Communications in Information and Systems
Volume 18, Number 4, 269–298, 2018

# Persistent similarity for biomolecular structure comparison

Kelin Xia

Biomolecular structure comparison not only reveals evolutionary relationships, but also sheds light on biological functional properties. However, traditional ways to calculate structure or sequence similarity, which always involve superposition or alignment, are computationally inefficient. In this paper, we propose a new method called persistent similarity, which is based on a newly-invented method in algebraic topology, known as persistent homology. Different from all previous topological methods, persistent homology is able to embed a geometric measurement into topological invariants, thus provides a bridge between geometry and topology. After that, the topological information derived from the persistent homology analysis can be uniquely represented by a series of one-dimensional (1D) persistent Betti functions (PBFs). In this way, any complicated biomolecular structure can be represented as several 1D PBFs, and persistent similarity is defined as the quotient of intersect areas and union areas of any two PBFs. If structures have no significant topological properties, a pseudo-barcode is introduced to insure a better comparison. Further, a multiscale biomolecular representation is introduced through the multiscale rigidity function. It naturally induces a multiscale persistent similarity. The multiscale persistent similarity enables an objective-oriented comparison. Stated differently, it facilitates the comparison of structures at any particular scale of interest. Finally, the proposed method is validated by four different cases. It is found that the persistent similarity can be used to describe the intrinsic similarities and differences between the structures very well. Particularly, it delivers one of the best results for isomer total curvature energy prediction.

## 1. Introduction

The most prominent feature of biological science in the $21^{st}$ century is its transition from an empirical, qualitative and phenomenological discipline to

a comprehensive, quantitative and predictive one. With the accumulation of gigantic structure and sequence data in Protein Data Bank[1], Gene Bank (GenBank)[2], and protein structure classification databanks CATH[3] and SCOP[4], revolutionary opportunities have arisen for data-driven advances in biological research. An essential component of quantitative biology is geometric analysis. Geometric measurements, algorithms and modeling offer a basis for molecular visualization, bridge the gap between experimental data from X-ray, NMR, and Cryo-electron microscopy, and theoretical models, and play a fundamental role in the analysis of biolmolecular structure, function, dynamics, and transport. Especially with the aid from increasingly powerful high performance computers, geometric analysis becomes more and more deeply involved in biological sciences. However, geometric invariants usually describe local features, such as distances, angles, curvatures, convexity, etc. As a consequence, geometric analysis tends to involve excessive irrelevant structure details and become computationally intractable, especially for macroproteins and protein complexes. A great promise comes from a newly founded area in big data analysis, known as topological data analysis (TDA). The essence of TDA is to employ concepts and algorithms from algebraic topology and computational topology to extract or identify intrinsic properties of the data. These intrinsic properties are topological invariants, which describe global features of the structure and are consistent under continuous deformation.

One of the most important tool in TAD is persistent homology (PH)[5–7]. Different from the traditional topological method, PH is able to embed a geometric measurement into topological invariants, thus provides a bridge between geometry and topology. Filtration is the key idea in PH. In a filtration process, a series of topological spaces are generated by a systematical variation of the filtration parameter. The Betti numbers for these simplicial complexes can be calculated. Their lifespans or persistent times are used as a geometric measurement[5, 6]. Various softwares, including JavaPlex [8], Perseus [9], Dipha [10], Dionysus [11], jHoles [12], GUDHI[13], etc[14], have been proposed, together with visualization methods, including persistent diagram[15], persistent barcode[16], and persistent landscape[17]. As a method deeply rooted in algebraic topology, persistent homology has demonstrated its great potential in data simplification and complexity reduction [5, 6]. It provides new opportunities for researchers from mathematics, computer sciences, computational biology, biomathematics, engineering, etc. Persistent homology has been used in a variety of fields, including shape recognition [18],network structure [19–21],image analysis [22–26], data analysis [27–31],chaotic dynamics verification [32, 33], computer

vision [24] and computational biology [34–36]. Recently, persistent homology has been used in analyzing fullerene molecules, proteins, DNAs and various other biomolecules [37–39]. A topological fingerprint is proposed to quantitatively analyze the biomolecular structures and functions. It is defined as consistent patterns of barcodes that have particular structure implications[37]. After that, a multiresolution and multidimensional persistent homology is introduced[40, 41] by incorporating a resolution/scale parameter into a rigidity density function. This model is able to focus the resolution on any scale of interest and is successfully used to study extremely large data from macroproteins or protein complexes. Further, based on the model, multiscale persistent functions are developed for biomolecular structure characterization[42]. Particularly, multiscale persistent entropy is successfully used in protein classification test[42]. Most recently, persistent homology is combined with machine learning tools to solve challenging classification and regression tasks in drug design[43–45], protein stability changes upon mutation[46, 47] and toxicity prediction[48]. For all these problems, topology based machine learning models have proved to be the state of the art. The great success of these models shows that specially-designed persistent homology models can retain critical chemical and biological information and provides a unique topological description of inter- and/or intra-molecular interactions of interest[47].

In this paper, we propose a new model of persistent similarity for a quantitative comparison between biomolecular structures. Biomolecular structure comparison is of great importance. It not only reveals evolutionary relationships, but also provides insights about biological functional properties. Even though various models of structure and sequence similarity have been proposed and widely used in biology[49], all these methods involve superposition or alignment at either global scale or common subregions. Computationally, algorithms for superposition or alignment require iterative searching and comparing, which can be very time-consuming, especially when many structures are considered simultaneously. Dramatically different from these models, our persistent similarity is based on topological characterization, thus free from structure or sequence alignment. More specifically, for each biomolecular structure, we can generate its topological representation, i.e., a series of barcodes. From these barcodes, we can define unique persistent Betti functions. These are simply one-dimensional continuous functions defined in exactly the same computational domain. In this way, evaluation of similarity between different structures is transferred into the comparison between one-dimensional functions. More importantly, since each biomolecular structure is associated with a unique barcode representation thus a

unique set of one-dimensional persistent Betti functions (PBFs), the comparison among various structures becomes much more efficient, as we only need to deal with similarity of one-dimensional PBFs. More importantly, we have introduced the multiscale persistent similarity, so that the structure comparison can be done at any scale of interest.

It has been noticed that "topological similarity" has been proposed for structure comparison[50–52] recently. In this model, structure similarity is directly measured from the persistent barcodes. The bottleneck distance and Wasserstein distance[53–56], which are widely used to measure distance between two sets of barcodes, are considered in this model. Our persistence similarity differs greatly from the model in several aspects. Firstly, we use the previously proposed PBFs [42]. These PBFs provide a unique representation of persistent barcode. Stated differently, there is a one to one relation between our PBFs and barcode representations. With these functions, the comparison between different barcodes becomes much more straightforward and efficient. Secondly, a multiscale persistent similarity is defined so that we can systematically compare the structure properties from various scales. Biomolecules, particularly macroproteins or protein complexes, are usually of multiple scales ranging from atom, residue, secondary structure, domain, protein monomer, etc. Different topological properties can be found at different scales. To be able to pinpoint to the right scale is of great importance for similarity comparison. In our model, multiscale rigidity functions are employed to represent the structures from various scales. And persistent similarities derived from them capture similarity information at different scales. Thirdly, we introduce a pseudo-barcode to deliver a more precise comparison in the special situation when one or several structure has no significant topological properties. For instance, if a structure has no $\beta_1$ barcodes while the others have, topological similarity between this structure and all the others will always be zero, no matter how many $\beta_1$ barcodes the others have. This ambiguity is avoided by the introduction of a pseudo-barcode into our persistent similarity. Fourthly, we introduce weight functions and kernel scales in our PBFs. These parameters give us more flexibility in defining the "significance" of the bars. It is found that for some biomolecular functions and properties, only some special barcodes matter while the others are irrelevant. And in this situation, our model can play an important role. It should be noticed that we deliberately avoid using the term of "topological similarity", because "topological similarity" is widely used in network modeling[57, 58]. The term "persistent similarity" captures the essence of the method and is

consistent with all previous notations including, persistent homology, persistent Betti number, persistent entropy[42], etc. Therefore, we believe it is a much better term to use.

