SCIENTIFIC REPORTS
nature research

OPEN

Weighted persistent homology for biomolecular data analysis

Zhenyu Meng[1], D. Vijay Anand[1], Yunpeng Lu[2], Jie Wu[3] & Kelin Xia[1,4*]

In this paper, we systematically review weighted persistent homology (WPH) models and their applications in biomolecular data analysis. Essentially, the weight value, which reflects physical, chemical and biological properties, can be assigned to vertices (atom centers), edges (bonds), or higher order simplexes (cluster of atoms), depending on the biomolecular structure, function, and dynamics properties. Further, we propose the first localized weighted persistent homology (LWPH). Inspired by the great success of element specific persistent homology (ESPH), we do not treat biomolecules as an inseparable system like all previous weighted models, instead we decompose them into a series of local domains, which may be overlapped with each other. The general persistent homology or weighted persistent homology analysis is then applied on each of these local domains. In this way, functional properties, that are embedded in local structures, can be revealed. Our model has been applied to systematically study DNA structures. It has been found that our LWPH based features can be used to successfully discriminate the A-, B-, and Z-types of DNA. More importantly, our LWPH based principal component analysis (PCA) model can identify two configurational states of DNA structures in ion liquid environment, which can be revealed only by the complicated helical coordinate system. The great consistence with the helical-coordinate model demonstrates that our model captures local structure variations so well that it is comparable with geometric models. Moreover, geometric measurements are usually defined in local regions. For instance, the helical-coordinate system is limited to one or two basepairs. However, our LWPH can quantitatively characterize structure information in regions or domains with arbitrary sizes and shapes, where traditional geometrical measurements fail.

The great advancement in biological sciences and technologies has led to the accumulation of unprecedented gigantic amount of biomolecular data. Generally speaking, biological data can be classified into several categories, including genomics, transcriptomics, proteomics and metabolomics, which are deposited in several major databanks. A quick look at these databanks gives us a general perspective of the great amount of biological data that is available. Currently, in GenBank, there are more than 100 million gene sequences, which is more than 1 billion bases. In protein data bank (PDB)[1], there are about 150,000 three-dimensional biomolecular structures. The availability of the tremendous amount of the biological data has posed unprecedented opportunities for researchers from all areas. With great opportunities come great challenges. The high dimensionality, complexity, and variety of the biological data have rendered most powerful traditional methods and models useless. Data analysis methods and models, including statistical learning, machine learning, data mining, manifold learning, graph/network models, topological data analysis (TDA), etc, have provided great promise in big data era and became more and more popular in bioinformatics and computational biology in the past two decades. Among these models, TDA has drawn special attention from mathematicians and computational scientists due to its unique characteristics. Unlike the general data analysis models, TDA studies topological invariants, which are global intrinsic structure properties. Roughly speaking, TDA can identify the "shape of the data", thus it works as a powerful tool for simplification and dimensionality reduction. The key component of TDA is persistent homology (PH), which is developed from computational topology and algebraic topology. By assigning a geometric measurement to topological invariants, PH provides a bridge between geometry and topology. Recently, PH based machine learning models have delivered one of the best results in protein-ligand binding affinity prediction, partition coefficients, and mutation-induced folding energy variation[2–7] and won champions in several categories in the recent D3R Grand Challenges[8], which is widely regarded as the most difficult challenge in drug design.

[1]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371, Singapore. [2]Division of Chemistry and Biological Chemistry, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371, Singapore. [3]School of Mathematical Sciences, Hebei Normal University, Shijiazhuang, Hebei, 050024, China. [4]School of Biological Sciences, Nanyang Technological University, Singapore, 637371, Singapore. *email: xiakelin@ntu.edu.sg

The great success of persistent homology based machine learning models depends on the multiscale topological features obtained from biomolecular structures[9–11]. Unlike all previous geometric and topological models, which either focus on local structure information or study qualitative global properties, persistent homology embeds the geometric information into the topological invariants, thus provides the first quantitative topological measurements. With the proposed filtration process, a series of nested simplicial complexes, which are encoded with structural topological information from different scales, are generated in PH. These simplicial complexes provide a multiscale topological profile of the structures. Topological features, such as individual components, holes, circles, and voids, can be evaluated from them. More importantly, some features persist while other die quickly during the filtration. The "lifespans" or "persisting times" provide a size measurement of the topological features[12–14]. With its unique power in data simplification and structure representation, PH has already been applied to various fields, including shape recognition[15], network structure[16–18], image analysis[19–23], data analysis[24–28], chaotic dynamics verification[29], computer vision[23], computational biology[30–37], amorphous material structures[38,39], etc. Many powerful softwares, including JavaPlex[40], Perseus[41], Dipha[42], Dionysus[43], jHoles[44], GUDHI[45], Ripser[46], PHAT[47], DIPHA[48], R-TDA package[49], etc, have been developed. The persistent times of topological features can be represented or visualized by several models, including persistent diagram (PD)[14], persistent barcode (PB)[50], persistent landscape[51,52], persistent image[53], persistent curves[54,55], etc. However, traditional persistent homology models use only one filtration parameter, significantly hindering their applications in revealing some heterogeneous properties. To overcome this problem, several multidimensional filtration PH models have been proposed. These models can greatly boost the performance of traditional PH models. Another important approach is to design weighted persistent homology (WPH). The essential idea of WPH is to introduce weight information, which reflects certain physical, chemical or biological properties, into the simplicial complex generation or homological generator calculation. In this way, the topological features, obtained from WPH, will characterize more heterogeneous biomolecular properties.

Generally speaking, the weighted persistent homology can be characterized into three major categories, vertex-weighted[35,56–60], edge-weighted[2,5,35,44,61,62], and simplex-weighted models[63–65]. For vertex-weighted models, a weight value is defined on each vertex. Among these methods, the weighted alpha complex is the first model that has been used in biomolecular structure characterization[58]. By assigning a weight value to each atom, a modified distance function can be proposed and further used to generate the weighted Voronoi cell. The weighted alpha complex is a subset of weighted Delaunay complex, which is the dual of weighted Voronoi diagram. Other vertex-weighted models consider different types of weighted distance functions[56,57,59]. In a weighted Vietoris-Rips and *Čech* complex model, a new distance function is proposed as a minimal value of the scaled Euclidean distances between the current position to all atoms[56]. The inverse of the distance function represents the union of balls centered at the atom, and naturally induces weighted *Čech* complexes through nerve theorem[56,59]. Weighted Vietoris-Rips complexes can be constructed by scaling the Euclidean distances between any two atoms by their weights[56]. Similarly, k-distance functions are proposed and can be used to produce weighted *Čech* complexes[59]. To characterize the multiscale properties of the biomolecules, a multiscale rigidity function is proposed[35,60,66]. Each atom is associated with a weighted kernel function with a scale parameter. A rigidity function is defined as the summation of all the kernel functions and can be used to generate a series of nested Morse complexes in persistent homology. Unlike vertex-weighted models, edge-weighted models usually specify unique weight values on edges[44,61]. Using weight value as a filtration parameter, Vietoris-Rips or clique complex can be defined as a maximal simplicial complex, whose 1-skeleton has weight values larger (or smaller) than a certain filtration value[57]. For weighted clique rank homology model[61], a network/graph, with a weight value on each edge, is considered. Clique complex can be defined on the subgraph composed of edges with weight larger than filtration value. A series of physics-aware models are proposed to characterize interactions within and between the biomolecules[2–8,35,62]. Various modified or generalized distance matrixes are used in these models. Further, simplex-weighted persistent homology models, based on weighted boundary operators, are proposed[63–65]. In these models, weight values are defined on simplexes in different dimensions. To ensure the consistence of the homology definition, weight values on different simplexes need to satisfy certain constraints or relations, so that a weighted boundary operation can be well-defined.

In this paper, we propose a localized weighted persistent homology LWPH. Our LWPH model is inspired by the recent great success of element specific persistent homology (ESPH) models as mentioned above. Unlike all previous weighted persistent homology models, which treat biomolecules as an inseparable system, ESPH decomposes the structure into a series of sub-structures made of certain type(s) of atoms. The subnetworks or subgraphs, especially those from protein-ligand complexes, have been proved to capture important biological properties, such as hydrophobic or hydrophilic interactions. In this way, ESPH models have delivered amazing results in biomolecular data analysis[2–8]. Further, our LWPH models are different from traditional persistent local homology (PLH)[67–72]. The PLH studies the relative homology groups between a topological space and its subspace. It is usually used to assess the local structure of a special point within a topological space. In our LWPH, the biomolecular structures and configurations are decomposed into a series of local domains, that may overlap with each other. The general persistent homology or weighted persistent homology analysis is then applied on each of these local domains. In this way, DNA local structure, dynamics and functional properties can be embedded in our LWPH models. Our model has been used in the analysis of DNAs. It has been found that our LWPH based features can be used to successfully discriminate A-, B-, and Z-types of DNA. More importantly, our LWPH based principal component analysis (PCA) model can identify two configurational states of a DNA system in ion liquid environment, which can only be revealed by the complicated helical coordinate representation. The great consistence with the helical-coordinate model demonstrates that our model captures the local structure variations so well that it is comparable with geometric models. Moreover, geometric measurements are usually defined in very local regions, for instance the helical-coordinate system is limited to one or two basepairs. However, our localized

weighted homology can quantitatively characterize structure information in a much larger domain, where traditional geometrical measurements fail.

Methods

In this section, we will provide a brief review of the weighted persistent homology models and their applications in biomolecular data analysis. After that, a detailed discussion of our localized weighted persistent homology model will be presented.

Weighted persistent homology. The essential idea of weighted persistent homology models is to introduce a specially-designed weight function/parameter that incorporates the biomolecular physical, chemical or biological properties, into the construction of simplicial complexes or homology generator evaluation. Generally speaking, all these WPH models can be classified into three types, including vertex-weighted[35,56–60], edge-weighted[2,5,35,44,61,62], and simplex-weighted models[63–65].

