scientific reports

Check for updates

OPEN

Persistent Dirac for molecular representation

Junjie Wee[1]✉, Ginestra Bianconi[2,3] & Kelin Xia[1]

Molecular representations are of fundamental importance for the modeling and analysing molecular systems. The successes in drug design and materials discovery have been greatly contributed by molecular representation models. In this paper, we present a computational framework for molecular representation that is mathematically rigorous and based on the persistent Dirac operator. The properties of the discrete weighted and unweighted Dirac matrix are systematically discussed, and the biological meanings of both homological and non-homological eigenvectors are studied. We also evaluate the impact of various weighting schemes on the weighted Dirac matrix. Additionally, a set of physical persistent attributes that characterize the persistence and variation of spectrum properties of Dirac matrices during a filtration process is proposed to be molecular fingerprints. Our persistent attributes are used to classify molecular configurations of nine different types of organic-inorganic halide perovskites. The combination of persistent attributes with gradient boosting tree model has achieved great success in molecular solvation free energy prediction. The results show that our model is effective in characterizing the molecular structures, demonstrating the power of our molecular representation and featurization approach.

Molecular representation and featurization play an essential role in physical as well as in data-driven learning models. The relationship between the structure and function of molecules is complex, and a comprehensive understanding of the structural properties is crucial to extract functional information. To establish explicit linear or nonlinear relationships between molecular structure and function, various quantitative structure-activity/property relationship (QSAR/QSPR) models have been developed[1,2]. Different molecular fingerprints have also been proposed for machine learning and deep learning models to predict molecular functions and properties[3–8]. Despite significant advances, the development of highly efficient descriptors remains a major challenge for QSAR/QSPR and learning models in the analysis of molecular data in the fields of materials, chemistry, and biology[1,2].

Graph models[9–18] are arguably the most widely used tools for molecular representations in molecular dynamics simulation, coarse-grained models, elastic network models, QSAR/QSPR, graph neural networks, etc. In general, a molecule (or a molecular complex) is modeled as a graph with each vertex representing an atom, an amino acid, a domain, or an entire molecule, and edge representing covalent-bond, non-covalent-bond, or more general interaction. However, graphs are designed for the characterization of pairwise interactions. To capture higher-order interactions, topological representations, such as multilayer networks[19], simplicial complexes[20–22], hypergraphs[23,24], etc, should be considered. Among them, multilayer networks have been used in the characterization of higher-order dynamics[25–28] and synchronization dynamics[29–31]. As a generalization of graphs, simplicial complexes are made not only of 0-simplices (nodes) and 1-simplices (edges), but also of higher-dimensional simplices, such as 2-simplices (triangles), 3-simplices (tetrahedron), etc. Note that higher-order networks and simplicial complexes can describe the many-body interactions beyond pairwise interactions. Hypergraphs are a further generalization of simplicial complexes. An hypergraph is composed of hyperedges, which are formed by a set of vertices. Recently, simplicial complexes and hypergraphs have been used in molecular representations and have allowed improved performance of drug design algorithms, in particular, in the protein-ligand binding affinity prediction.

Based on topological representations, molecular descriptors or fingerprints can be generated and further used as features for learning models. The recent emergence of topological data analysis (TDA)[32,33] and combinatorial Hodge theory-based molecular descriptors has had a significant impact on drug design. These models have been successful in various stages of drug design, such as predicting protein-ligand binding affinity[4,14,34–39], protein stability changes resulting from mutations[40,41], toxicity[42], solvation free energy[43,44], partition coefficient and aqueous solubility[45], and identifying binding pockets[46]. In comparison to traditional molecular representations, these models have demonstrated superior performance in the D3R Grand Challenge[47,48]. TDA's fundamental

[1]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore. [2]School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK. [3]The Alan Turing Institute, London NW1 2DB, UK. ✉email: weej0019@e.ntu.edu.sg

mathematical concept involves using persistent homology to extract topological information, tracking the change of homology generators of simplicial complexes over a filtration process. In particular, the topological invariant Betti numbers can be obtained from the kernel of combinatorial Hodge Laplacian (HL) matrix. Interestingly, the Forman Ricci curvature can be obtained via the Bochner–Weitzenböck decomposition of HL matrix[6]. The great success of TDA and combinatorial Hodge theory based molecular descriptors in learning models is due to their characterization of structures with intrinsic invariants, including Betti numbers and Ricci curvatures. These intrinsic descriptors are well defined mathematical observables that characterize fundamental topological and geometrical properties of real datasets, thus they have an excellent transferability for learning models.

Inspired by the success of Hodge Laplacian matrix in molecular sciences, here we propose the persistent Dirac based molecular representation and fingerprint. The discrete Dirac operator[49–55] is a first-order differential operator which can be interpreted as the square root of Hodge Laplace operator. This operator has been developed on graphs and simplicial complexes and used in TDA and for investigating dynamics of topological signals[50,56–58]. Moreover, the persistent Dirac model can be used in the quantum algorithm of persistent homology[52,53,59]. Here we present a rigorous mathematical theory for persistent Dirac through the commutative diagram of discrete Dirac operator over a filtration process. The commutative diagram is similar to the ones in persistent spectral graph[5,60], persistent Hodge Laplacian[61], and persistent sheaf Laplacian[61,62]. Further, we develop a series of persistent attributes from persistent Dirac, and use them as descriptors to characterize molecular structures.

Our work starts with a systematic study of the spectrum of the discrete Dirac matrix. In particular, we identify the geometric and topological properties of both non-homology and homology eigenvectors for molecular structures. We generalize these results to weighted simplicial complexes on top of which the weighted Dirac operator[63] is carefully defined. In particular, we analyse the influence of weighting schemes on the spectral properties of molecular structures. The persistent Dirac is then introduced and is employed for the clustering of molecular configurations from the molecular dynamic simulations of nine types of organic-inorganic halide perovskites (OIHP). By the comparison with several existing models, we show that our model is highly efficient in clustering the structure configurations. Further, the combination of persistent attributes with gradient boosting tree model has achieved great success in molecular solvation free energy prediction. This demonstrates the great potential of our persistent Dirac-based fingerprints in molecular representation and featurization.

The paper is organized as follows. Section "Methods" is devoted to the discrete Dirac models. It covers basic mathematical background such as simplicial complexes, chain groups, boundary operators, Hodge Laplacian. Thereafter, the section discusses the use of spectrum of discrete Dirac models for biomolecular representation and characterization. In section "Results", persistent Dirac model is present. The eigenspectrum information for (weighted) Dirac matrix and persistent attributes from persistent Dirac are discussed in detailed. The section ends with an application of the persistent Dirac based fingerprints on organic-inorganic halide perovskite (OIHP) classification and prediction of solvation free energy. The paper ends with a conclusion.

Methods

In this section, we discuss the discrete Dirac models, including discrete Dirac matrices and weighted Dirac matrices for biomolecular structure representation and characterization. Different from previous graph-based models, molecular structures are represented based on simplicial complexes, and algebraic tools from chain groups, homology groups, boundary operators and Dirac matrices, are used to reveal deeper geometric and topological properties.

Mathematical background for discrete Dirac models. *Simplicial complex.* Generally speaking, a simplicial complex can be viewed as a higher-dimensional generalization of graphs. A p-dimensional simplicial complex is formed by simplices of dimension up to p. Every p dimensional simplex consists of a set of $p+1$ vertices and this set can be viewed geometrically as a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex), etc.

Here and in the following we indicate with n_p the number of p-simplices belonging to the simplicial complex $\mathcal{K}$. The most commonly used simplicial complexes include Čech complex, Vietoris–Rips complex, Alpha complex, Cubical complex, Morse complex, etc.[64].

Two p-dimensional simplices σ_1 and σ_2 in a simplicial complex $\mathcal{K}$, are simplex neighbors if

(i) σ_1 and σ_2 share a $(p+1)$-simplex μ, that is, there exists a μ in $\mathcal{K}$ such that $\mu > \sigma_1$ and $\mu > \sigma_2$.
(ii) σ_1 and σ_2 share a $(p-1)$-simplex γ, that is, there exists a γ in $\mathcal{K}$ such that $\gamma < \sigma_1$ and $\gamma < \sigma_2$.

If either condition is satisfied, but both conditions do not hold at the same time, σ_1 and σ_2 are called parallel simplex neighbors. Here σ_1 and σ_2 are called upper adjacent neighbors and denoted as $\sigma_1 \frown \sigma_2$, if they satisfy condition (i). They are lower adjacent neighbors and denoted as $\sigma_1 \smile \sigma_2$ if they satisfy condition (ii).

Homology. In homology, a p-dimensional oriented simplex σ^p is the set of ordered $p+1$ nodes $[v_0, v_1, \ldots, v_p]$. For example, an oriented 1-simplex $\sigma^1 = [v_0, v_1]$ has the opposite sign of the oriented 1-simplex $[v_1, v_0]$. In other words,

$$[v_i, v_j] = -[v_j, v_i].$$

Similarly, this orientation can be written for higher-order simplices in the following way,

$$[v_0, v_1, \ldots, v_p] = (-1)^{\alpha(\pi)}[v_{\pi(0)}, v_{\pi(1)}, \ldots, v_{\pi(p)}],$$

where $\alpha(\pi)$ refers to the parity of the permutation π. In this paper, we consider the orientation induced by node labels, i.e. for every simplex in a simplicial complex, we assign a positive orientation to the one provided by the increasing set of node labels.

For an oriented simplicial complex $\mathcal{K}$, its p-dimensional chain group $C_p(\mathcal{K})$ is composed by linear combination of positively oriented p-simplices in $\mathcal{K}$. Let $[v_0, v_1, \ldots v_p]$ indicate the generic positively oriented p-simplex $\sigma^p \in \mathcal{K}$. We notice that the set of simplices σ_p constitute a basis for the p-dimensional chains $C_p(\mathcal{K})$. Therefore any p-chain $f_1 \in C_p(\mathcal{K})$ can be written in a unique way as

$$f_1 = \sum_{i=1}^{n_p} c_i \sigma^i. \tag{1}$$

The weighted boundary operator $\overline{\partial}_p : C_p \to C_{p-1}$ can be determined by its action on any given $\sigma^p \in \mathcal{K}$:

$$\overline{\partial}_p(\sigma^p) = a_p \sum_{i=0}^{p} (-1)^i [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_p].$$

Here a_p is a constant in $\mathbb{R}^+$ dependent on p and the boundary of p-simplex is made of $(p-1)$-simplices $[v_0, v_1, \ldots, \hat{v}_i, \ldots, v_p]$, where $\hat{v}_i$ means that v_i has been removed from the sequence $v_0, \ldots, v_p$. It is also well-known that $\overline{\partial}_{p-1}\overline{\partial}_p = 0$. The unweighted boundary operator can be obtained by setting $a_p = 1$. In other words, the unweighted boundary operator $\partial_p : C_p \to C_{p-1}$ for a given $\sigma^p \in \mathcal{K}$ is defined as

$$\partial_p(\sigma^p) = \sum_{i=0}^{p} (-1)^i [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_p].$$

For an oriented simplicial complex $\mathcal{K}$, its two oriented p-dimensional simplices σ_1 and σ_2 are similarly oriented and denoted as $\sigma_1 \sim \sigma_2$, if they are lower adjacent and have the same sign on the common lower $(p-1)$-simplex. Two simplex σ_1 and σ_2 are dissimilarly oriented and denoted as $\sigma_1 \nsim \sigma_2$, if they are lower adjacent but have different signs on the common lower $(p-1)$-simplex.

The p-th cycle group Z_p is defined as,

$$Z_p = \ker(\overline{\partial}_p) = \{c \in C_p | \overline{\partial}_p(c) = 0\},$$

and p-th boundary group B_p is,

$$B_p = \mathrm{im}(\overline{\partial}_{p+1}) = \{c \in C_p | \exists d \in C_{p+1} : c = \overline{\partial}_{p+1}(d)\}.$$

The p-th homology group is defined as $H_p = Z_p / B_p$. Its rank is p-th Betti number that satisfies

$$\beta_p = \mathrm{rank}\, H_p = \mathrm{rank}\, Z_p - \mathrm{rank}\, B_p.$$

With the boundary operators, we have chain complexes

$$\cdots \xrightarrow{\overline{\partial}_{p+2}} C_{p+1} \xrightarrow{\overline{\partial}_{p+1}} C_p \xrightarrow{\overline{\partial}_p} C_{p-1} \xrightarrow{\overline{\partial}_{p-1}} \cdots$$

The adjoint of $\overline{\partial}_p$, which is

$$\overline{\partial}_p^* : C_{p-1} \to C_p,$$

satisfies the inner product relation $\langle \overline{\partial}_p(f), g \rangle = \langle f, \overline{\partial}_p^*(g) \rangle$, for every $f \in C_p$, $g \in C_{p-1}$. It is used in the weighted Hodge Laplacian.

Weighted Hodge Laplacian and Hodge decomposition. The p-dimensional weighted Hodge Laplacian $\Delta_p : C_p \to C_p$ is defined as follows:

$$\Delta_p = \begin{cases} \overline{\partial}_1 \circ \overline{\partial}_1^*, & \text{if } p = 0. \\ \overline{\partial}_p^* \circ \overline{\partial}_p + \overline{\partial}_{p+1} \circ \overline{\partial}_{p+1}^*, & \text{if } p \geq 1. \end{cases}$$

The special case where $p = 0$ is the well-known graph Laplacian.

Computationally, the information for weighted boundary operators acting from finite dimensional chain groups C_p to C_{p-1} can be stored efficiently in matrix representations. As matrix representations, the weighted boundary operators and its adjoint satisfies $\overline{\partial}_p^\top = \overline{\partial}_p^*$.

