HHS Public Access

Author manuscript

J Comput Chem. Author manuscript; available in PMC 2016 July 30.

Published in final edited form as:

J Comput Chem. 2015 July 30; 36(20): 1502–1520. doi:10.1002/jcc.23953.

Multidimensional persistence in biomolecular data

Kelin Xia[1] and **Guo-Wei Wei**[1,2,3,*]

[1]Department of Mathematics, Michigan State University, MI 48824, USA

[2]Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA

[3]Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

Abstract

Persistent homology has emerged as a popular technique for the topological simplification of big data, including biomolecular data. Multidimensional persistence bears considerable promise to bridge the gap between geometry and topology. However, its practical and robust construction has been a challenge. We introduce two families of multidimensional persistence, namely pseudo-multidimensional persistence and multiscale multidimensional persistence. The former is generated via the repeated applications of persistent homology filtration to high dimensional data, such as results from molecular dynamics or partial differential equations. The latter is constructed via isotropic and anisotropic scales that create new simiplicial complexes and associated topological spaces. The utility, robustness and efficiency of the proposed topological methods are demonstrated via protein folding, protein flexibility analysis, the topological denoising of cryo-electron microscopy data, and the scale dependence of nano particles. Topological transition between partial folded and unfolded proteins has been observed in multidimensional persistence. The separation between noise topological signatures and molecular topological fingerprints is achieved by the Laplace-Beltrami flow. The multiscale multidimensional persistent homology reveals relative local features in Betti-0 invariants and the relatively global characteristics of Betti-1 and Betti-2 invariants.

Keywords

Multidimensional peristence; Multifiltration; Anisotropic filtration; Multiscale persitence; Portein folding; Protein flexibility; Topological denoising

1 Introduction

The rapid progress in science and technology has led to the explosion in biomolecular data. The past decade has witnessed a rapid growth in gene sequencing. Vast sequence databases are readily available for entire genomes of many bacteria, archaea and eukaryotes. The human genome decoding that originally took 10 years to process can be achieved in a few days nowadays. The Protein Data Bank (PDB) updates new structures on a daily basis and has accumulated more than one hundred thousand tertiary structures. The availability of

*Address correspondences to Guo-Wei Wei. wei@math.msu.edu.

these structural data enables the comparative study of evolutionary processes, gene-sequence based protein homology modeling of protein structures and the decryption of the structure-function relationship. The abundant protein sequence and structural information makes it possible to build up unprecedentedly comprehensive and accurate theoretical models. One of ultimate goals is to predict protein functions from known protein sequences and structures, which remains a fabulous challenge.

Fundamental laws of physics described in quantum mechanics (QM), molecular mechanism (MM), continuum mechanics, statistical mechanics, thermodynamics, etc. underpin most physical models of biomolecular systems. QM methods are indispensable for chemical reactions, enzymatic processes and protein degradations.[1, 2] MM approaches are able to elucidate the conformational landscapes of proteins.[3] However, both QM and MM involve an excessively large number of degrees of freedom and their application to real-time large-scale protein dynamics becomes prohibitively expensive. For instance, current computer simulations of protein folding take many months to come up with a very poor copy of what Nature administers perfectly within a tiny fraction of a second. One way to reduce the number of degrees of freedom is to employ time-independent approaches, such as normal mode analysis (NMA),[4–7] flexibility-rigidity index (FRI)[8, 9] and elastic network model (ENM),[10] including Gaussian network model (GNM)[11–13] and anisotropic network model (ANM).[14] Another way is to incorporate continuum descriptions in atomistic representation to construct multiscale models for large biological systems.[1, 2, 15–19] Implicit solvent models are some of the most popular approaches for solvation analysis.[20–29] Recently, differential geometry based multiscale models have been proposed for biomolecular structure, solvation, and transport.[30–33] The other way is to combine several atomic particles into one or a few pseudo atoms or beads in coarse-grained (CG) models.[34–37] This approach is efficient for biomolecular processes occurring at slow time scales and involving large length scales.

All of the aforementioned theoretical models share a common feature: they are geometry based approaches[38–40] and depend on geometric modeling methodologies.[41] Technically, these approaches utilize geometric information, namely, atomic coordinates, angles, distances, areas[40, 42, 43] and sometimes curvatures[44–46] as well as physical information, such as charges and their locations or distributions, for the mathematical modeling of biomolecular systems. Indeed, there is an increased importance in geometric modeling for biochemistry,[39] biophysics[47, 48] and bioengineering.[49, 50] Nevertheless, geometry based models are typically computationally expensive and become intractable for biomolecular processes such as protein folding, signal transduction, transcription and translation. Such a failure is often associated with massive data acquisition, curation, storage, search, sharing, transfer, analysis and visualization. The challenge originated from geometric modeling call for game-changing strategies, revolutionary theories and innovative methodologies.

Topological simplification offers an entirely different strategy for big data analysis. Topology deals with the connectivity of different components in a space and is able to classify independent entities, rings and higher dimensional holes within the space. Topology captures geometric properties that are independent of metrics or coordinates. Indeed, for many biological problems, including the opening or closing of ion channels, the association or disassociation of ligands, and the assembly or disassembly of proteins, it is the qualitative

topology, rather than the quantitative geometry that determines physical and biological functions. Therefore, there is a topology-function relationship in many biological processes[51] such that topology is of major concern.

In contrast to geometric tools which are frequently inundated with too much structural information to be computationally practical, Topological approaches often incur too much reduction of the geometric information. Indeed, a coffee mug is topologically equivalent to a doughnut. Therefore, topology is rarely used for quantitative modeling. Persistent homology is a new branch of topology that is able to bridge the gap between traditional geometry and topology and provide a potentially revolutionary approach to complex biomolecular systems. Unlike computational homology which gives rise to truly metric free or coordinate free representations, persistent homology is able to embed additionally geometric information into topological invariants via a filtration process so that “birth” and “death” of isolated components, circles, rings, loops, voids or cavities at all geometric scales can be measured.[52–54] As such, the filtration process create a multiscale representation of important topological features. Mathematically these topological features are described by simplicial complexes, i.e., topological spaces constructed by points, line segments, triangles, and their higher-dimensional counterparts. The basic concept of persistent homology was introduced by Frosini and Landi[55] and Robins,[56] independently. The first realization was due to Edelsbrunner et al.[52] The concept was genearlized by Zomorodian and Carlsson.[53] Many efficient computational algorithms have been proposed in the past decade.[57–61] Many methods have been developed for the geometric representation and visualization of topological invariants computed from persistent homology. Among them, the barcode representation[62] utilizes various horizontal line segments or bars to describe the “birth” and “death” of homology generators over the filtration process. Additionally, persistent diagram representation directly displays topological connectivity in the filtration process. The availability of efficient persistent homology tools[63, 64] has led to applications in a diverse fields, including image analysis,[65–68] image retrieval,[69] chaotic dynamics,[70, 71] complex network,[72, 73] sensor network,[74] data analysis,[75–79] computer vision,[67] shape recognition,[80] computational biology,[51, 81–83] and nano particles.[84, 85]

The most successful applications of persistent homology have been limited to topological characterization identification and analysis (CIA). Indeed, there is little persistent homology based physical or mathematical modeling and quantitative prediction in the literature. Recently, we have introduced persistent homology as unique means for the quantitative modeling and prediction of nano particles, proteins and other biomolecules.[51, 84] Molecular topological fingerprint (MTF), a recently introduced concept,[51] is utilized not only for the CIA, but also for revealing topology-function relationships in protein folding and protein flexibility. Persistent homology is found to provide excellent prediction of stability and curvature energies for hundreds of nano particles.[84, 85] More recently, we have proposed a systematical variational framework to construct objective-oriented persistent homology (OPH),[85] which is able to proactively extract desirable topological traits from complex data. An example realization of the OPH is achieved via differential geometry and Laplace-Beltrami flow.[85] Most recently, we have developed persistent homology based topological denoising method for noise removal in volumetric data from cryo-electron microscopy

(cryo-EM).[86] We have shown that persistent homology provides a powerful tool for solving ill-posed inverse problems in cryo-EM structure determination.[86]

However, one dimensional (1D) persistent homology has its inherent limitations. It is suitable for relatively simple systems described by one or a few parameters. The emergence of complexity in self-organizing biological systems frequently requires more comprehensive topological descriptions. Therefore, multidimensional persistent homology, or multidimensional persistence, becomes valuable for biological systems as well as many other complex systems. In principle, multidimensional persistence should be able to seamlessly bridge geometry and topology. Although multidimensional persistence bears great promise, its construction is non-trivial and elusive to the scientific community.[87] A major obstacle is that, theoretically, it has been proved there is no complete discrete representation for multidimensional persistent module analogous to one dimensional situation.[87] State differently, the persistent barcodes or persistent diagram representation is only available in one dimension filtration, no counterparts can be found in higher dimensions. Therefore, in higher dimensional filtration, incomplete discrete invariants that are computable, compact while still maintain important persistent information, are being considered.[87] Among them, a well-recognized one is persistent Betti numbers (PBNs),[52] which simply displays the histogram of Betti numbers over the filtration parameter. The PBN is also known as rank invariant[87] and size functions (0th homology).[55] A major merit of the PBN representation is its equivalent to the persistent barcodes in one dimension, which means that this special invariant is complete in 1D filtration. Also, it has been proved that PBN is stable in the constraint of certain marching distance.[88] A few mathematical algorithms have been proposed.[88–90] Multi-filtration has been used in pattern recognition or shape comparison.[55, 91, 92] Computationally, the realization of robust multidimensional persistent homology remains a challenge as algorithms proposed have to be topologically feasible, computationally efficient and practically useful.

The objective of this work is to introduce two classes of multidimensional persistence for biomolecular data. One class of multidimensional persistence is generated by repeated applications of 1D persistent homology to high-dimensional data, such as those from protein folding, molecular dynamics, geometric partial differential equations (PDEs), varied signal to noise ratios (SNRs), etc. The resulting high-dimensional persistent homology is a pseudo-multidimensional persistence. Another class of multidimensional persistence is created from a family of new simplicial complexes associated an isotropic scale or anisotropic scales. In general, scales behave in the same manner as wavelet scales do. They can focus on the certain features of the interest and/or defocus on undesirable characteristics. As a consequence, the proposed scale based isotropic and anisotropic filtrations give rise to new multiscale multidimensional persistence. We demonstrate the application of the proposed multidimensional persistence to a number of biomolecular and/or molecular systems, including protein flexibility analysis, protein folding characterization, topological denoising, noise removal from cryo-EM data, and analysis of fullerene molecules. Our multidimensional filtrations are carried out on three types of data formats, namely, point cloud data, matrix data and volumetric data. Therefore, the proposed methods can be easily applied to problems in other disciplines that involve similar data formats.

Our algorithm for multidimensional persistence is robust and straightforward. In a two-dimensional (2D) filtration, we fix one of the filtration parameters and perform the filtration on the second parameter to obtain PBNs. Then we systematically change the fixed parameter to sweep over its whole range, and stack all the PBNs together. This idea can be directly applied to three dimensional (3D) and higher dimensional filtrations. Essentially, we just repeat the 1D filtration over and over until the full ranges of other parameters are sampled. The PBNs are then glued together. This multidimensional persistent homology method can be applied to any other high dimensional data. In this work, point cloud data and matrix data are analyzed by using the JavaPlex.[63] Volumetric data are processed with the Perseus.[64]

The rest of this paper is organized as follows. In Section 2, we explore the multidimensional persistence in point cloud data for protein folding. We model the protein unfolding process by an all-atom steer molecular dynamics (SMD). We consider both an all-atom representation and a coarse-grained representation to analyze the SMD data. From our multifiltration analysis, it is found that PBNs associated with local hexagonal and pentagonal ring structures in protein residues are preserved during the unfolding process while those due to global rings and cavities diminish. Coarse-grained representation is able to directly capture the dramatic topological transition during the unfolding process. In Section 3, we investigate the multidimensional persistence in matrix data. The GNM Kirchhoff (or connectivity) matrix and FRI correlation matrix are analyzed by multidimensional persistent homology. The present approach is able to predict the optimal cutoff distance of the GNM and the optimal scale of the FRI algorithm for protein flexibility analysis. Section 4 is devoted to the multidimensional persistence in volumetric data. We analyze the multidimensional topological fingerprints of Gaussian noise and demonstrate the multidimensional topological denoising of synthetic data and cryo-EM data in conjugation with the Laplace-Beltrami flow method. Finally, we construct multiscale 2D and 3D persistent homology methods to analyze the intrinsic topological patterns of protein 2YGD and fullerene C_{60} molecule. This paper ends with a conclusion.

2 Multidimensional persistence in the point cloud data of protein folding

In this section, we reveal multidimensional persistence in point cloud data associated with protein folding process. It is commonly believed that after the translation from mRNA, unfolded polypeptide or random coil folds into a unique 3D structure which defines the protein function.[93] However, protein folding does not always lead to a unique 3D structure. Aggregated or misfolded proteins are often associated with sporadic neurodegenerative diseases, such as mad cow disease, Alzheimer's disease and Parkinson's disease. Currently, there is no efficient means to characterize disordered proteins or disordered aggregation, which is crucial to the understanding of the molecular mechanism of degenerative disease. In this section, we show that multidimensional persistence provides an efficient tool to characterize and visualize the orderliness of protein folding.

2.1 Protein folding/unfolding processes

The SMD is commonly used to generate elongated protein configurations from its nature state.[94–96] Our goal is to examine the associated changes in the protein topological invariants induced by SMD. There are three approaches to achieve SMD: high temperature,

constant force pulling, and constant velocity pulling.[94–96] Both implicit and explicit molecular dynamics can be used for SMD simulations. The mechanical properties of protein FN-III_{10} has been utilized to carefully design and valid SMD. Appropriate treatment of solvent environment in the implicit SMD is crucial. Typically, a large box which can hold the stretched protein is required, although the computational cost is relatively high.[97] In our study, a popular SMD simulation tool NAMD is employed to generate the partially folded and unfolded protein conformations. The procedure consists of two steps: the relaxation of the given structure and unfolding simulation with constant velocity pulling. In the first step, the protein structure is downloaded from the Protein Data Bank (PDB), which is the major reservoir for protein structures with atomic details. Then, the structure is prepared through the standard procedure, including adding missed hydrogen atoms, before it is solvated with a water box which has an extra 5Å layer, comparing with the initial minimal box that barely hold the protein structure.[98] The standard minimization and equilibration processes are carried out. We employ a total of 5000 time steps of equilibration iterations with the periodic boundary condition after 10000 time steps of initial energy minimization. In our simulations, we use a time increment of 2 femtoseconds (fs). We set SMDk=7. The results are recorded after each 50 time steps, i.e., one frame for each 0.1 picosecond (ps). We accumulate a total of 1000 frames or protein configurations, which are employed for our persistent homology filtration.