The paper is organized as following. In Section 2, we introduce the basic theory of persistent homology firstly. After that, we discuss a special persistent Betti function and use it to define our persistent similarity. Further, we generalize our persistent similarity to multiscale persistent similarity through a multiscale persistent homology model. Section 3 is dedicated to basic results and discussions. Four different cases are studied, including two similar nucleotide kinases, a series of protein structures, configurations from molecular dynamics simulation and fullerene $C_{44}$ isomers. The paper ends with a conclusion.

## 2. Method

A suitable definition of similarities between biomolecules is of great importance for structure and functional analysis. A similarity matrix can be directly used for the hierarchical classification of biomolecular structures and helps to reveal their evolutional and functional relations. In this section, we introduce a new similarity measurement based on persistent homology analysis.

### 2.1. Persistent homology

To avoid the heavy mathematical notations and present essential ideas more straightforwardly, we will only focus on the simplical complex with direct geometric implications. Further, the homology is calculated in $Z_2$ field and only Vietoris-Rips complex is considered in simplicial complex generation. Interested readers are referred to more detailed description in papers[5–7].

Generally speaking, homology is a mathematical representation of topological invariants, including connected components, circles, rings, channels, cavities, voids, etc. Persistent homology gives a geometric measurement, i.e., a size, to these invariants. Figures 1 and 2 illustrate the essential concepts used in persistent homology.

**2.1.1. Simplicial complex.** Simplices are the building block for a simplicial complex. A set of $k+1$ affinely independent points $v_0, v_1, v_2, \ldots, v_k$

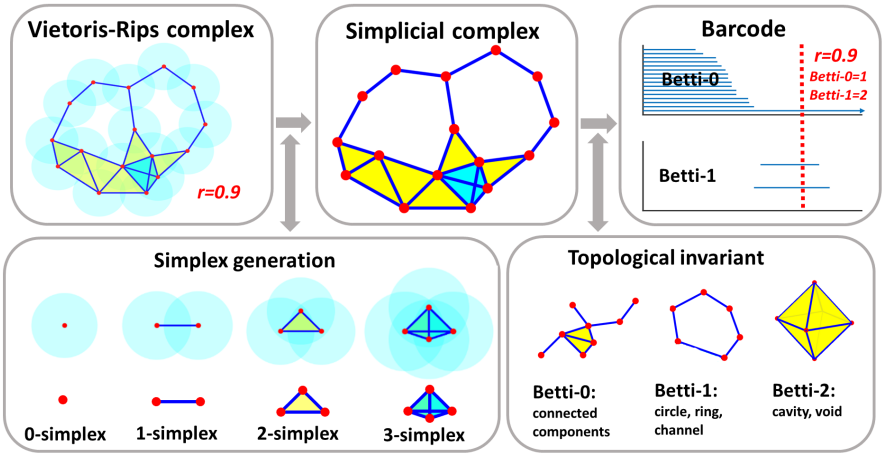


Figure 1: The illustration of basic concepts in persistent homology, including simplexes, simplicial complex, Vietoris-Rips complex, topological invariants. A filtration process is demonstrated in Figure 2.

can form a $k$-simplex $\sigma^k = \{v_0, v_1, v_2, \ldots, v_k\}$ as following,

$$\sigma^k = \left\{ \lambda_0 v_0 + \lambda_1 v_1 + \cdots + \lambda_k v_k \;\middle|\; \sum_{i=0}^{k} \lambda_i = 1; 0 \leq \lambda_i \leq 1, i = 0, 1, \ldots, k \right\}. \tag{1}$$

Geometrically, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex represents a tetrahedron, just as depicted in Figure 1. An $i$-dimensional face of $\sigma^k$ is the convex hull formed by $i+1$ vertices from $\sigma^k$ $(k > i)$. A simplicial complex $K$ is a finite set of simplices that satisfy two essential conditions, i.e., 1) any face of a simplex from $K$ is also in $K$; 2) the intersection of any two simplices in $K$ is either empty or shares one face. An oriented simplex is a simplex together with an orientation, i.e., ordering of its vertex set. We denote an oriented $k$-simplex as $[\sigma^k]$.

**2.1.2. Homology.** A $k$-chain $c$ is a linear combination of $k$-simplexes $c = \sum_i \alpha_i \sigma_i^k$ with $\{\alpha_i \in Z_2\}$. An Abelian group $C_k(K, \mathbb{Z}_2)$ is formed by the set of all $k$-chains from the simplicial complex $K$ together with addition operation (modulo-2). A boundary operator $\partial_k$ is defined as $\partial_k : C_k \to C_{k-1}$.

The boundary of an oriented $k$-simplex $[\sigma^k] = [v_0, v_1, v_2, \ldots, v_k]$ can be denoted as,

$$\partial_k[\sigma^k] = \sum_{i=0}^{k} [v_0, v_1, v_2, \ldots, \hat{v_i}, \ldots, v_k]. \tag{2}$$

Here $[v_0, v_1, v_2, \ldots, \hat{v_i}, \ldots, v_k]$ means a $(k-1)$ oriented simplex, which is generated by the elimination of vertex $v_i$. Further, one has $\partial_0 = 0$ and $\partial_{k-1}\partial_k = 0$. The $k$-th cycle group $Z_k$ and the $k$-th boundary group $B_k$ are the subgroups of $C_k$ and can be defined as,

$$Z_k = \text{Ker } \partial_k = \{c \in C_k \mid \partial_k c = 0\}, \tag{3}$$

$$B_k = \text{Im } \partial_{k+1} = \{c \in C_k \mid \exists d \in C_{k+1} : c = \partial_{k+1} d\}. \tag{4}$$

Their elements are called the $k$-th cycle and the $k$-th boundary, respectively. It can be noticed that $B_k \subseteq Z_k$, as the boundary of a boundary is always zero $\partial_{k-1}\partial_k = 0$. The $k$-th homology group $H_k$ is the quotient group generated by the $k$-th cycle group $Z_k$ and $k$-th boundary group $B_k$: $H_k = Z_k/B_k$. The rank of $k$-th homology group is called $k$-th Betti number and it can be calculated by

$$\beta_k = \text{rank } H_k = \text{rank } Z_k - \text{rank } B_k. \tag{5}$$

As indicated in Figure 1, the geometric meanings of Betti numbers in $\mathbb{R}^3$ are as following: $\beta_0$ represents the number of isolated components; $\beta_1$ is the number of one-dimensional loops, circles, or tunnels; $\beta_2$ describes the number of two-dimensional voids or holes. Together, the Betti number sequence $\{\beta_0, \beta_1, \beta_2\}$ describes the intrinsic topological properties of a system.

**2.1.3. Rips complex.** For a point set $X \in \mathbb{R}^N$, one defines a cover of closed balls centered at $x$ with radius $\epsilon$. A Rips simplex (or Vietoris-Rips simplex) $\sigma$ is generated if the largest distance between any of its vertices reaches $2\epsilon$. Figure 1 illustrates the generation of the Rips simplex.

**2.1.4. Filtration.** In the generation of Rips complex, a radius parameter $\epsilon$ is used. However, how to find the best suitable $\epsilon$ so that it can represent the underling space very well, has been a long standing problem. To solve this problem, the idea of filtration has been proposed[5]. As illustrated in Figure 2, instead of finding the best radius value, an ever-increasing $\epsilon$ value is used to generate a series of topological spaces. These topological spaces form

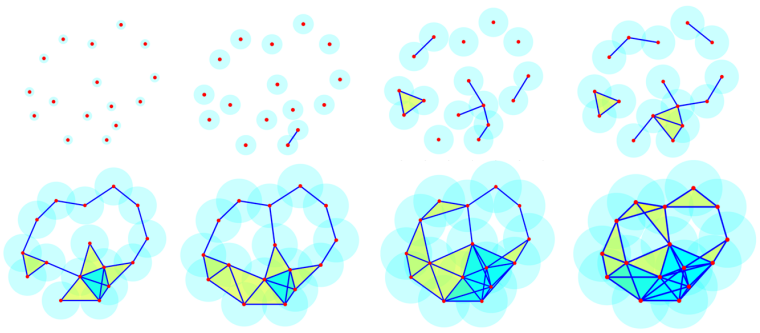

Figure 2: An illustration of a filtration process in persistent homology. During the filtration process, each point from the point cloud data is associated with a sphere, whose radius increases systematically. A series of topological spaces from different scales are generated. Based on them, a nested sequence of simplicial complexes can be obtained.

a nested sequence of complexes, and topological invariants can be calculated from them. Some topological invariants may last for a wide range of $\epsilon$ values, some may disappear very quickly when the $\epsilon$ value changes. In this way, these topological invariants have certain "lifespans", which can be viewed as an extra geometric measurement.