More specifically, let n_{p-1} and n_p be the number of $(p-1)$-simplices and p-simplices respectively in a simplicial complex $\mathcal{K}$. The $n_{p-1} \times n_p$ weighted boundary matrix $\overline{\mathbf{B}}_p$ has entries defined as follows:

$$\overline{\mathbf{B}}_p(i,j) = \begin{cases} a_p, & \text{if } \sigma_i^{p-1} < \sigma_j^p, \sigma_i^{p-1} \sim \sigma_j^p. \\ -a_p, & \text{if } \sigma_i^{p-1} < \sigma_j^p, \sigma_i^{p-1} \nsim \sigma_j^p. \\ 0, & \text{if } \sigma_i^{p-1} \not< \sigma_j^p. \end{cases}$$

where $1 \le i \le n_{p-1}$ and $1 \le j \le n_p$. Here, $\sigma_i^{p-1} < \sigma_j^p$ represents the i-th $(p-1)$-simplex σ_i^{p-1} is a face of j-th p-simplex σ_j^p and $\sigma_i^{p-1} \sim \sigma_j^p$ indicates the coefficient of σ_i^{p-1} in $\overline{\partial}_p(\sigma_j^p)$ is a_p. Likewise, $\sigma_i^{p-1} \not< \sigma_j^p$ means that σ_i^{p-1} is not a face of σ_j^p and $\sigma_i^{p-1} \nsim \sigma_j^p$ indicates that the coefficient of σ_i^{p-1} in $\overline{\partial}_p(\sigma_j^p)$ is $-a_p$.

Since the unweighted boundary operator $\partial_p = \frac{1}{a_p}\overline{\partial}_p$, note that an unweighted boundary matrix can be similarly written as

$$\mathbf{B}_p = \frac{1}{a_p}\overline{\mathbf{B}}_p. \tag{2}$$

Using the weighted boundary matrices, the lower and upper weighted Hodge Laplacians can be defined as $\overline{\mathbf{L}}_p^{\text{down}} = \overline{\mathbf{B}}_p^\top \overline{\mathbf{B}}_p$ and $\overline{\mathbf{L}}_p^{\text{up}} = \overline{\mathbf{B}}_{p+1}\overline{\mathbf{B}}_{p+1}^\top$ respectively. More specifically, the entries of $\overline{\mathbf{L}}_p^{\text{down}}$ $(p > 0)$ are as follows,

$$\overline{\mathbf{L}}_p^{\text{down}}(i,j) = \begin{cases} a_p^2(p+1), & i = j. \\ a_p^2, & i \ne j, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \sim \sigma_j^p. \\ -a_p^2, & i \ne j, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \nsim \sigma_j^p. \\ 0, & i \ne j \text{ and } \sigma_i^p \not\sim \sigma_j^p. \end{cases}$$

Note that all the entries of $\overline{\mathbf{L}}_0^{\text{down}}$ are zero since 0-simplices have no lower adjacent neighbors. Further, $\sigma_i^p \smile \sigma_j^p$ refers to σ_i^p and σ_j^p being lower adjacent neighbors while $\sigma_i^p \frown \sigma_j^p$ refers to σ_i^p and σ_j^p being upper adjacent neighbors.

It is important to observe that $\sigma_i^p \frown \sigma_j^p$ $(p > 0)$ also implies that σ_i^p and σ_j^p share a lower simplex σ^{p-1}. The case where $\sigma_i^p \sim \sigma_j^p$ refers to σ_i^p and σ_j^p sharing a common similar lower simplex σ^{p-1}. This means that the signs of coefficient of σ^{p-1} in $\overline{\partial}_p(\sigma_i^p)$ and $\overline{\partial}_p(\sigma_j^p)$ are the same. On the other hand, $\sigma_i^p \nsim \sigma_j^p$ refers to σ_i^p and σ_j^p sharing a common dissimilar lower simplex σ^{p-1}. This can be verified by checking the signs of coefficient of σ^{p-1} in $\overline{\partial}_p(\sigma_i^p)$ and $\overline{\partial}_p(\sigma_j^p)$ to be not the same.

Since 0-simplices have no lower adjacent neighbors, any two 0-simplices σ_i^0 and σ_j^0 that are upper adjacent neighbors will always satisfy $\sigma_i^0 \sim \sigma_j^0$ vacuously.

Hence, the matrix elements of the Hodge Laplacian $\overline{\mathbf{L}}_p^{\text{up}}$ are given by

$$\overline{\mathbf{L}}_p^{\text{up}}(i,j) = \begin{cases} a_{p+1}^2 d(\sigma_i^p), & i = j. \\ -a_{p+1}^2, & i \ne j, \sigma_i^p \frown \sigma_j^p, \sigma_i^p \sim \sigma_j^p. \\ a_{p+1}^2, & i \ne j, \sigma_i^p \frown \sigma_j^p, \sigma_i^p \nsim \sigma_j^p. \\ 0, & i \ne j \text{ and } \sigma_i^p \not\frown \sigma_j^p. \end{cases}$$

Here $d(\sigma_i^p)$ denotes the number of cofaces with dimension $p+1$ of simplex σ_i^p.

The p^{th} weighted combinatorial Laplacian $\overline{\mathbf{L}}_p$ is defined as $\overline{\mathbf{L}}_p = \overline{\mathbf{B}}_p^\top \overline{\mathbf{B}}_p + \overline{\mathbf{B}}_{p+1}\overline{\mathbf{B}}_{p+1}^\top$. Note that $\overline{\mathbf{L}}_0 = \overline{\mathbf{B}}_1\overline{\mathbf{B}}_1^\top$. The matrix elements of the Hodge Laplacians $\overline{\mathbf{L}}_p$ with $p = 0$ are given by

$$\overline{\mathbf{L}}_0(i,j) = \begin{cases} a_1^2 d(\sigma_i^0), & i = j. \\ -a_1^2, & i \ne j, \sigma_i^0 \frown \sigma_j^0. \\ 0, & i \ne j, \sigma_i^0 \not\frown \sigma_j^0. \end{cases}$$

while the matrix elements for $p > 0$ can be expressed as

$$\overline{\mathbf{L}}_p(i,j) = \begin{cases} a_{p+1}^2 d(\sigma_i^p) + a_p^2(p+1), & i = j. \\ a_p^2 - a_{p+1}^2, & i \ne j, \sigma_i^p \frown \sigma_j^p, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \sim \sigma_j^p. \\ a_{p+1}^2 - a_p^2, & i \ne j, \sigma_i^p \frown \sigma_j^p, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \nsim \sigma_j^p. \\ a_p^2, & i \ne j, \sigma_i^p \not\frown \sigma_j^p, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \sim \sigma_j^p. \\ -a_p^2, & i \ne j, \sigma_i^p \not\frown \sigma_j^p, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \nsim \sigma_j^p. \\ 0, & i \ne j \text{ and } \sigma_i^p \not\sim \sigma_j^p. \end{cases}$$

It follows from Eq. (2) that the lower and upper unweighted Hodge Laplacians can be written as $\mathbf{L}_p^{\text{down}} = \mathbf{B}_p^\top \mathbf{B}_p$ and $\mathbf{L}_p^{\text{up}} = \mathbf{B}_{p+1}\mathbf{B}_{p+1}^\top$ respectively. Hence, the pth unweighted combinatorial Laplacian $\mathbf{L}_p = \mathbf{L}_p^{\text{down}} + \mathbf{L}_p^{\text{up}}$ have elements given by

$$\mathbf{L}_0(i,j) = \begin{cases} d(\sigma_i^0), & i = j. \\ -1, & i \neq j, \sigma_i^0 \frown \sigma_j^0. \\ 0, & i \neq j, \sigma_i^0 \not\frown \sigma_j^0. \end{cases}$$

for $p = 0$ while for $p > 0$ the matrix elements of the Hodge Laplacian are given by

$$\mathbf{L}_p(i,j) = \begin{cases} d(\sigma_i^p) + p + 1, & i = j. \\ 1, & i \neq j, \sigma_i^p \not\sim \sigma_j^p, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \sim \sigma_j^p. \\ -1, & i \neq j, \sigma_i^p \not\sim \sigma_j^p, \sigma_i^p \smile \sigma_j^p, \sigma_i^p \nsim \sigma_j^p. \\ 0, & i \neq j \text{ and either } \sigma_i^p \frown \sigma_j^p \text{ or } \sigma_i^p \not\sim \sigma_j^p. \end{cases}$$

It is well-known that λ is a non-zero eigenvalue of $\overline{\mathbf{L}}_p$ if and only if λ is an non-zero eigenvalue of $\overline{\mathbf{L}}_p^{\text{down}}$ or $\overline{\mathbf{L}}_p^{\text{up}}$. The multiplicity of the zero eigenvalues of $\overline{\mathbf{L}}_p$ corresponds to the pth Betti number as follows,

$$\dim \ker \overline{\mathbf{L}}_p = \beta_p = \dim \ker \overline{\mathbf{L}}_p^{\text{down}} - \dim \text{im} \overline{\mathbf{L}}_p^{\text{up}}$$

where β_p is also the rank H_p[65] (see Appendix C).

Further, $\dim \ker \overline{\mathbf{L}}_p^{\text{down}}$ can be written as:

$$\begin{aligned} \dim \ker \overline{\mathbf{L}}_p^{\text{down}} &= \beta_p + \dim \text{im} \overline{\mathbf{L}}_p^{\text{up}} \\ &= \beta_p + \dim C_p - \dim \ker \overline{\mathbf{L}}_p^{\text{up}} \\ &= \beta_p + \dim C_p - \dim \ker \overline{\mathbf{B}}_{p+1}^{\top} \\ &= \beta_p + \text{rank } \overline{\mathbf{B}}_{p+1}^{\top}. \end{aligned} \tag{3}$$

The above Eq. (3) will be an important relation for persistent Dirac models in later sections.

Closely related to the Hodge Laplacian is the Hodge decomposition. The Hodge decomposition is an orthogonal decomposition of a vector field into gradient part, harmonic part and curl part. More formally, the Hodge decomposition states that a p-th chain group C_p of a simplicial complex $\mathcal{K}$ admits the following orthogonal direct sum decomposition:

$$C_p = \overbrace{\underbrace{\text{im}(\overline{\mathbf{B}}_{p+1})}_{\text{im } \overline{\mathbf{L}}_p^{\text{down}}} \oplus \underbrace{\ker(\overline{\mathbf{L}}_p)}_{\ker \overline{\mathbf{L}}_p^{\text{down}} = \ker \overline{\mathbf{B}}_p}}^{\ker \overline{\mathbf{L}}_p^{\text{up}} = \ker \overline{\mathbf{B}}_{p+1}^{\top}} \oplus \overbrace{\text{im}(\overline{\mathbf{B}}_p^{\top})}^{\text{im } \overline{\mathbf{L}}_p^{\text{up}}},$$

where $\ker(\overline{\mathbf{L}}_p) = \ker \overline{\mathbf{B}}_p \cap \ker \overline{\mathbf{B}}_{p+1}^{\top}$.

It is worth mentioning that such flows have also been extended to five component decompositions with edge and face vector fields[66], applied to the protein B-factor prediction problems via Hodge theory[8] and also in de Rham–Hodge biomolecular data analysis[9,67].

Discrete Dirac models. *Weighted Dirac matrix.* Recently, weighted Dirac matrices have been proposed based on a weighted simplicial complex[63]. For a d-dimensional weighted simplicial complex $\mathcal{K}$, let us define the $n_p \times n_p$ metric matrix $\mathbf{G}_p$ $(0 \leq p \leq d)$ to be a diagonal matrix with positive entries. For any two p-chains $f_1 = \sum_{i=1}^{n_p} c_i \sigma^i$ and $f_2 = \sum_{i=0}^{n_p} d_i \sigma^i$ in C_p, the matrix $\mathbf{G}_p$ can be used to define the weighted inner product

$$\langle f_1, f_2 \rangle = \sum_{i=1}^{n_p} \mathbf{G}_p(\sigma^i, \sigma^i) c_i d_i = (\mathbf{f}_1)^{\top} \mathbf{G}_p(\mathbf{f}_2), \tag{4}$$

where $(\mathbf{f}_1)^{\top} = [c_1, c_2, \ldots, c_p]$ and $(\mathbf{f}_2)^{\top} = [d_1, d_2, \ldots, d_p]$.