To facilitate our discussion, we define a weighted point set as (X, V) with $X = \{x_i|_{i=1,2,\ldots,N}\}$ and $V = \{v_i|_{i=1,2,\ldots,N}\}$. For each point x_i, a weight v_i is assigned to it. We use $d(x_i, x_j)$ to represent the Euclidean distances between two points x_i and x_j. Biologically, the weight V is usually chosen to be the radius, atom number, etc.

Vertex-weighted persistent homology. **Weighted alpha complex**: In weighted alpha complex[58], a weighted distance is defined as $d^{\alpha}(x, x_i) = \sqrt{v_i^2 + d(x, x_i)^2}$. The weighted Voronoi region or Voronoi cell can be defined as

$$VC_i = \{x | d^{\alpha}(x, x_i) \leq d^{\alpha}(x, x_j), \textit{ for all } i \neq j\}.$$

A t-weighted closed ball for x_i is defined as $\overline{B}_i^{\alpha}(t) = \{x | d^{\alpha}(x, x_i) \leq t\}$. The intersection of t-weighted closed balls and Voronoi cells is $R_i(t) = \overline{B}_i^{\alpha}(t) \cap VC_i$. In this way, the weighted alpha complex can be expressed as,

$$A(t) = \{\sigma | \bigcap_{x_i \in X} R_i(t) \neq 0\}.$$

Essentially, the weighted alpha complex is a subsect of weighted Delaunay complex, which is the dual of weighted Voronoi diagram.

Weighted Vietoris-Rips and Čech: Bell *et al.*, have proposed a weighted Vietoris-Rips model and a weighted Čech model[56]. The weighted *Čech* complex is defined as $\check{C}ech(X, V) = \mathcal{N}\{\overline{B}(x_i, v_i)|_{i=1,2,\ldots,N}\}$. Here $\overline{B}(x_i, v_i) = \{x | d(x, x_i) \leq v_i\}$ is the closed ball centered at x_i with radius v_i. The nerve $\mathcal{N}$ is the abstract simplicial complex from the closed balls. The weighted Vietoris-Rips Complex is defined as $VR(X, V) = \{\sigma \subset X | d(x_i, x_j) \leq v_i + v_j, \textit{ for all } x_i, x_j \in \sigma \textit{ with } x_i \neq x_j\}$.

For a filtration parameter $t \geq 0$, weighted *Čech* complex at scale t can be denoted as

$$\check{C}ech(X, V, t) = \mathcal{N}\{\overline{B}(x_i, tv_i);\ i = 1, 2, \ldots, N\}, \tag{1}$$

and the weighted Vietoris-Rips Complex at scale t can be expressed as

$$VR(X, V, t) = \{\sigma \subset X | d(x_i, x_j) \leq tv_i + tv_j, \textit{ for all } x_i, x_j \in \sigma \textit{ with } x_i \neq x_j\}.$$

Moreover, we can define a distance function as $f_{X,V}(x) = \min_{x_i \in X}\left\{\frac{d(x, x_i)}{v_i}\right\}$. In this way, we have the inverse function

$$f_{X,V}^{-1}([0, t]) = \bigcup_{x_i \in X} \overline{B}(x_i, tv_i), \tag{2}$$

and it is homotopy equivalent to $\check{C}ech(X, V, t)$ as in Eq. (1).

Computationally, the weighted Vietoris-Rips Complex[56] can be constructed by using a weighted distance matrix $M = \{M_{ij}|_{i,j=1,2,\ldots,N}\}$ with

$$M_{i,j} = \frac{d(x_i, x_j)}{v_i + v_j}.$$

Various softwares, such as JavaPlex[40], Perseus[41], Dipha[42], GUDHI[45], Ripser[46], PHAT[47], DIPHA[48], R-TDA package[49], can use distance matrix as their input data.

k-distance based model

Definition 1. *For any point set X and k is nonnegative integer, the k-distance*[57,59] *can be denoted as*

$$d_{X,k}^2(x) = \frac{1}{k} \sum_{x_i \in NN_X^k(x)} d^2(x, x_i)$$

with $NN_X^k(x)$ *denotes the k nearest neighbors in X to the point x.*

Further, it can be expressed as power distance as follows,

$$d_{X,k}^2(x) = \min \{d^2(x, \overline{x}) - w_{\overline{x}};\ \overline{x} \in Bary^k(X)\}.$$

Here $Bary^k(X)$ denotes the barycenters of any subsets of k points of X and $w_{\overline{x}} = -\frac{1}{k}\sum_{1\le i\le k} d^2(\overline{x}, x_i)$. Moreover, the sublevel sets of the k-distance $d_{X,k}$ are finite union of balls,

$$d_{X,k}^{-1}([0, t]) = \bigcup_{\overline{x}\in Bary^k(X)} B(\overline{x}, (t^2 + w_{\overline{x}})^{1/2}). \tag{3}$$

Similar to Eq. (2), the inverse of this distance function is homotopy equivalent to a weighted *Čech* complex, which is the nerve of the closed balls[59].

Rigidity function based models: Multiscale topological simplification models have been proposed[35,60,66]. The key part of these models is multiscale rigidity function,

$$\mu(x) = \sum_{i=1}^{N} \nu_i \Phi(d(x, x_i); \eta_i).$$

Here η_i is the scale parameter and $\Phi(d(x, x_i); \eta_i)$ can be chosen from any monotonically-decreasing functions, such as the generalized power-law equation,

$$\Phi(d(x, x_i); \eta_i, \nu) = \frac{1}{1 + \left(\frac{d(x,x_i)}{\eta_i}\right)^{\nu}}, \tag{4}$$

and the generalized exponential equation,

$$\Phi(d(x, x_i); \eta_i; \nu) = e^{-\left(\frac{d(x,x_i)}{\eta_i}\right)^{\nu}}. \tag{5}$$

Unlike previous distance functions in Eqs. (2) and (3), the inverse of rigidity function $\mu^{-1}([0, t])$ may not be expressed as a union of balls. However, it can generate Morse complexes. Computationally, the discrete Morse models can be used to evaluate the persistent homology.

Edge-weighted persistent homology. The essential idea for the edge-weighted persistent homology models is to assign a weight to each edge. With weight as filtration parameter, the Vietoris-Rips complex can be defined as the maximal simplicial complex whose 1-skeleton has weight values larger (or smaller) than the filtration value. Computationally, a weighted distance matrix is usually proposed. The filtration is achieved through the increasing (or decreasing) of the weighted distance value.

Weighted clique rank homology: The weighted clique rank homology is defined on weighted complex networks[44,61]. For weighted networks, each edge/link has a weight on it. The filtration goes from the largest weight to the lowest one. At each filtration value t, a subgraph composed of edges with weight larger than t is formed. Based on the subgraph, clique complex can be constructed. In this way, with the decrease of filtration value, a series of clique complexes are built and their homology and persistence can be calculated.

Physics-aware models: Recently, a series of new persistent homology models have been proposed to characterize the various physical interactions within and between the biomolecules[2,5,35,62]. In these models, the distance matrix between atoms is modified based on their physical properties, including covalent bonds, protein-ligand interactions, electrostatic interactions, etc. To avoid confusion, we call them as physics-aware persistent homology.

For a biomolecule or biomolecular complex, we denote their atomic coordinates as $X = \{x_i|_{i=1,2,\ldots,N}\}$, a distance matrix can be constructed as $M = \{M_{ij} = d(x_i, x_j)|_{i,j=1,2,\ldots,N}\}$. Various modified distance matrices are proposed to characterize different physical, chemical and biological properties of the biomolecular structure.

Definition 2. *Multi-level persistent homology model*[2] *considers a modified distance matrix as follows*,

$$M_{ij} = \begin{cases} d(x_i, x_j), & \text{if atoms } i \text{ and } j \text{ are not bonded;} \\ \infty, & \text{if atoms } i \text{ and } j \text{ are bonded.} \end{cases} \tag{6}$$

In computation, we can take ∞ as any value larger than the filtration size. More generally, we can define an n-th level matrix[2] as

$$M_{ij} = \begin{cases} \infty, & d(x_i, x_j) \le n; \\ d(x_i, x_j), & \text{otherwise.} \end{cases}$$

It has been found that when the modified matrices are employed, the barcode representation is significantly enriched and is able to capture the tiny structure perturbation between the conformations. Further, an interactive persistent homology model is proposed for protein-ligand binding analysis.

Definition 3. *An interactive persistent homology model is based on the revised distance matrix as follows*,

$$M_{ij} = \begin{cases} d(x_i, x_j), & \text{if atoms } i \text{ and } j \text{ are from different molecules;} \\ \infty, & \text{otherwise.} \end{cases} \tag{7}$$

In this way, interactions between two molecules, such as protein-protein, protein-DNA/RNA, protein-ligand, DNA/RNA-ligand, etc, can be incorporated into topological invariants.

Essentially, Physics-aware persistent homology models[2] are all based on the generalized matrix $M_{ij} = \Phi(x_i, x_j)$. Here $\Phi(x_i, x_j)$ can be any function properties, including van der Waals interaction, electrostatic potential, or any other generalized correlations.

Simplex-weighted persistent homology. **Weighted simplicial homology**: Weighted simplicial homology is a generalization of simplicial homology[63–65]. Every simplex has a weight in a ring R, and the boundary map is weighted accordingly. When all the simplices have the same weight $a \in R\backslash\{0\}$, the resulting weighted homology is the same as the usual simplicial homology. We list some of the key definitions and results below.

Definition 4. *A weighted simplicial complex is a pair (K, w) consisting of a simplicial complex K and a weight function w: $K \to R$, where R is a commutative ring, such that for any σ_1, σ_2 with $\sigma_1 \subseteq \sigma_2$, we have $w(\sigma_1)|w(\sigma_2)$.*

Theorem 1. *Let I be an ideal of a commutative ring R. Let (K, w) be a weighted simplicial complex, where w: $K \to R$ is a weight function. Then $K\backslash w^{-1}(I)$ is a simplicial subcomplex of K.*

For the definition of homology of weighted simplicial complexes, we require R to be an integral domain with unity 1.

