

Persistent de Rham-Hodge Laplacians in Eulerian representation for manifold topological learning

Zhe Su¹, Yiying Tong², *and Guo-Wei Wei^{1,3,4} †

¹Department of Mathematics,

Michigan State University, East Lansing, MI 48824, USA

²Department of Computer Science and Engineering,

Michigan State University, East Lansing, MI 48824, USA

³Department of Biochemistry and Molecular Biology,

Michigan State University, East Lansing, MI 48824, USA

⁴Department of Electrical and Computer Engineering,

Michigan State University, East Lansing, MI 48824, USA

Abstract

Recently, topological data analysis has become a trending topic in data science and engineering. However, the key technique of topological data analysis, i.e., persistent homology, is defined on point cloud data, which does not work directly for data on manifolds. Although earlier evolutionary de Rham-Hodge theory deals with data on manifolds, it is inconvenient for machine learning applications because of the numerical inconsistency caused by remeshing the involving manifolds in the Lagrangian representation. In this work, we introduce persistent de Rham-Hodge Laplacian, or persistent Hodge Laplacian (PHL) as an abbreviation, for manifold topological learning. Our PHLs are constructed in the Eulerian representation via structure-persevering Cartesian grids, avoiding the numerical inconsistency over the multiscale manifolds. To facilitate the manifold topological learning, we propose a persistent Hodge Laplacian learning algorithm for data on manifolds or volumetric data. As a proof-of-principle application of the proposed manifold topological learning model, we consider the prediction of protein-ligand binding affinities with two benchmark datasets. Our numerical experiments highlight the power and promise of the proposed method.

*Corresponding author. Email: ytong@msu.edu

†Corresponding author. Email: weig@msu.edu

1 Introduction

Recent years have witnessed a fast growth of topological data analysis (TDA) in data science and engineering [1]. The growth is driven by the great promise of topological approaches to real-world data that are distinguished from any other statistical, mathematical, physical, and engineering methods [2, 3]. Typically, TDA offers a multi-scale topological characterization of data, which is the case with persistent homology [4, 5], a key method employed in TDA. A major feature of persistent homology is its multi-scale analysis, which creates a family of topological spaces from the original data to track the topological persistence, i.e., the lifespan of topological invariants across scales [6, 7]. The other major feature of persistent homology is its topological description of a space (like connected components, loops, and voids) in terms of topological invariants, such as Betti numbers. As such, persistent homology-based TDA leads to much topological simplification of the geometric information in the data [8, 9]. Consequently, TDA typically works extremely well for data with intricate complexity [10, 11]. Unfortunately, for data without geometric complexes, TDA may give rise to an oversimplification of key geometric characteristics, leading to a less competitive approach.

For many years, persistent homology has been used in qualitative analysis, which is somewhat counterintuitive and unproductive for nonexperts. The power of persistent homology was not demonstrated until it was utilized in quantitative and predictive analysis via machine learning algorithms [12, 13]. Topological deep learning (TDL), coined in 2017 [14], was introduced to deal with large and intrinsically complex datasets using both persistent homology and deep neural networks. More recently, simplicial neural networks and other topological neural techniques have been applied in TDL to the design of neural network architecture. TDL has become an emerging paradigm in data science and machine learning [15]. However, an increasing concern associated with this rising popularity is whether TDL brings any practical benefit beyond its mathematical elegance. There are many applications where TDL has demonstrated superiority to other competitive methods [16]. Perhaps some of the most compelling examples are TDL’s dominant wining of D3R Grand Challenges, an annual worldwide competition series in computer-aided drug design [17, 18], its discovery of the mechanisms of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) evolution [19, 20], and its successful forecast of emerging dominant SARS-CoV-2 variants BA.2 [21] and BA.4/BA.5 about two months in advance [22].

It is interesting to understand why TDL (or TDA) was so successful in the aforementioned examples, but was not competitive in many other situations in the literature [23]. First, biomolecular data, which is intricately complex in their internal structures[10], was involved in the above compelling examples. As such, topological simplification was a productive process, whereas TDL leads to the severe loss of crucial geometric information in many other data that is relatively simple in their internal structures. Additionally, it was element-specific persistent homology, rather than the plain persistent homology, that was applied in the above examples. This approach captures physical and biological interactions in the biomolecular data [14]. In fact, in the forecast of emerging dominant SARS-CoV-2 variants BA.4/BA.5, persistent

Laplacian, rather than persistent homology, was utilized. This happens because persistent homology has many drawbacks or limitations [24]. First, the topological invariant extracted from persistent homology is qualitative, rather than quantitative. For example, the barcode from persistent homology does not distinguish a five-number from a six-number ring. Additionally, persistent homology is incapable of dealing with different elements in a point cloud, which is ineffective with the physics and chemistry of (bio)molecular data. Moreover, persistent homology cannot describe non-topological changes, i.e., homotopic shape evolution during the multi-scale (or filtration) analysis. Further, persistent homology is incapable of handling directed networks and digraphs, such as polarization, regulation, and control issues in applications. Finally, persistent homology is unable to characterize structured data, e.g., hypergraphs, directed networks, etc. These challenges call for innovative new topological methods.

To address these challenges, the persistent spectral graph, also known as persistent combinatorial Laplacian or persistent Laplacian (PL), was introduced in 2019 [25]. The harmonic spectra of PLs fully recover the topological invariants of persistent homology. However, the nonharmonic spectra of PLs capture the homotopic shape evolution during the multi-scale analysis that cannot be observed with persistent homology. Computational algorithms [26, 27] and mathematical analysis [28, 29] of PLs have been reported. In the past few years, much effort has been given to extend persistent Laplacian to further address other limitations of persistent homology [30], leading to persistent sheaf Laplacians [31], persistent path Laplacians, persistent hypergraph and hyperdigraph Laplacians [32], persistent directed flag Laplacians, persistent Mayer Laplacians, and persistent interaction Laplacians [24]. PLs have been shown to outperform persistent homology in many applications [22, 33].

However, defined on point cloud data, neither persistent homology nor PL can directly deal with two other commonly occurring data formats, namely, data on manifolds [34], such as electron density [35], cryogenic electron microscopy density, and computed tomography images [36], and curves embedded in the three-dimensional Euclidean space, such as knots, links, and tangles, and their generalizations [37, 38]. Multi-scale Gauss link integral [39] and evolutionary Khovanov homology have been proposed to deal with embedded curve data [40]. Evolutionary Khovanov homology integrates algebraic topology, geometric topology, and metric analysis for the first time. However, effective computational algorithms are needed for this approach to be widely used in practical applications.

To carry out manifold topological analysis of data on manifolds, the evolutionary de Rham-Hodge method was introduced [34]. This approach creates a family of multi-scale manifolds with boundaries from a given data and then builds evolutionary Hodge Laplacian operators on the multi-scale manifolds with appropriate boundary conditions. While originated from sharply different topological spaces, evolutionary Hodge Laplacian and PLs share the same algebraic structure and capture topological invariants in their harmonic spectra [41]. Case studies have been given to demonstrate evolutionary de Rham-Hodge theory-based manifold topological analysis of data on manifolds [34]. However, this approach was based on discrete exterior calculus [42, 43] or finite element exterior calculus [44] in the Lagrangian representation, which is not

efficient for multi-scale analysis and machine learning studies. Specifically, the regeneration of the evolving manifolds at different scales with different Lagrangian meshes causes numerical inconsistencies and becomes expensive for practical applications in machine learning studies. This challenge calls for new effective manifold topological analysis approaches for data on manifolds.

The objective of this work is to develop a persistent de Rham-Hodge theory on the Euler representation for manifold topological learning (MTL). To this end, we solve Hodge Laplacians on a pre-designed structure-persevering Cartesian grid for all scales to avoid numerical inconsistency. We construct a natural mapping of differential forms from a manifold with boundary embedded in \mathbb{R}^3 to a large manifold, use it to produce persistent cohomology mapping, and construct a persistent Hodge Laplacian with built-in boundary conditions. Our new approach draws on differential geometry, algebraic topology, partial differential equations, metric analysis, and numerical analysis. To give a proof-of-principle demonstration, we pair the proposed persistent de Rham-Hodge Laplacians with an effective machine learning algorithm to predict protein-ligand binding affinities. Based on two benchmark datasets in the Protein Data Bank (PDB), PDBbind v2007 and PDBbind v2016, we show that our MTL model gives rise to cutting-edge performance.

The rest of this paper is organized as follows: Section 2 offers a primer on the de Rham-Hodge theory on manifolds with boundaries; Section 3 presents our discretization for evolutionary de Rham-Hodge theory based on spectrum calculation of Laplacians associated with sublevel sets on Cartesian grids; Section 4 presents our construction for persistent de Rham-Hodge Laplacians both in the continuous setting and for given level set functions on Cartesian grids; Section 5 showcases preliminary studies on the applications of MTL; and Section 6 concludes the paper.

2 De Rham-Hodge Theory

Let M be an m -dimensional smooth, orientable, compact Riemannian manifold with boundary. Denote by $\Omega^k(M)$ the space of all differential k -forms on M , i.e., the space of all smooth antisymmetric covariant tensor fields on M of degree k . The *differential* d , also called exterior derivative, is the unique \mathbb{R} -linear mapping from the space of k -forms $\Omega^k(M)$ to the space of $(k+1)$ -forms $\Omega^{k+1}(M)$ satisfying the Leibniz rule with respect to the wedge product \wedge and the nilpotent property $dd = 0$. A key property of differential forms is that they can be integrated over any orientable k -submanifolds of M . For any oriented $(k+1)$ -submanifold $S \subset M$ with boundary ∂S , Stokes' theorem, as a generalization of the Newton-Leibniz rule, states that the integral of a differential k -form ω over ∂S is equal to the integral of its differential over S , i.e.,

$$\int_S d\omega = \int_{\partial S} \omega. \quad (1)$$

The differential d generalizes and unifies the classical operators in vector calculus, such as gradient ∇ , curl $\nabla \times$, and divergence $\nabla \cdot$ in \mathbb{R}^2 and \mathbb{R}^3 . For instance, in \mathbb{R}^3 , 0-forms and 3-forms can be identified with scalar fields, while 1-forms and 2-forms can be identified with vector fields. In this case, the differential d corresponds to the

gradient operator ∇ when applied to 0-forms, the curl operator $\nabla \times$ when applied to 1-forms, or the divergence operator $\nabla \cdot$ when applied 2-forms. The nilpotent property $dd = 0$ directly leads to the vector field analysis identities $\nabla \times \nabla = 0$ and $\nabla \cdot \nabla \times = 0$.

A differential form $\omega \in \Omega^k(M)$ is called *closed* if $d\omega = 0$, or *exact* if there is a $(k-1)$ -form $\zeta \in \Omega^{k-1}(M)$ such that $\omega = d\zeta$. Due to the property $dd = 0$, every exact form is closed. Thus, the differential d links the sequence of the spaces of differential forms on M into a chain complex

$$0 \longrightarrow \Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \cdots \xrightarrow{d} \Omega^{m-1}(M) \xrightarrow{d} \Omega^m(M) \rightarrow 0. \quad (2)$$

The k -th *de Rham cohomology* group, denoted by $H_{dR}^k(M)$, is then defined to be the k -th homology of this chain complex, i.e., the quotient space of closed k -forms modulo the space of exact k -forms, i.e.,

$$H_{dR}^k(M) = \frac{\ker(d : \Omega^k(M) \rightarrow \Omega^{k+1}(M))}{\text{im}(d : \Omega^{k-1}(M) \rightarrow \Omega^k(M))}. \quad (3)$$

The de Rham cohomology, by the de Rham theorem, is naturally isomorphic to the singular cohomology, and thus depends only on the manifold topology.

Let g be a Riemannian metric on M and $\langle \cdot, \cdot \rangle_g$ be the pointwise inner product induced by g on $\Omega^k(M)$. The *Hodge star* operator \star provides an isomorphism from the space of differential k -forms $\Omega^k(M)$ to the space of $(m-k)$ -forms $\Omega^{m-k}(M)$, defined by the following formula

$$\omega \wedge \star \eta = \langle \omega, \eta \rangle_g \mu_g, \quad (4)$$

where μ_g is the volume form on M induced by g . The Hodge L^2 -inner product on the space of k -forms $\Omega^k(M)$ can then be obtained by taking the integral of the formula (4)

$$(\omega, \eta) = \int_M \omega \wedge \star \eta. \quad (5)$$

The *codifferential* $\delta : \Omega^k(M) \rightarrow \Omega^{k-1}(M)$ is defined by

$$\delta = (-1)^{m(k-1)+1} \star d \star, \quad (6)$$

which also has the nilpotent property $\delta\delta = 0$. We call a differential form $\omega \in \Omega^k(M)$ *coclosed* if $\delta\omega = 0$, or *coexact* if there is a $(k+1)$ -form $\eta \in \Omega^{k+1}(M)$ such that $\omega = \delta\eta$. The codifferential δ , as the differential d , also extends the classical gradient, curl and divergence in vector calculus. In \mathbb{R}^3 , it corresponds to $-\nabla \cdot$, $\nabla \times$ and $-\nabla$ when applied to 1-forms, 2-forms and 3-forms, respectively.