2.2 All-atom and coarse-grained representations

Persistent homology analysis of proteins can be carried out either in an all-atom representation or in CG representations.[51] For the all-atom representation, various types of atoms, including O, C, N, S, P, etc., are all included and regarded as equally important in our computation. We deliberately ignore the Hydrogen atoms in our structure during the filtration analysis, as we found that they tend to contaminate our local protein fingerprints. The all-atom representation gives an atomic description of a given protein frame or configuration and is widely used in molecular dynamic simulation. In contrast, CG representations describe the protein molecule with the reduced number of degrees of freedom and are able to highlight important protein structure features. CG representations can be constructed in many ways. A standard coarse-grained representation of proteins used in our earlier topological analysis is to represent each amino acid by the corresponding C_α atom.[51] CG representations are efficient for describing large proteins and protein complexes and significantly reduce the cost of calculating topological invariants.[51]

Figure 1 demonstrates the persistence information for the all-atom representation and the C_α coarse-grained representation of 1UBQ relaxation structure (i.e., the initial structure for the unfolding process). Figures 1**a** and **b** are topological invariants from the all-atom representation without hydrogen atoms. In Fig. 1**a**, it can be observed that β_1 and β_2 barcodes are clearly divided into two unconnected regions: local region (from 1.6 to 2.7 Å for β_1 and from 2.4Å to 2.7Å for β_2) and global region (from 2.85Å to 6.7Å for β_1 and from 3.5Å to 6.7Å for β_2). Local region appears first during the filtration process and it is directly related to the hexagonal ring (HR) and pentagonal ring (PR) structures from the residues. As indicated in the zoomed-in regions enclosed by dotted red rectangles, there are 7 local β_1 bars and 3 β_2 bars, which are topological fingerprints for phenylalanine (one HR), tyrosine

(one HR), tryptophan (one HR and one PR), proline (one PR), and histidine (one PR). In Fig. 2**a**, we have 3 hexagonal rings (red color) corresponding to 3 local β_1 bars and 3 local β_2 bars. It is well known that hexagonal structures produce β_2 invariants in the Vietoris–*C*Rips complex based filtration.[51] The other 4 local β_1 bars are from pentagonal structures (blue color). Figures 2 **c** and **d** are results from the coarse-grained representation. It can be seen that there is barely any β_2 information for the initial structure. As protein unfolds, almost no cavities or holes are detected. Therefore, we only consider β_0 and β_1 invariants in the coarse-grained model.

2.3 Multidimensional persistence in protein folding process

In our protein folding analysis, we extract 1000 configurations over the unfolding process. For each configuration, we carry out the point cloud filtration, i.e., systematically increasing the radius of ball associated with each atom, and come up with three 1D PBN graphs for β_0, β_1, and β_2. We then stack 1000 PBN graphs of the same type, say all β_0 graphs, together. In this way, the final result can be stored in a 2D matrix with the row number indicating the filtration radius, the column number indicating the configuration, and the elements are PBN values. Figures 2 **a, b** and **c** demonstrate the unfolding of protein 1UBQ in the all atom representation without hydrogen atoms and the corresponding 2D persistence diagrams. In these subfigures, we highlight residual pentagonal rings and hexagonal rings in blue and red, respectively. These ring structures correspond to the local topological invariants as indicated in Fig. 1**a**. Figures 2 **d, e** and **f** depict 2D persistent homology analysis of the protein unfolding process. Because all the bond lengths are around 1.5 to 2.0 Å and do not change during the unfolding process, the 2D β_0 persistence shown in Fig. 2 **d** is relatively simple and consistent with the top panel in Fig. 1**b**. The 2D β_1 persistence shown in Fig. 2 **e** is very interesting. The local rings occurred from 1.6 to 2.7 Å are due to pentagonal and hexagonal structures in the residues and are persistent over the unfolding process. However, the numbers of β_1 invariants for global rings in the region from 2.85 to 6.7 Å vary dramatically during the unfolding process. Essentially, the SMD induced elongation of the polypeptide structure reduces the number of rings. Finally, the behavior of the β_2 invariants in Fig. 2 **f** is quite similar to that of the β_1. The local β_2 invariants occurred from 2.4 to 2.7 Å induced by the hexagonal structures[51] remain unchanged during the unfolding process, while the number of global β_2 invariants occurred from 3.5 to 6.7 Å rapidly decreases during the unfolding process. Especially, at 750th configuration and beyond, the number of β_2 invariants in global region has plummeted. The related PBNs for β_2 drop to zero abruptly, indicates that the protein has become completely unfolded. Indeed, there is an obvious topological transition in multidimensional β_1 persistence around 750th configuration as shown in Fig. 2 **e**. The global β_1 PBNs are dramatically reduced and their distribution regions are significantly narrowed for all configurations beyond 75 picosecond simulations, which is an evidence for solely intraresidue β_1 rings.

Having analyzed the multidimensional persistence in protein folding via the all-atom representation, it is interesting to further explore the same process and data set in the coarse-grained representation. Figure 3 illustrates our results. In Fig. 3 **a**, protein 1UBQ is plotted with all atoms except for hydrogen atoms. We use different colors to label different types of residues. The same structure is illustrated by the C_α based CG representation in Fig. 3 **b**. An

advantage of the CG model is that it simplifies topological relations by ignoring intraresidue topological invariants, while emphasizing interresidue topological features. Figures 3 **c** and **d** respectively depict 2D β_0 and β_1 invariants of the protein unfolding process. Compared with the all-atom results in Figs. 2 **d** and **e**, there are some unique properties. First, the CG analysis only emphasizes the global topological relations among residues and their evolution during the protein unfolding. Additionally, the 2D β_0 profile of the all-atom representation was a strict invariant over the time evolution as shown in Fig. 2 **d**, while that of CG model in Fig. 3 **c** varies obviously during the SMD simulation. The standard mean distance between two adjacent C_α atoms is about 3.8 Å, which can be enlarged under the pulling force of the SMD. The deviation from the mean residue distance indicates the strength of the pulling force. Finally, Fig. 3 **d** displays a clear topological transition from a partially folded state to a completely unfolded state at 75 picoseconds or 750th configuration.

As demonstrated in our earlier work,[51] one can establish a quantitative model based on the PBNs of β_1 to predict the relative folding energy and stability. The β_1 PBNs computed from the present CG representation are particularly suitable for this purpose. A similar quantitative model can be established to describe the orderliness of disordered proteins.[51] In Figure 4, we demonstrate the prediction of bond and total energies using β_0 and β_1 accumulated bar lengths, respectively. Basically, the PBNs for each individual configuration are added to deliver the accumulated bar lengths, which are then used to fit with the simulated results in a total of 1000 frames. It can be seen that the accumulated bar lengths of β_0 give a nice prediction of the bond energy, and the accumulated bar lengths of β_0 capture the essential properties of the total energy. For these two fittings, Pearson's correlation coefficients are 0.924 and 0.990, respectively. It can been seen these topological measurements capture the essential properties of the bond and total energies, and thus can be used to characterize the unfolding process.

In summary, multidimensional persistent homology analysis provides a wealth of information about protein folding and/or unfolding process including the number of atoms or residues, the numbers of hexagonal rings and pentagonal rings in the protein, bond lengths or residue distances, the strength of applied pulling force, the orderliness of disordered proteins, the relative folding energies, and topological translation from partially folded states to completely unfolded states. Therefore, multidimensional topological persistence is a powerful new tool for describing protein dynamics, protein folding and protein-protein interaction.

3 Multidimensional persistence in biological matrices

Having illustrated the construction of multidimensional topological persistence in point cloud data, we further demonstrate the development of multidimensional topological analysis of matrix data. To this end, we consider biomolecular matrices associated flexibility analysis. The proposed method can be similarly applied to other biological matrices.

3.1 Protein flexibility prediction

Geometry, electrostatics, and flexibility are some of the most important properties for a protein that determine its functions. The role of protein geometry and electrostatics has been

extensively studied in the literature. However, the importance of protein flexibility is often overlooked. An interesting argument is that it is the protein flexibility, not disorder, that is intrinsic to molecular recognition.[99] Protein flexibility can be defined as its ability to deform from the equilibrium state under external force. The external stimuli are omnipresent either in the cellular environment and in the lattice condition. In response, protein spontaneous fluctuations orchestrate with the Brownian dynamics in living cells or lattice dynamics in solid with its degree of fluctuations determined by both the strength of external stimuli and protein flexibility. It has been shown that the Gaussian network model (GNM) and the flexibility-rigidity index (FRI) are some of most successful methods for protein flexibility analysis.[8, 9] However, the performance of these methods depends on their parameters, namely, the cutoff distance of the GNM and the characteristic distance or the scale of the FRI. In this work, we develop matrix based multidimensional persistent homology methods to examine the optimal scale of FRI and optimal cutoff distance of the GNM. Brief descriptions are given to both methods to facilitate our persistent homology analysis.

Flexibility rigidity index—The FRI have been proposed as a matrix diagonalization free method for the flexibility analysis of biomolecules.[8, 9] The computational complexity of the fast FRI constructed by using the cell lists algorithm is of $O(N)$, with N being the number of particles.[9] In FRI, protein topological connectivity is measured by a correlation matrix. Consider a protein with N particles with coordinates given by $\{\mathbf{r}_j|\mathbf{r}_j \in \mathbb{R}^3, j = 1, 2, \cdots, N\}$. We denote $\|\mathbf{r}_i - \mathbf{r}_j\|$ the Euclidean distance between ith particle and the jth particle. For the ith particle, its correlation matrix element with the jth particles is given by $\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \sigma_j)$, where σ_j is the scale depending on the particle type. The correlation matrix element is a real-valued monotonically decreasing function satisfying

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|;\sigma_j)=1 \text{as} \|\mathbf{r}_i - \mathbf{r}_j\| \to 0 \quad (1)$$

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|;\sigma_j)=0 \text{as} \|\mathbf{r}_i - \mathbf{r}_j\| \to \infty. \quad (2)$$

The Delta sequences of the positive type discussed in an earlier work[100] are suitable choices. For example, one can select generalized exponential functions

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|;\sigma_j)=e^{-(\|\mathbf{r}_i-\mathbf{r}_j\|/\sigma_j)^\kappa}, \kappa>0 \quad (3)$$

and generalized Lorentz functions

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|;\sigma_j)=\frac{1}{1+(\|\mathbf{r}_i - \mathbf{r}_j\|/\sigma_j)^\upsilon}, \upsilon>0. \quad (4)$$

We have defined the atomic rigidity index μ_i for the ith particle as[8]

$$\mu_i=\sum_j^N w_j \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|;\sigma_j), \forall i=1, 2, \cdots, N. \quad (5)$$

where w_j is a particle type dependent weight function. The the atomic rigidity index has a straight forward physical interpretation, i.e., a strong connectivity leads to a high rigidity.

We also defined the atomic flexibility index as the inverse of the atomic rigidity index,

$$f_i=\frac{1}{\mu_i}, \forall i=1,2,\cdots,N. \quad (6)$$

The atomic flexibility indices $\{f_i\}$ are used to predict experimental B-factors or Debye-Waller factors via a linear regression.[8] The FRI theory has been intensively validated by a set of 365 proteins.[8, 9] It outperforms the GNM in terms of accuracy and efficiency.[8]

When we only consider one type of particles, say C_α atoms in a protein, we can set $w_j = 1$. Additionally, it is convenient to set $\sigma_j = \sigma$ for C_α based CG model. We use σ as a scale parameter in our multidimensional persistent homology analysis, which leads to a 2D persistent homology.

Elastic network model—The normal mode analysis (NMA)[4–7] is a well developed technique and is constructed based on the matrix diagonalization of MD force field. It can be employed to study, understand and characterize the mechanical aspects of the long-time scale dynamics. The computational complexity for the matrix diagonalization is typically of $O(N^3)$, where N is the number of matrix rows or particles. Elastic network model (ENM)[10] simplifies the MD force field by considering only the elastic interactions between nearby pairs of atoms. The Gaussian network model (GNM)[11–13] makes a further simplification by using the coarse-grained representation of a macromolecule. This coarse-grained representation ensures the computational efficiency. Yang et al.[101] have demonstrated that the GNM is about one order more efficient than most other matrix diagonalization based approaches. In fact, GNM is more accurate than the NMA.[9] It should be noticed that the GNM models can be further improved by the incorporation of information from crystalline structure, residual types, and co-factors.

The performance of GNM depends on its cutoff distance parameter, which allows only the nearby neighbor atoms within the cutoff distance to be considered in the elastic Hamiltonian. In this work, we construct multidimensional persistent homology based on the cutoff distance in the GNM. We further analyze the parameter dependence of the GNM by our 2D persistence.

3.2 Persistent homology analysis of optimal cutoff distance

Protein elastic network models, including the GNM, usually employ the coarse-grained representation and do not distinguish between different residues. Let us denote N the total number of C_α atoms in a protein, and $\|\mathbf{r}_i - \mathbf{r}_j\|$ the distance between ith and jth C_α atoms. To analyze the topological properties of protein elastic networks, we have introduced a new distance matrix $\mathbf{D} = \{D_{ij} | i = 1, 2, \cdots, N; j = 1, 2, \cdots, N\}$[51]

$$D_{ij}=\begin{cases} \|\mathbf{r}_i-\mathbf{r}_j\|, & \|\mathbf{r}_i-\mathbf{r}_j\|\le r_c; \\ d_\infty, & \|\mathbf{r}_i-\mathbf{r}_j\|>r_c, \end{cases} \quad (7)$$

where d_∞ is a sufficiently large value which is much larger than the final filtration size and r_c is a given cutoff distance. Here d_∞ is designed to ensure that atoms beyond the cutoff

distance r_c do not form any high order simplicial complex during the filtration process. Therefore, the resulting persistent homology shares the same topological connectivity with elastic network models. By systematically increasing the cutoff distance r_c, one can analyze the topological connectivity and performance of the GNM. Additionally, the cutoff distance (r_c) in Eq. (7) is also employed as the filtration parameter in our 2D persistent homology analysis of the GNM.

The performance of the GNM for the B-factor prediction and the multidimensional persistent homology analysis of protein 1PZ4 are plotted in Fig. 5. In Fig. 5 **a**, we compare the experimental B-factors and those predicted by the GNM with a cutoff distance 6.6 Å. The Pearson correlation coefficient for the prediction is 0.89. The GNM provides very good predictions except for the first three residues and the high flexibility around the 42nd residue. Figure 5 **b** shows the relation between correlation coefficient and cutoff distance. It can be seen that the largest correlation coefficients are obtained in the region when cutoff distance is in the range of 6Å to 9Å. Figures 5 **c** and **d** illustrate 2D β_0 and β_1 persistence, respectively. The *x*-axises are the cutoff distance r_c in filtration matrix (7), which is the major filtration parameter. The *y*-axises are the cutoff distance r_c in the GNM Kirchhoff matrix. The resulting β_0 and β_1 PBNs in the matrix representation have unique patterns which are highly symmetric along the diagonal lines. This symmetry, to a large extent, is duo to the way of forming the GNM Kirchhoff matrix. The 2D β_0 persistence has an obvious interpretation in terms of 113 residues. Interestingly, patterns in Fig. 5 **d** can be employed to explain the behavior of the correlation coefficients under different cutoff distances. To this end, we roughly divide Fig. 5 **d** into four regions according to the cutoff distance, i.e., (0Å, 4.5Å), (4.5Å, 5.8Å), (5.8Å, 9Å) and (9Å, 12Å). In the first region, the network is not well constructed. As the distance between two C_α atoms is around 3.8Å, there is only a cluster of isolated atoms when cutoff distance is smaller than 4.5 Å. Therefore, the corresponding GNMs do not give any reasonable prediction. In the second region, network structures begin to form. The number of 1D ring structures within these networks increases dramatically. It reaches its maximum when cutoff is about 5Å, and then drops quickly. This behavior means that many local small-sized loops are developed. The corresponding GNMs can capture certain local properties, however, they neglect the global networks and are unable to grab the essential characteristics of the protein. As a consequence, the correlation coefficients are quite poor. In the third region, constructed networks incorporate more and more large-sized loops or rings and the corresponding GNMs improve predictions. In the last region, local rings disappear while global rings are included in the network models. It is natural to assume that only when the constructed network includes all essential topological invariants that the corresponding GNM delivers the best prediction. However, this assumption turns out to be incorrect. As indicated in Fig. 5 **b**, the largest correlation coefficient is actually in the third region. The best cutoff distances are around 7Å to 9Å. This happens because in the GNM, equal weights are assigned to all elastic springs once spring lengths are within the cutoff distance. Thus, there is no discrimination between local and global ring structures.