Definition 5. *The weighted boundary map ∂_n: $C_n(K) \to C_{n-1}(K)$ is the map:*

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

where the face maps d_i are defined as:

$$d_i(\sigma) = [v_0, \ldots, \hat{v}_i, \ldots, v_n] \text{(deleting the vertex } v_i)$$

for any n-simplex $\sigma = [v_0, \ldots, v_n]$.

Theorem 2. *Let f: $K \to L$ be a simplicial map. Then $f_{\#}\partial = \partial f_{\#}$, where ∂ refers to the relevant weighted boundary map.*

Definition 6. *We define the weighted homology of a weighted simplicial complex to be*

$$H_n(K, w) = ker(\partial_n)/Im(\partial_{n+1}),$$

where ∂_n is the weighted boundary map.

Proposition 2.1. *Proposition. If all the simplices in (K, w) have the same weight $a \in R\backslash\{0\}$, the weighted homology functor is the same as the usual simplicial homology functor.*

Weighted persistent homology: Given a weighted filtered complex $(K, w) = \{(K^i, w)\}_i$, for the i-th complex K^i, we have the associated weighted boundary maps ∂_k^i and chain group C_k^i, cycle group Z_k^i, boundary group B_k^i, and homology group H_k^i for all integers i and k.

Definition 7. *The weighted boundary map ∂_i^k, where i denotes the filtration index, is the weighted boundary map of the i-th complex K^i. That is, ∂_k^i is the map ∂_k^i: $C_k(K_i, w) \to C_{k-1}(K_i, w)$. The chain group C_i^k is the group $C_k(K_i, w)$. The cycle group Z_k^i is the group $ker(\partial_i^k)$, while the boundary group B_i^k is the group $Im(\partial_{k+1}^i)$. The homology group H_k^i is the quotient group Z_k^i/B_k^i.*

Definition 8. *The p-persistent k-th homology group of $(K, w) = \{(K_i, w)\}_{i=0}$ is defined as*

$$H_k^{i,p}(K, w) := Z_k^i/(B_k^{i+p} \bigcap Z_k^i)$$

Localized weighted persistent homology. In all the above WPH models, weights are defined on the whole system to reveal the intrinsic global structural properties. Stated differently, all these models treat a biomolecule structure as an inseparable system, and explore their topological properties using the whole structure. In contrast, ESPH models[2–8] decompose a biomolecular structure into a series of sub-structures made of certain type(s) of atoms. It has been found that the generated subnetworks or subgraphs, especially those from protein-ligand complexes, can capture important biological properties, such as hydrophobic or hydrophilic interactions[2]. Different from WPH and ESPH models, persistent local homology considers the relative homology groups between a topological space and its subspace. Usually, persistent local homology focuses on the local structure around a special point in a certain topological space.

Motivated by the success of ESPH models, we introduce our localized weighted persistent homology. Instead of decomposing biomolecular structures by their atom types, we focus on local biomolecular regions or domains and study their topological properties. For a better introduction of our LWPH, we will briefly review ESPH and PLH first.

Element specific persistent homology: Biomolecules are made of various atoms with different properties. In protein, DNA or RNA, there are five common types of atoms, including *C*, *N*, *O*, *P*, and *S*, and several metal ions, such as *Fe*, *Mn*, *Zn*, etc. Ligands are small molecules, that interact with the protein, DNA or RNA. Other than the common five types of atoms, they may have some other unique atoms, such as *F*, *Cl*, *Br*, *I*, etc.

Generally speaking, ESPH is proposed to characterize the topological properties within the structure formed by one or several types of atoms[2–5,7]. Mathematically, it is achieved by assigning weight value 1 to the selected types of atoms, and weight value 0 to all the rest atoms. For instance, all *C* atoms from both protein and ligand can be selected to form *C*-networks or graphs. The topological properties of these structures characterize the

hydrophobic interactions between protein and ligands. Similarly, hydrophilic interactions can be well captured by networks from protein nitrogen atoms and ligand oxygen atoms.

One of the most important properties for ESPH is to generate a series of sub-structures from the biomolecule, and systematically explore their topological properties. These element based sub-structures reveal more structure information that is directly related to the biomolecular physical, chemical, and biological properties.

Persistent local homology. The persistent local homology[67–71] is based on the algebraic topological concept called local homology groups[72].

Definition 9. *If X is a space and if $x \in X$ is a point, then the local homology groups of X at x are the singular homology groups $H_k(X, X - x)$.*

From the excision theorem, we have the following Lemma.

Lemma 2.1. *Let X be a Hausdorff space and let $A \subset X$. If A contains a neighborhood of the point x, then $H_k(X, X - x) \simeq H_k(A, A - x)$. Therefore, for Hausdorff spaces X and Y, if $x \in X$ and $y \in Y$ have neighbourhoods U, V, respectively, such that (U, x) is homeomorphic to (V, y), then the local homology groups of X at x and of Y at y are isomorphic.*

Generally speaking, local homology is used to evaluate the local structure of a topological space. Persistent local homology can be used in dimension reduction and manifold dimension detection[69–71].

Localized weighted persistent homology: The essential idea of our LWPH is to focus on local regions, of biomolecular structures or configurations, that incorporate the important physical, chemical or biological information, and perform the persistent homology analysis on them. In many situations, there may be several domains that are of great interest and we need to perform LPH on each domain. More generally, we can decompose biomolecular structures into a series of (overlapping) domains. For each domain, we carry out LWPH on all (or certain type(s) of) atoms that are of particular interest. In this way, a series of local topological properties can be obtained and we call them localized topological fingerprints. To avoid confusion, if the general persistent homology is considered, we call it localized persistent homology (LPH). If a weighted persistent homology is considered, we call it localized weighted persistent homology (LWPH).

More specifically, for a biomolecule or biomolecular complex with atomic coordinates $X = \{x_i|_{i=1,2,\ldots,N}\}$. The coordinate set X can be decomposed into a series of domains X^I, with $X = \cup_{I=1}^{m} X^I$. Similar distance matrix M^I as in Eqs. (6) and (7) can be constructed on each of the domain. In this paper, we will focus on the study of DNA, which is made of paired nucleotides. We can consider the weighted distance matrix on the domain X^I as follows,

$$M_{ij}^I = \begin{cases} d(x_i, x_j), & x_i, x_j \in X^I, \text{ atoms } i \text{ and } j \text{ are from different nucleotide residues;} \\ \infty, & \text{otherwise.} \end{cases} \tag{8}$$

By using different weight values, LWPH can be designed to capture various local properties in biomolecular structures. The application of LPH and LWPH can be found in following sections.

It should be noticed that we still regard our LPH as a weighted persistent homology model. Essentially, we can set weight as a vector composed of only 1 and 0, and only atoms within the local region have weight values as 1.

Results

In this section, we discuss the application of our localized persistent homology and localized weighted persistent homology in the study of DNA structures. To avoid confusion, only the weight definition as in Eq. (8) is used in our LWPH. The persistent barcodes are used for the representation and visualization of LPH and LWPH results. The persistent Betti numbers (PBNs) are evaluated from barcodes. A systematical evaluation of PBNs under helical coordinates demonstrates the incorporation of the geometric information in our LPH and LWPH models. Further, we show that PCA of the feature vectors from LWPH based PBNs can be successfully used in the classification of the A-, B-, and Z-types of DNAs. Moreover, we explore DNA structure variations in both water and ion liquid (IL) environments using molecular dynamics. With LWPH based feature vectors, we can not only reveal the confinement effect of DNA configurations from water to IL environment, but also identify two DNA configurational states in IL environment. Detailed analysis shows that global-scale PCA models, including atom-coordinate-based PCA, common PH-based PCA, and ESPH-based PCA, all fail in clustering the DNA configurational states in IL. In contrast, LWPH based all-atom or selected-atom models are always able to characterize the DNA structure variations in IL environment. The LWPH results are highly consistent with the helical-coordinate based PCA model.

DNA local topological features. DNA molecule has a double helical structure composed of four types of nucleotides, i.e., A, T, G, and C. It has remarkably different scales, ranging from nucleobases, minor and major grooves, to larger structures like nucleosome, chromatin, and then to chromosome. We have proposed a multi-resolution PH model to characterize structural topological features or topological fingerprints of DNA structures in different scales[60,66]. However, the model focuses on the global DNA topology. In this section, we explore DNA topological features from local structures, i.e., DNA localized topological fingerprints. We study the barcodes for both LPH and LWPH models, i.e., one with traditional persistent homology and the other with the weighted persistent homology as in Eq. (8). Different local structures can be systematically studied. In the current paper, our focus is local topological features from different DNA base pairs or base steps.

