INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN BIOMEDICAL ENGINEERING
Int. J. Numer. Meth. Biomed. Engng. (2015); e02719
Published online 6 May 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cnm.2719

Persistent topology for cryo-EM data analysis

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*

SUMMARY

In this work, we introduce persistent homology for the analysis of cryo-electron microscopy (cryo-EM) density maps. We identify the topological fingerprint or topological signature of noise, which is widespread in cryo-EM data. For low signal-to-noise ratio (SNR) volumetric data, intrinsic topological features of biomolecular structures are indistinguishable from noise. To remove noise, we employ geometric flows that are found to preserve the intrinsic topological fingerprints of cryo-EM structures and diminish the topological signature of noise. In particular, persistent homology enables us to visualize the gradual separation of the topological fingerprints of cryo-EM structures from those of noise during the denoising process, which gives rise to a practical procedure for prescribing a noise threshold to extract cryo-EM structure information from noise contaminated data after certain iterations of the geometric flow equation. To further demonstrate the utility of persistent homology for cryo-EM data analysis, we consider a microtubule intermediate structure Electron Microscopy Data (EMD 1129). Three helix models, an alpha-tubulin monomer model, an alpha-tubulin and beta-tubulin model, and an alpha-tubulin and beta-tubulin dimer model, are constructed to fit the cryo-EM data. The least square fitting leads to similarly high correlation coefficients, which indicates that structure determination via optimization is an ill-posed inverse problem. However, these models have dramatically different topological fingerprints. Especially, linkages or connectivities that discriminate one model from another, play little role in the traditional density fitting or optimization but are very sensitive and crucial to topological fingerprints. The intrinsic topological features of the microtubule data are identified after topological denoising. By a comparison of the topological fingerprints of the original data and those of three models, we found that the third model is topologically favored. The present work offers persistent homology based new strategies for topological denoising and for resolving ill-posed inverse problems.

Received 23 December 2014; Revised 13 March 2015; Accepted 31 March 2015

KEY WORDS: cryo-EM; topological signature; geometric flow; topological denoising; topology-aided structure determination

1. INTRODUCTION

The quantitative understanding of structure, function, dynamics, and transport of biomolecules is a fundamental theme in contemporary life sciences. Geometric analysis and associated biophysical modeling have been the main workhorse in revealing the structure–function relationship of biomolecules and contribute enormously to the present understanding of biomolecular systems. However, biology encompasses over more than 20 orders of magnitude in time scales from electron transfer and ionization on the scale of femtoseconds to organism life spanning over tens of years, and over 15 orders of magnitude in spatial scales from electrons and nuclei to organisms. The intriguing complexity and extraordinarily large number of degrees of freedom of biological systems give rise to formidable challenges to their quantitative description and theoretical prediction. Most biological processes, such as signal transduction, gene regulation, DNA specification,

*Correspondence to: Guo-Wei Wei, Department of Mathematics Michigan State University, MI 48824, U.S.A.
†E-mail: wei@math.msu.edu

transcription, and post transcriptional modification, are essentially intractable for atomistic geometric analysis and biophysical simulations, let alone *ab-initio* quantum mechanical descriptions. Therefore, the complexity of biology and the need for its understanding offer an extraordinary opportunity for innovative theories, methodologies, algorithms, and tools.

The study of subcellular structures, organelles, and large multiprotein complexes has become one of the major trends in structural biology. Currently, some of the most powerful tools for the aforementioned systems are cryo-electron microscopy (cryo-EM) and cryo-electron tomography (cryo-ET), although other techniques, such as macromolecular X-ray crystallography, nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), multiangle light scattering, confocal laser-scanning microscopy, small angle scattering, ultra fast laser spectroscopy, and so on, are useful for structure determination in general [1–5]. In cryo-EM and cryo-ET experiments, samples are bombarded by electron beams at cryogenic temperatures to improve the signal-to-noise ratio (SNR). Three-dimensional (3D) cryo-EM maps are constructed from a large number of two-dimensional (2D) images using appropriate computational techniques, such as image segmentation, alignment, and classification. The working principle of cryo-ET is based on the projection (thin film) specimen scans collected from many different directions around one or two axes, and the Radon transform for the creation of three-dimensional (3D) images. One of major advantages of cryo-EM and cryo-ET is that they allow the imaging of specimens in their native environment. Another major advantage is their capability of providing 3D mapping of entire cellular proteomes together with their detailed interactions at nanometer or subnanometer resolution [1–4]. The resolution of cryo-EM maps has been improved dramatically in the past two decades, thanks to the technical advances in experimental hardware, noise reduction, and image segmentation techniques. By further taking the advantage of symmetric averaging, many cryo-EM based virus structures have already achieved a resolution that can be interpreted in terms of atomic models [6]. There have been a variety of elegant methods [7–12] and software packages in cryo-EM structural determination [13–19].

Most biological specimens are extremely radiation sensitive and can only sustain a limited electron dose of illumination. As a result, cryo-EM or cryo-ET images are inevitably of low SNR and limited resolution [5]. In fact, the SNRs of cryo-tomograms for subcellular structures, organelles, and large multi-protein complexes are typically in the neighborhood of 0.01 [5]. To make the situation worse, the image contrast, which depends on the difference between electron scattering cross sections of cellular components, is also very low in most biological systems. Consequently, cryo-ET and cryo-EM maps often do not contain adequate information to offer unambiguous atomic-scale structural reconstruction of biological specimens. Additional information obtained from other techniques, such as X-ray crystallography, NMR, and computer simulation, is indispensable to achieve subnanometer resolutions. However, for cryo-EM data that do not have much additional information obtained from other techniques, the determination of what proteins are involved can be a challenge, not to mention the subnanometer structural resolution.

To improve the SNR and image contrast of cryo-EM and cryo-ET data, a wide variety of denoising algorithms has been employed [20–27]. Standard techniques, such as bilateral filter [24–26] and iterative median filtering [27] have been utilized for noise reduction. Additionally, wavelets and related techniques have also been developed for cryo-EM and cryo-ET noise removing [20]. Moreover, anisotropic diffusion [21, 22] or Beltrami flow [23] approach has been proposed for cryo-EM and cryo-ET signal recovering. However, cryo-EM and cryo-ET data denoising is far from adequate and remains a challenge because of the extremely low SNRs and other technical complications [5, 28–30]. For example, one of difficulties is how to distinguish signal from noise in cryo-EM data. As a result, one does not know when to stop or how to apply a threshold in an iterative noise removing process. There is a pressing need for innovative approaches to further tackle this problem.

Recently, persistent homology has been advocated as a new approach for dealing with big data sets [31–34]. In general, persistent homology characterizes the geometric features with persistent topological invariants by defining a scale parameter relevant to topological events. The essential difference between the persistent homology and traditional topological approaches is that traditional topological approaches describe the topology of a given object in truly metric free or coordinate free representations, while persistent homology analyzes the persistence of the topological features of a given object via a filtration process, which creates a family of similar copies of the object at

DOI: 10.1002/cnm

different spatial resolutions. Technically, a series of nested simplicial complexes is constructed from a filtration process, which captures topological structures continuously over a range of spatial scales. The involved topological features are measured by their persistent intervals. Persistent homology is able to embed geometric information to topological invariants, so that 'birth' and 'death' of isolated components, circles, rings, loops, pockets, voids, or cavities at all geometric scales can be monitored by topological measurements. The basic concept of persistent homology was introduced by Frosini and Landi [35] and by Robins [36] in 1990s. Edelsbrunner *et al*. [37] introduced the first efficient computational algorithm, and Zomorodian and Carlsson [38] generalized the concept. A variety of elegant computational algorithms has been proposed to track topological variations during the filtration process [39–43]. Often, the persistent diagram can be visualized through barcodes [33], in which various horizontal line segments or bars are the homology generators lasted over filtration scales. It has been applied to a variety of domains, including image analysis [44–47], image retrieval [48], chaotic dynamics verification [49, 50], sensor network [51], complex network [52, 53], data analysis [32, 54–57], computer vision [46], shape recognition [58], and computational biology [59–61].

The concept of persistent homology has also been used for noise reduction. It is generally believed that short lifetime events (or bars) are of less importance and thus regarded as 'noise', while long lifetime ones are considered as 'topological signals' [62]; although this idea was challenged in our recent work [63]. In topological data analysis, pre-processing algorithms are needed to efficiently remove noise. Depending on the scale of a feature, a simple approach is to pick up a portion of landmark points as a representative of topological data [64]. The points can be chosen randomly, spatially evenly, or from extreme values. More generally, certain functions can be defined as a guidance for node selection to attenuate the noise effect, which is known as thresholding. Clustering algorithms with special kernel functions can also be employed to recover topological signal [62]. All of these methods can be viewed as a process of data sampling without losing the significant topological features. They rely heavily on the previous knowledge of the geometric or statistic information. In contrast, topological simplification [37, 65, 66], which is to remove the simplices and/or the topological attributes that do not survive certain threshold, focuses directly on the persistence of topological invariant. In contrast, Gaussian noise is known to generate a band of bars distributes over a wide range in the barcode representation [67]. Thanks to the pairing algorithm, persistence of a homology group is measured through an interval represented by a simplex pair. If the associated topological invariant is regarded less important, simplices related to this simplex pair are reordered. This approach, combined with Morse theory, proves to be a useful tool for denoising [65, 66], as it can alters the data locally in the region defined as noise. Additionally, statistical analysis has been carried out to provide confidence sets for persistence diagram. However, persistent homology has not been utilized for cryo-EM data noise reduction to our knowledge.

A large amount of experimental data for macroproteins and protein–protein complexes has been obtained from cryo-EM. To analyze these structural data, it is a routine procedure to fit them with the available high-resolution crystal structures of their component proteins. This approach has been shown to be efficient for analyzing many structures and has been integrated into many useful software packages such as Chimera [68]. However, this docking process is limited by data quality. For some low resolution data, which usually also suffer from low SNRs, there is enormous ambiguity in structure fitting or optimization, that is, a mathematically ill-posed inverse problem. Sometimes, high correlation coefficients can be attained simultaneously in many alternative structures, while none of them proves to be biologically meaningful. Basically, the fitting or optimization emphasizes more on capturing 'bulk' regions, which is reasonable as greater similarities in density distributions imply higher possibility. However, little attention is paid to certain small 'linkage' regions, which play important roles in biological system especially in macroproteins and protein–protein complexes. Different linkage parts generate different connectivity and thus directly influence biomolecular flexibility, rigidity, and even its functions. Because persistent homology is equally sensitive to both bulk regions and small linkage regions, it is able to make a critical judgment on the model selection in structure determination. However, nothing has been reported on persistent homology based solution to ill-posed inverse problems to our knowledge.

DOI: 10.1002/cnm

Although persistent homology has been applied to a variety of fields, the successful use of persistent homology is mostly limited to characterization, identification, and analysis (CIA). Indeed, persistent homology has seldom been employed for quantitative prediction. Recently, we have introduced molecular topological fingerprints (MTFs) based on persistent homology analysis of the topological invariants of biomolecules [63]. We have utilized MTFs to reveal the topology-function relationship of macromolecules. It was found that protein flexibility and folding stability are strongly correlated to protein topological connectivity, characterized by the persistence of topological invariants (i.e., accumulated bar lengths) [63]. Most recently, we have employed persistent homology to analyze the structure and function of nanomaterial, such as nanotubes and fullerenes. The MTFs are utilized to quantitatively predict total curvature energies of fullerene isomers [69].

The overall objective of this work is to explore the utility of persistent homology for cryo-EM analysis. First, we propose a topology-based algorithm for cryo-EM noise reduction and clean-up. We study the topological fingerprint or topological signature of noise and its relation to the topological fingerprint of cryo-EM structures. We reveal that the histograms of topological invariants of the Gaussian random noise have Gaussian distributions in the filtration parameter space. Contrary to the common belief that short barcode bars correspond to noise, it is found that there is an inverse relation between the SNR and the band widths of topological invariants, that is, the lower SNR, the larger noise barcode band width is. Therefore, at a low SNR, noise can produce long persisting topological invariants or bars in the barcode presentation. Moreover, for the cryo-EM data of low SNRs, intrinsic topological features of the biomolecular structure are hidden in the persistent barcodes of noise and indistinguishable from noise contribution. To recover the topological features of biomolecular structures, geometric flow equations are employed in the present work. It is interesting to note that topological features of biomolecular structures persist, while the topological fingerprint of noise moves to the right during the geometric flow iterations. As such, 'signal' and noise separate from each other during the geometric flow-based denoising process and make it possible to prescribe a precise noise threshold for the noise removal after certain iterations. We demonstrate the efficiency of our persistent homology controlled noise removal algorithm for both synthetic data and cryo-EM density maps.

Additionally, we introduce persistent homology as a new strategy for resolving the ill-posed inverse problem in cryo-EM structure determination. Although the structure determination of microtubule data EMD 1129 is used as an example, similar problems are widespread in other intermediate resolution and low resolution cryo-EM data. As EMD 1129 is contaminated by noise, a preprocess of denoising is carried out by using our persistent homology controlled geometric flow algorithm. A helix backbone is obtained for the microtubule intermediate structure. Based on the assumption that the voxels with high electron density values are the centers of tubulin proteins, we construct three different microtubule models, namely a monomer model, a two-monomer model, and a dimer model. We have found that all three models give rise to essentially the same high correlation coefficients, that is, 0.9601, 0.9607, and 0.9604, with the cryo-EM data. This ambiguity in structure fitting is very common with intermediate and low resolution data. Fortunately, after our topology-based noise removal, the topology fingerprint of microtubule data is very unique, which is true for cryo-EM data or data generated by using other molecular imaging modalities. It is interesting to note that, although three models offer the same correlation coefficients with the cryo-EM data, their topological fingerprints are dramatically different. It is found that the topological fingerprint of the microtubule intermediate structure (EMD 1129) can be captured only when two conditions are simultaneously satisfied: first, there must exist two different types of monomers and, additionally, two type of monomers from dimers. Therefore, based on topological fingerprint analysis, we can determine that only the third model is a correct model for microtubule data EMD 1129.

The rest of this paper is organized as follows. The essential methods and algorithms for geometric and topological modelings of biomolecular data are presented in Section 2. Approaches for geometric modeling, which are necessary for topological analysis, are briefly discussed. Methods for persistent homology analysis are described in detail. We illustrate the use of topological methods with both synthetic volumetric data and cryo-EM density maps. Their persistence of topological invariants is represented by barcodes. The geometric interpretation of the topological features is given. Section 3 is devoted to the persistent homology-based noise removal. The characterization

DOI: 10.1002/cnm

of Gaussian noise is carried out over a variety of SNRs to understand noise topological signature. Based on this understanding, we design a persistent homology monitored and controlled algorithm for noise removal, which is implemented via the geometric flow. Persistent homology guided denoising is applied to the analysis of a supramolecular filamentous complex. In Section 4, we demonstrate topology-aided structure determination of microtubule cryo-EM data. Several aspects are considered including helix backbone evaluation, coarse-grained modeling, and topology-aided structure design and evaluation. We show that topology is able to resolve ill-posed inverse problem. This paper ends with a conclusion.