Recall from Eq. (2) that the weighted boundary operator can be represented by a matrix $\overline{\mathbf{B}}_p = a_p \mathbf{B}_p$ where $\mathbf{B}_p$ is the unweighted boundary matrix. If $a_p = 1$, then $\overline{\mathbf{B}}_p$ reduces to the adjoint operator of $\mathbf{B}_p$. Formally, for any p-chain f and any $(p-1)$-chain g, the adjoint operator $\overline{\mathbf{B}}_p^*$ satisfies

$$\langle f, \overline{\mathbf{B}}_p^* g \rangle = \langle \overline{\mathbf{B}}_p f, g \rangle. \tag{5}$$

From the inner product relation (5), an explicit expression of $\overline{\mathbf{B}}_p^*$ can be deduced in terms of $\overline{\mathbf{B}}_p$ and the matrices $\mathbf{G}_p$. Based on the weighted inner product definition (4), this gives

$$(\mathbf{f})^\top \mathbf{G}_p \overline{\mathbf{B}}_p^* (\mathbf{g}) = (\mathbf{f})^\top \overline{\mathbf{B}}_p^\top \mathbf{G}_{p-1}(\mathbf{g}).$$

Since the expression is true for any arbitrary $\mathbf{f}$ and $\mathbf{g}$, this implies

$$\mathbf{G}_p \overline{\mathbf{B}}_p^* = \overline{\mathbf{B}}_p^\top \mathbf{G}_{p-1}.$$

Hence, the following becomes an explicit expression for the adjoint operator $\overline{\mathbf{B}}_p^*$:

$$\overline{\mathbf{B}}_p^* = \mathbf{G}_p^{-1} \overline{\mathbf{B}}_p^\top \mathbf{G}_{p-1}. \tag{6}$$

Here $\overline{\mathbf{B}}_p^*$[65] is the adjoint of the weighted boundary operator[10,68]. It is important to note that if the metric matrices $\mathbf{G}_p$ are the identity matrices, the above expression then reduces to the transpose of the boundary operator multiplied by the constant a_p,

$$\overline{\mathbf{B}}_p^* = \overline{\mathbf{B}}_p^\top = a_p \mathbf{B}_p^\top. \tag{7}$$

This also means the transpose of adjoint operator, i.e. $(\overline{\mathbf{B}}_p^*)^\top$, is equal to $\overline{\mathbf{B}}_p$ only if $\mathbf{G}_p$ are identity matrices. To see this, apply the transpose to both sides of Eq. (6) and obtain the expression

$$(\overline{\mathbf{B}}_p^*)^\top = \mathbf{G}_{p-1}^{-1} \overline{\mathbf{B}}_p \mathbf{G}_p = a_p \mathbf{G}_{p-1}^{-1} \mathbf{B}_p \mathbf{G}_p. \tag{8}$$

The matrices $(\overline{\mathbf{B}}_p^*)^\top$ and $\overline{\mathbf{B}}_p^*$ can then be used to construct the following weighted Dirac matrix (9).

For a simplicial complex $\mathcal{K}$ with $n_p \times n_{p-1}$ adjoint operators $\overline{\mathbf{B}}_p^*$ where n_{p-1} is the number of $(p-1)$-simplices and n_p is the number of p-simplices in $\mathcal{K}$, the weighted Dirac matrix $\overline{\mathbf{D}}_p$ is

$$\overline{\mathbf{D}}_p = \begin{bmatrix} 0_{\,n_0\times n_0} & (\overline{\mathbf{B}}_1^*)^\top & 0_{\,n_0\times n_2} & \cdots & 0_{\,n_0\times n_p} & 0_{\,n_0\times n_{p+1}} \\ \overline{\mathbf{B}}_1^* & 0_{\,n_1\times n_1} & (\overline{\mathbf{B}}_2^*)^\top & \cdots & 0_{\,n_1\times n_p} & 0_{\,n_1\times n_{p+1}} \\ 0_{\,n_2\times n_0} & \overline{\mathbf{B}}_2^* & 0_{\,n_2\times n_2} & \cdots & 0_{\,n_2\times n_p} & 0_{\,n_2\times n_{p+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{\,n_p\times n_0} & 0_{\,n_p\times n_1} & 0_{\,n_p\times n_2} & \cdots & 0_{\,n_p\times n_p} & (\overline{\mathbf{B}}_{p+1}^*)^\top \\ 0_{\,n_{p+1}\times n_0} & 0_{\,n_{p+1}\times n_1} & 0_{\,n_{p+1}\times n_2} & \cdots & \overline{\mathbf{B}}_{p+1}^* & 0_{\,n_{p+1}\times n_{p+1}} \end{bmatrix}. \tag{9}$$

In particular, we set $a_p = (p+1)^{-1/2}$ for all p up to the order of the simplicial complex and consider the matrices $\overline{\mathbf{B}}_p^*$ and $(\overline{\mathbf{B}}_p^*)^\top$ in Eq. (7) and (8) respectively. For $p = 2$, the weighted Dirac matrix from (9) becomes

$$\overline{\mathbf{D}}_1 = \begin{bmatrix} 0_{\,n_0\times n_0} & \mathbf{G}_0^{-1}\mathbf{B}_1\mathbf{G}_1/\sqrt{2} & 0_{\,n_0\times n_2} & 0_{\,n_0\times n_3} \\ \mathbf{B}_1^\top/\sqrt{2} & 0_{\,n_1\times n_1} & \mathbf{G}_1^{-1}\mathbf{B}_2\mathbf{G}_2/\sqrt{3} & 0_{\,n_1\times n_3} \\ 0_{\,n_2\times n_0} & \mathbf{B}_2^\top/\sqrt{3} & 0_{\,n_2\times n_2} & \mathbf{G}_2^{-1}\mathbf{B}_3\mathbf{G}_3/2 \\ 0_{\,n_3\times n_0} & 0_{\,n_3\times n_1} & \mathbf{B}_3^\top/2 & 0_{\,n_3\times n_3} \end{bmatrix}$$

This definition can be extended easily to higher dimensions. Note that $\mathbf{B}_p^\top/\sqrt{p+1}$ is the adjoint operator $\overline{\mathbf{B}}_p^*$ and $\mathbf{G}_{p-1}^{-1}\mathbf{B}_p\mathbf{G}_p/\sqrt{p+1}$ is equal to the transpose of $\overline{\mathbf{B}}_p^*$. Note that this definition of weighted Dirac is self-adjoint and with eigenvalues smaller than or equal to one. The square of the weighted Dirac also forms a diagonal block of metric Hodge Laplacian matrices

$$\overline{\mathbf{D}}_2^2 = \begin{bmatrix} \mathbf{L}_{[0]} & 0_{\,n_0\times n_1} & 0_{\,n_0\times n_2} & 0_{\,n_0\times n_3} \\ 0_{\,n_1\times n_0} & \mathbf{L}_{[1]} & 0_{\,n_1\times n_2} & 0_{\,n_1\times n_3} \\ 0_{\,n_2\times n_0} & 0_{\,n_2\times n_1} & \mathbf{L}_{[2]} & 0_{\,n_2\times n_3} \\ 0_{\,n_3\times n_0} & 0_{\,n_3\times n_1} & 0_{\,n_3\times n_2} & \mathbf{L}_3^{\text{down}} \end{bmatrix}$$

where the metric Hodge Laplacian matrices are defined as

$$\mathbf{L}_{[p]} = \mathbf{L}_{[p]}^{\text{down}} + \mathbf{L}_{[p]}^{\text{up}},$$

with

$$\mathbf{L}_{[p]}^{\text{down}} = \mathbf{B}_p^\top \mathbf{G}_{p-1}^{-1} \mathbf{B}_p \mathbf{G}_p/(p+1),$$
$$\mathbf{L}_{[p]}^{\text{up}} = \mathbf{G}_p^{-1} \mathbf{B}_{p+1} \mathbf{G}_{p+1} \mathbf{B}_{p+1}^\top/(p+2).$$

Depending on the matrices $\mathbf{G}_p$, the weighted Dirac matrix may not always be symmetric, despite its eigenspectrum can be shown to be always real (see Appendix I).

For the rest of the paper, the metric matrices $\mathbf{G}_p$ shall be defined with each metric value for a simplex σ^p to be dependent on its $(p+1)$-dimensional cofaces in the following way[63]:

$$\mathbf{G}_p(\sigma^p,\sigma^p)=\begin{cases} w_{\sigma^d}, & p=d \\ w_{\sigma^p}+\sum\limits_{\sigma^p<\sigma^{p+1}} \mathbf{G}_{p+1}(\sigma^{p+1},\sigma^{p+1}), & 0\le p<d. \end{cases}$$

Here, $w_{\sigma^p}>0$ is a positive weight on p-simplex σ^p, which can be related to physical, chemical and biological properties.

Discrete Dirac matrix. With the weighted Dirac matrix $\overline{\mathbf{D}}_p$, a discrete Dirac matrix is simply the special case of $\overline{\mathbf{D}}_p$ when $\mathbf{G}_p$ are identity matrices and $a_p=1$ for all $p\ge 1$.

Previously, a general Dirac matrix has been defined as[49,69,70]

$$\mathbf{D}_p(z)=\begin{bmatrix} 0_{n_p\times n_p} & z\mathbf{B}_{p+1} \\ \overline{z}\mathbf{B}_{p+1}^\top & 0_{n_{p+1}\times n_{p+1}} \end{bmatrix},$$

where $z\in\mathbb{C}$ such that $|z|=1$. Since $|z|=1$, the typical values of z occurs when $z=\overline{z}=1$ or $z=-\overline{z}=i$. In general, the parameter $z\in\mathbb{C}$ extends the real eigenvectors of $\mathbf{D}_p(z)$ to $\mathbb{C}$ while the eigenvalue remains unchanged. By taking the square of the Dirac operator, we have

$$\mathbf{D}_p^2(z)=\begin{bmatrix} \mathbf{L}_p^{\text{up}} & 0_{n_p\times n_{p+1}} \\ 0_{n_{p+1}\times n_p} & \mathbf{L}_{p+1}^{\text{down}} \end{bmatrix},$$

which implies that the eigenvalues of diagonal block real-valued Hodge-Laplacian matrices will also be the eigenvalues of $\mathbf{D}_p^2(z)$. Since the Hodge-Laplacians are positive semi-definite symmetric matrices, the eigenvalues of $\mathbf{D}_p^2(z)$ are non-negative as well. However, eigenvectors from the Dirac matrix may contain complex numbers.

For a simplicial complex $\mathcal{K}$ with $n_{p-1}\times n_p$ boundary matrices $\mathbf{B}_p$ where n_{p-1} is the number of $(p-1)$-simplices and n_p is the number of p-simplices in $\mathcal{K}$, the discrete Dirac matrix $\mathbf{D}_p$[70] is

$$\mathbf{D}_p=\begin{bmatrix} 0_{n_0\times n_0} & \mathbf{B}_1 & 0_{n_0\times n_2} & \cdots & 0_{n_0\times n_p} & 0_{n_0\times n_{p+1}} \\ \mathbf{B}_1^\top & 0_{n_1\times n_1} & \mathbf{B}_2 & \cdots & 0_{n_1\times n_p} & 0_{n_1\times n_{p+1}} \\ 0_{n_2\times n_0} & \mathbf{B}_2^\top & 0_{n_2\times n_2} & \cdots & 0_{n_2\times n_p} & 0_{n_2\times n_{p+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{n_p\times n_0} & 0_{n_p\times n_1} & 0_{n_p\times n_2} & \cdots & 0_{n_p\times n_p} & \mathbf{B}_{p+1} \\ 0_{n_{p+1}\times n_0} & 0_{n_{p+1}\times n_1} & 0_{n_{p+1}\times n_2} & \cdots & \mathbf{B}_{p+1}^\top & 0_{n_{p+1}\times n_{p+1}} \end{bmatrix}. \tag{10}$$

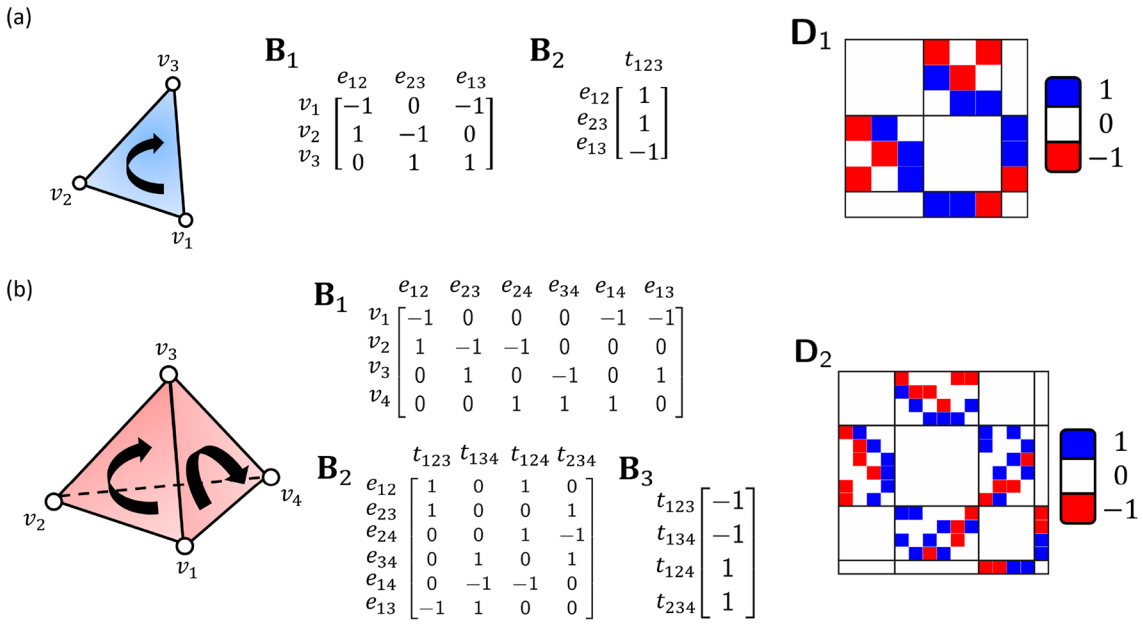


Figure 1. Illustration of constructions of (**a**) Discrete Dirac matrix $\mathbf{D}_1$ of a triangle and (**b**) Discrete Dirac matrix $\mathbf{D}_2$ of a tetrahedron along with its corresponding boundary matrices. The rows and columns of boundary matrices corresponds to a respective simplex each. For instance, in the boundary matrix $\mathbf{B}_2$ of (**a**), edge e_{12} is oriented similarly as t_{123}, hence having an entry 1 in the matrix. As the entries of Dirac operator either take a value of -1, 0 or 1, the entries of Dirac operators are color coded with blue indicating 1, white indicating 0 and red indicating -1.

It is of size $\sum_{i=0}^{p+1} n_i \times \sum_{i=0}^{p+1} n_i$.

