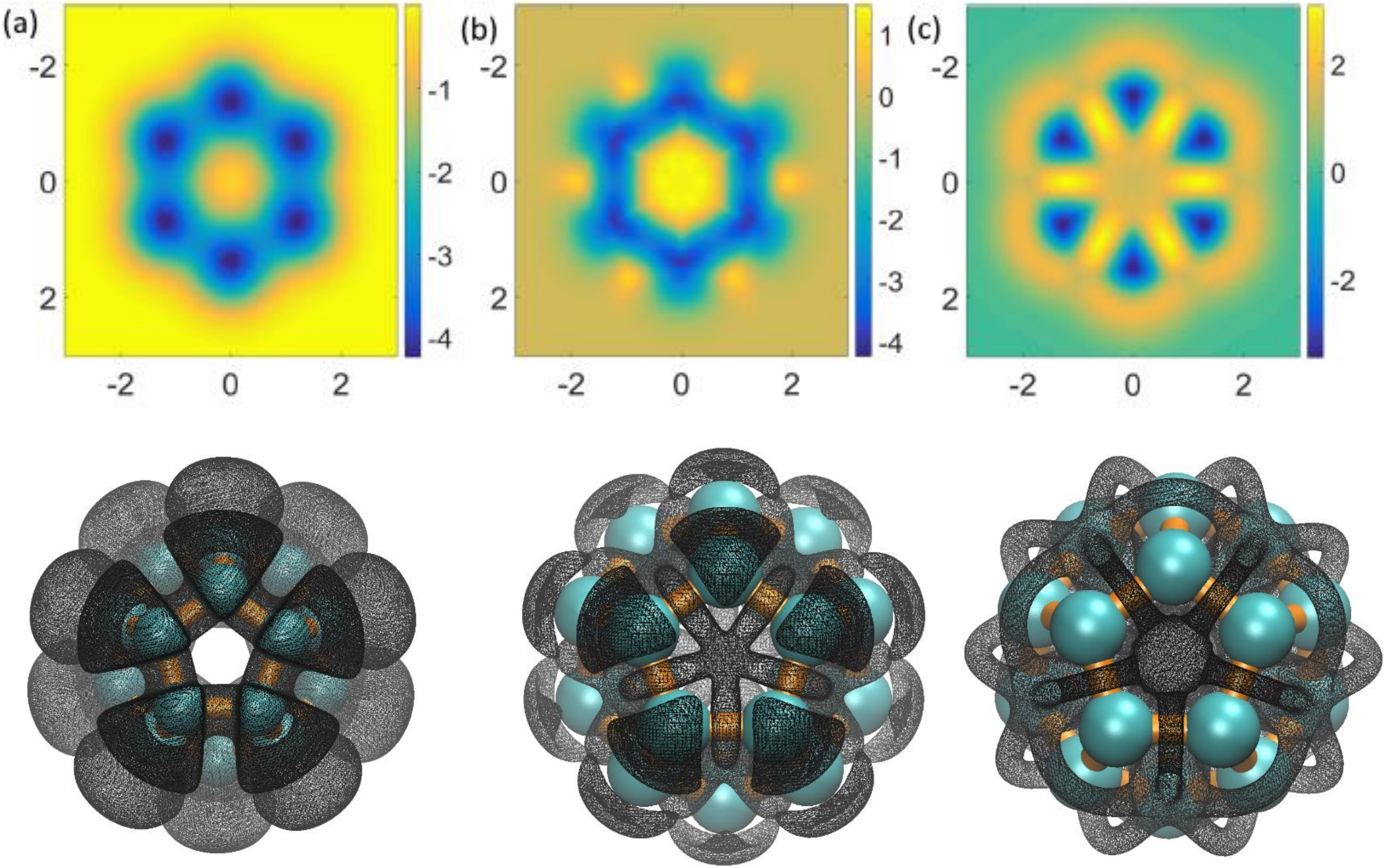


A Topological Data Analysis perspective on noncovalent interactions in relativistic calculations

Małgorzata Olejniczak[1] | André Severo Pereira Gomes[2] | Julien Tierny[3]

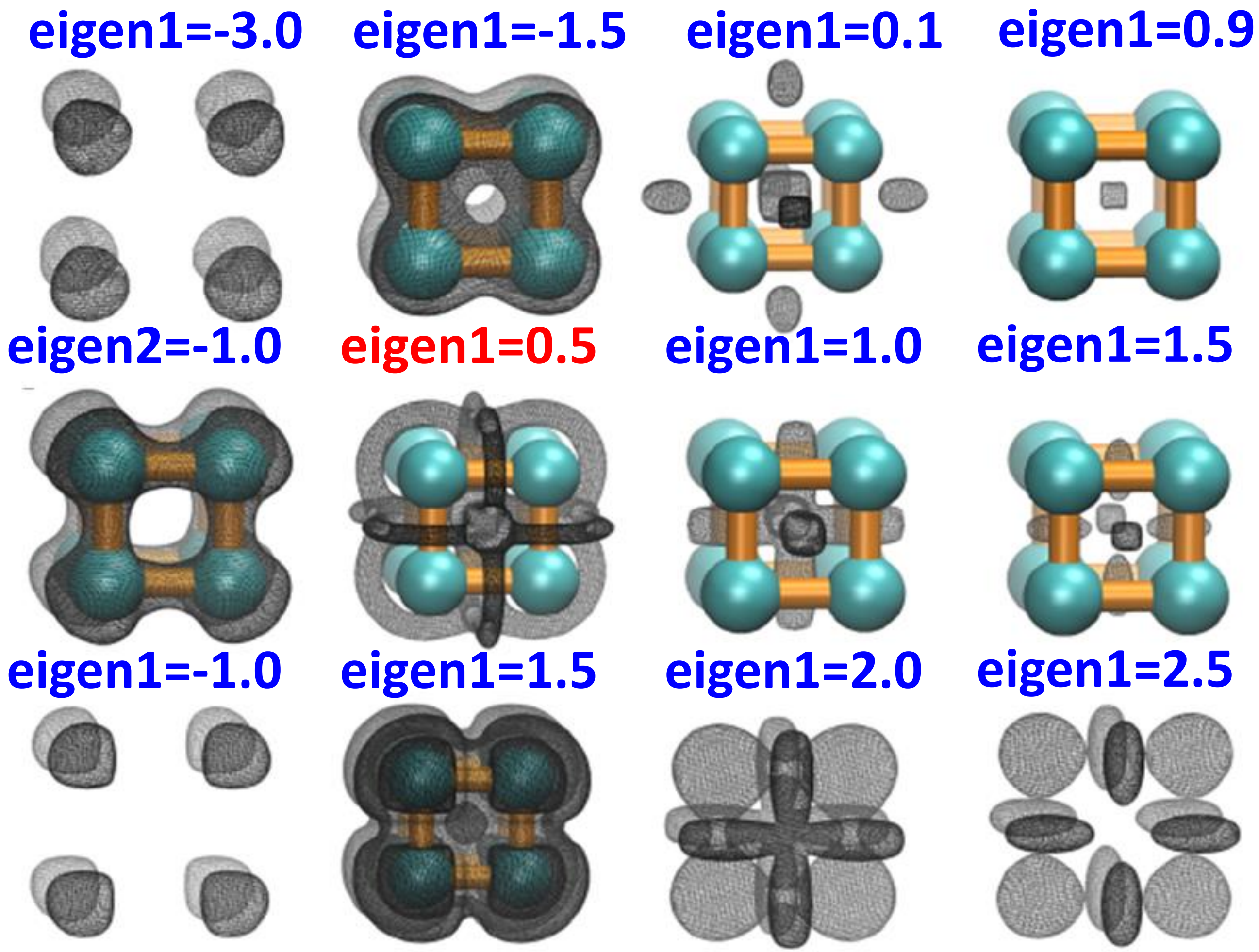


Isosurfaces of eigenvalues



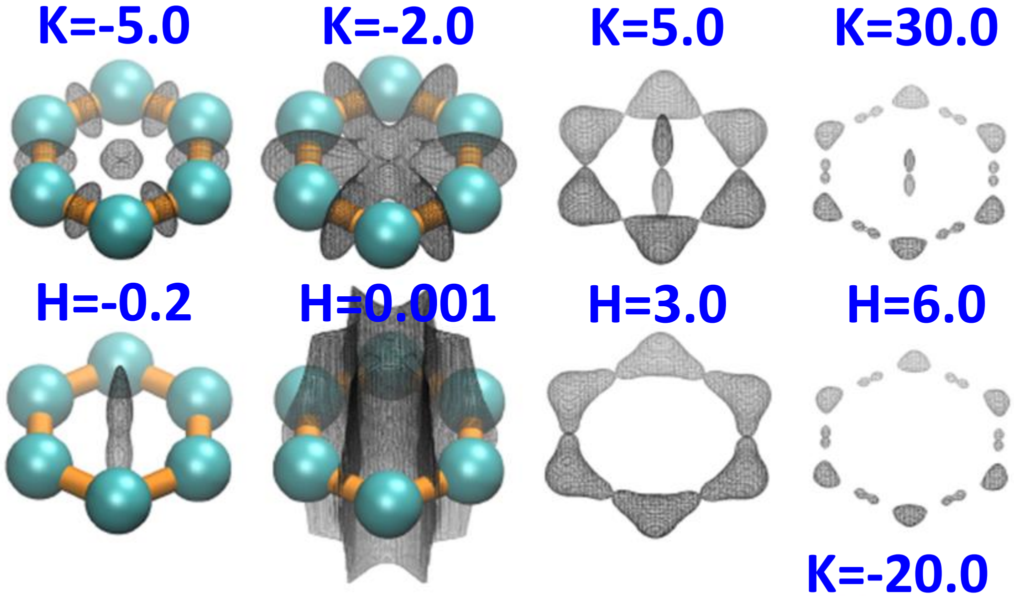
(Xia & Wei, arXiv, 2017)