2. GEOMETRIC AND TOPOLOGICAL MODELINGS OF BIOMOLECULAR DATA

Persistent homology has been utilized to analyze biomolecular data, which are collected by different experimental means, such as macromolecular X-ray crystallography, NMR, EPR, and so on. Because of their different origins, these data may be available in different formats, which requires appropriate topological tools for their analysis. Additionally, their quality, that is, resolution and SNR, varies from case to case; thus, a preprocessing may be required. Moreover, although biomolecular structures are not a prerequisite for persistent homology analysis, the understanding of biomolecular structure, function, and dynamics is crucial for the interpretation of topological results. As a consequence, appropriate geometric modeling [70] is carried out in a close association with topological analysis. Furthermore, information from geometric and topological modelings is, in turn, very valuable for data preprocessing and denoising. Finally, topological information is shown to be crucial for geometric modeling, structural determination, and ill-posed inverse problems.

2.1. Geometric modeling of biomolecules

Geometric modeling of biomolecules gives rise to their structural information, which is of paramount importance for biological understanding and structure–function relationship [70, 71]. Geometric modeling typically begins with experimental data. There are two major repositories, namely, Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB), for storing biomolecular experimental data. The PDB lists detailed information of atomic coordinates, occupancy, and Debye–Waller factor (or B-factor) of all the atoms in proteins, DNAs, RNAs, and their complexes. In the persistent homology terminology, PDB provides point cloud data for biomolecules. In contrast, the EMDB typically offers volumetric data, or density maps of large biomolecular systems, such as multi-proteins, subcellular structures, and organelles from cryo-EM at the molecular level resolution.

Molecular structures and their visualization can be generated by using a variety of molecular models, such as the atom and bond model of molecules [72], the van der Waals surface, the solvent-excluded surface (also known as molecular surface), and the solvent-accessible surface, have been proposed [73, 74]. These models have been widely applied to the analysis of biomolecular structure, function, and interaction, such as ligand–receptor binding, protein specification, drug design, macromolecular assembly, protein-nucleic acid and protein–protein interactions, and enzymatic mechanism [75]. Nevertheless, they admit geometric singularities, that is, tips, cusps, and self-intersecting surfaces that are troublesome in simulations [76–79] and *ad hoc* in physical foundation because electron density decays gradually at the molecular boundary [80, 81].

Recently, we have introduced the differential geometry theory of surfaces to address the aforementioned problems in biomolecular geometric modeling by curvature control PDEs [82], mean curvature flows [83, 84], and potential-driven geometric flows [85]. The minimization principle was utilized for the biomolecular surface construction. We have further generalized these ideas to incorporate multiscale and multiphysical descriptions of biomolecules [80, 81, 86, 87]. Our approaches have been adopted and/or generalized by many others [88–91].

Most recently, we have proposed flexibility and rigidity index (FRI) for flexibility analysis and B-factor prediction of proteins and other biomolecules [92, 93]. In the FRI, protein topological connectivity is measured by rigidity index and flexibility index. In particular, the rigidity index represents the protein density profile. Consider a protein with N atoms. Their locations are represented

DOI: 10.1002/cnm

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 space distance between the ith atom and the jth atom. We define a position ($\mathbf{r}$) dependent rigidity or density function [92, 93]

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

where $w_j(\mathbf{r})$ is an atom type-dependent weight function, and σ_j is a resolution parameter, which can be adjusted to focus on the scale of the interest. It plays the same role as the resolution in wavelet theory. Here, $\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \sigma_j)$ is a correlation kernel, which is, in general, a real-valued monotonically decreasing function satisfying

$$\Phi\left(\|\mathbf{r} - \mathbf{r}_j\|; \sigma_j\right) = 1 \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_j\| \to 0 \tag{2}$$

and

$$\Phi\left(\|\mathbf{r} - \mathbf{r}_j\|; \sigma_j\right) = 0 \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_j\| \to \infty. \tag{3}$$

Although delta sequences of the positive type discussed in earlier work [94] are all good choices, generalized exponential functions

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \sigma_j) = e^{-\left(\|\mathbf{r} - \mathbf{r}_j\|/\sigma_j\right)^{\kappa}}, \quad \kappa > 0 \tag{4}$$

and generalized Lorentz functions

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \sigma_j) = \frac{1}{1 + \left(\|\mathbf{r} - \mathbf{r}_j\|/\sigma_j\right)^{\upsilon}}, \quad \upsilon > 0 \tag{5}$$

have been commonly used in our recent work [92, 93]. We refer these general classes of volumetric surface definition as rigidity surfaces or density profiles. We use a surface extraction procedure, such as marching cubes, to extract a Lagrangian surface from their volumetric data, or the Eulerian representation of surface density profiles. Obviously, when $\kappa = 2$ in Equation (4), Equation (1) gives rise to a representation of Gaussian surfaces, which have many formulations [71, 92, 95–99]. In general, Gaussian surfaces are quite smooth and free of geometric singularity. The generation of Gaussian surfaces can be very fast and readily available in the Cartesian representation [97]. Other geometric modeling approaches include curvature analysis and symmetry analysis. Mean curvature and Gauss curvature can be estimated in both Lagrangian representations [70] and Eulerian representation [71]. Maximal and minimal principle curvatures can be utilized for drug-binding site prediction [71]. Symmetric analysis is frequently employed in biophysical modeling [97]. Utilizing symmetry leads to the reduction of number of genes, which is very common in viral complexes. Many protein complexes, such as microtubules, are highly symmetric as well.

Figure 1 illustrates surfaces extracted from density function Equation (1) with $w_j = 1, \kappa = 2$, and $\sigma_j = 0.5$Å. In this work, we use density function (1) as a mathematical model for cryo-EM density maps. A series of surfaces is plotted in Figure 1 to demonstrate some typical structures in the filtration procedure for fullerene C_{20} density function. The isovalues for Figures 1**a**–**c** are 0.7, 0.6, and 0.4, respectively. Figure 1**d** is a wire-frame surface representation of Figure 1**c**.

2.2. Topological modeling of biomolecules

Persistent homology theory and algorithm can be found in the literature [33, 37–41, 43] as well as our papers [63, 69]. In this work, we focus on electron density maps of macro-protein or protein–protein complexes available as volumetric data deposited in the EMDB. Unlike point-cloud data that are commonly studied with simplicial complex using Javaplex [100], volumetric data are usually analyzed by the discrete Morse theory. For all the density map-based volumetric data used in this work, we use the same filtration process that is built based on the decrement of the electron density

DOI: 10.1002/cnm

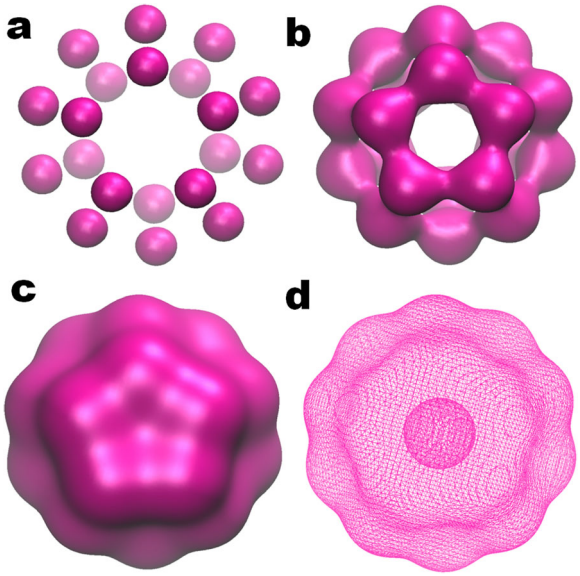

Figure 1. A series of surfaces extracted from fullerene C_{20} density function (1) with $w_j = 1, \kappa = 2$, and $\sigma_j = 0.5$Å. The isovalues for subfigures **a**, **b**, and **c** are 0.7, 0.6, and 0.4, respectively. Subfigure **d** is wire-frame surface representation of subfigure **c**. The Betti numbers can be directly obtained through identifying the numbers of connected components, circles or loops, and voids. In **a**, β_0 is 20, and β_1 and β_2 are 0; In **b**, $\beta_0 = 1$, and $\beta_1 = 11$ and $\beta_2 = 0$; In **c** and **d**, $\beta_0 = 1$, and $\beta_1 = 0$ and $\beta_2 = 1$. Attention should be paid to β_1 for the structure in **b**. The β_1 is not exactly equal to the total number (N_1) of the circles or loops; instead, it is equal to $N_1 - 1$. This is due to the fact that the corresponding homology group has a basis with only $N_1 - 1$ elements. Therefore, one of elements can be expressed as the 'linear combination' of the rest.

value. More specifically, cryo-EM density maps or density functions generated from Equation (1) are used as the filtration parameter. The filtration process goes from the highest density isovalue to the lowest one. After the filtration, the data are analyzed with Perseus [101]. We first consider a benchmark test to illustrate the persistent homology analysis of density function (1).

Topological persistence of C_{20} The Betti numbers can be directly obtained through identifying the number of connected components, circles or loops, and voids or holes. For example, in Figure 1 **a**, β_0 is 20, and β_1 and β_2 are 0. In Figure 1**b**, $\beta_0 = 1, \beta_1 = 11$, and $\beta_2 = 0$. In Figures 1**c** and **d**, $\beta_0 = 1, \beta_1 = 0$, and $\beta_2 = 1$. Attention should be paid to β_1 for the structure in Figure 1**b**. The β_1 is not exactly equal to the total number (N_1) of the circles or loops; instead, it equals to $N_1 - 1$. This is due to the reason that the corresponding homology group has a basis with only $N_1 - 1$ independent elements. Roughly speaking, one of the circles can be expressed as the 'linear combination' of the rest.

Barcodes provide a systematic representation of topological persistence [33]. Figure 2 shows the intrinsic topological patterns of fullerene C_{20}. Just the same as we counted earlier, there are 11 β_1 bars and only one β_2 bar. It should be noticed that β_0 bars do not emerge simultaneously, which is because of the discretization effect, namely, the density function is discretized with only a finite resolution.

Topological persistence of EMD 1776 After demonstrating the persistent homology analysis for the density function (1) of a known structure (C_{20}), we further consider realistic cryo-EM data, EMD 1776, which is for eye lens chaperone α-crystallin assemblies. Figure 3 depicts the surfaces extracted from different isovalues for EMD 1776. The isovalues for Figures 3**a**–**d** are 0.150, 0.100, 0.081, and 0.050, respectively. Similarly, Betti numbers can be directly obtained through counting the numbers of connected components, circles, and voids. In Figure 3**a**, β_0 is 12, and β_1 and β_2 are 0. In Figure 3**b**, $\beta_0 = 4, \beta_1 = 4$, and $\beta_2 = 0$. In Figure 3**c**, $\beta_0 = 1, \beta_1 = 13$, and $\beta_2 = 0$. Also in Figure 3**d**, one has $\beta_0 = 1, \beta_1 = 9$, and $\beta_2 = 0$. As discussed earlier, in Figures 3**c** and **d**, the β_1 value is $N_1 - 1$, rather than the total number (N_1) of the circles, because

DOI: 10.1002/cnm

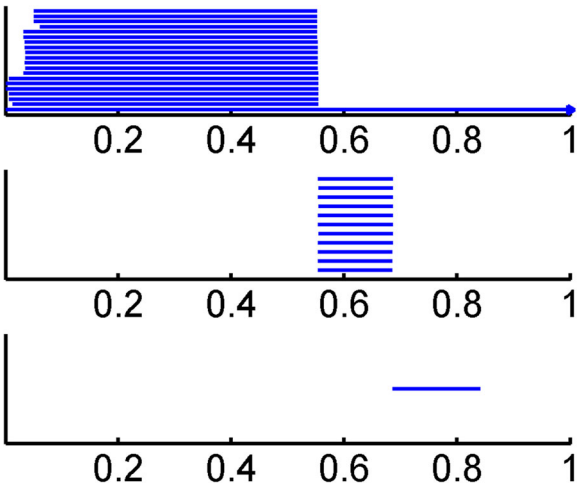

Figure 2. The intrinsic topological fingerprints of fullerene C_{20}. The top, middle, and bottom panels are for the barcodes of β_0, β_1, and β_2, respectively. The filtration process is based on the decrement of the density isovalues. The horizontal axis represents the rescaled density values, with 0 denoting the largest density value and 1 denoting the smallest one. The filtration barcodes demonstrate pentagonal structures and central void. It should be notice that β_0 bars do not emerges simultaneously, which is because of the Cartesian representation of data.

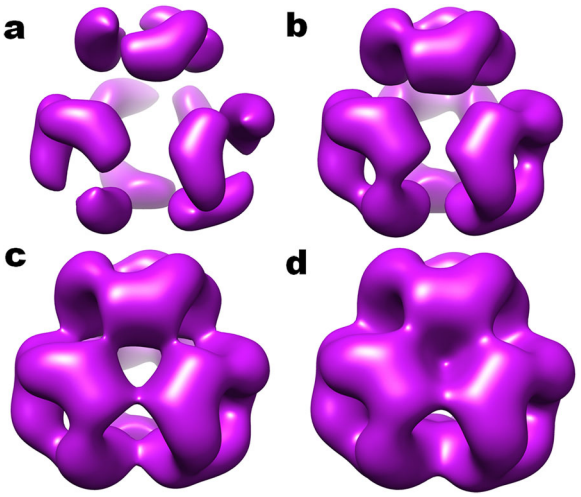

Figure 3. Surfaces extracted with different isovalues for EMD 1776. The isovalues for subfigures **a**, **b**, **c**, and **d** are 0.150, 0.100, 0.081, and 0.050, respectively. The Betti numbers can be directly obtained through identifying the number of connected components, circles or loops, and voids or holes. In **a**, β_0 is 12, and β_1 and β_2 are 0; In **b**, $\beta_0 = 4$, $\beta_1 = 4$, and $\beta_2 = 0$; In **c**, $\beta_0 = 1$, $\beta_1 = 13$, and $\beta_2 = 0$; In **d**, $\beta_0 = 1$, $\beta_1 = 9$, and $\beta_2 = 0$. In **c** and **d**, the β_1 is one count fewer than the total number (N_1) of the circles or loops because the corresponding homology group has a basis with only $N_1 - 1$ elements.

of $N_1 - 1$ independent elements in the corresponding homology group. The barcode representation is demonstrated in Figure 4, which is consistent with our analysis.

It should be noticed that we only consider the regions with density values larger than 0.03. In other words, the filtration goes from the largest value (0.28) to a threshold value (0.03). For density values smaller than 0.03, data suffer from lower SNRs as discussed in Section 3.2. Denoising techniques are indispensable for extracting more information from low isovalues. In the next section, we apply persistent homology to noise reduction and topological feature identification.