The *Hodge Laplacian* for differential forms is defined as $\Delta = d\delta + \delta d : \Omega^k(M) \rightarrow \Omega^k(M)$. Its kernel, consisting of all differential k -forms ω on M with $\Delta\omega = 0$, is called the space of *harmonic k -forms*. We denote by $\mathcal{H}_\Delta^k(M)$ the space of harmonic k -forms and by $\mathcal{H}^k(M)$ the space of k -forms that are both closed and coclosed, i.e., $\mathcal{H}^k(M) = \ker d \cap \ker \delta$. The latter space $\mathcal{H}^k(M)$, known as the space of harmonic k -fields, is in general only a subset of the space of harmonic forms $\mathcal{H}^k(M) \subset \mathcal{H}_\Delta^k(M)$, and

is infinite-dimensional [45]. However, in the case of closed manifolds where $\partial M = \emptyset$, the space of harmonic forms $\mathcal{H}_\Delta^k(M)$ reduces to the space $\mathcal{H}^k(M)$, as any harmonic form is both closed and coclosed. The result follows directly from the following formula

$$0 = (\Delta\omega, \omega) = ((d\delta + \delta d)\omega, \omega) = (d\omega, d\omega) + (\delta\omega, \delta\omega), \quad (7)$$

due to the L^2 -adjointness of the codifferential δ and the differential d on closed manifolds, i.e., $(d\omega, \eta) = (\omega, \delta\eta)$.

The classical Hodge decomposition theorem for closed manifolds states that the space of differential k -forms $\Omega^k(M)$ can be decomposed as

$$\Omega^k(M) = d\Omega^{k-1}(M) \oplus \delta\Omega^{k+1}(M) \oplus \mathcal{H}_\Delta^k(M). \quad (8)$$

These three subspaces are mutually orthogonal with respect to the inner product (5). Moreover, Hodge theorem identifies the harmonic space $\mathcal{H}_\Delta^k(M)$ with the k -th de Rham cohomology group $H_{dR}^k(M)$, which states that each harmonic form corresponds to exactly one equivalence class in $H_{dR}^k(M)$. Therefore, the harmonic space $\mathcal{H}_\Delta^k(M)$ is fully determined by the manifold topology, and is finite-dimensional with its dimension given by the Betti number $\dim \mathcal{H}_\Delta^k(M) = \beta_k$.

2.1 Hodge decomposition for manifolds with boundary

In the presence of a non-empty boundary ∂M , the two operators d and δ are not L^2 -adjoint, as integration by parts leads to

$$(d\omega, \eta) = (\omega, \delta\eta) + \int_{\partial M} \omega \wedge \star\eta, \quad (9)$$

which contains a boundary term that may not vanish, and thus the decomposed subspaces in (8) are not orthogonal. However, certain boundary conditions can be enforced, ensuring the adjointness of the differential d and the codifferential δ , thereby inducing an orthogonal decomposition of the space of differential forms.

The most common choices of boundary conditions ensuring the adjointness of d and δ are the normal (Dirichlet) and tangential (Neumann) boundary conditions. A differential form $\omega \in \Omega^k(M)$ is called *normal* (Dirichlet) if it gives zero when applied to tangent vectors of the boundary, or *tangential* (Neumann) if the same holds for its dual $\star\omega$ instead. Denote by $\Omega_n^k(M)$ the set of normal differential k -forms and by $\Omega_t^k(M)$ the set of tangential differential forms, i.e.,

$$\Omega_n^k(M) = \{\omega \in \Omega^k(M) \mid \omega|_{\partial M} = 0\} \quad (10)$$

$$\Omega_t^k(M) = \{\omega \in \Omega^k(M) \mid \star\omega|_{\partial M} = 0\}. \quad (11)$$

Following their definitions, the spaces $\Omega_n^k(M)$ and $\Omega_t^{m-k}(M)$ are isomorphic under the Hodge star operator \star , also known as the Hodge duality. Moreover, the differential d preserves the normal boundary conditions, while the codifferential δ preserves the tangential boundary conditions.

The Hodge-Morrey decomposition [46] states that there is a 3-component L^2 -orthogonal decomposition

$$\Omega^k(M) = d\Omega_n^{k-1}(M) \oplus \delta\Omega_t^{k+1}(M) \oplus \mathcal{H}^k(M), \quad (12)$$

The orthogonality of the decomposition directly comes from the adjointness of δ and d when enforcing the normal or tangential boundary conditions. For $\omega \in \Omega^k(M)$, there is a unique decomposition of ω given as follows:

$$\omega = d\alpha_n + \delta\beta_t + \eta, \quad (13)$$

where $\alpha_n \in \Omega_n^{k-1}(M)$, $\beta_t \in \Omega_t^{k+1}(M)$, and $\eta \in \mathcal{H}^k(M)$. Note that the potentials α_n and β_t are not uniquely determined as all $\alpha_n + d\eta$ and $\beta_t + \delta\gamma$ with any $\eta \in \Omega_n^{k-2}(M)$ and $\gamma \in \Omega_t^{k+2}(M)$ serve as potentials for the same components. However, the issue can be addressed by enforcing *gauge* conditions, such as

$$\delta\alpha_n = 0, \quad (14)$$

$$d\beta_t = 0. \quad (15)$$

The potentials α_n and β_t can then be uniquely determined by the following equations

$$\begin{cases} \Delta\alpha_n = \delta\omega \\ \Delta\beta_t = d\omega, \end{cases} \quad (16)$$

by resolving the (finite) rank deficiencies of Δ under these boundary conditions.

Remark 1 In the case that M is a closed manifold, i.e., $\partial M = \emptyset$, both the spaces $\Omega_n^k(M)$ and $\Omega_t^k(M)$ coincide with the space of differential forms $\Omega^k(M)$, and the space of harmonic fields is identical to the space of harmonic forms $\mathcal{H}^k(M) = \mathcal{H}_\Delta^k(M)$. The Hodge decomposition (12) then reduces to the classical Hodge decomposition (8) for closed manifolds.

Remark 2 The Hodge-Morrey decomposition (12) in the low dimensional Euclidean spaces \mathbb{R}^2 and \mathbb{R}^3 , often referred to as the Helmholtz-Hodge decomposition in vector calculus, states that any vector field \mathbf{v} defined on a compact domain can be orthogonality decomposed as

$$\mathbf{v} = \nabla f + \nabla \times \mathbf{u} + \mathbf{h}, \quad (17)$$

where f is a scalar potential that vanishes on the boundary of the domain, \mathbf{u} is a vector field orthogonal to the boundary, and \mathbf{h} is the harmonic vector field satisfying $\nabla \times \mathbf{h} = 0$ and $\nabla \cdot \mathbf{h} = 0$. The first component ∇f and the second component $\nabla \times \mathbf{u}$ are often called the curl-free and divergence-free parts of the vector field \mathbf{v} respectively. Note that in the presence of a boundary, the resulting scalar potential f is also called satisfying the normal boundary of 0-forms, and the vector field \mathbf{u} is called satisfying the tangential boundary condition of 2-forms, which are direct counterparts of the potentials α_n and β_t in (12). For a complete correspondence between scalar or vector fields, and differential forms under the normal and tangential boundary conditions, see [47].

The space of harmonic fields \mathcal{H}^k , in general, is infinite-dimensional, and thus has no direct correspondence with the cohomology of the manifold. However, as noted by [47], one can restrict to the space of normal harmonic fields, namely $\mathcal{H}_n^k(M) = \mathcal{H}^k(M) \cap \Omega_n^k(M)$, and the space of tangential harmonic fields, $\mathcal{H}_t^k(M) = \mathcal{H}^k(M) \cap \Omega_t^k(M)$. As a consequence of the de Rham map, these two subspaces $\mathcal{H}_n^k(M)$ and $\mathcal{H}_t^k(M)$ are fully determined by the topology of M : the space of normal harmonic fields $\mathcal{H}_n^k(M)$ is isomorphic to the relative de Rham cohomology $H_{dR}^k(M, \partial M)$, while the space of tangential harmonic fields $\mathcal{H}_t^k(M)$ is isomorphic to the absolute de Rham cohomology $H_{dR}^k(M)$ [48]. The two subspaces $\mathcal{H}_n^k(M)$ and $\mathcal{H}_t^k(M)$ are thus finite-dimensional, with dimensions given by the Betti numbers: $\dim \mathcal{H}_n^k(M) = \beta_{m-k}$ and $\dim \mathcal{H}_t^k(M) = \beta_k$. Furthermore, the kernels of the Hodge Laplacian Δ , when restricted to the space of normal forms $\Omega_n^k(M)$ and the space of tangential forms $\Omega_t^k(M)$ with gauge conditions on the boundary, can be identified to the space of normal harmonic fields and the space of tangential harmonic fields, respectively. Denote by Δ_n and Δ_t the restrictions of the Hodge Laplacian Δ on the space of normal fields $\Omega_n^k(M)$ satisfying Eq. (14) and the space of tangential fields $\Omega_t^k(M)$ satisfying Eq. (15), i.e., $\Delta_n : \Omega_n^k(M) \rightarrow \Omega^k(M)$ and $\Delta_t : \Omega_t^k(M) \rightarrow \Omega^k(M)$. Then immediately we have $\ker \Delta_n = \mathcal{H}^k(M) \cap \Omega_n^k(M) = \mathcal{H}_n^k(M)$ and $\ker \Delta_t = \mathcal{H}^k(M) \cap \Omega_t^k(M) = \mathcal{H}_t^k(M)$. The result follows directly from Eq. (7). These identifications, finally, enable us to study the topology of the underlying manifold M through the Hodge Laplacians on normal and tangential forms.

Remark 3 In fact, let $\mathcal{H}_{co}^k = \mathcal{H}^k(M) \cap \delta \Omega^{k+1}(M)$ and $\mathcal{H}_{ex}^k = \mathcal{H}^k(M) \cap d \Omega^{k-1}(M)$. The space of harmonic fields $\mathcal{H}^k(M)$ can be further orthogonally decomposed for smooth manifolds

$$\mathcal{H}^k(M) = \mathcal{H}_{co}^k(M) \oplus \mathcal{H}_n^k(M) \quad (18)$$

$$= \mathcal{H}_{ex}^k(M) \oplus \mathcal{H}_t^k(M), \quad (19)$$

which results in the Hodge-Morrey-Friedrichs decomposition given as follows

$$\Omega^k(M) = d\Omega_n^{k-1}(M) \oplus \delta\Omega_t^{k+1}(M) \oplus \mathcal{H}_{co}^k(M) \oplus \mathcal{H}_n^k(M) \quad (20)$$

$$= d\Omega_n^{k-1}(M) \oplus \delta\Omega_t^{k+1}(M) \oplus \mathcal{H}_{ex}^k(M) \oplus \mathcal{H}_t^k(M). \quad (21)$$

In particular, if M is a compact domain in Euclidean spaces, then there is a unique orthogonal 5-component decomposition

$$\Omega^k(M) = d\Omega_n^{k-1}(M) \oplus \delta\Omega_t^{k+1}(M) \oplus \mathcal{H}_n^k(M) \oplus \mathcal{H}_t^k(M) \oplus (d\Omega^{k-1}(M) \cap \delta\Omega^{k+1}(M)), \quad (22)$$

as the spaces $\mathcal{H}_n^k(M)$ and $\mathcal{H}_t^k(M)$ are L^2 -orthogonal, instead of just being transversal for compact manifolds in general [49]. Due to the correspondence between differential forms and vector fields in the low-dimensional Euclidean spaces, the implementation of this 5-component Hodge decomposition has been applied and implemented to the study of vector fields for surface triangle meshes, for tetrahedral meshes [47] and for regular Cartesian grids [50].

As we mainly focus on applications of compact domains in \mathbb{R}^3 , to study the geometric and topological information of the underlying manifolds, there are eight Laplacians to be considered, which are defined on the spaces of differential k -forms with $k = 0, 1, 2, 3$ satisfying either the normal or the tangential boundary conditions. However, thanks to the duality between the space of normal fields and tangential fields, the study of the spectra of these eight Laplacians reduces to that of four Laplacians

on one of the two types of boundary conditions, and finally to the singular spectra of three differential operators, applied to differential forms of degree $k = 0, 1, 2, 3$ [34]. Further details will be discussed in the next section for the discretization of Laplacians.

3 Discretization and construction of Laplacians

In this section, we elaborate on the discretization of the Hodge Laplacian and introduce the Boundary-Induced Graph (BIG) Laplacian for compact domains in low-dimensional Euclidean spaces. Although the theory works for 2D compact domains, we focus only, for the remainder of the paper, compact domains in \mathbb{R}^3 , as we target mainly 3D applications. We use DEC to discretize all differential operators and differential forms on regular Cartesian grids, as it allows for efficient and accurate numerical algorithms relying on just matrix algebra, while keeping the L^2 orthogonality between different components in Hodge decomposition. In addition, the constructed discrete differential operators and differential forms in DEC approximate their smooth analogs. For the characterization of the underlying manifold, we choose the Eulerian formulation, where the manifold is given as a sublevel set of a level set function defined on a regular Cartesian grid. Another common way, called the Lagrangian formulation, discretizes the manifold as simplicial meshes, i.e., triangular or tetrahedral meshes in 2D or 3D. The spectrum analysis of the Hodge Laplacians has been discussed in [47] for the Lagrangian formulation and in [50] for the Eulerian formulation. Compared to the Lagrangian case, the Eulerian representation uses vertices, edges, faces and cells all fixed in a Cartesian grid, which significantly simplifies the data structures and algorithms. The Hodge stars, in the latter case, are close to rescaled identity matrices. This fact simplifies the study of Hodge Laplacians to that of BIG Laplacians with no Hodge stars involved, and thus leads to algorithms with efficient computations.