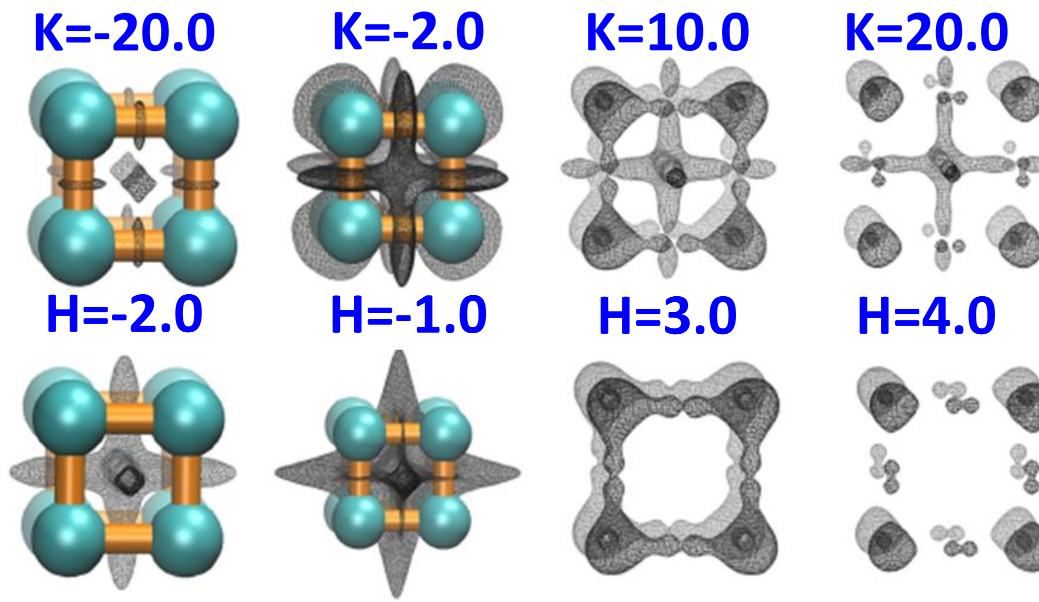
Isosurfaces of eigenvalues



	NCP	BCP	RCP	CCP
γ_1	N	N	P	P
γ_2	N	N; P-Loop	P	P
γ_3	N	P	P	P

Isosurfaces of Gaussian and mean curvature



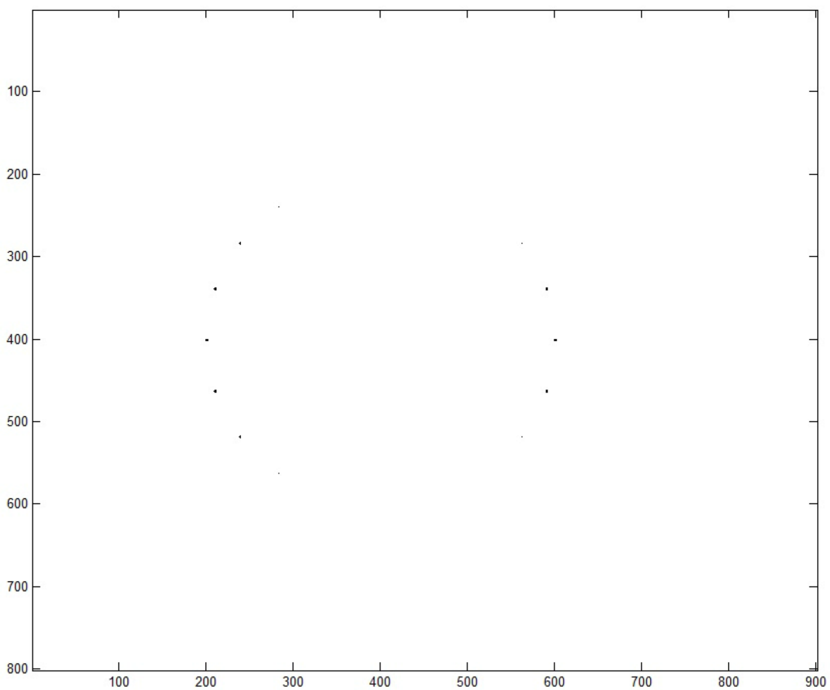


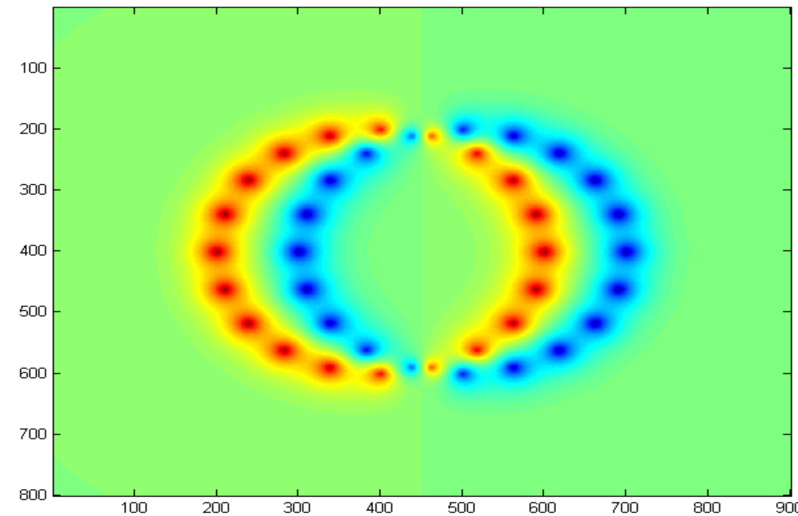
A-type and V-type means the isosurface is along with or vertical to atomic bonds or ring planes

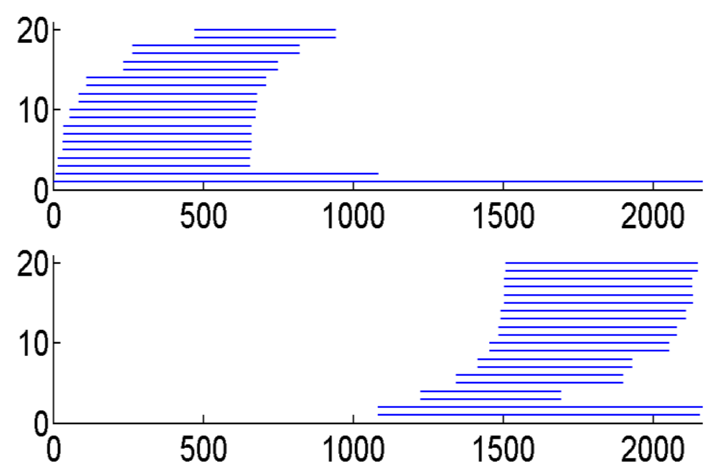
	NCP	BCP	RCP	CCP
K	P	A-type (P); V-type (N)	V-type (P); A-type (N)	P
H	P	A-type (P); V-type(N)	N	N
κ_1	P	P	A-type (P); V-type (N)	N
κ_2	P	A-type (P); V-type (N)	N	N

(Xia & Wei, arXiv, 2017)

Topic 6--Topological properties for electrostatic potential functions



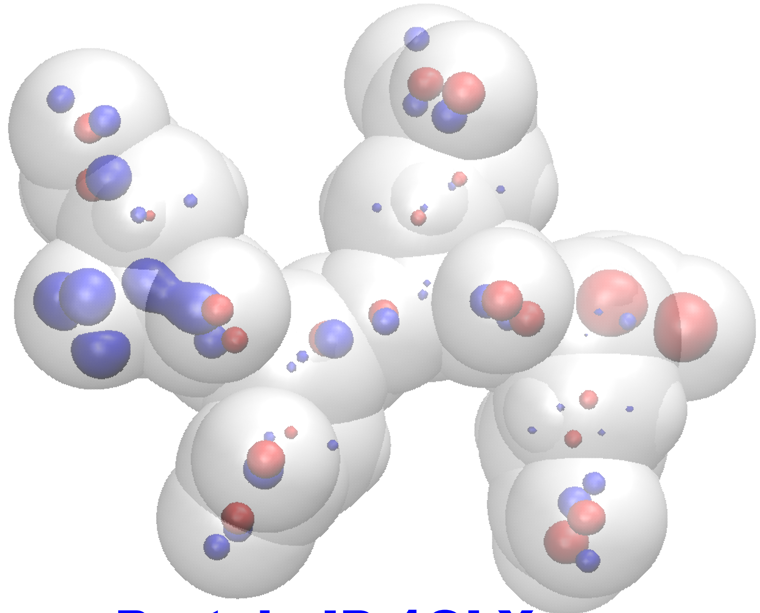




$$potential = \sum_i e^{-\left(\frac{|r-r_i|}{\sigma}\right)^2} - \sum_i e^{-\left(\frac{|r-r_i|}{\sigma}\right)^2}$$