Figure 1 shows a simple construction of discrete Dirac matrices (10) for a triangle and a tetrahedron. In Fig. 1a, the triangle is a 2-simplex and hence the largest Dirac operator is $\mathbf{D}_1$. On the other hand, the tetrahedron in Fig. 1b is a 3-simplex and thus the largest Dirac operator is $\mathbf{D}_2$.

Note that by taking the square of $\mathbf{D}_p$, one would obtain a matrix with diagonal blocks of unweighted combinatorial Hodge Laplacians as shown below.

$$\mathbf{D}_p^2 = \begin{bmatrix} \mathbf{L}_0 & 0_{\,n_0\times n_1} & 0_{\,n_0\times n_2} & \cdots & 0_{\,n_0\times n_p} & 0_{\,n_0\times n_{p+1}} \\ 0_{\,n_1\times n_0} & \mathbf{L}_1 & 0_{\,n_1\times n_2} & \cdots & 0_{\,n_1\times n_p} & 0_{\,n_1\times n_{p+1}} \\ 0_{\,n_2\times n_0} & 0_{\,n_2\times n_1} & \mathbf{L}_2 & \cdots & 0_{\,n_2\times n_p} & 0_{\,n_2\times n_{p+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{\,n_p\times n_0} & 0_{\,n_p\times n_1} & 0_{\,n_p\times n_2} & \cdots & \mathbf{L}_p & 0_{\,n_p\times n_{p+1}} \\ 0_{\,n_{p+1}\times n_0} & 0_{\,n_{p+1}\times n_1} & 0_{\,n_{p+1}\times n_2} & \cdots & 0_{\,n_{p+1}\times n_p} & \mathbf{L}_{p+1}^{\text{down}} \end{bmatrix},$$

where the unweighted Hodge Laplacian $\mathbf{L}_p$, is given by $\mathbf{L}_p = \mathbf{L}_p^{\text{down}} + \mathbf{L}_p^{\text{up}}$ with $\mathbf{L}_p^{\text{down}} = \mathbf{B}_p^\top \mathbf{B}_p$ and $\mathbf{L}_p^{\text{up}} = \mathbf{B}_{p+1}\mathbf{B}_{p+1}^\top$. In our case, the last term contains only $\mathbf{L}_{p+1}^{\text{down}}$.

Recall that $\mathbf{B}_{p+1}^\top \mathbf{B}_{p+1}$ is also known as the lower Hodge Laplacian $\mathbf{L}_{p+1}^{\text{down}}$ while $\mathbf{B}_{p+2}\mathbf{B}_{p+2}^\top$ is known as the upper Hodge Laplacian $\mathbf{L}_{p+1}^{\text{up}}$.

$$\mathbf{L}_{p+1} = \mathbf{L}_{p+1}^{\text{up}} + \mathbf{L}_{p+1}^{\text{down}}.$$

Spectrum of the discrete Dirac operator. *Spectral of Dirac matrix.* Let $\mathbf{Q}_p$ be the block diagonal matrix

$$\mathbf{Q}_p = \begin{bmatrix} \mathbf{I}_{n_0} & 0_{\,n_0\times n_1} & 0_{\,n_0\times n_2} & \cdots & 0_{\,n_0\times n_p} & 0_{\,n_0\times n_{p+1}} \\ 0_{\,n_1\times n_0} & -\mathbf{I}_{n_1} & 0_{\,n_1\times n_2} & \cdots & 0_{\,n_1\times n_p} & 0_{\,n_1\times n_{p+1}} \\ 0_{\,n_2\times n_0} & 0_{\,n_2\times n_1} & \mathbf{I}_{n_2} & \cdots & 0_{\,n_2\times n_p} & 0_{\,n_2\times n_{p+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{\,n_p\times n_0} & 0_{\,n_p\times n_1} & 0_{\,n_p\times n_2} & \cdots & (-1)^p\mathbf{I}_{n_p} & 0_{\,n_p\times n_{p+1}} \\ 0_{\,n_{p+1}\times n_0} & 0_{\,n_{p+1}\times n_1} & 0_{\,n_{p+1}\times n_2} & \cdots & 0_{\,n_{p+1}\times n_p} & (-1)^{p+1}\mathbf{I}_{n_{p+1}} \end{bmatrix},$$

where $\mathbf{I}_{n_p}$ denotes an $n_p \times n_p$ identity matrix and $\mathbf{Q}_p$ satisfies

$$\mathbf{Q}_p^2 = \mathbf{I}_{\sum_{i=0}^{p+1} n_i}.$$

The Dirac matrix satisfies the supersymmetry condition $\mathbf{D}_p\mathbf{Q}_p = -\mathbf{Q}_p\mathbf{D}_p$. That also means that the anti-commutator between the Dirac matrix $\mathbf{D}_p$ and block diagonal matrix $\mathbf{Q}_p$ vanishes. Further,

$$\begin{aligned} \mathbf{D}_p v = \lambda v &\iff \mathbf{Q}_p\mathbf{D}_p\mathbf{Q}_p v = -\mathbf{D}_p v = -\lambda v \\ &\iff -\mathbf{Q}_p\mathbf{D}_p v = -\lambda\mathbf{Q}_p v \\ &\iff \mathbf{D}_p\mathbf{Q}_p v = -\lambda\mathbf{Q}_p v, \end{aligned} \tag{11}$$

which implies that $\mathbf{Q}_p v$ is an eigenvector associated with the eigenvalue $-\lambda$.

Essentially, the above shows that the Dirac operator of a simplicial complex satisfies $\mathbf{D}_p v = \lambda v$ where λ is the eigenvalue associated with the eigenvector v if and only if

$$\mathbf{D}_p(\mathbf{Q}_p v) = -\lambda(\mathbf{Q}_p v),$$

where $\mathbf{Q}_p v$ is the eigenvector associated to the eigenvalue $-\lambda$.

Since λ (resp. $-\lambda$) is an eigenvalue of $\mathbf{D}_p$ with corresponding eigenvector v (resp. $\mathbf{Q}_p v$), then for any positive integer s, λ^s (resp. $(-\lambda)^s$) is an eigenvalue of $\mathbf{D}_p^s$ with corresponding eigenvector v (resp. $\mathbf{Q}_p v$). The detailed proof is in Appendix E.

Now, we consider the relationship between the eigenspectrum of $\mathbf{D}_p^2$ and $\mathbf{D}_p$. For the case of zero eigenvalues, $\mathbf{D}_p^2 v = 0$ naturally implies $\mathbf{D}_p v = 0$. Hence, $\mathbf{D}_p$ shares the same eigenvectors as $\mathbf{D}_p^2$ for zero eigenvalues. If λ^2 is a non-zero eigenvalue of $\mathbf{D}_p^2$ with eigenvector v, then we have the following possible cases for $\mathbf{D}_p$:

(i) λ is an eigenvalue of $\mathbf{D}_p$ with eigenvector $w = (\mathbf{D}_p + \lambda I)v$. i.e. $(\mathbf{D}_p - \lambda I)w = 0$.
(ii) $-\lambda$ is an eigenvalue of $\mathbf{D}_p$ with eigenvector $w = (\mathbf{D}_p - \lambda I)v$. i.e. $(\mathbf{D}_p + \lambda I)w = 0$.

It is easy to derive the above cases by considering $(\mathbf{D}_p^2 - \lambda^2 I)v = 0$. Then

$$(\mathbf{D}_p - \lambda I)(\mathbf{D}_p + \lambda I)v = 0. \tag{12}$$

Here, there are two possible cases since by (11), either λ or $-\lambda$ is the eigenvalue of $\mathbf{D}_p$. If $-\lambda$ is the eigenvalue of $\mathbf{D}_p$, then $(\mathbf{D}_p + \lambda I)w = 0$ for some non-zero eigenvector w. This implies that $(\mathbf{D}_p - \lambda I)w \neq 0$, otherwise it contradicts $(\mathbf{D}_p + \lambda I)w = 0$. Hence, this means that Eq. (12) can be rewritten as

$$(\mathbf{D}_p + \lambda I)w = 0,$$

where $w = (\mathbf{D}_p - \lambda I)v$ is a non-zero eigenvector for $\mathbf{D}_p$ with corresponding eigenvalue $-\lambda$.

Similarly, if λ is an eigenvalue of $\mathbf{D}_p$, then $(\mathbf{D}_p - \lambda I)w = 0$ for some non-zero eigenvector w. This implies that $(\mathbf{D}_p + \lambda I)w \neq 0$, otherwise it contradicts $(\mathbf{D}_p - \lambda I)w = 0$. Therefore, Eq. (12) can be rewritten as

$$(\mathbf{D}_p - \lambda I)w = 0,$$

where $w = (\mathbf{D}_p + \lambda I)v$ is a non-zero eigenvector for $\mathbf{D}_p$ with corresponding eigenvalue λ.

This leads us to the following relations connecting $\mathbf{D}_p$, $\mathbf{D}_p^2$ and $\mathbf{L}_k$ $(0 \leq k \leq p+1)$. For any $v \in \ker \mathbf{D}_p^2$,

$$\mathbf{D}_p^2 v = 0 \iff \begin{cases} \mathbf{L}_0 \mathbf{w}_0 = 0\,, & k = 0 \\ \mathbf{L}_k \mathbf{w}_k = 0\,, & 0 < k < p+1 \\ \mathbf{L}_{p+1}^{\text{down}} \mathbf{w}_{p+1} = 0\,, & k = p+1 \end{cases},$$

where $v = (\mathbf{w}_0^\top, \mathbf{w}_1^\top, \ldots, \mathbf{w}_{k-1}^\top, \mathbf{w}_k^\top, \mathbf{w}_{k+1}^\top, \ldots, \mathbf{w}_p^\top, \mathbf{w}_{p+1}^\top)^\top$. In other words, v is a vector consisting of block vectors $\mathbf{w}_k^\top$ for $0 \leq k \leq p+1$. This means that for every $0 \leq k \leq p$, $\mathbf{w}_k^\top \in \ker \mathbf{L}_k$. In the case where $k = p+1$, $\mathbf{w}_{p+1}^\top \in \ker \mathbf{L}_{p+1}^{\text{down}}$. We have,

$$(\mathbf{w}_0^\top, \mathbf{w}_1^\top, \ldots, \mathbf{w}_{p+1}^\top)^\top \in \ker \mathbf{L}_{p+1}^{\text{down}} \oplus \bigoplus_{k=0}^{p} \ker \mathbf{L}_k.$$

Note that for $\mathbf{w}_{p+1}^\top$, it is the eigenvector from the kernel of $\mathbf{L}_{p+1}^{\text{down}}$.

Hence, the kernel of $\mathbf{D}_p^2$ can be decomposed into a direct sum of kernels of $\mathbf{L}_k$ from $k = 0$ to $k = p+1$:

$$\ker \mathbf{D}_p^2 = \ker \mathbf{L}_{p+1}^{\text{down}} \oplus \bigoplus_{k=0}^{p} \ker \mathbf{L}_k.$$

Further, we have

$$\begin{aligned} \ker \mathbf{D}_p = \ker \mathbf{D}_p^2 &= \ker \mathbf{L}_{p+1}^{\text{down}} \oplus \bigoplus_{k=0}^{p} \ker \mathbf{L}_k \\ &\cong \ker \mathbf{L}_{p+1}^{\text{down}} \oplus \bigoplus_{k=0}^{p} H_k, \end{aligned} \tag{13}$$

where $\bigoplus_{k=0}^{p} H_k$ refers to the direct sum of homology groups.

Therefore, the eigenvectors of $\mathbf{D}_p$ reveal both k-th homology and k-th non-homology information within the structural data for all $0 \leq k \leq p+1$. Instead of eigendecomposing HL matrices for all $0 \leq k \leq p+1$, one can simply eigendecompose $\mathbf{D}_p$ to obtain all of the eigenspectrums. As the number of zero eigenvalues of $\mathbf{L}_{p+1}^{\text{down}}$ is the rank $\mathbf{B}_{p+2}^\top$ plus the $(p+1)$-th Betti number β_{p+1}, the multiplicity of zero eigenvalues in $\mathbf{D}_p$ is the rank $\mathbf{B}_{p+2}^\top$ plus the total sum of all the Betti numbers from dimension 0 to $p+1$. That is,

$$\dim \ker \mathbf{D}_p = \operatorname{rank} \mathbf{B}_{p+2}^\top + \sum_{k=0}^{p+1} \beta_k. \tag{14}$$

Mathematically, the eigenvectors corresponding to the zero eigenvalues are known as homology generators while those from non-zero eigenvalues are the non-homology generators. Both of them can be used in structural clustering. More specifically, the homology generators can be used for clustering structures based on their loop or circle components, while non-homology generators are related to the spectral clustering, in which communities and clusters are based on their distances. Figure 2 demonstrates the structural clustering with homology and non-homology generators for a protein (PDBID: 1AXC). We only consider the C_α atoms in structure. A Vietoris Rips complex is constructed by using a cutoff distance of 10Å. The Dirac matrix $\mathbf{D}_1$ and its eigenvalues and eigenvectors are calculated. As the non-zero eigenvalues of $\mathbf{D}_1$ come in pairs, it suffices to consider the eigenvectors corresponding to the positive eigenvalues. For all the non-negative eigenvalues of $\mathbf{D}_1$, the eigenvectors are arranged in ascending order according to its corresponding eigenvalues.