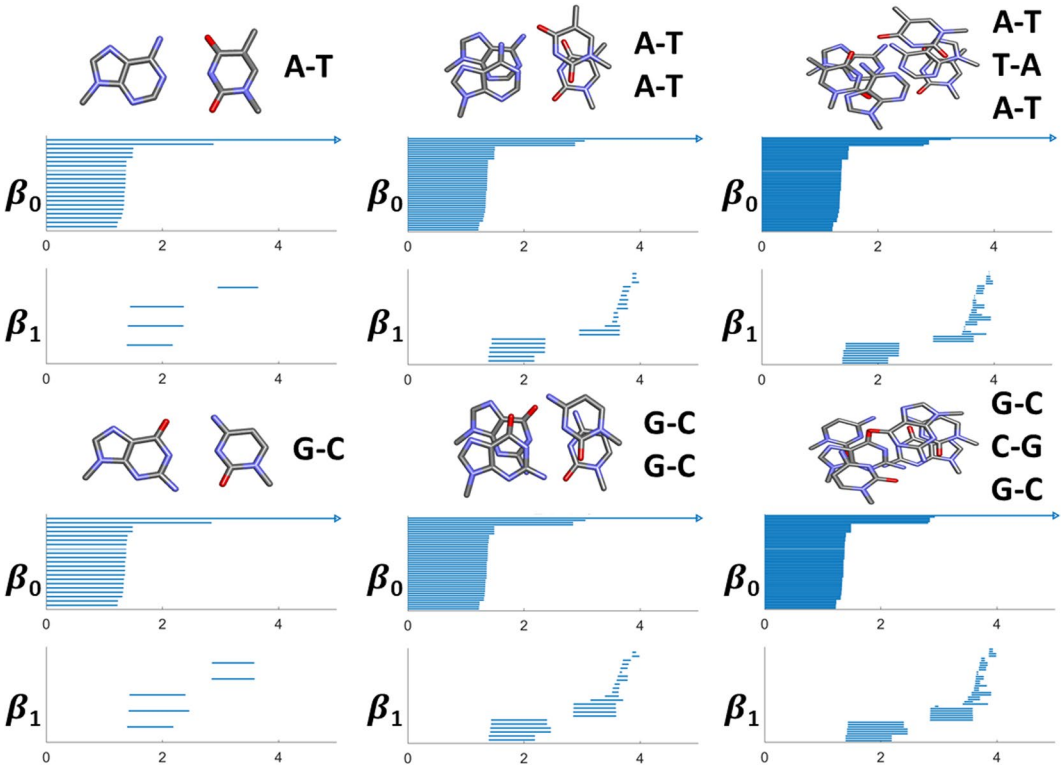

Figure 1. The LPH based barcodes for different combinations of DNA base pairs. It can be seen that A-T and G-C base pairs all have three local β_1 bars from around 1.4 Å to 2.4 Å. But they differ greatly in the global region, where A-T pair contributes one significant β_1 bar from around 2.9 Å to 3.6 Å, while G-C pair generates two. These barcode fingerprints characterize the intrinsic DNA structure properties, i.e., local and global loop/ring motifs.

To facilitate a better description, we introduce some basic notations. In general, results from PH can be represented as pairs of "birth" and "death" times, i.e., the filtration values for homology generators to appear and disappear. We denote them as follows,

$$L_k = \{l_j^k | l_j^k = b_j^k - a_j^k;\ k \in \mathbb{N};\ j \in \{1, 2, \ldots, N_k\}\},$$

here a_j^k, b_j^k, l_j^k to represent "birth", "death", and "persistence" for j-th generator of k-th dimensional Betti number, respectively. And N_k is the total number of k-th dimensional topological generators. Due to the limited number of atoms in a local structure, we only consider dimension k equals to 0 and 1. Further, different PH based functions are proposed for the visualization, representation and modeling of topological information[24,51,54,73]. Persistent Betti number, or Betti curve, is one of them. It is defined as the summation of all the k-th dimensional barcodes,

$$f_{\mathrm{PBN}}(x; L_k) = \sum_j \chi_{[a_j^k, b_j^k]}(x). \tag{9}$$

Function $\chi_{[a_j^k, b_j^k]}(x)$ is a step function, which equals to one in the region $[a_j^k,\ b_j^k]$ and zero otherwise.

As a simple one-dimensional function, PBN has been used in data analysis for dimensionality and complexity reduction. Moreover, PBN based feature vectors can be input into various machine learning models[2–8]. In this section, PBN based PCA models are used in DNA structure classification and trajectory clustering.

LPH and LWPH for DNA structure representation. As stated above, we consider two types of LPHs, one is LPH with Euclidean distance matrix and the other is LWPH with a similar weighted distance matrix as in Eq. (8). For LWPH, we assume the distance between two atoms from same nucleotide residue to be infinity, and the distance between two atoms from different residues to be their Euclidean distance. In this way, our LPH characterizes topology from covalent-bond-formed structure, while our LWPH reveals topological information on non-covalent bond properties. The DNA base-pair structures are generated by using the 3DNA software[74].

Figure 1 illustrates LPH barcodes for different combinations of base pairs. Shorter β_0 bars (with length around 1.0 Å to 1.5 Å) correspond to covalent bonds, longer β_0 bars with length around 2.8 Å correspond to the hydrogen bonds between paired bases. For β_1 bars, longer ones appear in earlier stage of filtration, i.e., from around 1.4 Å to 2.4 Å, correspond to sugar rings and nitrogenous base rings. The longer β_1 bars range from around 2.9 Å to 3.6 Å, representing loops between paired bases. It can be seen that each A-T pair only contributes one such longer β_1 bar, while each G-C pair contributes two.

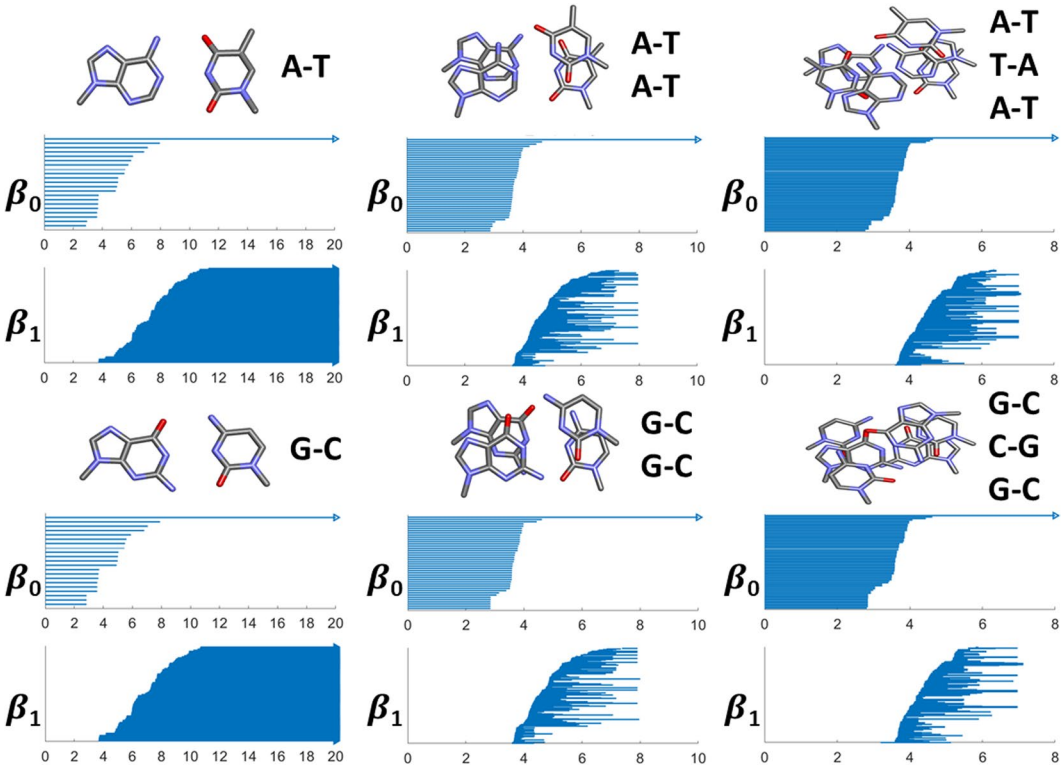

Figure 2. The LWPH based barcodes for different combinations of DNA base pairs. The weighted distance matrix in Eq. (8) is considered. Only the distances between two atoms from different nucleobases are set to be their Euclidean distance, while other distances are set to infinity. The shortest β_0 bars are distances between adjacent atoms from two bases. They characterize the hydrogen bonds between two nucleobases. For a single base pair (A-T or G-C) situation, the generated β_1 bars will never be "killed" as no 2-simplexes can be formed.

Figure 2 illustrates the corresponding LWPH barcodes for different combinations of base pairs. Similar to LPH results, the total number of β_0 bars is exactly the number of atoms, and shorter β_0 bars with length around 2.8 Å correspond to hydrogen bonds between paired bases. Different from LPH results, more hydrogen-bond related β_0 bars appear in LWPH barcode than LPH barcodes. And the lengths of β_0 bars for LWPH are systematically longer than those for LPH model, indicating that more long-range interactions related information are preserved in our LWPH model. Moreover, the β_1 barcodes for LWPH are much more complicated. Their geometric meanings are not as straightforward as LPH models. Generally speaking, the β_1 bar in LWPH represents loop or ring structure with edges between different nucleotides. In this way, a much larger amount of β_1 bars are generated. Moreover, when there are only two nucleotides (or one base pair), the β_1 bars persist forever.

In general, LPH and LWPH characterize different structure properties, the former is more about covalent-bond related topology while the latter is more about topology from non-covalent bonds. Physically, covalent bonds are much stronger than non-covalent bonds. In this way, LPH are relatively more "stable" and less sensitive to structure variations under thermal fluctuations. While LWPH are less "stable" and much easier to change if there is some external perturbations.

LPH and LWPH based DNA structural analysis. With the embedded geometric information, LPH and LWPH can be used in not only qualitative but also quantitative analysis of different structures. To assess LPH and LWPH based quantitative analysis, we systematically generate a series of DNA base-pair configurations using DNA helical coordinates. According to the Cambridge University Engineering Department Helix computation Scheme (CEHS)[75], the motion of a base pair or two neighbouring base pairs can be depicted by 12 helical parameters, including 6 one-base-step related parameters, i.e., shear, stretch, stagger, buckle, propeller and opening, and another 6 two-base-step related parameters, i.e., shift, slide, rise, tilt, roll and twist. For each parameter, we prepare 11 DNA structures, with parameter value taken equally from $\mu_i - 2\sigma_i$ to $\mu_i + 2\sigma_i$, using 3DNA[74]. Here μ_i, σ_i are the mean value and standard deviation of parameter i. And the rest of the helical parameters remain as constants, i.e., their mean values. The mean value and standard deviation of each parameter can be obtained from crystal structures[75]. In DNA helical coordinate evaluation, only the base atoms and $C1'$ of the sugar ring are considered. For a fair comparison, the same atoms are used in our LPH and LWPH models.

