Topological modeling of biomolecular data

Kelin Xia[1], Xin Feng[2], Zhixiong Zhao[1], Rundong Zhao[2], Yiying Tong[2] and Guo-wei Wei[1]*

[1]Department of Mathematics, Michigan State University, East Lansing, MI, 48824
[2]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824

Introduction

Biomolecules have enormous amount of topological information, which tightly connects with their biological functions. In this work, we reveal the topology-function relationship of biomolecules using the persistent homology analysis (PHA). We use fullerene molecules as an example and find out that our PHA is able to quantitatively predict fullerene stability. We introduce PHA to extract molecular topological fingerprints (MTFs) based on the persistence of molecular topological invariants, and utilize MTFs for biomolecular data analysis, including characterization, identification and analysis (CIA). We construct both all-atom and coarse-grained representations of MTFs. We further employ MTFs to characterize protein topological evolution during protein folding and quantitatively predict the protein folding stability. An excellent consistence between our persistent homology prediction and molecular dynamics simulation is found. Finally, we propose a multiscale, multiresolution and multidimensional persistent homology to match the resolution with the scale of interest so as to create a topological microscopy for the underlying data.

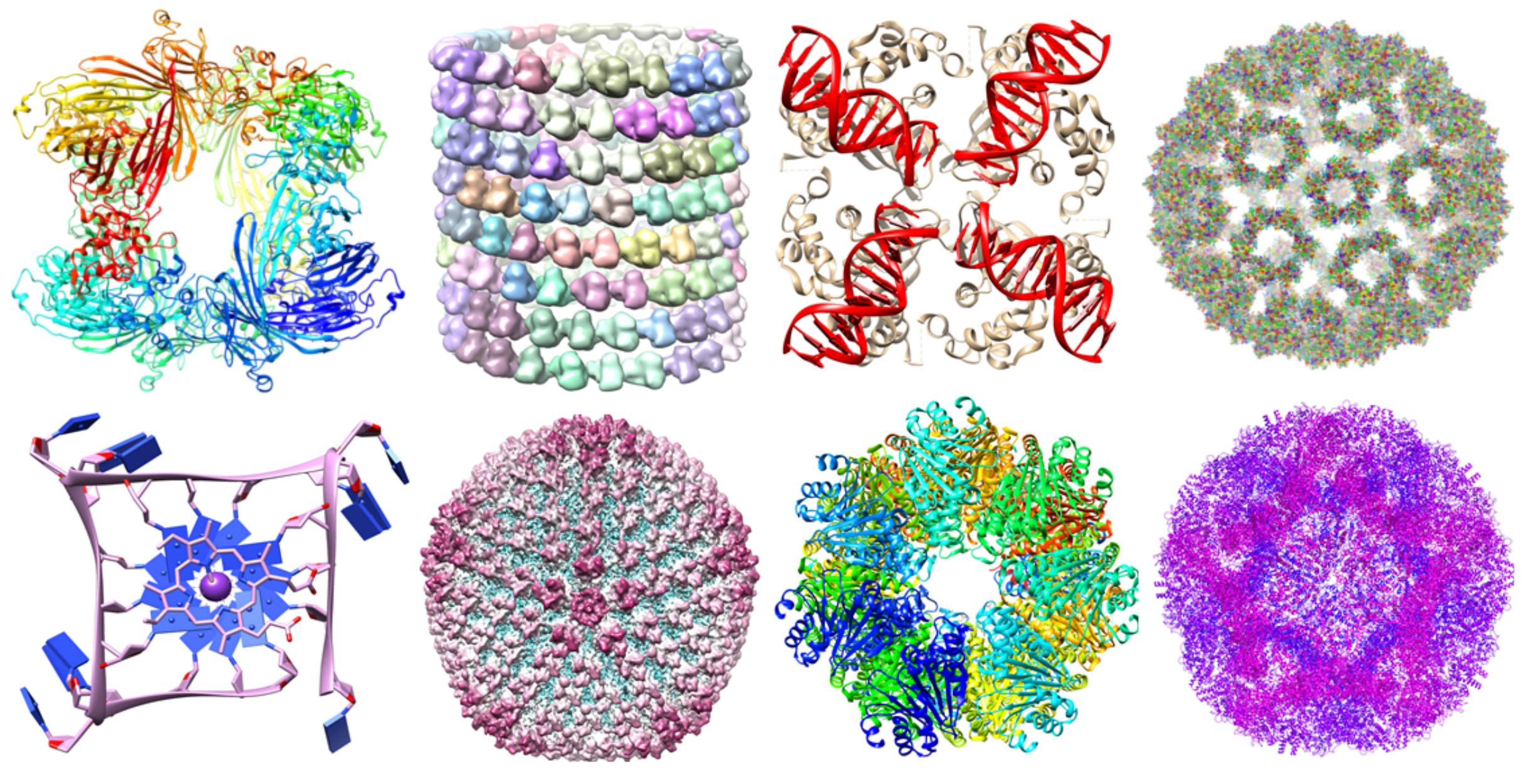


Figure: Enormous amount of topological information in biomolecular systems.

