APPLIED ALGEBRAIC AND GEOMETRIC TOPOLOGIES AND THEIR BIOLOGICAL APPLICATIONS

By

Li Shen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mathematics—Doctor of Philosophy

2025

ABSTRACT

Biological macromolecules display intricate geometric and topological organization that defies traditional descriptors based solely on atom-level coordinates or sequence information. This dissertation introduces an integrated framework that advances both computational algebraic and geometric topology to capture multiscale structure–function relationships in biomolecular data. In the algebraic domain, we expand persistent homology to higher-order N-chain complexes, producing generalized, efficiently computable descriptors; in the geometric domain, we develop a suite of multiscale invariants—including the multiscale Gauss linking integral, evolutionary Khovanov homology, and persistent Khovanov homology—to quantify entanglement in knot-type data. Applied to protein–ligand affinity prediction, DNA/RNA topological analysis, and macromolecular flexibility assessment, these tools yield interpretable features with competitive accuracy, underscoring the promise of topological approaches in contemporary biological research.

ACKNOWLEDGEMENTS

I begin by expressing my deepest and most heartfelt thanks to Professor Guo-Wei Wei, whose vision, rigor, and encouragement have shaped every stage of my doctoral journey. His ability to link abstract topology with concrete biological questions has been both inspiring and transformative for my research.

I am sincerely grateful to my committee members—Professor Yiyang Tong, Professor Moxun Tang, and Professor Ekaterina Rapinchuk—for their insightful feedback and steady guidance. Their thoughtful questions and expert advice have strengthened this dissertation and broadened my perspective.

I also wish to thank the many colleagues and friends I have met in the Wei Lab who made this journey both productive and enjoyable, particularly Wanying Bi, Jones Benjamin, Dong Chen, Jiahui Chen,Hongsong Feng, Nicole Hayes, Yuta Hozumi, Jian Jiang, Dilan Karagüler, Gengzhuo Liu, Jian Liu, Xiang Liu, Lulu Lu, Yuchi Qiu, Zhe Su, Faisal Suwayyid, Rui Wang, Xiaoqi Wei, Junjie Wee, Mushal Zia. Their collaboration, constructive discussions, and day-to-day support turned challenges into opportunities and enriched my graduate experience immeasurably.

I am further indebted to my external collaborators Fengling Li, Fengchun Lei, and Jie Wu for their expertise and generous cooperation on joint projects that expanded the scope and impact of this work.

Finally, I am profoundly grateful to my family for their unwavering love, patience, and encouragement. Their quiet strength and constant support have sustained me throughout this endeavor and made everything possible.

To everyone who has shared their time, expertise, and kindness along the way—thank you.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Computational topology has emerged as a powerful tool for analyzing the complex structures found in biological systems. The rapid growth of high-dimensional biological data—such as molecular conformations, protein–ligand complexes, and nucleic acid chains—poses substantial challenges to traditional geometric and statistical descriptors. These methods often fail to capture essential structural, multiscale, or topological features that underlie biological function and dynamics. In this dissertation, we present a comprehensive framework grounded in both computational algebraic topology and computational geometric topology, aiming to bridge mathematical theory and practical biological applications.

Persistent homology lies at the foundation of many advances in computational algebraic topology. It quantifies topological features across a filtration of simplicial complexes, offering robust and multiscale descriptors for complex data. The theory has proven useful in various biological tasks including molecular property prediction and mutation impact assessment, as shown in [1, 2, 3, 4, 5]. However, classical persistent homology is built on the standard differential condition $d^2 = 0$, which limits its expressiveness for encoding higher-order or cyclic interactions among simplices.

To overcome this limitation, we develop new methods based on N-complexes, where the differential satisfies $d^N = 0$. These generalized chain complexes, first introduced by Mayer [6] and later formalized by Spanier and Dubois-Violette [7, 8], form the basis for a new class of homology theories. By extending coefficients to N-th roots of unity, we construct Persistent Mayer Homology (PMH) and Persistent Mayer Laplacians (PMLs), which yield a family of topological descriptors indexed by degrees $q = 1, 2, \cdots, N-1$. These methods not only enrich the topological information captured but also reduce computational complexity compared to spectral approaches like persistent Laplacians [9, 4]. We establish theoretical stability for PMH and PMLs under metric perturbations and validate their practical effectiveness in tasks such as protein–ligand binding affinity prediction.

Beyond point-cloud topology, many biological structures—such as DNA helices, protein backbones, and molecular loops—are naturally modeled as curves, links, or tangles embedded

in three-dimensional space. These structures motivate a computational geometric topology perspective that captures both global and local entanglement. To our knowledge, this dissertation is among the first systematic efforts to harness geometric-topology techniques for data analysis, though we recognize the field is nascent and complementary approaches will continue to evolve. To this end, we propose the Multiscale Gauss Linking Integral (mGLI), which generalizes the classical Gauss linking number into a multiscale, quantitative descriptor. This invariant captures fine-grained entanglement features across length scales and has shown utility in applications such as protein flexibility analysis and drug screening [10].

Building further, we introduce Evolutionary Khovanov Homology (EKH), a homological categorification that tracks how knot or tangle diagrams evolve through sequences of crossing smoothings. Unlike traditional knot invariants, EKH incorporates a filtration structure to reveal topological transformations that occur across resolutions [11]. We also develop Persistent Khovanov Homology (PKH) for open tangles, overcoming challenges in defining persistent homology on knot-type data. By leveraging concepts from planar algebras and cobordism categories, we establish a theoretical foundation for persistent knot analysis.

Collectively, these developments in algebraic and geometric topology are implemented in computational pipelines and validated on biological datasets involving binding affinity prediction, molecular screening, and structural classification. Our methods consistently demonstrate interpretability, robustness, and predictive power. For instance, using topological features derived from PMH and mGLI, we achieved state-of-the-art performance in predicting protein–ligand binding strengths and in identifying structural features.

In summary, this dissertation presents a unified approach to computational topology in biology by expanding classical topological tools through persistent, multiscale, and categorified methods. By integrating algebraic and geometric topology into algorithmic frameworks, we provide biologically meaningful, mathematically rigorous, and computationally efficient tools for modeling the structure and dynamics of complex biomolecular systems.

CHAPTER 2

COMPUTATIONAL ALGEBRAIC TOPOLOGY IN BIOLOGICAL STUDIES

2.1 N-chain complex and Mayer homology

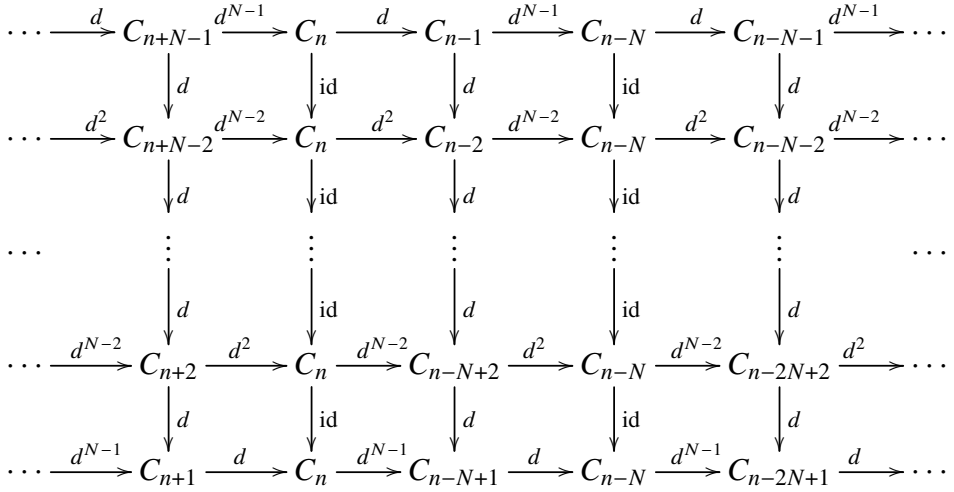
In this section, we review fundamental concepts, including the N-chain complex and Mayer homology. Moreover, for a given simplicial complex, it is possible to construct multiple N-chain complexes. We concentrate on a specific construction, which will be applied to our examples and dataset later on. Additionally, we introduce Laplacian operators on N-chain complexes. This section encompasses some properties of N-chain complexes and Mayer homology, along with examples of related computations. From now on, the ground field is assumed to be the field $\mathbb{K}$. The N-chain complex and Mayer homology can be also built on a commutative ring with unit.

2.1.1 Mayer homology

From now on, N is always an integer ≥ 2.

Definition 2.1.1. An *N-chain complex* consists of a graded $\mathbb{K}$-linear space $C_* = (C_n)_{n\geq 0}$, equipped with a linear map $d : C_* \to C_{*-1}$ of degree -1 satisfying $d^N = 0$. The linear map $d_* : C_* \to C_{*-1}$ is called the *N-differential (N-boundary operator)*.

The following diagram illustrates the N-differential within the N-chain complex. Each horizontal sequence represents a chain complex corresponding to stage q. The vertical sequences are given by the identity map (id) or by the N-differential d.



In particular, when $N = 2$, the N-chain complex reduces to the usual chain complex.

Definition 2.1.2. A *morphism* $f : (C_*, d) \to (C'_*, d')$ *of* N*-chain complexes* is a linear map of degree zero such that $f \circ d = d' \circ f$.

Let (C_*, d) be an N-chain complex. For each $1 \leq q \leq N-1$, the *space of the* q*-th* n*-cycles* is defined by $Z_{n,q} = \{x \in C_n | d^q x = 0\}$. The *space of the* q*-th* n*-boundaries* is given by $B_{n,q} = \{d^{N-q}x | x \in C_{N-q+n}\}$. It follows that $B_{n,q} \subseteq Z_{n,q}$. Let us denote $d_n : C_n \to C_{n-1}$. In particular, for $N = 3$, we can prove that $d_n C_n \subseteq B_{n-1,2}$, $d_n Z_{n,2} \subseteq Z_{n-1,1} \cap B_{n-1,2}$, $d_n Z_{n,1} = 0$, and $d_n B_{n,2} \subseteq B_{n-1,1}$. The *Mayer homology* of the N-chain complex (C_*, d) is defined as

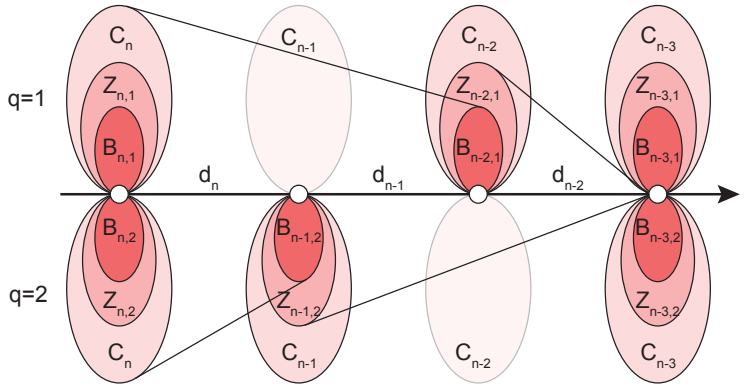


Figure 2.1 Illustration of the boundary operators and chain, cycle, and boundary groups of the N-chain complex for $N = 3$.

$$H_{n,q}(C_*, d) := Z_{n,q}/B_{n,q}, \quad n \geq 0. \tag{2.1.1}$$

The rank of $H_{n,q}(C_*, d)$ is defined as the *Mayer Betti number* of the N-chain complex (C_*, d). The idea of Mayer homology was first introduced by Mayer in 1942 [6]. In Mayer's paper, he constructed the N-chain complex on simplicial complexes over the field $\mathbb{Z}/p$. Here, p is a prime number. And the name of Mayer homology first appeared in [7], which showed the relationship between Mayer homology and the classical homology of simplicial complexes.

Example 2.1.3. Consider the graded vector space $\mathbb{Z}_3[x]$, with the grading $(\mathbb{Z}_3[x])_n = \mathbb{Z}_3 x^n$ and the basis $1, x, x^2, \ldots, x^k, \ldots$. Here, $\mathbb{Z}_3$ is the field with elements $0, 1, 2$ modulo 3. Consider the linear map $d : \mathbb{Z}_3[x] \to \mathbb{Z}_3[x]$ given by $dx^n = nx^{n-1}$ and $d(1) = 0$. It follows that $d^3 = 0$. By a straightforward calculation, we have

$$Z_{n,1} = B_{n,1} = \begin{cases} \mathbb{Z}_3 x^n, & n = 3k, k \in \mathbb{Z}_{\geq 0}; \\ 0, & \text{otherwise.} \end{cases}$$

$$Z_{n,2} = B_{n,2} = \begin{cases} \mathbb{Z}_3 x^n, & n = 3k, 3k+1, k \in \mathbb{Z}_{\geq 0}; \\ 0, & \text{otherwise.} \end{cases}$$

By definition, the Mayer homology is given by

$$H_{n,1}(\mathbb{Z}_3[x]) = H_{n,2}(\mathbb{Z}_3[x]) = 0, \quad n \geq 0.$$

Now, let $A_m = \mathbb{Z}_3\{1, x, \ldots, x^{3m+1}\}$ be the graded vector space generated by $1, x, \ldots, x^{3m+1}$. One

has

$$Z_{n,1} = \begin{cases} \mathbb{Z}_3 x^n, & n = 3k, k = 0, 1, \ldots, m; \\ 0, & \text{otherwise.} \end{cases}$$

$$Z_{n,2} = \begin{cases} \mathbb{Z}_3 x^n, & n = 3k, 3k+1, k = 0, 1, \ldots, m; \\ 0, & \text{otherwise.} \end{cases}$$

$$B_{n,1} = \begin{cases} \mathbb{Z}_3 x^n, & n = 3k, k = 0, 1, \ldots, m-1; \\ 0, & \text{otherwise.} \end{cases}$$

$$B_{n,2} = \begin{cases} \mathbb{Z}_3 x^n, & n = 3k, 3k+1, k = 0, 1, \ldots, m-1; \\ \mathbb{Z}_3 x^n, & n = 3m; \\ 0, & \text{otherwise.} \end{cases}$$

It follows that $H_{n,1}(A_m) = \begin{cases} \mathbb{Z}_3 x^n, & n = 3m; \\ 0, & \text{otherwise} \end{cases}$ and $H_{n,2}(A_m) = \begin{cases} \mathbb{Z}_3 x^n, & n = 3m+1; \\ 0, & \text{otherwise.} \end{cases}$

Let $f : (C_*, d) \to (C'_*, d')$ be a morphism of N-chain complexes. Since f commutes with the N-differential, it induces the morphism of Mayer homology

$$f_{*,q} : H_{*,q}(C_*, d) \to H_{*,q}(C'_*, d'), \quad [z] \mapsto [f(z)] \tag{2.1.2}$$

for any $1 \leq q \leq N-1$. Moreover, one has

Proposition 2.1.1. *([12, Proposition 1])* If $f_{*,1} : H_{*,1}(C_*, d) \to H_{*,1}(C'_*, d')$ and $f_{*,N-1} : H_{*,N-1}(C_*, d) \to H_{*,N-1}(C'_*, d')$ are isomorphisms, then $f_{*,q} : H_{*,q}(C_*, d) \to H_{*,q}(C'_*, d')$ is an isomorphism for any $1 \leq q \leq N-1$.