3.1 Discretization on entire grid

Denote by I_m a rectangular m -dimensional regular Cartesian grid with k -cells oriented according to their alignments with the coordinate axes. The entire grid I_m can be treated as a cell complex tessellating a rectangular domain in \mathbb{R}^m , where each k -cell is a k -dimensional hypercube with edge length ℓ . A continuous differential k -form ω on I_m , following the de Rham map, can be discretized by its integral value over each oriented k -cell σ_i , given as $W^i = \int_{\sigma_i} \omega$ [42]. The discrete differential on discrete k -forms of the grid I_m is then encoded by a sparse matrix D_k^I , which stores the signed incidence between $(k+1)$ -cells and k -cells and is given as the transpose of the cell boundary operator ∂_{k+1}^T on $(k+1)$ -cells following from Stokes' theorem $\int_{\sigma} d\omega = \int_{\partial\sigma} \omega$. An illustration of the chain complex formed by boundary operator ∂ for a simple grid complex with a single 2D cell can be seen in Fig. 1, which is a straightforward generalization of the chain complex on simplicial complexes. Note that the boundary of the boundary of a cell always results in a 0 chain, i.e., $\partial\partial = 0$, whose transpose immediately produces $D_{k+1}^I D_k^I = 0$, thus preserving the nilpotent property in the continuous setting.

The discrete Hodge star establishes a one-to-one correspondence between discrete k -forms on the primal grid I_m and discrete $(m-k)$ -forms on its dual grid, given as

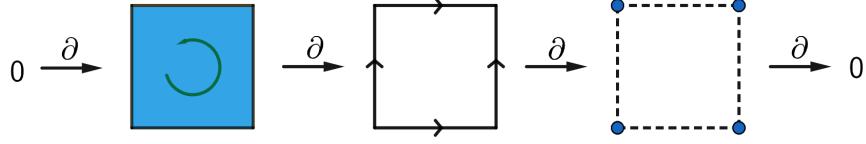


Fig. 1 The chain complex of a single-cell grid formed by the boundary operator: from the face, to its edges, and to their vertices.

the translated grid with grid points located at the m -cell centers of I_m , based on the following formula

$$\frac{1}{|\sigma_k|} \int_{\sigma_k} \omega \approx \frac{1}{|\star \sigma_k|} \int_{\star \sigma_k} \star \omega, \quad (23)$$

where $\star \sigma_k$ is the dual $(m-k)$ -cell formed by the dual grid points located at the centers of the primal m -cells incident to σ_k . See Fig. 2 for an illustration for the correspondences between the primal and dual cells in the Cartesian grid case. Following from the discretization of differential forms, this correspondence leads to a diagonal matrix S_k^I with diagonal entries given by the ratio between the volumes of the dual $(m-k)$ -cells and the primal k -cells, $\ell^{m-k}/\ell^k = \ell^{m-2k}$. The associated discrete Hodge L^2 -inner product (5) of two discrete k -forms V_k and W_k on grid I_m is then given by

$$(V_k, W_k)^I = V_k^T S_k^I W_k. \quad (24)$$

The discrete codifferential, by definition of its smooth counterpart (6), can be assembled from the discrete differential and Hodge star operators as $\delta_k^I = (S_{k-1}^I)^{-1} D_{k-1}^I S_k^I$. Note that the discrete counterpart of the Hodge Laplacian $\Delta = d\delta + \delta d$ by replacing the differential and codifferential operators results in a nonsymmetric matrix. Instead, we consider the counterpart of $\star \Delta$ as the discrete Hodge Laplacian given by

$$L_k^I = (D_k^I)^T S_{k+1}^I D_k^I + S_k^I D_{k-1}^I (S_{k-1}^I)^{-1} (D_{k-1}^I)^T S_k^I, \quad (25)$$

where the operators are considered to be null for $k < 0$ or $k > m$.

3.2 Discrete differential forms and operators on M

Compared to the case of simplicial or polygonal meshes, where the projection matrices to the interior can be straightforward to implement with the boundary elements explicitly labeled, modeling the manifold M as the volume bounded by a level set surface leads to delicate computation of the projection matrices. Note that the boundary of M using grid representation typically intersects with boundary k -cells instead of being its supersets. We restrict the computation to relevant cells by implementing the two types of boundary conditions through the inclusion or exclusion of the entire k -cells. We use the strategy as in [50] for the computation of projection matrices for each type of boundary condition: for the normal boundary condition, we include all cells if at least one of its vertices is inside or on the boundary of M , while for the

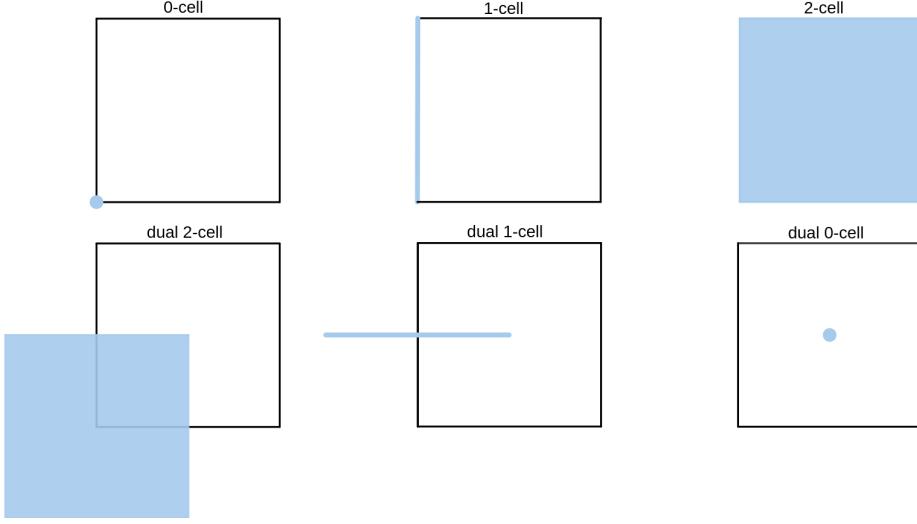


Fig. 2 An example of the primal and dual grid cells for the 2D case. The top row highlights the primal cells, and the bottom row presents their corresponding dual cells.

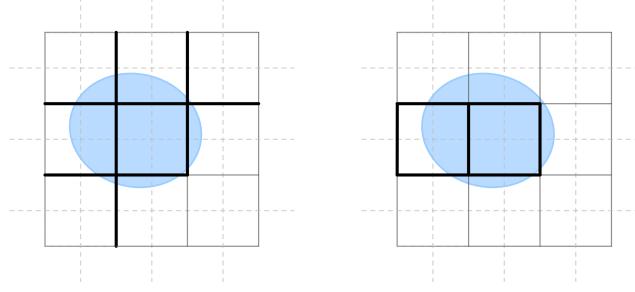


Fig. 3 Distinction of normal supports (left) and tangential supports (right) for primal 1-forms in a 2D Cartesian grid.

tangential boundary condition, we include all cells with at least one of the vertices of the corresponding dual cells is inside or on the boundary. We refer to the former set of cells as the normal support and the latter as the tangential support. In contrast to the mesh case, it is important to note that neither the normal nor the tangential support is necessarily a superset of the other. See Fig. 3 for one example showing the distinction of these two supports for 1-forms.

In the computation of the discrete Hodge star operators, it is essential to consider and incorporate the boundary conditions. Following the procedure in [50], we keep the dual cell volumes and adjust the primal cell volumes for normal boundary conditions, and do conversely for tangential boundary conditions with the primal cell volumes kept and the dual cell volumes changed. To be specific, when dealing with normal (resp. tangential) boundary conditions, we only compute the volume of the region of the

primal (resp. dual) k -cells within the boundary ∂M for the denominator (resp. numerator) of the ratio in the discrete Hodge star matrix, and leave the dual (resp. primal) cell volumes in the numerator (resp. denominator) unchanged. Each unaltered k -cell has a k -volume of ℓ^k . In addition, For numerical stability, we do not alter the volume of outside primal k -cells, and perturb the level set function evaluated at primal/dual grid points to have an absolute value above $\epsilon = 10^{-5}\ell$, which ensures well-behaved fractional k -volumes. We denote by $S_{k,n}^I$ and $S_{k,t}^I$ the diagonal Hodge star matrices defined on the entire grid I^m corresponding to the normal and tangential boundary conditions, respectively.

The projection matrix to the corresponding support, for each type of boundary condition, can be constructed from the identity matrices by eliminating the rows corresponding to k -cells outside the support. Denote by $P_{k,n}$ the projection matrix for k -cells onto the normal support and by $P_{k,t}$ the one onto the tangential support. We then obtain a new set of differential and Hodge star operators for M :

$$D_{k,n} = P_{k+1,n} D_k P_{k,n}^T, \quad S_{k,n} = P_{k,n} S_{k,n}^I P_{k,n}^T \quad (26)$$

$$D_{k,t} = P_{k+1,t} D_k P_{k,t}^T, \quad S_{k,t} = P_{k,t} S_{k,t}^I P_{k,t}^T \quad (27)$$

The nilpotent property $D_{k+1,n} D_{k,n} = 0$ and $D_{k+1,t} D_{k,t} = 0$ still holds for both boundary conditions due to $D_{k+1}^I D_k^I = 0$ and the following observations

$$P_{k+1,n}^T P_{k+1,n} D_k^I P_{k,n}^T = D_k^I P_{k,n}^T, \quad P_{k+1,t} D_k^I P_{k,t}^T P_{k,t} = P_{k+1,t} D_k^I. \quad (28)$$

The discrete Hodge L^2 -inner products of the two types of discrete k -forms on the manifold M for these two boundary conditions are then given by

$$(\xi^k, \zeta^k)^n = (\xi^k)^T S_{k,n} \zeta^k \quad (29)$$

$$(\xi^k, \zeta^k)^t = (\xi^k)^T S_{k,t} \zeta^k, \quad (30)$$

whose domains are the discrete $\Omega_n^k(M)$ and the discrete $\Omega_t^k(M)$ respectively. Finally, we assemble the two types of discrete Hodge Laplacians as in the mesh case:

$$L_{k,n} = D_{k,n}^T S_{k+1,n} D_{k,n} + S_{k,n} D_{k-1,n} S_{k-1,n}^{-1} D_{k-1,n}^T S_{k,n} \quad (31)$$

$$L_{k,t} = D_{k,t}^T S_{k+1,t} D_{k,t} + S_{k,t} D_{k-1,t} S_{k-1,t}^{-1} D_{k-1,t}^T S_{k,t}. \quad (32)$$

The null spaces of these discrete Hodge Laplacians, as in the continuous case, are fully determined by the topology of the underlying manifold M , since they only depend on the differential and projection matrices. The dimension of the kernel of $L_{k,n}$ is given by the Betti number β_{m-k} , while the dimension of the kernel of $L_{k,t}$ is given by β_k . Here the Betti number β_k presents directly the number of k -dimensional holes on the manifold M . For instance, β_0 gives the number of connected components, β_1 gives the number of tunnels, and β_2 provides the number of closed cavities, respectively. The spectra of these Laplacians, in addition, could be used to study the geometric information of the manifold. It is known that the non-zero eigenvalues of the Laplacians

provide rich insights into the shape of a manifold. For instance, the Fiedler value, defined as the smallest non-zero eigenvalue of a graph Laplacian, describes connectivity. As another example, the multiplicity of eigenvalues can reveal certain symmetries of the shape.

Remark 4 The two types of discrete Hodge Laplacians (31) not only provide rich geometrical and topological information of the underlying manifold, but also play a central role in the computation of the discrete Hodge decomposition (22) of differential forms for compact domains in 2D and 3D Euclidean spaces. In particular, they can be utilized, by resolving the rank deficiencies, to compute the potentials of the decomposed components in Hodge decomposition on normal or tangential support satisfying the corresponding boundary conditions. In addition, as the kernel sizes of Laplacians are finite, their eigenvectors corresponding to 0 eigenvalues, for each k , form a basis for the space of normal or tangential harmonic fields.

Note that the discrete Hodge stars in the Eulerian setting are almost identical to rescaled identity matrices. Therefore, the computations of the Hodge Laplacian can be further simplified by replacing the Hodge stars with identity matrices, leading to the definition of the Boundary-Induced Graph (BIG) Laplacians as follows:

$$L_{k,n}^B = D_{k,n}^T D_{k,n} + D_{k-1,n} D_{k-1,n}^T \quad (33)$$

$$L_{k,t}^B = D_{k,t}^T D_{k,t} + D_{k-1,t} D_{k-1,t}^T. \quad (34)$$

The BIG Laplacians were introduced in [41] for bounded domains to facilitate the comparison and contrast of the Hodge Laplacians and the combinatorial Laplacians. They preserve the Hodge Laplacian's capability to perform differential calculus but also retain the discrete nature of combinatorial Laplacians. The convergence of the spectra of the BIG Laplacians to Hodge Laplacians has been discussed in [41], showing that the spectra of (33) converge to those of Hodge Laplacians up to a scaling value ℓ^{-2} when enforcing the boundary conditions. This scaling value ℓ^{-2} is exactly the ratio between the missing scaling factor $\ell^{m-2(k+1)}$ in L_k and the missing factor ℓ^{m-2k} of S_k . As the BIG Laplacians produce results similar to those obtained from the discrete Hodge Laplacians with less computation, they can also be used to study the geometric and topological information of the underlying manifolds.