Figure 2a illustrates the loop/circle-based clustering using four one-dimensional (1D) homology generators. Note that these 1D homology generators $\mathbf{w}_1^\top$ are taken from the homology generators of $\mathbf{D}_1$ (with eigenvalues 0). More specifically, these 1D homology generators are defined by the 1-simplices. In Fig. 2a, a thick edge with dark blue color indicates large magnitude of the value, while a thinner edge with light blue color means the corresponding 1D homology generator has a value with small magnitude on this 1-simplex. It can be seen that

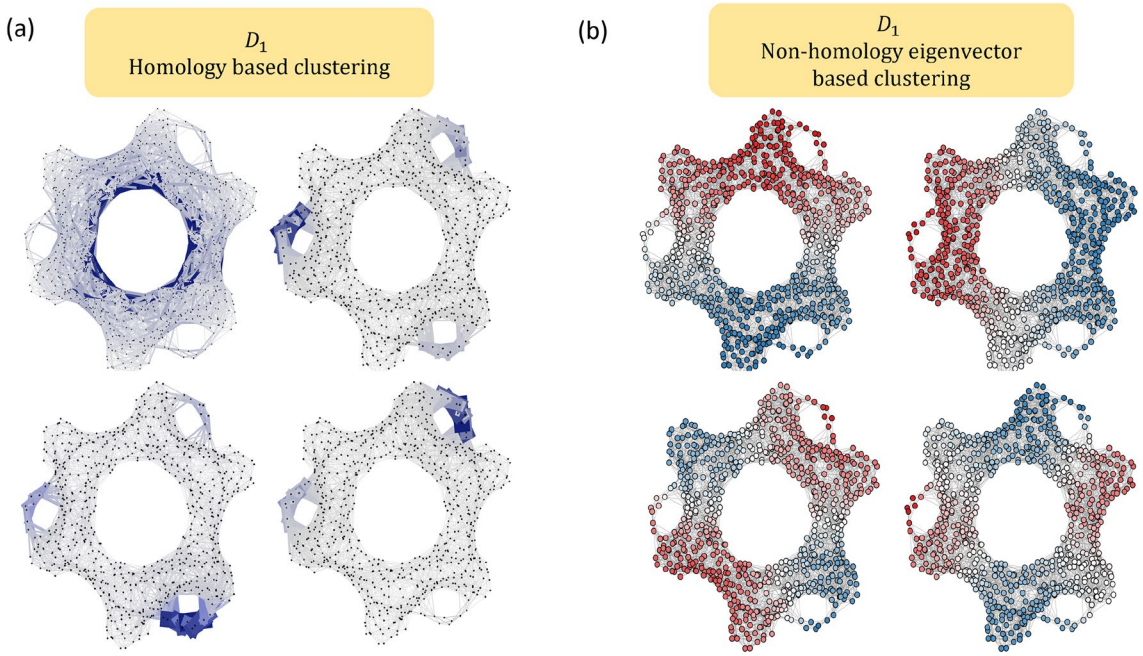


Figure 2. Illustration of loop/circle-based clustering using four one-dimensional (1D) homology generators (**a**) and spectral clustering using four zero-dimensional (0D) non-homology generators (**b**). The Dirac matrices $\mathbf{D}_1$ are generated from the Vietoris Rips complex of the C_α atoms in PDBID: 1AXC at 10Å. (**a**) Here 1D homology generators $\mathbf{w}_1^\top$ are taken from the homology generators of $\mathbf{D}_1$ with eigenvalues as 0. A thick edge with dark blue color indicates large magnitude of the value, while a thinner edge with light blue color means the corresponding the 1D homology generator has a value with small magnitude on this 1-simplex. Each 1D homology generator forms an individual loop or circle. (**b**) The four 0D non-homology generators $\mathbf{w}_0^\top$ are taken from the non-homology generators of $\mathbf{D}_1$ with the four smallest positive eigenvalues. Note that these 0D non-homology generators are defined on nodes (0-simplices). Nodes with negative values are colored in red while nodes with positive values are of blue color. It can be seen that the nodes in the structure can be naturally clustered into groups based on the signs of these 0D non-homology generators.

edges with large magnitudes are the 1-simplices that form circles or loops. Each 1D homology generator forms an individual loop or circle. In this way, 1D homology generators can be used for loop/circle-based clustering of molecular structures.

Figure 2b illustrates the spectral clustering using four zero-dimensional (0D) non-homology generators. The four 0D non-homology generators $\mathbf{w}_0^\top$ are taken from the non-homology generators of $\mathbf{D}_1$ with the four smallest positive eigenvalues. Note that these 0D non-homology generators are defined on nodes (0-simplices). In Fig. 2b, nodes with negative values are colored in red while nodes with positive values are of blue color. It can be seen that the nodes in the structure can be naturally clustered into groups based on the signs of these 0D non-homology generators. This approach is known as spectral clustering and widely used in data analysis. It should be noticed that using the higher order Dirac matrices, we can cluster not only nodes (0-simplices), but also higher dimensional simplices.

Spectral of weighted Dirac matrix. The weighted Dirac matrix has different spectral properties based on the different weighting schemes. Figure 3 illustrates the spectrum of the weighted Dirac matrix defined from the guanine molecule structure (using all-atom representation). We construct an unweighted Vietoris Rips complex using a cutoff distance of 1.2Å. The discrete Dirac matrix $\mathbf{D}_1$ can be computed using (10). The discrete Dirac matrix $\mathbf{D}_1$ is eigendecomposed to obtain its eigenvalues and eigenvectors. Moreover, a weighted simplicial complex is constructed by assigning simplex σ with different weight w_σ. The metric matrices $\mathbf{G}_p$ are computed and weighted Dirac matrix $\overline{\mathbf{D}}_1$ can then be constructed. Figure 3 shows the homology generators and Fiedler vector for an unweighted simplicial complex and three different weighted simplicial complexes. Among the three weighted simplicial complexes, Fig. 3b shows a weighted simplicial complex where all weights w_σ are equal to 1. Two modified weighted simplicial complexes are constructed by modifying the weights of edge e_1 ranging from 10 and 0.01 with all the other weights kept unchanged. With the same underlying simplicial complex, they share the same three homology generators, one 1D component and two 2D circles. Figure 3 shows the corresponding eigenvectors for these homology generators. The magnitude of the eigenvectors are represented by the thickness and darkness. An edge (or vertex) with thicker lines and darker blue color indicates a larger magnitude.

In general, the weight of a simplex has an inverse effect on the corresponding element of the homology eigenvectors (i.e., homology generators). When the simplex has a smaller weight, the corresponding element of the homology eigenvectors has larger magnitude. Similar patterns also appear in non-homology generators.

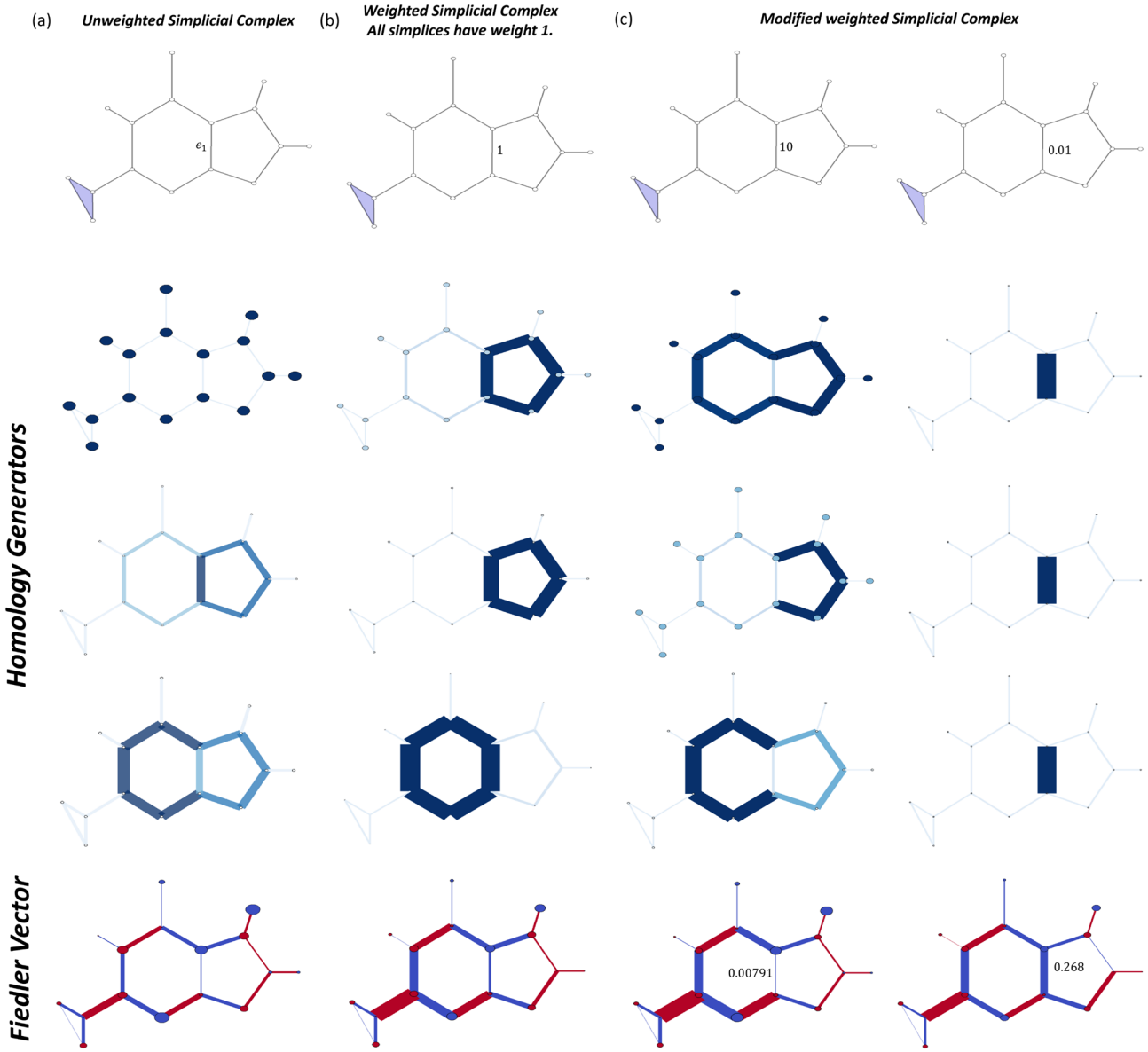


Figure 3. Illustration of three homology generators and Fiedler vector from (**a**) discrete Dirac matrix and (**b**,**c**) weighted Dirac matrix (from weighted simplicial complexes). For the discrete Dirac matrix, the three homology generators represents one 1D component and two 2D circles. By assigning simplex σ with different weight w_σs, three weighted simplicial complexes are constructed in (**b**) and (**c**). In (**b**), the weighted simplicial complex consists of all weights w_σ equal to 1. (**c**) Shows two weighted simplicial complexes by changing the weights of edge e_1 from 1 to 10 and 0.01 while the rest of weights remain unchanged. The magnitude of the homology generators are influenced by these weights and are reflected based on their thickness and darkness. For the homology generators, the edges (or vertices) are thicker and in darker blue color if they have a larger magnitude. Similarly, the edges and vertices are colored in red/blue if their elements in the Fiedler vectors have positive/negative sign. The magnitudes of their values in Fiedler vectors are represented by the thickness of edges and size of vertices.

Fig. 3 illustrates the Fiedler vectors (i.e., eigenvector corresponding to the first smallest non-zero eigenvalue) of the nonweighted and weighted simplicial complexes. Simplices are colored in red/blue if the element of the non-homology eigenvectors has value positive/negative. The thickness of simplices represents the magnitude of their values in non-homology generators. It can be seen clearly that the weight of a simplex has an inverse effect on its magnitude of the values of eigenvectors.

Results

Persistent Dirac. *Mathematical foundation for persistent Dirac analysis.* Recently, persistent Laplacian and persistent sheaf Laplacians have been developed[61,62]. Their essential idea is to explore the persistence of spectral information during the filtration process. Here we develop the rigorous mathematical framework for persistent Dirac.

Let $(\mathbb{R}, \leq)$ be a category of real numbers with morphisms given by $a \to b$ for any $a \leq b$. A functor $\mathcal{F} : (\mathbb{R}, \leq) \to \mathbf{Simp}$ gives a filtration of simplicial complexes of finite type, i.e. $\mathcal{F}$ maps from a category of real numbers to a category of simplicial complexes of finite type. For any two real numbers $a \leq b$, the functor $\mathcal{F}$ satisfies the inclusion

$$\mathcal{F}(a) \hookrightarrow \mathcal{F}(b),$$

which induces a morphism of chain complexes

$$C_*(\mathcal{F}(a), \mathbb{R}) \hookrightarrow C_*(\mathcal{F}(b), \mathbb{R}).$$

Let $\mathcal{F}(\infty) = \bigcup_{a\in\mathbb{R}} \mathcal{F}(a)$ and $C_* = C_*(\mathcal{F}(\infty), \mathbb{R})$. Note that C_* can be endowed with an innerproduct $\langle \cdot, \cdot \rangle$. Further, a subspace $C_*(\mathcal{F}(a), \mathbb{R})$ would inherit the inner product structure of C_* and a boundary operator given by the restriction

$$\partial_p^a = \partial_p|_{C_p(\mathcal{F}(a),\mathbb{R})} : C_p(\mathcal{F}(a), \mathbb{R}) \to C_{p-1}(\mathcal{F}(a), \mathbb{R}).$$

Here ∂_* is the boundary operator of C_*. For convenience, we shall write $C_p^a = C_p(\mathcal{F}(a), \mathbb{R})$. For a pair of simplicial complexes $\mathcal{F}(a) \subset \mathcal{F}(b)$, we consider the inclusion map $\iota : \mathcal{F}(a) \hookrightarrow \mathcal{F}(b)$. For $p \in \mathbb{N}$, the subspace

$$C_p^{a,b} := \{x \in C_p^b : \partial_p^b(x) \in C_{p-1}^a\} \subseteq C_p^b,$$

which consists of the p-chains in C_p^b such that their images are under the boundary operator ∂_p^b in the subspace C_{p-1}^a of C_{p-1}^b. Also, we have a linear operator

$$\partial_p^{a,b} = \partial_p^b|_{C_p^{a,b}} : C_p^{a,b} \to C_{p-1}^a,$$

which induces an adjoint operator

$$(\partial_p^{a,b})^* : C_{p-1}^a \to C_p^{a,b}$$

with respect to the inner product $\langle \cdot, \cdot \rangle$.