The above proposition shows that if $f_{*,q} : H_{*,1}(C_*, d) \to H_{*,1}(C'_*, d')$ is an isomorphism for $q = 1, N-1$, then it is an isomorphism for any $1 \leq q \leq N-1$. There are various distinctive properties associated with Mayer homology. For instance, it has been demonstrated in [12] that there exists an isomorphism of linear spaces, $H_{*,q}(C_*, d) \cong H_{*,N-q}(C_*, d)$. However, it does not have to be $H_{n,q}(C_*, d) \cong H_{n,N-q}(C_*, d)$ for a given n.

Let **Nchain** be the category of N-chain complexes, whose objects are the N-chain complexes, and whose morphisms are the morphisms of N-chain complexes. Let $\mathbf{Vec}_{\mathbb{K}}$ be the category of vector spaces over $\mathbb{K}$. Then we have the following proposition.

Proposition 2.1.2. The Mayer homology $H_{*,q} : \mathbf{Nchain} \to \mathbf{Vec}_{\mathbb{K}}$ is a functor for $1 \leq q \leq N-1$.

Proof. For morphisms $f : (C_*, d) \to (C'_*, d')$ and $g : (C'_*, d') \to (C''_*, d'')$ of N-chain complexes, one has

$$g_{*,q} f_{*,q}([z]) = g_{*,q}([f(z)]) = [gf(z)] = (g \circ f)_{*,q}([z]). \tag{2.1.3}$$

Here, $z \in H_{*,q}(C_*, d)$. The left can be verified step by step. □

It is worth noting that the functorial property of Mayer homology is crucial for us to develop the persistence for Mayer homology. More specifically, morphisms at the N-chain level can always induce morphisms at the homology level. Indeed, we also require the functorial property that maps the morphisms at the simplicial complex level to morphisms at the N-chain level.

The N-chain complex is a kind of generalization of the usual chain complex by changing the boundary operator by an N-boundary operator. Other than the homology of N-chain complexes, the homotopy for N-chain complexes can be also built. More precisely, two morphisms $f, g : (C_*, d) \to (C'_*, d')$ of N-chain complexes are *homotopic* if there exist linear maps $h_k : (C_*, d) \to (C'_{*+1}, d')$ of degree 1 for $0 \leq k \leq N-1$ such that $f - g = \sum_{k=0}^{N-1} h_k d^k$. If $f, g : (C_*, d) \to (C'_*, d')$ are N-chain homotopic, then they induce the same morphism of Mayer homology, i.e., $f_{*,q} = g_{*,q}$ for $1 \leq q \leq N-1$.

2.1.2 N-chain complex on simplicial complexes

From now on, for the sake of simplicity, we will always consider the case where N is a prime number, and the field $\mathbb{K}$ is taken to be the complex number field $\mathbb{C}$. Let $\xi = e^{2\pi\sqrt{-1}/N}$ be the primitive N-th root of unity. It follows that $\sum_{i=0}^{N-1} \xi^i = 0$. Moreover, $\sum_{i=0}^{k} \xi^i \neq 0$ for any $0 \leq k \leq N-2$.

Let K be a simplicial complex. Let $C_n(K;\mathbb{C})$ be the linear space generated by the n-simplices of K over $\mathbb{C}$. Consider the linear map $d_n : C_n(K;\mathbb{C}) \to C_{n-1}(K;\mathbb{C})$ given by

$$d_n\langle v_0, v_1, \ldots, v_n\rangle = \sum_{i=0}^{n} \xi^i \langle v_0, \ldots, \hat{v_i}, \ldots, v_n\rangle, \quad n \geq 1 \tag{2.1.4}$$

and $d_0 = 0$. Then $d : C_*(K;\mathbb{C}) \to C_*(K;\mathbb{C})$ is a linear map of degree -1. Moreover, we have

Lemma 2.1.3. $d^N = 0$.

Proof. Let $\partial_i : K_n \to K_{n-1}, \langle v_0, v_1, \ldots, v_n\rangle \mapsto \langle v_0, \ldots, \hat{v_i}, \ldots, v_n\rangle$ denote the i-th face map of simplicial complex K. If $n < N$, we have $d^N = 0$. For $r \leq n$, by induction, we can prove

$$d^r = \left(\prod_{k=1}^{r}(1 + \xi + \cdots + \xi^{k-1})\right) \sum_{j_1<\cdots<j_r} \xi^{j_1+\cdots+j_r-\frac{r(r-1)}{2}} \partial_{j_1} \cdots \partial_{j_r}. \tag{2.1.5}$$

Note that $1 + \xi + \cdots + \xi^{N-1} = 0$. It follows that $d^N = 0$. □

Then the construction $(C_*(K;\mathbb{C}), d)$ is an N-chain complex. There are various ways to construct N-chain complexes on a simplicial complex, and these different constructions lead to different Mayer homology [12]. In this work, we will study the N-chain complex constructed above. The N-chain complex $(C_*(K;\mathbb{C}), d)$ is over the field $\mathbb{C}$, which is more computationally feasible. In addition, we can consider the inner product structure on the N-chain complex $(C_*(K;\mathbb{C}), d)$, which leads to the Laplacians on the N-chain complex.

For $1 \leq q \leq N-1$, the *Mayer homology of the simplicial complex K* is defined by

$$H_{n,q}(K;\mathbb{C}) := H_{n,q}(C_*(K;\mathbb{C}), d), \quad n \geq 0. \tag{2.1.6}$$

The Betti numbers corresponding to the Mayer homology are called the *Mayer Betti numbers* of simplicial complex, denoted by $\beta_{n,q}$.

Proposition 2.1.4. The construction $C_*(-;\mathbb{C}) : \mathbf{Cpx} \to \mathbf{Nchain}$ is a functor from the category of simplicial complexes to the category of N-chain.

Proof. Let $\phi : K \to L$ be a morphism of simplicial complexes. The induced morphism

$$C_*(\phi) : (C_*(K;\mathbb{C}), d_K) \to (C_*(L;\mathbb{C}), d_L)$$

of N-chain complexes is given by $C_*(\phi)(\sigma) = \phi(\sigma)$. Indeed, for any $\sigma = \langle v_0, v_1, \ldots, v_n\rangle$, we have

$$dC_*(\phi)(\sigma) = \sum_{i=0}^{n} \xi^i \langle \phi(v_0), \ldots, \phi(\hat{v}_i), \ldots, \phi(v_n)\rangle = \phi(\sum_{i=0}^{n} \xi^i \langle v_0, \ldots, \hat{v}_i, \ldots, v_n\rangle) = C_*(\phi)(d\sigma). \tag{2.1.7}$$

Obviously, $C_*(\phi)$ preserves identity. The desired result follows. □

Corollary 2.1.5. The Mayer homology $H_{*,q}(-;\mathbb{C}) : \mathbf{Cpx} \to \mathbf{Vec}_{\mathbb{K}}$ is a functor from the category of simplicial complexes to the category of vector spaces over $\mathbb{K}$.

Proof. It is a directed corollary of Proposition 2.1.2 and Proposition 2.1.4. □

The generalized Mayer homology contains the information of the usual simplicial homology. It is worth noting that the Mayer homology here is different from the simplicial homology. Thus, we can obtain additional topological information from the Mayer homology defined above.

Lemma 2.1.6. Let $M_{n,q}$ be the representation matrix of $d_{n,q} = d_{n-q+1} \cdots d_{n-1} d_n : C_n(K;\mathbb{C}) \to C_{n-q}(K;\mathbb{C})$. Then we have

$$\beta_{n,q} = \dim C_n(K;\mathbb{C}) - \operatorname{rank}(M_{n,q}) - \operatorname{rank}(M_{n+N-q,N-q}). \tag{2.1.8}$$

Proof. Consider the short exact sequence

$$0 \longrightarrow Z_{n,q} \hookrightarrow C_n(K;\mathbb{C}) \xrightarrow{d_{n,q}} B_{n-q,N-q} \longrightarrow 0. \tag{2.1.9}$$

Indeed, we have the decomposition

$$C_n(K;\mathbb{C}) \cong Z_{n,q} \oplus B_{n-q,N-q} \cong H_{n,q}(K;\mathbb{C}) \oplus B_{n,q} \oplus B_{n-q,N-q}. \tag{2.1.10}$$

Note that $\operatorname{rank}(M_{n,q}) = \dim B_{n-q,N-q}$. It follows that $\dim B_{n,q} = \operatorname{rank}(M_{n+N-q,N-q})$. Thus we have

$$\dim C_n(K;\mathbb{C}) = \beta_{n,q} + \operatorname{rank}(M_{n+N-q,N-q}) + \operatorname{rank}(M_{n,q}). \tag{2.1.11}$$

The desired result follows. □

Example 2.1.4. Consider the simplicial complex $\Delta[3]$ with the simplices

$$\begin{aligned}&\{0\},\{1\},\{2\},\{3\},\{0,1\},\{0,2\},\{0,3\},\{1,2\},\{1,3\},\{2,3\},\\&\{0,1,2\},\{0,1,3\},\{0,2,3\},\{1,2,3\},\{0,1,2,3\}.\end{aligned} \tag{2.1.12}$$

Consider the 3-chain complex $C_*(\Delta[3];\mathbb{C})$ with the 3-boundary operator given by

$$\begin{aligned}d_3\{0,1,2,3\} &= \{1,2,3\}+\xi\{0,2,3\}+\xi^2\{0,1,3\}+\{0,1,2\},\\ d_2\{0,1,2\} &= \{1,2\}+\xi\{0,2\}+\xi^2\{0,1\},\\ d_2\{0,1,3\} &= \{1,3\}+\xi\{0,3\}+\xi^2\{0,1\},\\ d_2\{0,2,3\} &= \{2,3\}+\xi\{0,3\}+\xi^2\{0,2\},\\ d_2\{1,2,3\} &= \{2,3\}+\xi\{1,3\}+\xi^2\{1,2\}\end{aligned} \tag{2.1.13}$$

and $d_1\{v,w\}=\{w\}+\xi\{v\}$ for $0\le v<w\le 3$. The representation matrices of d_1, d_2 and d_3 with the simplices as basis are given by

$$B_1=\begin{pmatrix}\xi&1&0&0\\ \xi&0&1&0\\ \xi&0&0&1\\ 0&\xi&1&0\\ 0&\xi&0&1\\ 0&0&\xi&1\end{pmatrix},\quad B_2=\begin{pmatrix}\xi^2&\xi&0&1&0&0\\ \xi^2&0&\xi&0&1&0\\ 0&\xi^2&\xi&0&0&1\\ 0&0&0&\xi^2&\xi&1\end{pmatrix},\quad B_3=\begin{pmatrix}1&\xi^2&\xi&1\end{pmatrix}. \tag{2.1.14}$$

The representation matrices of d_1d_2 and d_2d_3 are listed as follows.

$$B_2B_1=\begin{pmatrix}-\xi&-1&-\xi^2&0\\ -\xi&-1&0&-\xi^2\\ -\xi&0&-1&-\xi^2\\ 0&-\xi&-1&-\xi^2\end{pmatrix},\quad B_3B_2=\begin{pmatrix}-1&-\xi^2&-\xi&-\xi&-1&-\xi^2\end{pmatrix}.$$

Moreover, have have that $B_3B_2B_1=\mathbf{O}_{4\times4}$, which shows that $d^3=0$ on $C_*(\Delta[3];\mathbb{C})$. On the other

hand, a straightforward calculation shows that

$$\begin{aligned}
&Z_{3,1} = Z_{3,2} = Z_{2,1} = B_{2,1} = 0,\\
&Z_{2,2} = B_{2,2} = \operatorname{span}\{\{1,2,3\} + \xi\{0,2,3\} + \xi^2\{0,1,3\} + \{0,1,2\}\},\\
&Z_{1,1} = \operatorname{span}\{\{0,2\} - \{0,3\} - \{1,2\} + \{1,3\}, \xi\{0,1\} - \xi\{0,2\} - \{1,3\} + \{2,3\}\},\\
&B_{1,1} = \operatorname{span}\{\xi\{0,1\} + \{0,2\} + \xi^2\{0,3\} + \xi^2\{1,2\} + \xi\{1,3\} + \{2,3\}\},\\
&Z_{1,2} = \operatorname{span}\{\{0,1\},\{0,2\},\{0,3\},\{1,2\},\{1,3\},\{2,3\}\},\\
&B_{1,2} = \operatorname{span}\{\{1,2\} + \xi\{0,2\} + \xi^2\{0,1\}, \{1,3\} + \xi\{0,3\} + \xi^2\{0,1\},\\
&\qquad\qquad \{2,3\} + \xi\{0,3\} + \xi^2\{0,2\}, \{2,3\} + \xi\{1,3\} + \xi^2\{1,2\}\},\\
&Z_{0,1} = \operatorname{span}\{\{0\},\{1\},\{2\},\{3\}\},\\
&B_{0,1} = \operatorname{span}\{\{0\} - \{1\}, \{1\} - \{2\}, \{2\} - \{3\}\},\\
&Z_{0,2} = B_{0,2} = \operatorname{span}\{\{0\},\{1\},\{2\},\{3\}\}.
\end{aligned} \tag{2.1.15}$$

By definition, one has

$$H_{3,1}(\Delta[3];\mathbb{C}) = H_{3,2}(\Delta[3];\mathbb{C}) = H_{2,2}(\Delta[3];\mathbb{C}) = H_{2,1}(\Delta[3];\mathbb{C}) = H_{0,2}(\Delta[3];\mathbb{C}) = 0 \tag{2.1.16}$$

and

$$H_{1,1}(\Delta[3];\mathbb{C}) \cong \mathbb{C}, \quad H_{1,2}(\Delta[3];\mathbb{C}) \cong \mathbb{C}^2, \quad H_{0,1}(\Delta[3];\mathbb{C}) \cong \mathbb{C}. \tag{2.1.17}$$

However, the simplicial homology of $\Delta[3]$ is $H_n(\Delta[3];\mathbb{C}) = \begin{cases} \mathbb{C}, & n = 0; \\ 0, & \text{otherwise.} \end{cases}$ This indicates that even for contractible spaces, Mayer homology may not be trivial.