We apply our LPH and LWPH on these series of DNA local structures and check if the quantitative structure variations are reflected in their barcodes. To facilitate a systematical comparison, we consider the PBN function in Eq. (9). For each helical parameter, we take the natural logarithm of all its PBN functions and stack them together to form a two-dimensional image. Note that we systematically add 1 to all PBN functions to avoid

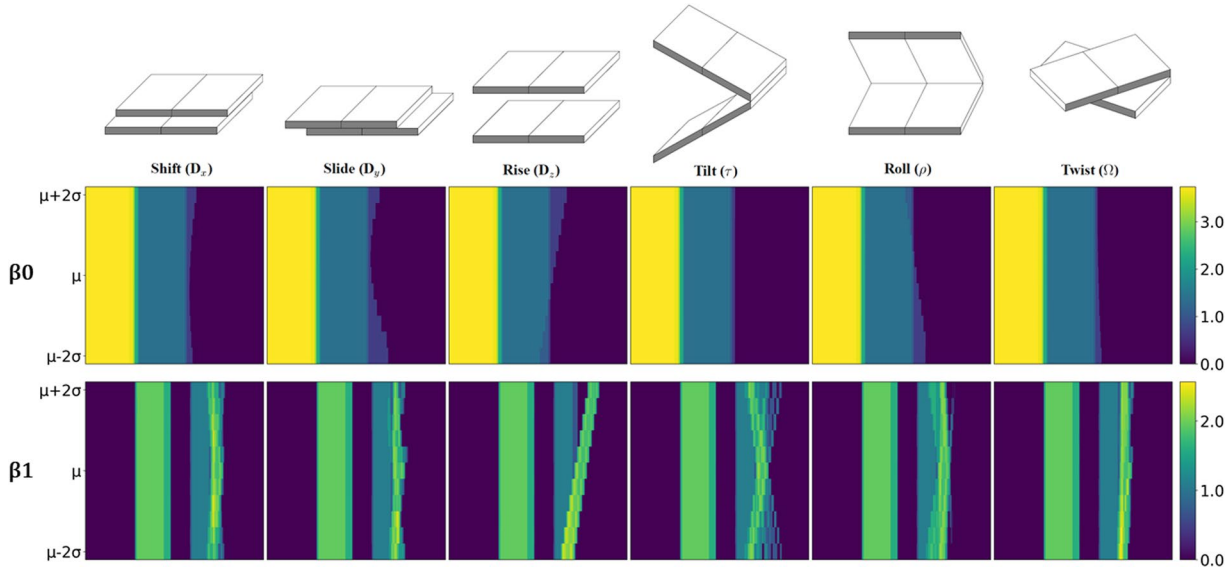

Figure 3. The LPH based PBN image representation for two-base-step (AT/AT) at different helical parameter values. In the i-th PBN image, we systematically change the i-th helical parameter value from $\mu_i - 2\sigma_i$ to $\mu_i + 2\sigma_i$, with all other helical parameters remaining as constants, to deliver a series of base-step structures. PBN can be calculated for each two-base-step structure and all of them stacked together to form a two-dimensional image. It can be seen that, both β_0 (upper figures) and β_1 (lower figures) PBN functions vary with the change of helical parameter value. The change of β_1 PBN functions seem to be more dramatic. Note that the color values are logarithm of PBNs.

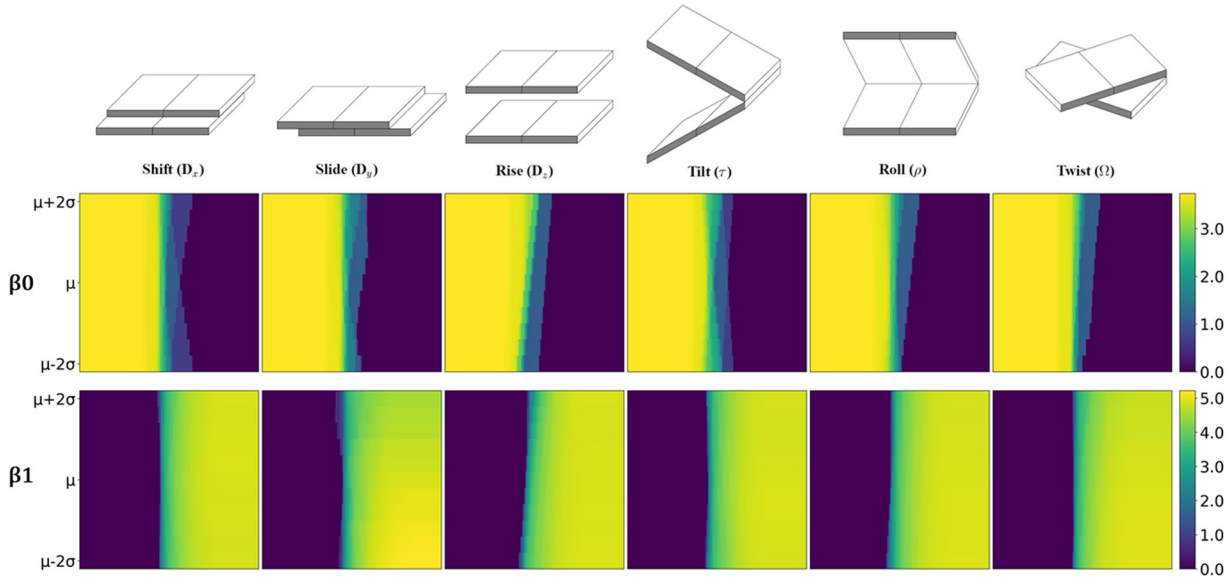

Figure 4. The LWPH based PBN image representation for each two-base-step (AT/AT) at different helical parameter values. Base-step structures are prepared in the same way as in Fig. 3. It can be seen that, similar to LPH based PBNs, LWPH based β_0 (upper figures) and β_1 (lower figures) PBN functions vary with the change of helical parameter value. However, the change of β_0 PBN functions seem is more dramatic in LWPH models.

computational problem (from ln 0). Figures 3 and 4 illustrate the results of our LPH and LWPH for two-base-step parameters, respectively. The AT/AT base steps are considered. It can be seen that instead of remaining unchanged for all helical parameter values, both β_0 and β_1 PBN functions for LPH and LWPH models vary greatly, indicating that both models are sensitive to subtle structure variations. More specifically, in LPH based PBN, β_1 functions seem to have comparably larger variations than β_0 functions. In LWPH based PBN, β_0 functions seem to have greater variations than β_1 functions. We also check the β_0 and β_1 PBN functions for LPH and LWPH models for one base step. The results are demonstrated in Figs. 9 and 10 in Supplementary. Both functions in those two models show variations with the change of helical parameter value.

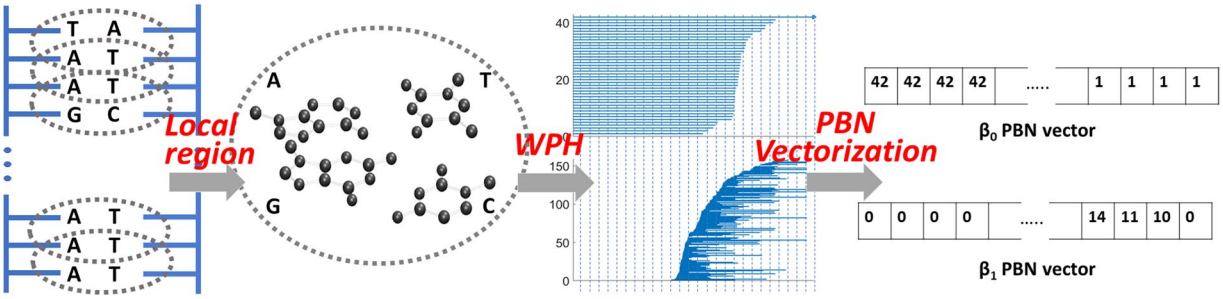

Figure 5. LPH and LWPH based DNA featurization. In our DNA cases, the local region is defined as the two adjacent base-steps. The common Euclidean distance matrix is considered in LPH, while the weighted distance matrix as in Eq. (8) is used in LWPH. From LPH or LWPH, PBNs can be calculated and then discretized into vectors, which can be used as a representation of DNA structures.

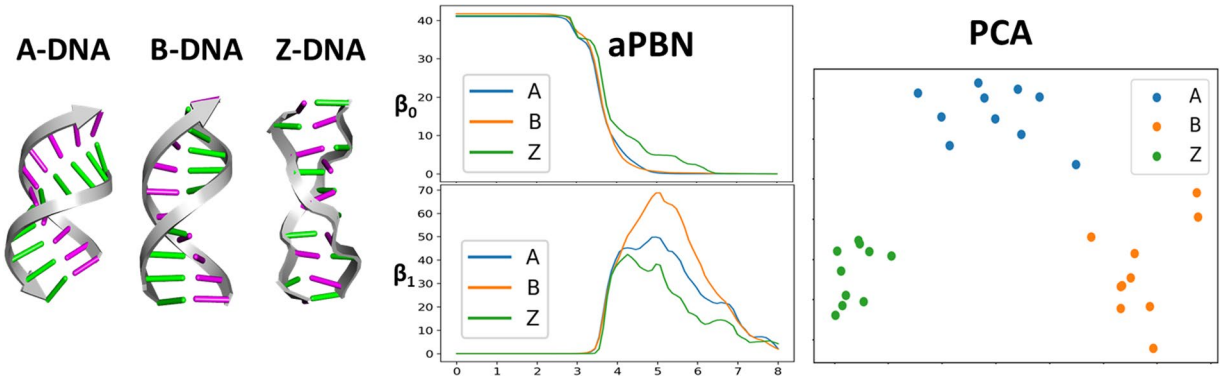

Figure 6. LWPH based classification of three DNA types, i.e., A-DNA, B-DNA, and Z-DNA. The average persistent Betti number (aPBN) from our LWPH for three types of DNAs. We discretize the aPBN equally into a series of numbers and use these values as features for PCA. It can be seen that LWPH based aPBN and PCA results can clearly discriminate the three DNA types.

Since PBN functions from LPH and LWPH are sensitive to the structure variations, we can use them as measurements for DNA structure, function and dynamics analysis. In the following sections, PBN and PBN based features are used in the classification of DNA types and clustering of DNA trajectories.