3.3 Persistent homology analysis of the FRI scale

Unlike GNM which utilizes a cutoff distance, the FRI theory employs a scale or characteristic distance σ in its correlation kernel. The scale has a similar function as the scale

in wavelet theory, and thus it emphasizes the contribution from the given scale. The FRI scale has a direct impact in the accuracy of protein B-factor prediction. Similar to the optimal cutoff distance in the GNM, the best FRI scale varies from protein to protein, although an optimal value can be found based on a statistical average over hundreds of proteins.[8, 9] In the present work, we use the scale as an additional variable to construct multidimensional persistent homology.

In our recent work, we have introduced a FRI based filtration method to convert the point cloud data into matrix data.[51] In this approach, we construct a new filtration matrix $\mathbf{M} = \{M_{ij} | i = 1, 2, \cdots, N; j = 1, 2, \cdots, N\}$

$$M_{ij} = \begin{cases} 1 - \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|;\sigma), & i \neq j, \\ 0, & i = j, \end{cases} \quad (8)$$

where $0 \leq \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \sigma) \leq 1$ is defined in Eqs. (1) and (2). To avoid any confusion, we simply use the exponential kernel with parameter $\kappa = 2$ in the present work.

The performance of the FRI B-factor prediction and the multidimensional persistence of protein 2MCM are illustrated in Fig. 6. The filtration matrices are constructed as $M_{ij} = 1.0 - e^{-\left(\frac{\|r_i - r_j\|}{\sigma}\right)^2}$. The comparison of experimental B-factors and predicted B-factors with the scale $\sigma = 9.2$Å is given in Fig. 6 **a**. The Pearson correlation coefficient is 0.81 for the prediction. Figure 6 **b**, shows the relation between the correlation coefficient and the scale. It is seen that the largest correlation coefficients are obtained when the scale is in the range of 5Å to 15Å. Figures 6 **c** and **d** demonstrate respectively β_0 and β_1 2D persistence. Unlike the GNM results shown in Fig. 5 where different cutoff distances lead to dramatic changes in network structures, the FRI connectivity shown in Fig. 6 **c** increases gradually as σ increases. For all $\sigma > 3$Å, the maximal β_1 values can reach 40 as shown in Fig. 6 **d**. However, in the region of 5Å< σ <15Å, 1D rings are established over a wide range of the matrix values, which implies a wide range of distances. The balance of the global and local rings gives rise to excellent FRI B-factor predictions.

In fact, a persistent homology based quantitative model can be established in terms of accumulated bar length.[51] Essentially, if all the PBNs are added up at each scale, the accumulated PBNs give rise a good prediction of the optimal scale range. State differently, the plot of the accumulated PBNs versus the scale will have a similar shape as the curve in Fig. 6 **b**.

4 Multidimensional persistence in volumetric data

Volumetric data are widely available in science and engineering. In biology, density information, such as the experimental data from cryo-EM,[86, 102] geometric flow based molecular hypersurface[31, 33, 42, 85] and electrostatic potential,[44, 103] are typically described in volumetric form. These volumetric data can be filtrated directly in terms of isovalues in

persistent homology analysis. Basically, the locations of the same density value form an isosurface. The discrete Morse theory can then be used to generate cell complexes. Additionally, we have developed techniques[51] to convert point cloud data from X-ray crystallography into the volumetric form by using the rigidity function or density in our FRI algorithm.[86] Specifically, the atomic rigidity index μ_i in Eq. (5) can be generalized to a position (**r**) dependent rigidity function or density[8, 9]

$$\mu(\mathbf{r})=\sum_{j=1}^{N} w_j(\mathbf{r})\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\sigma_j). \quad (9)$$

Volumetric multidimensional persistence can be constructed in many different ways. Because w_j and σ_j are 2*N* independent variables, it is feasible to construct 2*N* + 1-dimensional persistence for an *N*-atom biomolecule. Here the additional dimension is due to the filtration over the density $\mu(\mathbf{r})$. If we set $w_j = 1$ and $\sigma_j = \sigma$, we can construct genuine 2D persistence by filtration over two independent variables, i.e., σ and density.

In this work, we also demonstrate the construction of pseudo-multidimensional persistence. Since noise and denoising are two important issues in volumetric data, we develop methods for pseudo-multidimensional topological representation of noise and pseudo-multidimensional topological denoising.

4.1 Multidimensional topological fingerprints and topological denoising

To analyze the topological signature of noise, we make a case study on Gaussian noise, which is perhaps the most commonly occurred noise. The Gaussian white noise is a set of random events satisfying the normal distribution

$$n(t)=\frac{A_n}{\sqrt{2\pi}\sigma_n}e^{-\frac{(t-\mu_n)^2}{2\sigma_n^2}}, \quad (10)$$

where A_n, μ_n and σ_n are the amplitude, mean value and standard deviation of the noise, respectively. The strength of Gaussian white noise can be characterized by the signal to noise ratio (SNR) defined as SNR = μ_s / σ_n, where μ_s is the mean value of signal. We generate noise polluted volumetric data by adding different levels of Gaussian white noise to the original data.

We employ fullerene C_{20} as an example to illustrate the multidimensional topological fingerprints of noise. The rigidity density of C_{20} is given by

$$\mu(\mathbf{r})=\sum_{j=1}^{20} e^{-2\|\mathbf{r}-\mathbf{r}_j\|}. \quad (11)$$

The noisy data and multifiltration results are demonstrated in Fig. 7. We plot the noisy data of C_{20} with three SNRs, 1, 10 and 100 in Figs. 7 **a$_1$–a$_3$**. The persistent barcodes of C_{20} have 20 β_0 bars, 11 β_1 bars and one β_2 bar. Figures 7 **b$_1$–b$_3$** are respectively 2D β_0, β_1 and β_2 persistent homology. In these figures, the vertical axises are the SNR values, which are varied over the range of 1.0 to 100.0. The horizontal axises represent the density isovalues

(i.e., the main filtration parameter). In these cases, the designed filtration goes from the highest density value around 2.0 to the lowest density about −1.0. The negative values are introduced by the Gaussian noise. The resulting PBNs are plotted in the natural logarithm scale as indicated by the color bars.

First of all, the topological fingerprints of C_{20} stand out in Figs. 7 **b_1–b_3** and demonstrate some invariant features as the SNR increases. In Figure 7 **b_1**, the rectangle-like region is due to the twenty isolated parts in C_{20}. Similarly, the rectangle-like region in Figs. 7 **b_2** and **b_3** represents the 12 rings and the central void of the C_{20} structure. These rectangle patterns are the intrinsic topological fingerprints of C_{20}. In Figs. 7 **b_1–b_3**, noise topological signatures dominate the counts of Betti numbers, particularly when the SNR is smaller than 30. For example, β_2 spectrum near the density value of 0.4 is essentially indistinguishable from noise induced cavities.

We have recently proposed topological denoising as a new strategy for topology-controlled noise reduction of synthetic, natural and experimental data.[86] Our essential idea is to couple noise reduction with persistent homology analysis. Since persistent topology is extremely sensitive to the noise, the strength of noise signature can be monitored by persistent homology in a denoising process. As a result, one can make optimal decisions on number of deniosing iterations. It was found that contrary to popular belief, noise can have very long lifetimes in the barcode representation,[86] while short lived features are part of molecular topological fingerprints.[51] In the present work, we introduce 2D topological denoising methods. To this end, we present a brief review of the Laplace-Beltrami flow based denoising approach.

Laplace-Beltrami flow—One of efficient approaches for noise reduction in signals, images and data is geometric analysis, which combines differential geometry and differential equations. The resulting geometric PDEs have become very popular in applied mathematics and computer science in the past two decades.[104–106] Wei introduced some of the first families of high-order geometric PDEs for image analysis[107]

$$\frac{\partial u(\mathbf{r},t)}{\partial t} = -\sum_q \nabla \cdot \mathbf{j}_q + e(u(\mathbf{r},t), |\nabla u(\mathbf{r},t)|, t), q = 0, 1, 2, \cdots \quad (12)$$

where the nonlinear hyperux term $\mathbf{j}_q$ is given by

$$\mathbf{j}_q = -d_q(u(\mathbf{r},t), |\nabla u(\mathbf{r},t)|, t)\nabla\nabla^{2q}u(\mathbf{r},t), q = 0, 1, 2, \cdots \quad (13)$$

where $\mathbf{r} \in \mathbb{R}^n$, $\nabla = \frac{\partial}{\partial \mathbf{r}}$, $u(\mathbf{r}, t)$ is the processed signal, image or data, $d_q(u(\mathbf{r}, t), |\nabla u(\mathbf{r}, t)|, t)$ are edge or gradient sensitive diffusion coefficients and $e(u(\mathbf{r}, t), |\nabla u(\mathbf{r}, t)|, t)$ is a nonlinear operator. Denote $X(\mathbf{r})$ the original noise data and set the initial input $u(\mathbf{r}, 0) = X(\mathbf{r})$. There are many ways to choose hyperdiffusion coefficients $d_q(u, |\nabla u|, t)$ in Eq. (13). For example, one can use the exponential form

$$d_q(u(\mathbf{r},t),|\nabla u(\mathbf{r},t)|,t)=d_{q0}\exp\left[-\frac{|\nabla u|^\kappa}{\sigma_q^\kappa}\right],\kappa>0, \quad (14)$$

where d_{q0} is chosen as a constant with value depended on the noise level, and σ_0 and σ_1 are local statistical variance of u and ∇u

$$\sigma_q^2(\mathbf{r})=\overline{|\nabla^q u-\overline{\nabla^q u}|^2}(q=0,1). \quad (15)$$

Here the notation $\overline{Y(\mathbf{r})}$ represents the local average of $Y(\mathbf{r})$ centered at position $\mathbf{r}$. The existence and uniqueness of high-order geometric PDEs were investigated in the literature.[108–111] Recently, we have proposed differential geometry based objective oriented persistent homology to enhance or preserve desirable traits in the original data during the filtration process and then automatically detect or extract the corresponding topological features from the data.[85] From the point of view of signal processing, the above high order geometric PDEs are designed as low-pass filters. Geometric PDE based high-pass filters was pioneered by Wei and Jia by coupling two nonlinear geometric PDEs.[112] Recently, this approach has been generalized to a new formalism, the PDE transform, for signal, image and data analysis.[40, 113–115]

Apart from their application to images,[107, 116, 117] high order geometric PDEs have also been modified for macromolecular surface formation and evolution,[43]

$$\frac{\partial S}{\partial t}=(-1)^q\sqrt{g(|\nabla\nabla^{2q}S|)}\nabla\cdot\left(\frac{\nabla(\nabla^{2q}S)}{\sqrt{g(|\nabla\nabla^{2q}S|)}}\right)+P(S,|\nabla S|), \quad (16)$$

where S is the hypersurface function, $g(|\nabla\nabla^{2q}S|) = 1+|\nabla\nabla^{2q}S|^2$ is the generalized Gram determinant and P is a generalized potential term. When $q = 0$ and $P = 0$, a Laplace-Beltrami equation is obtained,[42]

$$\frac{\partial S}{\partial t}=|\nabla S|\nabla\cdot\left(\frac{\nabla S}{|\nabla S|}\right). \quad (17)$$

We employe this Laplace-Beltrami equation for the noise removal in this work.

Computationally, the finite different method is used to discretize the Laplace-Beltrami equation in 3D. Suitable time interval δt and grid spacing h are required to ensure the stability and accuracy. To avoid confusion and control the noise reduction process systematically, we simply ignore the voxel spacing in different data sets and employ a set of unified parameters of $\delta t = 5.0E – 6$ and $h = 0.01$ in our computation. The intensity of noise reduction is then described by the duration of time integration or the number of iterations of Eq. (17).

Topological fingerprint identification—From Figures 7 **$\mathbf{b_1}$–$\mathbf{b_3}$**, it can been seen that, with the increase of SNR, the intrinsic topological properties emerge and persist. Persistent

patterns can be seen in the PBN representation. It is interesting to know whether the topological persistence of the signal is a feature in the denoising process.

Figures 7 **c_1–c_3** depict the topological invariants of contaminated fullerene C_{20} over the Laplace-Beltrami flow based denoising process. The fullerene C_{20} rigidity density is generated by using Eq. (11). The noise is added according to Eq. (10) with the SNR of 1.0. The Laplace Beltrami equation (17) is solved with time stepping $\delta t = 5.0E - 6$ and spatial spacing $h = 0.01$. Figures 7 **c_1–c_3** illustrate respectively the β_0, β_1 and β_2 persistent homology analysis of the denoising process. The filtration goes from density 2.0 to −1.0 (the negative values are due to the added noise). A total of 200 denoising iterations is applied to the noisy data. The PBNs are plotted in the natural logarithm scale. It can be seen that after about 40 denoising iterations, the noise intensity has been reduced dramatically. Indeed, the intrinsic topological features of C_{20} emerge and persist. It appears that the bandwidths of C_{20} PBNs reduce during the denoising process. However, such a bandwidth reduction is due to the fact that there is a dramatic density reduction during the denoising precess, particularly at the early stage of the denoising. In fact, the accumulated Betti numbers of C_{20} do not change and stay stable. It should be noted the color bar denotes the natural logarithm of PBNs values added by 1. The comparison between Figures 7 **b_1–b_3** and **c_1–c_3** demonstrates clearly the noise reduction effect in various iteration steps. It provides a criteria to distinguish between the intrinsic topological properties and noise in denoising process.

Having demonstrated the construction of 2D persistence for topological denoising, we further apply this new technique for the analysis of noisy cryo-EM data of a microtubule (EMD 1129).[102] Figures 8 **a, b**, and **c** are surfaces extracted from denoising data with the numbers of iterations of 1, 100 and 200, respectively. A common isovalues of 15.0 used to extract surfaces in these plots. It is seen that the denoising process reduces not only the noise, but also the density, which leads to the shift in the topological distribution. Figures 8 **d, e** and **f** are respectively the 2D β_0, β_1 and β_2 persistence. The filtrations in horizontal axises go from density 45 to 0. In Figures 8 **d, e** and **f**, vertical axises are the numbers of iterations. A total of 300 iterations is employed for integrating Eq. (10) with time stepping $\delta t = 2.0E - 6$ and spatial spacing $h = 0.01$. Color bar values represent the natural logarithm of PBNs. It can be seen that after about 100 denoising iterations, the noise intensity has been dramatically reduced. Persistent behavior can be observed in β_0, β_1 and β_2. This persistent behavior is a manifest of the intrinsic topological features of the micortubule structure.