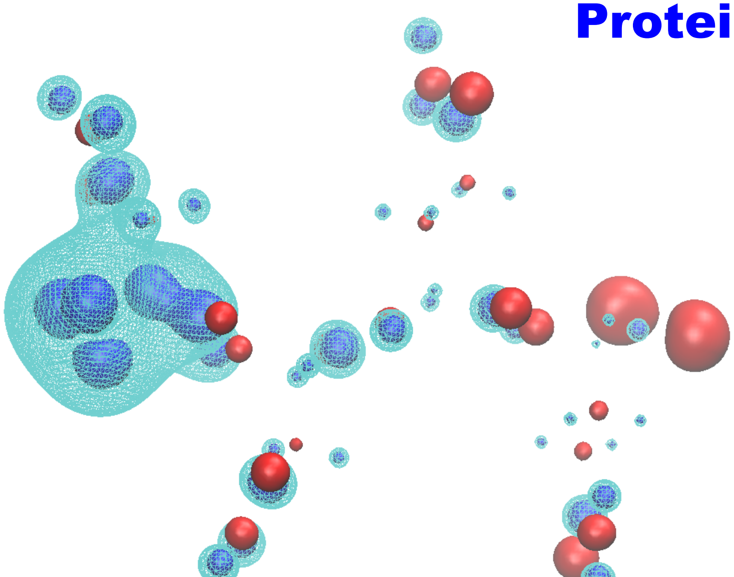
Protein electrostatic potential

$$\Phi = \sum_j \frac{q_j}{\varepsilon_0 \, | r - r_j |}$$

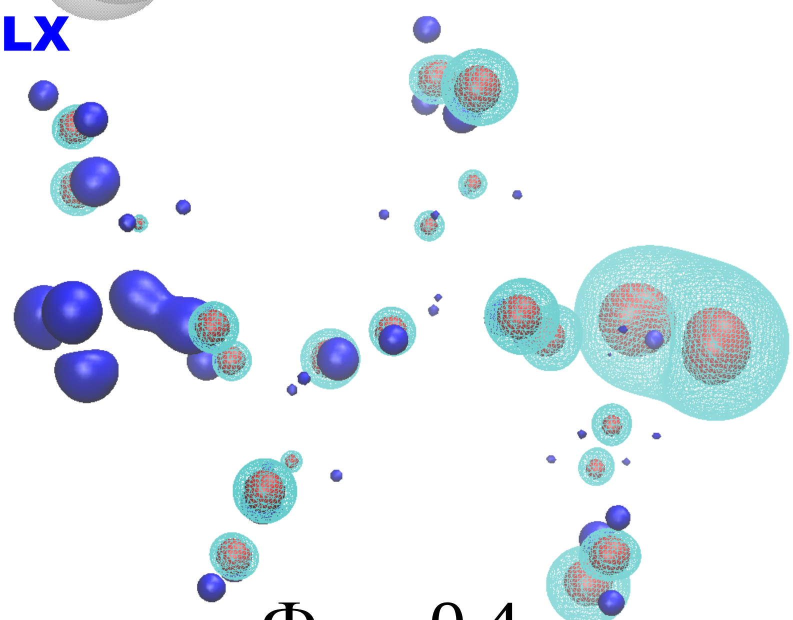


Protein ID:1OLX

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

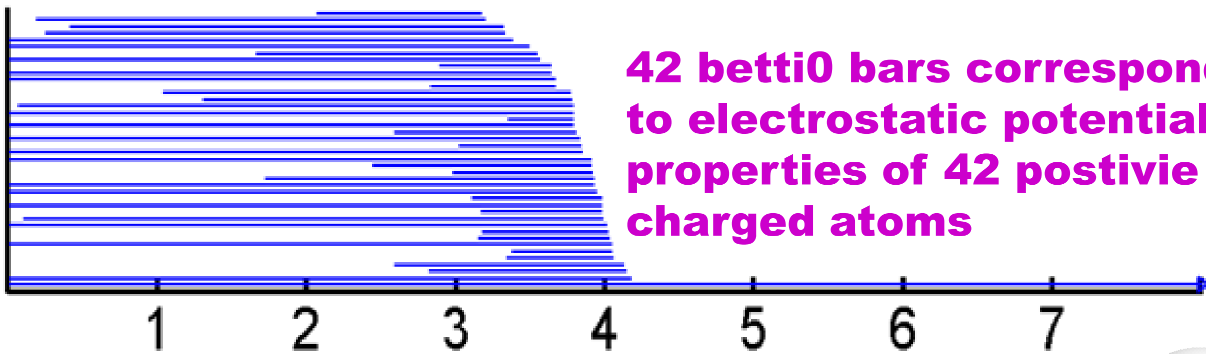


$\Phi = 0.4$

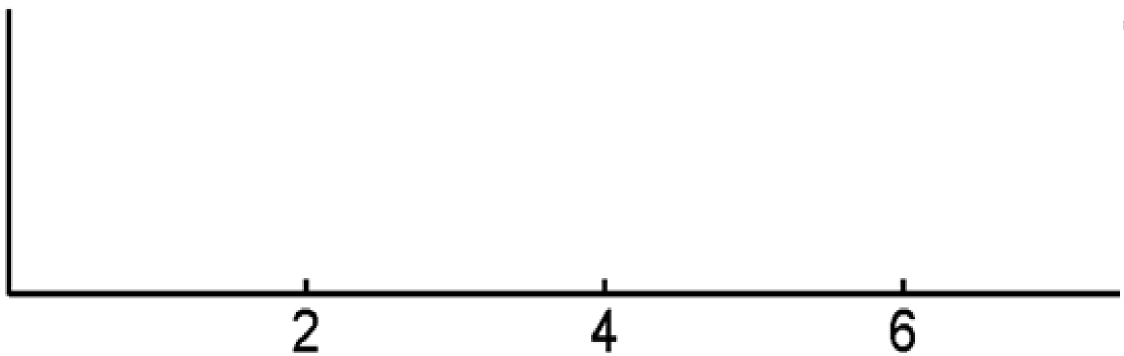


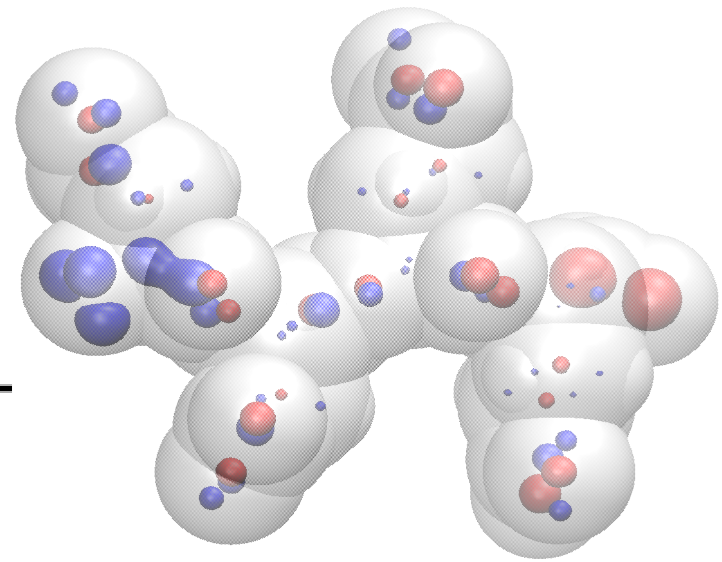
$\Phi = -0.4$

Persistent homology analysis for electrostatic potential

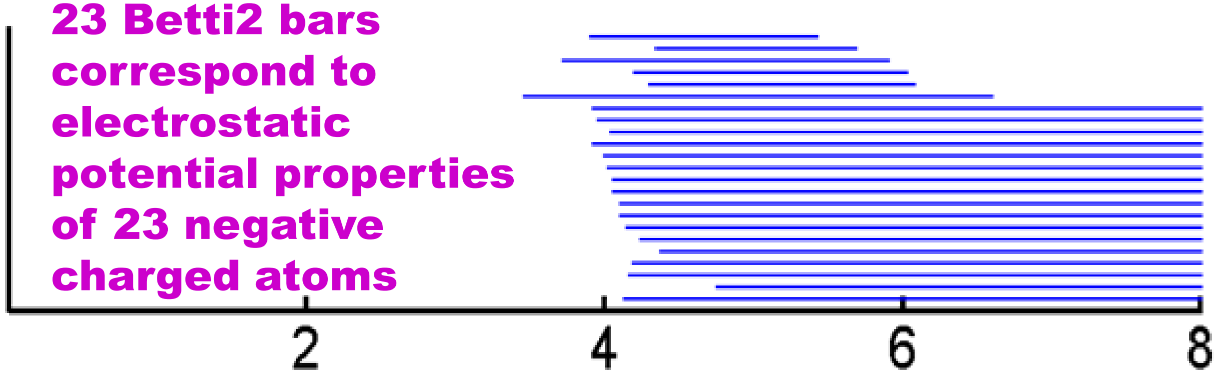


42 betti0 bars correspond to electrostatic potential properties of 42 postivie charged atoms