Persistent homology theory

Homology utilizes a topological space with an algebraic group representation to characterize topological features, such as isolated components, circles, holes and void. Persistent homology further embeds geometric information to topological invariants through a filtration process, in which a series of nested simplicial complexes is constructed and topological structures are characterized continuously over a range of scales.

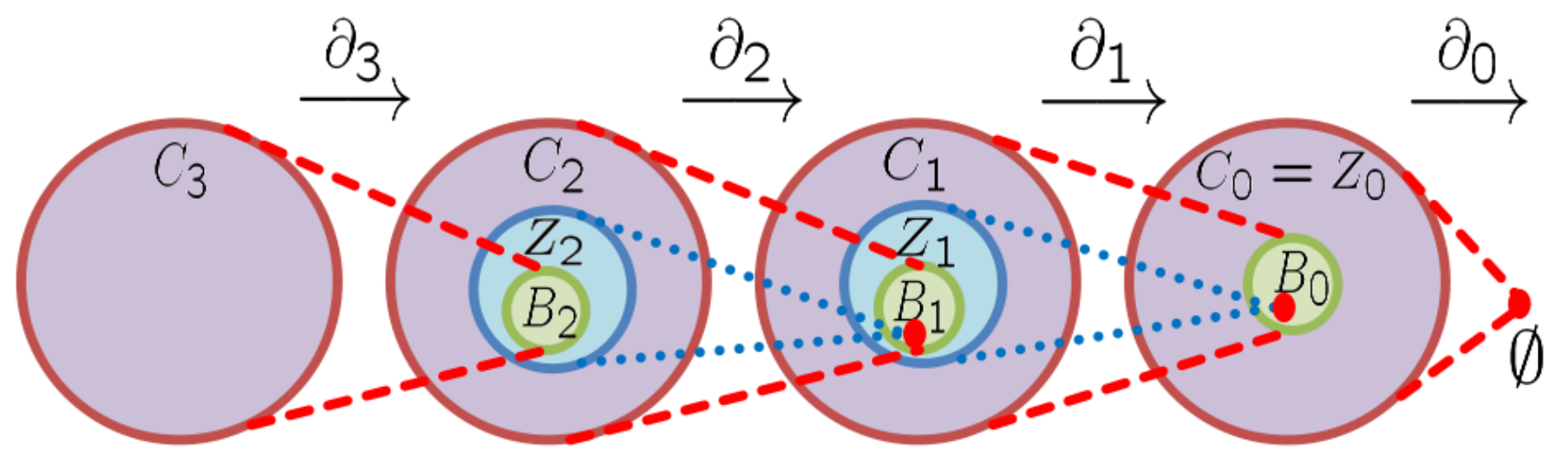


Figure: Illustration of boundary operators, and chain, cycle and boundary groups. Red dots stand for empty sets.

- **Fundamental theorem of finitely generated abelian groups:**

$$H_k = Z \oplus \cdots \oplus Z \oplus Z_{p_1} \oplus \cdots \oplus Z_{p_n} = Z^{\beta_k} \oplus Z_{p_1} \oplus \cdots \oplus Z_{p_n}. \tag{1}$$