**2.1.5. Persistent homology.** The filtration can be described as a nested sequence of complexes,

$$K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K. \tag{6}$$

And the $p$-persistent $k$-th homology group at filtration time $i$ can be represented as

$$H_k^{i,p} = Z_k^i / \left( B_k^{i+p} \bigcap Z_k^i \right). \tag{7}$$

Essentially, persistence gives a geometric measurement of the topological invariant.

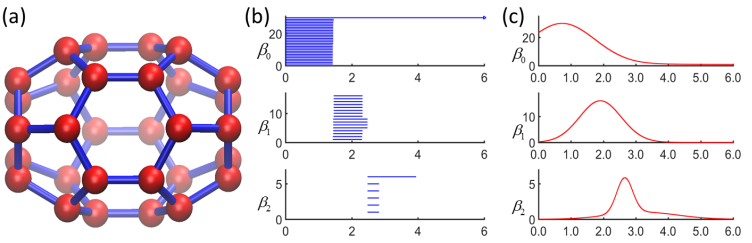


Figure 3: The illustration of fullerence $C_{30}$ molecule structure, barcodes and PBFs. (**a**) The cage structure of fullerence $C_{30}$. (**b**) The barcode representation. Each bar represents a homology generator and has its unique chemical or physical properties. For $\beta_0$ bars, they are related to chemical bonds. For $\beta_1$ bars, they represent pentagon and hexagon rings in the molecule. For $\beta_2$ bars, they describe cavity or cage structures. (**c**) Fullerence $C_{30}$ PBFs for $\beta_0, \beta_1$ and $\beta_2$.

### 2.2. Persistent similarity

The results from persistent homology analysis can be represented as following,

$$\{L_{k,j} = [a_{k,j}, b_{k,j}] | k = 0, 1, 2; j = 1, 2, 3, \ldots, N_k\}, \tag{8}$$

where parameter $k$ is the dimension of Betti number $\beta_k$, parameter $j$ indicates the $j$-th homology generator and $N_k$ is the number of $\beta_k$ generator. Further, we define the set of $k$-th dimension homology generators as,

$$L_k = \{L_{k,j}, j = 1, 2, 3, \ldots, N_k\}, \quad k = 0, 1, 2.$$

To visualize the persistent homology results, we use the barcode plot as illustrated in Figure 3 (**b**). For fullerence $C_{30}$ barcodes, the length of $\beta_0$ bars corresponds to the bond length. The number of $\beta_0$ bars is exactly the total number of atoms in the molecule. Further, the pentagon and hexagon ring structures are captured by $\beta_1$ bars. The cage structure of fullerene $C_{30}$ is described by the longest $\beta_2$ bar. More generally, for a chemical structure, each bar from its barcode has an unique structural and physical meaning[38, 40, 59]. These physical and chemical implications of the barcodes are very important for the understanding of biomolecular functions.

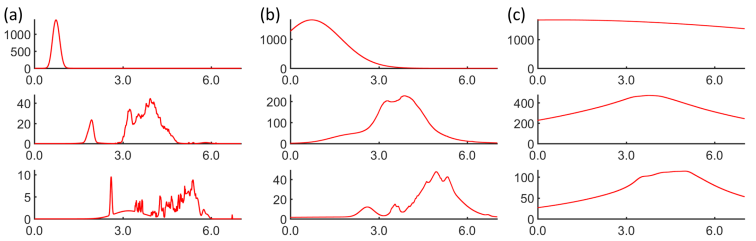


Figure 4: Illustration of protein 1AKY (all-atom-without-hydrogen model) PBFs generated with different resolution values. Here the resolution parameter $\omega$ are chosen as 0.1 Å**(a)**, 1.0 Å**(b)** and 10.0 Å**(c)**. It can be seen that the PBFs differ greatly in different resolutions.

**2.2.1. Persistent Betti function (PBF).** Based on the barcode, we can build up various models to further explore the biomolecular structure, flexibility, function and dynamics[40, 42]. To uniquely represent the barcode, we define a persistent Betti function as,

$$f(x;L_k)=\sum_j \upsilon_{k,j} e^{-\left(\frac{x-\frac{b_{k,j}+a_{k,j}}{2}}{\omega(b_{k,j}-a_{k,j})}\right)^{\kappa}}, \quad \kappa>0; k=0,1,2; x\in[0,r_f] \tag{9}$$

where $\upsilon_{k,j}$ are weight function for $j$-th barcode of $\beta_k$. Parameter $\omega$ is a resolution parameter and $r_f$ is filtration ending time.

It should be noticed that the sequence of the bars in the barcodes is not unique. For visualization, we usually arrange the barcodes by their birth times. For biomolecules, each bar or each type of bars has its unique structural, physical and chemical implications. So we can assign or define a weight value $\upsilon_{k,j}$ for each or each type of bar if needed. In this paper, unless stated otherwise, the weight function and resolution parameter are all chosen as 1, i.e., $\omega=1$ and $\upsilon_{k,j}=1$ for all $k$ and $j$.

There are different ways to represent barcodes as function, such as persistent Betti number, persistent landscapes [60], etc. The PBF provides a unique transformation of persistent barcodes into 1D continuous functions. There is a strict one-to-one correlation between barcodes and PBFs. In this way, any complicated biomolecular structures can be uniquely represented by three 1D PBFs, thus dramatically reduce the dimensionality and complexity in structure comparison. The resolution parameter gives extra degree

of freedom in similarity evaluation. Figure 5 illustrates the influence of resolution value to PBFs. Three different resolution values are considered in protein 1AKY. It can be seen that with the enlargement of $\omega$ values, the local "fluctuations" in PBFs are gradually smoothed out. In general, when resolution is low, i.e., a large $\omega$ value, long bars will dominate the behavior of the PBF. In contrast, when resolution is high, i.e., a small $\omega$ value, the effect of short bars, which are generally considered as noise, will be reflected in the PBF. To emphasize again, we only consider $\omega = 1$ in the current paper.

**2.2.2. Persistent similarity.** For two different biomolecular structures denoted as $F_1$ and $F_2$, if their PBFs are $\{f(x; L_k^1), k = 0, 1, 2\}$ and $\{f(x; L_k^2), k = 0, 1, 2\}$. Regions below these functions are denoted as $S_k^1 = \{(x, y)|0 \leq y(x) \leq f(x; L_k^1); 0 \leq x \leq r_f\}$ and $S_k^2 = \{(x, y)|0 \leq y(x) \leq f(x; L_k^2); 0 \leq x \leq r_f\}$, their persistent similarity can be defined as,

$$P_k(F_1, F_2) = \frac{\text{Area}(S_k^1 \bigcap S_k^2)}{\text{Area}(S_k^1 \bigcup S_k^2)}, \quad k = 0, 1, 2. \tag{10}$$

A suitable filtration ending time is not unique. Usually, it is chosen as the smallest value, after which no significant topological properties will appear. The above definition of persistent similarity is equivalent to,

$$P_k(F_1, F_2) = \frac{\int_0^{r_f} \min\{f(x; L_k^1), f(x; L_k^2)\}dx}{\int_0^{r_f} \max\{f(x; L_k^1), f(x; L_k^2)\}dx}, \quad k = 0, 1, 2. \tag{11}$$

The higher dimensional topological invariants such as $\beta_1$ and $\beta_2$ can be zero in small molecules. In this situation, if it is compared with other molecules with $\beta_1$ barcodes, the similarity is always zero, no matter how many $\beta_1$ barcodes in the other structures or how long are these $\beta_1$ barcodes. This ambiguity can bring troubles in structure comparison. To overcome this problem, we introduce a pseudo-bar into the PBFs and definite a new PBF as,

$$f^{\text{pseudo}}(x; L_k) = \upsilon_{k,0} + \sum_j \upsilon_{k,j} e^{-\left(\frac{x - \frac{b_{k,j} + a_{k,j}}{2}}{\omega(b_{k,j} - a_{k,j})}\right)^{\kappa}}, \quad \kappa > 0; k = 0, 1, 2. \tag{12}$$

Essentially, a small weight value $\upsilon_{k,0}$ is introduced to avoid the situation when PBF is a zero function. In this way, we can significantly reduce the ambiguity.