Local topological feature based DNA classification and clustering. In this section, we study local topological feature based DNA classification and clustering. Essentially, topological features are extracted from PBNs, which are generated from LPH or LWPH. Figure 5 illustrates the LPH and LWPH based DNA featurization. Note that the results from LPH and LWPH can also be represented as persistent diagram[14], persistent barcode[50], persistent landscape[51,52], persistent image[53], persistent curves[55], etc. Based on these representations, other featurization forms can also be considered[76]. In this paper, we focus on PBN based featurization.

Classification of three typical DNA forms. We consider three types of DNA structures, including A-, B- and Z-forms. We randomly pick 10 PDB files from each form of DNA and the PDB IDs are shown in Table 2 in Supplementary. In LPH and LWPH, the same atom combination of each base step is chosen as in the above case, i.e., base atoms and $C1'$ of the sugar ring. The PBN function is calculated for each base step. To systematically compare the PBNs for three types of DNA forms, we summarize all the PBNs from the same DNA form and then compute the average.

The results from LWPH are demonstrated in Fig. 6. It can be seen that the three PBN profiles have very different β_1, particularly on the filtration range from 4.0 Å to 7.0 Å. Further, we consider the PCA for DNA classification. For each DNA structure, we define a vector made of the average β_0 and β_1 PBN values equally taken from 2.0 Å to 8.0 Å with an interval 0.1 Å. In this way, a feature vector with 120 elements is defined for all 30 DNA structures. The PCA results are demonstrated in Fig. 6. Here x-axis and y-axis represent the first and second eigenvectors (principal components), respectively. It can be seen that three forms of DNA locate in the different regions with clear boundary, which further confirms that LWPH based features can distinguish the subtle conformational deviation. Figure 11 in Supplementary shows the results from LPH. In comparison with LWPH, LPH based PBNs show no obvious difference, and PBN based PCA does not classify the dataset into three individual clusters.

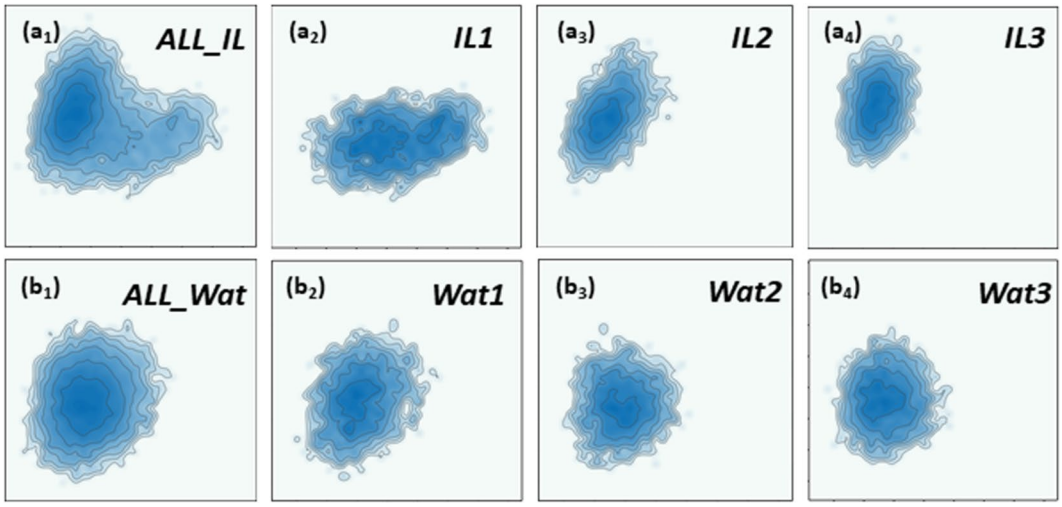

Figure 7. The contour map generated from our LWPH based PCA models for DNA configurations in IL and WAT environments. The x-axis and y-axis are the first and second principal components. ($\mathbf{a_1}$) The DNA configuration ensemble for all three trajectories for IL. ($\mathbf{a_2}$–$\mathbf{a_4}$) Three DNA configuration trajectories from the MD simulation with IL. ($\mathbf{b_1}$) The DNA configuration ensemble for all three trajectories in water environment. ($\mathbf{b_2}$–$\mathbf{b_4}$) Three DNA configuration trajectories from the MD simulation using water solution. It can be seen clearly that areas of contour map in IL are much smaller than in water, indicating the confinement effect of the DNA configurations in IL. Further, contour graph for IL1 ($\mathbf{a_2}$) shows a clear difference of that for IL2 ($\mathbf{a_3}$) and IL3 ($\mathbf{a_4}$), meaning there is a subtle change of the ion-DNA binding mode in trajectory IL1.

Clustering of DNA conformations in different environments. We have demonstrated the DNA structure classification with our LWPH. However, A-, B- and Z- forms of DNAs are static and have relatively "large" configurational differences. In the following, we consider a more challenging case. That is the clustering of the molecular dynamics simulations of the same DNA molecule in different solvent environments.

Molecular dynamics setting. A brief introduction of the MD procedure is presented as follows. The initial structure of 16-mer DNA duplex is prepared using 3DNA[74] and centered in a cubic box. Two different solution environments are used, including ion liquid (IL) and water (WAT). For IL environment, 600 $BMIM^+$ and 600 BF_4^- are firstly inserted and the box is then solvated with TIP3P water and Na^+. For WAT environment, the box is directly solvated with water and Na^+. After a 100 ps thermostat and 100 ps barostat, the system then goes through a 100 ns product MD. Under each environment setting, we conduct 3 repeated MD simulations, so we obtain 6 trajectories in total. We denoted them as IL1-3 (trajectory 1 to 3 in IL) and WAT1-3 (trajectory 1 to 3 in water). All the simulations are conducted using GROMACS 4.6 package[77]. In our data analysis, 5000 sample frames evenly sampled from the last 10 ns trajectory of 6 simulations are considered. The detailed MD simulation setting and parameters can be found in the related paper[78].

Weighted persistent homology modeling. For DNA conformation clustering, we extract 13 non-terminal DNA base steps for each frame of the simulation data. Similar to DNA-type classification case, for each base step, we construct a 120-element PBN feature from the LWPH. Then, we concatenate all 13 sets of PBN values together into feature vector for each DNA configuration. This LWPH based feature vector is used in the PCA of DNA trajectories in different environments. More specifically, a covariance matrix of feature vectors for all the frames is built up, and further eigen-decomposed into principal components (eigenvectors). The first two eigenvectors construct a plane and all feature vectors are projected to it. For comparison, in both IL and water, we apply PCA not only on three individual MD trajectories, but also the ensemble made of all three trajectories together. The results are demonstrated in Fig. 7. The projected points are illustrated as their contour values for a better visualization. To avoid confusion, Fig. 7(a_1–a_4) are for DNA trajectories in IL environment. Among them, Fig. 7(a_1) is for the ensemble of all three trajectories. Figure 7(a_2–a_4) are for the three trajectories, respectively. Figure 7(b_1–b_4) are for DNA trajectories in water environment. Among them, Fig. 7(b_2–b_4) are for the three trajectories, respectively, and Fig. 7(b_1) is for the ensemble of all three trajectories.

Several unique properties can be seen from Fig. 7. Firstly, the confinement effect, i.e., the reduction of distribution area, can be clearly observed in IL environment. In fact, we can count the area of the distribution in IL and water and the results are listed in Table 1. It can be seen that, the areas of distribution map in IL are up to 2/3 of that in water, the smaller area confirms the fluctuation of DNA in IL is greatly attenuated. Secondly, contour graphs for IL and WAT show significantly different patterns. For IL solution, two centers can be clearly identified from contour graph in Fig. 7(a_1). Among them, one locates on upper left and the other locates on right part. In contrast, for water solution, only one center can be found and its position differs greatly from the ones of IL systems. The huge difference indicates the change of DNA conformations in different environments. Further, for water solution, all contour graphs have nearly the same distribution, indicating that all three trajectories behaved

	IL1	IL2	IL3	Wat1	Wat2	Wat3
AC-PCA	212	167	173	307	301	268
HP-PCA	—	217	203	350	352	341
LPH-PCA	340	321	333	345	357	354
LWPH-PCA	—	211	186	354	295	313

Table 1. The confinement effect in ion liquid (IL) solution. The area of distribution of each MD trajectory repeat. The area is counted as the number of grids with population larger than 5% of the largest population grid. Four different methods are considered, including the atom-coordinate-based PCA (AC-PCA), the helical-parameter-based PCA (HP-PCA), the LPH-based PCA (LPH-PCA), and LWPH-based PCA (LWPH-PCA). We have not listed the area for IL1 from HP-PCA and LWPH-PCA, as dynamic transition is observed in these representations and the area cannot be treated as the measurement of degree of conformational fluctuation.