Betti number can be simply calculated by

$$\beta_k = \text{rank } H_k = \text{rank } Z_k - \text{rank } B_k. \tag{2}$$

- **Filtration process:**

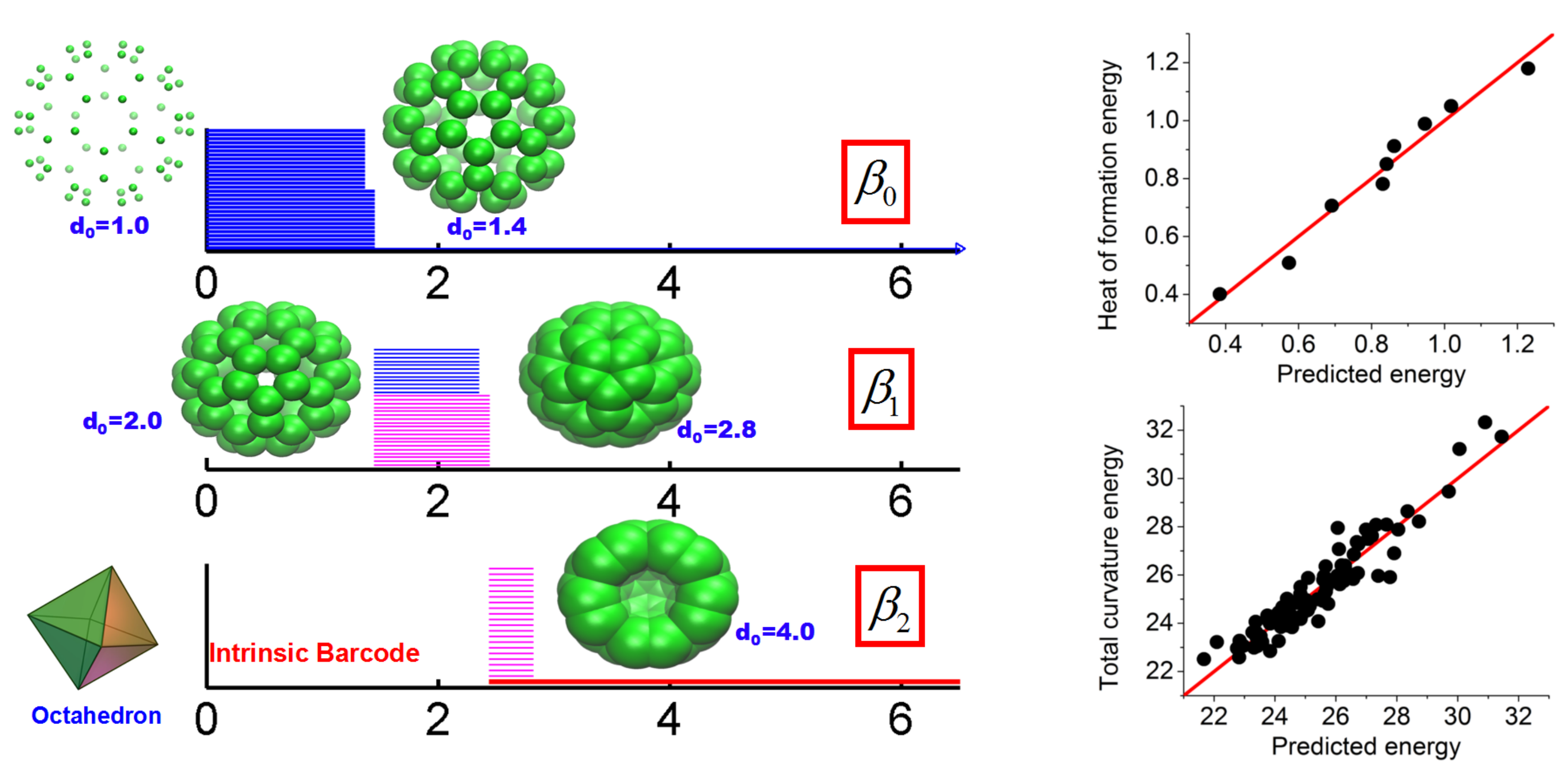
$$\varnothing = K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K. \tag{3}$$

- **Persistence:** The p-persistent k-th homology group K^i is

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \bigcap Z_k^i). \tag{4}$$

Fullerene stability analysis

Fullerene molecules have carbon-cage structures, which contain only pentagonal and hexagonal rings. Using the Vietoris-Rips complex, the persistence of Betti numbers (i.e., ranks of homology groups), including β_0, β_1 and β_2, provides the information regarding length of the bond, width of the pentagon and hexagon ring, and size of the central void.



Molecular topological fingerprints

Persistent homology is used to extract MTFs based on barcode representation. MTFs are utilized for protein CIA.