Let $n_p^{a,b} := \dim(C_p^{a,b})$. Then following commutative diagram is thus induced by ι.

$$\begin{array}{ccccccccc}
\cdots & \xrightarrow{\partial^a} & C_{p-1}^a & \xrightarrow{\partial^a} & C_p^a & \xrightarrow{\partial^a} & C_{p+1}^a & \xrightarrow{\partial^a} & \cdots \\
 & & \uparrow \iota & & \uparrow \iota & & \uparrow \iota & & \\
\cdots & \xrightarrow{\partial^b} & C_{p-1}^b & \xrightarrow{\partial^b} & C_p^b & \xrightarrow{\partial^b} & C_{p+1}^b & \xrightarrow{\partial^b} & \cdots
\end{array}$$

Notice that $\partial_p^{a,b}$ is a restriction to $C_p^{a,b}$ in order to obtain the "diagonal" operators $\partial_p^{a,b} : C_p^{a,b} \to C_{p-1}^a$. Similarly, with a restriction to $C_p^{a,b}$, we can then define the p-dimensional boundary matrices $\mathbf{B}_p^{a,b}$ which consists of every entry value of $\partial_p^{a,b}$.

The persistent Dirac operator $\mathbf{D}_p^{a,b}$ can then be written as follows.

$$\mathbf{D}_p^{a,b} = \begin{bmatrix}
0_{\,n_0\times n_0} & \mathbf{B}_1^{a,b} & 0_{\,n_0\times n_2} & \cdots & 0_{\,n_0\times n_p} & 0_{\,n_0\times n_{p+1}} \\
(\mathbf{B}_1^{a,b})^\top & 0_{\,n_1\times n_1} & \mathbf{B}_2^{a,b} & \cdots & 0_{\,n_1\times n_p} & 0_{\,n_1\times n_{p+1}} \\
0_{\,n_2\times n_0} & (\mathbf{B}_2^{a,b})^\top & 0_{\,n_2\times n_2} & \cdots & 0_{\,n_2\times n_p} & 0_{\,n_2\times n_{p+1}} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0_{\,n_p\times n_0} & 0_{\,n_p\times n_1} & 0_{\,n_p\times n_2} & \cdots & 0_{\,n_p\times n_p} & \mathbf{B}_p^{a,b} \\
0_{\,n_{p+1}\times n_0} & 0_{\,n_{p+1}\times n_1} & 0_{\,n_{p+1}\times n_2} & \cdots & (\mathbf{B}_p^{a,b})^\top & 0_{\,n_{p+1}\times n_{p+1}}
\end{bmatrix}.$$

The maps and spaces are also illustrated in the diagram below

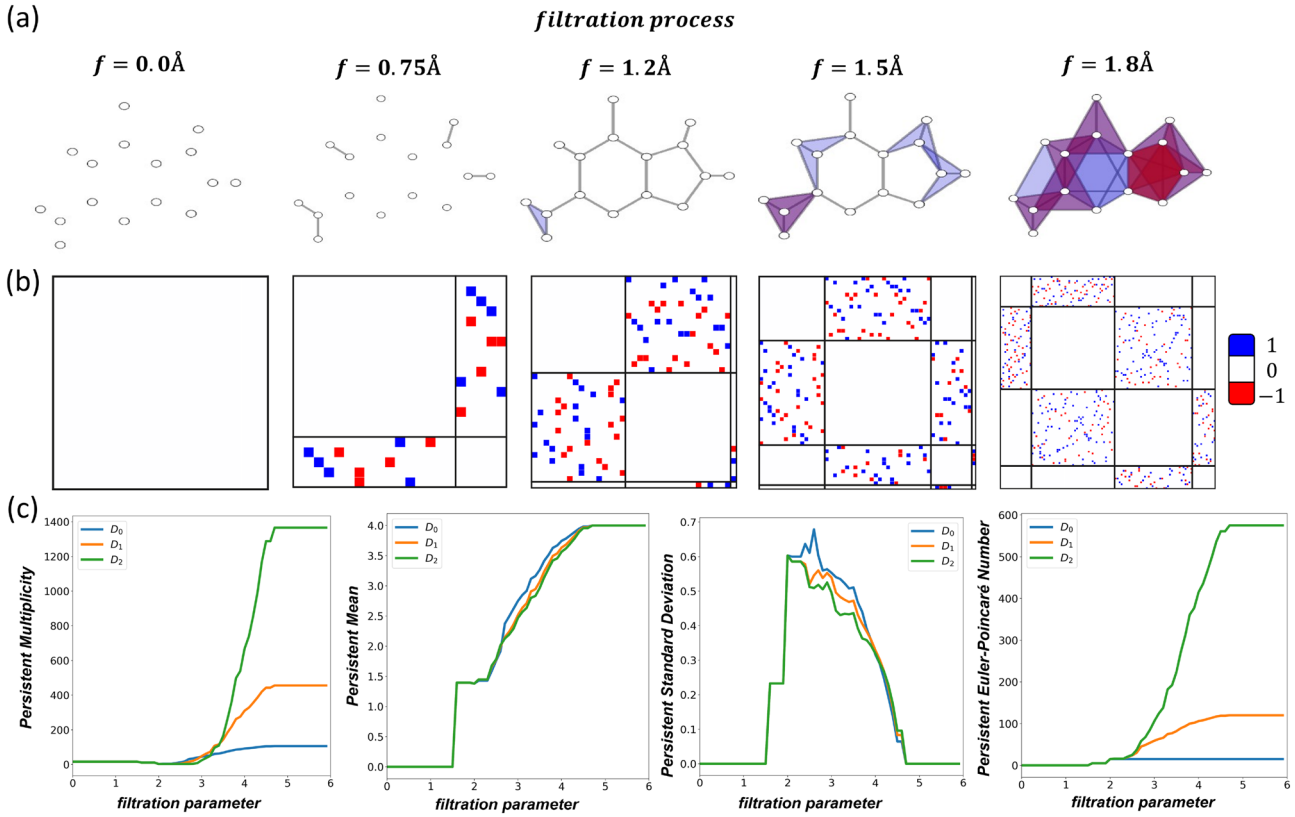
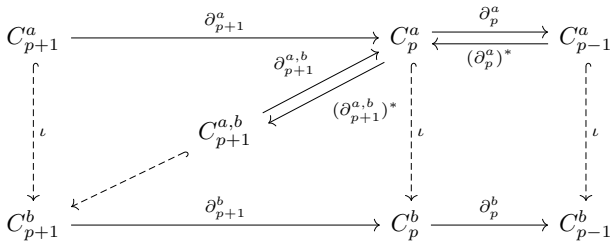


Figure 4. Illustration of the filtration process of the guanine molecule (**a**), its associated Dirac matrices (**b**), and persistent attributes (**c**). In the filtration process, more simplices are formed in simplicial complex and the size of Dirac matrix increases. The eigenspectrum of Dirac matrices changes in the filtration process. The changes in eigenspectrum are being converted into a series of 12 statistical and combinatorial attributes (i–xii). One of the statistical attribute, persistent multiplicity, provides quantitative analysis to the change in zero eigenvalues of Dirac matrices while the remaining 11 persistent attributes are derived from the non-zero eigenvalues.



Further, the p-th persistent Hodge Laplacian can be defined as

$$\mathbf{L}_p^{a,b} = \begin{cases} \mathbf{B}_1^{a,b}(\mathbf{B}_1^{a,b})^\top, & p = 0 \\ (\mathbf{B}_p^{a,b})^\top \mathbf{B}_p^{a,b} + \mathbf{B}_{p+1}^{a,b}(\mathbf{B}_{p+1}^{a,b})^\top, & p > 0. \end{cases}$$

Similarly, the matrices $(\mathbf{B}_p^{a,b})^\top \mathbf{B}_p^{a,b}$ and $\mathbf{B}_{p+1}^{a,b}(\mathbf{B}_{p+1}^{a,b})^\top$ are the p-th persistent lower and upper Hodge Laplacians $(\mathbf{L}_{p+1}^{\text{down}})^{a,b}$ and $(\mathbf{L}_{p+1}^{\text{up}})^{a,b}$ respectively. Based on (13), the following result shows that the nullity of p-th persistent Dirac operator equals to the rank of $\mathbf{B}_{p+2}^\top$ plus the sum of k-th persistent Betti numbers, where $0 \le k \le p+1$.

$$\ker \mathbf{D}_p^{a,b} = \ker(\mathbf{D}_p^{a,b})^2 = \ker(\mathbf{L}_{p+1}^{\text{down}})^{a,b} \oplus \bigoplus_{k=0}^{p} \ker \mathbf{L}_k^{a,b}$$
$$\cong \ker(\mathbf{L}_{p+1}^{\text{down}})^{a,b} \oplus \bigoplus_{k=0}^{p} (H_k)^{a,b},$$

where $\bigoplus_{k=0}^{p}(H_k)^{a,b}$ refers to the direct sum of (a, b)-persistent homology groups. The (a, b)-persistent homology groups characterizes the homology generators that are born at time a and survive to time b.

Figure 4 illustrates the persistent Dirac analysis of the guanine molecule (using all-atom representation). More specifically, Fig. 4a shows the Vietoris–Rips complex of the guanine molecule when filtration parameter $f = 0.0$Å, 0.75Å, 1.2Å, 1.5Å and 1.8Å. In particular, triangles first appear around 1.2Å and tetrahedron starts to appear at 1.5Å. Figure 4b shows the corresponding Dirac matrix $\mathbf{D}_2$. The size of the Dirac matrix $\mathbf{D}_2$ consistently increases during the filtration process.

Persistent attributes. For any Dirac matrix, its non-zero eigenvalues come in pairs. Each pair contains one negative eigenvalue and one positive counterpart. For the set of all its positive eigenvalues, a Dirac Zeta function can be defined as follows[69],

$$\zeta(s) = \sum_{j=1}^{n} \frac{1}{\lambda_j^s} = \sum_{j=1}^{n} e^{-s\log\lambda_j}, s \in \mathbb{C}.$$

Here $\zeta(-m) = \sum_{i=1}^{n} \lambda_i^m, m \in \mathbb{Z}$ is the m-th spectral moments of Dirac matrices and $\zeta(-1)$ is the Laplacian graph energy. Another way to define Dirac Zeta function is to consider its negative eigenvalues by replacing the λ_j^{-s} with $(1 + e^{-i\pi s})|\lambda_j|^{-s}$. Here λ_j can be negative. For instance, $\zeta(2) = 2\sum_{j=1} \lambda_j^{-2}$.

Furthermore, the q-Dirac complexity of a simplicial complex $\mathcal{K}$ can be defined as

$$c_q(\mathbf{D}_p) = \prod_{\substack{\lambda_j \neq 0 \\ \lambda_j \in \sigma(\mathbf{D}_p)}} \lambda_j^q.$$

The case where $q = 1$ is previously introduced by Knill[70]. $c_1(\mathbf{D}_p)$ is equal to the product of all non-zero eigenvalues in spectra of $\mathbf{D}_p$ since the non-zero eigenvalues come in pairs. The number of non-zero eigenvalues pairs in $\mathbf{D}_p$ is the (signless) Euler–Poincaré number defined as follows,

$$\ell = \frac{1}{2}\sum_{k=0}^{p+1} n_k - \frac{1}{2}\dim\ker \mathbf{D}_p$$

where n_k is the number of k-simplices and $\dim\ker \mathbf{D}_p$ is the multiplicity of zero eigenvalues of $\mathbf{D}_p$.

Using Eq. (14), ℓ can be computed as follows:

$$\ell = \frac{1}{2}\sum_{k=0}^{p+1}(n_k - \beta_k) - \frac{1}{2}\text{rank}\,\mathbf{B}_{p+2}^{\top}. \tag{15}$$

Interestingly, the spanning tree number, introduced as one of the spectral indices in molecular descriptors[1], can be written as

$$t(\mathbf{D}_p) = \frac{1}{2}\log(c_1(\mathbf{D}_p)) - \log(\ell + 1),$$

Alternatively, $t(\mathbf{D}_p) = \log\left[\frac{1}{\ell+1} \cdot \sqrt{c_1(\mathbf{D}_p)}\right]$.

To summarize, we consolidate and consider 11 statistical and combinatorial attributes as molecular descriptors for each given set of positive eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_\ell\}$ where ℓ is the number of non-zero eigenvalue pairs:

(i) $\min\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, also known as the Fiedler value.
(ii) $\max\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$
(iii) $\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i = \frac{1}{n}\zeta(-1)$.
(iv) Standard Deviation
(v) Laplacian Graph Energy $\zeta(-1)$.
(vi) (Signless) Euler–Poincaré Number (number of non-zero eigenvalue pairs) ℓ
(vii) Generalised Mean Graph Energy $\sum_{i=1}^{n} \frac{|\lambda_i - \bar{\lambda}|}{n}$.
(viii) Spectral 2nd Moment $\zeta(-2)$.
(ix) $\zeta(2) = 2\sum_{j=1}^{n} \lambda_j^{-2}$.
(x) Quasi-Wiener Index $(n + 1)\zeta(1)$.
(xi) Spanning Tree Number $t(\mathbf{D}_p)$.