4.2 Multiscale multidimensional persistence

In this section, we demonstrate the construction of multiscale multidimensional persistent homology. To this end, we consider protein 2YGD in our multiscale 2D persistence analysis. The fullerene C_{60}, whose topological properties have been analyzed in our earlier work,[51] is used as an example to illustrate our multiscale high-dimensional persistence.

Multiscale 2D persistence—We generate volumetric density data of protein 2YGD by using the exponential kernel function

$$\mu(\mathbf{r})=\sum_{j=1}^{N} w_j e^{-\frac{\|\mathbf{r}-\mathbf{r}_j\|}{\sigma}}, \quad (18)$$

where the resolution σ is utilized as a multiscale parameter and will be varied from 0.7Å to 14.7Å. Weight w_j is chosen as the atomic number of the jth atom. We linearly rescale the density value to region [0,1] using expression $\mu(\mathbf{r})^s=\frac{\mu(\mathbf{r})}{\mu_{\max}}$. Here $\mu(\mathbf{r})^s$ is the rescaled density value. Here $\mu_{\max}$ is the largest density value in the original data. For each given scale, we carry out the density value based filtration of protein 2YGD. Our results are depicted in Fig. 9. The structure of protein 2YGD is plotted in Fig. 9 **a**.

The structure of protein 2YGD exhibits dramatically different scales ranging from atom, residue, secondary-structure, domain to entire protein. Figure 9 illustrates the topological representation of this multiscale structure. Generally speaking, we can roughly divide results of β_0, β_1 and β_2 into three parts according to the resolution parameter σ. The first part is when σ is smaller than 3 Å. In this region, the topological properties related to the local structures, i.e., atoms or intra-residues, are well captured. The second part is the region when σ is larger than 3 Å and smaller than 7 Å. With the increase of the resolution value, local structures gradually disappear, more global type of structures, i.e., inter-residual and domain, begins to emerge. The rest region belongs to the third part, in which, only the global backbone structure of the protein 2YGD is captured. We can seen that the PBNs in this region are comparably consistent. In β_0, we have 4 individual components corresponding to the four major domains in the protein. In β_1, the PBNs are majorly 9 and 4, representing the 6 large ring and 4 small ring pattern in the structure. Finally the PBN is 1 for β_2, this captures the central void in the protein.

Multiscale high-dimensional persistence—Having demonstrated the construction of 2D topological persistence in a number of ways, we pursue to the development of 3D persistence. Obviously, there are a variety of ways that one can construct 3D or multidimensional persistent homology. For example, 3D persistent homology can be generated by the combination of scale, time and the matrix filtration, the combination of scale, time and density filtration, and the combination of scale, SNR and density filtration. In the present work, we illustrate 3D persistent homology by using anisotropic scales or anisotropic filtrations, which give rise to truly multidimensional simplicial complexes and truly multidimensional persistent homology. For simplicity, we take fullerene C_{60} as an example to illustrate our approach.

We define the density of the fullerene C_{60} by a multiscale function,

$$\mu(\mathbf{r})=\sum_{j=1}^{60} \frac{1.0}{1.0+\sqrt{\left(\frac{x-x_j}{\sigma_j^x}\right)^2+\left(\frac{y-y_j}{\sigma_j^y}\right)^2+\left(\frac{z-z_j}{\sigma_j^z}\right)^2}}, \quad (19)$$

where (x_j, y_j, z_j) are the atomic coordinates of C_{60} molecule and σ_j^x, σ_j^y and σ_j^z are 180 independent scales. Obviously, each of these scales can vary independently. Therefore,

together with the density, these scales are able to deliver 181-dimensional filtrations. However, the visualization of such a high-dimensional persistent homology will be a problem, not to mention its physical meaning. To reduce the dimensionality, we set $\sigma_j^x=\sigma^x$, $\sigma_j^y=\sigma^y$ and $\sigma_j^z=\sigma^z$, which leads to four-dimensional (4D) persistent homology. To further reduce the dimensionality, we set $\sigma^x = \sigma^y$ to end up with 3D persistence.

Unlike the isotropic filtration created by an isotropic scale, the anisotropic filtration creates a family of distorted "molecules" for topological analysis. For the highly symmetry C_{60} molecule, these distorted versions are not very physical by themselves. However, C_{60} is a good choice for illustrating and analyzing our methodology, because any distortion is due to the method. On the other hand, the method itself is meaningful due to the fact that most molecules are not symmetric and have anisotropic shapes or anisotropic thermal fluctuations. Figure 10 depicts anisotropic C_{60} molecules generated by different combinations of $\sigma^x = \sigma^y$ and σ^z according to Eq. (19). Figures 10 **a** and **b** are obtained with $\sigma^x = \sigma^y = 0.2$ Å and $\sigma^z = 0.5$ Å at the isovalue of 0.4. There is an elongation along the *z* axis. Figures 10 **c** and **d** are generated with $\sigma^x = \sigma^y = 0.5$ Å and $\sigma^z = 0.2$ Å at the isovalue of 1.0. In this case, there is an obvious compression in the *z*-direction.

Topologically, the anisotropic filtration systematically creates a family of truly multidimensional simplicial complexes which would be difficult to imagine otherwise in the 3D space. Figure 11 illustrates the multiscale 3D persistent homology of C_{60} molecule. The molecular structure is presented in Fig. 11 **a** with $\sigma^x = \sigma^y = \sigma^z = 0.5$ Å at the isovalue of 1.5. For the 2D persistent homology, the variation of PBNs over two axises can be represented by different color schemes. However, the visualization of PBNs in 3D is not trivial. Figures 11 **b, c** and **d** are respectively multiscale 3D β_0, β_1 and β_2 persistence. Here the *x*-axises represent the density value (i.e., the main filtration parameter). The *y*-axises denote σ^z and the *z*-axises are for $\sigma^x = \sigma^y$. The distributions of two PBNs, $\beta_0 = 4$ and $\beta_0 = 50$ are plotted with blue dots and red dots respectively in Fig. 11 **b**. It is seen that PBNs of β_0 are mainly distributed at small *fix* and σ^z scales. In Fig. 11 **c**, we depict the distributions of $\beta_1 = 3$ and $\beta_1 = 20$ with blue dots and red dots, respectively. As the scales increase, the PBNs of β_1 first increase then decay. Finally, the distributions of $\beta_2 = 1$ and $\beta_2 = 2$ are illustrated with blue dots and red dots, respectively in Fig. 11 **d**. As the cavity of C_{60} is relatively global, the values of $\beta_2 = 1$ is seen to locate at relatively large scales.

5 Conclusion

Recently, persistent homology, a new branch of topology, has gained considerable popularity for computational application in big data simplification. It generates a one-parameter family of topological spaces via filtration such that topological invariants can be measured at a variety of geometric scales. As a result, persistent homology is able to bridge the gap between geometry and topology. However, one-dimensional (1D) persistent homology has its limitation to represent high dimensional complex data. Multidimensional persistence, a generalization of 1D persistent homology to a multidimensional one, provides a new promise for big data analysis. Nevertheless, the realization and construction of robust multidimensional persistence have been a challenge.

In this work, we introduce two types of multidimensional persistence. The first type is called pseudo-multidimensional persistence, which is generated by the repeated applications of 1D persistent homology to high-dimensional data, such as results from molecular dynamics simulation, partial differential equations (PDEs), molecular surface evolution, video data sets, etc. The other type of multidimensional persistence is constructed by appropriate multifiltration processes. Specifically, cutoff distance and scale are introduced as new filtration variables to create multifiltration and multidimensional persistence. The scale of flexibility-rigidity index (FRI)[8, 9] behaves in the same manner as the wavelet scale. It serves as an independent filtration variable and controls the formation of simplicial complexes and the corresponding topological spaces. As a result, the FRI scale creates truly multiscale multidimensional persistent homology, in conjugation with the matrix value variable or the density variable. We have developed genuine two-dimensional (2D) persistent homology. By using anisotropic scales, in which the scale in each spatial direction can vary independently, we can construct four dimensional (4D) persistent homology. A protocol is prescribed for the construction of arbitrarily high dimensional persistence. Concrete numerical example is given to three-dimensional (3D) persistence.

We have demonstrated the utility, established the robustness and explored the efficiency of the proposed multidimensional persistence by its applications to a wide range of biomolecular systems. First, we have constructed pseudo-multidimensional persistence for the protein unfolding process. It is shown that local topological features such as pentagonal and hexagonal rings in the amino acid residues are preserved during the unfolding process, whereas global topological invariants diminish over the unfolding process. Topological transition from folded or partially folded proteins to unfolded proteins can be clearly identified in the 2D persistence. We show that the β_0 persistence also provides an indication of the strength of applied pulling forces in the steer molecular dynamics. Additionally, we have analyzed the optimal cutoff distance of the Gaussian network model (GNM) and the optimal scale of the FRI theory by using 2D persistence. We have revealed the relationship between the topological connectivity in terms of Betti numbers and the performance of the GNM and the FRI for the prediction of protein Debye-Waller factors. Moreover, we have utilized 2D persistence to illustrate the topological signature of Gaussian noise. The efficiency of Laplace-Beltrami flow based topological denoising is studied by the present 2D persistence. We show that the topological invariants of C_{20}, especially β_2, persist during the denoising process, whereas the topological invariants of noising diminish during the denoising process. Similar results are also observed for the topological denoising of cryo-electron microscopy (cryo-EM) data. Finally, we have employed multiscale multidimensional persistence to investigate the topological behavior of protein 2YGD. We reveal its multiscale structure properties in the our 2D persistence. We also consider the C_{60} over anisotropic scale variations. This study unveils that β_0 invariants are intrinsically local, while β_1 and β_2 invariants are relatively global.

Multidimensional persistence techniques have been developed for three types of data formats, i.e., point cloud data, matrix data and volumetric data. We have also illustrated conversion of point cloud data to matrix and volumetric data via the FRI theory. Therefore, the proposed methodology can be directly applied to other biomolecular systems, biological networks, and diverse other disciplines.

Acknowledgments

This work was supported in part by NSF grants DMS-1160352 and IIS-1302285, NIH Grant R01GM-090208 and MSU Center for Mathematical Molecular Biosciences initiative. The authors acknowledge the Mathematical Biosciences Institute for hosting valuable workshops.

References

1. Cui Q. Combining implicit solvation models with hybrid quantum mechanical/molecular mechanical methods: A critical test with glycine. Journal of Chemical Physics. 2002; 117(10):4720.
2. Zhang Y, Yu H, Qin JH, Lin BC. A microuidic dna computing processor for gene expression analysis and gene drug synthesisn. Biomicrofluidics. 2009; 3(044105)
3. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977; 267:585–590. [PubMed: 301613]
4. Go N, Noguti T, Nishikawa T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. 1983; 80:3696–3700. [PubMed: 6574507]
5. Tasumi M, Takenchi H, Ataka S, Dwidedi AM, Krimm S. Normal vibrations of proteins: Glucagon. Biopolymers. 1982; 21:711–714. [PubMed: 7066480]
6. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 1983; 4:187–217.
7. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 1985; 181(3):423–447. [PubMed: 2580101]
8. Xia KL, Opron K, Wei GW. Multiscale multiphysics and multidomain models — Flexibility and rigidity. Journal of Chemical Physics. 2013; 139:194109. [PubMed: 24320318]
9. Opron K, Xia KL, Wei GW. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. Journal of Chemical Physics. 2014; 140:234105. [PubMed: 24952521]
10. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 1996; 77:1905–1908. [PubMed: 10063201]
11. Flory PJ. Statistical thermodynamics of random networks. Proc. Roy. Soc. Lond. A. 1976; 351:351–378.
12. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design. 1997; 2:173–181. [PubMed: 9218955]
13. Bahar I, Atilgan AR, Demirel MC, Erman B. Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability. Phys. Rev. Lett. 1998; 80:2733–2736.
14. Atilgan AR, Durrell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. 2001; 80:505–515. [PubMed: 11159421]
15. Warshel A, Papazyan A. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. Current Opinion in Structural Biology. 1998; 8(2):211–217. [PubMed: 9631295]
16. Sharp KA, Honig B. Electrostatic interactions in macromolecules - theory and applications. Annual Review of Biophysics and Biophysical Chemistry. 1990; 19:301–332.
17. Tully-Smith DM, Reiss H. Further development of scaled particle theory of rigid sphere fluids. Journal of Chemical Physics. 1970; 53(10):4015–4025.
18. Tian WF, Zhao Shan. A fast ADI algorithm for geometric flow equations in biomolecular surface generations. International Journal for Numerical Methods in Biomedical Engineering. 2014; 30:490–516. [PubMed: 24574191]
19. Geng W, Wei GW. Multiscale molecular dynamics using the matched interface and boundary method. J Comput. Phys. 2011; 230(2):435–457. [PubMed: 21088761]
20. Holst, Michael J. Multilevel Methods for the Poisson-Boltzmann Equation. Urbana-Champaign: University of Illinois at Urbana-Champaign, Numerical Computing Group; 1993.

21. Baker NA. Poisson-Boltzmann methods for biomolecular electrostatics. Methods in Enzymology. 2004; 383:94–118. [PubMed: 15063648]
22. Dong F, Olsen B, Baker NA. Computational methods for biomolecular electrostatics. Methods in Cell Biology. 2008; 84:843–870. [PubMed: 17964951]
23. Boschitsch AH, Fenley MO. Hybrid boundary element and finite difference method for solving the nonlinear Poisson-Boltzmann equation. Journal of Computational Chemistry. 2004; 25(7):935–955. [PubMed: 15027106]
24. Bertonati C, Honig B, Alexov E. Poisson-Boltzmann calculations of nonspecific salt effects on protein-protein binding free energy. Biophysical Journal. 2007; 92:1891–1899. [PubMed: 17208980]
25. Georgescu RE, Alexov EG, Gunner MR. Combining conformational flexibility and continuum electrostatics for calculating pKas in proteins. Biophysical Journal. 2002; 83(4):1731–1748. [PubMed: 12324397]
26. Feig M, Brooks CL III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol. 2004; 14:217–224. [PubMed: 15093837]
27. Rocchia W, Alexov E, Honig B. Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions. J. Phys. Chem. 2001; 105:6507–6514.
28. Chen J, Brooks CL III. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. Physical Chemistry Chemical Physics. 2008; 10:471–481. [PubMed: 18183310]
29. Zhou YC, Feig M, Wei GW. Highly accurate biomolecular electrostatics in continuum dielectric environments. Journal of Computational Chemistry. 2008; 29:87–97. [PubMed: 17508411]
30. Wei GW. Differential geometry based multiscale models. Bulletin of Mathematical Biology. 2010; 72:1562–1622. [PubMed: 20169418]
31. Wei, Guo-Wei; Zheng, Qiong; Chen, Zhan; Xia, Kelin. Variational multiscale models for charge transport. SIAM Review. 2012; 54(4):699–754. [PubMed: 23172978]
32. Chen, Duan; Chen, Zhan; Wei, GW. Quantum dynamics in continuum for proton transport II: Variational solvent-solute interface. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:25–51. [PubMed: 22328970]
33. Wei, Guo-Wei. Multiscale, multiphysics and multidomain models I: Basic theory. Journal of Theoretical and Computational Chemistry. 2013; 12(8):1341006.
34. Rzepiela AJ, Schafer LV, Goga N, Risselada HJ, De Vries AH, Marrink SJ. Software news and update reconstruction of atomistic details from coarse-grained structures. Journal of Computational Chemistry. 2010; 31:1333–1343. [PubMed: 20087907]
35. Smith A, Hall CK. Alpha-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. Proteins. 2001; 44:344–360. [PubMed: 11455608]
36. Ding F, Borreguero JM, Buldyrey SV, Stanley HE, Dokholyan NV. Mechanism for the alpha-helix to beta-hairpin transition. J Am Chem Soc. 2003; 53:220–228.
37. Paci E, Vendruscolo M, Karplus M. Validity of go models: Comparison with a solvent-shielded empirical energy decomposition. Biophys J. 2002; 83:3032–3038. [PubMed: 12496075]
38. Yu ZY, Holst M, Cheng Y, McCammon JA. Feature-preserving adaptive mesh generation for molecular shape modeling and simulation. Journal of Molecular Graphics and Modeling. 2008; 26:1370–1380.
39. Feng, Xin; Xia, Kelin; Tong, Yiying; Wei, Guo-Wei. Geometric modeling of subcellular structures, organelles and large multiprotein complexes. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:1198–1223. [PubMed: 23212797]
40. Zheng Q, Yang SY, Wei GW. Molecular surface generation using PDE transform. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:291–316. [PubMed: 22582140]
41. Sazonov, Igor; Nithiarasu, Perumal. Semi-automatic surface and volume mesh generation for subject-specific biomedical geometries. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:133–157. [PubMed: 25830210]