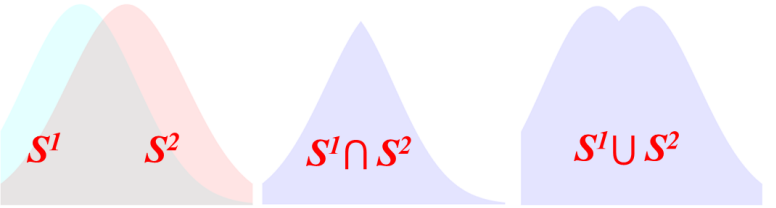


Figure 5: Illustration of the definition of persistent similarity. Here $S^1$ and $S^2$ represent the region under two different persistent Betti functions, respectively. The persistent similarity $P$ is defined to be the quotient between the intersect area and the union area, i.e., $P = \frac{\text{Area}(S^1 \bigcap S^2)}{\text{Area}(S^1 \bigcup S^2)}$.

### 2.3. Multiscale persistent similarity

The structure of biomolecules is hierarchical and built up with components from various scales. To capture multiscale structure properties, we have proposed a multiresolution/multiscale persistent homology[40, 61]. The key of our model is a multiscale density function, which is derived from the previous flexibility and rigidity index (FRI) method[62–67]. A resolution parameter is incorporated into the density function. And by turning its value, we can generate a series of density functions from different scales. More details will be discussed below.

**2.3.1. Multiscale rigidity function.** For a data set with $N$ entries, which can be physical elements like atoms, residues, domains, monomer proteins or data components like points, pixels and voxels, if one assumes their generalized coordinates are $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$, a multiscale rigidity function of the data can be expressed as,

$$\mu(\mathbf{r}, \eta) = \sum_j^N w_j \Phi(\| \mathbf{r} - \mathbf{r}_j \|; \eta) \tag{13}$$

where $w_j$ is the weight parameter, which is usually chosen as the atomic number. For example, its value is 6 for carbon atom and 8 for oxygen atom. The parameter $\eta$ is the resolution or scale parameter. The function $\Phi(\| \mathbf{r} - \mathbf{r}_j \|; \eta)$ is a kernel function. Commonly used kernel functions are generalized exponential functions,

$$\Phi(\| \mathbf{r} - \mathbf{r}_j \|; \eta, \kappa_0) = e^{-(\|\mathbf{r}-\mathbf{r}_j\|/\eta)^{\kappa_0}}, \quad \kappa_0 > 0. \tag{14}$$

It can be noticed that the larger the $\eta$ value, the lower the resolution is. A multiscale geometric model can be naturally derived from our multiscale rigidity functions. An example can be found in Figure 11.

**2.3.2. Multiscale persistent homology.** Based on the multiscale rigidity function, we have proposed a multiscale persistent homology[40, 61]. In this model, we linearly rescale all the rigidity function values to the region $[0, 1]$ using formula

$$\mu^s(\mathbf{r}, \eta) = 1.0 - \frac{\mu(\mathbf{r}, \eta)}{\mu_{\max}(\eta)}. \tag{15}$$

Here $\mu(\mathbf{r}, \eta)$ and $\mu^s(\mathbf{r}, \eta)$ are the original and normalized rigidity function, $\mu_{\max}(\eta)$ is the maximum value of the original rigidity function.

We can perform the persistent homology analysis on these normalized rigidity functions. The filtration parameter is chosen as the isovalue (level-set value or contour value). More specifically, a molecular surface can be generated from an isovalue. With the continuous variation of its values, a series of molecular surfaces will be generated. Based on these surfaces, a nested sequence of Morse complexes are generated. In this way, persistent homology analysis can be employed[15, 68, 69].

**2.3.3. Multiscale persistent similarity.** A series of barcodes from various scales are generated in the multiscale persistent homology and can be represented as follows,

$$\{L_{k,j}(\eta) = [a_{k,j}(\eta), b_{k,j}(\eta)] | k = 0, 1, 2; j = 1, 2, 3, \ldots, N_k(\eta)\}. \tag{16}$$

Similar to the previous definition, parameter $k$ is the dimension of Betti number $\beta_k$, parameter $j$ indicates $j$-th barcode and $N_k$ is the number of $\beta_k$ barcodes. And the sets of barcodes in the $k$-th dimension is represented as,

$$L_k(\eta) = \{L_{k,j}(\eta), j = 1, 2, 3, \ldots, N_k(\eta)\}, \quad k = 0, 1, 2$$

Further the multiscale persistent Betti function is represented as,

$$f(x; L_k(\eta)) = \sum_j \upsilon_{k,j}(\eta) e^{-\left(\frac{x - \frac{b_{k,j}(\eta) + a_{k,j}(\eta)}{2}}{\omega(\eta)(b_{k,j}(\eta) - a_{k,j}(\eta))}\right)^{\kappa}}, \quad \kappa > 0, k = 0, 1, 2. \tag{17}$$

Again $\upsilon_{k,j}(\eta)$ is the weight function for $j$-th barcode of $\beta_k$, and parameter $\omega(\eta)$ is the resolution or scale parameter. The multiscale persistent similarity

between structures $F_1$ and $F_2$ can be defined as

$$P_k(F_1, F_2, \eta) = \frac{\int \min\{f(x; L_k^1(\eta)), f(x; L_k^2(\eta))\}}{\int \max\{f(x; L_k^1(\eta)), f(x; L_k^2(\eta))\}}, \quad k = 0, 1, 2. \tag{18}$$

The multiscale persistent similarity enable us to compare the structure properties from various scales.

## 3. Results and discussions

In this section, we validate our persistent similarity method using four different cases. In the first case, we consider two nucleotide kinases 1AKY and 1GKY with similar structures. We calculate their persistent similarities for both all-atom model and $C_\alpha$ coarse-grained model. We find that the calculated persistent similarities are around 0.8, indicating structure similarities between two structures. In the second case, a series of configurations of protein 2KIX are considered. These structures are very similar to each other with very small variations. The persistent similarities for both $\beta_0$ and $\beta_1$ are very large, indicating a strong structural consistence between all these frames. The third case is devoted to the validation of multiscale persistent similarity. We consider four configurations obtained from the steered molecular dynamic simulation of protein Titin. Persistent similarities are evaluated from two different scales. For local scale models, high persistent values are obtained, meaning that local structures are very consistent during the simulation. For global scale models, a dramatic reduction of persistent similarity values are observed, suggesting that the global properties of these structures vary greatly during the simulation. Our results are highly consistent with structure properties in the unfolding process. The last case is employed for the study of fullerene $C_{44}$ isomers. Previously, we have found that fullerene total curvature energies are largely determined by the longest $\beta_2$ bar[38]. To further explore properties of the total curvature energy, we use special weight parameters and define our persistent similarity on this particular bar. We find that our persistent similarity delivers one of the best results for isomer total curvature energy prediction. To void confusion, in the first three cases, the weight function and resolution parameter in the PBFs are all chosen as 1, i.e., $\omega = 1$ and $\upsilon_{k,j} = 1$ for all $k$ and $j$.

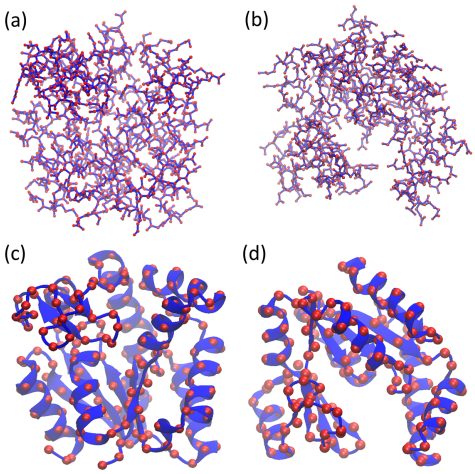


Figure 6: The all-atom-without-hydrogen model and coarse-grained model of Kinase proteins 1AKY and 1GKY. Each red point represents an atom. (**a**) and (**b**) All-atom-without-hydrogen models for 1AKY (left) and 1GKY (right). (**c**) and (**d**) Coarse-grained models of 1AKY (left) and 1GKY (right).