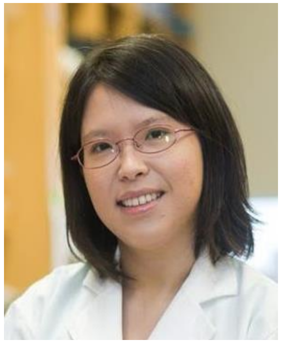


23 Betti2 bars correspond to electrostatic potential properties of 23 negative charged atoms

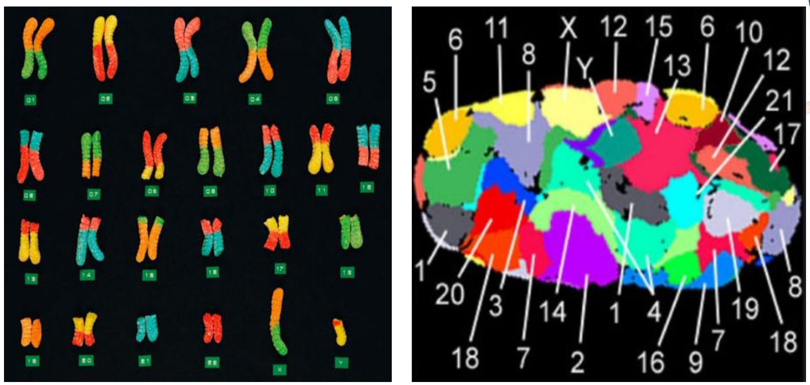


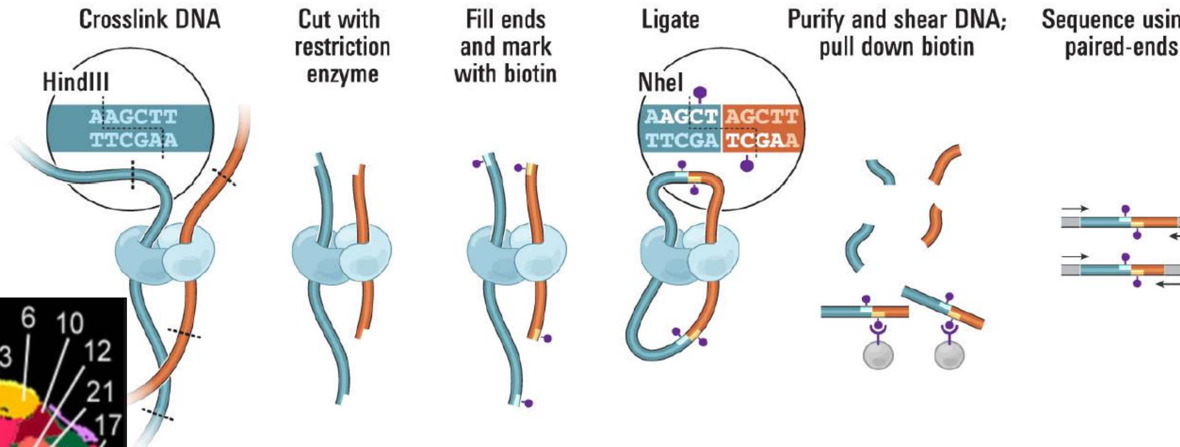
Extra topic--Hi-C Data analysis

Collaborator
Jiajie Peng
CS, NWPU



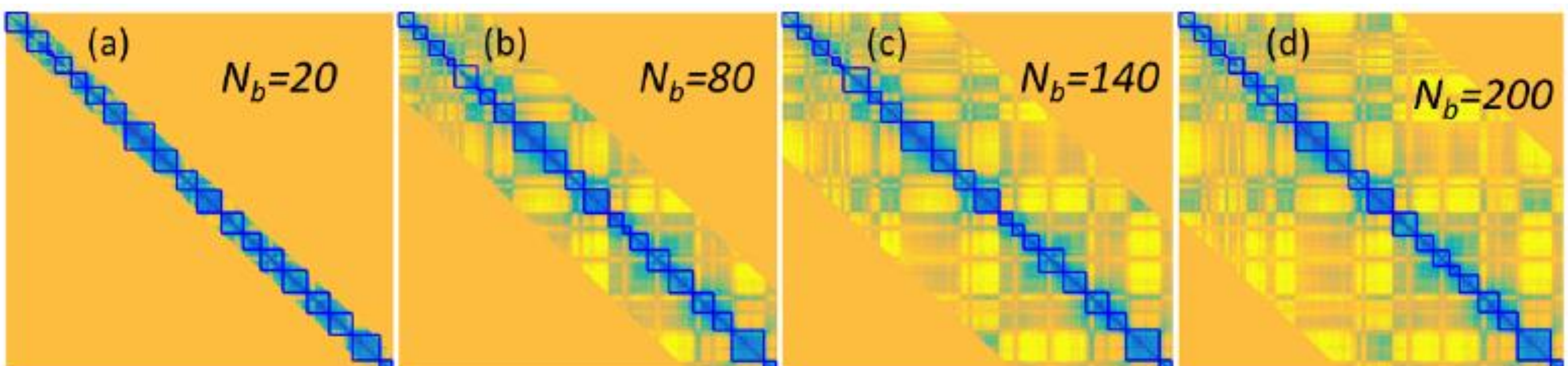
Collaborator
Melissa Fullwood
SBS, NTU





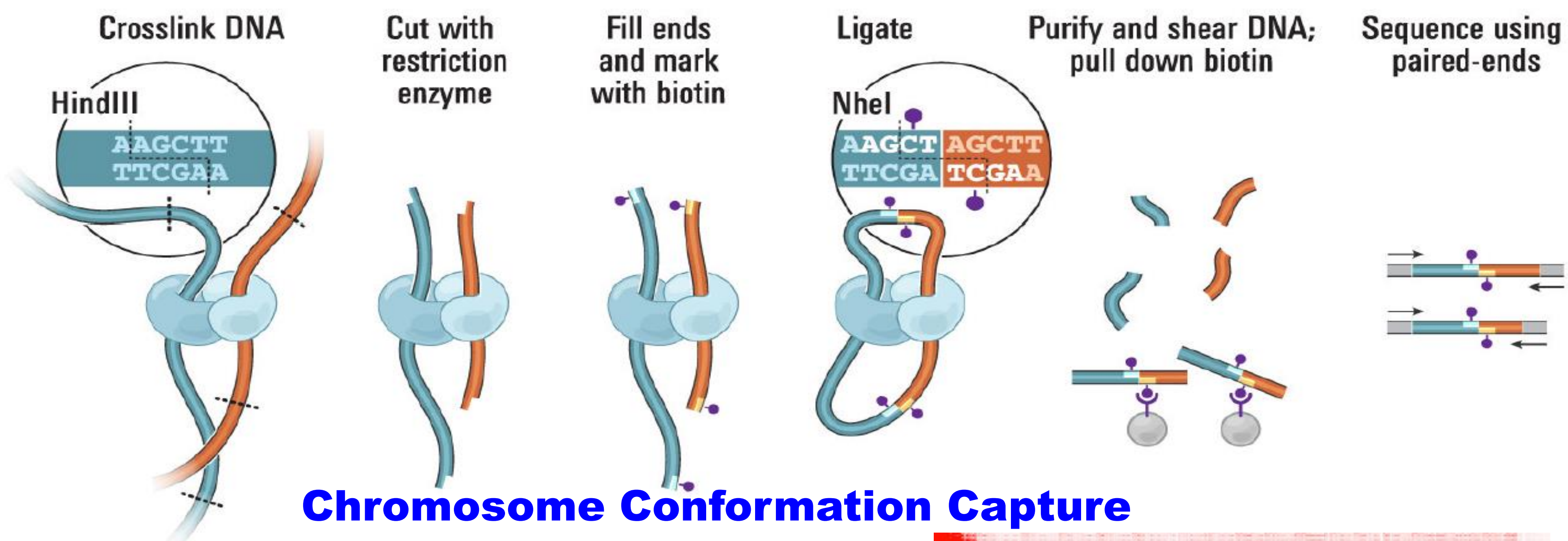
Chromosome Conformation Capture

A multiscale spectral graph model for Hi-C data analysis

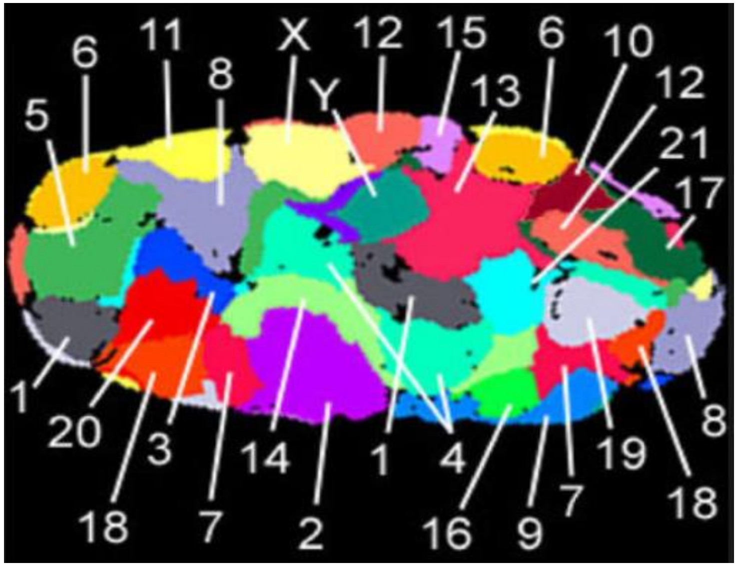


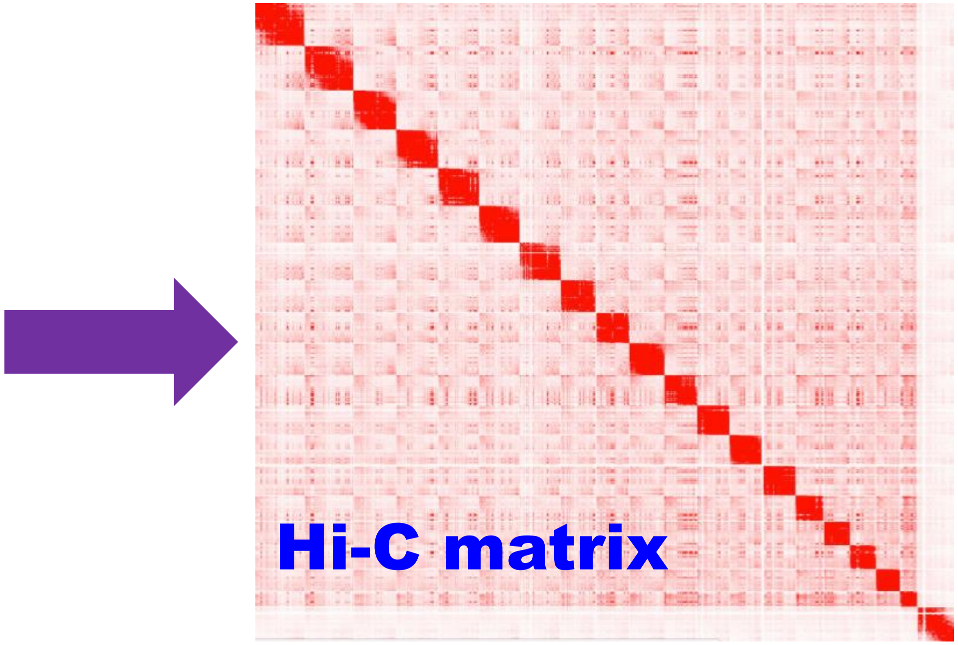
(Xia, Plos one, 2018)