Note that the dual grid is also a Cartesian grid staggered with the primal grid by a replacement of $\ell/2$ in all three axial directions of the Cartesian coordinates. For the study of the spectra of these Laplacians, one only needs to implement one type of boundary condition, for instance, the normal boundary condition, as $L_{k,n}$ defined on the primal grid with normal boundary conditions is equivalent to $L_{m-k,t}$ defined on its dual grid with tangential boundary conditions.

3.3 Topology-preserving construction of Laplacians

Note that, on the grid, the Hodge Laplacians and the BIG Laplacians are of the same sparsity patterns. For simplicity in exposition when discussing the spectrum analysis

of the Laplacians, we let L_k be a generic Laplacian matrix of the form

$$L_k = D_k^T S_{k+1} D_k + S_k D_{k-1} S_{k-1}^{-1} D_{k-1}^T S_k. \quad (35)$$

Here the Laplacian L_k can be interpreted, under choices of boundary conditions and Hodge star accuracy, as either a Hodge Laplacian, or BIG Laplacian (with S_k set to identity), under tangential or normal boundary condition. The eigenvalues and eigenvectors of L_k can be solved by considering the generalized eigenvalue problem

$$L_k W = \lambda S_k W, \quad (36)$$

where λ is an eigenvalue, and W is the associated eigenvector. To analyze the results, we perform the following transformation in the space of discrete forms: $\bar{D}_k = S_{k+1}^{1/2} D_k S_k^{-1/2}$, $\bar{L}_k = S_k^{-1/2} L_k S_k^{-1/2}$ and $\bar{W} = S_k^{1/2} W$. Rewriting the formulas above yields a simplified form of the Laplacian

$$\bar{L}_k = \bar{D}_k^T \bar{D}_k + \bar{D}_{k-1} \bar{D}_{k-1}^T, \quad (37)$$

and a regular eigenvalue problem:

$$\bar{L}_k \bar{W} = \lambda \bar{W}. \quad (38)$$

Note that the property $\bar{D}_k \bar{D}_{k-1} = 0$ is preserved. As the non-zero eigenvalues of $\bar{D}_k^T \bar{D}_k$ and $\bar{D}_k \bar{D}_k^T$ for each k are the same, given by the squared non-zero singular values of the discrete differential \bar{D}_k , and each Laplacian \bar{L}_k is just the combination of $\bar{D}_k^T \bar{D}_k$ and $\bar{D}_{k-1} \bar{D}_{k-1}^T$, the entire spectrum of the Laplacians can thus be studied through the singular values of discrete differentials. Let

$$\bar{D}_k = U_{k+1} \Sigma_k V_k^T \quad (39)$$

be the singular value decomposition of \bar{D}_k , where U_{k+1} and V_k are orthogonal matrices and Σ_k is a rectangular diagonal matrix with diagonal values given by the singular values of \bar{D}_k . It follows immediately from $\bar{D}_k \bar{D}_{k-1} = 0$ that

$$\Sigma_k V_k^T U_k \Sigma_{k-1} = 0. \quad (40)$$

Therefore, the columns of V_k corresponding to non-zero singular values of \bar{D}_k are orthogonal to columns of U_k associated with non-zero singular values of \bar{D}_{k-1} . In addition, it follows from

$$L_k = V_k \Sigma_k^2 V_k^T + U_k \Sigma_{k-1}^2 U_k^T \quad (41)$$

that the spectrum of \bar{L}_k is given by the union of squared non-zero singular values of \bar{D}_k and \bar{D}_{k-1} , and 0, with the multiplicity of 0 given by the k -th Betti numbers. The columns of U_k and V_k corresponding to non-zero singular values, together with the set of harmonic forms, span the entire space of differential k -forms.

In the case that $\dim(M) = 3$, for each type of boundary condition, we have four Laplacians of different degrees in total $k = 0, 1, 2, 3$:

$$\bar{L}_0 = \bar{D}_0^T \bar{D}_0 \quad (42)$$

$$\bar{L}_1 = \bar{D}_1^T \bar{D}_1 + \bar{D}_0 \bar{D}_0^T \quad (43)$$

$$\bar{L}_2 = \bar{D}_2^T \bar{D}_2 + \bar{D}_1 \bar{D}_1^T \quad (44)$$

$$\bar{L}_3 = \bar{D}_2 \bar{D}_2^T. \quad (45)$$

Due to the aforementioned discussion on the spectrum of Laplacians and the duality of the normal and tangential boundary conditions, the spectral analysis of all Laplacians can be reduced to the singular spectra analysis of the three discrete differentials \bar{D}_0 , \bar{D}_1 , and \bar{D}_2 with one type of boundary conditions. Note that the numerical evaluation of the singular values of these differentials, in the simplicial mesh case, may differ for the two types of boundary conditions, as the DoF for normal k -forms and tangent $m - k$ forms are different. However, in the Cartesian representation, they are strictly equivalent to each other by shifting the grid in all directions of the axis by $\ell/2$, so long as M is at least one grid spacing away from the boundary of the grid.

For the computation of the spectra of the Laplacians, we choose the normal boundary condition. The spectra of all Laplacians $\bar{L}_{k,n}$ for compact domains in \mathbb{R}^3 can be finally decomposed into three distinct parts: the squared singular values of the gradient of tangential scalar fields, denoted by T , the squared singular values of the gradient of normal scalar fields, denoted by N , and the squared singular values of the curl of tangential curl fields, denoted by C .

4 Persistent de Rham-Hodge Laplacians

In this section, we present the construction of the persistent de Rham-Hodge Laplacian on differentiable manifolds, which is based on the filtration of manifolds induced by varying a single parameter (the filtration parameter). The spectra of Laplacians carry rich topological and geometric information of a manifold. However, a single manifold might not provide enough information in applications like feature extraction for machine learning analysis. As such, instead of studying just a single manifold, one could examine the spectra of a family of manifolds by adjusting the filtration parameter. The spectra of the Laplacians from this family of manifolds could provide much more information than by considering just one, as the topology and geometry could change for different parameters. This single-parameter family of manifolds, called the evolution of manifolds, was first introduced in [51] based on tetrahedral meshes. We briefly recap the background.

The formal definition of the evolving manifold is given by a one-parameter family of immersions $F_c = F(\cdot, c)$ with $F : B \times [a, b] \rightarrow N$ being a smooth map, where B is called the base manifold, N is the ambient manifold, and $c \in [a, b]$ is a real parameter within the interval. In practice, the most common way to define the evolution of manifolds without specifying B is through a level set function by adjusting the isovalue. Given a function $f : N \rightarrow [a, b]$, then in our case, we consider the sublevel sets $M = \{x \in$

$N \mid f(x) \leq c \}$ with boundary given as $\partial M = \{x \in N \mid f(x) = c\}$ for $c \in [a, b]$. A sequence of manifolds can then be obtained by considering evenly distributed isovales of the function f with the inclusion map

$$M_0 \hookrightarrow M_1 \hookrightarrow M_2 \hookrightarrow \cdots \hookrightarrow M_{s-1} \hookrightarrow M_s, \quad (46)$$

where each M_l is given as the sublevel set corresponding to c_l with $a \leq c_0 < c_1 < \cdots < c_s \leq b$. To ensure that M_l is a manifold, we assume that the function f is a Morse function on N , and none of the c_l 's corresponds to a critical value of the function f , i.e., $f^{-1}(c_l)$ does not contain any critical points. This is always possible as the set of Morse functions on a compact manifold is dense in the space of smooth functions, and their critical points are isolated, nondegenerate, and finite for compact manifolds. Thus we can always perturb any input function slightly to avoid critical isovales in $\{c_l, l = 0, 1, \dots, s\}$. In addition, we assume that for each l , $M_{l,l+1} = \overline{M_{l+1} \setminus M_l} = \{x \in N \mid f(x) \in [c_l, c_{l+1}]\}$ contains at most one critical point, which can be realized by refining the parameter sequence. Note that both M_l and $M_{l,l+1}$ are compact. By Morse theory, if $M_{l,l+1}$ contains no critical points, M_l is diffeomorphic to M_{l+1} . The retraction from M_{l+1} to M_l can be easily constructed by considering a flow along the gradient of the function. As M_{l+1} is homotopic to M_l in this case, there is no topological change happening between (c_l, c_{l+1}) . For the other case when there is exactly one critical point in $M_{l,l+1}$, the manifold M_{l+1} is homotopic to M_l with a k -cell attached, where k is the index of the critical point, defined to be the dimension of the largest subspace on which the Hessian $\text{Hess}(f)(x)$ is negative definite. The topological change of the sublevel sets occurs precisely at the critical values of the level set function. Depending on the type of the critical points, i.e., local minimum, saddle points, and local maximum, the topology changes in different ways. In general, a local maximum has the full index m , a local minimum has index 0, while saddle points have indices strictly between 0 and m . In the case of \mathbb{R}^3 , the occurrences of minima and maxima correspond to the birth of the 0-th generators and the death of the 2nd homology generators respectively, while the occurrences of 1-saddle points correspond to the birth of 1st homology generators or the death of the 0-th homology generators, and those of 2-saddle points correspond to the birth of 2nd homology generators or the death of 1st homology generators.

4.1 Persistent harmonic forms

As the de Rham complex depends on the topology, it can also be extended to the filtration of manifolds. Due to the duality of the normal and tangential boundary conditions, without loss of generality, one may focus on the space of normal differential forms. Given $M_l \hookrightarrow M_{l+1}$, we then need to construct a map from the space of normal k -forms $\Omega_n^k(M_l)$ to the space of normal k -forms $\Omega_n^k(M_{l+1})$, that extends each normal k -form on M_l to a normal k -form on M_{l+1} . Let $\omega \in \Omega_n^k(M_l)$. The idea is to utilize the boundary condition of ω on M_l and extend the forms $\omega|_{\partial M_l}$ to *exact* normal forms on the domain $M_{l,l+1}$ with certain boundary conditions on $\partial M_{l,l+1} = \partial M_l \cup \partial M_{l+1}$. Then the combination $\bar{\omega}$ defines a normal k -form on the manifold M_{l+1} . Note however that $\delta\bar{\omega}$ is only 0 in M_l , so the extension of $\omega \in \ker \delta$ may no longer be in $\ker \delta$ on M_{l+1} .

To be specific, we consider the biharmonic equation $\Delta^2\zeta = \Delta(\Delta\zeta) = 0$ on $M_{l,l+1}$ with both Dirichlet and Neumann boundary conditions to ensure the smoothness of $d\zeta$ with ω through ∂M_l . Note that $d\zeta$ satisfies the normal boundary condition on $M_{l,l+1}$. Let $\bar{\omega}$ be the extension of ω on M_{l+1} with $\bar{\omega} = \omega$ on M_l and $\bar{\omega} = d\zeta$ on $M_{l,l+1}$. It follows that $\bar{\omega} \in \Omega_n^k(M_{l+1})$ as it satisfies the normal boundary condition $\bar{\omega}|_{\partial M_{l+1}} = \zeta|_{\partial M_{l+1}} = 0$.

While the biharmonic equation produces a smooth extension, in practice, it is more efficient to consider the harmonic extension with the boundary condition $\Delta\zeta = 0$ with the boundary condition of $\star d\zeta = \star\omega$ on ∂M_l and the typical normal form boundary condition on ∂M_{l+1} . The solution, by [45, Theorem 3.4.10], is unique. The resulting $\bar{\omega}$ is continuous but nonsmooth as $\delta\bar{\omega}$ may lead to a Dirac distribution on ∂M_l when $M_{l,l+1}$ induces a topological change. For instance, for a harmonic normal 1-form ω on a spherical shell M_l with M_{l+1} turning into a solid ball, the biharmonic extension would create a uniform divergence $\delta\bar{\omega}$ in $M_{l,l+1}$, whereas the harmonic extension creates a thin layer of nonzero divergence $\delta\bar{\omega}$ near the part of ∂M_l around the cavity in the middle. Thus, the harmonic extension serves the same purpose in reducing the kernel of δ .