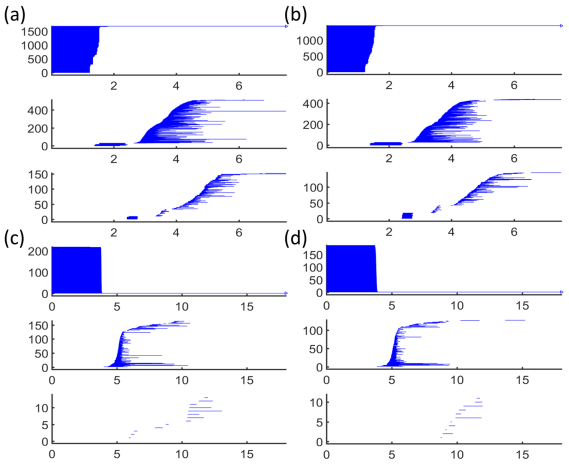


Figure 7: Persistent barcodes for Kinase 1AKY and 1GKY models (as in Figure 6). (**a**) and (**b**) Barcodes for all-atom-without-hydrogen model of 1AKY and 1GKY, respectively. (**c**) and (**d**) Barcodes for coarse-grained model of 1AKY and 1GKY, respectively.

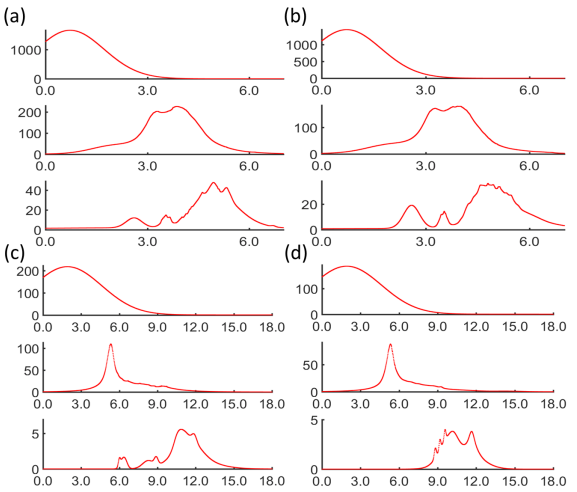


Figure 8: PBFs for Kinase 1AKY and 1GKY models in Figure 6. (**a**) and (**b**) PBFs for all-atom-without-hydrogen models of 1AKY and 1GKY, respectively. (**c**) and (**d**) PBFs for coarse-grained models of 1AKY and 1GKY, respectively. For the two all-atom models, their persistence similarities of $\beta_0$, $\beta_1$ and $\beta_2$ are 0.860, 0.819 and 0.795, respectively. For the two coarse-grained models, their persistence similarities of $\beta_0$, $\beta_1$ and $\beta_2$ are 0.857, 0.770 and 0.485, respectively.

### 3.1. Case 1: Two similar nucleotide kinases

In the first case, we consider two nucleotide kinases (1AKY and 1GKY) used in structural alignment [70]. The all-atom-without-hydrogen model and $C_\alpha$ coarse-grained model, as illustrated in Figure 6, are used for persistent similarity evaluation. These two proteins share some similar regions, like the $\alpha$-helixes on the left boundary and the $\beta$-sheets in the middle. To quantitatively measure their structural similarity, we employ persistent homology analysis and generate barcodes for both structures in two representations. Figure 7 (**a**) and (**b**) are barcodes of all-atom-without-hydrogen models for 1AKY and 1GKY, respectively. As stated above, the length of short $\beta_0$ bars represents chemical bond length. And the number of $\beta_0$ bars is the number of atoms in the system. In this way, from $\beta_0$ bars, chemical components of the structure can be understood. More chemical implications can be learned

from $\beta_1$ bars. Previously, we have found that, short $\beta_1$ bars located in local region around 2.0 Å represent pentagon and hexagon rings in aromatic residues [40, 59]. Particularly, the hexagon rings can further manifest themselves in local $\beta_2$ bars. The global structure properties captured by $\beta_1$ bars appear much later in the filtration. For all-atom-without-hydrogen model, there is clear separation of local and global type of $\beta_1$ bars in both $\beta_1$ and $\beta_2$ barcodes. Figure 7 (**c**) and (**d**) are barcodes for $C_\alpha$ coarse-grained modeling of 1AKY and 1GKY, respectively. One can see that the length of all short $\beta_0$ bars are around 3.8 Å, i.e., the distance between the two adjacent $C_\alpha$ atoms. Moreover, all the three types of barcodes are dramatically reduced. Particularly the $\beta_2$ barcodes.

With these barcodes, we can generate the persistent Betti functions. Figure 8 illustrates PBFs for the two proteins in the same sequence as Figures 6 and 7. For all-atom-without-hydrogen models, the persistent similarities for $\beta_0$, $\beta_1$ and $\beta_2$ are 0.860, 0.819 and 0.795, respectively. For $C_\alpha$ coarse-grained model, the persistence similarities for $\beta_0$, $\beta_1$ and $\beta_2$ are 0.857, 0.770 and 0.485, respectively. It can be seen that for $\beta_0$ and $\beta_1$, the persistent similarities for the two models are very close. This indicates the robustness of our persistent similarity. Further, persistent similarities for $\beta_2$ differ a lot. This is due to the fact that there are very few $\beta_2$ bars in the coarse-grained representation for both models. Very limited structural information is captured in $\beta_2$ bars, so that they do not deliver a good representation of the whole structure. In later examples, we only consider the $\beta_0$ and $\beta_1$ persistent similarity for coarse-grained models.

### 3.2. Case 2: NMR configurations

In the second case, we consider M-crystallin structures in calcium free form (PDB ID: 2KIX). There are totally twenty configurations in the PDB data and all of them are very similar to each other with only small variations. The coarse-grained representation is considered and we illustrate ten configurations (out of twenty) in Figure 9. Further, we calculate the persistent similarities for $\beta_0$ and $\beta_1$, and demonstrate the results for all twenty configurations in Figure 10. It can be seen that all these structures share a high persistent similarity. For $\beta_0$, the persistent similarities are all around 1.000. For $\beta_1$, the persistent similarities vary from 0.8 to 1.0. And the smallest persistent similarity is about 0.802.

From comparison of persistent similarity values of Case 1 and Case 2, one can see that the persistent similarity gives a very reasonable evaluation

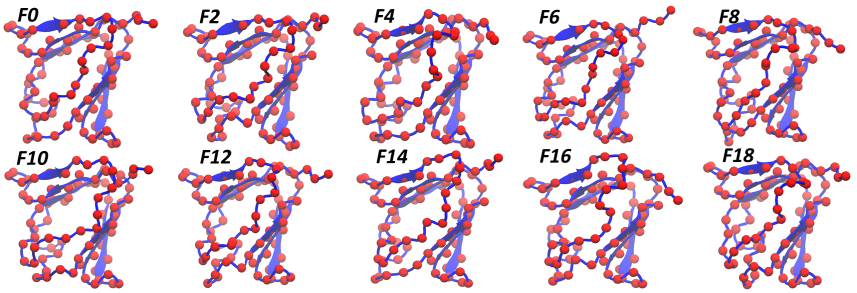


Figure 9: The coarse-grained models of protein 2KIX structures. There are totally twenty configurations (denoted as F1 to F20). We take ten different configurations among them. It can be seen that they all have very similar structures.

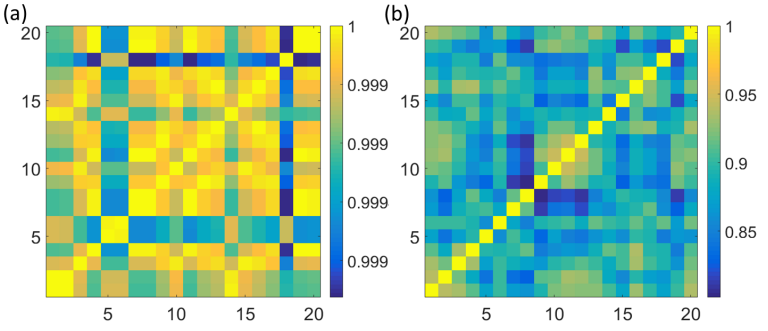


Figure 10: The persistent similarities between twenty different configurations of protein 2KIX. (**a**) Betti-0 persistent similarities. (**b**) Betti-1 persistent similarities. It can be seen that the values for Betti-0 persistent similarities are all close to 1.0.

of the structure similarity. For structures with same atoms and same chemical bonds, the $\beta_0$ persistent homology values are all around 0.999. This is consistent with chemical implications of $\beta_0$ bars. Further, the $\beta_1$ persistent similarities between 2K1X structures are all larger the ones between 1AKY and 1GKY. This is reasonable, as 2K1X NMR configurations are highly consistent with only small variations.