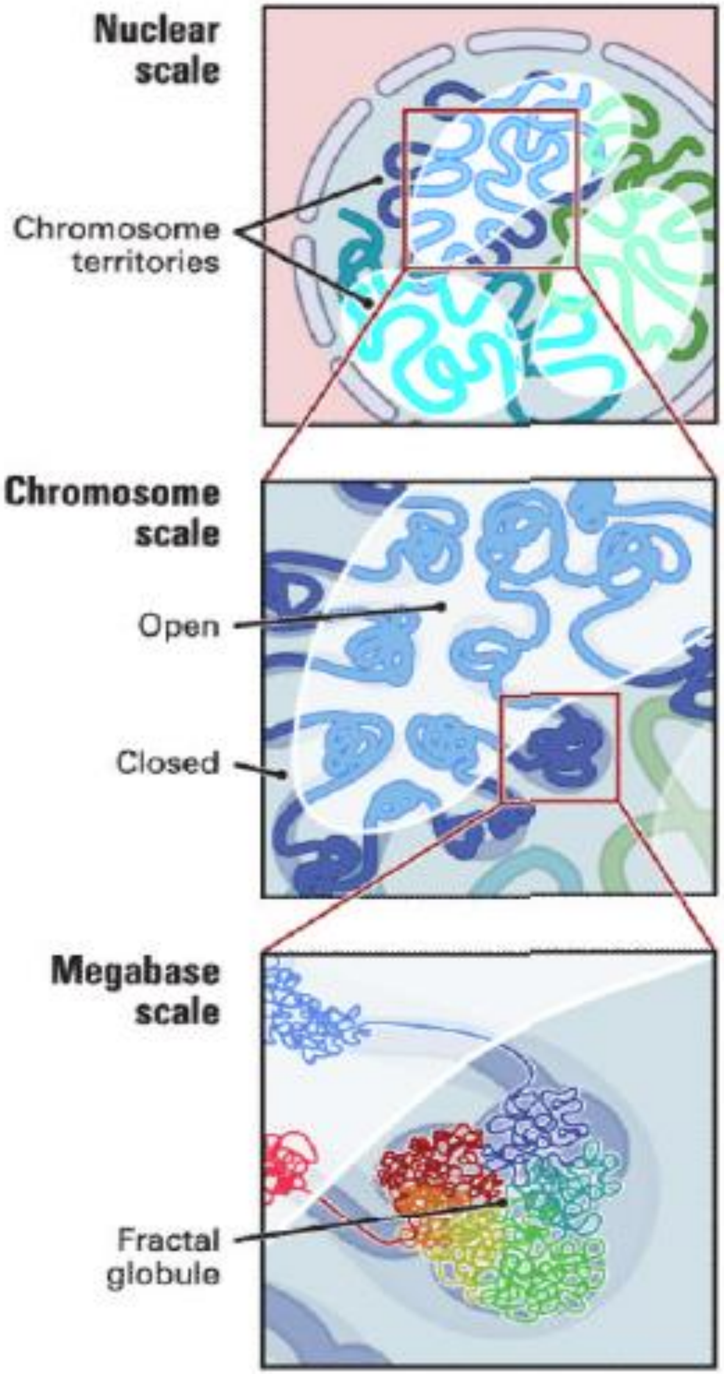
Hi-C data analysis



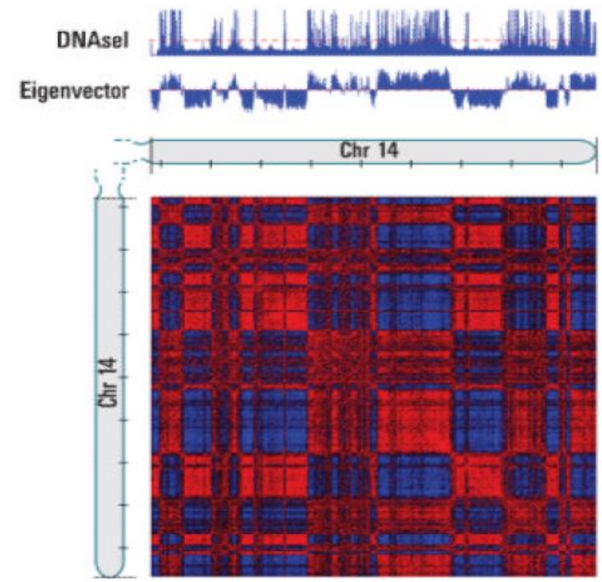
Chromosome Conformation Capture



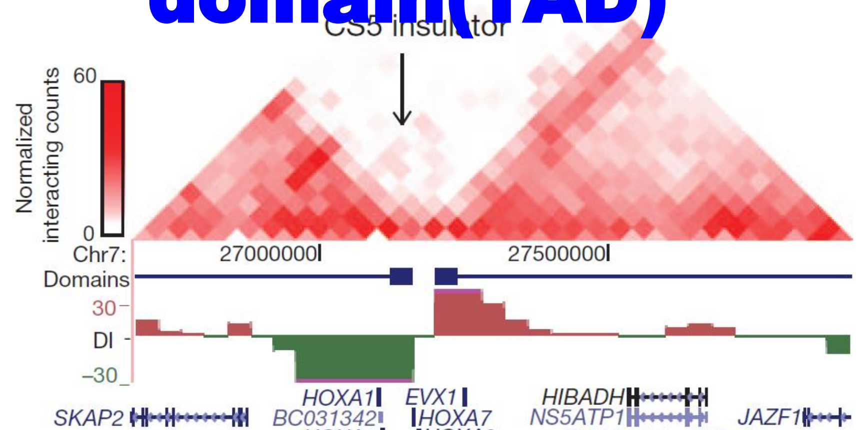




Genomic compartment



Topological associated domain(TAD)