Example 2.1.5. Many common geometric shapes can be viewed as simplicial complexes through simplicial triangulations. In this example, we compute the Mayer Betti numbers for the simplicial complexes $\Delta[3]$, $\partial\Delta[3]$, and a hexagon. Additionally, we perform simplicial triangulations for the Möbius strip, torus, and octahedron, and calculate the Mayer Betti numbers for these simplicial complexes. The simplicial complex $\partial\Delta[3]$ has the simplices listed as follows:

$$\{0\},\{1\},\{2\},\{3\},\{0,1\},\{0,2\},\{0,3\},\{1,2\},\{1,3\},\{2,3\},\{0,1,2\},\{0,1,3\},\{0,2,3\},\{1,2,3\}. \tag{2.1.18}$$

A hexagon is a simplicial complex with the simplices listed as follows:

$$\{0\},\{1\},\{2\},\{3\},\{4\},\{5\},\{0,1\},\{1,2\},\{2,3\},\{3,4\},\{4,5\},\{0,5\}. \tag{2.1.19}$$

Now, we provide simplicial triangulations for the Möbius strip, torus, and octahedron, and compute the corresponding Mayer Betti numbers.

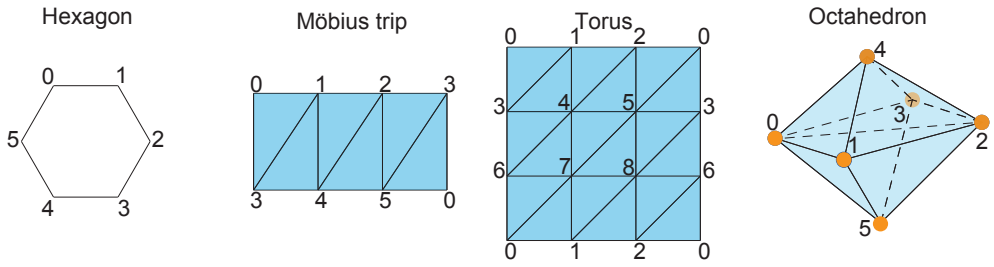


Figure 2.2 The simplicial triangulations of the Möbius strip, hexagon, torus, and octahedron.

The simplicial triangulations of the Möbius strip, torus, and octahedron are shown in Figure 2.2.

simplicial complexes	$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{2,1}$	$\beta_{0,2}$	$\beta_{1,2}$	$\beta_{2,2}$
$\Delta[3]$	1	1	0	0	2	0
$\partial\Delta[3]$	1	2	0	0	2	1
Hexagon	6	0	0	0	6	0
Möbius trip	1	6	0	0	6	1
Torus	1	18	0	0	9	10
Octahedron	1	3	1	0	2	3

Table 2.1 The Mayer Betti numbers for the simplicial complexes $\Delta[3]$, $\partial\Delta[3]$, a hexagon, and the simplicial triangulations of the Möbius strip, torus, and octahedron.

Using our algorithm's computations, Mayer Betti numbers can be obtained, as illustrated in Table 2.1.

2.1.3 The Mayer Laplacians on N-chain complexes

Now, let K be a simplicial complex. Then we have a chain complex $(C_*(K;\mathbb{C}),d)$. One can endow $C_*(K;\mathbb{C})$ with an inner product given by

$$\langle \lambda\sigma, \mu\tau\rangle = \begin{cases} \lambda\cdot\overline{\mu}, & \sigma=\tau; \\ 0, & \text{otherwise.} \end{cases} \tag{2.1.20}$$

Here, $\lambda, \mu \in \mathbb{C}$, and $\overline{\mu}$ is the complex conjugate of μ. Consider the adjoint operator d^* of d, i.e., $\langle dx, y\rangle = \langle x, d^* y\rangle$ for any $x, y \in C_*(K;\mathbb{C})$. Note that

$$\langle d^q x, y\rangle = \langle d^{q-1}x, d^* y\rangle = \cdots = \langle x, (d^*)^q y\rangle. \tag{2.1.21}$$

By the definiteness of inner product, one has $(d^q)^* = (d^*)^q$. For $1 \leq q \leq N-1$, the *Mayer Laplacian* $\Delta_{*,q} : C_*(K;\mathbb{C}) \to C_*(K;\mathbb{C})$ is defined as

$$\Delta_{*,q} := (d^q)^* \circ d^q + d^{N-q} \circ (d^{N-q})^*. \tag{2.1.22}$$

Choose the simplices of K as an orthogonal basis of the N-chain complex $C_*(K;\mathbb{C})$ over $\mathbb{C}$. Let B be the representation matrix of the linear operator $d : C_*(K;\mathbb{C}) \to C_{*-1}(K;\mathbb{C})$ with respect to the chosen orthogonal basis under left multiplication. Then the representation matrix of $\Delta_{*,q}$ is given by

$$L_q = B^q(\overline{B}^q)^T + (\overline{B}^{N-q})^T B^{N-q}. \tag{2.1.23}$$

Here, $\overline{B}^T$ is the conjugate transpose or Hermitian transpose matrix of B. For the graded case, the Mayer Laplacian $\Delta_{n,q} : C_n(K;\mathbb{C}) \to C_n(K;\mathbb{C})$ is given by

$$\Delta_{n,q} = (d_n)^* \circ \cdots \circ (d_{n-q+1})^* \circ d_{n-q+1} \circ \cdots \circ d_n + d_{n+1} \circ \cdots \circ d_{n+N-q} \circ (d_{n+N-q})^* \circ \cdots \circ (d_{n+1})^*. \tag{2.1.24}$$

Here, $d_n : C_n(K;\mathbb{C}) \to C_{n-1}(K;\mathbb{C})$ is the operator of d restricted to $C_n(K;\mathbb{C})$. Let B_n be the representation matrix of d_n with respect to the chosen orthogonal basis, and the representation matrix of $\Delta_{n,q}$ is given by

$$L_{n,q} = B_n \cdots B_{n-q+1}\overline{B_{n-q+1}}^T \cdots \overline{B_n}^T + \overline{B_{n+1}}^T \cdots \overline{B_{n+N-q}}^T B_{n+N-q} \cdots B_{n+1}. \tag{2.1.25}$$

Here, B_n is a complex matrix and $\overline{B_n}^T$ is the conjugate transpose of B_n.

Proposition 2.1.7. The Laplacian $\Delta_{n,q}$ on $C_n(K;\mathbb{C})$ is a self-adjoint and non-negative definite operator.

The proof of Proposition 2.1.7 is a straightforward verification, one can refer to [13]. It is worth noting that even over the complex number field $\mathbb{C}$, the eigenvalues of the Laplacian operator are non-negative.

Proposition 2.1.8. For any n and $1 \leq q \leq N-1$, we have $\dim \ker \Delta_{n,q} = \beta_{n,q}$.

Proof. It is a classic result. One can obtain a detailed proof in a [14]. □

Example 2.1.6. Let us compute the Mayer Laplacians on $\partial\Delta[3]$. We can obtain the N-chain complex $C_*(\partial\Delta[3];\mathbb{C})$ with the differential given by $d_0 = 0$,

$$d_1 \begin{pmatrix} \{0,1\} \\ \{0,2\} \\ \{0,3\} \\ \{1,2\} \\ \{1,3\} \\ \{2,3\} \end{pmatrix} = \begin{pmatrix} \xi & 1 & 0 & 0 \\ \xi & 0 & 1 & 0 \\ \xi & 0 & 0 & 1 \\ 0 & \xi & 1 & 0 \\ 0 & \xi & 0 & 1 \\ 0 & 0 & \xi & 1 \end{pmatrix} \begin{pmatrix} \{0\} \\ \{1\} \\ \{2\} \\ \{3\} \end{pmatrix} \tag{2.1.26}$$

and

$$d_2 \begin{pmatrix} \{0,1,2\} \\ \{0,1,3\} \\ \{0,2,3\} \\ \{1,2,3\} \end{pmatrix} = \begin{pmatrix} \xi^2 & \xi & 0 & 1 & 0 & 0 \\ \xi^2 & 0 & \xi & 0 & 1 & 0 \\ 0 & \xi^2 & \xi & 0 & 0 & 1 \\ 0 & 0 & 0 & \xi^2 & \xi & 1 \end{pmatrix} \begin{pmatrix} \{0,1\} \\ \{0,2\} \\ \{0,3\} \\ \{1,2\} \\ \{1,3\} \\ \{2,3\} \end{pmatrix}. \tag{2.1.27}$$

We denote the representation matrix of d_n by B_n. Observe that $B_0 = B_3 = \mathbf{O}$. It follows that

$$L_{0,1} = \begin{pmatrix} 3 & 2\xi^2 & -1 & 2\xi \\ 2\xi & 3 & 2\xi^2 & -1 \\ -1 & 2\xi & 3 & 2\xi^2 \\ 2\xi^2 & -1 & 2\xi & 3 \end{pmatrix}, \quad L_{0,2} = \begin{pmatrix} 3 & \xi^2 & \xi^2 & \xi^2 \\ \xi & 3 & \xi^2 & \xi^2 \\ \xi & \xi & 3 & \xi^2 \\ \xi & \xi & \xi & 3 \end{pmatrix}, \tag{2.1.28}$$

$$L_{1,1} = \begin{pmatrix} 2 & 1 & 1 & \xi^2 & \xi^2 & 0 \\ 1 & 2 & 1 & 1 & 0 & \xi^2 \\ 1 & 1 & 2 & 0 & 1 & 1 \\ \xi & 1 & 0 & 2 & 1 & \xi^2 \\ \xi & 0 & 1 & 1 & 2 & 1 \\ 0 & \xi & 1 & \xi & 1 & 2 \end{pmatrix}, \quad L_{1,2} = \begin{pmatrix} 2 & \xi^2 & \xi^2 & \xi & \xi & 0 \\ \xi & 2 & \xi^2 & \xi^2 & 0 & \xi \\ \xi & \xi & 2 & 0 & \xi^2 & \xi^2 \\ \xi^2 & \xi & 0 & 2 & \xi^2 & \xi \\ \xi^2 & 0 & \xi & \xi & 2 & \xi^2 \\ 0 & \xi^2 & \xi & \xi^2 & \xi & 2 \end{pmatrix}. \tag{2.1.29}$$

The spectra of $L_{0,1}$, $L_{0,2}$, $L_{1,1}$, and $L_{1,2}$ are

$$\begin{aligned} \mathbf{Spec}(L_{0,1}) &= \{0, 4 - 2\sqrt{3}, 4, 4 + 2\sqrt{3}\}, \\ \mathbf{Spec}(L_{0,2}) &= \{2 - \sqrt{3}, 3, 5, 2 + \sqrt{3}\}, \\ \mathbf{Spec}(L_{1,1}) &= \{0, 0, 2 - \sqrt{3}, 3, 5, 2 + \sqrt{3}\}, \\ \mathbf{Spec}(L_{1,2}) &= \{0, 0, 2 - \sqrt{3}, 3, 2 + \sqrt{3}, 5\}. \end{aligned} \tag{2.1.30}$$

Let $\omega(\Delta_{n,q})$ denote the number of zero eigenvalues of the operator $\Delta_{n,q}$. It is worth noting that $\omega(\Delta_{0,1}) = 1$, $\omega(\Delta_{0,2}) = 0$, $\omega(\Delta_{1,1}) = 2$, $\omega(\Delta_{2,2}) = 2$. This is consistent with the Betti numbers corresponding to Table 2.1.