Denote by $\mathcal{I}_{l,1}$ the map from $\Omega_n^k(M_l)$ to $\Omega_n^k(M_{l+1})$ sending ω to $\bar{\omega}$. Note that $(d \circ \mathcal{I}_{l,1})(\omega)$ is 0 on $M_{l,l+1}$ and thus the same as the extension of the differential of a normal form $d\omega$ on $M_{l,l+1}$, i.e., $d \circ \mathcal{I}_{l,1} = \mathcal{I}_{l,1} \circ d$. It follows that there is a commutative diagram

$$\begin{array}{ccccccc} \Omega_n^0(M_0) & \xrightarrow{d} & \Omega_n^1(M_0) & \xrightarrow{d} & \Omega_n^2(M_0) & \xrightarrow{d} & \Omega_n^3(M_0) \\ \downarrow \mathcal{I}_{0,1}^0 & & \downarrow \mathcal{I}_{0,1}^1 & & \downarrow \mathcal{I}_{0,1}^2 & & \downarrow \mathcal{I}_{0,1}^3 \\ \Omega_n^0(M_1) & \xrightarrow{d} & \Omega_n^1(M_1) & \xrightarrow{d} & \Omega_n^2(M_1) & \xrightarrow{d} & \Omega_n^3(M_1) \\ \downarrow \mathcal{I}_{1,1}^0 & & \downarrow \mathcal{I}_{1,1}^1 & & \downarrow \mathcal{I}_{1,1}^2 & & \downarrow \mathcal{I}_{1,1}^3 \\ \Omega_n^0(M_2) & \xrightarrow{d} & \Omega_n^1(M_2) & \xrightarrow{d} & \Omega_n^2(M_2) & \xrightarrow{d} & \Omega_n^3(M_2) \\ \downarrow \mathcal{I}_{2,1}^0 & & \downarrow \mathcal{I}_{2,1}^1 & & \downarrow \mathcal{I}_{2,1}^2 & & \downarrow \mathcal{I}_{2,1}^3 \\ \dots & & \dots & & \dots & & \dots \end{array}$$

where the horizontal direction gives the de Rham complex and the vertical direction shows the filtration-induced extensions.

Next, we introduce the p -persistent Hodge Laplacian. Let $\mathcal{I}_{l,p} = \mathcal{I}_{l+p-1,1} \circ \dots \circ \mathcal{I}_{l,1}$, which then gives an extension map from the space of normal forms on M_l to the space of normal forms on M_{l+p} . We have the following commutative diagram

$$\begin{array}{ccccc} & & \Omega_n^k(M_l) & & \\ & \nearrow \tilde{d}_{l,p}^{k-1} & & \xrightleftharpoons[d_l^k]{\delta_l^{k+1}} & \Omega_n^{k+1}(M_l) \\ & \searrow \tilde{\delta}_{l,p}^k & & & \\ \Omega_n^{k-1}(M_{l+p}) & \xrightleftharpoons[\delta_{l+p}^k]{d_{l+p}^{k-1}} & \Omega_n^k(M_{l+p}) & \xrightleftharpoons{R_{l,p}} & I_{l,p} \end{array} .$$

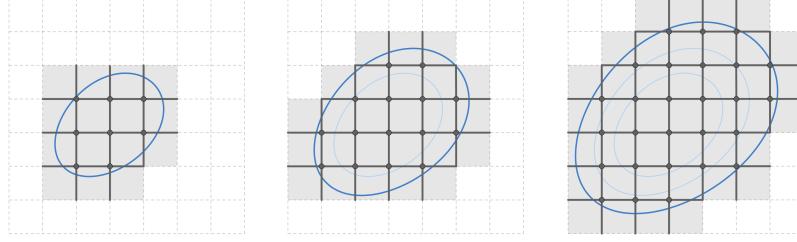


Fig. 4 An example of a nested sequence of sub-cell complexes in a 2D Cartesian grid under the normal boundary condition, illustrating the inclusion of normal supports for 0, 1, and 2 discrete differential forms for an evolution of manifolds. Here the manifolds are represented by the bounded regions of the blue isocurves of a level set function.

Here d_l, δ_l denotes the differential and codifferential on $\Omega^k(M_l)$, d_{l+p}, δ_{l+p} denotes the differential and codifferential on $\Omega^k(M_{l+p})$, respectively, and $\mathcal{R}_{l,p}$ is the projection of differential forms in $\Omega_n^k(M_{l+p})$ to the space spanned by the harmonic extensions followed by the restriction to M_l . Let $\tilde{\delta}_{l,p} = \delta_{l+p} \circ \mathcal{I}_{l,p}$ and $\tilde{d}_{l,p} = \mathcal{R}_{l,p} \circ d_{l+p}$. By the construction of the extension, we have $(\tilde{\delta}_{l,p}\omega, \eta) = (\omega, \tilde{d}_{l,p}\eta)$, i.e., $\tilde{\delta}_{l,p}$ are adjoint to $\tilde{d}_{l,p}$. We then define the p -persistent Hodge Laplacian operator $\Delta_{n,l}^p : \Omega_n^k(M_l) \rightarrow \Omega_n^k(M_l)$ as follows

$$\Delta_{n,l}^p = \tilde{d}_{l,p}\tilde{\delta}_{l,p} + \delta_l d_l. \quad (47)$$

It is easy to see that when $p = 0$, the p -persistent Hodge Laplacian gives exactly the usual Hodge Laplacian $\Delta_{n,l} : \Omega_n^k(M_l) \rightarrow \Omega_n^k(M_l)$ restricted to the space of normal forms. We then define the p -persistent normal harmonic fields as the kernel of the p -persistent Hodge Laplacian $\mathcal{H}_n^{k,p} = \ker \Delta_{n,l}^p$, which can be identified with the space $\ker \tilde{\delta}_{l,p} \cap \ker d_l$. Note that by the extension construction and $\mathcal{R}_{l,p} \circ \mathcal{I}_{l,p} = \text{Id}$, one can see that $\ker \tilde{\delta}_{l,p} \subset \ker \delta_l$ gets smaller as p increases, which confirms that fewer cohomology generators persist longer.

4.2 Discretization of p -persistent de Rham cohomology

The regular Cartesian grid allows one to define persistent graph Laplacian on manifolds in the same way as persistent graph Laplacian [25]. It also allows defining persistent Hodge Laplacian in a consistent way, with the inclusion of nontrivial Hodge stars.

Recall that the discrete differential k -forms can be seen as a k -cochain, i.e., a linear mapping from the chain space C_k to \mathbb{R} that sends a k -chain $c_k = \sum_i a_i \sigma_i$ to $\int_{c_k} \omega = \sum_i a_i W_i$, where $W_i = \int_{\sigma_i} \omega$ is the integral of a smooth k -form ω over the k -cell σ_i .

By varying the isovalue of the level set function f , we can get a sequence of cell complexes given as nested sequences of sub-cell complexes of K satisfying the normal boundary conditions.

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_{s-1} \subset K_s = K. \quad (48)$$

See Fig. 4 for an example of such a nested sequence of sub-cell complexes in a 2D Cartesian grid. Denote by $\mathcal{C}^k(K_l)$ the space of discrete k -forms on subcomplex K_l with $0 \leq l \leq s$. Note that $K_l \subset K_{l+1}$. A discrete k -form on K_l can be easily extended to K_{l+1} by solving the discrete Laplace equation with the above boundary conditions for values on every k -cells in $K_{l,l+1} = \text{Cl}(K_{l+1} \setminus K_l)$, the closure of the difference complex. We denote this extension map as $I_{l,1} : \mathcal{C}^k(K_l) \rightarrow \mathcal{C}^k(K_{l+1})$ and by $I_{l,p} = I_{l+p-1,1} \circ I_{l+p-2,1} \circ \dots \circ I_{l,1} : \mathcal{C}^k(K_l) \rightarrow \mathcal{C}^k(K_{l+p})$ the extension mapping from the space of discrete k -forms on K_l to the space of discrete k -forms on K_{l+p} , which may also be constructed directly by solving the Laplace equation on $K_{l,l+p} = \text{Cl}(K_{l+p} \setminus K_l)$. With this extension mapping, the space of discrete k -forms on K_l can be seen as a subspace of discrete k -forms on K_{l+p} .

A sequence of the discrete de Rham cochain complexes can be defined as follows:

$$\begin{array}{ccccccc}
\mathcal{C}^0(K_0) & \xrightarrow{D_0^0} & \mathcal{C}^1(K_0) & \xrightarrow{D_0^1} & \dots & \xrightarrow{D_0^{k-1}} & \mathcal{C}^k(K_0) \xrightarrow{D_0^k} \mathcal{C}^{k+1}(K_0) \xrightarrow{D_0^{k+1}} \dots \\
\downarrow I_{0,1} & \delta_0^1 & \downarrow I_{0,1} & \delta_0^2 & & \delta_0^k & \downarrow I_{0,1} \\
\mathcal{C}^0(K_1) & \xrightarrow{D_1^0} & \mathcal{C}^1(K_1) & \xrightarrow{D_1^1} & \dots & \xrightarrow{D_1^{k-1}} & \mathcal{C}^k(K_1) \xrightarrow{D_1^k} \mathcal{C}^{k+1}(K_1) \xrightarrow{D_1^{k+1}} \dots \\
\downarrow I_{1,1} & \delta_1^1 & \downarrow I_{1,1} & \delta_1^2 & & \delta_1^k & \downarrow I_{1,1} \\
\mathcal{C}^1(K_2) & \xrightarrow{D_2^0} & \mathcal{C}^1(K_2) & \xrightarrow{D_2^1} & \dots & \xrightarrow{D_2^{k-1}} & \mathcal{C}^k(K_2) \xrightarrow{D_2^k} \mathcal{C}^{k+1}(K_2) \xrightarrow{D_2^{k+1}} \dots \\
\downarrow I_{2,1} & \delta_2^1 & \downarrow I_{2,1} & \delta_2^2 & & \delta_2^k & \downarrow I_{2,1} \\
& \dots & & & & \dots & \dots
\end{array}$$

where $D_l^k : \mathcal{C}^{k+1}(K_l) \rightarrow \mathcal{C}^k(K_l)$ denotes the discrete differential operator, and $\delta_l^k : \mathcal{C}^k(K_l) \rightarrow \mathcal{C}^{k-1}(K_l)$ denotes the discrete codifferential operator on K_l .

To define the persistent discrete Hodge Laplacian, we construct the discrete counterparts of $\tilde{d}_{l,p}$ and $\tilde{\delta}_{l,p}$ in the previous section.

Denote by $\delta_{l,p}^{k+1,n} : \mathcal{C}^{k+1}(K_l) \rightarrow \mathcal{C}_{l,p}^k$ the operator given as $\delta_{l,p}^{k,n} = \delta_{l+p}^k I_{l,p}^{k,n}$, where $\delta_{l+p}^{k,n}$ is the previously defined discrete operator for K_{l+p} and $I_{l,p}^{k,n}$ is the discrete harmonic extension operator defined next. Assuming $K_{l,l+p}$ contains few k -cells, the harmonic extension is then constructed by the linear system $L_{K_{l,l+p}}^{k-1,n} \zeta = 0$, and shifting all $\star d\zeta$ values in the overlap of supports of K_l and $K_{l,l+p}$ to the right-hand side and replacing them with a rescaling of $\star \omega$ based on the k -volume within each support. More specifically, the resulting system is $\tilde{L}_{K_{l,l+p}}^{k-1,n} \tilde{\zeta} = -S^{k-1,n} \delta_{\partial K_l}^k \omega$, where $\tilde{L}_{K_{l,l+p}}^{k-1,n}$ is the Laplace operator applied to a form $\tilde{\zeta}$ defined on $K_{l,l+p} \setminus \partial K_l$, and $\delta_{\partial K_l}^k$ is the boundary codifferential operator that uses the values of ω on ∂K_l to evaluate the neighboring $(k-1)$ -cells in $K_{l,l+p} \setminus \partial K_l$.

The resulting extension operator

$$I_{l,p} = \begin{pmatrix} \text{Id}_{K_l} \\ -D_{K_{l,l+p}}^k (\tilde{L}_{K_{l,l+p}}^{k,n})^{-1} S^{k,n} \delta_{\partial K_l}^k \end{pmatrix},$$

where Id_{K_l} is the identity matrix in K_l up to a rescaling in the boundary, provides the combination of ω in K_l and $d\tilde{\zeta}$ in $K_{l,l+p} \setminus \partial K_l$, when applied to ω . The matrix corresponding to $I_{l,p}$ is dense for rows corresponding to cells in $K_{l,l+p}$ but diagonal for rows corresponding to cells in K_l . Note that $\delta_{\partial K_l}$ is not necessarily 0 for coclosed ω , but is 0 for coexact ω .

The adjoint operator of $\delta_{l,p}^{k+1,n}$ defines $D_{l,p}^k$. In the following, we drop most of the subscripts for clarity. Recall that $(\omega, d\eta) = (\tilde{\delta}\omega, \eta)$ can be discretized as

$$[W]^T S[\tilde{D}E] = [S^{-1} D^T S I_{l,p} W]^T S[E]$$

with W and E as discrete versions of ω and η . Thus $\tilde{D} = S^{-1} I_{l,p}^T S D$, from which we may recognize the restriction operator as $R = S^{-1} I_{l,p}^T S$. This restriction operator can be seen as the L_2 -projection onto the space formed by all harmonic extensions from $\Omega_n^K(M_l)$.