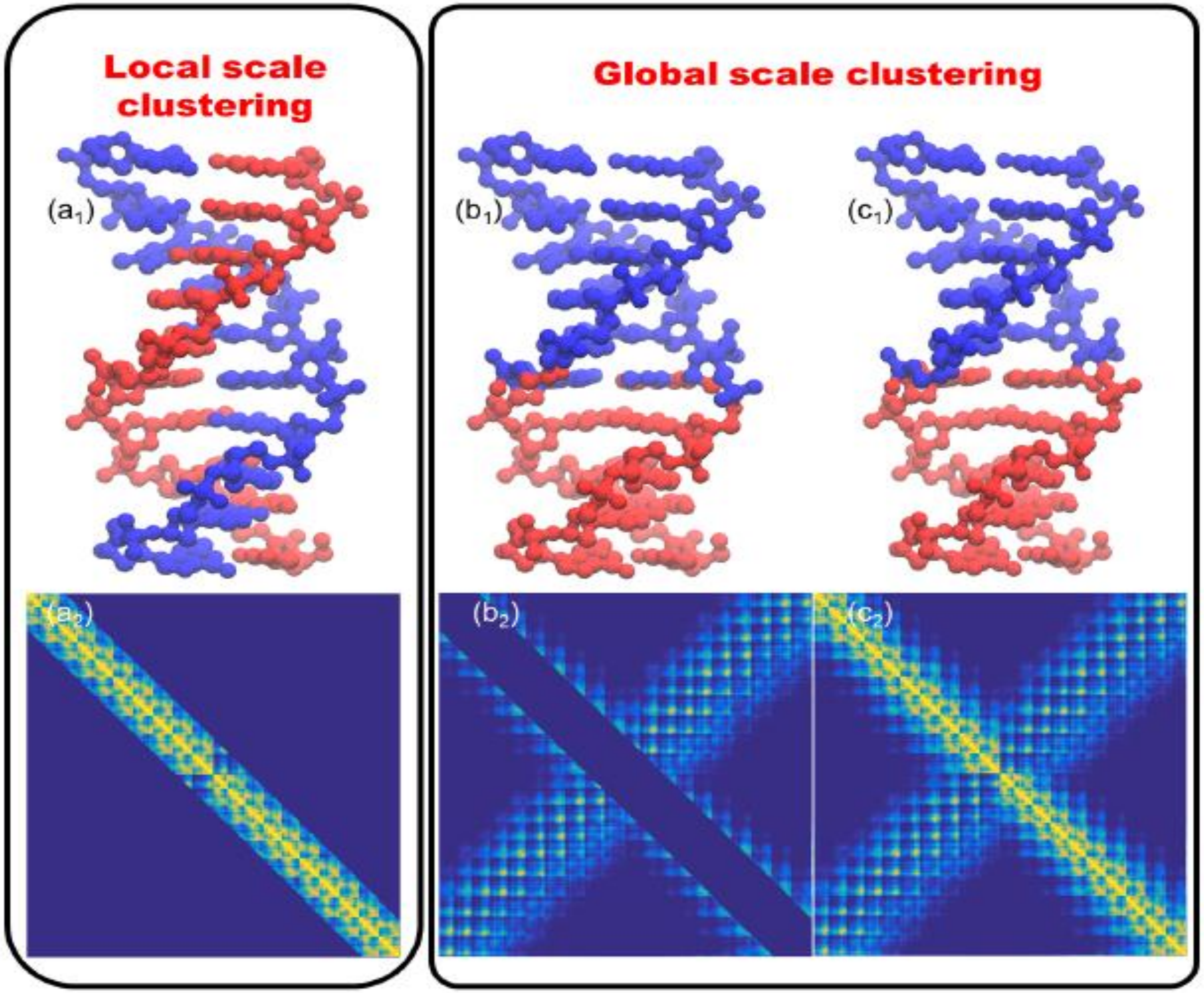


Megabase-sized local chromatin domain

Sequence-based multiscale models

Kelin Xia, PLOS ONE, 2018

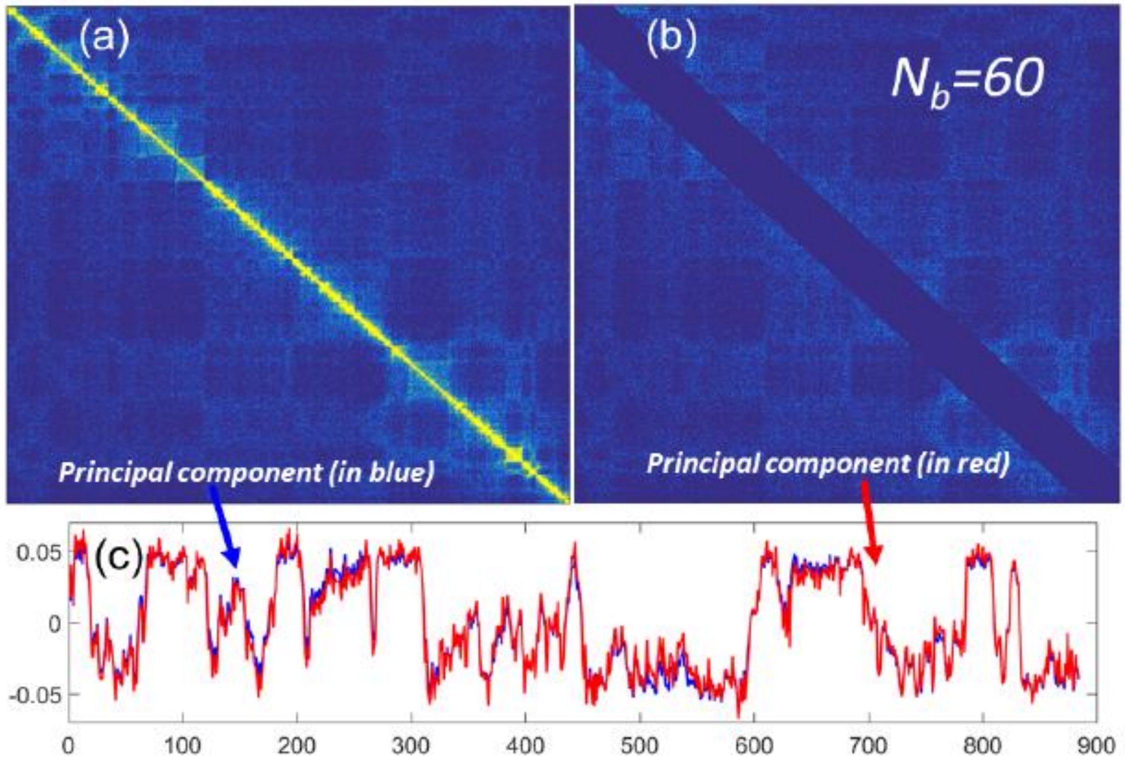
Optimize sequence information



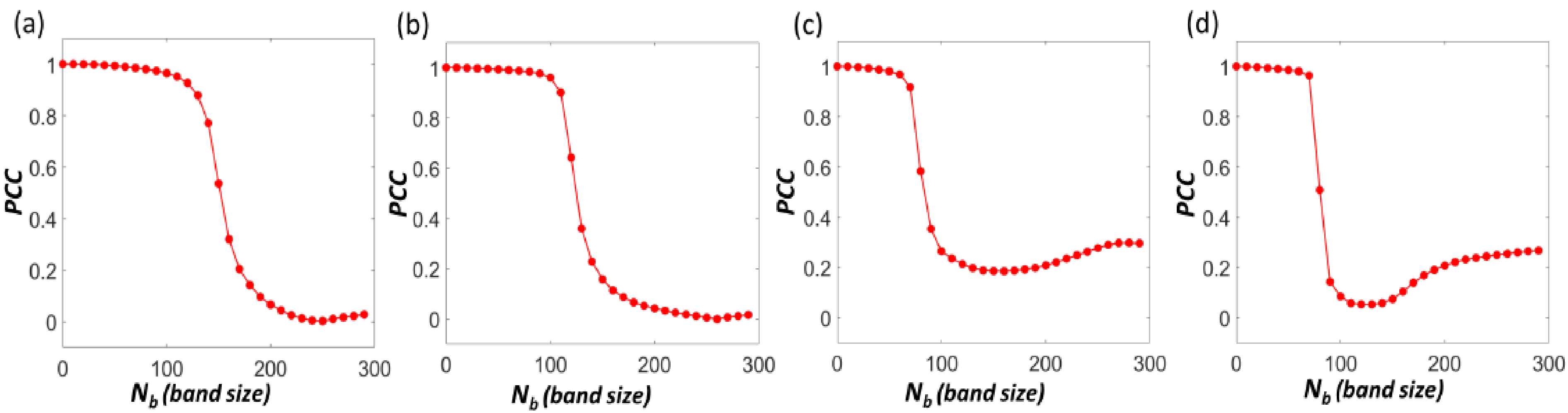
Optimize spatial information

TAD **Genomic compartment**

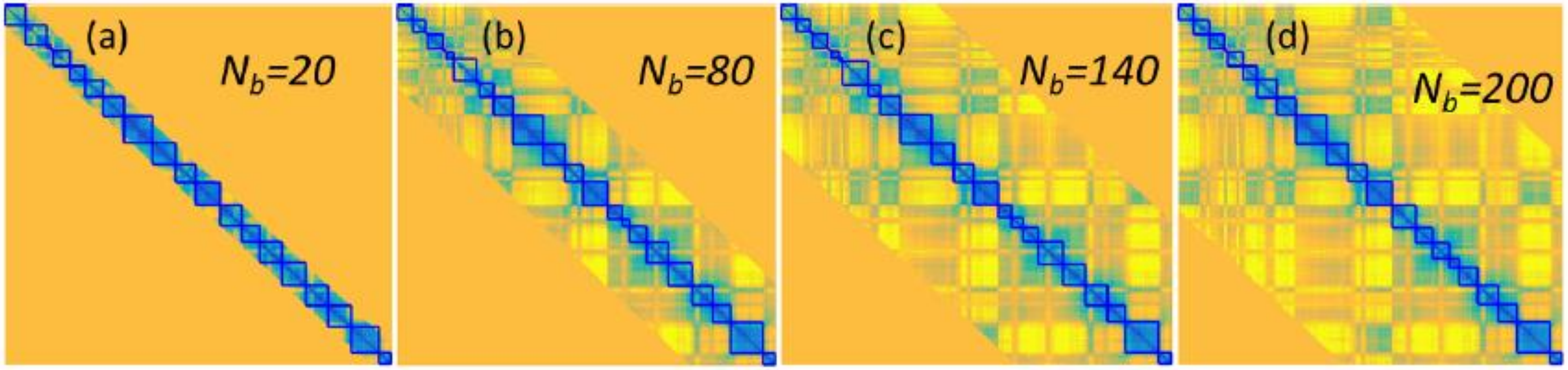
Genomic compartment analysis



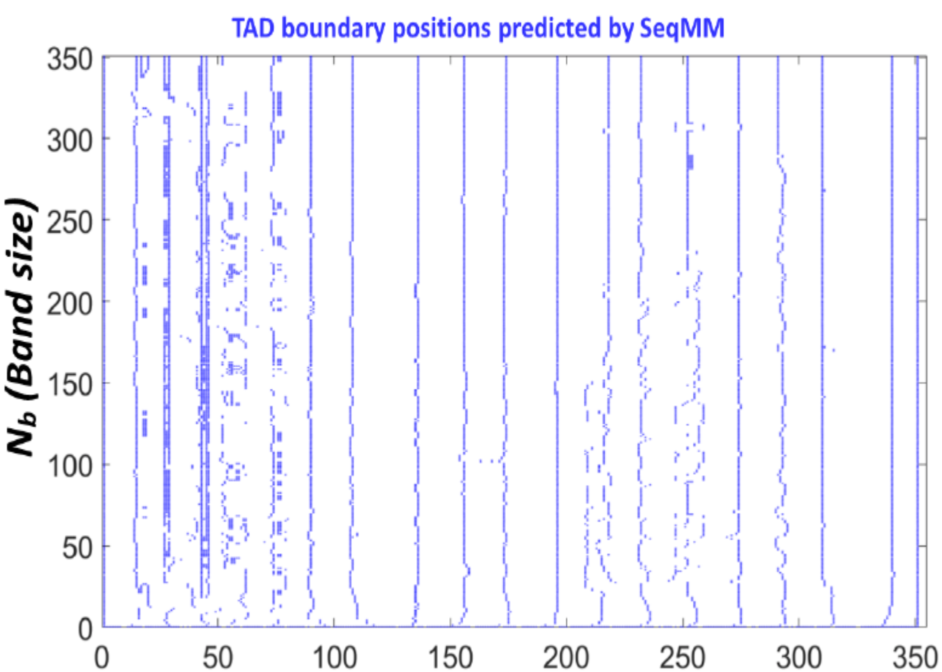
GM06990
Chromosome 14
Resolution:100kb



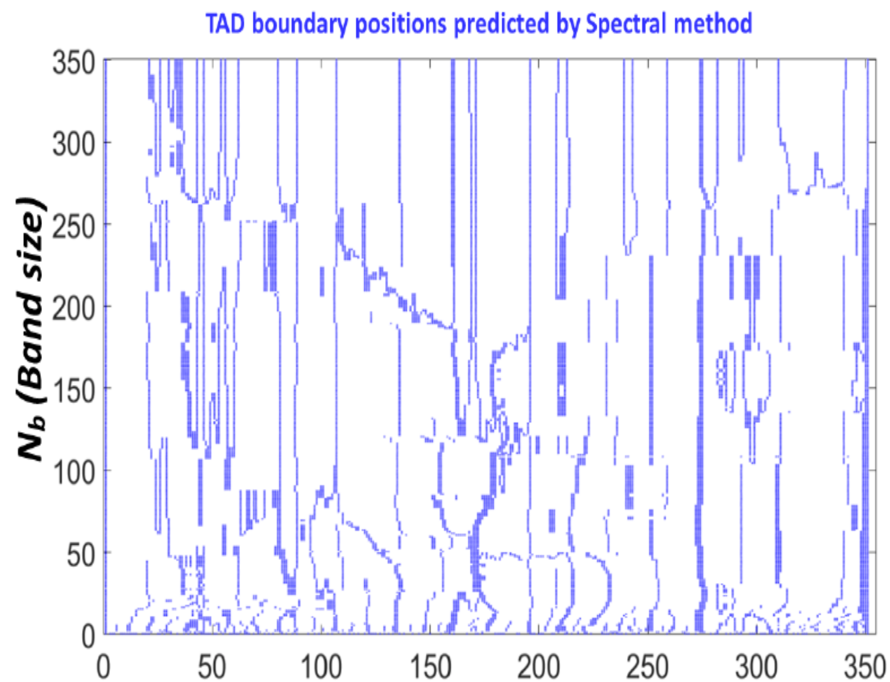