DOI: 10.1002/cnm

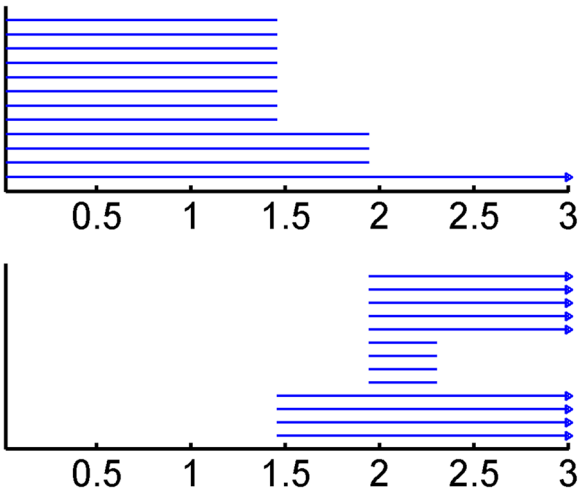

Figure 4. The intrinsic topological patterns of the EMD 1776 structure. The top and bottom panels are for the barcodes of β_0 and β_1, respectively. The filtration process is based on the decrement of density isovalues. The filtration goes from the largest isovalue (0.28) to the isovalue threshold (0.03), which have been rescaled to 0 and 3, respectively.

3. PERSISTENT HOMOLOGY BASED NOISE REDUCTION

In this section, we present persistent homology-based cryo-EM data noise reduction, which is a crucial process in cryo-EM analysis. Protein data in EMDB are mostly obtained by cryo-EM. As discussed earlier, the cryo-EM data suffer from low resolution and low SNR. Therefore, a denoising process is always a necessity before carrying out geometric modeling and/or topological analysis. We focus on noise reduction based on persistent homology analysis. Specifically, we employ persistent homology to discriminate signal from noise and utilize this information for denoising thresholding. In our benchmark test, we assume a known object is contaminated with Gaussian noise. Geometric flows proposed in our earlier work [85, 102] are used for noise reduction of realistic cryo-EM data.

3.1. Topological fingerprints of Gaussian noise

We first analyze the topological fingerprint or topological signature of noise. We use Gaussian noise as an example for the present study. Other noise can be analyzed in a similar manner. The Gaussian white noise is generated by randomly selecting values from a normal distribution,

$$n(t) = \frac{A_n}{\sqrt{2\pi}\sigma_n} e^{-\frac{(t-\mu_n)^2}{2\sigma_n^2}}, \tag{6}$$

where A_n, μ_n and σ_n are the amplitude, mean value, and standard deviation of the noise, respectively. We denote μ_n as the mean value of signal. Then the degree of noise contamination can be described by SNR; SNR $= \mu_n/\sigma_n$. Please note that the present definition of SNR is by no means unique. Based on the physical properties of the signal, SNR can be defined in terms of average power, amplitude, variance of the signal, and so on. In our discussion, the signal information represented in volumetric data can be easily analyzed. We generate the noise data with specified SNR by adding suitable amplitude of Gaussian white noise. Stated differently, the noise contaminated data are generated by adding different strengths of Gaussian white noise to the original data, that is, the density function or density map. Then, the corresponding persistent barcode patterns are analyzed.

The density function described in Equation (1) with the generalized exponential kernel shown in Equation (4) is used to simulate the density of fullerene C_{20}. We choose $\kappa = 1.0$ and $\sigma = 0.5$Å in our study. Figure 5 demonstrates the barcode representation for contaminated fullerene C_{20} data with different SNRs. The SNRs for Figures 5**a**–**d** are 0.1, 1.0, 10.0, and 100.0, respectively. It can be seen from Figures 5**a** and **b**, that, when SNR is low, that is, SNR $= 0.1$ or 1.0, fullerene atoms

DOI: 10.1002/cnm

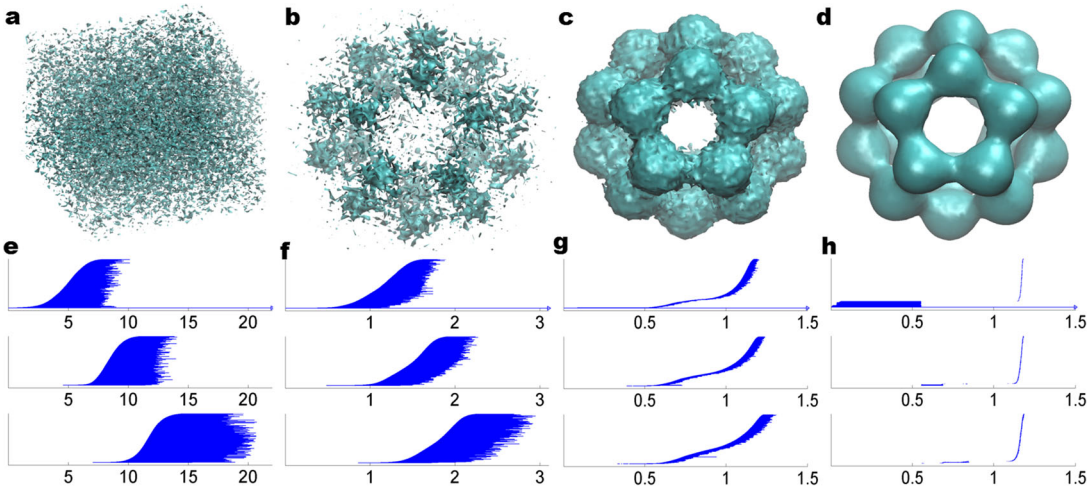

Figure 5. The barcode representation for contaminated fullerene C_{20} data with different SNRs. The SNRs for **a**, **b**, **c**, and **d** are 0.1, 1.0, 10.0, and 100.0, respectively, and isovalues used for their visualization are 4.00, 1.00, 0.60, and 0.60, respectively. The corresponding barcodes are given in **e**, **f**, **g**, and **h**, respectively. It should be noticed that the filtration always goes from the largest density value to the smallest, and the horizontal axis in our barcodes represents the rescaled density values. The top, middle, and bottom panels are for the barcodes of β_0, β_1, and β_2, respectively. It can be seen from the barcodes that, when SNR is low, that is, SNR $= 0.1$ or 1.0, fullerene molecule is invisible. At a low noise level, molecular pattern emerges. More importantly, the persistence of the pseudo-topological structure is directly related to the SNR. In the barcode representation, noise tends to induce a continuous band (or stripe) of bars, of which the width or relative persistent length is determined by the magnitude of the noise.

are invisible. When the SNR is increased in Figures 5**c** and **d**, the molecular intrinsic patterns begin to emerge.

The topological signature of the aforementioned four cases can be analyzed as shown in Figures 5**e**–**h**. Again, it should be noticed that in all our density filtration, the filtration always goes from the largest density value to the smallest one. The horizontal axis in our barcodes represents the rescaled density values. First, we note that all of three topological invariants, namely, β_0, β_1, and β_2, are very sensitive to and essentially dominated by noise. Additionally, noise gives rise to continuous bands (or stripes) of bars in all the three topological invariants. Moreover, noise provides similar numbers of bars in β_0, β_1, and β_2 panels. Furthermore, contradicting to the common belief that noise only contributes to short lived bars, the noise bars can be very long. The average band length of noise bars proportions to noise magnitude. In fact, in our recent work [63], we regard all bars, including short lived bars, of a protein as molecular topological fingerprints and as being of equal importance. Finally, at low SNRs (i. e., Figures 5**e** and **f**), the bars of three invariants maintain Gaussian-like distributions with respect to the isovalue filtration (i. e., the x-axis) as shown in Figure 6. However, at relatively high SNRs, bars do not have the Gaussian-like distributions because of a relatively high C_{20} density (Figure 6).

To further analyze the topological signature of noise, we consider a protein structure made of only beta sheets obtained from PDB 2GR8 (Protein Data Bank ID 2GR8). Again, its total density distribution is approximated by using the density function given in Equation (1), realized by using the generalized exponential kernel of Equation (4) with $\kappa = 1.0$ and $\sigma = 2.0$Å. Figure 7 shows our results. The SNRs for Figures 7**a**–**d** are 0.1, 0.5, 1.0, and 10.0, respectively. The corresponding barcodes are given in Figures7**e**–**h**. The topological signature of noise in Figure 7 is quite similar to that in Figure 6. Essentially, Gaussian noise induces continuous bands (or stripes) of bars, whose width or relative persistent length is determined by noise intensity.

As demonstrated in the aforementioned examples, barcodes present a topological description of Gaussian noise. The related band width of the noise can be used to assess noise magnitude. This qualitative description can be used as a guidance for topological noise reduction and topological fingerprint identification.

DOI: 10.1002/cnm

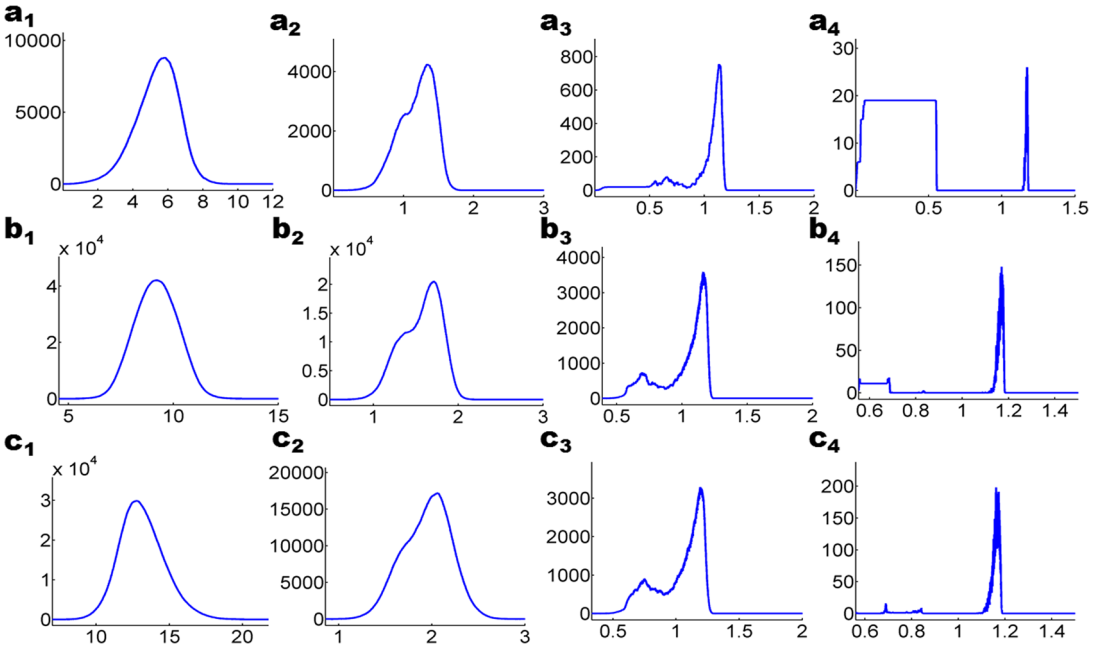

Figure 6. The histograms of topological invariants over the filtration process for contaminated fullerene C_{20} data with different SNRs. The $\mathbf{a}_i$, $\mathbf{b}_i$, and $\mathbf{c}_i$ rows are the counts of β_0, β_1, and β_2, respectively, where subscripts $i = 1, 2, 3$, and 4 correspond to SNRs 0.1, 1.0, 10.0, and 100.0, respectively. It can be seen that, as the SNR increases, barcodes for noise and signal gradually separate from each other. Because the Gaussian noise is used, the noise parts typically assume Gaussian distributions in shape.

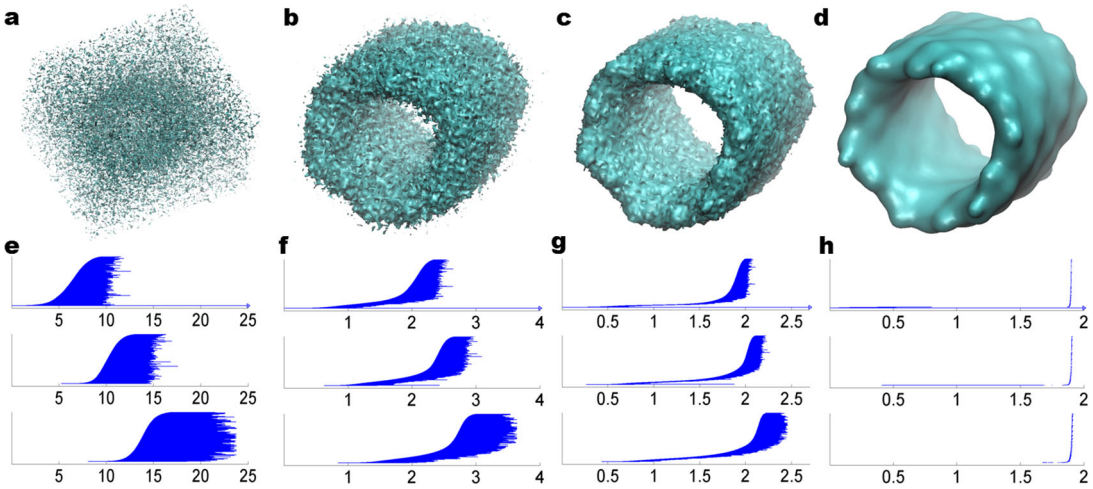

Figure 7. The barcodes representation for PDB 2GR8 (Protein Data Bank ID 2GR8) beta segment with different SNRs. The SNRs for **a**, **b**, **c**, and **d** are 0.1, 0.5, 1.0, and 10.0, respectively; isovalues used for their visualization are 5.00, 1.00, 1.00, and 1.00, respectively. The corresponding barcodes are presented in **e**, **f**, **g**, and **h**, respectively. The top, middle, and bottom panels are for the barcodes of β_0, β_1, and β_2, respectively. It can be seen from the barcodes that, when the SNR is large, that is, SNR $= 0.1$ or 0.5, the original topological properties are blurred. When the noise effect dwindles, the intrinsic patterns begin to emerge. More importantly, the persistence of the pseudo-topological structure is directly related to the SNR. In the barcode representation, noise tends to induce a continuous band (or stripe) of bars, of which the width or relative persistent length is determined by the magnitude of the noise.