Note that in this case, we immediately $\delta_{l,p}^k \delta_l^{k+1} = 0$, since the extension operator will generate $\tilde{\zeta} = 0$ for any coexact form $\omega = \delta\beta$ on K_l as the right-hand side of the associated linear system essentially corresponds to $\delta\delta\beta = 0$. From the adjoint version, we have $D_l^k D_{l,p}^{k-1} = 0$, and thus the following commutative diagram

$$\begin{array}{ccccc} & & \mathcal{C}^k(K_l) & & \\ & \nearrow D_{l,p}^{k-1} & \xrightarrow{D_l^k} & \xleftarrow{\delta_l^{k+1}} & \mathcal{C}^{k+1}(K_l) \\ \mathcal{C}^{k-1}(K_{l+p}) & \xleftarrow[\delta_{l+p}^k]{D_{l+p}^{k-1}} & R_{l,p} & \uparrow I_{l,p} & \\ & \searrow \delta_{l,p}^k & & & \end{array} .$$

The discrete p -persistent Hodge Laplacian is then given as follows

$$L_{l,p}^k = D_{l,p}^{k-1} \delta_{l,p}^k + \delta_l^{k+1} D_l^k, \quad (49)$$

and the discrete p -persistent BIG Laplacian is

$$L_{l,p}^k = D_{l,p}^{k-1} (D_{l,p}^{k-1})^T + (D_l^k)^T D_l^k. \quad (50)$$

We now present some examples of evolving manifolds and show results for the spectral analysis of their persistent Laplacians. In particular, we focus on the changes of Betti numbers β_0 , β_1 and β_2 and the first non-zero eigenvalues λ_1^T , λ_1^C and λ_1^N of the 0-persistent BIG Laplacians in the set T, C and N, respectively, as introduced in Sec. 3.3. Four models are considered, including the Bimba model, the kitten model, a genus-3 model, and a four-ball model. For each model, we show on the top row snapshots of evolving manifolds at five evenly spaced isovalue intervals in a chosen interval, and on the bottom row the changes in Betti numbers and the first non-zero eigenvalues λ_1^T , λ_1^C and λ_1^N . All the evolving manifolds are generated using isovalue intervals of the signed distance function (SDF) from the original surface model, given as the 0-isosurface of

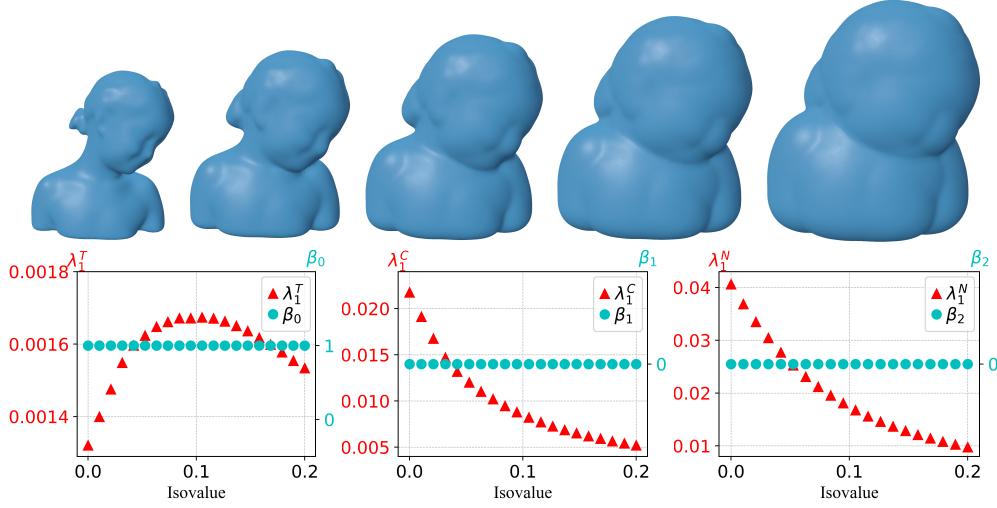


Fig. 5 First row: Snapshots of evolving manifolds for the Bimba model. Second row: Changes in Betti numbers β_0 , β_1 , β_2 and the first non-zero eigenvalues in T, C, N along 20 evenly spaced isovalue intervals from 0 to 0.2. Here the first shape in the top first row corresponds to isovalue 0 and the last shape in the first row corresponds to isovalue 0.2. λ_1^T , λ_1^C and λ_1^N are the first non-zero eigenvalues in the set T, C, N, respectively. The signed distance function generated from the original Bimba model is used as the level set function.

the SDF. As we show below, these values from the evolution of manifolds provide rich information than considering just a single manifold. The discontinuity of these variables indicates the topological changes occurring during the evolution process, and the monotonicity of these non-zero eigenvalues reveals the geometric changes.

The results for the Bimba model are presented in Fig. 5 with an isovalue interval $[0, 0.2]$. As there is no topological change happening in the evolution process, all Betti numbers β_0 , β_1 and β_2 remain constant, and λ_1^T , λ_1^C and λ_1^N are continuous throughout the whole process. Both λ_1^C and λ_1^N decrease as the isovalue increases.

Fig. 6 illustrates the results for the kitten mode with one tunnel formed by its tail. The isovalue interval $[0, 8]$ is considered. One can see all variables are continuous during the evolution process except that β_1 and λ_1^C both drop at the same isovalue, where β_1 changes from 1 to 0. This happens due to the disappearance of the tunnel. In addition, λ_1^T increases at the beginning, and then slows down its rate of increase at the isovalue after the tunnel disappears, and λ_1^C and λ_1^N decrease during the evolution process.

Note that there are also tunnels in the evolving manifolds for the genes-3 model, as we expected, a similar phenomenon can also be observed in Fig. 7 for the change of the Betti numbers and the first non-zero eigenvalues. The isovalue interval $[0.1, 4]$ is considered for this model. The disappearance of the three tunnels leads to a drop of β_1 from 3 to 0 and also a drop of λ_1^C . λ_1^T initially increases, and then changes its behavior to decrease after the tunnels vanish. The evolution process results in a decrease in λ_1^C and λ_1^N , just as the previous two models.

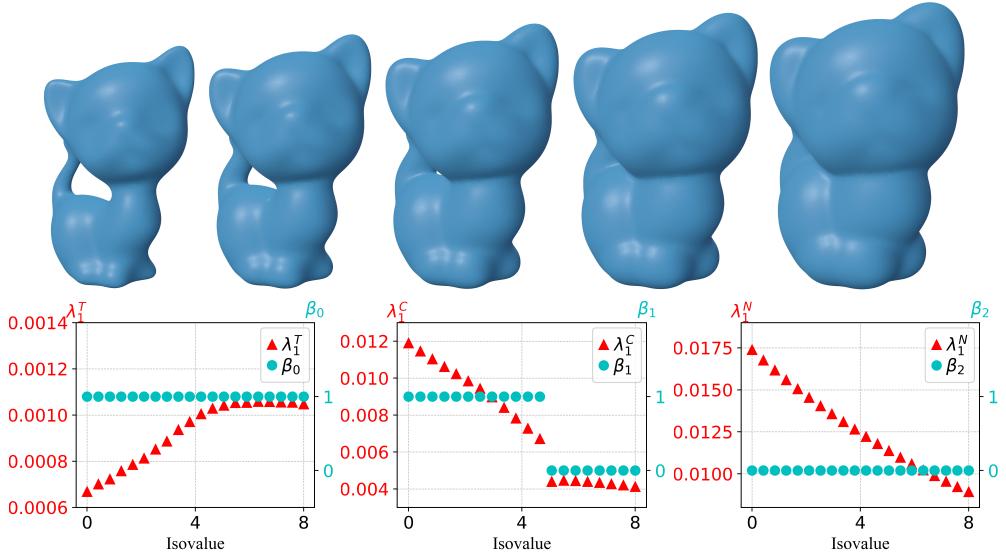


Fig. 6 First row: Snapshots of evolving manifolds for the kitten model. Second row: Changes in Betti numbers β_0 , β_1 , β_2 , and the first non-zero eigenvalues in T, C, N along 20 evenly spaced isovalues from 0 to 8. Here the first shape in the top first row corresponds to isovalue 0 and the last shape in the first row corresponds to isovalue 8. λ_1^T , λ_1^C and λ_1^N are the first non-zero eigenvalues in the set T, C, N, respectively. The signed distance function generated from the original Kitten model is used as the level set function.

The evolving process of the four-ball model with isovalue interval [2, 3.84], see Fig. 8, leads to discontinuities in all Betti numbers and the first non-zero eigenvalues. As the four separate components merge in the evolution, β_0 changes from 4 to 1, along with a drop in λ_1^T at the same isovalue. In addition, β_1 increases from 0 to 3 due to the appearance of three tunnels when the merge happens and then decreases to 0 after the disappearance of all tunnels. The non-zero eigenvalue λ_1^C has a drop that occurs when the tunnel vanishes, however, it is continuous when the tunnels are formed. This suggests that the continuity of λ_1^C is only related to the death but not the birth of tunnels. One can also observe a slowdown in the rate of change of λ_1^T following the disappearance of all tunnels. As the isolate increases further, a cavity occurs in the manifold, resulting in an increase of β_2 from 0 to 1 and finally a decrease from 1 to 0 after the cavity disappears. This topological change can also be observed in λ_1^N , where λ_1^N becomes non-differentiable.

As illustrated by these models, changes in Betti numbers β_0 , β_1 and β_2 and the first non-zero eigenvalues λ_1^T , λ_1^C and λ_1^N not only reflect the changes in topology, but also characterizes the changes in geometry for the evolution of manifolds. The rich information revealed by these variables leads to potential applications in various topological data analysis tasks.

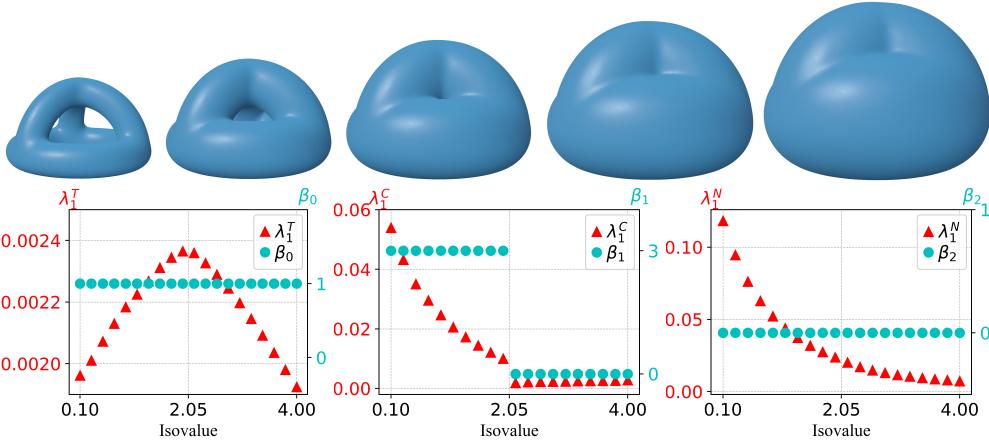


Fig. 7 First row: Snapshots of evolving manifolds for a genus 3 model. Second row: Changes in Betti numbers β_0 , β_1 , β_2 and the first non-zero eigenvalues in T, C, N along 20 evenly spaced isovales from 0.1 to 4. Here the first shape in the top first row corresponds to isovale 0.1 and the last shape in the first row corresponds to isovale 4. λ_1^T , λ_1^C and λ_1^N are the first non-zero eigenvalues in the set T, C, N, respectively. The signed distance function generated from a genus 3 shape is used as the level set function.

5 Proof-of-Principle Experimentation

In this section, we carry out a proof-of-principle experimental demonstration of the proposed persistent de Rham-Hodge theory. In this approach, the problem is defined on manifolds with boundaries. Appropriate boundary conditions are implemented to match actual topological dimensions. The resulting persistent Hodge Laplacians are solved to deliver the corresponding series of eigenvectors and eigenvalues at various scales. In this approach, we use these eigenvalues for machine learning predictions of protein-ligand binding affinity. The binding affinity describes the strength of protein-ligand interactions for each protein-ligand complex.

We consider two benchmark datasets, PDBbind-v2007 and PDBbind-v2016 [52], to demonstrate the effectiveness of our framework in capturing the topological features of protein-ligand complexes. The datasets can be downloaded from <http://pdbind.org.cn/>. These two PDBbind datasets provide collections of biomolecular complexes in Protein Data Bank (PDB) with experimentally measured binding affinity for each protein-ligand complex, and are commonly used in various studies such as drug-discovery or molecular recognition, etc [33, 52–57]. We aim to build a machine learning model, by utilizing the topological and geometric features of the protein-ligand complexes generated using our persistent Hodge Laplacian (PHL) framework as inputs, for predicting the protein-ligand binding affinities.