TAD analysis



IMR90, cell chromosome 22, Resolution:100kb

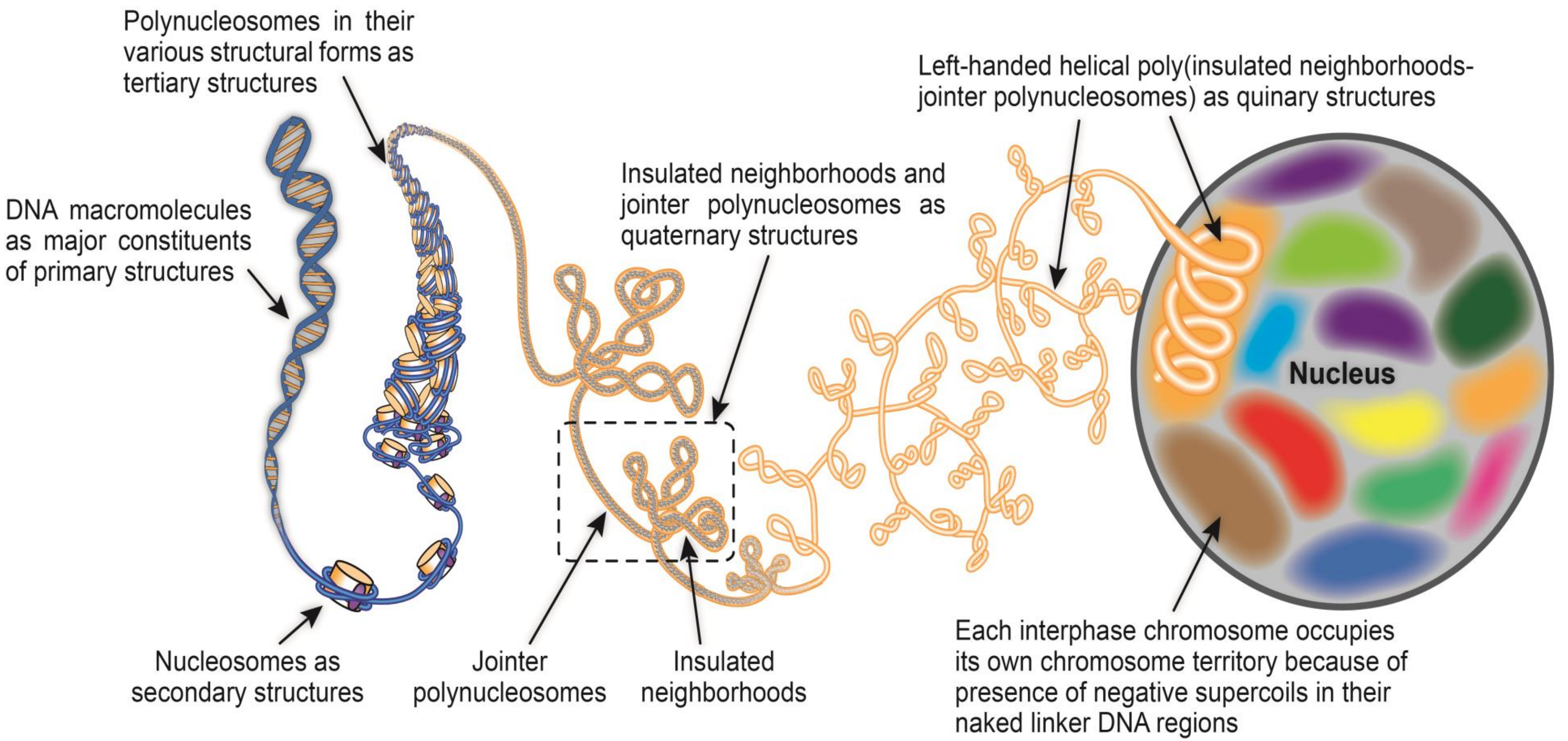


Our SeqMM



Chen’s spectral method

Multiscale Knots

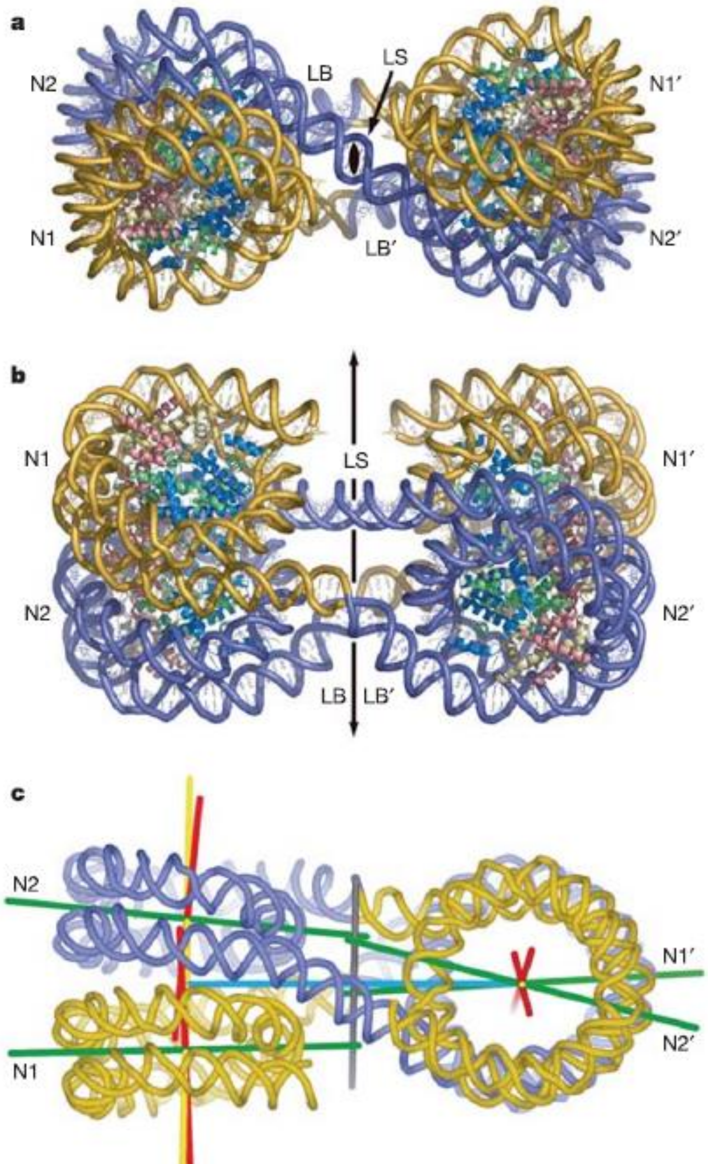


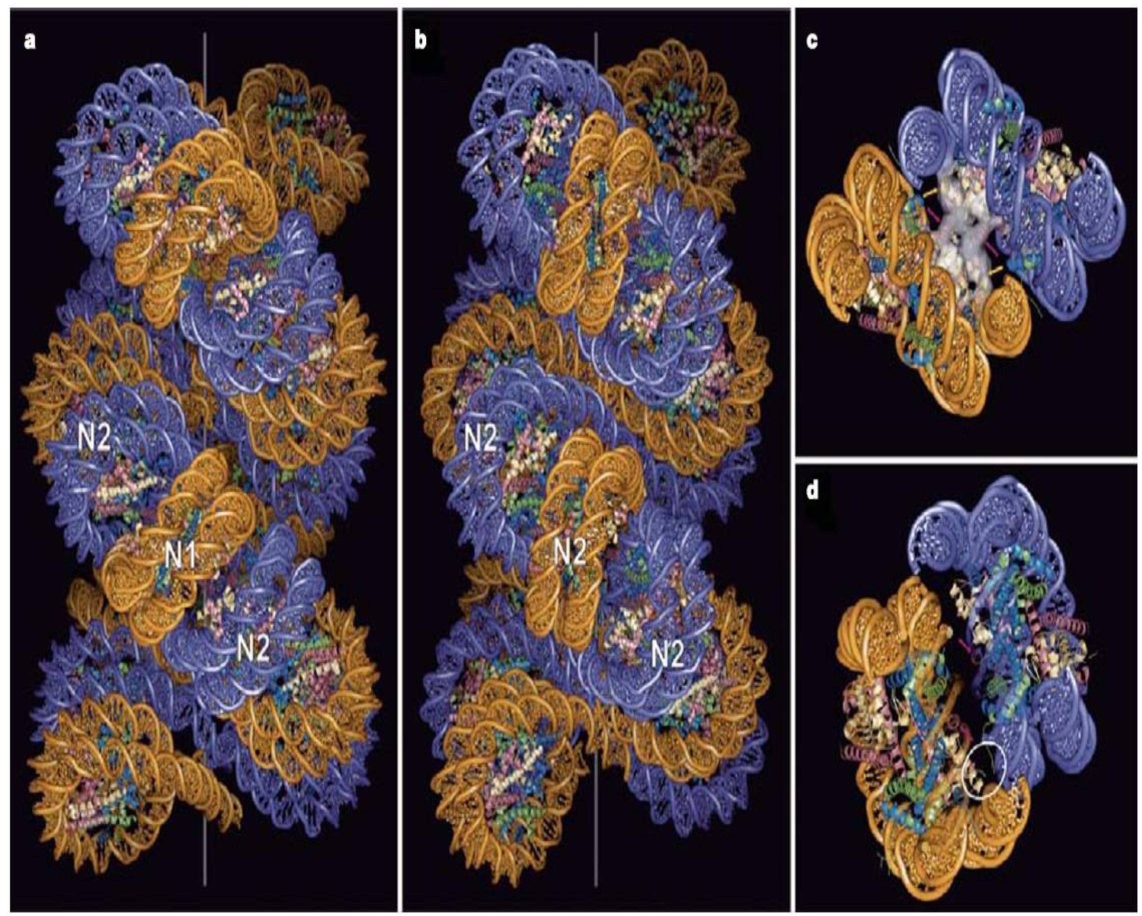
Supercoiling Theory and Model of Chromosomal Structures in Eukaryotic Cells

Hao Zhang, Tianhu Li*

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

Thomas Schalch[1], Sylwia Duda[1], David F. Sargent[1] & Timothy J. Richmond[1]