DOI: 10.1002/cnm

3.2. Topological denoising

Geometric flows Geometric PDEs offer an efficient approach for noise reduction. High-order geometric PDEs were first introduced for edge-preserving image restoration in 1999, and have a general form [102]

$$\frac{\partial u(\mathbf{r},t)}{\partial t} = -\sum_q \nabla\cdot\mathbf{j}_q + e(u(\mathbf{r},t),|\nabla u(\mathbf{r},t)|,t), \quad q = 0,1,2,\cdots, \tag{7}$$

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

$$\mathbf{j}_q = -d_q(u(\mathbf{r},t),|\nabla u(\mathbf{r},t)|,t)\nabla\nabla^{2q}u(\mathbf{r},t), \quad q = 0,1,2,\cdots, \tag{8}$$

where $\mathbf{r}\in\mathbb{R}^n, \nabla = \frac{\partial}{\partial\mathbf{r}}, u(\mathbf{r},t)$ is the processed image function, $d_q(u(\mathbf{r},t),|\nabla u(\mathbf{r},t)|,t)$ are edge-sensitive diffusion coefficients, and $e(u(\mathbf{r},t),|\nabla u(\mathbf{r},t)|,t)$ is a nonlinear operator. The original noise data $X(\mathbf{r})$ is used as the initial input $u(\mathbf{r},0) = X(\mathbf{r})$. The hyper-diffusion coefficients $d_q(u,|\nabla u|,t)$ in Equation (8) can also be chosen as the Gaussian form

$$d_q(u(\mathbf{r},t),|\nabla u(\mathbf{r},t)|,t) = d_{q0}\exp\left[-\frac{|\nabla u|^2}{2\sigma_q^2}\right], \tag{9}$$

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

$$\sigma_q^2(\mathbf{r}) = \overline{|\nabla^q u - \overline{\nabla^q u}|^2} \quad (q = 0,1). \tag{10}$$

Here, the notation $\overline{Y(\mathbf{r})}$ represents the local average of $Y(\mathbf{r})$ centered at position $\mathbf{r}$.

High-order geometric PDEs have many practical applications [102–104]. They have been specifically modified for molecular surface formation and evolution [85] as

$$\frac{\partial S}{\partial t} = (-1)^q\sqrt{g(|\nabla\nabla^{2q}S|)}\nabla\cdot\left(\frac{\nabla(\nabla^{2q}S)}{\sqrt{g(|\nabla\nabla^{2q}S|)}}\right) + P(S,|\nabla S|), \tag{11}$$

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 [84],

$$\frac{\partial S}{\partial t} = |\nabla S|\nabla\cdot\left(\frac{\nabla S}{|\nabla S|}\right). \tag{12}$$

We employ this Laplace–Beltrami equation for the noise reduction in this paper.

Topological fingerprint identification Computationally, the finite different method is used to discretize Equation (12). Suitable time stepping Δt and grid spacing h are needed. For cryo-EM data, its voxel spacing is related to the data resolution and varies greatly. For example, The voxel spacings of EMD1776, EMD1229, and EMD5729 are 1.69 Å, 4.00 Å, and 4.16 Å, respectively. In our simulated fullerene C_{20} , C_{60}, and PDB 2GR8 examples, the voxel spacings are 0.06 Å, 0.08 Å, and 0.40 Å, respectively. To avoid confusion and control the noise-reduction process systematically, we simply ignore the voxel spacing and use unified parameters $\Delta t = 1.0E-5$ and $h = 1.0E-2$. The intensity of denoising is then described by the number of iteration steps for solving the governing Equation (12).

The noise-reduction effectiveness is commonly validated by a visual comparison with the original results. Quantitative assessment usually proves to be difficult, as noise and signal or image information can be tightly entangled. In this section, we propose a topological method for monitoring the evolution of relative behaviors of noise and signal during the denoising process. More specifically, persistent barcodes from a series of denoising data are compared. The noise signature and topological fingerprint in these barcodes are carefully studied. The present emphasis is on the topological fingerprint identification.

DOI: 10.1002/cnm

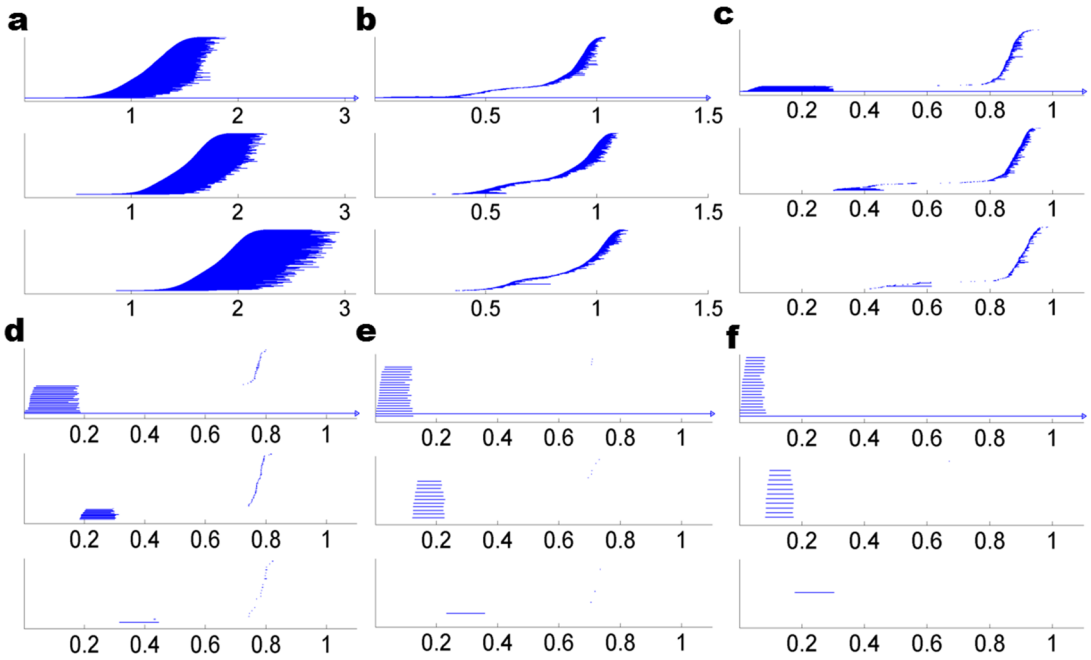

Figure 8. The barcodes representation for denoising contaminated fullerene C_{20} with SNR 1.0. The barcodes for fullerene C_{20} with SNR 1.0 is demonstrated in **a**. The denoising steps for **b**, **c**, **d**, **e**, and **f** are 20, 40, 60, 80, and 100, respectively. The noise induced topological invariants have been gradually weakened and finally eradicated. Compared with the original noise-polluted barcodes in **a**, the noise effect has been enormously scaled down after only 20 steps of denoising as indicated in **b**. In **c**, there is a clear separation between the intrinsic topological features of fullerene C_{20} and noise induced topological invariants. From **d** to **f**, the noise effect is further reduced, and we are able to identify a persistent barcode pattern, which is an indication of the intrinsic topological invariants of fullerene C_{20}. It should also be noticed that the denoising process fades noise intensity; therefore, noise-induced topological patterns gradually move to the right of filtration parameter and finally disappear.

During the denoising process, topological features of both signal and noise are constantly evolving. To quantitatively analyze their behaviors, three cases with Gaussian white noise and a case with Cryo-EM data are considered. The first case is a noise contaminated fullerene C_{20} with SNR 1.0. The persistent barcode results are demonstrated in Figure 8. We vary the number of denoising steps in our study. For Figures 8**b**–**f**, the numbers of iterations are 20, 40, 60, 80, and 100, respectively. The noise induced topological persistence has been gradually weakened and finally cleaned up. Compared with the original noise-polluted barcodes in Figure 8**a**, the noise effect has been enormously scaled down after only 10 steps of denoising as indicated in Figure 8**b**. In Figure 8**c**, there is a clear separation between the intrinsic topological persistence of fullerene C_{20} and noise induced topological persistence. From Figures 8**d**–**f**, the noise effect is further weakened; thus, we are able to identify a consistent barcode pattern, which is an indication of the intrinsic topological invariants of fullerene C_{20}. It should also be noticed that, because noise intensity diminishes during the denoising process, noise induced topological patterns gradually shift to the right of the filtration parameter and eventually disappear.

The second case is a noise contaminated protein segment from PDB 2GR8 with SNR 1.0. Its topological behavior under the denoising process is illustrated in Figure 9. In this case, the noise induced topological invariants have been gradually weakened but not eradicated. Compared with the original noise-polluted barcodes in Figure 9**a**, the noise persistence has been enormously reduced after 10 steps of denoising as indicated in Figure 9**b**. From Figures 9**b**–**f**, there is a clear separation between the intrinsic topological invariants of protein segment and noise-induced topological invariants. As the noise effect is continuously weakened, we are able to identify a persistent barcode pattern, which is an indication of the intrinsic topological invariants of the protein segment.

DOI: 10.1002/cnm

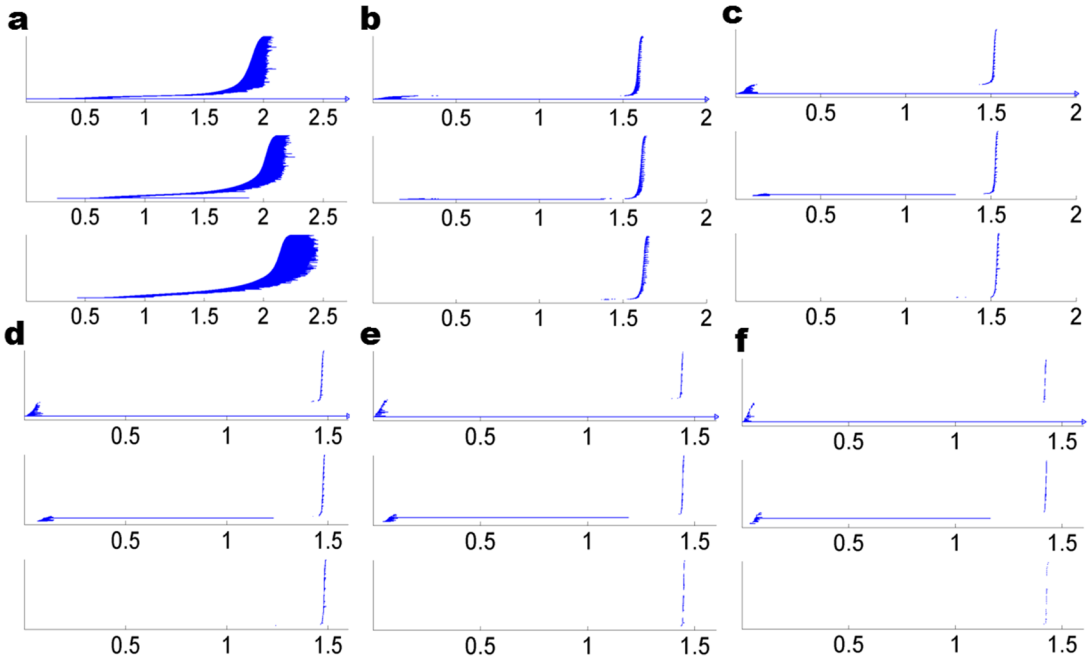

Figure 9. The barcodes representation for the noise reduction of contaminated PDB 2GR8 beta segment with SNR 1.0. The barcodes for PDB 2GR8 with SNR 1.0 is demonstrated in **a**. The denoising iteration steps for **b**, **c**, **d**, **e**, and **f** are 10, 20, 30, 40, and 50, respectively. In this case, the noise induced topological invariants have been gradually weakened but not eradicated. Compared with the original noise-polluted barcodes in **a**, the noise effect has been enormously scaled down after 10 steps of denoising as indicated in **b**. From **c** to **f**, there is a clear separation between the intrinsic topological features of protein segment and noise induced topological invariants. The noise effect is continuously reduced, and we are able to identify a persistent barcode pattern, which is an indication of the intrinsic topological invariants of the protein segment. However, unlike the fullerene C_{20}, further denoising of PDB 2GR8 data will remove both the noise and intrinsic structure-related topological information.

However, unlike the situation for fullerene C_{20}, further denoising will remove both the noise and intrinsic structure-related topological information.

Finally, we consider a more realistic example, that is, EMD 1776, obtained from cryo-EM. For EMD 1776, when the isovalue goes to around 0.020, noise begins to emerge. Figure 10 depicts noise in EMD 1776 data. The isovalues for Figures 10**a**–**d** are 0.020, 0.010, 0.005, and 0.000, respectively. The geometric flow-based denoising method is employed. Persistent homology results are demonstrated in Figure 11. The numbers of denoising steps in Figures 11**a**–**f** are 10, 40, 80, 120, 160, and 200, respectively. In this case, the noise induced topological invariants have been gradually weakened but have not been cleaned up.

From the aforementioned analysis, some common features can be unveiled. First, signal-related topological features tend to be buried near the left end of the filtration parameter during the denoising process. Second, topological features corresponding to signal and noise begin to separate as the denoising procedure advances. Third, intrinsic topological invariants associated with the 'signal' are essentially preserved over the denoising process. These features can be used to guide the evaluation and thresholding of the denoising process.

When the persistent features in the barcode representation are identified, one can retrieve the intrinsic barcodes of the signal by simply setting up a noise threshold and removing all barcodes with length less than it [105]. Figure 12 demonstrates this technique. Figure 12**a** shows the barcodes after 20 iterations of the noisy fullerene C_{20} data as given in Figure 8**b**. Figure 12**b** illustrates the barcodes after 20 iterations of the noisy PDB 2GR8 beta segment data as given in Figure 9**c**. Figure 12**c** depicts the barcodes after 40 iterations of the noisy EMD1776 data as in presented Figure 11**b**. Figures 12**d**–**f** display the barcodes of the aforementioned three cases with a noise threshold of 0.1, that is, removing all barcodes with their lengths shorter than 0.1.

DOI: 10.1002/cnm

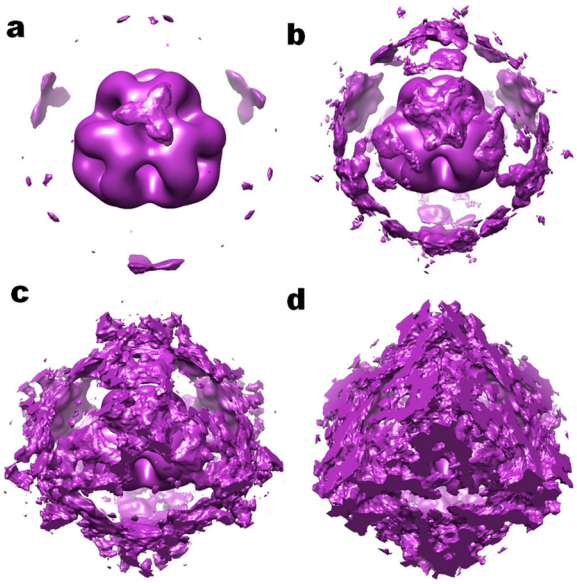

Figure 10. The original noise in EMD 1776 data. The isovalues for **a**, **b**, **c**, and **d** are 0.020, 0.010, 0.005, and 0.000, respectively.