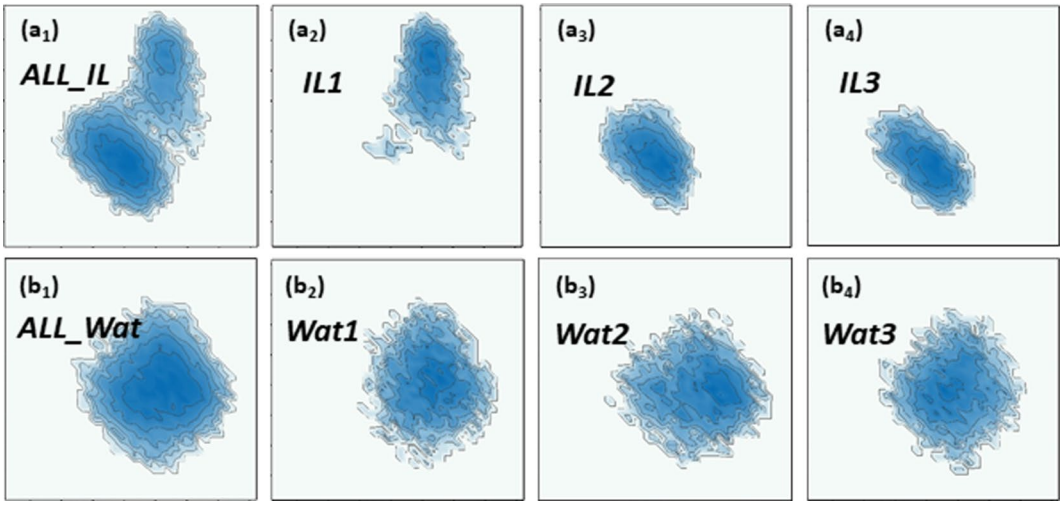

Figure 8. The contour map generated from helical-parameter based PCA models for DNA configurations in IL and WAT environments. The x-axis and y-axis are the first and second principal components. ($\mathbf{a}_1$) The DNA configuration ensemble for all three trajectories for IL. ($\mathbf{a}_2$–$\mathbf{a}_4$) Three DNA configuration trajectories from the MD simulation with IL. ($\mathbf{b}_1$) The DNA configuration ensemble for all three trajectories in water environment. ($\mathbf{b}_2$–$\mathbf{b}_4$) Three DNA configuration trajectories from the MD simulation using water solution. The same confinement effect and two center distribution of DNA configurations in IL environment as in Fig. 7 are observed.

quite similarly. In contrast, for IL solution, contour graph for IL1 shows a clear difference of those for IL2 and IL3, meaning there is a subtle change of ion-DNA binding mode in trajectory IL1. Figure 12 in Supplementary shows the observed variation of ion-DNA binding in different metastates in IL1. Our LWPH based PCA is sensitive enough to capture this subtle structure variation.

Our LWPH based results are highly consistent with previous results from the helical parameter based model[78]. As demonstrated in Fig. 8, similar confinement effect is also observed in CEHS based contour graphs for IL solution. Further, CEHS results show two different centers, while WAT contour graphs have only one center. This means that, CEHS model also captures the subtle change of ion-DNA binding mode in trajectory IL1. To further check if our LWPH and CEHS models identify the same type of ion-DNA configurational changes, we decompose the contour graphs of IL1 into 10 separated subgraphs, each of them represent the DNA trajectories in 1 ns. We can clearly identify four center regions from these subgraphs, and they are consistent with results from HP based PCA. This further indicates that our LWPH based models are highly sensitive to DNA local structural variations. The results are demonstrated in Figs. 13 and 14 in Supplementary.

Further, it should be noted that the local DNA structure variations in IL1 cannot be captured by the general global models. As demonstrated in Fig. 15 in Supplementary, the general atom-coordinates based PCA fails to capture the DNA structure variation in IL1, even thought it manages to preserve the confinement effect (details in Table 1). Similarly, LPH based model also fails to reveal the variation. Moreover, it even cannot reveal the confinement effect of IL environment. Details can be found in Table 1 and Fig. 16 in Supplementary. This is largely due to the reason that LPH based model focuses more on the covalent bonds and its related structures. Even though different combination of atoms are considered, both coordinate-based PCA and LPH-based PCA are unable to identify the structure variation. In contrast, not only the general LWPH model works, we also construct different LWPH models by taking different combinations of backbone atoms and base atoms at local scale. For instance, according to CEHS scheme, we can take $C8$, $C4$, $N1$ and $C1'$ of purine base and $N3$, $C6$ and $C1'$ of pyrimidine base, as shown in Fig. 17 in Supplementary. These selected atoms based LWPH can also capture very well the

confinement effect and ion-DNA configurational changes. The corresponding trajectories also show great consistence with both helical-parameter based and LWPH based results. Figures 18 and 19 in Supplementary show the corresponding results.

Lastly, our LWPH based models are very flexible and easy to be combined with machine learning models. Traditional CEHS helical coordinate systems work well only for one and two base steps. They tend to fail if the local structure variation is induced by adjacent three or more base steps. Moreover, CEHS models are only suitable for DNA or RNA and cannot be used in proteins or other biomolecules. In comparison, our LWPH are more general and can be used for any local structures from DNAs, RNAs, proteins, biomolecular complexes, or biomolecular assemblies. Another important property of LWPH based feature vectors are that they are unit free and can be used to compare different-sized local structures. In this way, these feature vectors are extremely suitable for machine learning models.

Conclusion Remarks

In this paper, we discuss weighted persistent homology models and their applications in biomolecular structure, function, and dynamics analysis. We briefly review all the WPH approaches, including vertex-weighted, edge-weighted, and simplex-weighted models. Essentially, weight values, which reflects physical, chemical and biological properties, are assigned to vertices (atom centers), edges (bonds), or higher order simplexes (cluster of atoms), depending on the biomolecular structure, function, and dynamics properties.

Further, we propose the first localized persistent homology and localized weighted persistent homology and apply them in the DNA structure classification and clustering. Our LPH and LWPH models are inspired by the great success of element specific persistent homology. In our models, biomolecules are not treated as an inseparable system, instead they are decomposed into a series of local domains, which may overlap with each other. The general persistent homology or weighted persistent homology analysis is then applied on each of these local domains. In this way, functional properties, that embedded in localized topological invariants, can be revealed. Our models characterize structural variations at any level and provide a new featurization of biomolecules for machine learning models.

Received: 30 September 2019; Accepted: 29 November 2019;
Published online: 07 February 2020

References

1. Berman, H. M. *et al*. The Protein Data Bank. *Nucleic acids research* **28**(1), 35–242 (2000).
2. Cang, Z. X., Mu, L. & Wei, G. W. Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring And Virtual Screening. *PLoS computational biology* **14**(1), e1005929 (2018).
3. Cang, Z. X. & Wei, G. W. Analysis and Prediction of Protein Folding Energy Changes Upon Mutation by Element Specific Persistent Homology. *Bioinformatics* **33**(22), 3549–3557 (2017).
4. Cang, Z. X. & Wei, G. W. Integration of Element Specific Persistent Homology and Machine Learning for Protein-Ligand Binding Affinity Prediction. *International journal for numerical methods in biomedical engineering*, page, https://doi.org/10.1002/cnm.2914 (2017).
5. Cang, Z. X. & Wei, G. W. TopologyNet: Topology Based Deep Convolutional And Multi-Task Neural Networks for Biomolecular Property Predictions. *PLOS Computational Biology* **13**(7), e1005690 (2017).
6. Nguyen, D. D., Xiao, T., Wang, M. L. & Wei, G. W. Rigidity Strengthening: A Mechanism for Protein–Ligand Binding. *Journal of chemical information and modeling* **57**(7), 1715–1721 (2017).
7. Wu, K. D. & Wei, G. W. Quantitative Toxicity Prediction Using Topology Based Multi-Task Deep Neural Networks. *Journal of chemical information and modeling*, page, https://doi.org/10.1021/acs.jcim.7b00558 (2018).
8. Nguyen, D. D. *et al*. Wei. Mathematical Deep Learning for Pose and Binding Affinity Prediction and Ranking in D3R Grand Challenges. *Journal of computer-aided molecular design* **33**(1), 71–82 (2019).
9. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological Persistence and Simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002).
10. Zomorodian, A. & Carlsson, G. Computing Persistent Homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).
11. Zomorodian, A. & Carlsson, G. Localized Homology. *Computational Geometry - Theory and Applications* **41**(3), 126–148 (2008).
12. Dey, T. K., Li, K. Y., Sun, J. & David, C. S. Computing Geometry Aware Handle and Tunnel Loops in 3d Models. *ACM Trans. Graph.* **27** (2008).
13. Dey, T. K. & Wang, Y. S. Reeb graphs: Approximation and Persistence. *Discrete and Computational Geometry* **49**(1), 46–73 (2013).
14. Mischaikow, K. & Nanda, V. Morse Theory for Filtrations and Efficient Computation of Persistent Homology. *Discrete and Computational Geometry* **50**(2), 330–353 (2013).
15. 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* **11**, 499–527 (2011).
16. Horak, D., Maletic, S. & Rajkovic, M. Persistent Homology of Complex Networks. *Journal of Statistical Mechanics: Theory and Experiment* **2009**(03), P03034 (2009).
17. Lee, H., Kang, H., Chung, M. K., Kim, B. & Lee, D. S. Persistent Brain Network Homology from The Perspective of Dendrogram. *Medical Imaging, IEEE Transactions on* **31**(12), 2267–2277 (Dec 2012).
18. Silva, V. D. & Ghrist, R. Blind Swarms for Coverage in 2-d. In *Proceedings of Robotics: Science and Systems*, page 01 (2005).
19. Bendich, P., Edelsbrunner, H. & Kerber, M. Computing Robustness and Persistence for Images. *IEEE Transactions on Visualization and Computer Graphics* **16**, 1251–1260 (2010).
20. Carlsson, G., Ishkhanov, T., Silva, V. & Zomorodian, A. On The Local Behavior of Spaces of Natural Images. *International Journal of Computer Vision* **76**(1), 1–12 (2008).
21. Frosini, P. & Landi, C. Persistent Betti numbers for A Noise Tolerant Shape-Based Approach to Image Retrieval. *Pattern Recognition Letters* **34**(8), 863–872 (2013).
22. Pachauri, D., Hinrichs, C., Chung, M. K., Johnson, S. C. & Singh, V. Topology-Based Kernels with Application to Inference Problems in Alzheimer's Disease. *Medical Imaging, IEEE Transactions on* **30**(10), 1760–1770 (2011).
23. Singh, G. *et al*. Topological Analysis of Population Activity in Visual Cortex. *Journal of Vision* **8**(8) (2008).
24. Carlsson, G. Topology and Data. *Am. Math. Soc* **46**(2), 255–308 (2009).
25. Liu, X., Xie, Z. & Yi, D. Y. A Fast Algorithm for Constructing Topological Structure in Large Data. *Homology, Homotopy and Applications* **14**, 221–238 (2012).