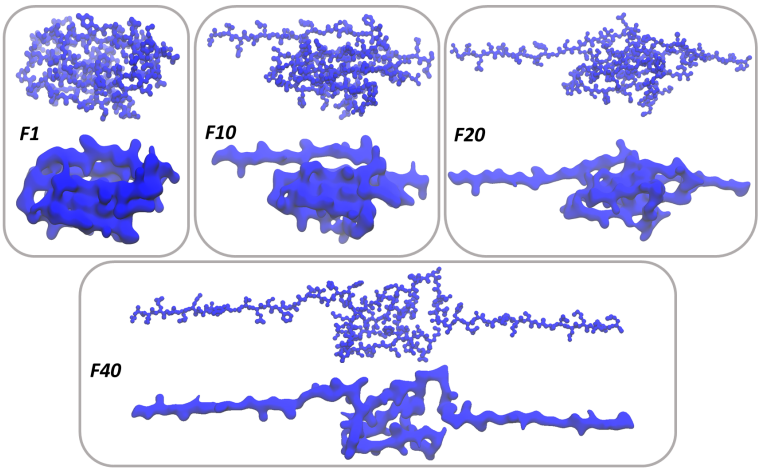


Figure 11: The Titin protein configurations derived from steered dynamic simulations. There are totally 89 frames from simulation trajectory. We choose only frames 1, 10, 20 and 40, denoted as $F1$, $F10$, $F20$ and $F40$, respectively. Two different scale parameters, i.e., $\eta = 0.6$ Å and 2.0Å, are considered in the upper and lower subfigures, respectively.

### 3.3. Case 3: Steered dynamic simulation

The third case is devoted for multiscale persistent similarity. We consider Titin I91 configurations extracted from a steered molecular dynamic simulation (`http://www.ks.uiuc.edu/Training/Tutorials/science/timeline/timeline-tutorial-files/`). Four out of eighty-seven frames are used for multiscale similarity analysis, including frame 1, 10, 20 and 40 (denoted as $F1$ to $F40$). The unfolding process goes from $F1$ to $F40$. We choose the generalized exponential kernel in Eq.(14) with parameter $\kappa_0 = 2$ and two different resolution values $\eta = 0.6$ Å and 2.0 Å. The density data is generated with grid spacing value 0.3 Å. Figure 11 illustrates the multiscale rigidity functions for the four configurations. For each subfigure, the structure on the upper part is generated with resolution value $\eta = 0.6$ Å and low part is generated with $\eta = 2.0$ Å. Further, we depict the barcodes for different configurations in Figure 12. Again, the upper subfigures are for resolution value $\eta = 0.6$ Å and low subfigures are for $\eta = 2.0$ Å. It can be seen that, barcodes obtained from $\eta = 0.6$Å density data are much more consistent. Their barcode lengths, total numbers, and general patterns show a great similarity. In

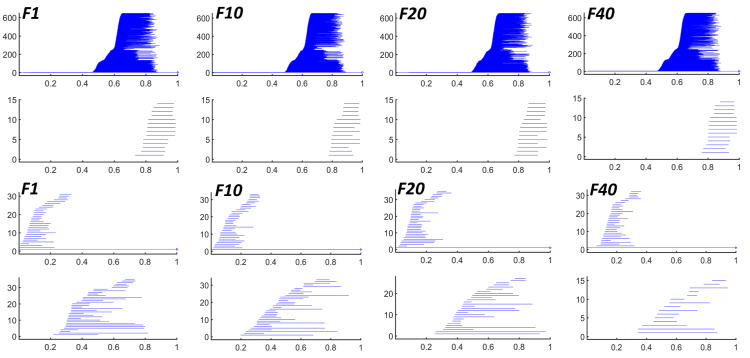


Figure 12: Persistent barcodes of the four Titin configurations at two different scales. The upper and lower four subfigures are results for models with $\eta = 0.6$ Å and 2.0Å, respectively. It can be seen that for local scale models (upper subfigures), the barcodes are very consistent. While in global scale models (lower subfigures), barcodes vary dramatically.

contrast, when the resolution value is $\eta = 2.0$ Å, we begin to observe large variations, particularly in $\beta_1$. Actually, the total numbers of $\beta_1$ bars keep decreasing during the unfolding process. To have a more quantitative comparison, we calculate the multiscale persistent similarities between these configurations. For any two Frames $F_a$ and $F_b$, we denote their persistent similarities as $P(F_a, F_b, \eta) = (P_0(F_a, F_b, \eta), P_1(F_a, F_b, \eta))$. When $\eta = 0.6$ Å, we have $P(F1, F10, 0.6) = (0.940, 0.915)$, $P(F1, F20, 0.6) = (0.915, 0.924)$ and $P(F1, F40, 0.6) = (0.967, 0.975)$. When $\eta = 2.0$ Å, we have $P(F1, F10, 2.0) = (0.832, 0.822)$, $P(F1, F20, 2.0) = (0.828, 0.695)$ and $P(F1, F40, 2.0) = (0.573, 0.413)$. It can be seen that at high resolutions, persistent similarities between configurations are relatively large. This is due to the reason that only local structure properties, including atomic numbers, pentagon and hexagon rings, are captured in barcodes[40, 61]. At lower resolutions, persistent similarities are relatively small. This means the global properties between these configurations differ greatly. Further, the similarity value decreases systematically when comparison goes from $F10$ to $F20$, then to $F40$, indicating a gradual derivation from the original structure. We only demonstrate the results for two different scales. In general, we can systematically change the resolution value and use the multiscale persistent similarity to compare the structure properties at different scales.

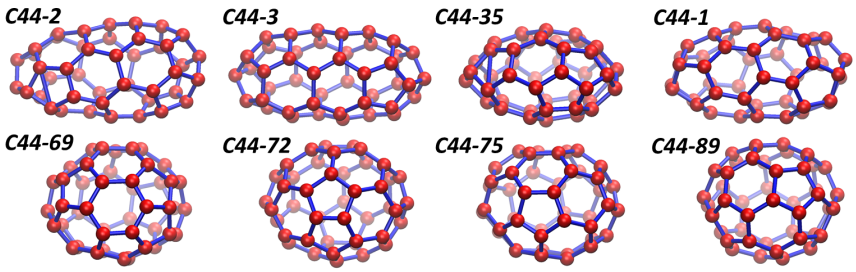


Figure 13: The illustration of eight different fullerene $C_{44}$ isomer structures. Fullerene $C_{44}$ has totally 89 isomers. These isomers have different total curvature energies. we have demonstrated four isomer structures with the largest total curvature energies in the upper figures. From the large energies to small ones, their indexes are 2, 3, 35 and 1, respectively. we have illustrated four isomer structures with the smallest total curvature energies in the lower figures. Again from the large energies to small ones, their indexes are 69, 72, 75 and 89, respectively.

It is worth mentioning that even though we have relatively large persistent similarities for higher resolution cases, their persistent similarity is not equal to 1.0. Theoretically, we should have same $\beta_0$ and $\beta_1$ persistent similarity values. However, due to computational constraints, we only use a grid spacing of 0.3 Å. In this way, the highest rigidity values for frame $F1$, $F10$ ,$F20$ and $F40$ are 15.01, 15.53, 15.03 and 15.30, respectively. The variations induce inconsistence in the normalized rigidity function and further into the barcode results.