Example 2.1.7. Now, we will compute the Mayer Laplacians of the hexagon. As described in Example 2.1.5, the 3-chain of a hexagon is a graded vector space with the corresponding 3-differential given by

$$d_1 \begin{pmatrix} \{0,1\} \\ \{1,2\} \\ \{2,3\} \\ \{3,4\} \\ \{4,5\} \\ \{0,5\} \end{pmatrix} = \begin{pmatrix} \xi & 1 & 0 & 0 & 0 & 0 \\ 0 & \xi & 1 & 0 & 0 & 0 \\ 0 & 0 & \xi & 1 & 0 & 0 \\ 0 & 0 & 0 & \xi & 1 & 0 \\ 0 & 0 & 0 & 0 & \xi & 1 \\ \xi & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \{0\} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{pmatrix} \tag{2.1.31}$$

and $d_n = 0$ for $n \neq 1$. The calculation for $N = 3$ is shown in Table 2.2. For the case $N = 5$, we have

n, q	$n = 0, q = 1$	$n = 0, q = 2$	$n = 1, q = 1$	$n = 1, q = 2$
$L_{n,q}$	$\mathbf{O}_{6\times 6}$	$\begin{pmatrix} 2 & \xi^2 & 0 & 0 & 0 & 1 \\ \xi & 2 & \xi^2 & 0 & 0 & 0 \\ 0 & \xi & 2 & \xi^2 & 0 & 0 \\ 0 & 0 & \xi & 2 & \xi^2 & 0 \\ 0 & 0 & 0 & \xi & 2 & 1 \\ 1 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & \xi^2 & 0 & 0 & 0 & 1 \\ \xi & 2 & \xi^2 & 0 & 0 & 0 \\ 0 & \xi & 2 & \xi^2 & 0 & 0 \\ 0 & 0 & \xi & 2 & \xi^2 & 0 \\ 0 & 0 & 0 & \xi & 2 & 1 \\ 1 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$	$\mathbf{O}_{6\times 6}$
$\beta_{n,q}$	6	0	0	6
$\mathbf{Spec}(L_{n,q})$	{0,0,0,0,0,0}	{0.12,0.47,1.65,2.35,3.53,3.88}	{0.12,0.47,1.65,2.35,3.53,3.88}	{0,0,0,0,0,0}

Table 2.2 Illustration of Mayer Laplacians for $N = 3$.

the corresponding 5-differential given by

$$d_1 \begin{pmatrix} \{0,1\} \\ \{1,2\} \\ \{2,3\} \\ \{3,4\} \\ \{4,5\} \\ \{0,5\} \end{pmatrix} = \begin{pmatrix} \xi_5 & 1 & 0 & 0 & 0 & 0 \\ 0 & \xi_5 & 1 & 0 & 0 & 0 \\ 0 & 0 & \xi_5 & 1 & 0 & 0 \\ 0 & 0 & 0 & \xi_5 & 1 & 0 \\ 0 & 0 & 0 & 0 & \xi_5 & 1 \\ \xi_5 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \{0\} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{pmatrix} \tag{2.1.32}$$

and $d_n = 0$ for $n \neq 1$. Here, ξ_5 is the primitive 5-th root of unity. The calculated result at this point is shown in Table 2.3. Our calculations demonstrate that the eigenvalues are consistently non-negative.

n, q	$n = 0, q = 1$	$n = 0, q = 2$	$n = 0, q = 3$	$n = 0, q = 4$
$L_{n,q}$	$\mathbf{O}_{6\times 6}$	$\mathbf{O}_{6\times 6}$	$\mathbf{O}_{6\times 6}$	$\begin{pmatrix} 2 & \xi_5^4 & 0 & 0 & 0 & 1 \\ \xi_5 & 2 & \xi_5^4 & 0 & 0 & 0 \\ 0 & \xi_5 & 2 & \xi_5^4 & 0 & 0 \\ 0 & 0 & \xi_5 & 2 & \xi_5^4 & 0 \\ 0 & 0 & 0 & \xi_5 & 2 & 1 \\ 1 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$
$\beta_{n,q}$	6	6	6	0
$\mathbf{Spec}(L_{n,q})$	{0,0,0,0,0,0}	{0,0,0,0,0,0}	{0,0,0,0,0,0}	{0.04,0.66,1.38,2.62,3.34,3.96}

n, q	$n = 1, q = 1$	$n = 1, q = 2$	$n = 1, q = 3$	$n = 1, q = 4$
$L_{n,q}$	$\begin{pmatrix} 2 & \xi_5^4 & 0 & 0 & 0 & 1 \\ \xi_5 & 2 & \xi_5^4 & 0 & 0 & 0 \\ 0 & \xi_5 & 2 & \xi_5^4 & 0 & 0 \\ 0 & 0 & \xi_5 & 2 & \xi_5^4 & 0 \\ 0 & 0 & 0 & \xi_5 & 2 & 1 \\ 1 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$	$\mathbf{O}_{6\times 6}$	$\mathbf{O}_{6\times 6}$	$\mathbf{O}_{6\times 6}$
$\beta_{n,q}$	0	6	6	6
$\mathbf{Spec}(L_{n,q})$	{0.04,0.66,1.38,2.62,3.34,3.96}	{0,0,0,0,0,0}	{0,0,0,0,0,0}	{0,0,0,0,0,0}

Table 2.3 Illustration of Mayer Laplacians for $N = 5$.

Moreover, the number of zero eigenvalues of Laplacians coincides with the corresponding Mayer

Betti numbers.

In an intuitive sense, the Mayer homology and Mayer Laplacian of a complex reflect connections between simplices at different dimensions. The corresponding Betti numbers reveal the topological cycles representing interactions between simplices of different dimensions, whereas the eigenvalues of the Laplacian operator deconstruct the connectivity between simplices of various dimensions. These relationships are more intricate and subtle, extending beyond what traditional simplicial homology theory can capture.

2.2 Persistence on Mayer features

In this section, we will explore the persistent versions of Mayer homology and Mayer Laplacians. Since Mayer homology and Mayer Laplacians provide information different from the usual simplicial homology and Laplacian, investigating Mayer features is highly meaningful for our study of the topological characteristics and geometric structure of data. From now on, the ground field is taken to be the complex number field $\mathbb{C}$. Besides, we always consider the case that N is a prime number for the sake of simplicity.

2.2.1 Persistent Mayer homology

Let K be a simplicial complex, and let $f : K \to \mathbb{R}$ be a real-valued function defined on K such that $f(\sigma) \leq f(\tau)$ for every face σ of τ in K. For each real number a, we can obtain a sub complex $K_a = \{\sigma \in K | f(\sigma) \leq a\}$ of K. Moreover, for real numbers $a \leq b$, one has $K_a \subseteq K_b$. Thus, we can obtain a filtration of simplicial complexes

$$K_{a_1} \subseteq K_{a_2} \subseteq \cdots \subseteq K_{a_m} \tag{2.2.1}$$

for real numbers $a_1 < a_1 < \cdots < a_m$. By Proposition 2.1.4, we have a sequence of N-chain complexes

$$C_*(K_{a_1}; \mathbb{C}) \to C_*(K_{a_2}; \mathbb{C}) \to \cdots \to C_*(K_{a_m}; \mathbb{C}). \tag{2.2.2}$$

By Proposition 2.1.2, this induces a sequence of Mayer homology

$$H_{*,q}(K_{a_1}; \mathbb{C}) \to H_{*,q}(K_{a_2}; \mathbb{C}) \to \cdots \to H_{*,q}(K_{a_m}; \mathbb{C}) \tag{2.2.3}$$

for any $1 \leq q \leq N-1$. For any real numbers $a \leq b$ and $1 \leq q \leq N-1$, the (a,b)-*persistent Mayer homology* is defined by

$$H_{n,q}^{a,b} := \operatorname{im}\left(H_{n,q}(K_a;\mathbb{C}) \to H_{n,q}(K_b;\mathbb{C})\right), \quad n \geq 0. \tag{2.2.4}$$

The rank of $H_{n,q}^{a,b}$ is the (a,b)-persistent Betti numbers. The persistent Betti numbers can also be visualized using a persistence diagram or barcode. It is worth noting that for each $1 \leq q \leq N-1$, we can obtain a persistence diagram, which means that the persistent Mayer homology contains more information than the usual persistent homology. Moreover, the fundamental theorems of persistent homology are also applicable to persistent Mayer homology.

Let $\{K_{a_i}\}_{i\geq 1}$ be a filtration of simplicial complexes. For each $i \geq 1$, we have the map $x : H_{*,q}(K_{a_i};\mathbb{C}) \to H_{*,q}(K_{a_{i+1}};\mathbb{C})$ induced by $i \to i+1$. Consider the persistent homology, denoted as $\mathbf{H}_q = \bigoplus_{i=1}^{\infty} H_{*,q}(K_{a_i};\mathbb{C})$, which encapsulates homological information from all time steps. Then one has a map $x : \mathbf{H}_q \to \mathbf{H}_q$, where x map a generator at a_i to a generator at a_{i+1}. Let $\mathbb{C}[x]$ be a polynomial ring over the complex number field $\mathbb{C}$. The space $\mathbf{H}_q$ is a left $\mathbb{C}[x]$-module given by

$$\mathbb{C}[x] \times \mathbf{H}_q \to \mathbf{H}_q, \quad (f(x), \alpha) \mapsto f(x)(\alpha). \tag{2.2.5}$$

Moreover, the module structure theorem for persistent Mayer homology is established as follows.

Theorem 2.2.1. For a filtration of finite simplicial complexes $\{K_{a_i}\}_{i\geq 1}$, the corresponding persistent Mayer homology $\mathbf{H}_q$ has a decomposition as $\mathbb{C}[x]$-module

$$\mathbf{H}_q \cong \left(\bigoplus_t \mathbb{C}[x]\cdot\alpha_{b_t}\right) \oplus \left(\bigoplus_s \mathbb{C}[x]/x^{c_s}\cdot\beta_{b_s}\right). \tag{2.2.6}$$

The proof of the above theorem is essentially a replica of the standard persistent homology structure theorem. Similarly, the generators in the free part, denoted as α_{b_t}, refer to those generators born at time b_t and persist until infinity, while β_{b_s} represents the generators born at time b_s and dead at time $b_s + c_s$. Similarly, we can define the barcode for persistent Mayer homology and give the fundamental characterization theorem for barcodes.

2.2.2 Wasserstein distance for Mayer persistence diagrams

Recall that the r-th Wasserstein distance of persistence diagrams is defined by

$$W_r(\mathcal{D}, \mathcal{D}') = \inf_{\gamma:\mathcal{D}\to\mathcal{D}'} \left(\sum_{x\in\mathcal{D}} \|x - \gamma(x)\|_s^r \right)^{1/r}, \tag{2.2.7}$$

where $\mathcal{D}, \mathcal{D}'$ are persistence diagrams, $\|\cdot\|_s$ denotes the L_s-distance on a persistence diagram, and the infimum is taken over all matchings between $\mathcal{D}$ and $\mathcal{D}'$.

In the context of a filtration of simplicial complexes, a family of persistence diagrams $\mathcal{D}_1, \ldots, \mathcal{D}_{N-1}$ can be obtained for the persistent Mayer homology concerning the p-boundary operator. This collection is referred to as the Mayer persistence diagram. To formalize the relationship between these diagrams, we introduce the r-th Wasserstein distance for Mayer persistence diagrams, defined by

$$W_r(\{\mathcal{D}_q\}_{1\le q\le N-1}, \{\mathcal{D}'_q\}_{1\le q\le N-1}) = \left(\sum_{q=1}^{N-1} W_r(\mathcal{D}_q, \mathcal{D}'_q)^r \right)^{1/r}. \tag{2.2.8}$$

The case where $r = \infty$ is notably well-known. In this scenario, the Wasserstein distance reduces to the bottleneck distance:

$$d_B(\{\mathcal{D}_q\}_{1\le q\le N-1}, \{\mathcal{D}'_q\}_{1\le q\le N-1}) = \sup_{1\le q\le N-1} \inf_{\gamma:\mathcal{D}_q\to\mathcal{D}'_q} \sup_{x\in\mathcal{D}_q} |x - \gamma(x)|. \tag{2.2.9}$$

The real number field $\mathbb{R}$ can be regarded as a poset category with the real numbers as objects and the binary relations $\le$ as morphisms. Recall that an $\mathbb{R}$-indexed diagram $\mathcal{F}$ in a category $\mathfrak{C}$ is a functor $\mathcal{F} : \mathbb{R} \to \mathfrak{C}$ from the poset category $\mathbb{R}$ to the category $\mathfrak{C}$. Let $\mathcal{F}^{\mathbb{R}}$ be the category of $\mathbb{R}$-indexed diagrams in $\mathfrak{C}$. Let $\Sigma : \mathcal{F}^{\mathbb{R}} \to \mathcal{F}^{\mathbb{R}}$ be a functor on the category of $\mathbb{R}$-indexed diagrams given by $(\Sigma^\varepsilon\mathcal{F})(a) = \mathcal{F}(a + \varepsilon)$.

Definition 2.2.1. Let $\mathcal{F}$ and $\mathcal{G}$ be two $\mathbb{R}$-indexed diagrams in a category $\mathfrak{C}$. We say $\mathcal{F}$ and $\mathcal{G}$ are ε-interleaved if there are natural transformations $\Phi : \mathcal{F} \to \Sigma^\varepsilon\mathcal{G}$ and $\Psi : \mathcal{G} \to \Sigma^\varepsilon\mathcal{F}$ such that $(\Sigma^\varepsilon\Psi) \circ \Phi = \Sigma^{2\varepsilon}|_{\mathcal{F}}$ and $(\Sigma^\varepsilon\Phi) \circ \Psi = \Sigma^{2\varepsilon}|_{\mathcal{G}}$.

Definition 2.2.2. Let $\mathcal{F}$ and $\mathcal{G}$ be two $\mathbb{R}$-indexed diagrams in a category $\mathfrak{C}$. The interleaving distance between $\mathcal{F}$ and $\mathcal{G}$ is defined by

$$d_I(\mathcal{F}, \mathcal{G}) = \inf\{\varepsilon \ge 0 | \mathcal{F} \text{ and } \mathcal{G} \text{ are } \varepsilon\text{-interleaved}\}. \tag{2.2.10}$$

Let f, g be two real-valued functions defined on a simplicial complex K. Then one has two filtrations of simplicial complexes. Let $\|f-g\|_\infty = \sup_{\sigma\in K} |f(\sigma)-g(\sigma)|$. Let $\mathcal{D}_q(K,f)$ and $\mathcal{D}_q(K,g)$ be the persistence diagrams of K filtered by f and g, respectively. We have the following result.