26. Niyogi, P., Smale, S. & Weinberger, S. A Topological View of Unsupervised Learning from Noisy Data. *SIAM Journal on Computing* **40**, 646–663 (2011).
27. Rieck, B., Mara, H. & Leitte, H. Multivariate Data Analysis Using Persistence-Based Filtering and Topological signatures. *IEEE Transactions on Visualization and Computer Graphics* **18**, 2382–2391 (2012).
28. Wang, B., Summa, B., Pascucci, V. & Vejdemo-Johansson, M. Branching and Circular Features in High Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics* **17**, 1902–1911 (2011).
29. Mischaikow, K., Mrozek, M., Reiss, J. & Szymczak, A. Construction of Symbolic Dynamics from Experimental Time Series. *Physical Review Letters* **82**, 1144–1147 (1999).
30. Gameiro, M. *et al*. Topological Measurement of Protein Compressibility Via Persistence Diagrams. *preprint* (2013).
31. Kasson, P. M. *et al*. Persistent Voids A New Structural Metric for Membrane Fusion. *Bioinformatics* **23**, 1753–1759 (2007).
32. Wang, B. & Wei, G. W. Object-Oriented Persistent Homology. *Journal of Computational Physics* **305**, 276–299 (2016).
33. Xia, K. L., Feng, X., Tong, Y. Y. & Wei, G. W. Persistent Homology for The Quantitative Prediction of Fullerene Stability. *Journal of Computational Chemsitry* **36**, 408–422 (2015).
34. Xia, K. L. & Wei, G. W. Persistent Homology Analysis of Protein Structure, Flexibility and Folding. *International Journal for Numerical Methods in Biomedical Engineerings* **30**, 814–844 (2014).
35. Xia, K. L. & Wei, G. W. Multidimensional Persistence in Biomolecular Data. *Journal Computational Chemistry* **36**, 1502–1520 (2015).
36. Xia, K. L. & Wei, G. W. Persistent Topology for Cryo-EM Data Analysis. *International Journal for Numerical Methods in Biomedical Engineering* **31**, e02719 (2015).
37. Yao, Y. *et al*. Topological Methods for Exploring Low-Density States in Biomolecular Folding Pathways. *The Journal of Chemical Physics* **130**, 144115 (2009).
38. Hiraoka, Y. *et al*. Hierarchical Structures of Amorphous Solids Characterized by Persistent Homology. *Proceedings of the National Academy of Sciences* **113**(26), 7035–7040 (2016).
39. Saadatfar, M., Takeuchi, H., Robins, V., Francois, N. & Hiraoka, Y. Pore Configuration Landscape of Granular Crystallization. *Nature communications* **8**, 15082 (2017).
40. Tausz, A., Vejdemo-Johansson, M. & Adams, H. Javaplex: A Research Software Package for Persistent (co)Homology. Software available at http://code.google.com/p/javaplex (2011).
41. Nanda, V. Perseus: The Persistent Homology Software. Software Available at, http://www.sas.upenn.edu/~vnanda/perseus.
42. Bauer, U., Kerber, M. & Reininghaus, J. Distributed Computation of Persistent Homology. *Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments* (*ALENEX*) (2014).
43. Dionysus: The Persistent Homology Software. Software Available at http://www.mrzv.org/software/dionysus.
44. Binchi, J., Merelli, E., Rucco, M., Petri, G. & Vaccarino, F. Jholes: A Tool for Understanding Biological Complex Networks Via Clique Weight Rank Persistent Homology. *Electronic Notes in Theoretical Computer Science* **306**, 5–18 (2014).
45. Maria, C. Filtered Complexes. In *GUDHI User and Reference Manual* (GUDHI Editorial Board, 2015).
46. Bauer, U. Ripser: a lean C++ code for The Computation of Vietoris-Rips Persistence Barcodes. *Software available at* https://github.com/Ripser/ripser (2017).
47. Bauer, U., Kerber, M., Reininghaus, J. & Wagner, H. PHAT–Persistent Homology Algorithms Toolbox. In *International Congress on Mathematical Software*, pages 137–143 (Springer, 2014).
48. Bauer, U., Kerber, M. & Reininghaus, J. Distributed Computation of Persistent Homology. In *2014 proceedings of the sixteenth workshop on algorithm engineering and experiments* (*ALENEX*), pages 31–38 (SIAM, 2014).
49. Fasy, B. T., Kim, J., Lecci, F. & Maria, C. Introduction to The r Package tda. *arXiv preprint arXiv:1411.1830* (2014).
50. Ghrist, R. Barcodes: The Persistent Topology of Data. *Bulletin of the American Mathematical Society* **45**(1), 61–75 (2008).
51. Bubenik, P. Statistical Topological Data Analysis Using Persistence Landscapes. *The Journal of Machine Learning Research* **16**(1), 77–102 (2015).
52. Bubenik, P. & Kim, P. T. A Statistical Approach to Persistent Homology. *Homology, Homotopy and Applications* **19**, 337–362 (2007).
53. Adams, H. *et al*. Persistence Images: A Stable Vector Representation of Persistent Homology. *The Journal of Machine Learning Research* **18**(1), 218–252 (2017).
54. Chung, Y. M., Hu, C. S., Lawson, A. & Smyth, C. TopoResNet: A Hybrid Deep Learning Architecture and its Application to Skin Lesion Classification (2019).
55. Chung, Y. M. & Lawson, A. Persistence curves: A Canonical Framework for Summarizing Persistence Diagrams. (2019).
56. Bell, G., Lawson, A., Martin, J., Rudzinski, J. & Smyth, C. Weighted Persistent Homology. *arXiv preprint arXiv:1709.00097* (2017).
57. Buchet, M., Chazal, F., Oudot, S. Y. & Sheehy, D. R. Efficient and Robust Persistent Homology For Measures. *Computational Geometry* **58**, 70–96 (2016).
58. Edelsbrunner, H. *Weighted Alpha Shapes*, volume 92 (University of Illinois at Urbana-Champaign, Department of Computer Science, 1992).
59. Guibas, L., Morozov, D. & Mérigot, Q. Witnessed k-Distance. *Discrete & Computational Geometry* **49**(1), 22–45 (2013).
60. Xia, K. L., Zhao, Z. X. & Wei, G. W. Multiresolution Persistent Homology for Excessively Large Biomolecular Datasets. *The Journal of chemical physics* **143**(13), 10B603_1 (2015).
61. Petri, G., Scolamiero, M., Donato, I. & Vaccarino, F. Topological Strata of Weighted Complex Networks. *PloS one* **8**(6), e66506 (2013).
62. Xia, K. L. & Wei, G. W. Persistent Homology Analysis of Protein Structure, Flexibility, And Folding. *International journal for numerical methods in biomedical engineering* **30**(8), 814–844 (2014).
63. Dawson, R. J. M. Homology of Weighted Simplicial Complexes. *Cahiers de Topologie et Géométrie Différentielle Catégoriques* **31**(3), 229–243 (1990).
64. Ren, S. Q., Wu, C. Y. & Wu, J. Weighted Persistent Homology. *Rocky Mountain Journal of Mathematics* **48**(8), 2661–2687 (2018).
65. Wu, C. Y., Ren, S. Q., Wu, J. & Xia, K. L. Weighted (co) Homology and Weighted Laplacian. *arXiv preprint arXiv:1804.06990* (2018).
66. Xia, K. L., Zhao, Z. X. & Wei, G. W. Multiresolution topological Simplification. *Journal Computational Biology* **22**, 1–5 (2015).
67. Ahmed, M., Fasy, B. T. & Wenk, C. Local Persistent Homology Based Distance Between Maps. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52 (ACM, 2014).
68. Bendich, P., Cohen-Steiner, D., Edelsbrunner, H., Harer, J. & Morozov, D. Inferring Local Homology from Sampled Stratified Spaces. In *Foundations of Computer Science*, *2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 536–546 (IEEE, 2007).
69. Bendich, P., Gasparovic, E., Harer, J., Izmailov, R. & Ness, L. Multi-Scale Local Shape Analysis and Feature Selection in Machine Learning Applications. In *Neural Networks* (*IJCNN*), *2015 International Joint Conference on*, pages 1–8 (IEEE, 2015).
70. Bendich, P., Wang, B. & Mukherjee, S. Local Homology Transfer and Stratification Learning. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1355–1370 (SIAM, 2012).
71. Fasy, B. T. & Wang, B. Exploring Persistent Local Homology in Topological Data Analysis. In *Acoustics*, *Speech and Signal Processing* (*ICASSP*), *2016 IEEE International Conference on*, pages 6430–6434 (IEEE, 2016).
72. Munkres, J. R. *Elements of algebraic topology* (CRC Press, 2018).
73. Chintakunta, H., Gentimis, T., Gonzalez-Diaz, R., Jimenez, M. J. & Krim, H. An Entropy-Based Persistence Barcode. *Pattern Recognition* **48**(2), 391–401 (2015).
74. Lu, X. J. & Olson, W. K. 3DNA: A Software Package for The Analysis, Rebuilding and Visualization Of Three-Dimensional Nucleic Acid Structures. *Nucleic acids research* **31**(17), 5108–5121 (2003).

75. Lu, X. J., El Hassan, M. A. & Hunter, C. A. Structure and Conformation of Helical Nucleic Acids: Analysis Program (SCHNAaP). *Journal of molecular biology* **273**(3), 668–680 (1997).
76. Pun, C. S., Xia, K. L. & Lee, S. X. Persistent-Homology-Based Machine Learning and its Applications–A Survey. *arXiv preprint arXiv:1811.00252* (2018).
77. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of chemical theory and computation* **4**(3), 435–447 (2008).
78. Meng, Z. Y., Kubar, T., Mu, Y. G. & Shao, F. W. A Molecular Dynamics-Quantum Mechanics Theoretical Study of DNA-Mediated Charge Transport in Hydrated Ionic Liquids. *Journal of chemical theory and computation* **14**(5), 2733–2742 (2018).

Acknowledgements

This work was supported in part by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG31/18, Tier 2 MOE2018-T2-1-033.

Author contributions

K.X. and Z.M. contributed the algorithm design. K.X. wrote the main manuscript text. V.D.A., Y.L. and J.W. reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55660-3.

Correspondence and requests for materials should be addressed to K.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.