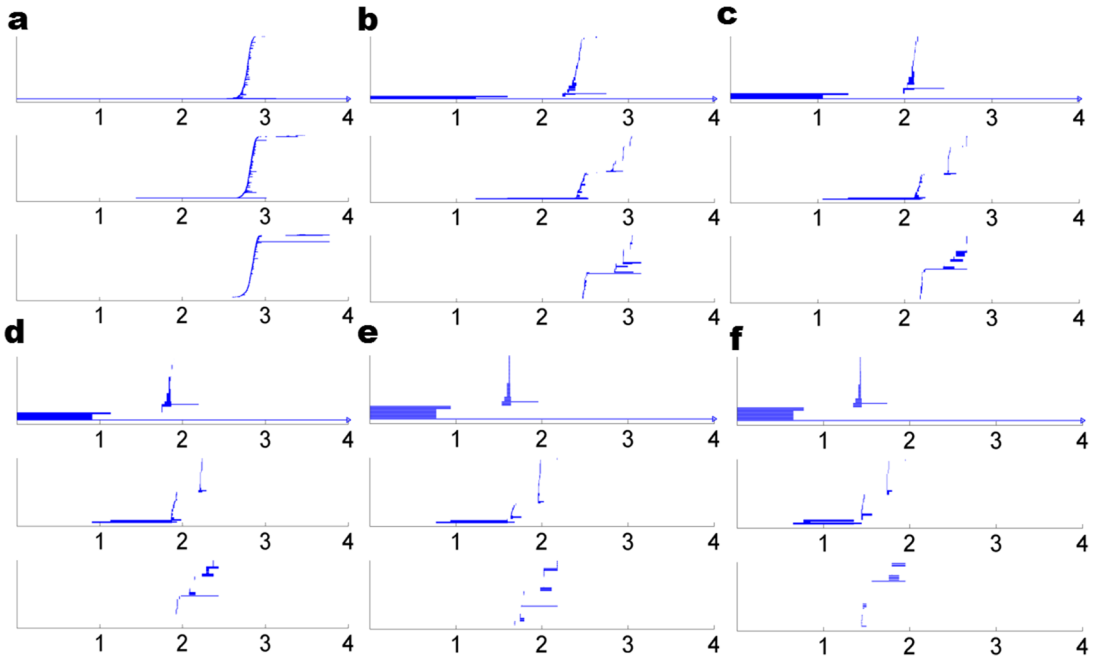

Figure 11. The barcodes representation for the noise removal of contaminated EMD 1776 with SNR 1.0. The denoising steps for **a**, **b**, **c**, **d**, **e**, and **f** are 10, 40, 80, 120, 160, and 200, respectively. In this case, the noise induced topological invariants have been gradually weakened but not eradicated. Compared with the original noise-polluted barcodes in Figure 7 **b**, the noise effect has been enormously reduced after 10 steps of denoising as indicated in **a**. From **b** to **f**, there is a clear separation between the intrinsic topological features of protein segment and noise induced topological invariants. The noise effect continuously wanes, and we are able to identify a persistent barcode pattern, which is an indication of the intrinsic topological invariants of the protein segment. However, unlike the fullerene C_{20}, further denoising will remove both the noise and intrinsic structure-related topological information.

DOI: 10.1002/cnm

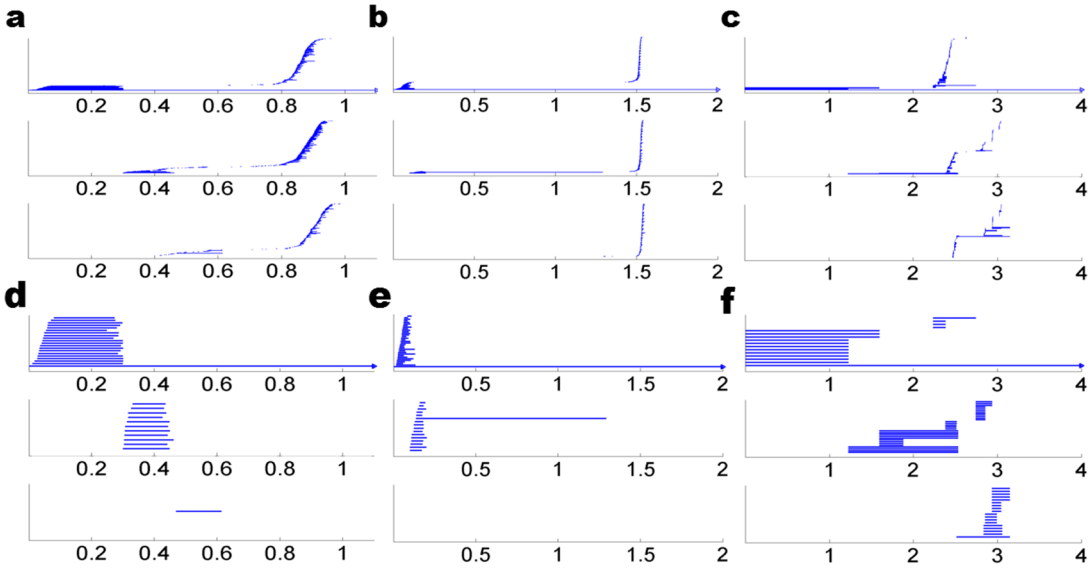

Figure 12. Retrieved barcode patterns through noise thresholding for three systems. Subfigure **a** is the barcode for denoising fullerene C_{20} data with 20 iterations as in Figure 8**b**. Subfigure **b** is the barcode for denoising PDB 2GR8 beta segment data with 20 iterations as in Figure 9**c**. Subfigure **c** is the barcode for denoising EMD 1776 data with 40 iterations as in Figure 11**b**. Barcodes in **d**, **e**, and **f** are all obtained respectively from **a**, **b**, and **c** by setting noise threshold as 0.1, that is, removing all barcodes with their lengths shorter than 0.1.

Another importance aspect is that for cryo-EM data, we just need to consider voxels with isovalue larger than a certain threshold. For instance, in EMD 1776 data, noise begins to emerge when the isovalue goes down to about 0.020 as indicated in Figure 10. As the isovalue decreases, more noise emerges. More importantly, in most cryo-EM data, isovalue can even go below 0.0. These special voxels, as far as we know, do not really represent the desirable biomolecular structure or, more specifically, do not directly reflect the desirable biomolecular structure. Usually, a recommended isovalue is specified for each cryo-EM data. The meaningful structure information can be only derived from voxels with value near this specified isovalue. From our persistent analysis, we believe that all the voxels with isovalue larger than certain threshold can be related to their inner structure properties. In order not to overlook certain potential structure pattern, in our persistent analysis, we just ignore all voxels with isovalue smaller than 0.0. This is usually done by assigning all negative isovalues to 0.0.

3.3. Case study: EMD 5729

Finally, we consider EMD 5729, a supramolecular filamentous complex [106], to demonstrate the application of topological denoising in cryo-EM data analysis. We first process the EMD 5729 data with 20 steps of noise reduction using our geometric flow method. The resulting data are illustrated in Figure 13 with four different isovalues.

Figure 14 depicts barcodes computed for EMD 5729. The β_0 pattern in Figure 14**a** shows a large number of bars of similar lengths, which indicates only one type of protein monomers. Four relatively long persistent bars in the highlighted circle indicate that there are four major pieces in the structure. The β_1 panel appears to be heavily contaminated by noise, which suggests the necessity for a denoising process. The denoised β_0 topological persistent patterns in Figures 14**b**–**d** confirm that only one type of bars can be found, which means that there is only one type of polymer monomers. Additionally, four relatively long β_0 bars confirm that there are four polymer chains. The β_1 bars in Figures 14**b**–**d** are relatively consistent over the denoising process and have very similar lengths, which suggest these polymers are evenly distributed and form certain tunnel type of global structures with high symmetry. The long β_1 bar in Figure 14**d** indicates a large cylinder structure. Indeed, Figure 13 shows that protein monomers form four helix polymers and then bind together to result in a hollow cylinder structure [106].

DOI: 10.1002/cnm

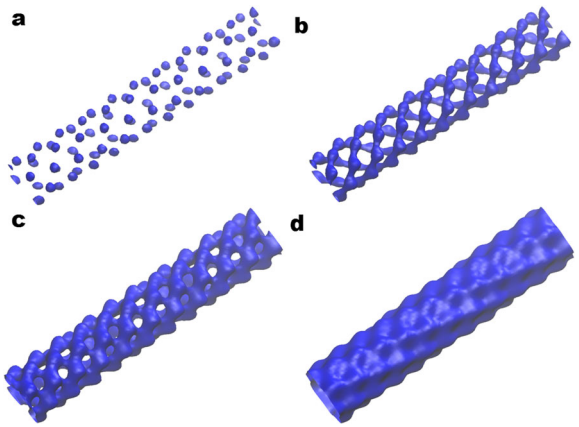

Figure 13. Isosurfaces of EMD 5729 data after 20 steps of noise reduction. Images in **a**, **b**, **c** , and **d** are extracted from isovalues 0.35, 0.30, 0.25, and 0.10, respectively.

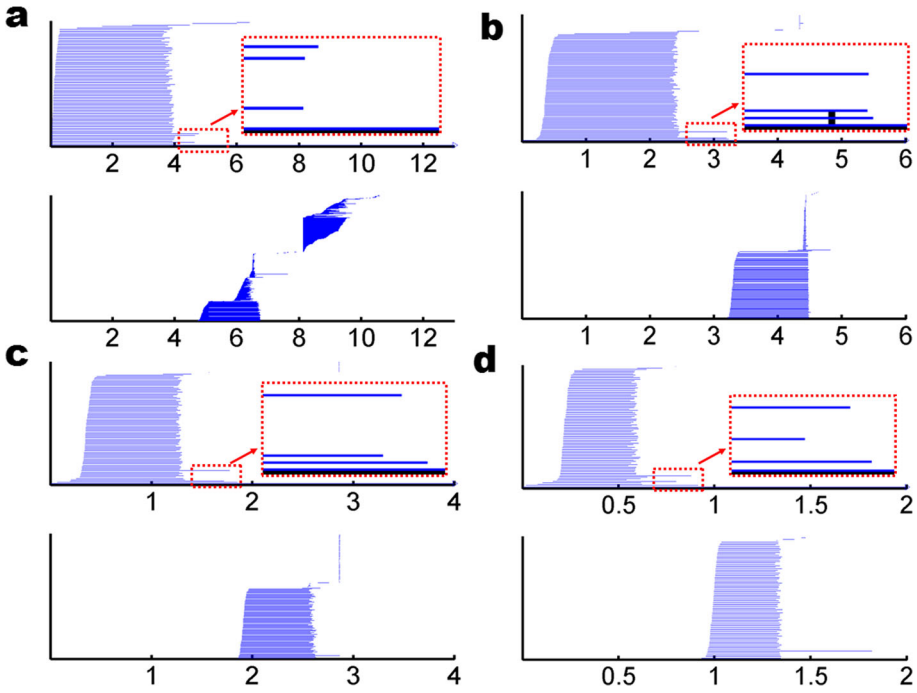

Figure 14. The β_0 and β_1 barcodes for EMD 5729 in the noise reduction process. The barcodes for the original data are shown in **a**. Here, **b**, **c**, and **d** are barcodes after 10, 20, and 30 steps of noise reduction, respectively. Four relatively long β_0 bars in the highlighted circles indicate that there are four polymer strands in the structure.

4. PERSISTENT HOMOLOGY ANALYSIS OF MICROTUBULE

In this section, persistent homology and topological denoising approaches are applied to the microtubule cryo-EM data analysis.

4.1. Microtubule structure EMD 1129

Microtubule is a cytoskeleton component of eukaryotic cells. It plays important roles in maintaining the structure or geometric shape of the cell, supporting intracellular transport, and facilitating cell division (mitosis and meiosis) [107]. Microtubule has a long hollow cylinder structure made up of polymerized α-tubulin and β- tubulin dimers. These hetero-dimers bind head to tail into

DOI: 10.1002/cnm

protofilaments, which further combine with each other in a parallel manner. The hollow cylinder structure of microtubule is finalized by attaching about 13 protofilaments with each other side by side. Although the crystal structures of α-tubulin and β-tubulin are available, experimental microtubule structure data are usually in low resolution and inadequate to separate between neither α-tubulin and β-tubulin nor intra-dimer interface and inter-dimer interface. Recently, a microtubule intermediate structure data (EMD 1129) with a 12 Angstrom resolution was obtained [108]. Based on these data, we demonstrate how to make use of persistent homology and topological denoising methods to aid the modeling of cryo-EM structures.

4.2. Coarse-grained models for microtubule

For cryo-EM data of low resolution or intermediate resolution, it is well-known that atomic scale models are unreliable. As such, coarse-grained models in terms of residues or even proteins can be useful. In this work, we propose a coarse-grained model for microtubule.

The EMD 1129 data seriously suffer from noise as demonstrated in Figure 15**a**. To build up a coarse-grained model, a topological denoising process as discussed in the previous section is employed. Surfaces extracted from denoising data are illustrated in Figure 15. It can be seen that, after ten iterations, the noise intensity is dramatically reduced and the basic geometry of the structure is preserved. More iterations are considered because of the requirement of persistent homology analysis, which will be discussed later in Section 4.3.

Based on the denoising data processed with 10 iterations, we analyze the structure of this microtubule intermediate and build up coarse-grained models. Through the observation of different isosurfaces from the data, it can be found that this microtubule intermediate has a unique helix backbone configuration. We assume that the center of each component protein has the largest electron density value. Through a threshold value of 31.2, we are able to identify centers of these component proteins; the helix backbone is thus constructed. Figure 16**a** demonstrates the construction of our theoretical model with two types of protein monomers. A helix function as illustrated in Figure 16**b** is parametrized based on fitting with these marked positions. It is seen that in each circle, there are about 12 blue color nodes, representing 12 protein monomers of the same type, which we denote as type 'I'. However, there are 12 type 'II' monomers missing in this model as they have a slightly lower electron density value. These type 'II' protein monomers are further accounted by adding 12 black nodes evenly distributed on the helix curve so that they can pair up with type 'I' protein

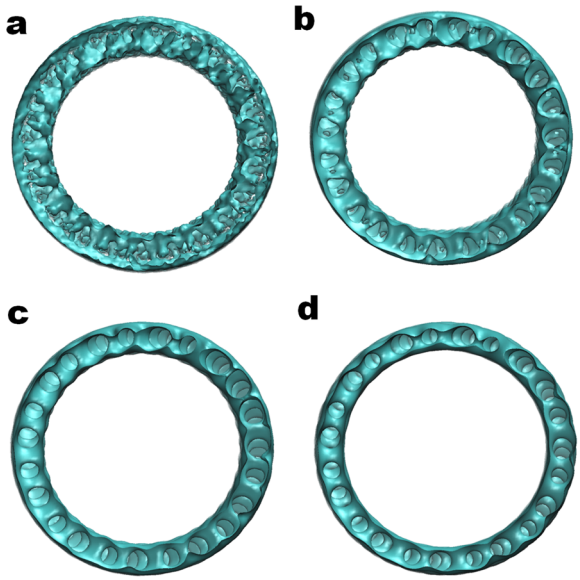

Figure 15. EMD 1129 data preprocessing. **a**: Surface extracted from original data with isovalue 16. **b**, **c**, and **d** are surfaces extracted from denoising data with 10, 20, and 40 iterations, respectively.