42. Bates PW, Wei GW, Zhao Shan. Minimal molecular surfaces and their applications. Journal of Computational Chemistry. 2008; 29(3):380–391. [PubMed: 17591718]
43. Bates PW, Chen Z, Sun YH, Wei GW, Zhao S. Geometric and potential driving formation and evolution of biomolecular surfaces. J. Math. Biol. 2009; 59:193–231. [PubMed: 18941751]
44. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models I: Eulerian formulation. J. Comput. Phys. 2010; 229:8231–8258. [PubMed: 20938489]
45. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models II: Lagrangian formulation. J. Math. Biol. 2011; 63:1139–1200. [PubMed: 21279359]
46. Chen Z, Zhao Shan, Chun J, Thomas DG, Baker NA, Bates PB, Wei GW. Variational approach for nonpolar solvation analysis. Journal of Chemical Physics. 2012; 137(084101)
47. Feng X, Xia KL, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules II: lagrangian representation. Journal of Computational Chemistry. 2013; 34:2100–2120. [PubMed: 23813599]
48. Xia KL, Feng X, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules i: Cartesian representation. Journal of Computational Physics. 2014; 275:912–936.
49. Boileau E, Bevan RLT, Sazonov I, Rees MI, Nithiarasu P. Flow-induced atp release in patient-specific arterial geometries - a comparative study of computational models. International Journal for Numerical Methods in Engineering. 2013; 29:1038–1056.
50. Mikhal J, Kroon DJ, Slump CH, Geurts BJ. Flow prediction in cerebral aneurysms based on geometry reconstruction from 3d rotational angiography. International Journal for Numerical Methods in Engineering. 2013; 29:777–805.
51. Xia KL, Wei GW. Persistent homology analysis of protein structure, flexibility and folding. International Journal for Numerical Methods in Biomedical Engineerings. 2014; 30:814–844.
52. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. Discrete Comput. Geom. 2002; 28:511–533.
53. Zomorodian A, Carlsson G. Computing persistent homology. Discrete Comput. Geom. 2005; 33:249–274.
54. Zomorodian, Afra; Carlsson, Gunnar. Localized homology. Computational Geometry - Theory and Applications. 2008; 41(3):126–148.
55. Frosini, Patrizio; Landi, Claudia. Size theory as a topological tool for computer vision. Pattern Recognition and Image Analysis. 1999; 9(4):596–603.
56. Robins, Vanessa. Towards computing homology from finite approximations. Topology Proceedings. 1999; 24:503–532.
57. Bubenik, Peter; Kim, Peter T. A statistical approach to persistent homology. Homology, Homotopy and Applications. 2007; 19:337–362.
58. Edelsbrunner, Herbert; Harer, John. Computational topology: an introduction. American Mathematical Soc. 2010
59. Dey TK, Li KY, Sun J, David CS. Computing geometry aware handle and tunnel loops in 3d models. ACM Trans. Graph. 2008; 27
60. Dey, Tamal K.; Wang, YS. Reeb graphs: Approximation and persistence. Discrete and Computational Geometry. 2013; 49(1):46–73.
61. Mischaikow K, Nanda V. Morse theory for filtrations and efficient computation of persistent homology. Discrete and Computational Geometry. 2013; 50(2):330–353.
62. Ghrist R. Barcodes: The persistent topology of data. Bull. Amer. Math. Soc. 2008; 45:61–75.
63. Tausz, Andrew; Vejdemo-Johansson, Mikael; Adams, Henry. Javaplex: A research software package for persistent (co)homology. 2011 Software available at http://code.google.com/p/javaplex.
64. Nanda, Vidit. Perseus: the persistent homology software. Software available at http://www.sas.upenn.edu/~vnanda/perseus.
65. Carlsson G, Ishkhanov T, Silva V, Zomorodian A. On the local behavior of spaces of natural images. International Journal of Computer Vision. 2008; 76(1):1–12.

66. Pachauri D, Hinrichs C, Chung MK, Johnson SC, Singh V. Topology-based kernels with application to inference problems in alzheimer's disease. Medical Imaging, IEEE Transactions on. 2011 Oct; 30(10):1760–1770.
67. Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL. Topological analysis of population activity in visual cortex. Journal of Vision. 2008; 8(8)
68. Bendich, Paul; Edelsbrunner, Herbert; Kerber, Michael. Computing robustness and persistence for images. IEEE Transactions on Visualization and Computer Graphics. 2010; 16:1251–1260. [PubMed: 20975165]
69. Frosini, Patrizio; Landi, Claudia. Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters. 2013; 34:863–872.
70. Mischaikow K, Mrozek M, Reiss J, Szymczak A. Construction of symbolic dynamics from experimental time series. Physical Review Letters. 1999; 82:1144–1147.
71. Kaczynski, T.; Mischaikow, K.; Mrozek, M. Computational homology. Springer-Verlag; 2004.
72. Lee H, Kang H, Chung MK, Kim B, Lee DS. Persistent brain network homology from the perspective of dendrogram. Medical Imaging, IEEE Transactions on. 2012 Dec; 31(12):2267–2277.
73. Horak D, Maletic S, Rajkovic M. Persistent homology of complex networks. Journal of Statistical Mechanics: Theory and Experiment. 2009; 2009(03):P03034.
74. Silva VD, Ghrist R. Blind swarms for coverage in 2-d. Proceedings of Robotics: Science and Systems. 2005:01.
75. Carlsson G. Topology and data. Am. Math. Soc. 2009; 46(2):255–308.
76. Niyogi P, Smale S, Weinberger S. A topological view of unsupervised learning from noisy data. SIAM Journal on Computing. 2011; 40:646–663.
77. Wang, Bei; Summa, Brian; Pascucci, Valerio; Vejdemo-Johansson, M. Branching and circular features in high dimensional data. IEEE Transactions on Visualization and Computer Graphics. 2011; 17:1902–1911. [PubMed: 22034307]
78. Rieck, Bastian; Mara, Hubert; Leitte, Heike. Multivariate data analysis using persistence-based filtering and topological signatures. IEEE Transactions on Visualization and Computer Graphics. 2012; 18:2382–2391.
79. Liu, Xu; Xie, Zheng; Yi, Dongyun. A fast algorithm for constructing topological structure in large data. Homology, Homotopy and Applications. 2012; 14:221–238.
80. Di Fabio, Barbara; Landi, Claudia. A mayer-vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. Foundations of Computational Mathematics. 2011; 11:499–527.
81. Kasson PM, Zomorodian A, Park S, Singhal N, Guibas LJ, Pande VS. Persistent voids a new structural metric for membrane fusion. Bioinformatics. 2007; 23:1753–1759. [PubMed: 17488753]
82. Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V. Topological measurement of protein compressibility via persistence diagrams. preprint. 2013
83. Dabaghian Y, Memoli F, Frank L, Carlsson G. A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput Biol. 2012; 8(8):e1002581. [PubMed: 22912564]
84. Xia KL, Feng X, Tong YY, Wei GW. Persistent homology for the quantitative prediction of fullerene stability. Journal of Computational Chemsitry. 2015; 36:408–422.
85. Wang, Bao; Wei, GW. Objective-oriented Persistent Homology. ArXiv e-prints. 2014 Dec.
86. Xia KL, Wei GW. Persistent topology for cryo-EM data analysis. ArXiv e-prints. 2014 Dec.
87. Carlsson G, Zomorodian A. The theory of multidimensional persistence. Discrete Computational Geometry. 2009; 42(1):71–93.
88. Cerri, A.; Landi, C. Discrete Geometry for Computer Imagery. Springer; 2013. The persistence space in multidimensional persistent homology; p. 180-191.
89. Carlsson, G.; Singh, G.; Zomorodian, A. Algorithms and computation. Springer; 2009. Computing multidimensional persistence; p. 730-739.

90. Cohen-Steiner, D.; Edelsbrunner, H.; Morozov, D. Vines and vineyards by updating persistence in linear time; Proceedings of the twenty-second annual symposium on Computational geometry ACM; 2006. p. 119-126.

91. Biasotti S, De Floriani L, Falcidieno B, Frosini P, Giorgi D, Landi C, Papaleo L, Spagnuolo M. Describing shapes by geometrical-topological properties of real functions. ACM Computing Surveys. 2008; 40(4):12.

92. Cerri A, Fabio B, Ferri M, Frosini P, Landi C. Betti numbers in multidimensional persistent homology are stable functions. Mathematical Methods in the Applied Sciences. 2013; 36(12): 1543–1557.

93. Anfinsen CB. Einfluss der configuration auf die wirkung den. Science. 1973; 181:223–230. [PubMed: 4124164]

94. Paci E, Karplus M. Unfolding proteins by external forces and temperature: The importance of topology and energetics. Proceedings of the National Academy of Sciences. 2000; 97:6521–6526.

95. Hui L, Isralewitz B, Krammer A, Vogel V, Schulten K. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. Biophysical Journal. 1998; 75:662–671. [PubMed: 9675168]

96. Srivastava A, Granek R. Cooperativity in thermal and force-induced protein unfolding: integration of crack propagation and network elasticity models. Phys. Rev. Lett. 2013; 110(138101):1–5.

97. Gao M, Craig D, Vogel V, Schulten K. Identifying unfolding intermediates of $f_n - iii_{10}$ by steered molecular dynamics. J. Mol. Biol. 2002; 323:939–950. [PubMed: 12417205]

98. Xia KL, Wei GW. Molecular nonlinear dynamics and protein thermal uncertainty quantification. Chaos. 2014; 24:013103. [PubMed: 24697365]

99. Janin J, Sternberg MJ. Protein flexibility, not disorder, is intrinsic to molecular recognition. F1000 Biology Reports. 2013; 5(2):1–7. [PubMed: 23361308]

100. Wei GW. Wavelets generated by using discrete singular convolution kernels. Journal of Physics A: Mathematical and General. 2000; 33:8577–8596.

101. Yang LW, Chng CP. Coarse-grained models reveal functional dynamics-I. elastic network models-theories, comparisons and perspectives. Bioinformatics and Biology Insights. 2008; 2:25–45. [PubMed: 19812764]

102. Nogales E, Wang HW. Structural intermediates in microtubule assembly and disassembly: how and why? Current opinion in cell biology. 2006; 18(2):179–184. [PubMed: 16495041]

103. Chen D, Guo WW. Modeling and simulation of electronic structure, material interface and random doping in nano-electronic devices. J. Comput. Phys. 2010; 229:4431–4460. [PubMed: 20396650]

104. Mumford, David; Shah, Jayant. Optimal approximations by piecewise smooth functions and associated variational problems. Communications on Pure and Applied Mathematics. 1989; 42(5):577–685.

105. Willmore, TJ. Riemannian Geometry. USA: Oxford University Press; 1997.

106. Osher S, Sethian JA. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. Journal of computational physics. 1988; 79(1):12–49.

107. Wei GW. Generalized Perona-Malik equation for image restoration. IEEE Signal Processing Lett. 1999; 6:165–167.

108. Greer JB, Bertozzi AL. H-1 solutions of a class of fourth order nonlinear equations for image processing. Discrete and Continuous Dynamical Systems. 2004; 10:349–366.

109. Greer JB, Bertozzi AL. Traveling wave solutions of fourth order pdes for image processingl. SIAM Journal on Mathematics Analysis. 2004; 36:38–68.

110. Xu M, Zhou SL. Existence and uniqueness of weak solutions for a fourth-order nonlinear parabolic equation. Journal of Mathematical Analysis and Applications. 2007; 325(1):636–654.

111. Jin ZM, Yang XP. Strong solutions for the generalized perona-malik equation for image restoration. Nonlinear Analysis-Theory Methods and Applications. 2010; 73:1077–1084.

112. Wei GW, Jia YQ. Synchronization-based image edge detection. Europhysics Letters. 2002; 59(6): 814–819.

113. Wang Y, Wei GW, Yang Si-Yang. Partial differential equation transform – Variational formulation and Fourier analysis. International Journal for Numerical Methods in Biomedical Engineering. 2011; 27:1996–2020. [PubMed: 22207904]

114. Wang Y, Wei GW, Yang Si-Yang. Mode decomposition evolution equations. Journal of Scientific Computing. 2012; 50:495–518. [PubMed: 22408289]

115. Wang Y, Wei GW, Yang Si-Yang. Selective extraction of entangled textures via adaptive pde transform. International Journal in Biomedical Imaging. 2012; 2012 Article ID 958142.

116. Lysaker M, Lundervold A, Tai XC. Noise removal using fourth-order partial differential equation with application to medical magnetic resonance images in space and time. IEEE Transactions on Image Processing. 2003; 12(12):1579–1590. [PubMed: 18244712]

117. Gilboa G, Sochen N, Zeevi YY. Image sharpening by flows based on triple well potentials. Journal of Mathematical Imaging and Vision. 2004; 20(1–2):121–131.