### 3.4. Case 4: Fullerene $C_{44}$ isomers

In the last case, we consider fullerene $C_{44}$ isomers and their total curvature energies. The fullerene $C_{44}$ isomers and energy data can be downloaded from webpage (`http://www.nanotube.msu.edu/fullerene/fullerene.php?C=44`). There are totally 89 isomers. Eight special isomer structures are chosen and illustrated in Figure 13. Among them, four isomers on the upper figures are of the largest total curvature energies. Their indexes are 2, 3, 35 and 1, respectively. Four isomer structures with the smallest total curvature energies are depicted in the lower figures. Their indexes are 69, 72, 75 and 89,

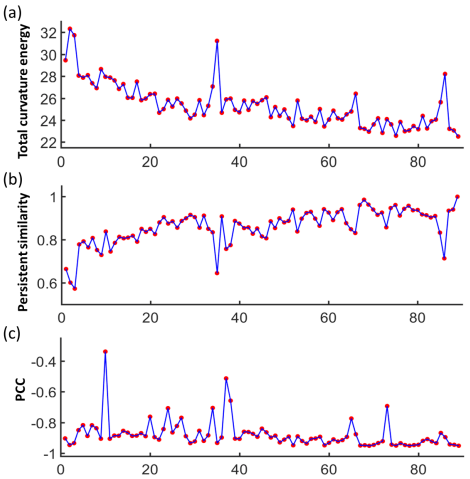


Figure 14: The comparison between total curvature energies and persistent similarities of 89 fullerene $C_{44}$ isomers. (**a**) is the total curvature energies of 89 fullerene $C_{44}$ isomers. (**b**) is the persistent similarities between all isomers with the isomer 89. (**c**) The Pearson correlation coefficient (PCC) of persistent similarities and curvature energy differences. For each isomer, we calculate its persistent similarities with all isomers including itself, then we compare these similarity values with the corresponding curvature energy differences (absolute value) to get PCC values.

respectively. It is found that the isomer total curvature energy is highly related to the regularity of isomer cage structure[38]. The longest $\beta_2$ barcode, representing the cage size, has been found to be linearly related to these energies[38]. In this case, we further explore the relation of structure similarities and total curvature energy differences. In our PBFs, only the weight for the longest $\beta_2$ barcode is chosen as 1.0, and all the other weights are defined as 0.0. The scale parameter is still chosen as 1, i.e., $w = 1$. And we only consider the $\beta_2$ PBFs. We calculate the similarity between isomer $C_{44}$-89 and all other $C_{44}$ isomers. Then we compare the persistent similarity with the total curvature energy. The results are demonstrated in Figures 14 (**a**) and (**b**). It can be seen that there is an inverse relation between them. Actually, the absolute value of Pearson correlation coefficient (PCC) between them is 0.952. This result is better that distance filtration [38] and density filtration results [39], and is as good as correlation matrix results[38].

Further, we change the reference isomer from $C_{44}$-89 to other isomers and recalculate PCCs between new persistent similarity values and new curvature energy differences. To avoid confusion, energy differences are taken as absolute difference values. The new PCCs are illustrated in Figure 14 (**c**). It can be seen that most PCCs have absolute values larger than 0.80. More interesting, higher absolute PCCs are from reference isomers with more extreme curvature energies (either very large or very small). If the reference isomer is of intermediate curvature energy, there will be a small absolute PCC value. This is due to the reason that our persistent similarity measures only the "absolute" different. To measure the intrinsic differences more accurately, one should always use extreme cases as references.

## 4. Conclusion

In this paper, we introduce a persistent similarity model for structure comparison. Based on persistent homology, our persistent similarity can deliver a quantitative comparison of the intrinsic topological properties between two structures. In our model, a persistent Betti function (PBF) is used to represent the barcodes into a series of one-dimensional functions. The similarity is defined as the ratio of intersection areas and union areas between any two 1D PBFs. In order to avoid the ambiguity of comparing structures with no significant topological properties, a pseudo-barcode is introduced. Further, to facilitate the comparison of structure properties at different scales, multiscale persistent similarity is considered. Finally, our persistent similarity model is validated with several test examples. It is found that our persistent similarity can be used to describe the intrinsic similarities and differences between biomolecular structures very well.

The proposed persistent similarity has several unique properties. Firstly, with the representation of structures in 1D PBFs, the comparison between various structures can be done very efficiently. In our persistent similarity, any complicated biomolecular structure is reduced to several simple 1D PBFs for comparison. Secondly, the multiscale persistent similarity enables an objective-oriented comparison. In our model, a multiscale biomolecular representation is considered and the associated persistent similarity can be used to compare structures at any particular scale of interest. Thirdly, a pseudo-barcode is introduced to deliver a more precise comparison when structures have no significant topological properties. In future, we will explore the application of persistent similarity in protein structure classification [71] and combine it with machine learning methods.

## Acknowledgments

This work was supported in part by Nanyang Technological University Startup Grant M4081842.110 and Singapore Ministry of Education Academic Research fund Tier 1 M401110000.

## References

[1] H. M. Berman, J. Westbrook, Z. K. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The protein data bank*, Nucleic Acids Research **28** (2000), no. 1, 35–242.

[2] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, *Genbank*, Nucleic Acids Research **41** (2012), D1:D36–D42.

[3] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, *CATH–a hierarchic classification of protein domain structures*, Structure **5** (1997), no. 8, 1093–1109.

[4] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *SCOP: a structural classification of proteins database for the investigation of sequences and structures*, Journal of Molecular Biology **247** (1995), no. 4, 536–540.

[5] H. Edelsbrunner, D. Letscher, and A. Zomorodian, *Topological persistence and simplification*, Discrete Comput. Geom. **28** (2002), 511–533.

[6] A. Zomorodian and G. Carlsson, *Computing persistent homology*, Discrete Comput. Geom. **33** (2005), 249–274.

[7] A. Zomorodian and G. Carlsson, *Localized homology*, Computational Geometry — Theory and Applications, **41** (2008), no. 3, 126–148.

[8] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams, *Javaplex: A research software package for persistent (co)homology*, software available at `http://code.google.com/p/javaplex`, 2011.

[9] Vidit Nanda, *Perseus: the persistent homology software*, software available at `http://www.sas.upenn.edu/~vnanda/perseus`.

[10] U. Bauer, M. Kerber, and J. Reininghaus, *Distributed computation of persistent homology*, Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX), 2014.

[11] Dionysus: the persistent homology software. Software available at `http://www.mrzv.org/software/dionysus`.

[12] J. Binchi, E. Merelli, M. Rucco, G. Petri, and F. Vaccarino, *jholes: A tool for understanding biological complex networks via clique weight rank persistent homology*, Electronic Notes in Theoretical Computer Science **306** (2014), 5–18.

[13] C. Maria, *Filtered complexes*, in: GUDHI User and Reference Manual. GUDHI Editorial Board, 2015.

[14] B. T. Fasy, J. Kim, F. Lecci, and C. Maria, *Introduction to the r package tda*, `arXiv:1411.1830`, (2014).

[15] K. Mischaikow and V. Nanda, *Morse theory for filtrations and efficient computation of persistent homology*, Discrete and Computational Geometry **50** (2013), no. 2, 330–353.

[16] R. Ghrist, *Barcodes: the persistent topology of data*, Bulletin of the American Mathematical Society **45** (2008), no. 1, 61–75.

[17] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, Journal of Machine Learning Research **16** (2015), no. 1, 77–102.

[18] B. Di Fabio and C. Landi, *A Mayer-Vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions*, Foundations of Computational Mathematics **11** (2011), 499–527.

[19] V. D. Silva and R Ghrist, *Blind swarms for coverage in 2-d*, in: Proceedings of Robotics: Science and Systems, page 01, 2005.

[20] H Lee, H. Kang, M. K. Chung, B. Kim, and D. S. Lee, *Persistent brain network homology from the perspective of dendrogram*, Medical Imaging, IEEE Transactions on **31** (2012), no. 12, 2267–2277.

[21] D. Horak, S Maletic, and M. Rajkovic, *Persistent homology of complex networks*, Journal of Statistical Mechanics: Theory and Experiment **2009** (2009), no. 03, P03034.

[22] G. Carlsson, T. Ishkhanov, V. Silva, and A. Zomorodian, *On the local behavior of spaces of natural images*, International Journal of Computer Vision **76** (2008), no. 1, 1–12.

[23] D. Pachauri, C. Hinrichs, M.K. Chung, S.C. Johnson, and V. Singh, *Topology-based kernels with application to inference problems in*

*alzheimer's disease*, IEEE Transactions on Medical Imaging **30** (2011), no. 10, 1760–1770.

[24] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, and D. L. Ringach, *Topological analysis of population activity in visual cortex*, Journal of Vision **8** (2008), no. 8.

[25] P. Bendich, H. Edelsbrunner, and M. Kerber, *Computing robustness and persistence for images*, IEEE Transactions on Visualization and Computer Graphics **16** (2010), 1251–1260.