DOI: 10.1002/cnm

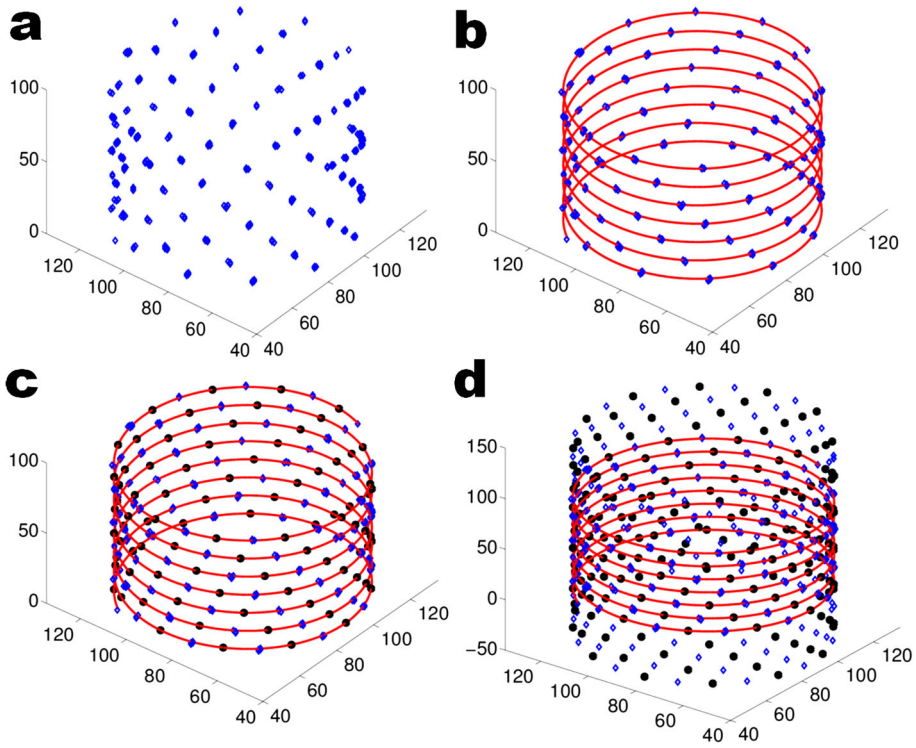

Figure 16. The helix backbone reconstruction for the theoretical model with two types of protein monomers. In **a**, positions of voxels with isovalue larger than 31.2 are marked with solid blue diamonds for EMD 1129 denoising data after 10 iterations. A helix backbone structure can be clearly identified. In **b**, a helix function is parametrized based on fitting with marked positions in **a**. In each circle, there are 12 independent blue color nodes, which represent the same type protein monomers (we call them Type 'I' monomers). In **c**, 12 black color nodes are added evenly on helix curve in each circle, representing Type 'II' protein monomers. Together with the blue nodes, they are used to compare with the EMD1129 experimental data, in which there are 24 proteins in each circle. In **d**, three and four more layers of proteins are added to two ends of the helix curve to eliminate the boundary effect in our density function.

monomers as demonstrated in Figure 16**c**. Finally, to avoid the boundary effect in our density function evaluation, about three layers are added to the top and bottom parts of the helix structure.

To avoid the complexities and present our persistent homology analysis directly and clearly, a simple coarse-grained model is considered. Basically, we use ellipsoids to represent protein monomers. The density function of our microtubule intermediate model can be expressed as

$$\rho(x,y,z)=\sum_i W_i e^{-\left[\left(\frac{x-x_i}{\sigma_i^x}\right)^2+\left(\frac{y-y_i}{\sigma_i^y}\right)^2+\left(\frac{z-z_i}{\sigma_i^z}\right)^2\right]}, \tag{13}$$

where $\rho(x,y,z)$ is the density function of the model, parameter W_i is the weight coefficient, and parameters σ_i^x, σ_i^y, and σ_i^z are ellipsoid resolutions. Coordinates (x_i, y_i, z_i) denote the positions of protein monomer centers. To eliminate the boundary effect, the simulated models incorporate extra protein elements as illustrated in Figure 16**d**. All the aforementioned parameters are optimized by the least-square fitting using the denoising data. It is found that, in any xy-cross section, the electron density of microtubule tightly concentrates in a highly symmetric ring-band region as shown in Figure 17. This ring-band region can be characterized by an inner circles and an outer circle. These circles share the same center at grid position (86, 86), and their radii are 37 and 48 voxels, respectively. As discussed in the literature [109], only the regions that have sufficiently large density values should be included in the fitting. In the present work, the fitting region is limited to the region within two dash-line circles in the xy-cross section as illustrated in Figure 17.

DOI: 10.1002/cnm

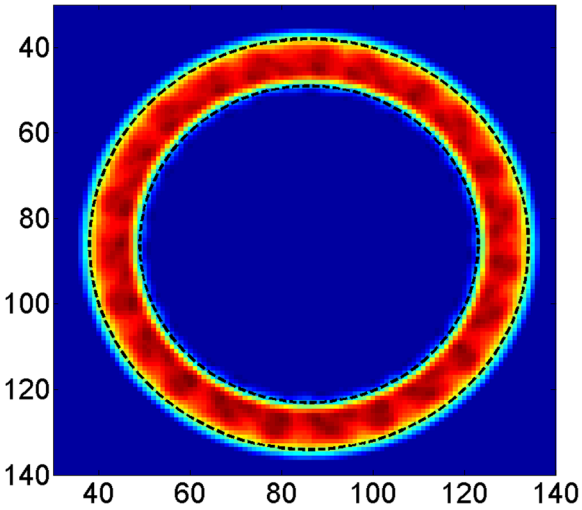

Figure 17. Illustration of the fitting region. The average isovalues over the z-axis, that is, $\frac{1}{n_z}\sum_k \rho^e(x_i, y_j, z_k)$, are given. The fitted region is indicated by two dash lines, that is, the inner circle and the outer circle. Two circles share the same center point coordinate (86, 86), and their radii are 37 and 48 voxels, respectively.

In 3D, a thick-layer cylinder region that encompasses the structure is considered as the fitting region and is denoted as V.

To obtain the optimized fitting parameters, two evaluation coefficients, that is, cross-correlation coefficients (CCF)

$$\mathrm{CCF} = \frac{\sum_{j\in V}\rho_j^e \sum_{j\in V}\rho_j}{\sqrt{\sum_{j\in V}\left(\rho_j^e\right)^2 \sum_{j\in V}(\rho_j)^2}}, \tag{14}$$

and correlation coefficients (CF)

$$\mathrm{CF} = \frac{\sum_{j\in V}\left(\rho_j^e - \bar{\rho}_j^e\right)\sum_{j\in V}(\rho_j - \bar{\rho}_j)}{\sqrt{\sum_{j\in V}\left(\rho_j^e - \bar{\rho}_j^e\right)^2 \sum_{j\in V}(\rho_j - \bar{\rho}_j)^2}} \tag{15}$$

are used [109]. Here, $\rho_j = \rho(x_j, y_j, z_j)$, ρ_j^e is the experimental electron-density value at (x_j, y_j, z_j) after 10 denoising iterations, $\bar{\rho}$ denotes the average of ρ, and parameter V is the fitting region as stated above.

Based on the helix backbone configuration, three theoretical models are constructed for microtubule structure. Using the least-square fitting, we determine the fitting parameters, that is, W_i and σ in Equation (13) for these models. Results are evaluated by aforementioned CCF and CF criteria. In the first model, only one type of ellipsoids is used, that is, one type of monomers with $W_i = 42$ for all protein monomers. The second model has two types of monomers (Figure 16). Two types of ellipsoids with $w_1 = 42$ and $w_2 = 38$ are considered by setting $\{W_i; W_i = w_1 \text{ or } w_2\}$. The third model is dimer one. In this model, location modification is considered to generate dimers by shifting the type 'II' protein monomers simultaneously along the helix backbone closer to type 'I' protein monomers slightly. In this manner, we discriminate between intra-dimer and inter-dimer distances. All ellipsoids are parametrized uniformly by assuming $\sigma_i^x = 24, \sigma_i^y = 24$, and $\sigma_i^z = 22$. The unit for resolution parameters is Angstrom (Å), and voxel spacing is 4Å. The results are illustrated in Figure 18. All three models look similarly. Actually, the CCFs for the three models are 0.9601, 0.9607, and 0.9604, respectively. The CFs for them are 0.7392, 0.7436, and 0.7662, respectively. The difference in these coefficients is very small. Therefore, we cannot determine with a good confidence that one model is definitely better than others. We therefore encounter a standard ill-posed inverse problem by using the structural optimization approach.

DOI: 10.1002/cnm

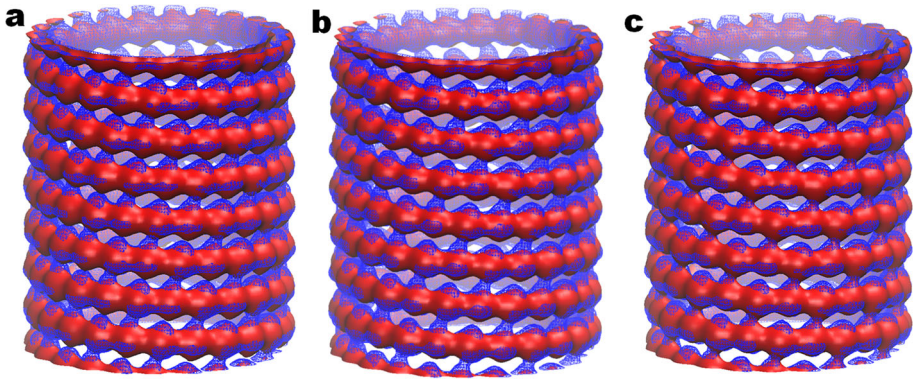

Figure 18. Three theoretical models for microtubule structures constructed from fitting the experimental data using proteins on helix backbone curve. Based on the coarse-grained representation, we use an ellipsoid to represent a protein monomer. In **a**, only one type of ellipsoids is used. In **b**, two types of ellipsoids with different weight functions are used. In **c**, two types of ellipsoids with two different weight functions and modified locations are considered. The isosurfaces in **a** and **b** look similar to each other. However, they have dramatically different topological behaviors. In **c**, we systematically shift type 'II' monomers to form dimmers with type 'I' monomers. The blue meshed surfaces are obtained from the denoising data, and red solid surfaces are computed from the corresponding three theoretical models.

4.3. Persistent homology based microtubule model evaluation

However, if we pay attention to the β_1 patterns of holes formed between the upper and lower helix circles as illustrated in Figure 18**c**, it can been seen that only the third model preserves these features. These linkage properties are directly related to biomolecular flexibility and functional properties. Therefore, it is important for us to characterize and capture them in our models. Unfortunately, least square optimization approach is insensitive to these little structural characteristics. Therefore, techniques that are sensitive to geometric variations are required to guide the structure determination. We show that persistent homology is a desirable technique for detecting geometric defects in the rest of this section.

To understand how the persistent homology can be employed to guide our model construction and evaluation, we investigate the topological fingerprint of the microtubule intermediate structure. As described earlier, because of the noise, a denoising process is required. The geometric flow-based denoising algorithm is used with different numbers of iterations. The denoising data are carefully analyzed and the topological persistence results are demonstrated in Figure 19. It can be seen that a special pattern begin to emerge when the number of iterations approaches 20. More specifically, groups of bars in both β_0 and β_1 panels appear with distinctive persisting patterns.

In the β_0 panel of Figure 19, bars can be roughly grouped into three parts from the top to the bottom, that is, an irregular 'hair-like' part on the top, a narrow regular 'body' part in the middle, and a large regular 'base' part in the bottom. Topologically, these parts represent different components in the microtubule intermediate structure. The irregular 'hair-like' part corresponds to the partial monomer structures located on the top and the bottom boundaries of the structure. As can be seen in Figure 18, each monomer has lost part of the structure at the boundary regions. The regular 'body' and 'base' parts are basically related to two types of monomers in the middle region where the structure is free of boundary effect. From the barcodes, it can be seen that 'body' part has a later 'birth' time and earlier 'death' time compared with the 'base' barcode part. This is due to the reason that this type of monomers has relatively lower electron density. As the filtration is defined to go from highest electron density values to lowest ones, their corresponding barcodes appear later. Their earlier death time, however, is due to the reason that they form dimers with the other type of monomers represented by the 'base' barcode part. It can be derived from these nonuniform behavior that monomers are not equally distributed along the helix backbone structure. Instead, two adjacent different types of monomers form a dimer first; then all these dimers simultaneously connect with each other as the filtration goes on. Moreover, from the analysis in the previous section, it is obvious

DOI: 10.1002/cnm

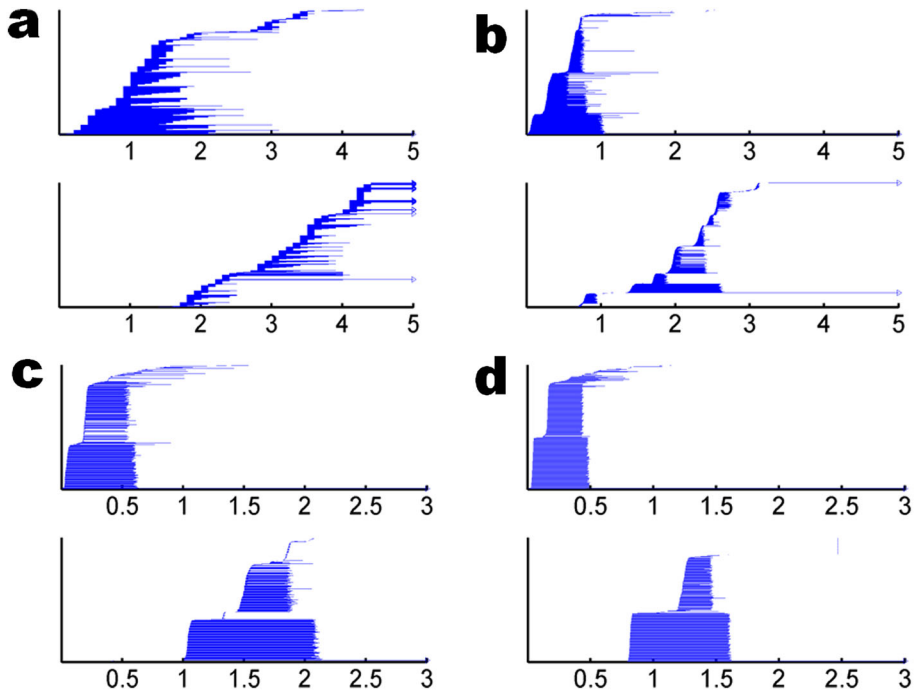

Figure 19. Topological persistence of β_0 (top row) and β_1 (bottom row) generated from original and preprocessed EMD 1129 data. **a** is the barcodes of the original EMD1129 data. **b**, **c**, and **d** are barcodes for EMD 1129 data after 10, 20 and 40 denoising iterations. A special pattern, that is, two individual bands of bars in both β_0 and β_1, persists in **c** and **d**.

to see that the 'body' and 'base' parts are topological representations of type 'II' monomer and type 'I' monomer, respectively.