In addition to the 11 statistical attributes, we also consider the persistent multiplicity of zero eigenvalues.

(xii) Persistent Multiplicity of zero eigenvalues.

Figure 4c shows the persistent multiplicity, persistent mean, persistent standard deviation and persistent (signless) Euler–Poincaré number for the filtration of guanine molecule. Further information such as the persistent multiplicities of $\mathbf{L}_k$ ($0 \leq k \leq 2$) and $\mathbf{L}_k^{\text{down}}$ ($1 \leq k \leq 3$) can be found in Appendix G. Recall that the persistent multiplicity is equivalent to the persistent Betti number. Here, the persistent multiplicity and persistent (signless) Euler–Poincaré number of $\mathbf{D}_p$ can be quantitatively analysed by comparing the persistent multiplicity of $\mathbf{L}_{p+1}^{\text{down}}$ and the k-th persistent Betti numbers for $0 \leq k \leq p$. It can be seen that these persistent attributes change with the filtration value. Each variation of the persistent attribute indicates a certain change in the simplicial complex.

At the very start of the filtration, there are 16 isolated atoms which means that there are 16 connected components. Hence, the persistent multiplicity of $\mathbf{L}_0$ is 16 since $\beta_0 = 16$. As all other Betti numbers are zero and there are no higher order simplices present at the start of the filtration, $\mathbf{D}_0$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are all-zero 16×16 matrices. Therefore, the persistent multiplicity of $\mathbf{D}_0$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are all equal to 16. Using Eq. (15), the persistent (signless) Euler–Poincaré number is zero.

As filtration parameter f increases, the size of $\mathbf{D}_0$, $\mathbf{D}_1$ and $\mathbf{D}_2$ matrix increases as well. This differs from the Hodge Laplacian matrix $\mathbf{L}_0$, whose size remains unchanged.

At filtration size 4.7Å, a complete simplicial complex is achieved, i.e., any $p+1$ vertices will form a p-simplex. When this happens, the size of $\mathbf{D}_p$ no longer increases any further. Here, the size of $\mathbf{D}_0$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are distinct. The size of $\mathbf{D}_0$ is 136×136 since $\frac{16\times15}{2}$ (no. of 1-simplices) + 16 (no. of 0-simplices) = 136. Similarly, the sizes of $\mathbf{D}_1$ and $\mathbf{D}_2$ are 696×696 and 2516×2516 respectively. Furthermore, the persistent multiplicity of $\mathbf{D}_0$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are also distinct. Using Eq. (14), the persistent multiplicity of $\mathbf{D}_0$ is 105 (persistent multiplicity of $\mathbf{L}_1^{\text{down}}$) and 1 (0-dimensional persistent Betti number) which sums up to 106. Since the persistent multiplicity of $\mathbf{L}_1$ (see Appendix G) is zero, then Eq. (3) implies that the rank of $\mathbf{B}_2^\top$ is 105. In addition, the persistent multiplicity of $\mathbf{D}_1$ and $\mathbf{D}_2$ are 456 and 1366 respectively. Based on the non-zero eigenvalues, the persistent (signless) Euler–Poincaré number of $\mathbf{D}_0$, $\mathbf{D}_1$ and $\mathbf{D}_2$ is 15, 120 and 575 due to Eq. (15).

Persistent Dirac for molecular structure representation. Recently, a series of persistent models, including persistent homology, persistent spectral, persistent Ricci curvature, and persistent Laplacian, have demonstrated their great power in molecular representations[3,6,38,71]. They have consistently outperformed traditional graph-based models in various tasks of drug design. Here we study the representation capability of Persistent Dirac in molecular data analysis.

We consider the organic-inorganic halide perovskite (OIHP) dataset. More specifically, three kinds of Methylammonium lead halides ($MAPbX_3$, X=Cl, Br, I), i.e., orthorhombic, tetragonal, and cubic phase of $MAPbX_3$ are used. For each kind, there are 3 types of X atoms, including chlorine Cl, bromine Br and iodine I. The molecular dynamic simulations are systematically carried out on these molecular structures with the initial configurations based on pre-defined crystal cell parameters. For each $MAPbX_3$ structure, 1000 configurations are equally sampled from its MD simulation trajectory and the last 500 configurations, which represent stable structures, are selected for the test of our persistent Dirac model. Essentially, a total of 4500 configurations from the 9 types of $MAPbX_3$ structures are mixed together and our persistent Dirac based molecular fingerprint is used in the clustering of these configurations.

Computationally, our persistent Dirac is generated based on Alpha complex and the filtration parameter is the circumradius. More specifically, for each frame, an Alpha complex is constructed based on Delauney triangulation and circumradius of the simplex. The Dirac matrices $\mathbf{D}_0$ and $\mathbf{D}_1$ are computed from 1Å to 6.5Å

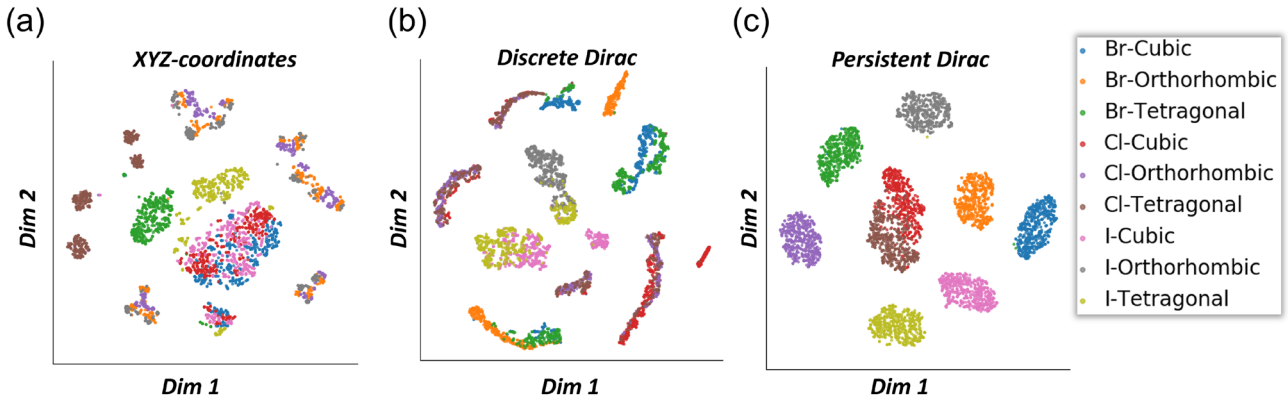


Figure 5. The clustering of 9 types of OIHP molecular dynamics (MD) trajectories. Three feature generation schemes are considered, including (**a**) *XYZ*-coordinates, (**b**) Discrete Dirac at 3.5Å and (**c**) Persistent Dirac. Each trajectory contains 1000 configurations and *t*-SNE model is used for clustering (of the last 500 configurations at equilibrium). The x-axis and y-axis are the two principal components obtained from the *t*-SNE model.

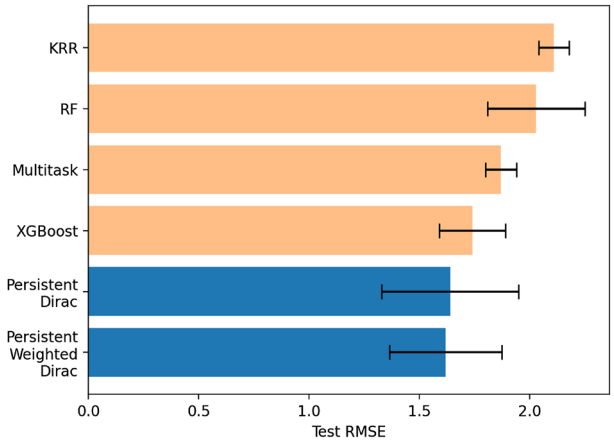


Figure 6. The comparison of persistent Dirac and persistent weighted Dirac models with conventional methods from MoleculeNet[74] on the FreeSolv database. Both types of persistent Dirac models performed better than existing conventional methods such as MoleculeNet's XGBoost (1.74 ± 0.15 kcal/mol), Multitask (1.87 ± 0.07 kcal/mol), Random Forest (RF) (2.03 ± 0.22 kcal/mol) and Kernel Ridge Regression (KRR) (2.11 ± 0.07 kcal/mol) models. Note that a lower RMSE value indicates better result.

with stepsize 0.25Å throughout the filtration process. Hence, the eigenvalues of $\mathbf{D}_0$ and $\mathbf{D}_1$ each contribute to 12 statistical attributes (i–xii) for 23 timesteps per frame. The feature size sums up to 552. By considering with and without hydrogen atoms, the total feature size for persistent Dirac is $552 \times 2 = 1104$. Likewise, for coordinate-only model, the input features are *xyz*-coordinates of all the atoms. Since each structure consists of 553 atoms, the feature size is of $553 \times 3 = 1659$. For the discrete Dirac model, the feature size is 552. The clustering of these $MAPbX_3$ structures is then studied using unsupervised learning models, in particular *t*-distributed stochastic neighbor embedding (*t*-SNE).

Figure 5 illustrates the comparison of the clustering results from three different models, including coordinate-only model (*XYZ*-coordinate) (a), discrete Dirac (b), and persistent Dirac (c). It can be seen that our persistent Dirac model demonstrates better capabilities in characterizing the intrinsic structure information and discriminating the 9 types of OIHPs clearly. In our persistent Dirac model, the filtration process at various scales provided the geometrical information needed to balance the topological information. The combination of topological and geometrical information contributes to the success of our persistent Dirac model in OIHP clustering. Figure 5b shows the performance of Dirac matrix related statistical attributes at filtration value 3.5Å. Even though it shows certain clustering effects, the overall performance is not as good as persistent Dirac. Additional clustering tests are performed for discrete Dirac model at 3Å and 4Å in Appendix F. Similarly, statistical attributes of discrete Dirac model at a single scale fail to distinguish the 9 types of OIHPs.

Persistent Dirac for solvation free energy prediction

In order to further validate the capabilities of persistent Dirac models, we perform a preliminary test of our persistent Dirac model on the Free Solvation (FreeSolv) database[72]. The FreeSolv database contains 643 SMILES sequences for small molecules and their solvation free energy values in water[73]. The FreeSolv database is also one of the physical chemistry benchmark tasks in MoleculeNet[74]. Recently, structural information has been generated from 643 SMILES sequences and applied in graph-based methods to improve overall solvation energy predictions[7]. Using the structural information, we consider three atom subsets, i.e. (A): all atoms, (B): all atoms except hydrogen, and (C): all atoms except hydrogen and carbon, for our persistent Dirac model. For (A), we generate the discrete Dirac matrices $\mathbf{D}_0$ and $\mathbf{D}_1$ using the Alpha complex for the filtration process from 0Å to 12Å with stepsize 0.1Å. Similarly, we generate the discrete Dirac matrices for (B) and (C) using the Rips complex for the filtration process from 0Å to 12Å with stepsize 0.1Å. We compute the 12 statistical attributes (i–xii) from the eigenvalues of $\mathbf{D}_0$ and $\mathbf{D}_1$ for all the 120 timesteps. From (A), (B) and (C), the total feature size sums up to $120 \times 12 \times 3 \times 2 = 8640$. Following the training and testing procedure in MoleculeNet[74], we apply the random 80/10/10 split and our persistent Dirac features act as input features for a XGBoost model. Specifically, the XGBoost model is used with hyperparameters: n_estimators=20000, eta=0.1, max_depth=7, subsample=0.4, colsample_bytree=0.8. After repeating the training process 50 times and taking the mean value of the root mean squared error (RMSE) from the test predictions, our results showed that the persistent Dirac features with XGBoost achieved mean RMSE of 1.64 ± 0.31 kcal/mol. This result is better than conventional methods from MoleculeNet[74] as shown in Fig. 6.

Furthermore, we consider a specially-designed persistent weighted Dirac model. We define the weight of each 0-simplex (or atom) as the magnitude of electrostatic charge of the atom, weight of each 1-simplex (edge) as the Euclidean distance (in Å) between the two connected atoms and the weight of each 2-simplex (triangle) as the area of the triangle (in $Å^3$). The area of the triangle can be approximated using Heron's formula. By generating the weighted Dirac matrices $\overline{\mathbf{D}}_0$ and $\overline{\mathbf{D}}_1$ along the same 120 timesteps, 12 statistical attributes (i–xii) are then computed from the eigenvalues to form our persistent weighted Dirac features. Figure 6 shows that our

persistent weighted Dirac model produced a slightly lower mean RMSE of 1.62 ± 0.26 kcal/mol as compared to the persistent Dirac model. This may suggest that the additional weight information incorporated into the weighted Dirac matrices increases the effectiveness of our approach.

Conclusion

Molecular representations are essential to the modeling and analysis of molecular systems. Motivated by the great success of persistent Hodge Laplacian, we develop the first persistent Dirac-based molecular representation and fingerprint. A rigorous theoretical framework for persistent Dirac is introduced through the commutative diagram of discrete Dirac operator over a filtration process. Moreover, a series of persistent attributes, which characterize the persistence and variations of the eigenspectrum of Dirac matrices, are proposed and further used as molecular fingerprints. The eigenspectrum properties of discrete Dirac matrices have been studied, in particular, the geometric and topological properties of both non-homology and homology eigenvectors. We also consider weighted Dirac model and the influence of weighting schemes on eigenspectrum information. Finally, our persistent Dirac-based models have been used in the clustering of molecular configurations from nine types of organic-inorganic halide perovskites (OIHPs). This work could open new perspectives for the use of persistent Dirac-based molecular fingerprints. We hope that this can inspire future interdisciplinary work between Dirac operators and machine learning along OIHPs or other relevant research directions. An interesting direction for further exploration would be the use of non-symmetric persistent Dirac features in predicting biological, chemical and physical properties in biomolecular data. For instance, further exploration in the use of non-symmetric persistent Dirac features can be considered in the prediction of energy bandgap and other material properties in OIHPs[75].