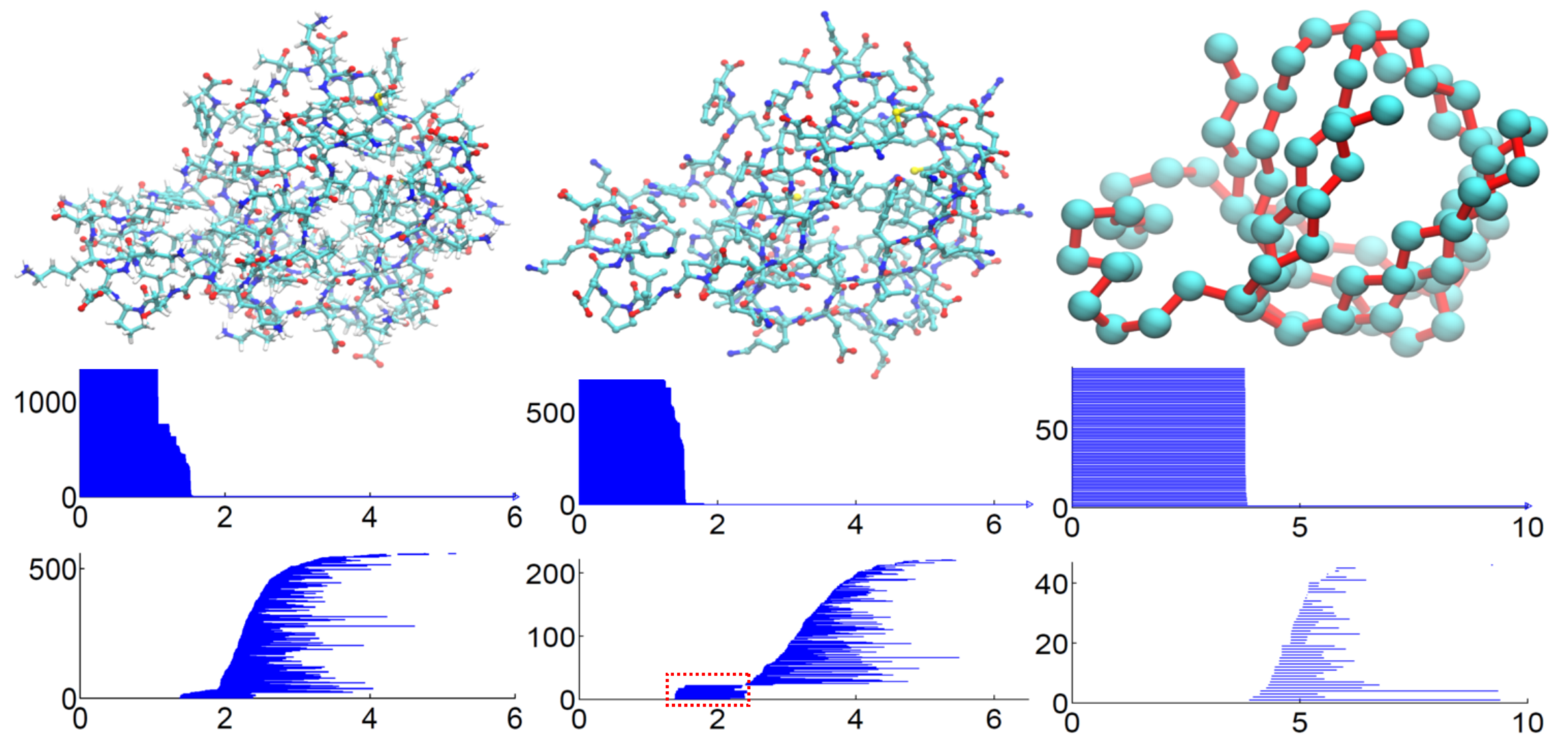


Figure: The identification of MTFs from two all-atom representations, i.e., hydrogen-included and non-hydrogen, and coarse-grain representation.