For the β_1 panel of Figure 19, there also exists a consistent pattern when the denoising process passing a certain stage. Two distinctive types of barcodes can be identified in the fingerprint, that is, a shorter band of barcodes on the top and a longer band of bars on the bottom. Topologically, these β_1 bars correspond to the rings formed between two adjacent helix circles of monomers or dimers. During the filtration, dimers are formed between type 'I' and type 'II' monomers; soon after that, all dimers connect with each other and form the helix backbone. As the filtration goes on, type 'I' monomers from the upper helix circle first connect with type 'II' monomers at the lower circle. Geometrically, this means six monomers, three ('I–II–I') from the upper layer and three ('II–I–II') from the lower layer, form a circle. As the filtration goes further, this circle evolves into two circles when two middle monomers on two layers also connect. However, these two circles do not disappear simultaneously. Instead, one persists longer than the other. This entire process generates the unique topological fingerprint in β_1 barcodes.

The topological fingerprint extracted from the denoising process can be used to guide the construction and evaluation of our microtubule models. To this end, we analyze the topological features of three theoretical models. Our persistent homology results for three models are demonstrated in Figures 20**a**–**c**, respectively. It can be seen that all the three models are able to capture the irregular 'hair' region in their β_0 barcode chart. From the topological point of view, the first model is the poorest one. It fails to capture the regular fingerprint patterns in both β_0 and β_1 panels of the original cryo-EM structure in Figure 19**d**. Using two different weight functions to represent two types of monomers, the second model delivers a relatively better topological result. It is able to preserve part of the difference between type 'I' and type 'II' barcodes in the β_0 panel. In β_1 panel, some nonuniform barcodes emerge. The persistent homology results are further improved in the third model when the intra-dimer and inter-dimer interactions are considered. In our third model, fingerprint patterns of the cryo-EM structure in both β_0 and β_1 panels of Figure 19**d** are

DOI: 10.1002/cnm

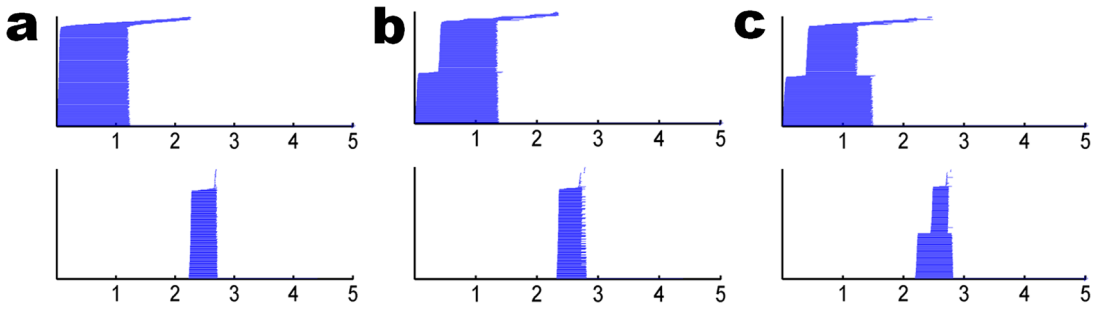

Figure 20. Topological fingerprints of β_0 (top row) and β_1 (bottom row) generated from three theoretical models as depicted in Figure 18. **a** is the barcodes of the first model with only one type of monomers. **b** is the barcodes of the second fitted model with only two types of monomers using different weight functions. **c** is barcodes of the third model with two types of monomers using different weight functions and modified locations. It can be seen that only the final model is able to capture the topological properties of the intrinsic topological fingerprints of cryo-EM data in Figure 19**d**.

essentially recovered by those of Figure 20**c**. Even though their resolutions are different, their shapes are strikingly similar.

4.4. Discussion

The essential topological features that are associated with major topological transitions of the original cryo-EM structure are illustrated in Figures 21$\mathbf{a_1}$–$\mathbf{a_4}$. As shown in Figures 21$\mathbf{b_1}$–$\mathbf{b_4}$, these features have been well preserved during the denoising process. Our best predicted model is depicted in Figures 21$\mathbf{c_1}$–$\mathbf{c_4}$. In these figure labels, subscripts 1, 2, 3, and 4 denote four topological transition stages in the filtration process, namely, hetero-dimmer formation, large circles formation, evolution of one large circle into two circles, and finally, death of one of two circles. By the comparison of denoising results (Figures b_1, b_2, b_3, and a_4) with original structures (Figures a_1, a_2, a_3, and a_4), it is seen that, in the noise reduction process, although some local geometric and topological details are removed, fundamental topological characteristics are well preserved. As illustrated in Figure 19, using the persistent homology description, these fundamental topological characteristics are well preserved in topological persistence patterns, which are further identified as fingerprints of the microtubule intermediate structure. We believe that topological fingerprints are crucial to the CIA of the biological structure. As demonstrated in Figures 21$\mathbf{c_1}$–$\mathbf{c_4}$, once our model successfully reproduces the topological fingerprints, the simulated structure is able to capture the essential topological characteristics of the original one. Moreover, through the analysis in Section 4.3, it can be seen that to reproduce the topological fingerprint of EMD 1129, two conditions are essential. The first is the creation of two types of monomers. The second is the differentiation of intra-dimers and inter-dimers. Biologically, these requirements mean the following: (1) there are two types of monomers, that is, α-tubulin monomers and β-tubulin monomers; and (2) intra-dimers and inter-dimers should behave differently from hetero-dimers.

It also should be noticed that a higher correlation coefficient may not guarantee the success of the model, especially when the original data is of low resolution and low SNR. As can be seen in Section 4.2, our three theoretical models have very similar fitting coefficients. The second model even has a slightly higher cross-correlation coefficient. However, only the third model is able to reproduce the essential topological features of the original cryo-EM data. This happens as topological invariants, that is, connected components, circles, loops, holes, or void, tend to be very sensitive to 'tiny' linkage parts, which are almost negligible in the density fitting process, compared with the major body part. We believe these linage parts play important roles in biological system especially in macro-proteins and protein–protein complexes. Different linkage parts generate different connectivity and thus can directly influence biomolecular flexibility, rigidity, and even its functions. By associating topological features with geometric measurements, our persistent homology analysis is able to distinguish these connectivity parts. Therefore, persistent homology is able to play a unique role in protein design, model evaluation, and structure determination.

DOI: 10.1002/cnm

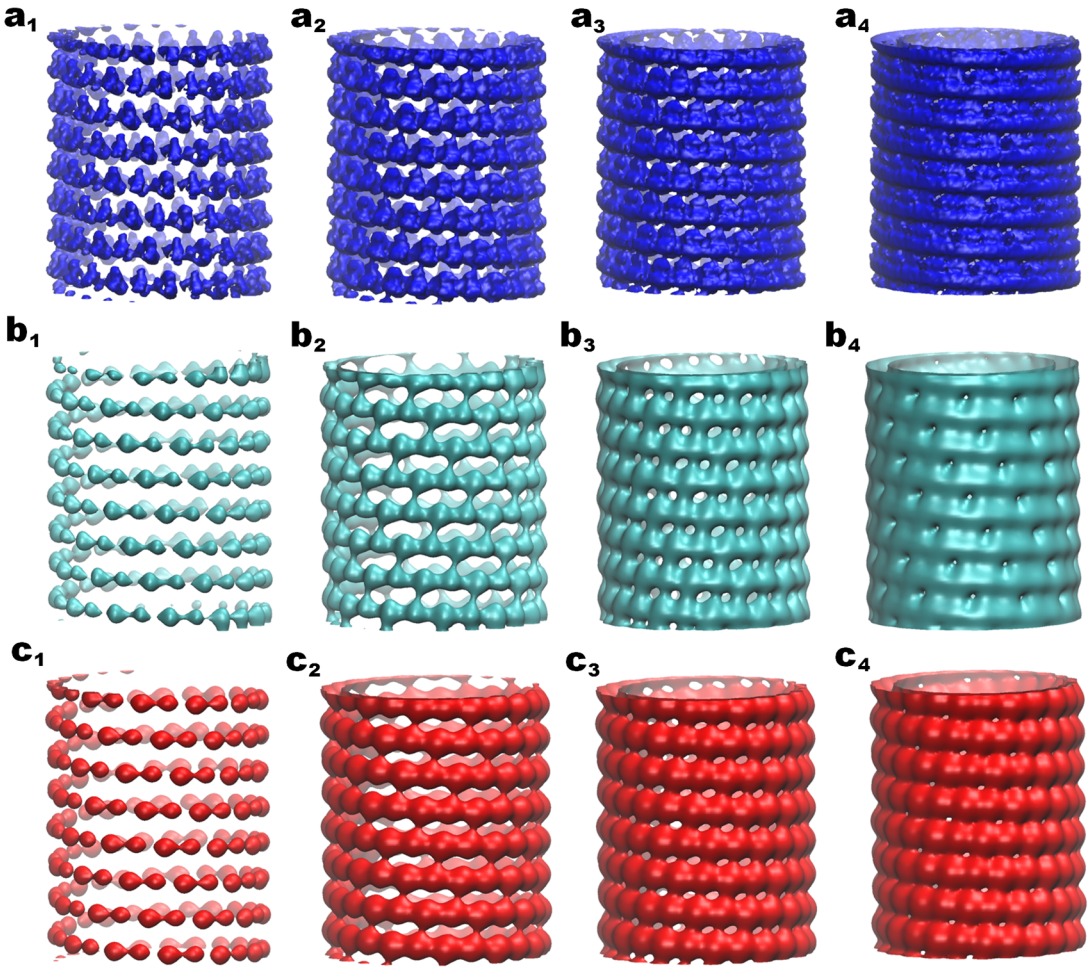

Figure 21. The topological transitions of microtubule geometry. Here, **a**, **b**, and **c** are the original data, denoising data (40 iterations) and theoretical model (the third model), respectively. The subscripts 1 to 4 represent four topological transitions during the filtration process, that is, hetero-dimmer formation, large circles formation, evolution of each large circles into two circles, and finally, death of one of two circles. It can be seen that these topological transitions are well-preserved in our denoising data and theoretical model, which explains the excellent topological consistency between them.

Inverse problems arise in many branches of science, engineering, and technology, including medical imaging, molecular imaging, computer vision, computer-aided tomography, acoustic tomography, geophysical imaging, remote sensing, nondestructive testing, to name only a few. Density fitting via the variational principle or optimization is a common practice in many inverse problems. This approach emphasizes the matching of high-density regions, while it is insensitive to the mismatch in the low density parts. Consequently, inverse problems are ill-posed in the sense that small errors in the data may lead to entirely different solutions. Persistent homology maps out the topological fingerprints of the data over the whole range of density variations and is very sensitive in detecting geometric defects or structural mismatch at various density levels. Therefore, persistent homology has a great potential for resolving ill-posed inverse problems.

5. CONCLUSION

Cryo-electron microscopy (Cryo-EM) is a major workhorse for the investigation of subcellular structures, organelles, and large multiprotein complexes. However, cryo-EM techniques and algorithms are far from mature because of limited data quality and/or sample stability, low signal to noise

DOI: 10.1002/cnm

ratio (SNR), low resolution, and the high complexity of the underlying biological structures. Persistent homology is a new branch of topology that is known for its potential in the classification, identification and analysis (CIA) of big data. In this work, persistent homology is, for the first time, employed for the cryo-EM data CIA.

Methods and algorithms for the geometric and topological modeling are presented. Here, geometric modeling, such the generation of density maps for proteins or other molecules, is employed to create known data sets for validating topological modeling algorithms. We demonstrate that cryo-EM density maps and fullerene density data can be effectively analyzed by using persistent homology.

Because topology is very sensitive to noise, the understanding of the topological signature of noise is a must in cryo-EM CIA. We first investigate the topological fingerprint of Gaussian noise. We reveal that for the Gaussian noise, its topological invariants, that is, β_0, β_1, and β_2 numbers, all exhibit the Gaussian distribution in the filtration space, that is, the space of volumetric density isovalues. At a low SNR, signal and noise are inseparable in the filtration space. However, after denoising with the geometric flow method, there is clear separation between signal and noise for various topological invariants. As such, a simple threshold can be prescribed to effectively remove noise. For the case of low SNR, the understanding of noise characteristic in the filtration space enables us to use persistent homology as an efficient means to monitor and control the noise removal process. This new strategy for noise reduction is called topological denoising.

Persistent homology has been applied to the theoretical modeling of a microtubule structure (EMD 1129). The backbone of the microtubule has a helix structure. Based on the helix structure, we propose three theoretical models. The first model assumes that protein monomers form the helix structure. The second model adopts two types of protein monomers evenly distributed along helix chain. The last model utilizes a series of protein dimers along the helix chain. These models are fitted with experimental data by the least square optimization method. It is found that all the three models give rise to similar high correlation coefficients with the experimental data, which indicates that the structural optimization is ill-posed. However, the topological fingerprints of three models are dramatically different. In the denoising process, the cryo-EM data of the microtubule structure demonstrate a persistent pattern that can be recognized as the intrinsic topological fingerprint of the microtubule structure. By the careful examination of the fingerprint, we reveal two essential topological characteristics that discriminate the protein dimers from the monomers. As such, we conclude that only the third model, that is, the protein dimer model, is able to capture the intrinsic topological characteristics of the cryo-EM structure and must be the best model for the experimental data. It is believed that the present work offers a novel topology-based strategy for resolving ill-posed inverse problems.

ACKNOWLEDGEMENTS

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. Nickell S, Kofler C, Leis AP, Baumeister W. A visual approach to proteomics. *Nature Reviews Molecular Cell Biology* 2006; **7**:225–230.
2. Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. *Nature* 2007; **450**:973–982.
3. Leis A, Rockel B, Andrees L, Baumeister W. Visualizing cells at the nanoscale. *Trends in Biochemical Sciences* 2009; **34**:60–70.
4. Tocheva EI, Li Z, Jensen GJ. Electron cryotomography. *Cold Spring Harbor Perspectives in Biology* 2010; **2**:A003442.
5. Volkmann N. Methods for segmentation and interpretation of electron tomographic reconstructions. *Methods Enzymol* 2010; **483**:31–46.
6. Kühlbrandt W. Cryo-EM enters a new era. *eLife* 2014; **3**:1–4.