Theorem 2.2.2. Let K be a finite complex. Then $d_B(\{\mathcal{D}_q(K,f)\}_{1\leq q\leq N-1}, \{\mathcal{D}_q(K,g)\}_{1\leq q\leq N-1}) \leq \|f-g\|_\infty$.

Proof. We construct the proof based on the concepts developed in [15, 16, 17]. We consider Mayer persistent homology as the entities in the category $\mathbf{Vec}^{\mathbb{R}}$ of diagrams in the vector spaces category indexed by $\mathbb{R}$. Similarly, we regard Mayer persistence diagrams as the entities in the category $\mathbf{Mch}^{\mathbb{R}}$ of diagrams in the matching category indexed by $\mathbb{R}$. By [16, Theorem 1.7] and [16, Proposition 4.3], one has

$$d_B(\mathcal{D}_q(K,f), \mathcal{D}_q(K,g)) = d_I(\mathbf{H}_q(K,f), \mathbf{H}_q(K,g)) \tag{2.2.11}$$

Here, d_I denotes the interleaving distance for diagrams indexed by $\mathbb{R}$. For (K,f), we have a diagram $K^f : \mathbb{R} \to \mathbf{Simp}$ in the category of simplicial complexes given by $K^f_a = \{\sigma \in K | f(\sigma) \leq a\}$. Let $\varepsilon = \|f-g\|_\infty$. Then there are inclusions of simplicial complexes $K^f_a \hookrightarrow K^g_{a+\varepsilon}$ and $K^g_a \hookrightarrow K^f_{a+\varepsilon}$ for any real number a. Thus one has natural transformations $\Phi : K^f_\bullet \hookrightarrow K^g_{\bullet+\varepsilon}$ and $\Psi : K^g_\bullet \hookrightarrow K^f_{\bullet+\varepsilon}$ of $\mathbb{R}$-indexed diagrams. Here, $K_\bullet(a) = K_a$. By construction, we have

$$(\Sigma^\varepsilon \Psi) \circ \Phi = \Sigma^{2\varepsilon}|_{K^f_\bullet}. \tag{2.2.12}$$

Here, $\Sigma^\varepsilon\Psi : K^g_{\bullet+\varepsilon} \hookrightarrow K^f_{\bullet+2\varepsilon}$ is given by $(\Sigma^\varepsilon\Psi)(K^g_{\bullet+\varepsilon})(a) = K^f_{a+2\varepsilon}$ and $\Sigma^{2\varepsilon}|_{K^f_\bullet} : K^f_\bullet \to K^f_{\bullet+2\varepsilon}$ is given by $\Sigma^{2\varepsilon}|_{K^f_\bullet}(K^f_\bullet)(a) = K^f_{a+2\varepsilon}$. Similarly, one has $(\Sigma^\varepsilon\Phi)\circ\Psi = \Sigma^{2\varepsilon}|_{K^g_\bullet}$. It follows that K^f and K^g are ε-interleaved. By definition, we have $d_I(K^f, K^g) \leq \varepsilon$. By [15, Proposition 3.6] and Corollary 2.1.5, we have

$$d_I(\mathbf{H}_q(K,f), \mathbf{H}_q(K,g)) \leq d_I(K^f, K^g) \leq \varepsilon. \tag{2.2.13}$$

It follows that

$$d_B(\mathcal{D}_q(K,f), \mathcal{D}_q(K,g)) \leq d_I(K^f, K^g) \leq \varepsilon. \tag{2.2.14}$$

By the definition of bottleneck distance, one has

$$d_B(\{\mathcal{D}_q(K,f)\}_{1\le q\le N-1},\{\mathcal{D}_q(K,g)\}_{1\le q\le N-1})\le \|f-g\|_\infty. \tag{2.2.15}$$

The desired result follows. □

The aforementioned conclusion establishes the stability of persistent Mayer Betti numbers under the bottleneck distance. This guarantees that the persistence of Mayer Betti numbers is a steadfast and resilient topological feature, resistant to noise.

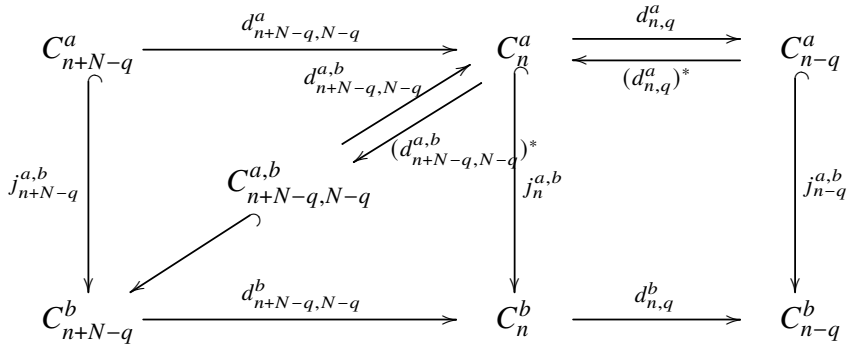
2.2.3 Persistent Mayer Laplacians

Let $\{K_{a_i}\}_{i\ge 1}$ be a filtration of simplicial complexes. Endow $C_*(K_{a_m};\mathbb{C})$ with an inner product structure over $\mathbb{C}$. Consequently, as subspaces, each $C_*(K_{a_i};\mathbb{C})$ inherits the inner product structure of $C_*(K_{a_m};\mathbb{C})$.

Consider the inclusion $j_{a,b}: K_a\to K_b$ of simplicial complexes. By Proposition 2.1.4, we have a morphism $C_*(j_{a,b}): C_*(K_a;\mathbb{C})\to C_*(K_b;\mathbb{C})$ of N-chain complexes. For the sake of simplicity, we denote $C_n^a=C_n(K_a;\mathbb{C})$ with the corresponding Mayer differential d_n^a, and denote $j_n^{a,b}=C_n(j_{a,b})$. Moreover, we denote $d_{n,q}^a=d_{n-q+1}^a\cdots d_{n-1}^a d_n^a: C_n^a\to C_{n-q}^a$. Let

$$C_{n,q}^{a,b}=\{x\in C_n^b \mid d_{n,q}^b x\in C_{n-q}^a\},\quad 1\le q\le N-1. \tag{2.2.16}$$

It follows that $C_{n,q}^{a,b}$ is a subspace of C_n^b with the subspace inner product. Besides, we have a linear map $d_{n,q}^{a,b}: C_{n,q}^{a,b}\to C_{n-q}^a$ given by $d_{n,q}^{a,b}(x)=d_{n,q}^b x$.



(2.2.17)

The (a,b)-*persistent Mayer Laplacian* $\Delta_{n,q}^{a,b}: C_n^a\to C_n^a$ is defined by

$$\Delta_{n,q}^{a,b}:=(d_{n,q}^a)^*\circ d_{n,q}^a+d_{n+N-q,N-q}^{a,b}\circ(d_{n+N-q,N-q}^{a,b})^*. \tag{2.2.18}$$

In particular, if $n < q$, the persistent Mayer Laplacian is reduced to $\Delta_{n,q}^{a,b} = d_{n+N-q,N-q}^{a,b} \circ (d_{n+N-q,N-q}^{a,b})^*$. We arrange the positive eigenvalues of $\Delta_{n,q}^{a,b}$ in ascending order as follows:

$$\lambda_{n,q}^{a,b}(1), \lambda_{n,q}^{a,b}(2), \ldots, \lambda_{n,q}^{a,b}(r), \tag{2.2.19}$$

where r is the number of positive eigenvalues. Specifically, $\lambda_{n,q}^{a,b}(1)$ denotes the smallest positive eigenvalue, serving as the spectral gap and bearing close relevance to the Cheeger constant in geometry.

Recall that for simplicial homology, the harmonic component of the persistent Laplacian and persistent homology are isomorphic. Similarly, the harmonic component of the persistent Mayer Laplacian and persistent Mayer homology are also isomorphic. This is presented follows.

Theorem 2.2.3. For any $a \leq b$, we have an isomorphism $\ker \Delta_{n,q}^{a,b} \cong H_{n,q}^{a,b}$, where $n \geq 0$ and $1 \leq q \leq N-1$.

Proof. Note that $d_{n,q}^{a} \circ d_{n+N-q,N-q}^{a,b} = 0$. The result follows from [14, Proposition 3.1]. □

The above theorem indicates that, within the Mayer homology theory, the persistent Mayer Laplacian contains more information than persistent Mayer homology. The persistent Mayer Laplacian reflects the geometric characteristics of complexes. It can be easily proven that the eigenvalues of the persistent Mayer Laplacian are non-negative. We arrange the positive eigenvalues in ascending order, denoting them as $\lambda_{n,q}(1), \ldots, \lambda_{n,q}(r)$. Here, r is the number of positive eigenvalues. Typically, attention is often focused on the smallest positive eigenvalue, the largest positive eigenvalue, the average value of eigenvalues, and similar information. In this paper, our examples and applications will involve computing the smallest eigenvalue.

2.2.4 Mayer features on Vietoris-Rips complexes

Let X be a finite set of points embedded in Euclidean space. It is always possible to construct a filtration of simplicial complexes. Common constructions include Vietoris-Rips complexes, alpha complexes, cubical complexes, and others. These complexes offer diverse topological descriptions for datasets. Now, we will focus on exploring the Mayer features on Vietoris-Rips complexes.

Given a real number ϵ, the Vietoris-Rips complex on X is given by the simplicial complex

$$\mathcal{VR}_\epsilon = \{\sigma \subseteq X | \text{every pair of points in } \sigma \text{ has a distance not larger than } \epsilon\}. \tag{2.2.20}$$

From the Vietoris-Rips complex, one can derive the N-chain complex $C_*(\mathcal{VR}_\epsilon; \mathbb{C})$. Furthermore, for any real numbers $\epsilon \leq \epsilon'$, the inclusion $\mathcal{VR}_\epsilon \hookrightarrow \mathcal{VR}_{\epsilon'}$ induces the inclusion $C_*(\mathcal{VR}_\epsilon; \mathbb{C}) \hookrightarrow C_*(\mathcal{VR}_{\epsilon'}; \mathbb{C})$ of N-chain complexes. It leads to the persistent Mayer homology

$$H_{n,q}^{\epsilon,\epsilon'} = \text{im}\,(H_{n,q}(\mathcal{VR}_\epsilon; \mathbb{C}) \to H_{n,q}(\mathcal{VR}_{\epsilon'}; \mathbb{C})), \quad n \geq 0. \tag{2.2.21}$$

and the persistent Mayer Laplacian based on the Vietoris-Rips complexes, serving as the primary tool in our work.

Example 2.2.3. Consider the example where X_1 consists of the following seven points on a plane

$$(0,0), (1,1), (1,-1), (2,1), (2.5,1.5), (2.5,0.5), (3,1). \tag{2.2.22}$$

Here, we exhibits a visualization of some of the corresponding Vietoris-Rips complexes in

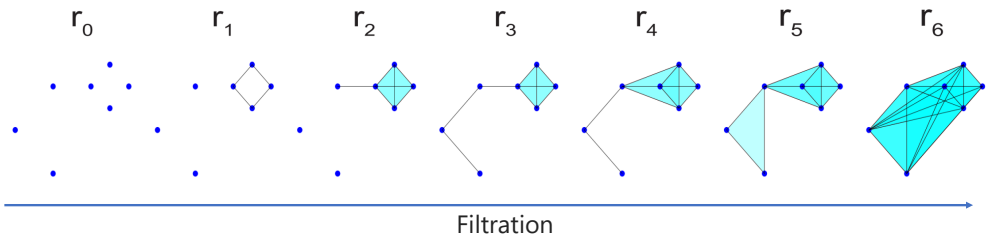


Figure 2.3 Illustration of the Vietoris-Rips complexes at different filtration radius for pointset X_1. Note that for the point set X_1 in this example, we can obtain a maximum of 12 Vietoris-Rips complexes with different filtration radius. For simplicity, we have omitted 5 complexes between r_5 and r_6.

Figure 2.3, labeled by their filtration radius, namely r_0 to r_6, respectively. In this example, the topological features we employed from the Mayer features include the Betti numbers at dimension 0 and 1. We display comparisons of calculation results of the persistent Mayer homology of the Vietoris-Rips complexes derived from the set X with different N values.

We first compare the case $N = 2$ with $N = 3$, shown in Figure 2.4. The $N = 2$ case, which also represents the classical persistent Betti numbers, exhibit fewer topological features than the

persistent Mayer Betti numbers for $N = 3$ case. Specifically, the classical ($N = 2$) persistent homology can yield non-trivial Betti numbers for dimensional 0 and 1 at filtration radius r_0,r_1,r_2, and r_1, respectively. In contrast, for $N = 3$ case, the persistent Mayer homology reveals non-trivial Mayer Betti number 0 at r_0 ($q = 1$ and $q = 2$), r_1 ($q = 1$ and $q = 2$), r_2 ($q = 1$ and $q = 2$), r_3 ($q = 1$), r_4 ($q = 1$), r_5 ($q = 1$), and r_6 ($q = 1$). Additionally, the $N = 3$ case yields non-trivial Mayer Betti number 1 at r_1 ($q = 1$ and $q = 2$), r_2 ($q = 1$ and $q = 2$), r_3 ($q = 1$ and $q = 2$), r_4 ($q = 1$ and $q = 2$), r_5 ($q = 1$ and $q = 2$), and r_6 ($q = 1$).