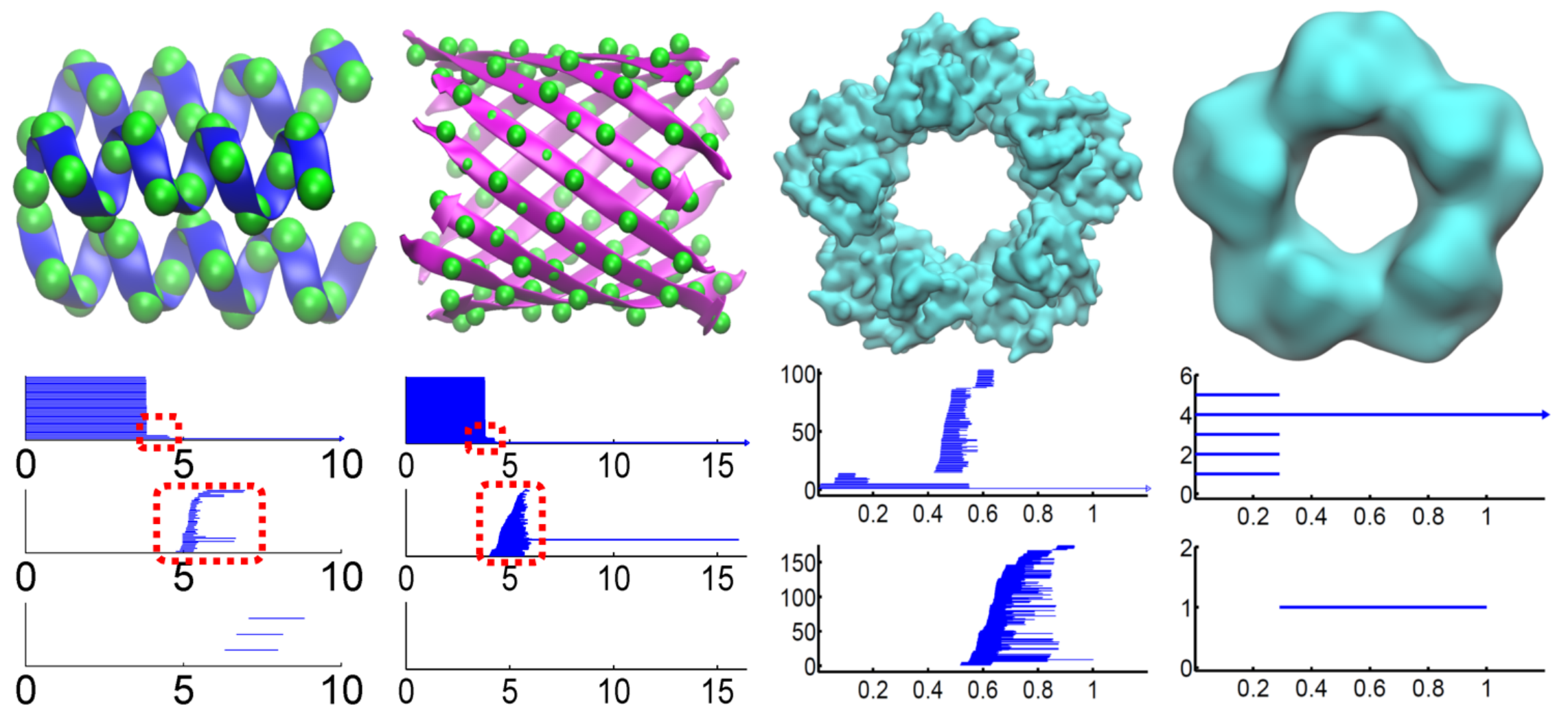


Figure: Various types of MTFs.

Cryo-EM data analysis

Cryo-electron microscopy (Cryo-EM) data is usually suffered from low signal to noise ratio (SNR). A geometric flow based denoising process is used to reveal the intrinsic topological invariants.

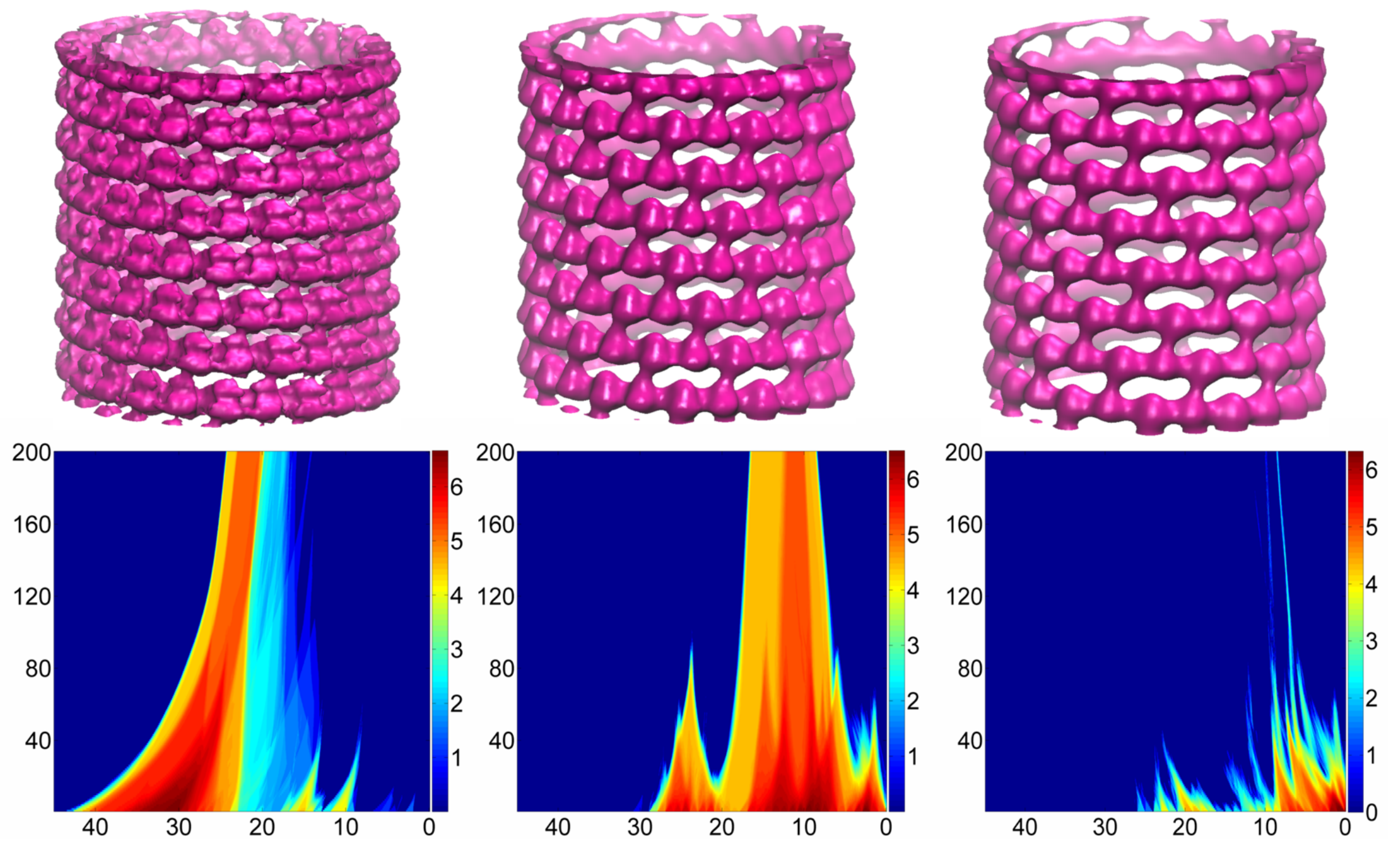


Figure: Persistent Betti number (PBN) representation of a noise reduction process. A consistent pattern can be observed after 100 iteration steps.