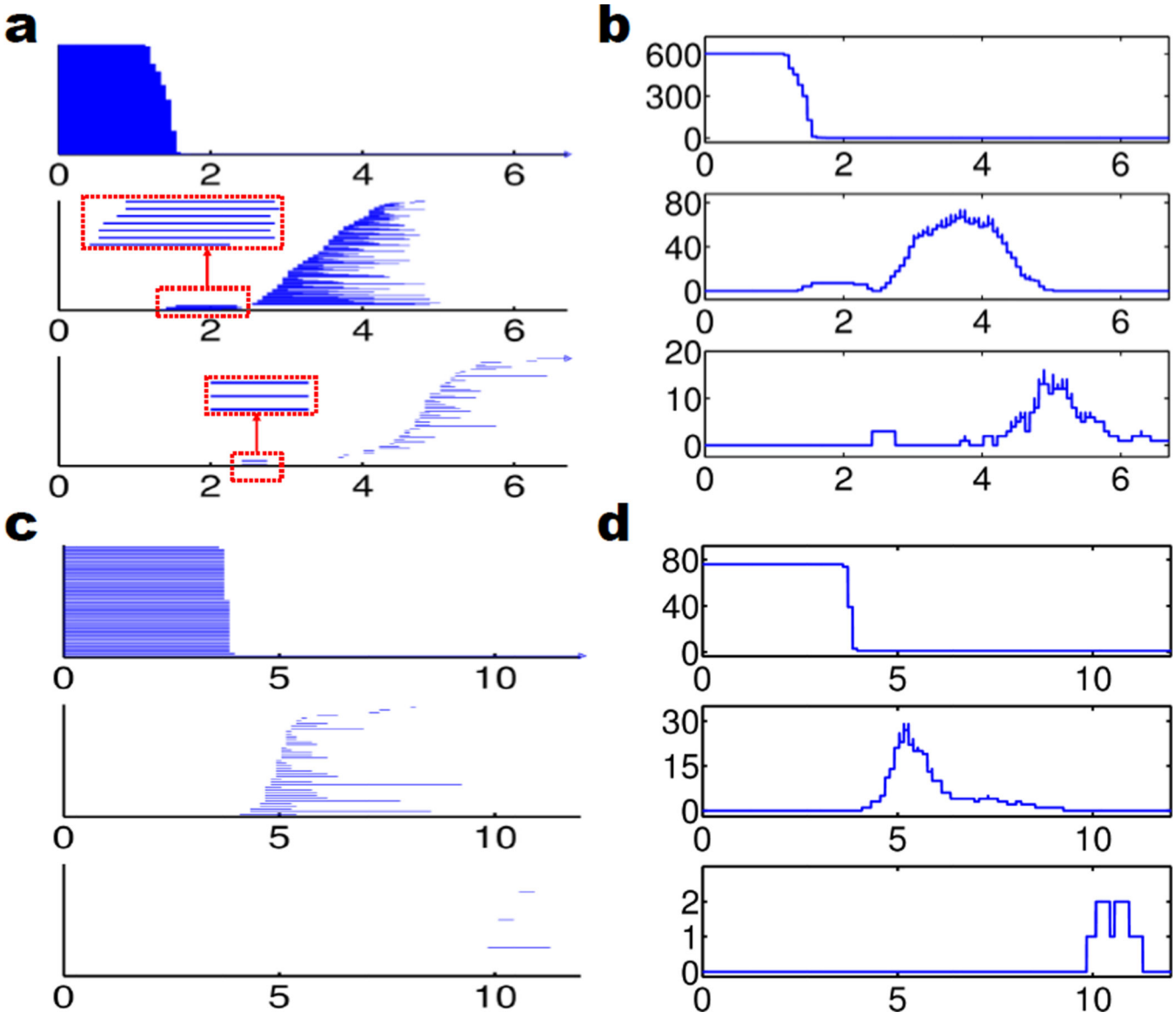

Figure 1.

Persistent barcodes and PBNs of 1UBQ structure. **a** Persistent barcodes for the all-atom representation without hydrogen atoms; **b** PBNs for the all-atom representation without hydrogen atoms; **c** Persistent barcodes for the CG representation; **d** PBNs for the CG representation. In each subfigure, β_0, β_1 and β_2 are displayed in the top, middle and bottom panels, respectively. In all subfigures, horizontal axises label the filtration radius (Å). Vertical axises in **b** and **d** are the numbers of topological invariants.

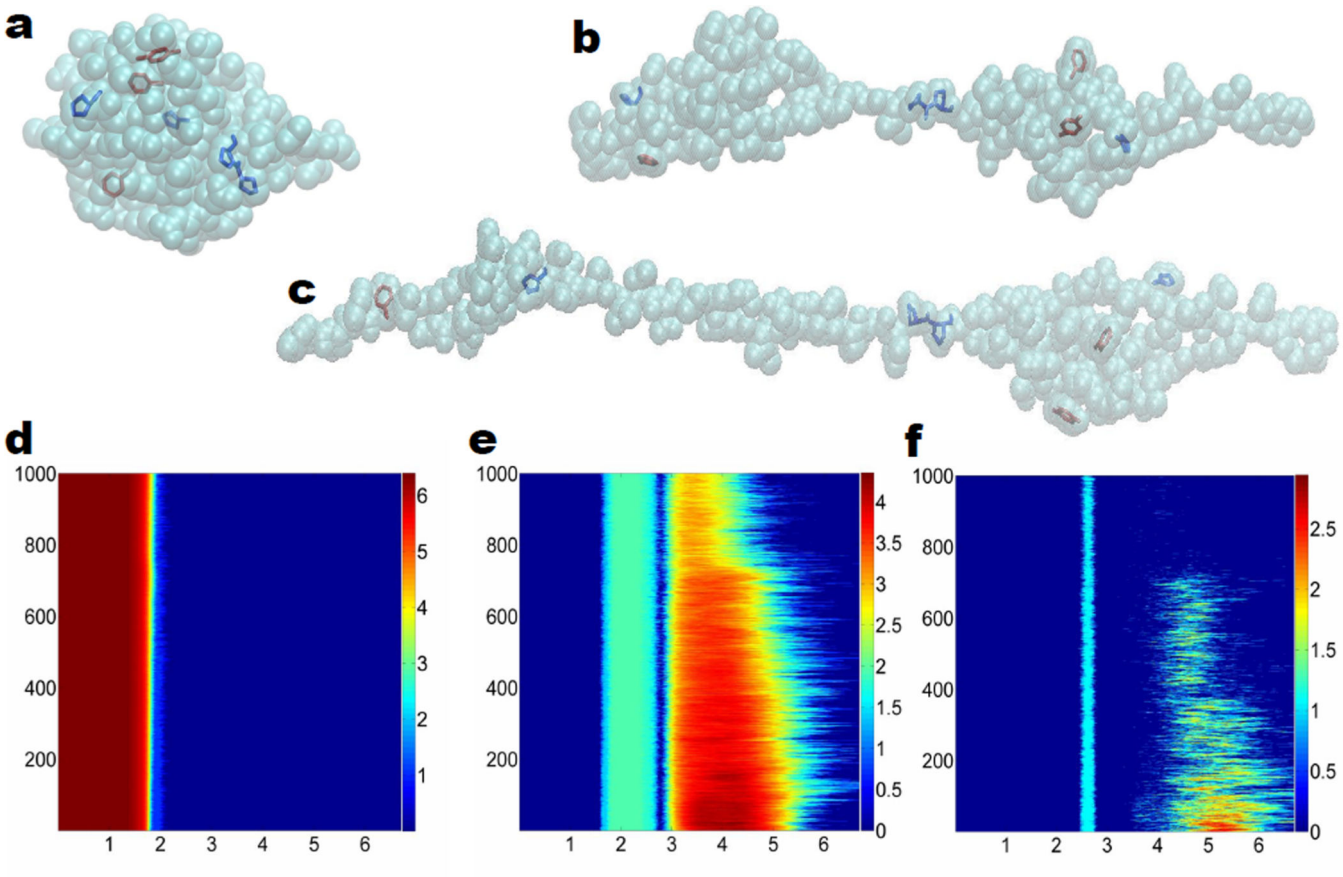

Figure 2.

The unfolding of protein 1UBQ and the corresponding multidimensional persistence. **a** All atom representation of the relaxed structure without hydrogen atoms; **b** All atom representation of the unfolded structure at the 300th frame; **c** All atom representation of the unfolded structure at the 500th frame; **d** 2D β_0 persistence; **e** 2D β_1 persistence; **f** 2D β_2 persistence. In subfigures **d, e** and **f**, horizontal axises label the filtration radius (Å) and the vertical axises are the configuration index. Color bars denote the natural logarithms of PBNs. We systematically add 1 to all PBNs to avoid the possible logarithm of 0.

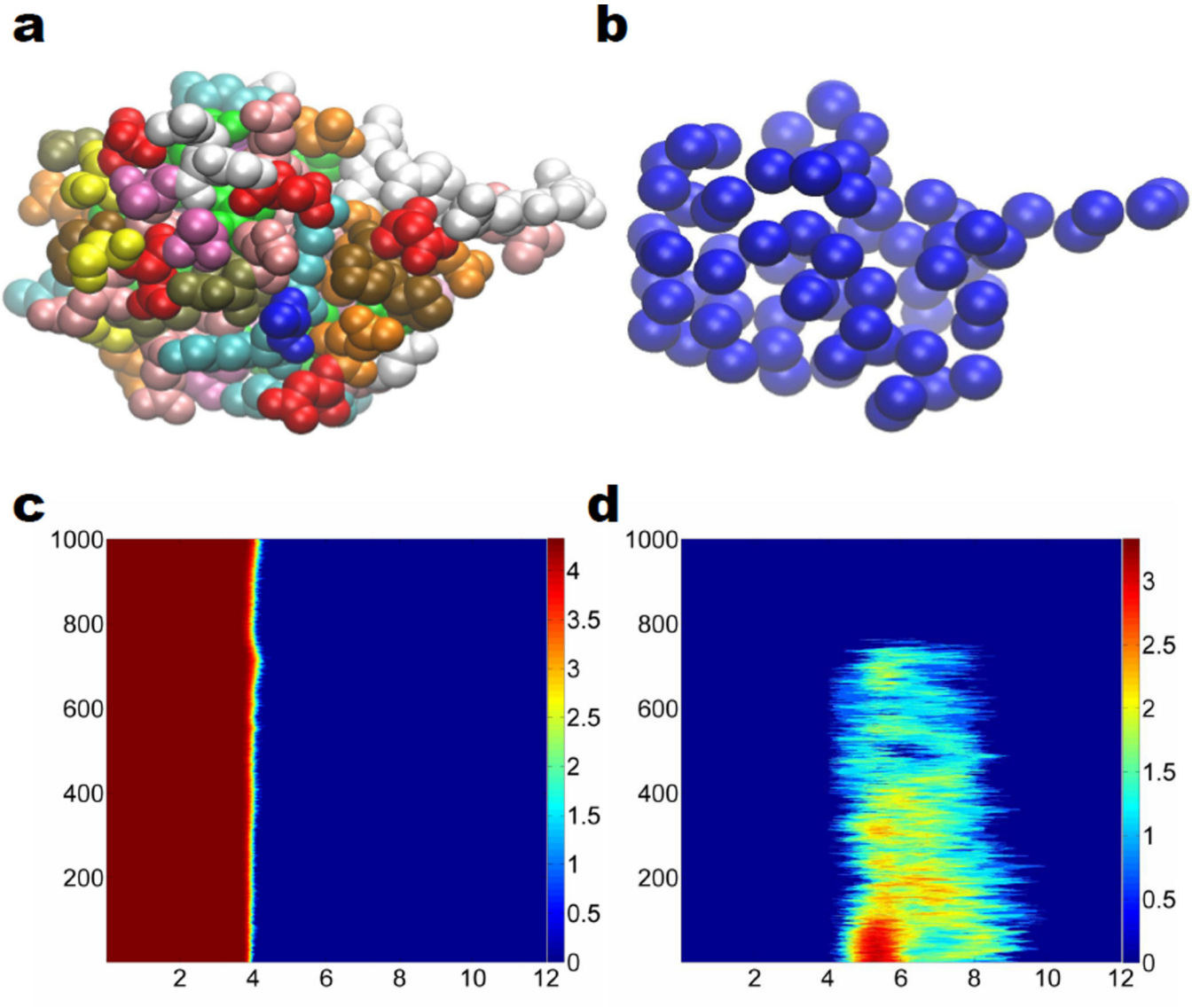

Figure 3.

Coarse-grained representation of the unfolding of protein 1UBQ and the corresponding multidimensional persistence. **a** All atom representation of the relaxed structure without hydrogen atoms; **b** Coarse-grain representation of the relaxed structure without hydrogen atoms; **c** 2D β_0 persistence; **d** 2D β_1 persistence. The color in subfigure **a** denotes different residues. In subfigures **c**, and **d**, horizontal axises label the filtration radius (Å) and the vertical axises are the protein configuration index. Color bars denote the natural logarithms of PBNs. We systematically add 1 to all PBNs to avoid the possible logarithm of 0.

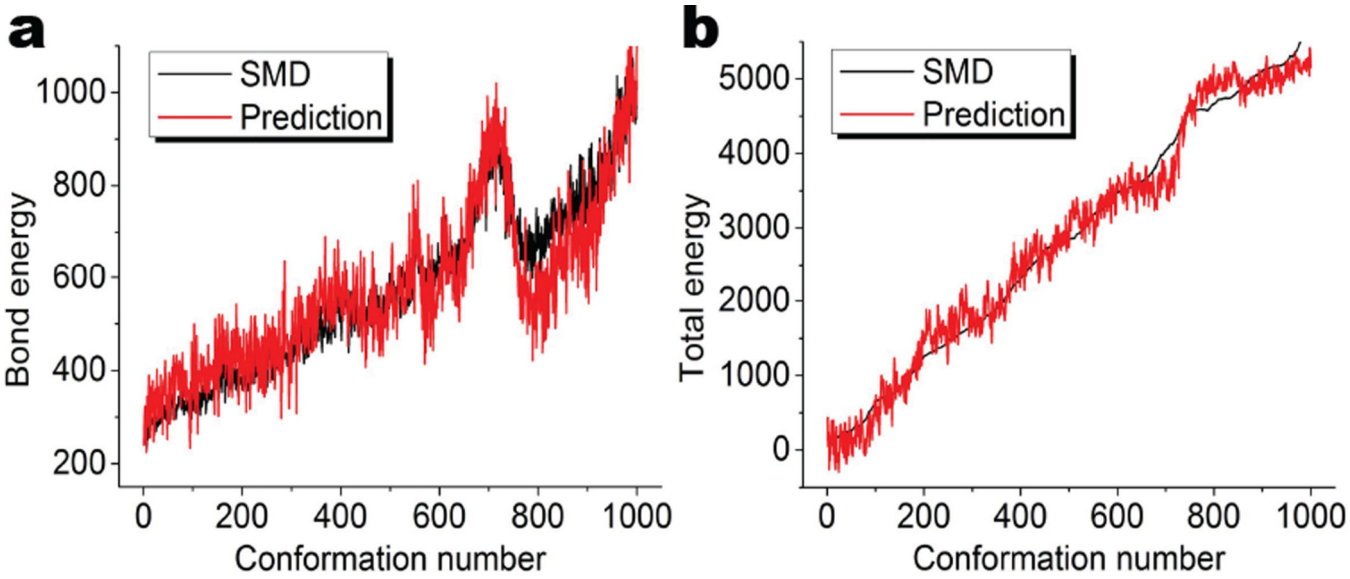

Figure 4.

The prediction of bond and total energy with β_0 and β_1 accumulated bar lengths, respectively. **a** The quantitative comparison of bond energies of β_0 predictions and steered molecular dynamic results. The horizontal axis labels the configuration number and the vertical axis is the bond energy (kcal/mol). The Pearson's correlation coefficient is 0.924. **b** The comparison of total energies of β_1 predictions and steered molecular dynamic results. The horizontal axis labels the configuration number and the vertical axis is the total energy (kcal/mol). The Pearson's correlation coefficient is 0.990. The accumulated bar length for each configuration is calculated by the summation of all the corresponding PBNs for the configuration. It can been seen these topological measurements capture the essential properties of the bond and total energies, and thus can be used to characterize the unfolding process.