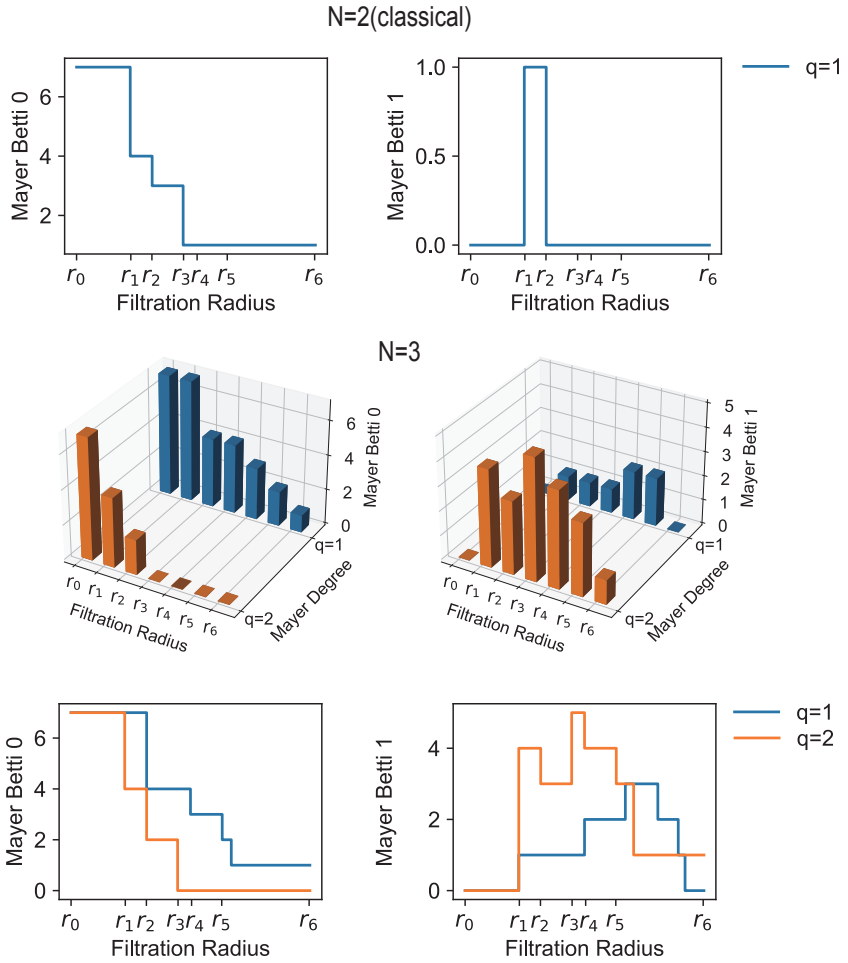


Figure 2.4 Comparison of persistent Betti numbers between the cases $N = 2$, $N = 3$.

While in other cases, such as $N = 5$, and $N = 7$, more topological features are encompassed. As illustrated in Figure 2.5, we consistently observe $N - 1$ Betti curves, each reflecting distinct topological information. To provide a more accurate description of the information content in

the Betti curves obtained for different values of N, we conducted a statistical analysis of the variations in Betti 0 and Betti 1 for different values of N, shown in Table 2.4. We observe that with the increase in the value of N, the quantities of Betti 0 variations and Betti 1 variations strictly and positively increase. The increasing effect is more pronounced for Betti 1, indicating that, unlike the information obtained from the classical persistent homology of Rips complexes, the one-dimensional information provided by persistent Mayer homology also plays a crucial role.

Additionally, it is noteworthy that the average Betti variation in Table 2.4 indicates that, for the majority of cases, increasing the value of N not only results in obtaining more Betti curves but also enhances the topological information of each Betti curve. The only exception is the case of Betti 0 for $N = 7$. This is primarily due to the fact that the point set considered in this example contains only 7 points, leading to a sparse existence of high-dimensional simplices in the corresponding Vietoris-Rips complex. In Mayer homology, Betti 0 variation implies that 0-dimensional simplices are killed by some higher-dimensional simplices. If the number of higher-dimensional simplices is too sparse, the difficulty of eliminating 0-dimensional simplices increases, leading to a reduction in the quantity of variations. However, in application scenarios, the number of points in the point set is generally much larger than the value of N. In such cases, we can typically expect an increase in the average Betti variations.

N value	Betti 0 variations	Avg. Betti 0 variations	Betti 1 variations	Avg. Betti 1 variations
2	3	3	2	2
3	7	3.5	12	6
5	15	3.75	33	8.25
7	17	2.83	54	9

Table 2.4 A statistics of the Mayer Betti curves variation for different N value.

Example 2.2.4. In this example, we show the comparison of Betti numbers and the smallest eigenvalues for the non-harmonic components of the Laplacians for the case $N = 5$. Here, we consider example where points are distributed on the vertices of a three-dimensional cube. Let X_2 be a set with points given by

$$(0, 0, 1.3), (0, 0, -1), (0, 1, 0), (0, -1, 0), (1, 0, 0), (-1, 0, 0). \tag{2.2.23}$$

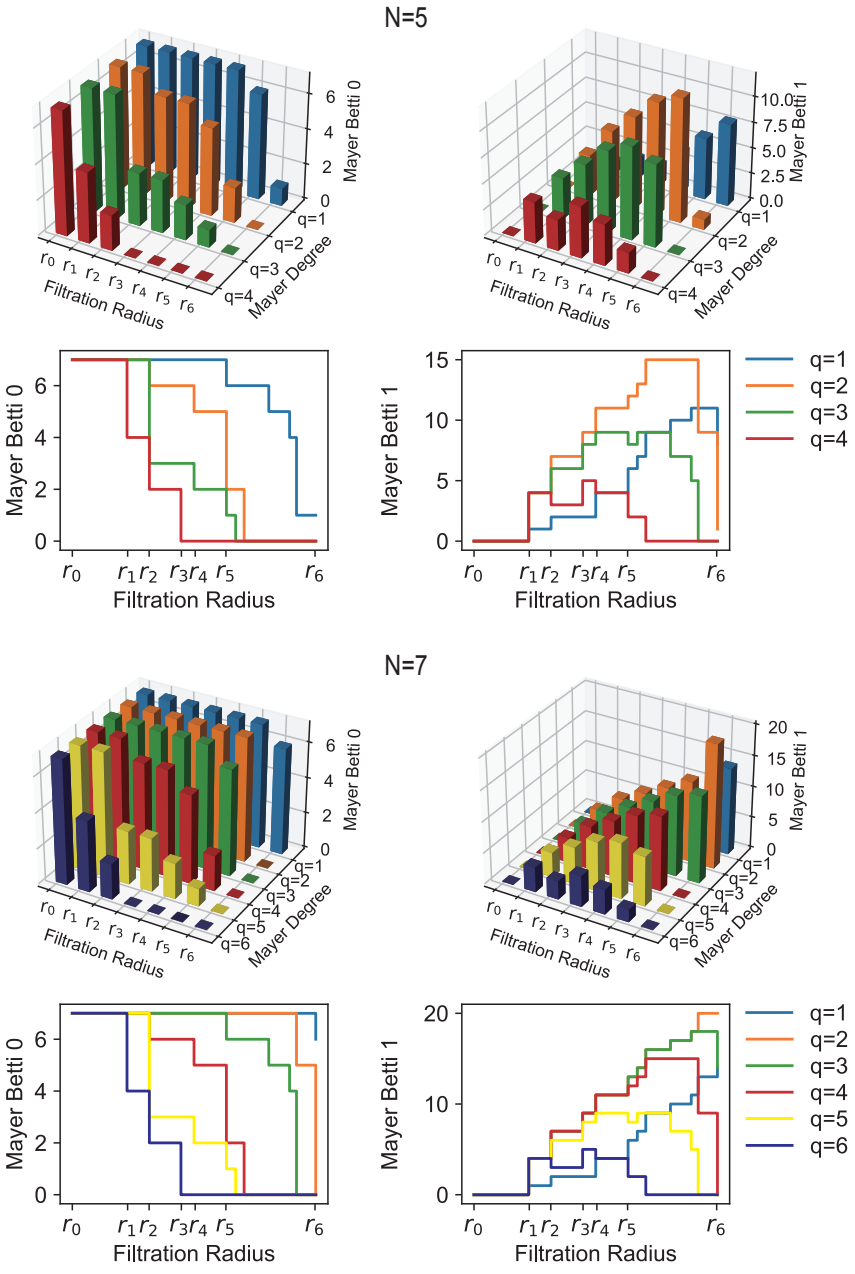


Figure 2.5 Illustration of persistent Betti numbers between the cases $N = 5$, $N = 7$. The Mayer degree, denoted by q, refers to the stage of Mayer homology.

Figure 2.6 shows the visualization of the Vietoris-Rips complexes.

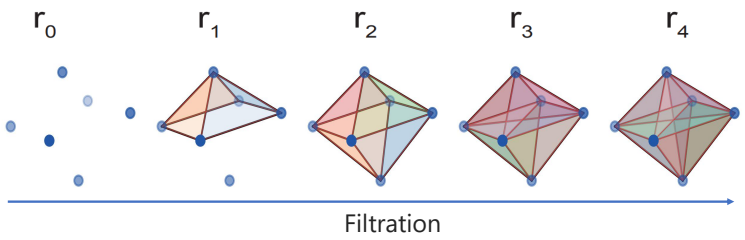


Figure 2.6 Illustration of the Vietoris-Rips complexes at different filtration radius for pointset X_2.

We are interested to know whether persistent Mayer Laplacian detects more geometric variations than persistent Mayer homology in characterizing data. To this end, we compare the persistent Betti numbers and the smallest non-zero eigenvalues of persistent Mayer Laplacians derived from X_2 for the case $N = 2$, $N = 3$, and $N = 5$ as shown in Figure 2.7, Figure 2.8, and Figure 2.9, respectively. Since the harmonic spectra of persistent Mayer Laplacians fully recovery the topological information of persistent Mayer homology, attention is given to whether Mayer Laplacian's non-zero eigenvalue can detect additional variations compared to Mayer Betti numbers. Our results are summarized in Table 2.5. After comparison, we observe that the classical ($N = 2$) Laplacian's nonharmonic spectra can detect more variations in both dimension 0 and 1. While Mayer Laplacian's first nonzero eigenvalue is superior in dimension 0 for all $N = 3$ cases, and $N = 5, q = 2$, $N = 5, q = 3$, $N = 5, q = 4$ cases, and in dimension 1 for $N = 3, q = 2$, $N = 5, q = 1$, and $N = 5, q = 4$ cases. It performs on par with Mayer Betti number in dimension 0 for $N = 5, q = 1$, in dimension 1 for $N = 3, q = 1$. In addition, Mayer Laplacian's first nonzero eigenvalue captures fewer variations than Mayer Betti number does in dimension 1 for $N = 5, q = 2$ and $N = 5, q = 3$. In summary, Mayer Laplacian exhibits superior performance compared to Mayer Betti numbers, confirming that persistent Mayer Laplacian indeed provides richer information compared to persistent Mayer homology.

A more detailed analysis reveals that the reason for the use of Mayer Laplacian lies in its inability to detect the variations from r_0 to r_1 and from r_1 to r_2 in the 1-dimensional case for $N = 5, q = 2$ and $N = 5, q = 3$. In both of these scenarios, the smallest eigenvalues of persistent Laplacians are

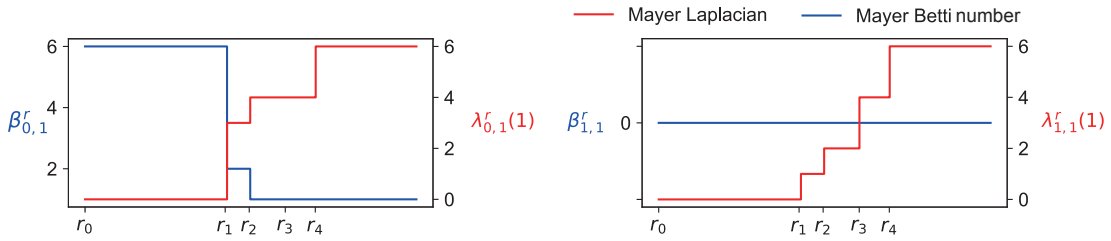


Figure 2.7 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for the case that $N = 2$ (classical). The blue curves denote the Betti curves, while the red curves represent changes of the smallest eigenvalues. The notion $\beta^r_{n,q}$ denotes the n-dimensional Betti number at stage q of the Vietoris-Rips complex at distance r. The notion $\lambda^r_{n,q}(1)$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta^r_{n,q}$ at distance parameter r.

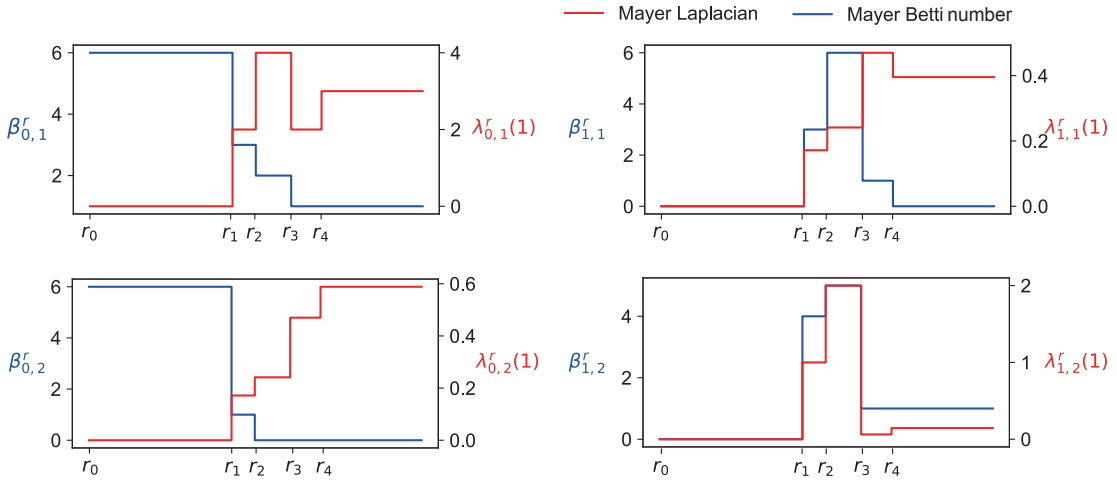


Figure 2.8 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for the case that $N = 3$. The blue curves denote the Betti curves, while the red curves represent changes of the smallest eigenvalues. The notion $\beta^r_{n,q}$ denotes the n-dimensional Betti number at stage q of the Vietoris-Rips complex at distance r. The notion $\lambda^r_{n,q}(1)$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta^r_{n,q}$ at filtration parameter r.