Two types of microtubule models have similarly high correlation coefficient but with dramatically different MTFs.

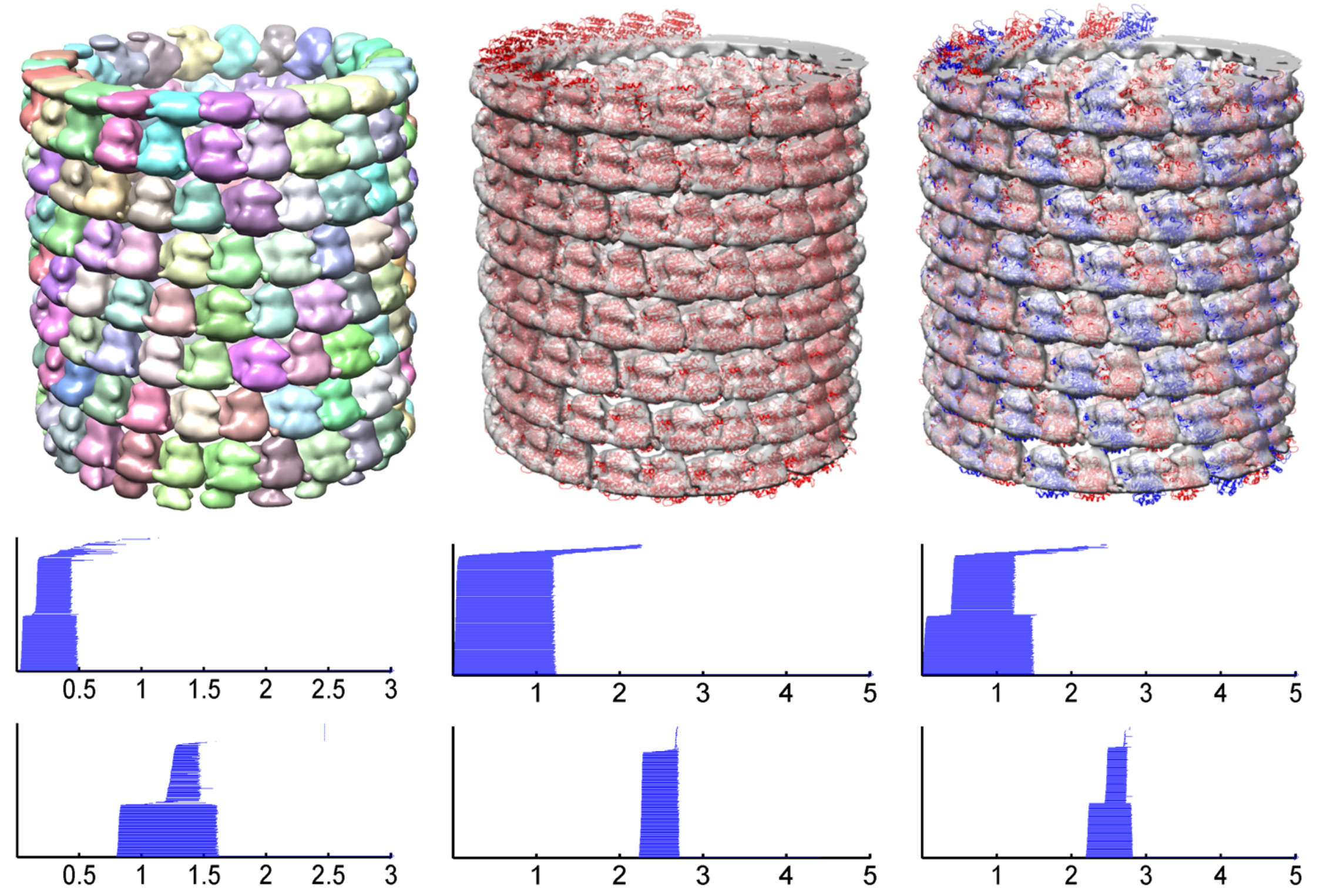


Figure: MTFs for microtubule experimental data and two fitting models, i.e., monomer model and dimer model, indicating that the latter model is topologically correct.

Protein folding analysis

We use steered molecular dynamics to simulate the unfolding process of PDB 1UBQ. To analyze its topological evolution, PBNs are calculated for every configurations and then stacked together in sequence to deliver a multidimensional persistence diagram.