The biomolecular complexes in each PDBbind dataset are organized into three sets, including a general set, a refined set and a core set, with each set being a superset of the next. In our experiments, for each dataset, we use the refined set, excluding

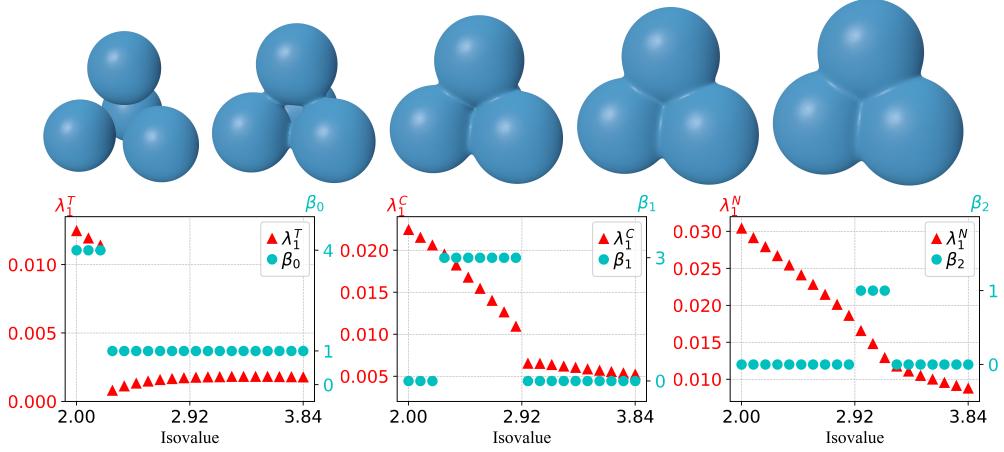


Fig. 8 First row: Snapshots of evolving manifolds for a four-ball model. Second row: Changes in Betti numbers β_0 , β_1 , β_2 and the first non-zero eigenvalues in T, C, N along 20 evenly spaced isovalues from 2 to 3.84. Here the first shape in the top first row corresponds to isovalue 2 and the last shape in the first row corresponds to isovalue 3.84. λ_1^T , λ_1^C and λ_1^N are the first non-zero eigenvalues in the set T, C, N, respectively. The signed distance function generated from four separate balls centered at the vertices of a tetrahedron is used as the level set function.

the core set, to train the predictive model for the binding affinities of the protein-ligand complexes in the core set. The PDBbind-v2007 dataset contains a total of 1,300 complexes with 1,105 in the refined set and 195 in the core set, while the PDBbind-v2016 dataset has a total of 4,057 complexes with 3,767 in the refined set and 290 in the PDBbind core set.

5.1 Element specific discrete to continuum mapping

The original datasets contain atomic names and coordinates, which are the so-called point cloud data. To generate manifold representations, we carry out the discrete to continuum mapping using the flexibility and rigidity index [58]. To compute the topological feature of each protein-ligand complex for the machine learning model, we use the element-specific approach [14]. Specifically, we consider the pairwise interactions between element types that are commonly found in proteins and ligands, including Hydrogen (H), Carbon (C), Nitrogen (N), Oxygen (O), and Sulfur (S) in proteins, and Hydrogen (H), Carbon (C), Nitrogen (N), Oxygen (O), Sulfur (S), Phosphorus (P), Fluorine (F), Chlorine (Cl), Bromine (Br), and Iodine (I) in ligands. These interactions result in a total of 50 pairs of atom types for each protein-ligand complex [14]. However, due to the absence of H in most proteins, we reduce the number of atom pairs to 40 in practice, ignoring the element H in all proteins. These 40 atom pairs, formed by atom types {C, N, O, S} in proteins, and atom types {H, C, N, O, S, P, F, Cl, Br, I} in ligands, along with their xyz coordinates, are used to generate the topological features for each protein-ligand complex. In this paper, all atom-pair complexes are determined by a cutoff distance 12Å from the ligand.

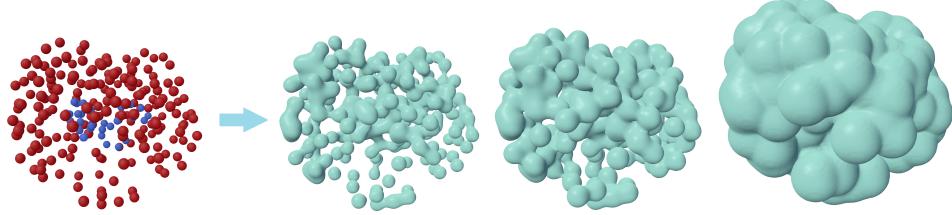


Fig. 9 Left: atoms in the atom pair of type OH in protein-ligand complex 4mnt, with O shown in red and H in blue. Right: a filtration of manifold for this atom pair complex at 3 different isovales with level set function Eq. (51).

Let $\{\mathbf{x}_i^\alpha, i = 1, \dots, s\}$ be the location coordinates of all s atoms in an atom pair, where α denotes the atom type of the atom either in the protein or in the ligand. For this atom pair, a level set function can then be obtained by considering the negative sum of Gaussian density functions defined at the xyz coordinates of all atoms, given as

$$\rho(\mathbf{x}, \tau) = - \sum_{i=1}^s \exp \left(- \left(\frac{\|\mathbf{x} - \mathbf{x}_i^\alpha\|}{\tau r_i^\alpha} \right)^2 \right), \quad (51)$$

where $\|\mathbf{x} - \mathbf{x}_i^\alpha\|$ is the Euclidean distance from position \mathbf{x} to the location \mathbf{x}_i^α of the i -th atom, τ is a scalar value, and r_i^α is the van der Waals radius of the i -th atom, determined by the atom type α . Given an isovalue c , the sublevel set

$$M = \{\mathbf{x} \mid \rho(\mathbf{x}, \tau) \leq c\} \quad (52)$$

defines a compact manifold in \mathbb{R}^3 with its boundary given by the isosurface $\partial M = \{\mathbf{x} \mid \rho(\mathbf{x}, \tau) = c\}$. A filtration of a manifold for the atom pair can then be obtained by choosing a list of evenly spaced isovales of this level set function (51). Let $c_1 < c_2 < \dots < c_s$ be such isovales. We have their corresponding sublevel sets given as follows

$$M_1 \subset M_2 \subset \dots \subset M_s, \quad (53)$$

where M_i is the compact manifold associated to isovale c_i . In Fig. 9 we present one example of the resulting filtration of manifolds at 3 different isovales for atom pair OH in protein-ligand complex 4tmn. Note that the function (51) is a special case of the flexibility rigidity index (FRI) density function [58], which has been shown computationally stable in converting discrete point cloud representations to continuous embeddings, and been used for generating protein boundary surfaces [34] and interactive manifolds [58]. Therefore, one can also make other reasonable choices of FRI density functions to generate the filtration of manifolds.

5.2 Machine learning feature extraction

In the computation of the Laplacians, one can ideally choose a common Cartesian grid such that it contains all manifolds of interest for all protein-ligand complexes, which

ensures that all Laplacians are computed consistently, making their spectra comparable for different complexes. However, as atoms are spread out in the space for different atom pairs, we need to use a sufficiently large grid with a fine resolution for accurate computation of Laplacians, which significantly increases the computational load. Instead, we consider, for each type of atom pairs, a fixed Cartesian grid, regardless of the types of protein-ligand complexes. This approach also ensures that the topological features are comparable for different protein-ligand complexes, as all spectra are computed in a same grid for all atom pairs of the same type. For simplicity, we choose a fixed grid spacing for all Cartesian grids across different atom pairs.

We consider 9 evenly spaced isovalue in the interval $[-0.5, -0.001]$ for all level set functions, which provide 9 compact manifolds for each atom pair. Note that the level set function (51) is always less than 0 and approaches 0 as the norm of \mathbf{x} increases. This interval is chosen as isovalue greater than -0.001 result in no change on the 0-th Betti number β_0 of manifolds for most atom pairs, and isovalue smaller than -0.5 leads to high computational cost, as finer grids are necessary to resolve those isosurfaces. To ensure that the computation of Laplacians is accurate and no topological information is missing due to numerical errors caused by low resolution, we require that at least 8 grid cells of the Cartesian grid are contained in each connected component of a manifold. We compute, for each manifold, the BIG Laplacian $L_{3,n}$ under the normal boundary condition, for which the number of its 0 eigenvalues gives the 0-th Betti number β_0 . We then use the 0-th Betti number β_0 and the first k non-zero eigenvalues of $L_{3,n}$, as the topological feature for the manifold. These $k+1$ features for each of the 9 compact manifolds for each atom pair, amount to $(k+1) \times 9 \times 40$ topological features for each protein-ligand complex. While we only used 9 isovalue within this interval for generating the manifolds in our experiments, more isovalue can be considered, which gives a filtration of more manifolds for each atom pair, and finally leads to more topological features for each protein-ligand complex.

The spectra of the 0-th Laplacian, which in our case corresponds to $L_{3,n}$ under the normal boundary condition, have proven effective and successful in many machine learning tasks [25, 33, 52, 53]. While the Laplacians of other orders could also be used for generating more topological features, we utilize, in this preliminary test, only the spectra of $L_{3,n}$ as features for the protein-ligand complexes in the machine learning model due to the computation efficiency. The results, as shown in Sec.5.4, indicate that these features are sufficient to validate our framework in the machine learning task for predicting the protein-ligand binding affinities.

5.3 Machine learning algorithm

The machine learning models for predicting protein-ligand binding affinities often fall into two categories depending on the type of input data: complex-based or sequence-based models. The complex-based methods are trained using features obtained from the 3D protein-ligand complexes, while the sequence-based models learn from the one-dimensional protein sequences and the ligand simplified molecular-input line-entry system (SMILES) strings. In our experiments, besides the topological features from the 3D protein-ligand complexes, we incorporate protein-ligand features obtained from sequence-based models to build consensus models. To be specific, we make use of the

Table 1 Model performance on PDBbind-v2007 and PDBbind-v2016 benchmarks

	Method	PCC	RMSE (kcal/mol)
PDBbind-v2007	PHL	0.794	2.066
	TF	0.795	2.006
	Consensus	0.826	1.954
PDBbind-v2016	PHL	0.808	1.863
	TF	0.836	1.716
	Consensus	0.849	1.728

Abbreviations: PCC, Pearson correlation coefficient; RMSE, root mean squared error.

recent pre-trained transformer protein language model Evolutionary Scale Modeling-2 (ESM-2) [59], and the pre-trained Transformer-CPZ model [60] for generating the protein and ligand features, respectively, and use their concatenation as inputs for the binding affinity prediction. The residue embeddings from the last layer of the pre-trained ESM-2 model esm.pretrained.esm2_t33_650M_UR50D are used as the protein features, while the embeddings from the last layer of the pre-trained Transformer-CPZ model chembl27_pubchem_zinc_512 are used as the ligand features.

With the topological features and the embedding features obtained from ESM-2 and Transformer-CPZ, we employ the Gradient Boosting Regressor (GBR) module from Scikit-learn 1.4.2 for predicting the protein-ligand binding affinities. We then use the consensus prediction from these models as the final results. The GBR parameters used in our experiments are: n_estimators=10,000, max_depth=5, min_samples_split=5, learning_rate=0.005, loss=squared_error, subsample=0.5, and max_features=sqrt. Changing these parameters does not result in significant differences. To address the randomness from the machine learning algorithm, we repeat each modeling process 20 times with different random seeds, and use the average of predictive results. The Pearson correlation coefficients (PCC) are used as the evaluation metric to assess the performance of our proposed models.

5.4 Experimental Results

The number of topological features for each protein-ligand complex, as in Sec. 5.2, is given by $(1+k) \times 9 \times 40$, where k denotes the number of the first k non-zero eigenvalues of the Laplacians. To find the optimal parameter k leading to the best performance of predictive modules, we carry out the five-fold cross-validation on the training set of each PDBbind dataset with varying values of k based on the average of PCC values. The results indicate that the optimal PCC values for the PDBbind-v2007 and PDBbind-v2016 training sets can be achieved when $k = 5$ and $k = 10$, respectively. For the PDBbind-v2007 training set, the PCC value is 0.709 and the RMSE value is 2.049, while for the PDBbind-v2016 training set, the PCC value is 0.748 and the RMSE value of 1.812. These choices of k result in a total of 2,160 topological features

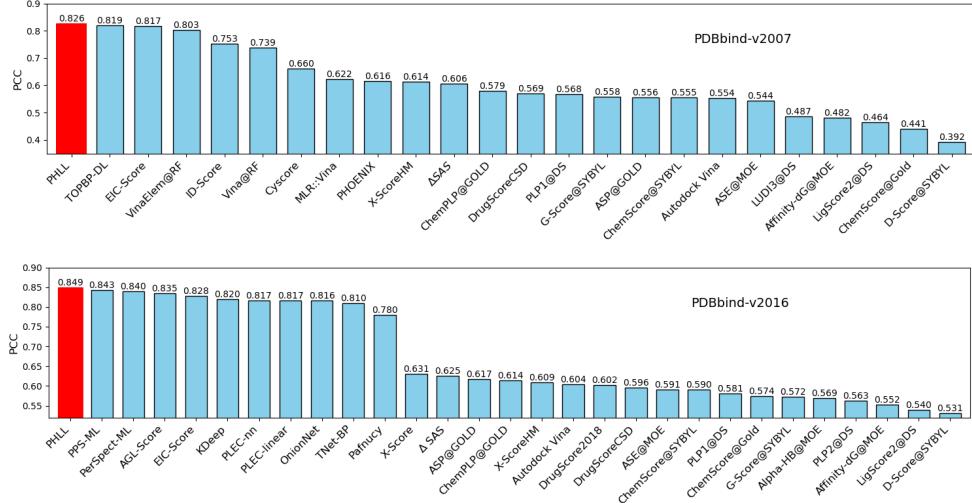


Fig. 10 Performance comparison of the proposed model with other machine learning models for the two PDBbind datasets. The results of the proposed model (PHLL) are in red. The results of other methods are adapted from Refs. [33, 52–55, 57]

for each protein-ligand complex in the PDBbind-v2007 dataset and 3,960 topological features for each protein-ligand complex in the PDBbind-v2016 dataset. These topological features, along with the concatenated protein-ligand features from ESM-2 and Transformer-CPZ, are then used as inputs of the gradient-boosting regressor for binding affinity prediction.