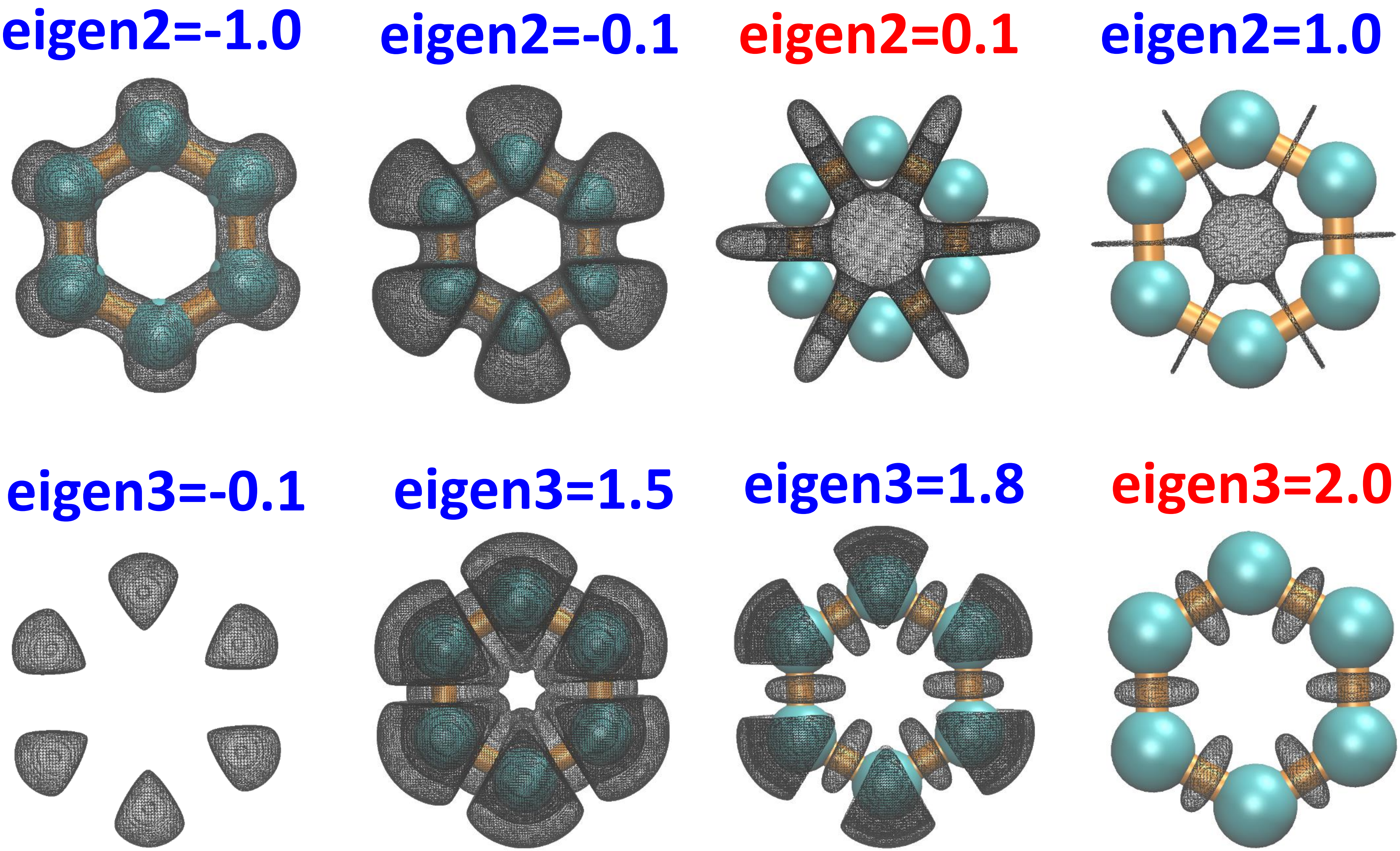


Group members

Grant support

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

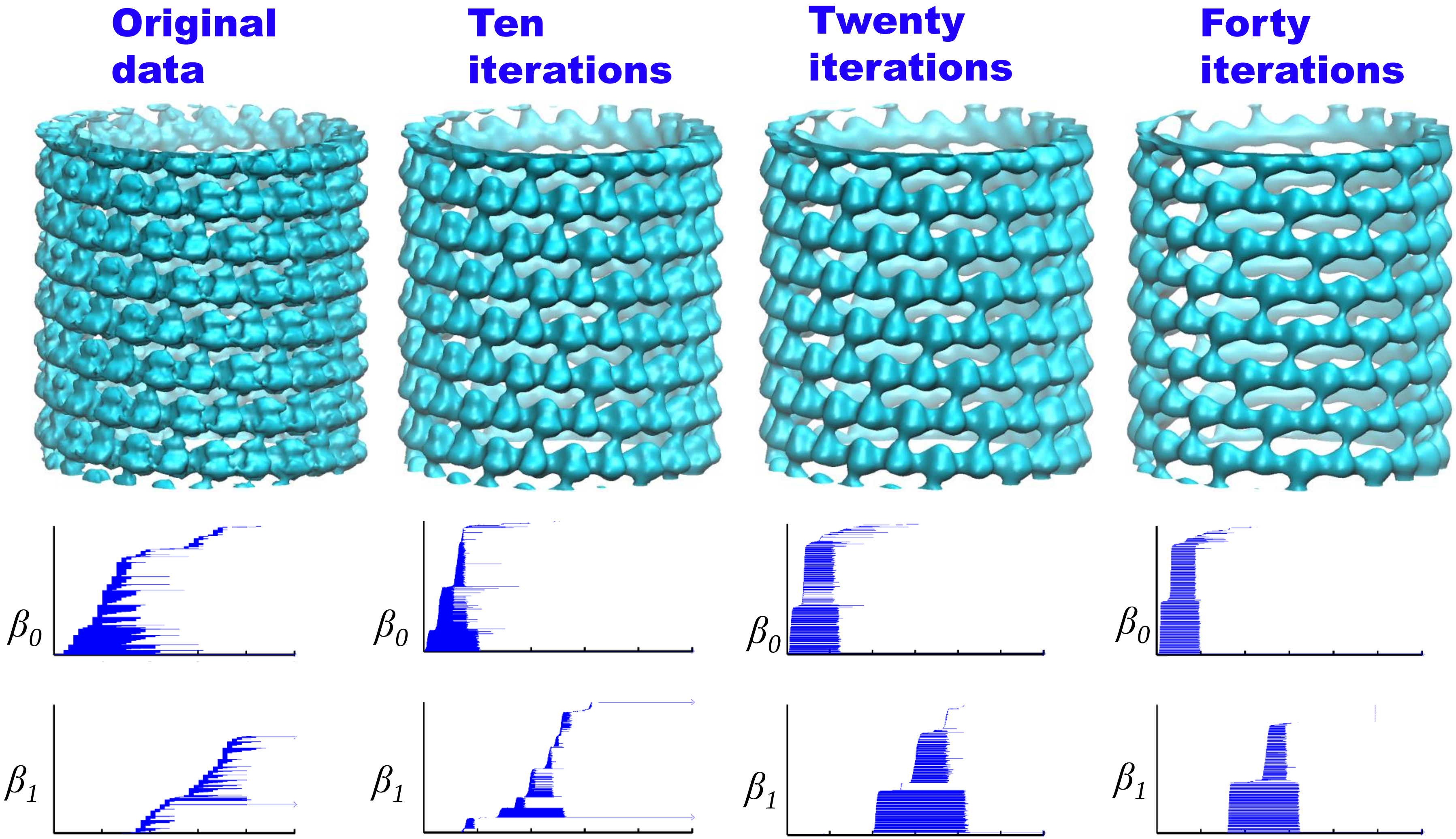
Isosurfaces of eigenvalues



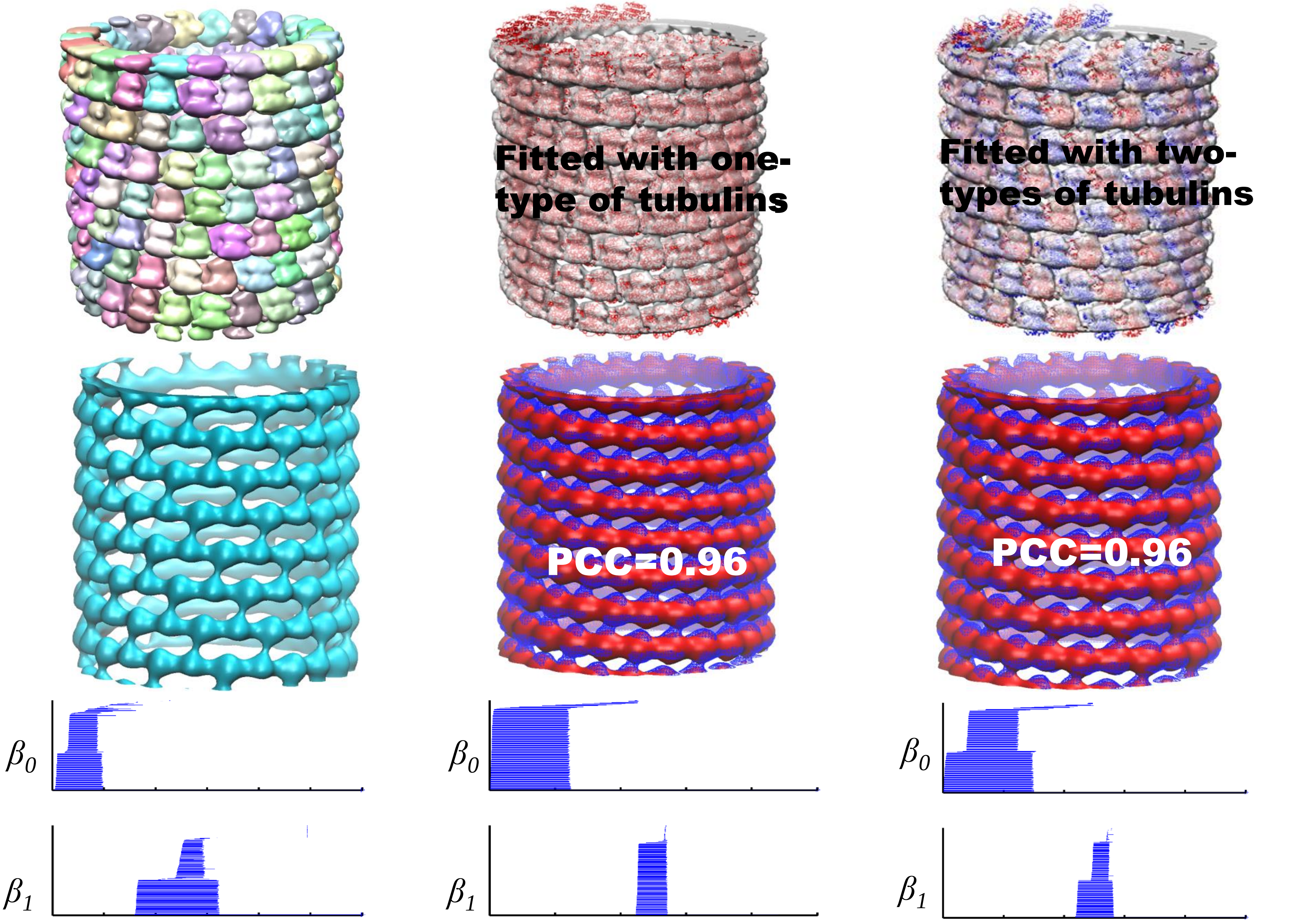
Topic--PHA for ill-posed inverse problems

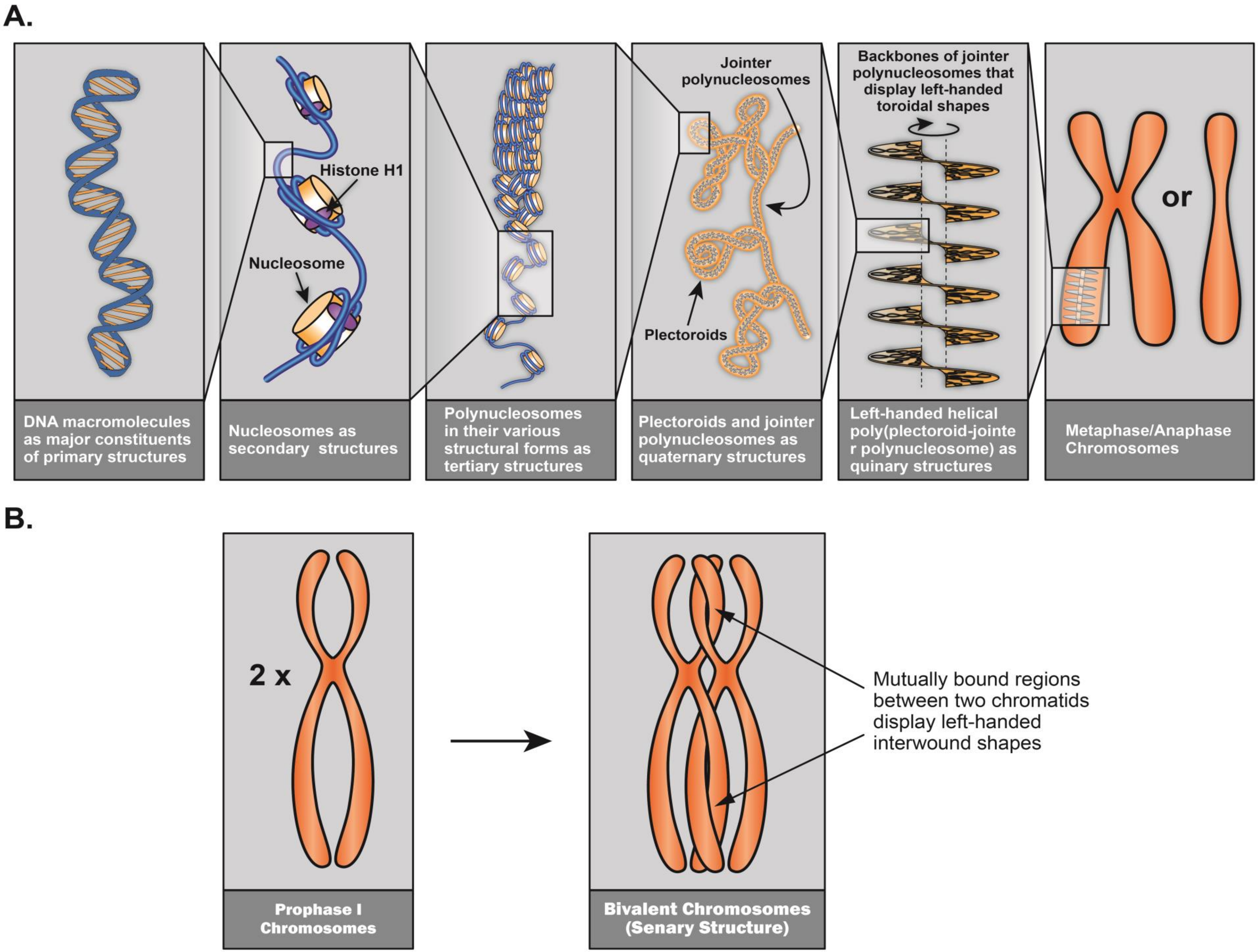
(Xia & Wei, IJNMBE, 2015)

Microtubule (EMD1129)



PHA for ill-posed inverse problems




A.
Jointer polynucleosomes
Backbones of jointer polynucleosomes that display left-handed toroidal shapes
Histone H1
Nucleosome
Plectoroids
or
DNA macromolecules as major constituents of primary structures
Nucleosomes as secondary structures
Polynucleosomes in their various structural forms as tertiary structures
Plectoroids and jointer polynucleosomes as quaternary structures
Left-handed helical poly(plectoroid-jointer polynucleosome) as quinary structures
Metaphase/Anaphase Chromosomes
B.
2 x
Prophase I Chromosomes
Bivalent Chromosomes (Senary Structure)
Mutually bound regions between two chromatids display left-handed interwound shapes