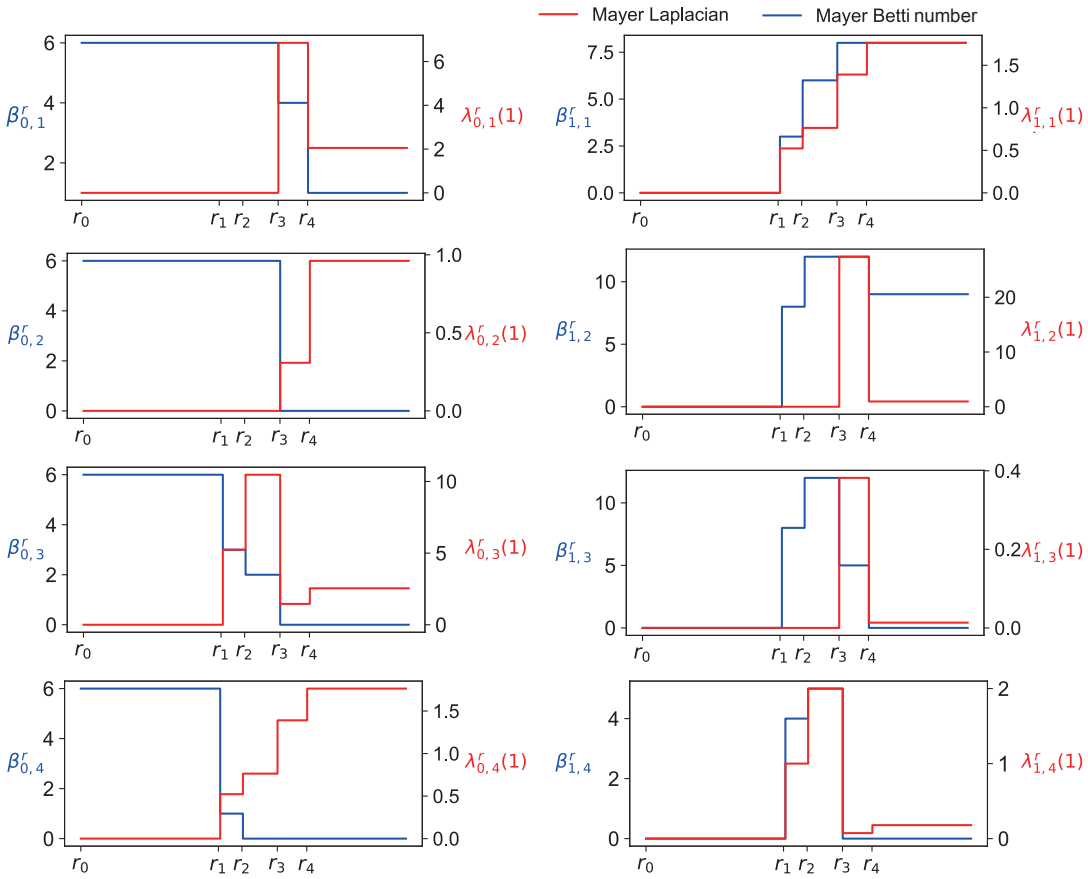


Figure 2.9 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for the case that $N = 5$. The blue curves denote the Betti curves, while the red curves represent changes of the smallest eigenvalues. The notion $\beta^r_{n,q}$ denotes the n-dimensional Betti number at stage q of the Vietoris-Rips complex at distance r. The notion $\lambda^r_{n,q}(1)$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta^r_{n,q}$ at filtration parameter r.

consistently 0. This indicates that, in these cases, all 1-dimensional simplices precisely serve as representatives of some Mayer homology classes. Therefore, we believe that while persistent Mayer Laplacian's first eigenvalue can offer more information compared to persistent Mayer homology, it is not sufficient to replace the latter. The combination of both harmonic and non-harmonic spectra is necessary to achieve better results in practical applications.

Mayer features	$N=2$ $q=1$	$N=3$ $q=1$	$N=3$ $q=2$	$N=5$ $q=1$	$N=5$ $q=2$	$N=5$ $q=3$	$N=5$ $q=4$
$\beta_{0,q}$	2	3	2	2	1	3	2
$\lambda_{0,q}(1)$	3	4	4	2	2	4	4
$\beta_{1,q}$	0	4	3	3	3	4	3
$\lambda_{1,q}(1)$	4	4	4	4	2	2	4

Table 2.5 A comparison of variation detection of the Mayer Betti numbers with the Mayer Laplacian's first non-zero eigenvalues for $N = 2, 3$, and 5.

2.2.5 Applications

In this section, we will compute the persistent Mayer Betti numbers and spectral gaps of Mayer Laplacians for fullerene C_{60} and cucurbit[7]uril CB7. We use the atomic coordinates of molecules as spatial points to construct the Vietoris-Rips complex, and then build an N-chain complex on it. Typically, we consider the cases $N = 2$, $N = 3$, and $N = 5$. Here, N represents the integer that $d^N = 0$. We focus on the Mayer Betti numbers denoted as $\beta_{n,q}$ and the smallest positive eigenvalues of Mayer Laplacians (spectral gaps) denoted as $\lambda_{n,q}(1)$. In this work, n denotes the dimension of Mayer homology or Mayer Laplacians, and we always compute the Mayer Betti numbers and the spectral gaps of Mayer Laplacians for dimensions 0 and 1. The parameter q refers to the subscript of Mayer homology or Mayer Laplacians, representing the q-th stage, where $1 \leq q \leq N-1$. Specifically, for the case of $N = 2$, we obtain the usual simplicial homology and its corresponding Laplacian, where q can only take the value of 1. This implies that for a given dimension n, there is only one homology group and one Laplacian operator.

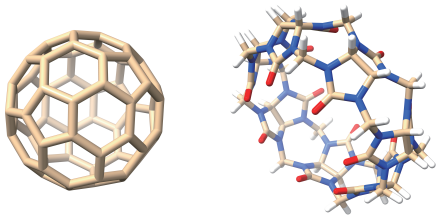


Figure 2.10 Structures of the fullerene C_{60} (Left) and the cucurbit[7]uril CB7 (Right).

In the depicted 3D structure showcased in Figure 2.10, the fullerene C_{60} is presented as a carbon molecule with a distinctive soccer ball-like arrangement, comprising 60 carbon points.

In contrast, the macrocyclic compound cucurbit[7]uril (CB7) is intricately composed of 126 points, encompassing carbon, hydrogen, oxygen, and nitrogen atoms. Given the more symmetrical and concise configuration of C_{60} in comparison to the complex structure of CB7, an effective featurization method is anticipated to reveal more nuanced patterns for CB7.

In Figure 2.11 and Figure 2.12, as well as Figure 2.13 and Figure 2.14, distinct colors represent the numerical values of different Betti numbers and spectral gaps. The structural differences between C_{60} and CB7 are readily apparent from the comparisons in Figure 2.11 with Figure 2.13, and Figure 2.12 with Figure 2.14. The persistent Mayer Betti numbers and persistent Mayer Laplacians of CB7 display more intricate patterns, and the critical points of variation in these patterns involve a broader range of filtration radius. This highlights the potential of persistent Mayer homology and persistent Mayer Laplacian as highly effective tools for featuring molecular structures.

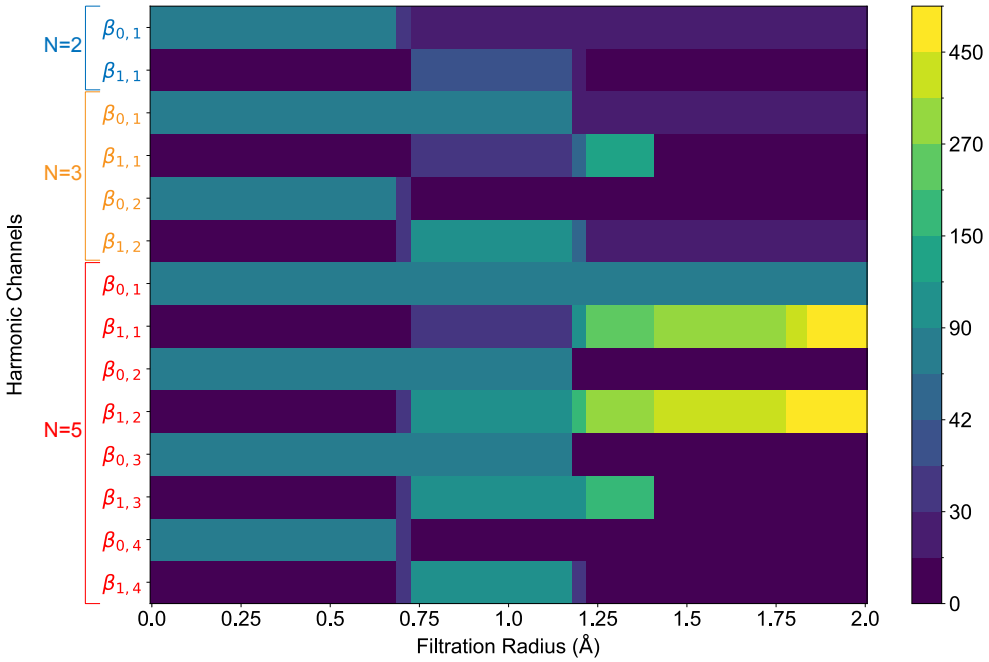


Figure 2.11 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for fullerene C_{60} in cases where $N = 2$, $N = 3$, and $N = 5$. Here, $\beta_{n,q}$ denotes the n-dimensional Betti number at stage q for a given distance parameter. Similarly, $\lambda_{n,q}$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta_{n,q}$ at a given distance parameter.

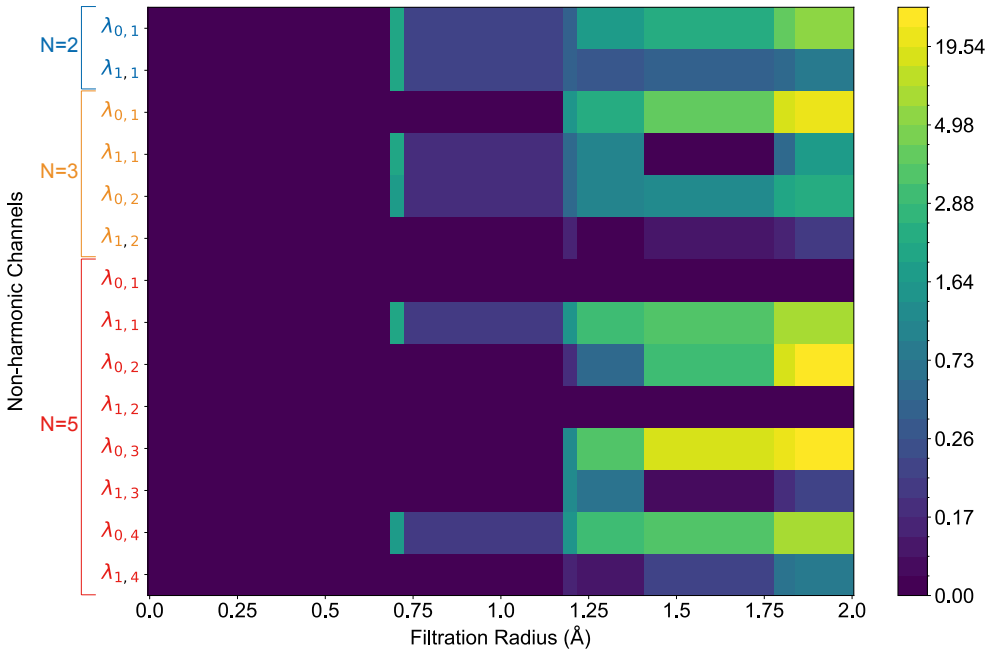


Figure 2.12 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for fullerene C_{60} in cases where $N = 2$, $N = 3$, and $N = 5$. Here, $\beta_{n,q}$ denotes the n-dimensional Betti number at stage q for a given distance parameter. Similarly, $\lambda_{n,q}$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta_{n,q}$ at a given distance parameter.

In the above calculations, for convenience, we computed the persistent Betti numbers and persistent spectral gaps of the 3-skeleton of the Vietoris-Rips complex. However, this does not hinder us from obtaining the topological and geometric characteristics of the structure. In the figures, we observe that for the case of $N = 2$, the Betti numbers provide relatively limited information, while the spectral gaps can complement the geometric information. For the cases of $N = 3$ and $N = 5$, the information contained in the Betti numbers alone is already comparable to the combined information of Betti numbers and spectral gaps for the $N = 2$ case. This implies that, for larger values of N, computing Mayer Betti numbers alone is sufficient to capture the sum of harmonic and non-harmonic information present in the $N = 2$ case. Generally, computing Betti numbers is much faster than solving for spectral gaps, providing a more efficient approach for calculating geometric features.