[26] P. Frosini and C. Landi, *Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval*, Pattern Recognition Letters **34** (2013), no. 8, 863–872.

[27] G. Carlsson, *Topology and data*, Am. Math. Soc. **46** (2009), no. 2, 255–308.

[28] P. Niyogi, S. Smale, and S. Weinberger, *A topological view of unsupervised learning from noisy data*, SIAM Journal on Computing **40** (2011), 646–663.

[29] B. Wang, B. Summa, V. Pascucci, and M. Vejdemo-Johansson, *Branching and circular features in high dimensional data*, IEEE Transactions on Visualization and Computer Graphics **17** (2011), 1902–1911.

[30] B. Rieck, H. Mara, and H. Leitte, *Multivariate data analysis using persistence-based filtering and topological signatures*, IEEE Transactions on Visualization and Computer Graphics **18** (2012), 2382–2391.

[31] X. Liu, Z. Xie, and D. Y. Yi, *A fast algorithm for constructing topological structure in large data*, Homology, Homotopy and Applications **14** (2012), 221–238.

[32] K. Mischaikow, M Mrozek, J. Reiss, and A. Szymczak, *Construction of symbolic dynamics from experimental time series*, Physical Review Letters **82** (1999), 1144–1147.

[33] T. Kaczynski, K. Mischaikow, and M. Mrozek, Computational Homology, Springer-Verlag, 2004.

[34] P. M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. J. Guibas, and V. S. Pande, *Persistent voids a new structural metric for membrane fusion*, Bioinformatics **23** (2007), 1753–1759.

[35] Y. Yao, J. Sun, X. H. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, and G. Carlsson, *Topological methods for exploring low-density states in biomolecular folding pathways*, The Journal of Chemical Physics **130** (2009), 144115.

[36] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda, *Topological measurement of protein compressibility via persistence diagrams*, preprint (2013).

[37] K. L. Xia and G. W. Wei, *Persistent homology analysis of protein structure, flexibility and folding*, International Journal for Numerical Methods in Biomedical Engineerings **30** (2014), 814–844.

[38] K. L. Xia, X. Feng, Y. Y. Tong, and G. W. Wei, *Persistent homology for the quantitative prediction of fullerene stability*, Journal of Computational Chemsitry **36** (2015), 408–422.

[39] B. Wang and G. W. Wei, *Object-oriented persistent homology*, Journal of Computational Physics **305** (2016), 276–299.

[40] K. L. Xia and G. W. Wei, *Multidimensional persistence in biomolecular data*, Journal Computational Chemistry **36** (2015), 1502–1520.

[41] K. L. Xia and G. W. Wei, *Persistent topology for cryo-EM data analysis*, International Journal for Numerical Methods in Biomedical Engineering **31** (2015), e02719.

[42] K. L. Xia, Z. M. Li, and L. Mu, *Multiscale persistent functions for biomolecular structure characterization*, Bulletin of Mathematical Biology, revised **80** (2017), no. 1, 1–31.

[43] Z. X. Cang and G. W. Wei, *TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions*, PLOS Computational Biology **13** (2017), no. 7, e1005690.

[44] Z. X. Cang and G. W. Wei, *Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction*, International Journal for Numerical Methods in Biomedical Engineering, page `10.1002/cnm.2914`, 2017.

[45] D. D. Nguyen, T. Xiao, M. L. Wang, and G. W. Wei, *Rigidity strengthening: A mechanism for protein–ligand binding*, Journal of Chemical Information and Modeling **57** (2017), no. 7, 1715–1721.

[46] Z. X. Cang and G. W. Wei, *Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology*, Bioinformatics **33** (2017), no. 22, 3549–3557.

[47] Z. X. Cang, L. Mu, and G. W. Wei, *Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening*, PLoS Computational Biology **14** (2018), no. 1, e1005929.

[48] K. D. Wu and G. W. Wei, *Quantitative toxicity prediction using topology based multi-task deep neural networks*, Journal of Chemical Information and Modeling, page `10.1021/acs.jcim.7b00558`, 2018.

[49] P. Koehl, *Protein structure similarities*, Current Opinion in Structural Biology **11** (2001), no. 3, 348–353.

[50] G. Máté, A. Hofmann, N. Wenzel, and D. W. Heermann, *A topological similarity measure for proteins*, Biochimica et Biophysica Acta (BBA)-Biomembranes **1838** (2014), no. 4, 1180–1190.

[51] C. J. Feinauer, A. Hofmann, S. Goldt, L. Liu, G. Mate, and D. W. Heermann, *Zinc finger proteins and the 3D organization of chromosomes*, Advances in Protein Chemistry and Structual Biology **90** (2013), 67–117.

[52] G. Máté and D. W. Heermann, *Statistical analysis of protein ensembles*, Frontiers in Physics **2** (2014), 20.

[53] A. Collins, A. Zomorodian, G. Carlsson, and L. J. Guibas, *A barcode shape descriptor for curve point cloud data*, Computers and Graphics **28** (2004), no. 6, 881–894.

[54] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of persistence diagrams*, Discrete & Computational Geometry **37** (2007), no. 1, 103–120.

[55] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko, *Lipschitz functions have $l_p$-stable persistence*, Foundations of Computational Mathematics **10** (2010), no. 2, 127–139.

[56] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, The Journal of Machine Learning Research **16** (2015), no. 1, 77–102.

[57] S. Erten, G. Bebek, and M. Koyutürk, *Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks*, Journal of Computational Biology **18** (2011), no. 11, 1561–1574.

[58] C. Lei and J. Ruan, *A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity*, Bioinformatics **29** (2013), no. 3, 355–364.

[59] K. L. Xia and G. W. Wei, *Persistent homology analysis of protein structure, flexibility, and folding*, International Journal for Numerical Methods in Biomedical Engineering **30** (2014), no. 8, 814–844.

[60] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo, *Using persistent homology and dynamical distances to analyze protein binding*, Stat. Appl. Genet. Mol. Biol. **15** (2016), no. 1, 19–38.

[61] K. L. Xia, Z. X. Zhao, and G. W. Wei, *Multiresolution topological simplification*, Journal Computational Biology **22** (2015), 1–5.

[62] K. L. Xia, K. Opron, and G. W. Wei, *Multiscale multiphysics and multidomain models — flexibility and rigidity*, Journal of Chemical Physics **139** (2013), 194109.

[63] K. Opron, K. L. Xia, and G. W. Wei, *Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis*, Journal of Chemical Physics **140** (2014), 234105.

[64] K. Opron, K. L. Xia, and G.W. Wei, *Communication: Capturing protein multiscale thermal fluctuations*, The Journal of Chemical Physics **142** (2015), no. 21, 211101.

[65] K. Opron, K. L. Xia, Z. F. Burton, and G. W. Wei, *Flexibility-rigidity index for protein–nucleic acid flexibility and fluctuation analysis*, Journal of Computational Chemistry **37** (2016), no. 14, 1283–1295.

[66] K. L. Xia, K. Opron, and G. W. Wei, *Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (manm)*, The Journal of Chemical Physics **143** (2015), no. 20, 204106.

[67] D. Nguyen, K. L. Xia, and G. W. Wei, *Generalized flexibility-rigidity index*, The Journal of Chemical Physics **144** (2016), no. 23, 234106.

[68] Shaun Harker, Konstantin Mischaikow, Marian Mrozek, Vidit N, Hubert Wagner, and Mateusz Juda, *The efficiency of a homology algorithm based on discrete morse theory and coreductions*, Proceeding of the 3rd International Workshop on Computational Topology in Image Context, Image A (2010), pages 41–47.

[69] Konstantin Mischaikow and Vidit Nanda, *Morse theory for filtrations and efficient computation of persistent homology*, Discrete Comput. Geom. **50** (2013), no. 2, 330–353.

[70] J. C. Gelly, A. P. Joseph, N. Srinivasan, and A. G. de Brevern, *iPBA: a tool for protein structure comparison using sequence alignment strategies*, Nucleic Acids Research **39** (2011), suppl 2, W18–W23.

[71] P. Røgen and B. Fain, *Automatic classification of protein structure by using Gauss integrals*, Proceedings of the National Academy of Sciences **100** (2003), no. 1, 119–124.

Kelin Xia:
Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University, Singapore 637371
and School of Biological Sciences
Nanyang Technological University, Singapore 637371
*E-mail address*: xiakelin@ntu.edu.sg