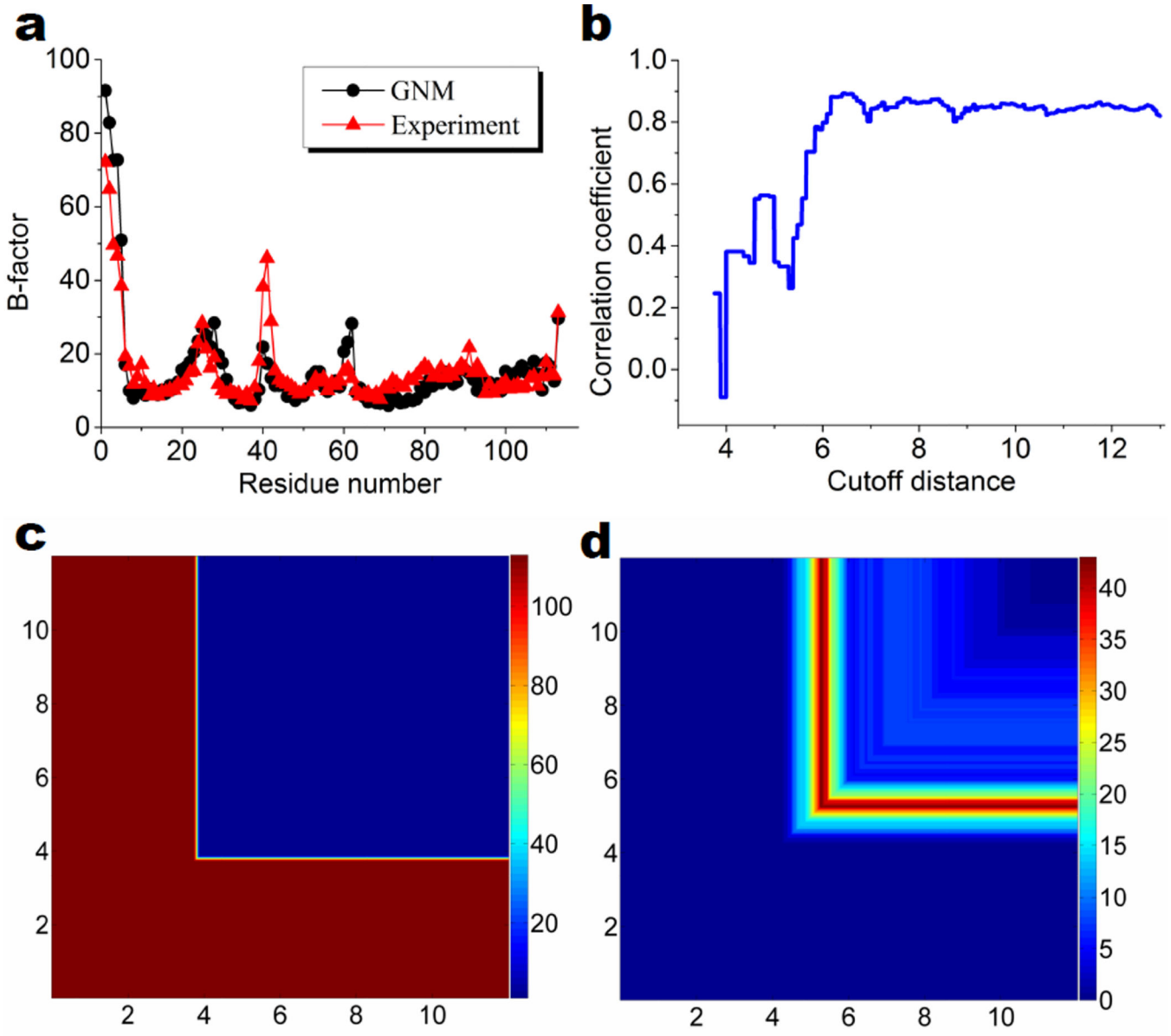

Figure 5.

Performance of GNM and multidimensional persistence of protein 1PZ4. **a** Comparison of the GNM prediction at r_c = 6.6Å and experimental B-factors; **b** Correlation coefficient vs cutoff distance (r_c) for the GNM; **c** 2D β_0 persistence; **d** 2D β_1 persistence. In **c** and **d**, the horizontal axis is the cutoff distance r_c in filtration matrix (7) and the vertical axis is the cutoff distance r_c in the GNM. The color bars represent PBNs.

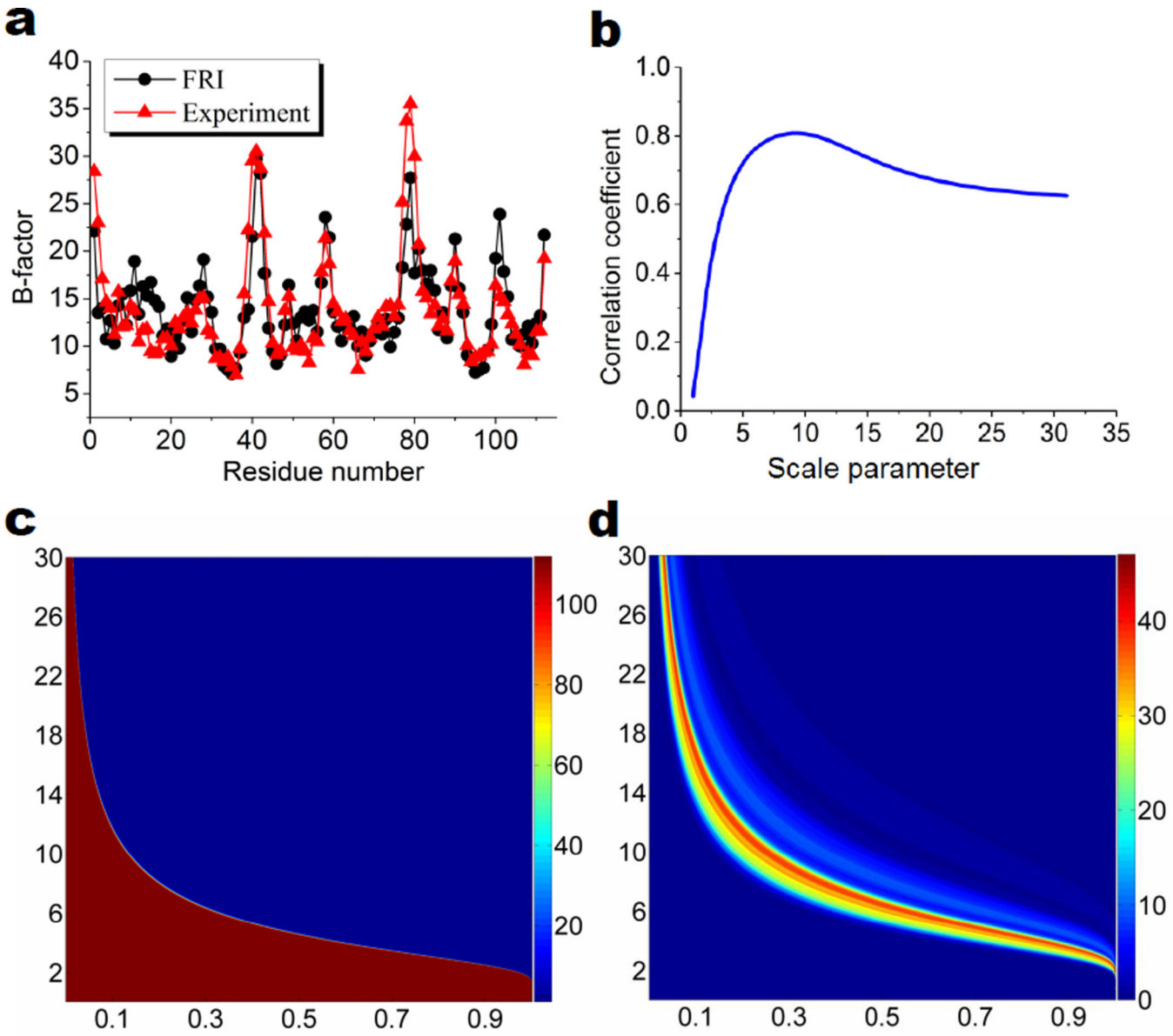

Figure 6.

Performance of the FRI and multidimensional persistence of protein 2MCM. **a** Comparison of the FRI prediction at σ = 9.2Å and experimental B-factors; **b** Correlation coefficient vs scale (σ) for the FRI; **c** 2D β_0 persistence; **d** 2D β_1 persistence. In **c** and **d**, the horizontal axis is in the FRI filtration matrix value M_{ij} (8) and the vertical axis is the scale (σ) in terms of Å in the FRI. The color bars represent PBNs.

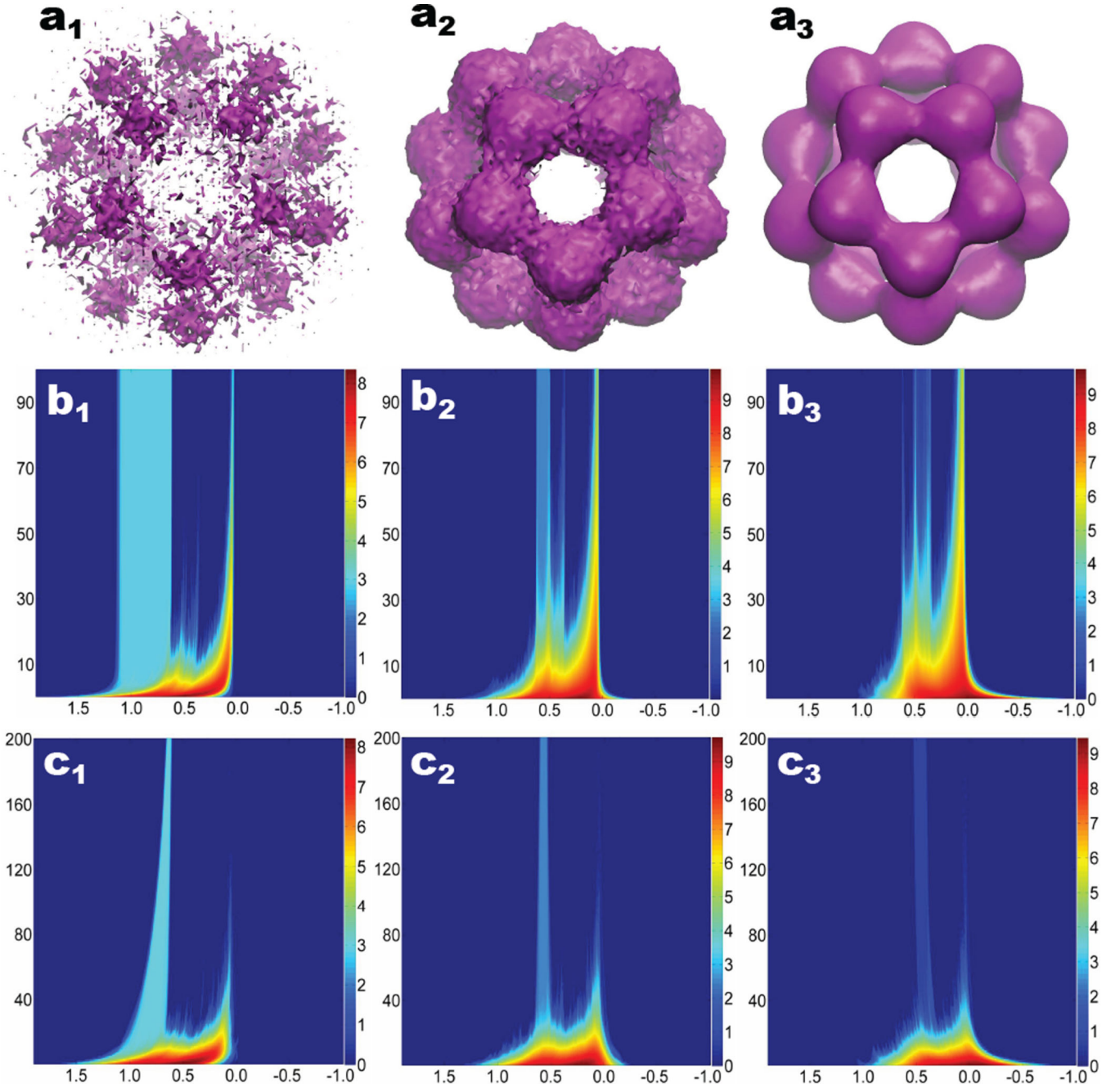

Figure 7.

Illustration of the Gaussian noise contaminated fullerene C_{20} data, their multidimensional persistence and multidimensional topological denoising. **a_1–a_3**: The noisy C_{20} at the SNRs of 1, 10, and 100, respectively; **b_1–b_3**: The 2D persistence representations of β_0, β_1, and β_2 respectively, for Gaussian noise contaminated fullerene C_{20} data. The horizontal axises represent the density values (i.e., the main filtration parameter). The vertical axises are the SNR. Color bars denote the natural logarithm of PBNs. We systematically add 1 to all PBNs to avoid the possible logarithm of 0; **c_1–c_3**: The 2D persistence representations of β_0, β_1, and β_2 respectively, for denoising contaminated fullerene C_{20} data with SNR 1.0. The horizontal

axises represent the density values. The vertical axises represent the number of iterations. A total of 200 iterations is employed. Color bars denote the natural logarithm of PBNs added by 1.

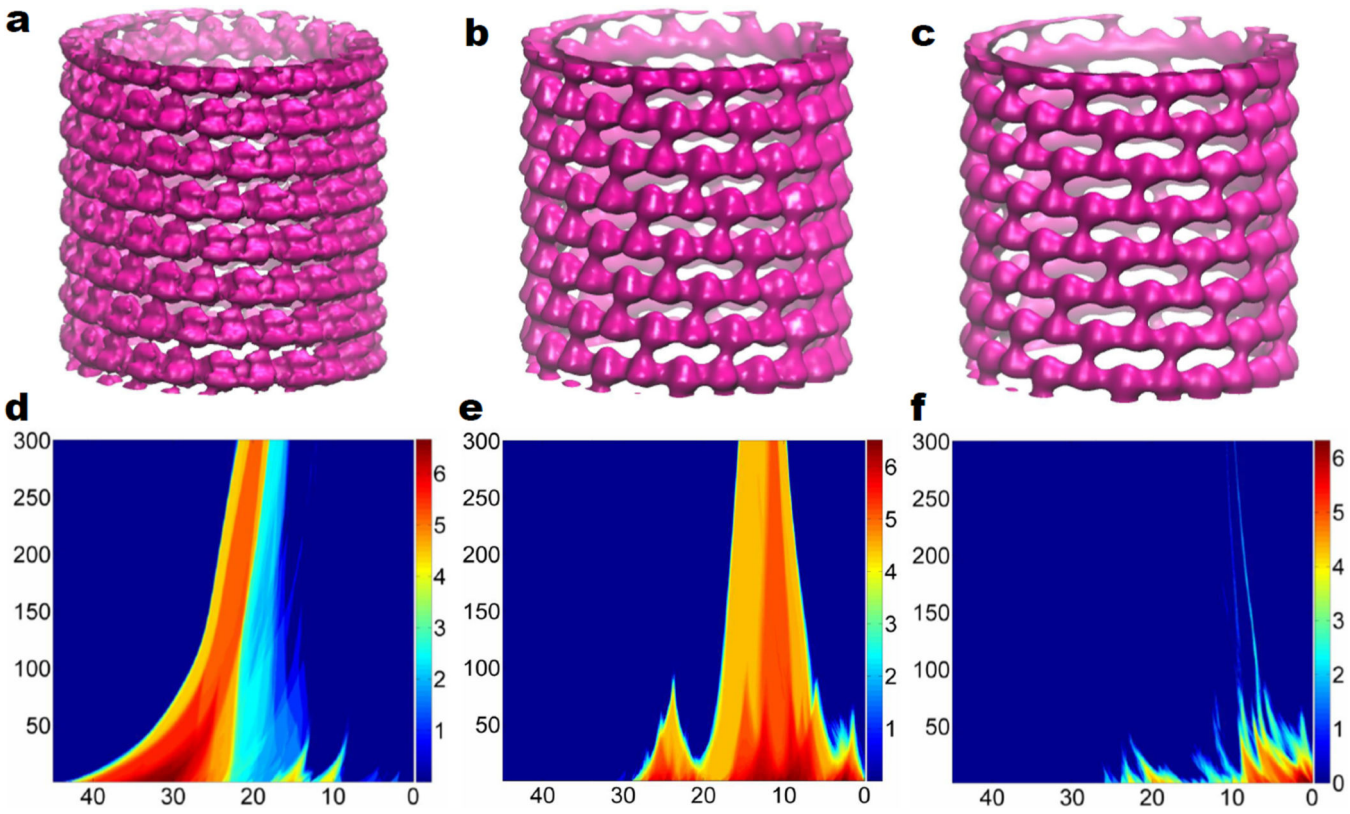

Figure 8.