In Table 1, we report the average PCC values and the average root mean squared error (RMSE) of our models on the test set for each PDBbind dataset using only the topological features from PHL, the model using only the transformer features (TF), and the consensus module using both types of features. With the incorporation of topological features, one can see a significant improvement in PCC values when using the proposed consensus model for each dataset, compared to the model using only TF features. The best performance is achieved when using the consensus model, yielding a PCC value of 0.826 with RMSE given as 1.954 for PDBbind-v2007 and 0.849 with RMSE 1.728 for PDBbind-v2016. In addition, we present the Pearson correlation coefficients obtained from our model and those in the previous studies, with results from [33, 53–55, 57]. As illustrated in Fig. 10, our model outperforms all the other models for the two PDBbind datasets. These results demonstrate the utility and effectiveness of our method in capturing the topological features.

6 Conclusion

Although there has had tremendous success of topological data analysis (TDA) [16, 18], particularly, topological deep learning (TDL) on point cloud data [14, 15], there are few methods for the topological analysis of data on manifolds or manifold topological analysis [42]. To fill this gap, we presented a new method, persistent Hodge Laplacian

(PHL) in the Eulerian representation, for manifold topological learning (MTL) of real-world data on manifolds. PHL differs from existing state-of-the-art TDA methods on point clouds in the sense that the proposed PHL is defined on manifolds, for which the traditional TDA methods do not work. Additionally, PHL extends our earlier evolutionary de Rham-Hodge theory constructed on the Lagrangian representation [34] to the Eulerian representation, which avoids numerical inconsistency over multiscale manifolds. We offer two discrete Hodge stars that mimic the continuous operator and developed both a continuous theory for mapping of normal forms across manifolds in a filtration to enable persistent cohomology analysis and the associated topology-persevering discrete construction on Cartesian grids. A proof-of-principle test on two benchmark datasets validates our MTL model, highlighting its simplicity and promise for the predictions of data on manifolds.

The popularity of TDA is facilitated by effective software packages, such as JavaPlex [61], Perseus [2], Ripser [62], etc. The further development of efficient PHL software is an important task. The computational efficiency has not been studied in this work. Algorithm acceleration and parallel and GPU architecture are to be explored. Further experimental validations of manifold topological learning are also needed.

Acknowledgments

This work was supported in part by NIH grants R01GM126189, R01AI164266, and R35GM148196, National Science Foundation grants DMS2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, Michigan State University Research Foundation, and Bristol-Myers Squibb 65109. The authors thank Dr. Hongsong Feng for his help on transformer embeddings.

References

- [1] Wasserman, L.: Topological data analysis. *Annual Review of Statistics and Its Application* **5**(1), 501–532 (2018)
- [2] Mischaikow, K., Nanda, V.: Morse theory for filtrations and efficient computation of persistent homology. *Discrete & Computational Geometry* **50**(2), 330–353 (2013)
- [3] Carlsson, G.: Topology and data. *Bulletin of the American Mathematical Society* **46**(2), 255–308 (2009)
- [4] Edelsbrunner, H., Harer, J., *et al.*: Persistent homology-a survey. *Contemporary mathematics* **453**(26), 257–282 (2008)
- [5] Zomorodian, A., Carlsson, G.: Computing persistent homology. *Discrete & Computational Geometry* **33**(2), 249–274 (2005)

- [6] Ghrist, R.: Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society* **45**(1), 61–75 (2008)
- [7] Bubenik, P., *et al.*: Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16**(1), 77–102 (2015)
- [8] Dey, T.K., Fan, F., Wang, Y.: Computing topological persistence for simplicial maps. In: *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*, pp. 345–354 (2014)
- [9] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., Ziegelmeier, L.: Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research* **18**(8), 1–35 (2017)
- [10] Xia, K., 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)
- [11] Townsend, J., Micucci, C.P., Hymel, J.H., Maroulas, V., Vogiatzis, K.D.: Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature communications* **11**(1), 3230 (2020)
- [12] MacPherson, R., Schweinhart, B.: Measuring shape with topology. *Journal of Mathematical Physics* **53**(7) (2012)
- [13] Cang, Z., Mu, L., Wu, K., Opron, K., Xia, K., Wei, G.-W.: A topological approach for protein classification. *Computational and Mathematical Biophysics* **3**(1) (2015)
- [14] Cang, Z., Wei, G.-W.: Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS computational biology* **13**(7), 1005690 (2017)
- [15] Papamarkou, T., Birdal, T., Bronstein, M.M., Carlsson, G.E., Curry, J., Gao, Y., Hajij, M., Kwitt, R., Lio, P., Di Lorenzo, P., *et al.*: Position: Topological deep learning is the new frontier for relational learning. In: *Forty-first International Conference on Machine Learning* (2024)
- [16] Nguyen, D.D., Cang, Z., Wei, G.-W.: A review of mathematical representations of biomolecular data. *Physical Chemistry Chemical Physics* **22**(8), 4343–4367 (2020)
- [17] Nguyen, D.D., Gao, K., Wang, M., Wei, G.-W.: Mathdl: mathematical deep learning for d3r grand challenge 4. *Journal of computer-aided molecular design* **34**(2), 131–147 (2020)

- [18] Nguyen, D.D., Cang, Z., Wu, K., Wang, M., Cao, Y., Wei, G.-W.: Mathematical deep learning for pose and binding affinity prediction and ranking in d3r grand challenges. *Journal of computer-aided molecular design* **33**, 71–82 (2019)
- [19] Chen, J., Wang, R., Wang, M., Wei, G.-W.: Mutations strengthened sars-cov-2 infectivity. *Journal of molecular biology* **432**(19), 5212–5226 (2020)
- [20] Wang, R., Chen, J., Wei, G.-W.: Mechanisms of sars-cov-2 evolution revealing vaccine-resistant mutations in europe and america. *The journal of physical chemistry letters* **12**(49), 11850–11857 (2021)
- [21] Chen, J., Wei, G.-W.: Omicron ba. 2 (b. 1.1. 529.2): high potential for becoming the next dominant variant. *The journal of physical chemistry letters* **13**(17), 3840–3849 (2022)
- [22] Chen, J., Qiu, Y., Wang, R., Wei, G.-W.: Persistent laplacian projected omicron ba. 4 and ba. 5 to become new dominating variants. *Computers in Biology and Medicine* **151**, 106262 (2022)
- [23] Pun, C.S., Xia, K., Lee, S.X.: Persistent-homology-based machine learning and its applications—a survey. *arXiv preprint arXiv:1811.00252* (2018)
- [24] Wei, X., Wei, G.-W.: Persistent topological laplacians—a survey. *arXiv preprint arXiv:2312.07563* (2023)
- [25] Wang, R., Nguyen, D.D., Wei, G.-W.: Persistent spectral graph. *International Journal for Numerical Methods in Biomedical Engineering* **36**(9), 3376 (2020)
- [26] Wang, R., Zhao, R., Ribando-Gros, E., Chen, J., Tong, Y., Wei, G.-W.: Hermes: Persistent spectral graph software. *Foundations of Data Science* **3**(1), 67–97 (2020)
- [27] Dong, R.: A faster algorithm of up persistent laplacian over non-branching simplicial complexes. *arXiv preprint arXiv:2408.16741* (2024)
- [28] Liu, J., Li, J., Wu, J.: The algebraic stability for persistent laplacians. *arXiv preprint arXiv:2302.03902* (2023)
- [29] Mémoli, F., Wan, Z., Wang, Y.: Persistent laplacians: Properties, algorithms and implications. *SIAM Journal on Mathematics of Data Science* **4**(2), 858–884 (2022)
- [30] Gülen, A.B., Mémoli, F., Wan, Z., Wang, Y.: A generalization of the persistent laplacian to simplicial maps. *arXiv preprint arXiv:2302.03771* (2023)
- [31] Wei, X., Wei, G.-W.: Persistent sheaf laplacian. *Foundations of data science* (Springfield, Mo.), 10–39342024033 (2024)
- [32] Liu, X., Feng, H., Wu, J., Xia, K.: Persistent spectral hypergraph based machine

- learning (psh-ml) for protein-ligand binding affinity prediction. *Briefings in Bioinformatics* **22**(5), 127 (2021)
- [33] Meng, Z., Xia, K.: Persistent spectral-based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science advances* **7**(19), 5329 (2021)
 - [34] Chen, J., Zhao, R., Tong, Y., Wei, G.-W.: Evolutionary de rham-hodge method. *Discrete and continuous dynamical systems. Series B* **26**(7), 3785 (2021)
 - [35] Yang, W., Parr, R.G.: Electron density, kohn–sham frontier orbitals, and fukui functions. *The Journal of Chemical Physics* **81**(6), 2862–2863 (1984)
 - [36] Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G.: Low-dose ct via convolutional neural network. *Biomedical optics express* **8**(2), 679–694 (2017)
 - [37] Khovanov, M.: A categorification of the jones polynomial. *Duke Mathematical Journal* **101**(3), 359–426 (2000)
 - [38] Panagiotou, E., Millett, K.C., Atzberger, P.J.: Topological methods for polymeric materials: characterizing the relationship between polymer entanglement and viscoelasticity. *Polymers* **11**(3), 437 (2019)
 - [39] Shen, L., Feng, H., Li, F., Lei, F., Wu, J., Wei, G.-W.: Knot data analysis using multiscale gauss link integral. *Proceedings of the National Academy of Sciences* (accepted, 2024)
 - [40] Shen, L., Liu, J., Wei, G.-W.: Evolutionary khovanov homology. *AIMS Mathematics* (accepted 2024)
 - [41] Ribando-Gros, E., Wang, R., Chen, J., Tong, Y., Wei, G.-W.: Combinatorial and hodge laplacians: Similarity and difference. *SIAM Review* **66**(3), 575–601 (2024)
 - [42] Desbrun, M., Kanso, E., Tong, Y.: Discrete differential forms for computational modeling. In: *ACM SIGGRAPH 2006 Courses*, pp. 39–54 (2006)
 - [43] Dodziuk, J.: Finite-difference approach to the hodge theory of harmonic forms. *American Journal of Mathematics* **98**(1), 79–104 (1976)
 - [44] Arnold, D.N., Falk, R.S., Winther, R.: Finite element exterior calculus, homological techniques, and applications. *Acta numerica* **15**, 1–155 (2006)
 - [45] Schwarz, G.: *Hodge decomposition - A method for solving boundary value problems*. Springer (2006)
 - [46] Morrey, C.B.: A variational method in the theory of harmonic integrals, ii. *American Journal of Mathematics* **78**(1), 137–170 (1956)

- [47] Zhao, R., Desbrun, M., Wei, G.-W., Tong, Y.: 3d hodge decompositions of edge- and face-based vector fields. *ACM Transactions on Graphics (TOG)* **38**(6), 1–13 (2019)
- [48] Friedrichs, K.O.: Differential forms on riemannian manifolds. *Communications on Pure and Applied Mathematics* **8**(4), 551–590 (1955)
- [49] Shonkwiler, C.: Poincaré duality angles on riemannian manifolds with boundary. PhD thesis, University of Pennsylvania (2009)
- [50] Su, Z., Tong, Y., Wei, G.-W.: Hodge decomposition of single-cell rna velocity. *Journal of chemical information and modeling* **64**(8), 3558–3568 (2024)
- [51] Chen, J., Zhao, R., Tong, Y., Wei, G.-W.: Evolutionary de rham-hodge method. *iscrete and Continuous Dynamical Systems - B* **26**(7), 3785–3821 (2021)
- [52] Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., Wang, R.: Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research* **50**(2), 302–309 (2017)
- [53] Cang, Z., 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* **34**(2), 2914 (2018)
- [54] Cang, Z., Mu, L., Wei, G.-W.: Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology* **14**(1), 1005929 (2018)
- [55] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., Wang, R.: Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling* **59**(2), 895–913 (2018)
- [56] Francoeur, P.G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R.B., Snyder, I., Koes, D.R.: Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling* **60**(9), 4200–4215 (2020)
- [57] Liu, R., Liu, X., Wu, J.: Persistent path-spectral (PPS) based machine learning for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling* **63**(3), 1066–1075 (2023)
- [58] Nguyen, D.D., Wei, G.-W.: Dg-gl: Differential geometry-based geometric learning of molecular datasets. *International journal for numerical methods in biomedical engineering* **35**(3), 3179 (2019)
- [59] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., *et al.*: Language models of protein sequences

- at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022**, 500902 (2022)
- [60] Chen, D., Zheng, J., Wei, G.-W., Pan, F.: Extracting predictive representations from hundreds of millions of molecules. *The journal of physical chemistry letters* **12**(44), 10793–10801 (2021)
 - [61] Adams, H., Tausz, A., Vejdemo-Johansson, M.: Javaplex: A research software package for persistent (co) homology. In: *Mathematical Software–ICMS 2014: 4th International Congress, Seoul, South Korea, August 5–9, 2014. Proceedings* 4, pp. 129–136 (2014). Springer
 - [62] Bauer, U.: Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology* **5**(3), 391–423 (2021)