DOI: 10.1002/cnm

7. Abeysinghe SS, Baker M, Chiu W, Tao J. Segmentation-free skeletonization of grayscale volumes for shape understanding. *IEEE International Conference on Shape Modeling and Applications*, SMI Stony Brook, N.Y, 2012; 63–71.
8. Biswas A, Si D, Al Nasr K, Ranjan D, Zubair M, He J. Improved effeciency in Cryo-EM secondary structure topology determination from inaccurate data. *Journal of Bioinformatics and Computational Biology* 2012; **10**:1242006.
9. Ju T, Baker ML, Chiu W. Computing a family of skeletons of volumetric models for shape description. *Computer-Aided Design* 2007; **39**:352–360.
10. Baker ML, Jiang W, Wedemeyer WJ, Rixon FJ, Baker D, Chiu W. Ab initio modeling of the herpesvirus VP26 core domain assessed by cryoEM density. *PLoS Computational Biology* 2006; **2**:e146.
11. Sun W, He J. Native secondary structure topology has near minimum contact energy among all possible geometrically constrained topolgies. *Proteins: Structure, Function and Bioinformatics* 2009; **77**:159–173.
12. Lu Y, He J. Deriving topology and sequence alignment for helix skeleton in low resolution protein density maps. *Journal of Bionformatics and Computational Biology* 2008; **8**:183–201.
13. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 2004; **25**: 1605–1612.
14. Zhenga SQ, Keszthelyi B, Branlunda E, Lyleb JM, Braunfelda MB, Sedatb JW, Agarda DA. UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction. *Journal of Structural Biology* 2007; **157**:138–147.
15. Amat F, Moussavi F, Comolli LR, Elidan G, Downing KH, Horowitz M. Markov random field based automatic image alignment for electron tomography. *Journal of Structural Biology* 2008; **161**:260–275.
16. Hrabe T, Chen Y, Pfeffer S, Cuellar LK, Mangold AV, Forster F. PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *Journal of Structural Biology* 2012; **178**:178–188.
17. Kremer JR, Mastronarde DN, McIntosh JR. Computer visualization of three-dimensional image data using IMOD. *Journal of Structural Biology* 1996; **116**:71–76.
18. Ress D, Harlow ML, Schwarz M, Marshall RM, McMahan UJ. Automatic acquisition of fiducial markers and alignment of images in tilt series for electron tomography. *J Electron Microsc (Tokyo)* 1999; **48**:277–287.
19. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. *Methods Enzymol* 2004; **383**:66–93.
20. Stoschek A, Hegerl R. Denoising of electron tomographic reconstructions using multiscale transformations. *Journal of Structural Biology* 1997; **120**:257–265.
21. Frangakis AS, Hegerl R. Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *Journal of Structural Biology* 2001; **135**:239–250.
22. Fernandez JJ, Li S. An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms. *Journal of Structural Biology* 2003; **144**:152–161.
23. Fernandez JJ. Tomobflow: feature-preserving noise filtering for electron tomography. *BMC Bioinformatics* 2009; **178**:1–10.
24. Tomasi C, Manduchi R. Bilateral filtering for gray and color images. *Proceedings of the 1998 IEEE International Conference on Computer Vision*; **98**:839–846.
25. Jiang W, Baker ML, Wu Q, Bajaj C, Chiu W. Applications of a bilateral denoising filter in biological electron microscopy. *Journal of Structural Biology* 2003; **144**:114 –122.
26. Pantelic RS, Rothnagel CY, Huang R, Muller D, Woolford D, Landsberg MJ, McDowall A, Pailthorpe B, Young PR, Banks J, Hankamer B, Erickssson G. The discriminative bilateral filter: an enhanced denoising filter for electron microscopy data. *Journal of Structural Biology* 2006; **155**:395–408.
27. van der Heide P, Xu XP, Marsh BJ, Hanein D, Volkmann N. Efficient automatic noise reduction of electron tomographic reconstructions based on iterative median filtering. *Journal of Structural Biology* 2007; **158**:196–204.
28. Tsai K, Ma J, Ye D, Wu J. Curvelet processing of MRI for local image enhancement. *International Journal for Numerical Methods in Biomedical Engineering* 2012; **28**:661–677.
29. Pan M, Tang J, Rong Q, Zhang F. Medical image registration using modified iterative closest points. *International Journal for Numerical Methods in Biomedical Engineering* 2011; **27**:1150–1166.
30. Radaelli AG, Peiro J. On the segmentation of vascular geometries from medical images. *International Journal for Numerical Methods in Biomedical Engineering* 2010; **26**:3–34.
31. Fujishiro I, Takeshima Y, Azuma T, Takahashi S. Volume data mining using 3D field topology analysis. *IEEE Computer Graphics and Applications* 2000; **20**(5):46–51.
32. Carlsson G. Topology and data. *Bulletin of the American Mathematical Society* 2009; **46**(2):255–308.
33. Ghrist R. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society* 2008; **45**: 61–75.
34. Doraiswamy H, Natarajan V. Computing Reeb graphs as a union of contour trees. *IEEE Transactions on Visualization and Computer Graphics* 2013; **19**:249–262.
35. Frosini P, Landi C. Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis* 1999; **9**(4):596–603.
36. Robins V. Towards computing homology from finite approximations. *Topology proceedings* 1999; **24**:503–532.

DOI: 10.1002/cnm

37. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete & Computational Geometry* 2002; **28**:511–533.
38. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete & Computational Geometry* 2005; **33**: 249–274.
39. Bubenik P, Kim PT. A statistical approach to persistent homology. *Homology, Homotopy and Applications* 2007; **19**:337–362.
40. Edelsbrunner H, Harer J. *Computational topology: an introduction*. American Mathematical Soc., Providence, 2010.
41. Dey TK, Li KY, Sun J, David CS. Computing geometry aware handle and tunnel loops in 3D models. *Association for Computing Machinery Transactions on Graph* 2008; **27**.
42. Dey TK, Wang YS. Reeb graphs: approximation and persistence. *Discrete and Computational Geometry* 2013; **49**(1):46–73.
43. Mischaikow K, Nanda V. Morse theory for filtrations and efficient computation of persistent homology. *Discrete and Computational Geometry* 2013; **50**(2):330–353.
44. 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.
45. 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; **30**(10):1760–1770.
46. 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).
47. Bendich P, Edelsbrunner H, Kerber M. Computing robustness and persistence for images. *IEEE Transactions on Visualization and Computer Graphics* 2010; **16**:1251–1260.
48. Frosini Patrizio, Landi Claudia. Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. *Pattern Recognition Letters* 2013; **34**:863–872.
49. Mischaikow K, Mrozek M, Reiss J, Szymczak A. Construction of symbolic dynamics from experimental time series. *Physical Review Letters* 1999; **82**:1144–1147.
50. Kaczynski T, Mischaikow K, Mrozek M. *Computational Homology*. Springer-Verlag, 2004.
51. Silva VD, Ghrist R. Blind swarms for coverage in 2-D. *In proceedings of robotics: Science and systems*, Cambridge, USA, 2005; 01.
52. Lee H, Kang H, Chung MK, Kim B, Lee DS. Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging* 2012; **31**(12):2267–2277.
53. Horak D, Maletic S, Rajkovic M. Persistent homology of complex networks. *Journal of Statistical Mechanics: Theory and Experiment* 2009; **2009**(03):P03034.
54. Niyogi P, Smale S, Weinberger S. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing* 2011; **40**:646–663.
55. Wang B, Summa B, Pascucci V, Vejdemo-Johansson M. Branching and circular features in high dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 2011; **17**:1902–1911.
56. Rieck B, Mara H, Leitte H. Multivariate data analysis using persistence-based filtering and topological signatures. *IEEE Transactions on Visualization and Computer Graphics* 2012; **18**:2382–2391.
57. Liu X, Xie Z, Yi D. A fast algorithm for constructing topological structure in large data. *Homology, Homotopy and Applications* 2012; **14**:221–238.
58. Di Fabio B, Landi C. 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.
59. 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.
60. Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics* 2012:1–17.
61. Dabaghian Y, Memoli F, Frank L, Carlsson G. A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS Comput Biol* 201208; **8**(8):e1002581.
62. Kloke J, Carlsson G. Topological de-noising: strengthening the topological signal. *ArXiv Preprint ArXiv:0910.5947* 2009.
63. 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.
64. De Silva V, Carlsson G. Topological estimation using witness complexes. *Proceedings of the First Eurographics Conference on Point-based Graphics*, Aire-la-Ville, Switzerland, 2004; 157–166. Eurographics Association.
65. Günther D, Jacobson A, Reininghaus J, Seidel HP, Sorkine-Hornung O, Weinkauf T. Fast and memory-efficient topological denoising of 2D and 3D scalar fields. *IEEE Transactions on Visualization and Computer Graphics* 2014; **20**:12.
66. Bauer U, Schönlieb CB, Wardetzky M. Total variation meets topological persistence: a first encounter. *AIP Conference Proceedings* 2010; **1281**(1):1022.
67. Adler RJ, Bobrowski O, Borman MS, Subag E, Weinberger S. Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applications–a Festschrift for Lawrence D. Brown*, Vol. 6. Institute of Mathematical Statistics, 2010; 124–143.
68. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 2004; **25**(13): 1605–1612.

DOI: 10.1002/cnm

69. 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.
70. Feng X, Xia KL, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules II: Lagrangian representation. *Journal of Computational Chemistry* 2013; **34**:2100–2120.
71. Xia KL, Feng X, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules I: Cartesian representation. *Journal of Computational Physics* 2014; **275**:912–936.
72. Corey RB, Pauling L. Molecular models of amino acids, peptides and proteins. *Review of Scientific Instruments* 1953; **24**:621–627.
73. Lee B, Richards F M. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology* 1971; **55**(3):379–400.
74. Richards FM. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering* 1977; **6**(1):151–176.
75. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *Journal of Computational Chemistry* 2002; **23**:128 –137.
76. Connolly ML. Depth buffer algorithms for molecular modeling. *Journal of Molecular Graphics* 1985; **3**:19–24.
77. Eisenhaber F, Argos P. Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. *Journal of Computational Chemistry* 1993; **14**:1272–1280.
78. Gogonea V, Osawa EE. Implementation of solvent effect in molecular mechanics. 1. Model development and analytical algorithm for the solvent - accessible surface area. *Supramolecular Chemistry* 1994; **3**:303–317.
79. Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* 1996; **38**:305–320.
80. Wei GW. Differential geometry based multiscale models. *Bulletin of Mathematical Biology* 2010; **72**:1562 –1622.
81. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models I: Eulerian formulation. *Journal of Computational Physics* 2010; **229**:8231–8258.
82. Wei GW, Sun YH, Zhou YC, Feig M. Molecular multiresolution surfaces. *arXiv:math-ph* 2005:1 –11.
83. Bates P, Wei GW, Zhao S. The minimal molecular surface. *arXiv:q-bio* 2006; **[q-bio.BM]**.
84. Bates P, Wei GW, Zhao S. Minimal molecular surfaces and their applications. *Journal of Computational Chemistry* 2008; **29**(3):380–91.
85. Bates P, Chen Z, Sun YH, Wei GW, Zhao S. Geometric and potential driving formation and evolution of biomolecular surfaces. *Journal of Mathematical Biology* 2009; **59**:193–231.
86. Wei GW, Zheng Q, Chen Z, Xia K. Variational multiscale models for charge transport. *SIAM Review* 2012; **54**(4):699 –754.
87. Wei GW. Multiscale, multiphysics and multidomain models I: Basic theory. *Journal of Theoretical and Computational Chemistry* 2013; **12**(8):1341006.
88. Xu G, Pan Q, Bajaj CL. Discrete surface modeling using partial differential equations. *Computer Aided Geometric Design* 2006; **23**(2):125–145.
89. Cheng LT, Dzubiella J, McCammon AJ, Li B. Application of the level-set method to the implicit solvation of nonpolar molecules. *Journal of Chemical Physics* 2007; **127**(8).
90. Zhao S. Pseudo-time-coupled nonlinear models for biomolecular surface representation and solvation analysis. *International Journal for Numerical Methods in Biomedical Engineering* 2011; **27**:1964–1981.
91. Zhao S. Operator splitting adi schemes for pseudo-time coupled nonlinear solvation simulations. *Journal of Computational Physics* 2014; **257**:1000 –1021.
92. Xia KL, Opron K, Wei GW. Multiscale multiphysics and multidomain models — flexibility and rigidity. *Journal of Chemical Physics* 2013; **139**:194109.
93. 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.
94. Wei GW. Wavelets generated by using discrete singular convolution kernels. *Journal of Physics A: Mathematical and General* 2000; **33**:8577 –8596.
95. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and completeness of highly likely single domain protein structures. *Proceedings of the National Academy of Sciences of the United States of America* 2006; **103**:2605–2610.
96. 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.
97. Zheng Q, Yang SY, Wei GW. Molecular surface generation using PDE transform. *International Journal for Numerical Methods in Biomedical Engineering* 2012; **28**:291–316.
98. Xia KL, Wei GW. A Galerkin formulation of the MIB method for three dimensional elliptic interface problems. *Computers and Mathematics with Applications* 2014; **68**:719–745.
99. Xia KL, Wei GW. A stochastic model for protein flexibility analysis. *Physical Review E* 2013; **88**:062709.
100. Tausz A, Vejdemo-Johansson M, Adams H. *Javaplex: A Research Software Package for Persistent (Co)homology*, 2011. (Available from: http://code.google.com/p/javaplex).
101. Nanda Vidit. *Perseus: the Persistent Homology Software*. (Available from: http://www.sas.upenn.edu/~vnanda/perseus).
102. Wei GW. Generalized Perona–Malik equation for image restoration. *IEEE Signal Processing Letters* 1999; **6**: 165–167.

DOI: 10.1002/cnm

103. 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.
104. 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.
105. Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A. Confidence sets for persistence diagrams. *The Annals of Statistics* 2014; **42**(6):2301–2339.
106. Qiao Q, Yang CH, Zheng C, Fontán L, David L, Yu X, Bracken C, Rosen M, Melnick A, Egelman EH, Wu H. Structural architecture of the CARMA1/Bcl10/MALT1 signalosome: nucleation-induced filamentous assembly. *Molecular Cell* 2013; **51**(6):766–779.
107. Nogales E, Wang HW. Structural intermediates in microtubule assembly and disassembly: how and why? *Current Opinion in Cell Biology* 2006; **18**(2):179–184.
108. Wang WH, Nogales E. Nucleotide-dependent bending flexibility of tubulin regulates microtubule assembly. *Nature* 2005; **435**:911–915.
109. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein structure fitting and refinement guided by cryo-EM density. *Structure* 2008; **16**(2):295–307.

DOI: 10.1002/cnm