Multidimensional topological denoising for EMD 1129 data of a microtubule structure. **a** Denoising data after one iteration; **b** Denoising data after 100 iteration; **c** Denoising data after 200 iteration; **d** 2D β_0 persistence; **e** 2D β_1 persistence; **f** 2D β_2 persistence. Isosurfaces in **a, b** and **c** are extracted at isovalue 15.0. In **d, e** and **f**, the horizontal axises are density isovalues (i.e., the main filtration parameter). The vertical axises represent the number of iterations. Color bars denote the natural logarithm of PBNs values added by 1.

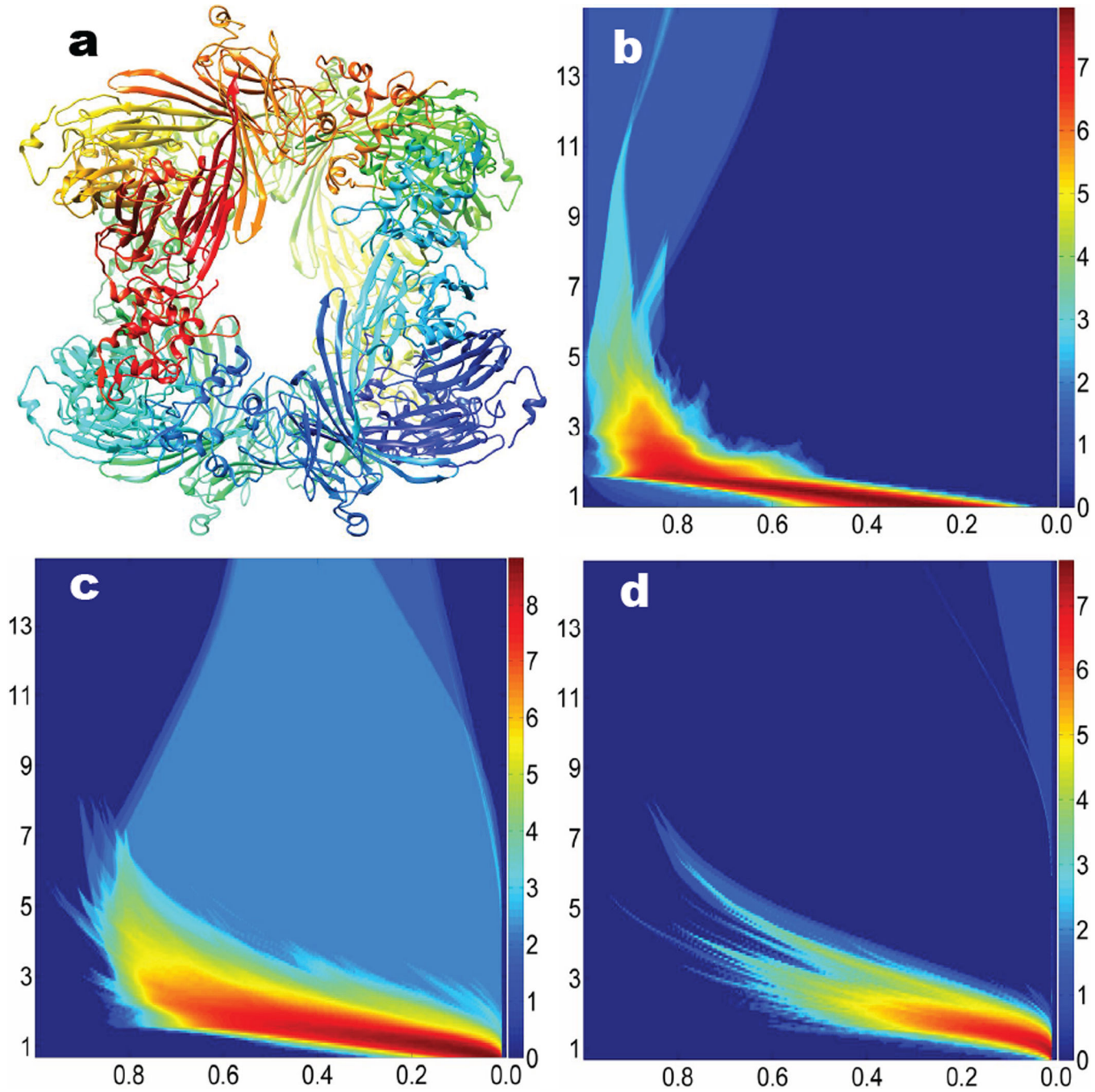

Figure 9.

The multiscale multidimensional persistence of the protein 2YGD. **a** The structure of protein 2YGD; **b** 2D β_0 persistence; **c** 2D β_1 persistence; **d** 2D β_2 persistence. In **b, c** and **d**, the horizontal axises are the density isovalues (i.e., the main filtration parameter). The vertical axises represent the scale (Å). natural logarithm of PBNs values added by 1.

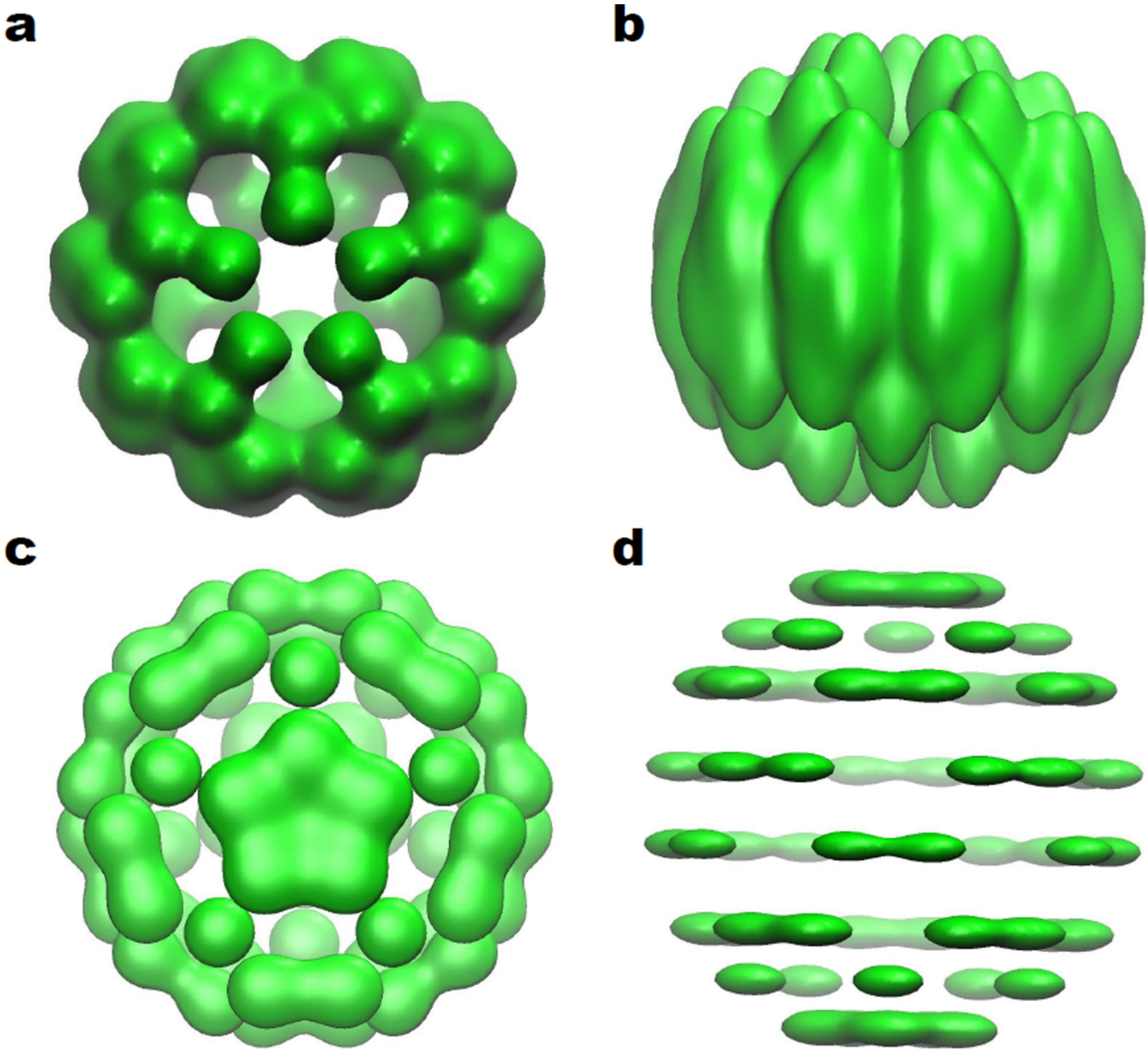

Figure 10.

Multidimensional anisotropic filtration of C_{60}. **a** The *z*-direction view of C_{60} with $\sigma^x = \sigma^y =$ 0.2 Å and $\sigma^z = 0.5$ Å at the isovalue of 0.4; **b** The *x*-direction view of C_{60} with $\sigma^x = \sigma^y = 0.2$ Å and $\sigma^z = 0.5$ Å at the isovalue of 0.4; **c** The *z*-direction view of C_{60} with $\sigma^x = \sigma^y = 0.5$ Å and $\sigma^z = 0.2$ Å at the isovalue of 1.0; **d** The *x*-direction view of C_{60} with $\sigma^x = \sigma^y = 0.5$ Å and $\sigma^z = 0.2$ Å at the isovalue of 1.0.

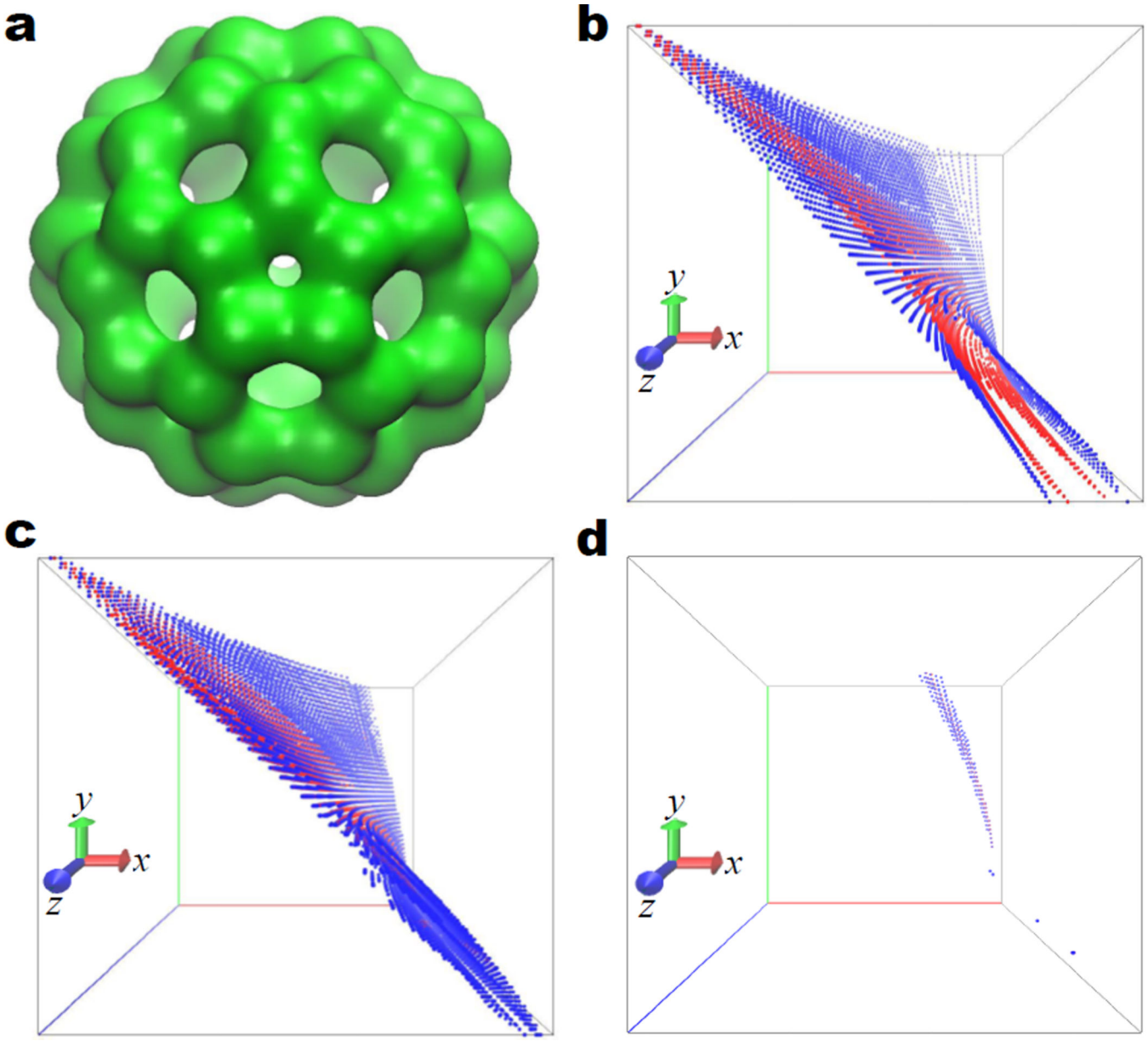

Figure 11.

The C_{60} molecule and its multiscale 3D persistence. **a** C_{60} molecule obtained with $\sigma^x = \sigma^y = \sigma^z = 0.5$ Å at the isovalue of 1.5; **b** 3D β_0 persistence; **c** 3D β_1 persistence; **d** 3D β_2 persistence. In **b, c** and **d**, the *x*-axises label the density value (i.e., the main filtration parameter), the *y*-axises denote σ^z and the *z*-axises represent $\sigma^x = \sigma^y$. The blue and red dots denote $\beta_0 = 4$ and 50 respectively in **b**, $\beta_1 = 3$ and 20 respectively in **c**, and $\beta_2 = 1$ and 2 respectively in **d**.