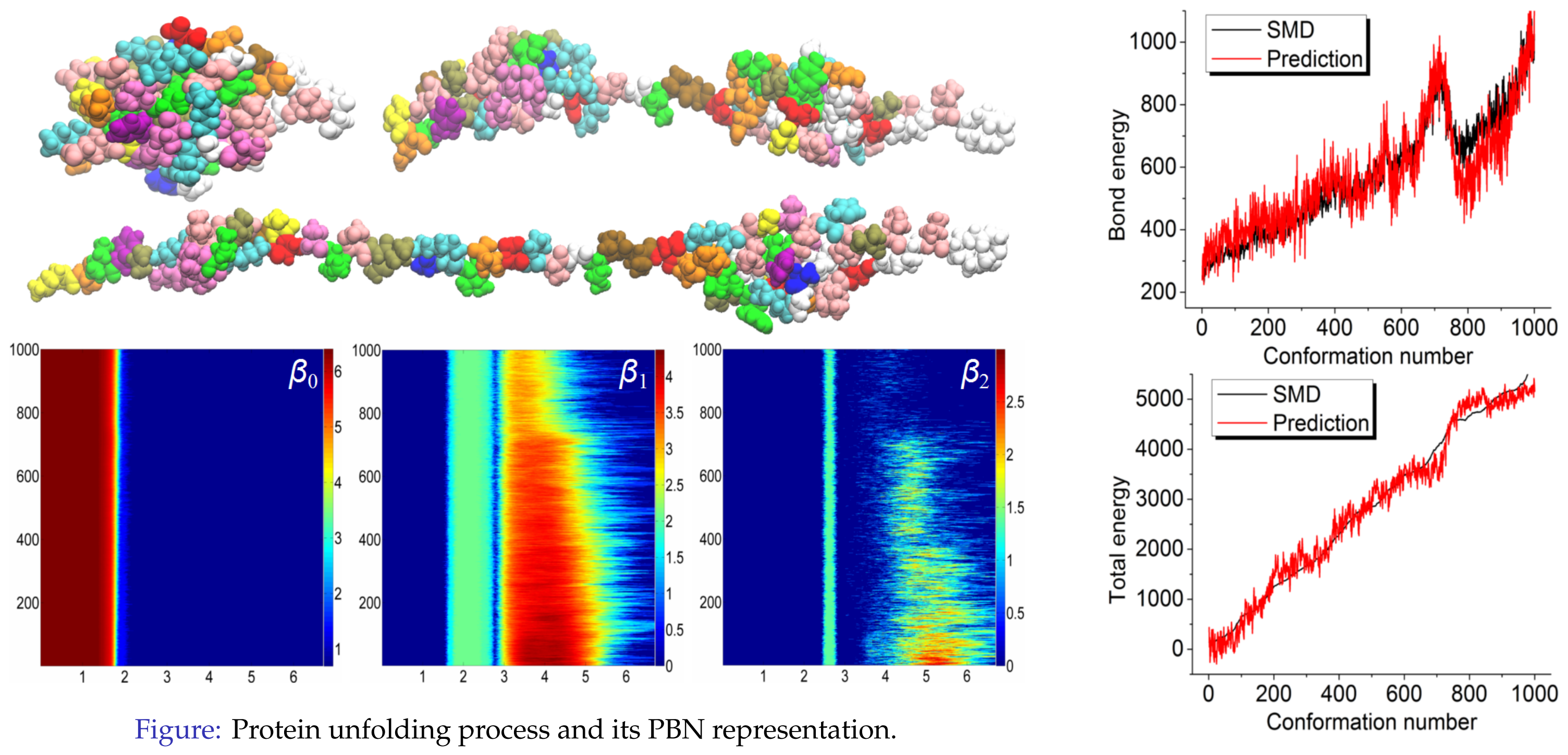


Figure: Protein unfolding process and its PBN representation.

The accumulated bar lengths of β_0 and β_1 are used to quantitatively predict bond energy and total energy during the unfolding process. The Pearson's correlation coefficients are 0.924 and 0.990, respectively.

Multiscale, multiresolution and multidimensional topology

To describe multiscale properties, we introduce a rigidity density function with a resolution parameter based on our FRI theory. When systematically changing the resolution, a series of geometric and topological representations of the underlying system will be produced. Our multiresolution PHA can match the resolution to the scale of interest so as to create a topological microscopy for the underlying data.