Data avaliability

The OIHP dataset generated and persistent Dirac codes during and/or analysed during the current study are available in https://github.com/ExpectozJJ/Persistent-Dirac-Models. The 3D coordinates of the molecular structures from FreeSolv database and the solvation free energy values are available in https://weilab.math.msu.edu/DataLibrary/3D/Downloads/FreeSolv.zip.

Received: 20 March 2023; Accepted: 28 June 2023
Published online: 11 July 2023

References

1. Puzyn, T., Leszczynski, J. & Cronin, M. T. *Recent Advances in QSAR Studies: Methods and Applications* Vol. 8 (Springer Science & Business Media, 2010).
2. Lo, Y. C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**(8), 1538–1546 (2018).
3. Wee, J. J. & Xia, K. Ollivier persistent Ricci curvature-based machine learning for the protein–ligand binding affinity prediction. *J. Chem. Inf. Model.* **61**(4), 1617–1626 (2021).
4. Liu, X., Feng, H., Wu, J. & Xia, K. Persistent spectral hypergraph based machine learning (PSH-ML) for protein–ligand binding affinity prediction. *Brief. Bioinform.* **22**(5), bbab127 (2021).
5. Wang, R., Nguyen, D. D. & Wei, G.-W. Persistent spectral graph. *Int. J. Numer. Methods Biomed. Eng.*, e3376 (2020).
6. Wee, J. J. & Xia, K. Forman persistent Ricci curvature (FPRC)-based machine learning models for protein-ligand binding affinity prediction. *Brief. Bioinform.* **22**(6), bbab136 (2021).
7. Chen, D. *et al.* Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* **12**(1), 1–9 (2021).
8. Chen, J., Zhao, R., Tong, Y. & Wei, G.-W. Evolutionary de Rham–Hodge method. *Discrete Contin. Dyn. Syst. Ser. B* **26**(7), 3785 (2021).
9. Wei, R. K. J., Wee, J., Laurent, V. E. & Xia, K. Hodge theory-based biomolecular data analysis. *Sci. Rep.* **12**(1), 1–16 (2022).
10. Meng, Z. Y., Anand, D. V., Lu, Y. P., Wu, J. & Xia, K. L. Weighted persistent homology for biomolecular data analysis. *Sci. Rep.* **10**(1), 1–15 (2020).
11. Anand, D. V., Meng, Z. Y., Xia, K. L. & Mu, Y. G. Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis. *Sci. Rep.* **10**(1), 1–17 (2020).
12. Xia, F. & Lu, L. Y. Multiscale coarse-graining via normal mode analysis. *J. Chem. Theory Comput.* **8**(11), 4797–4806 (2012).
13. Xia, K. L., Zhao, Z. X. & Wei, G. W. Multiresolution persistent homology for excessively large biomolecular datasets. *J. Chem. Phys.* **143**(13), 10B6031 (2015).
14. Nguyen, D. D. & Wei, G. W. AGL-Score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *J. Chem. Inf. Model.* **59**(7), 3291–3304 (2019).
15. Xia, K. L., Opron, K. & Wei, G. W. Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM). *J. Chem. Phys.* **143**(20), 204106 (2015).
16. Xia, K. L. Multiscale virtual particle based elastic network model (MVP-ENM) for normal mode analysis of large-sized biomolecules. *Phys. Chem. Chem. Phys.* **20**(1), 658–669 (2018).
17. Berrone, S., Santa, F. D., Mastropietro, A., Pieraccini, S. & Vaccarino, F. Graph informed deep learning for uncertainty quantification in discrete fracture networks. In *Proceedings of SIMAI 2020+21* (2021).
18. Berrone, S., Santa, F. D., Mastropietro, A., Pieraccini, S. & Vaccarino, F. Graph-informed neural networks for regressions on graph-structured data. *Mathematics* **10**(5), 786 (2022).
19. Bianconi, G. *Multilayer Networks: Structure and Function* (Oxford University Press, 2018).
20. Bianconi, G. *Higher-Order Networks* (Cambridge University Press, 2021).
21. Petri, G., Scolamiero, M., Donato, I. & Vaccarino, F. Topological strata of weighted complex networks. *PLoS One* **8**(6), e66506 (2013).
22. Petri, G., Scolamiero, M., Donato, I. & Vaccarino, F. Networks and cycles: A persistent homology approach to complex networks. In *Proceedings of the European Conference on Complex Systems 2012*, 93–99 (Springer, 2013).
23. Barbensi, A., Yoon, H. R., Madsen, C. D., Ajayi, D. O., Stumpf, M. P. H. & Harrington, H. A. Hypergraphs for multiscale cycles in structured data. arXiv preprint arXiv:2210.07545 (2022).
24. Bick, C., Gross, E., Harrington, H. A. & Schaub, M. T. What are higher-order networks? arXiv preprint arXiv:2104.11329 (2021).

25. Torres, J. J. & Bianconi, G. Simplicial complexes: Higher-order spectral dimension and dynamics. *J. Phys. Complex.* **1**(1), 015002 (2020).
26. Millán, A. P., Torres, J. J. & Bianconi, G. Explosive higher-order Kuramoto dynamics on simplicial complexes. *Phys. Rev. Lett.* **124**(21), 218301 (2020).
27. Ghorbanchian, R., Restrepo, J. G., Torres, J. J. & Bianconi, G. Higher-order simplicial synchronization of coupled topological signals. *Commun. Phys.* **4**(1), 1–13 (2021).
28. Calmon, L., Restrepo, J. G., Torres, J. J. & Bianconi, G. Dirac synchronization is rhythmic and explosive. *Commun. Phys.* **5**(1), 253 (2022).
29. Wu, Z., Menichetti, G., Rahmede, C. & Bianconi, G. Emergent complex network geometry. *Sci. Rep.***5**(1), 1–12 (2015).
30. Bianconi, G. & Rahmede, C. Network geometry with flavor: From complexity to quantum geometry. *Phys. Rev. E* **93**(3), 032315 (2016).
31. Bianconi, G. & Rahmede, C. Emergent hyperbolic network geometry. *Sci. Rep.***7**(1), 1–9 (2017).
32. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002).
33. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).
34. Cang, Z. X. & Wei, G. W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **13**(7), e1005690 (2017).
35. Cang, Z. X. & Wei, G. W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int. J. Numer. Methods Biomed. Eng.*https://doi.org/10.1002/cnm.2914 *(2017)*.
36. Nguyen, D. D., Xiao, T., Wang, M. L. & Wei, G. W. Rigidity strengthening: A mechanism for protein–ligand binding. *J. Chem. Inf. Model.* **57**(7), 1715–1721 (2017).
37. Cang, Z. X. & Wei, G. W. Integration of element specific persistent homology and machine learning for protein–ligand binding affinity prediction. *Int. J. Numer. Methods Biomed. Eng.* **34**(2), e2914 (2018).
38. Meng, Z. & Xia, K. Persistent spectral-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **7**(19), eabc5329 (2021).
39. Liu, X., Wang, X. J., Wu, J. & Xia, K. L. Hypergraph based persistent cohomology (HPC) for molecular representations in drug design. *Brief. Bioinform.* **22**, bba411 (2021).
40. 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).
41. Cang, Z. X., Mu, L. & Wei, G. W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **14**(1), e1005929 (2018).
42. Wu, K. D. & Wei, G. W. Quantitative toxicity prediction using topology based multi-task deep neural networks. *J. Chem. Inf. Model.*https://doi.org/10.1021/acs.jcim.7b00558 *(2018)*.
43. Wang, B., Zhao, Z. X. & Wei, G. W. Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *J. Chem. Phys.* **145**(12), 124110 (2016).
44. Wang, B., Wang, C. Z., Wu, K. D. & Wei, G. W. Breaking the polar-nonpolar division in solvation free energy prediction. *J. Comput. Chem.* **39**(4), 217–233 (2018).
45. Wu, K. D., Zhao, Z. X., Wang, R. X. & Wei, G. W. TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J. Comput. Chem.* **39**(20), 1444–1454 (2018).
46. Zhao, R. D., Cang, Z. X., Tong, Y. Y. & Wei, G. W. Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics* **34**(17), i830–i837 (2018).
47. Nguyen, D. D. *et al.* Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *J. Comput. Aided Mol. Design* **33**(1), 71–82 (2019).
48. Nguyen, D. D., Gao, K. F., Wang, M. L. & Wei, G. W. MathDL: Mathematical deep learning for D3R Grand Challenge 4. *J. Comput. Aided Mol. Design* **34**, 131–147 (2019).
49. Bianconi, G. The topological Dirac equation of networks and simplicial complexes. *J. Phys. Complex.* **2**(3), 035022 (2021).
50. Calmon, L., Schaub, M. T. & Bianconi, G. Dirac signal processing of higher-order topological signals. arXiv preprint arXiv:2301.10137 (2023).
51. Post, O. First order approach and index theorems for discrete and metric graphs. *Ann. Henri Poincaré* **10**, 823–866 (2009).
52. Lloyd, S., Garnerone, S. & Zanardi, P. Quantum algorithms for topological and geometric analysis of data. *Nat. Commun.* **7**(1), 1–7 (2016).
53. Ameneyro, B., Maroulas, V. & Siopsis, G. Quantum persistent homology. arXiv preprint arXiv:2202.12965 (2022).
54. Crane, K., Pinkall, U. & Schröder, P. Spin transformations of discrete surfaces. In *ACM SIGGRAPH 2011 papers*, 1–10 (ACM, 2011).
55. Bianconi, G. Dirac gauge theory for topological spinors in 3 + 1 dimensional networks. *J. Phys. A Math. Theor.* **56**, 275001 (2023).
56. Giambagli, L., Calmon, L., Muolo, R., Carletti, T. & Bianconi, G. Diffusion-driven instability of topological signals coupled by the Dirac operator. *Phys. Rev. E* **106**(6), 064314 (2022).
57. Calmon, L., Krishnagopal, S. & Bianconi, G. Local Dirac synchronization on networks. *Chaos Interdiscip. J. Nonlinear Sci.***33**(3) (2023).
58. Calmon, L., Restrepo, J. G., Torres, J. J. & Bianconi, G. Dirac synchronization is rhythmic and explosive. *Commun. Phys.* **5**(1), 253 (2022).
59. Ameneyro, B., Siopsis, G. & Maroulas, V. Quantum persistent homology for time series. In *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*, 387–392 (IEEE, 2022).
60. Wang, R. *et al.* Hermes: Persistent spectral graph software. *Found. Data Sci.* **3**(1), 67 (2021).
61. Mémoli, F., Wan, Z. & Wang, Y. Persistent Laplacians: Properties, algorithms and implications. *SIAM J. Math. Data Sci.* **4**(2), 858–884 (2022).
62. Wei, X. & Wei, G.-W. Persistent sheaf Laplacians. arXiv preprint arXiv:2112.10906 (2021).
63. Baccini, F., Geraci, F. & Bianconi, G. Weighted simplicial complexes and their representation power of higher-order network data and topology. *Phys. Rev. E* **106**(3), 034319 (2022).
64. Vaccarino, F., Fugacci, U. & Scaramuccia, S. Persistent homology: A topological tool for higher-interaction systems. In *Higher-Order Systems*, 97–139 (Springer, 2022).
65. Horak, D. & Jost, J. Spectra of combinatorial Laplace operators on simplicial complexes. *Adv. Math.* **244**, 303–336 (2013).
66. Zhao, R., Desbrun, M., Wei, G.-W. & Tong, Y. 3D Hodge decompositions of edge-and face-based vector fields. *ACM Trans. Graph. (TOG)* **38**(6), 1–13 (2019).
67. Zhao, R., Wang, M., Chen, J., Tong, Y. & Wei, G.-W. The de Rham–Hodge analysis and modeling of biomolecules. *Bull. Math. Biol.* **82**(8), 1–38 (2020).
68. Wu, C. Y., Ren, S. Q., Wu, J. & Xia, K. L. Weighted (co)homology and weighted Laplacian. *Sci. China Math.* (2018).
69. Knill, O. The Dirac operator of a graph. arXiv preprint arXiv:1306.2166 (2013).
70. Knill, O. The McKean–Singer formula in graph theory. arXiv preprint arXiv:1301.1408 (2013).
71. Wee, J. J. & Xia, K. Persistent spectral based ensemble learning (PerSpect-EL) for protein-protein binding affinity prediction. *Brief. Bioinform.* **23**(2), bbac024 (2022).

72. Mobley, D. L. & Guthrie, J. P. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Design* **28**, 711–720 (2014).
73. Mobley, D. L., Wymer, K. L., Lim, N. M. & Guthrie, J. P. Blind prediction of solvation free energies from the SAMPL4 challenge. *J. Comput. Aided Mol. Design* **28**, 135–150 (2014).
74. Wu, Z. *et al.* MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9**(2), 513–530 (2018).
75. Anand, D. V., Xu, Q., Wee, J. J., Xia, K. & Sum, T. C. Topological feature engineering for machine learning based halide perovskite materials design. *npj Comput. Mater.* **8**(1), 203 (2022).

Acknowledgements

This work was supported in part by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG109/19 and Tier 2 MOE2018-T2-1-033.

Author contributions

K.X. and G.B. designed the research. K.X. and J.W. performed research, analyzed data and wrote the initial draft of the paper. All authors revised the paper subsequently.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37853-z.

Correspondence and requests for materials should be addressed to J.W.

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.