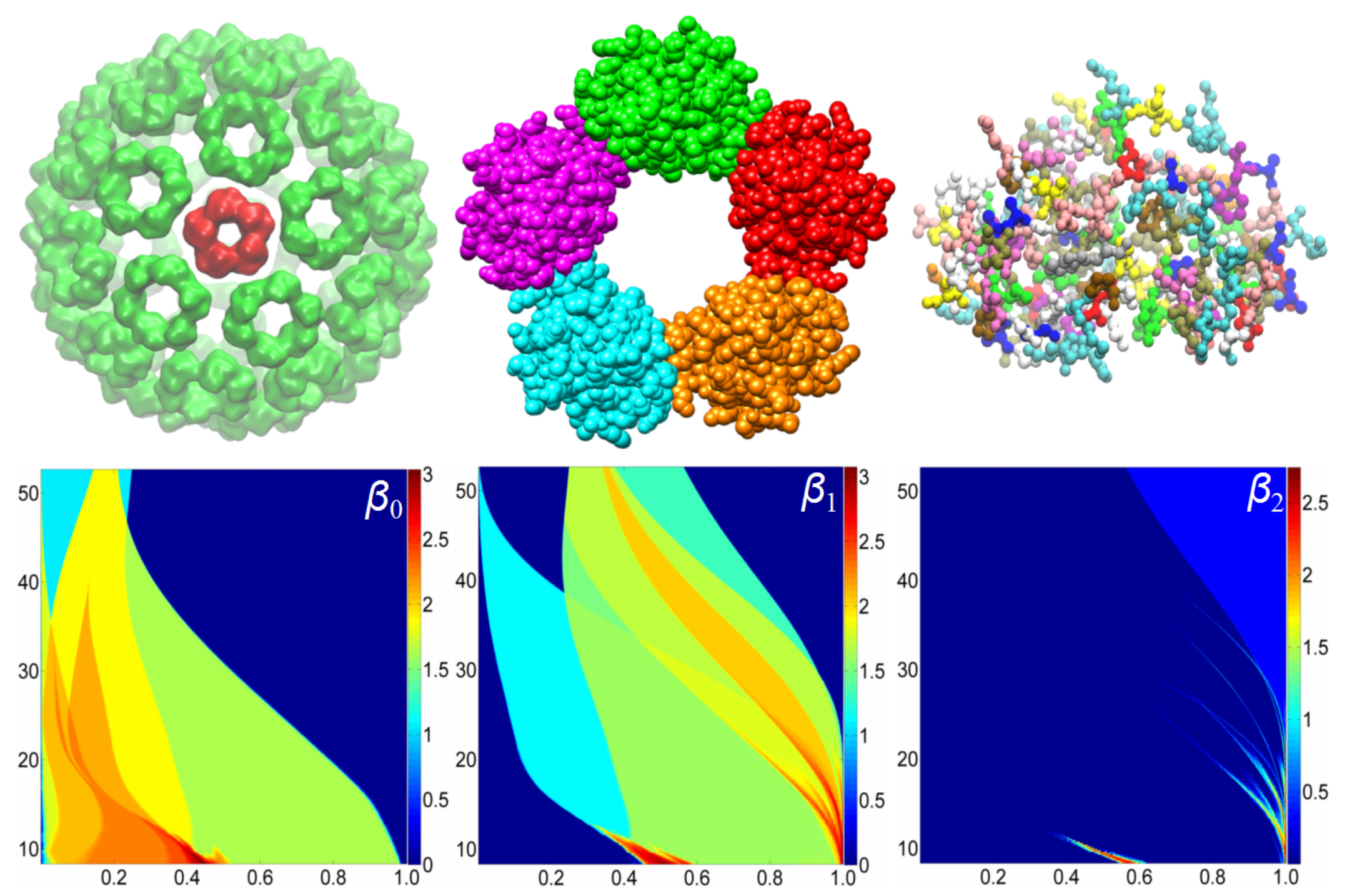


Figure: Different scales within protein complex 1DYL.

Conclusion

We introduce PHA for quantitative modeling of the biomolecular systems, and establish biomolecular topology-function relationship through the study of protein flexibility, rigidity, folding and multiscale structure properties. Based on biomolecular data obtained from different sources, i.e., Protein Data Bank and Cryo-EM Data Bank, and in different representations, i.e., point-cloud and density data, we employ various ways to represent biomolecular structures and construct simplicial complexes for our PHA. Moreover, we introduce a new framework of multiresolution and multidimensional persistent topology. As a result, we can systematically analyze a dynamic process and identify related intrinsic topological properties. Finally, by varying the resolution, we are able to deliver a full geometric and topological "spectrum" for biomolecular analysis.

References

- 1. Kelin Xia and Guo-Wei Wei, "Persistent homology analysis of protein structure, flexibility and folding", IJNMBE, 30, 814-844 (2014)
- 2. Kelin Xia, Xin Feng, Yiying Tong and Guo-Wei Wei, "Persistent homology for the quantitative prediction of fullerene stability", JCC, 36, 408-422 (2015)
- 3. Kelin Xia and Guo-Wei Wei, "Multidimensional persistence in biomolecular data", accepted, JCC (2015)
- 4. Kelin Xia and Guo-Wei Wei, "Persistent topology for cryo-EM data analysis", accepted, IJNMBE (2015)
- 5. Kelin Xia, Zhixiong Zhao and Guo-Wei Wei, "Multiresolution topological simplification", accepted, JCB (2015)

Acknowledgement

This work was partially supported by NSF grants DMS-1160352 and IIS-1302285, NIH Grant R01GM-090208 and MSU Center for Mathematical Molecular Biosciences initiative.

http://www.math.msu.edu/~wei/ http://www.cse.msu.edu/~ytong/ http://www.msu.edu/~xiakelin/ wei@math.msu.edu ytong@msu.edu xiakelin@msu.edu