Despite the calculation cost of persistent Mayer Laplacian, which should be approximately $N - 1$ times that of the classical persistent Laplacian if we omit some of matrix multiplications,

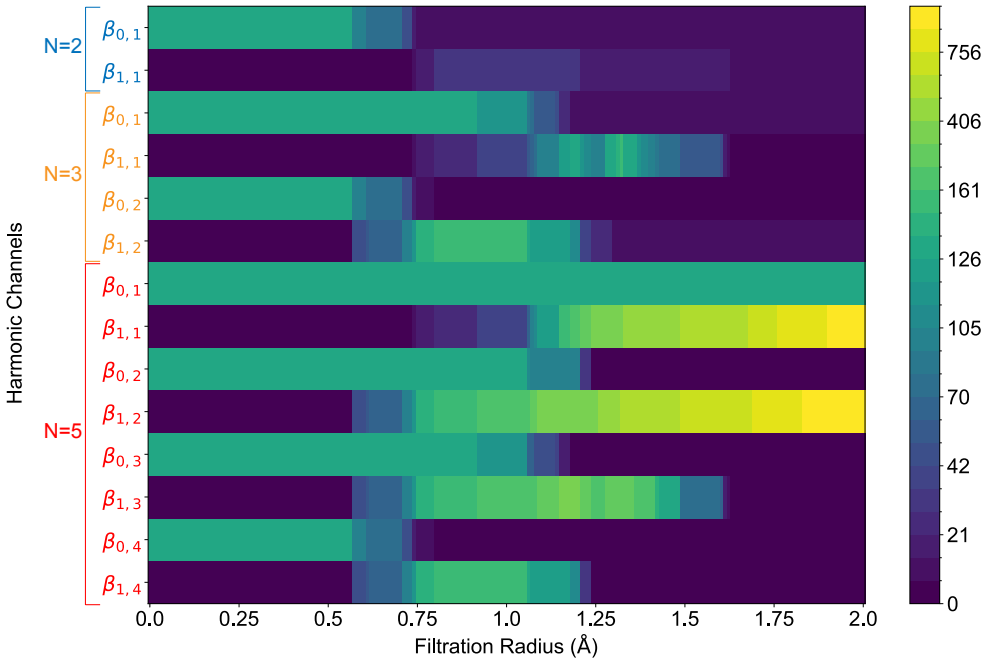


Figure 2.13 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for cucurbit[7]uril CB7 in cases where $N = 2$, $N = 3$, and $N = 5$. Here, $\beta_{n,q}$ denotes the n-dimensional Betti number at stage q for a given distance parameter. Similarly, $\lambda_{n,q}$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta_{n,q}$ at a given distance parameter.

the persistent Mayer homology and persistent Mayer Laplacian, from an applied perspective, successfully provide practical multichannel featurization technique. As in applications, it is essential to obtain effective features of sufficient dimensionality before engaging in machine learning tasks, especially when dealing with datasets containing thousands or even millions of samples.

Traditional persistent homology and persistent Laplacian methods can only increase the feature dimensionality by adding more filtrations. This approach faces two main challenges. Firstly, there is an upper limit to the number of filtrations that can be added, and the computational cost becomes prohibitively high when dealing large filtration. Secondly, even with an increased number of filtrations, it does not guarantee the acquisition of useful information. This issue significantly impacts persistent homology, especially in higher dimensions (1-dimensional and above). In such scenarios, to obtain the desired features, it is common to divide the data into subgroups based on the physical understanding. For example, element-specific persistent homology considers different types of elements in the data [3]. Persistent Laplacians not only consider the smallest positive

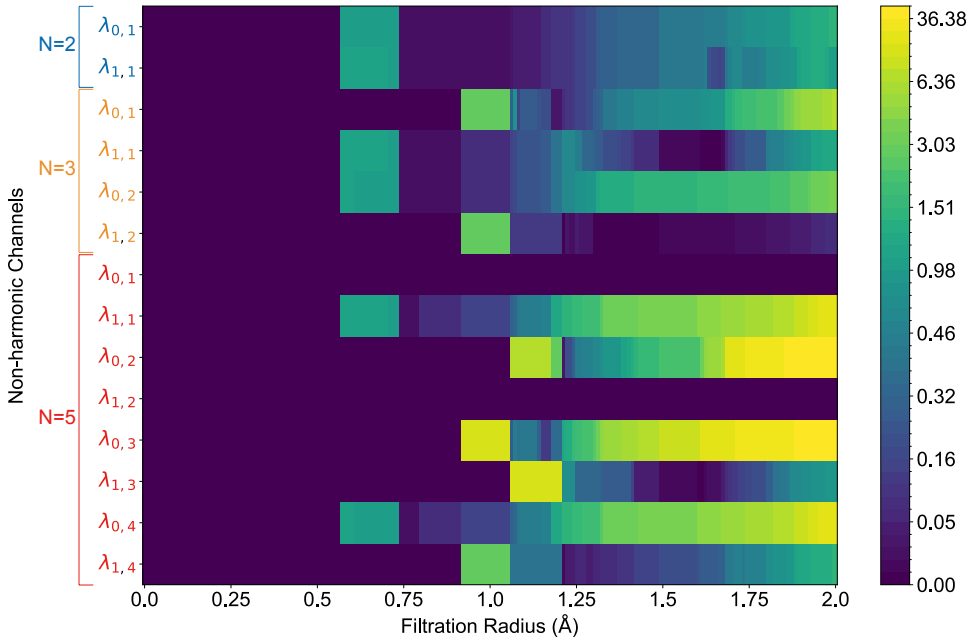


Figure 2.14 Comparison of persistent Betti numbers and the smallest positive eigenvalues of persistent Laplacians for cucurbit[7]uril CB7 in cases where $N = 2$, $N = 3$, and $N = 5$. Here, $\beta_{n,q}$ denotes the n-dimensional Betti number at stage q for a given distance parameter. Similarly, $\lambda_{n,q}$ represents the smallest eigenvalue of the non-harmonic component of the Laplacian $\Delta_{n,q}$ at a given distance parameter.

eigenvalue but also take into account the largest eigenvalue and some statistical measures of the positive eigenvalues [18].

Persistent Mayer homology and persistent Mayer Laplacian possess Mayer degrees, serving as an additional dimension. By selecting specific values of N, we can effortlessly expand the feature dimensionality by a factor of $N - 1$. Moreover, as the value of N increases, each Mayer degree can have additional effective filtration choices for its corresponding features. As shown in Figure 2.11 and Figure 2.13, more patterns in the persistent Mayer Betti numbers as N increases.

2.3 Mayer-homology learning prediction of protein-ligand binding affinities

As mentioned early, Mayer homology of simplicial complex reduces to simplicial homology when N is taken to 2. We will begin with a brief review of simplicial complexes, the classical homology of simplicial complexes, and then generalize the discussion to Mayer homology.

Simplicial complex is a well-known topological model in data science, with notable examples including the Vietoris-Rips complex, Čech complex, and Alpha complex. A simplicial complex

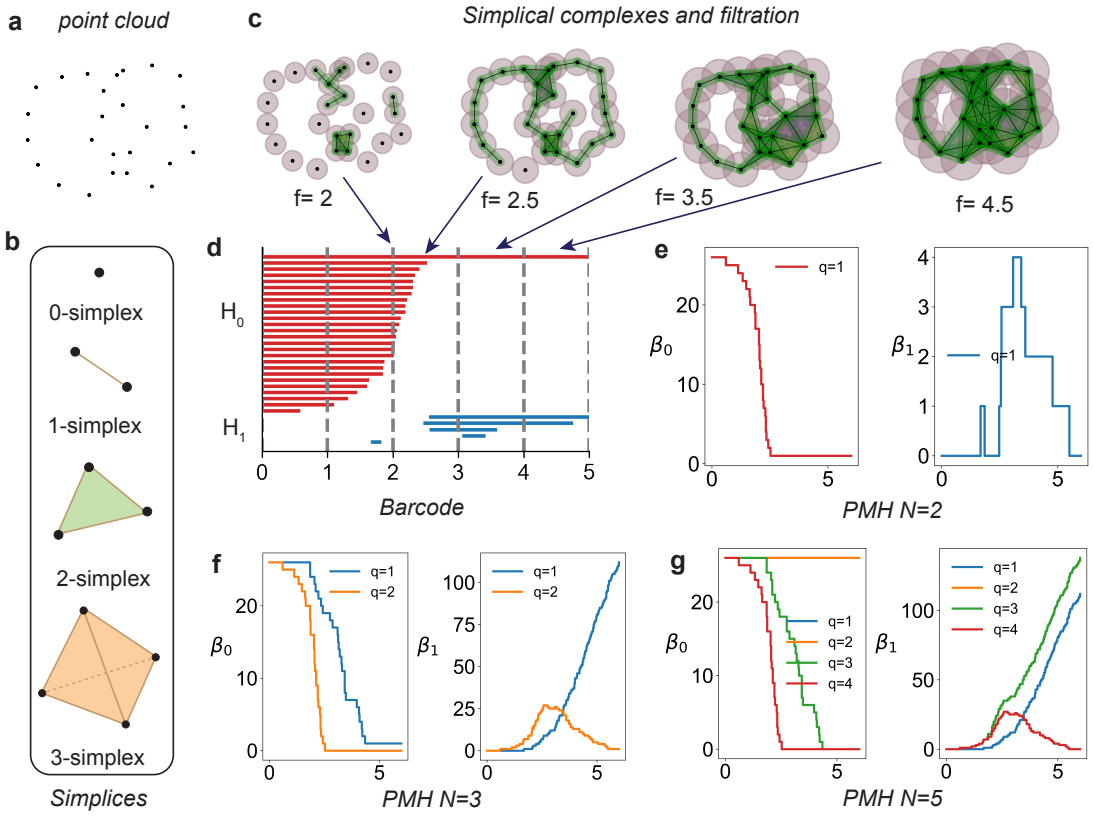


Figure 2.15 The persistent Mayer homology representation for a point cloud based on VR complex. a: A 2D point cloud. b: The representation of simplices in dimension $n = 0, 1, 2, 3$. c: A filtration of simplicial complexes obtained from the point cloud. d: The barcode of dimension 0 and 1 corresponding to the filtration process in c. The filtration parameter is defined to be the diameter of circles around given points. e: The Betti numbers β_0 and β_1 calculated from persistent Mayer homology (PMH) for $N = 2$. f: The Betti numbers $\beta_{0,q}$ and $\beta_{1,q}$ calculated from persistent Mayer homology (PMH) for $N = 3$ ($q = 1, 2$). g: The Betti numbers $\beta_{0,q}$ and $\beta_{1,q}$ calculated from persistent Mayer homology (PMH) for $N = 5$ ($q = 1, 2, 3, 4$). The curves for $\beta_{0,1}$ and $\beta_{0,2}$ coincide. The curves for $\beta_{1,2}$ and $\beta_{1,3}$ coincide.

is composed of a collection of simplices following specific combinatorial rules. An n-simplex is the convex hull formed by $n + 1$ geometrically independent points. For example, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle (with a solid interior), and a 3-simplex is a solid tetrahedron, as illustrated in Figure 2.15b.

The key idea of persistent homology is to introduce multi-scale information, which is provided by the filtration of simplicial complexes. For a given point cloud data set, the most common filtration of simplicial complexes is the Vietoris-Rips (VR) complex, as illustrated in Figure 2.15c. Topological features at different scales exhibit a certain kind of persistence, meaning that homology

generators at smaller scales may persist as homology generators at larger scales, thereby giving rise to persistent homology generators. The scale at which a generator is born is referred to as its birth time, while the scale at which it disappears is known as its death time. The topological features of persistent homology are represented by bars that record the birth and death times of homology generators, as shown in Figure 2.15d, corresponding to the barcode of the filtration of simplicial complexes in Figure 2.15d.

Unlike classical homology theories, the Mayer homology theory explored in this study has a generalized differential $d^N = 0$ with an integer $N \geq 2$ on the N-chain complex. This approach allows us to obtain a family of homology groups $H_{n,q}(K)$ for a simplicial complex, where n is the dimension and $1 \leq q \leq N-1$ corresponds to the Mayer degree. The homology groups $H_{n,q}(K)$ are referred to as Mayer homology. The Betti numbers associated with Mayer homology are termed the *Mayer Betti numbers* of the simplicial complex, denoted by $\beta_{n,q}$. For $N = 2$, the Mayer degree q can only be $q = 1$, which means that for a fixed dimension n, there is only one homology group, which is consistent with the usual homology groups of a simplicial complex. For general N, Mayer homology reveals more information than classical homology, offering potentially valuable geometric and topological features for applications. Beyond contributing to a unified mathematical framework for homology theory, Mayer homology and the associated Betti numbers provide valuable tools for analyzing the topological space of a given data set.

The Betti numbers for each simplicial complex are recorded in the barcode diagram shown in Figure 2.15d. For example, the number of red lines in Figure 2.15d at a filtration parameter of 2 corresponds to β_0. The Betti number $\beta_n : [0, +\infty) \to \mathbb{N}$ can be regarded as a function with the filtration parameter as its variable. Such a function is referred to as a *Betti curve*. Figure 2.15e shows the Betti curves for $N = 2$, with the red line representing the Betti curve β_0 and the blue line representing the Betti curve β_1. Additionally, Figure 2.15f and Figure 2.15g present the Betti curves for Mayer homology with $N = 3$ and $N = 5$, respectively. Each plot contains multiple curves because, in the case of Mayer homology, $\beta_{n,q}$ forms a curve for each $1 \leq q \leq N-1$. It is worth noting that when $N = 5$, the $\beta_{0,1}$ and $\beta_{0,2}$ align with each other and the $\beta_{1,2}$ and $\beta_{1,3}$ align with each

other as shown in Figure 2.15g. The comparison of these figures highlights the richer topological and geometric features of Mayer Betti numbers.

2.3.1 PMH-based element interactive molecular representation

Atomic coordinates in molecules can be viewed as point cloud data. Persistent Mayer homology is well-suited for characterizing molecular structures, and a multiscale topological representation can be obtained through a filtration process. The resulting persistent features effectively capture the hierarchical and multiscale properties of biomolecular structures and interactions. Various intramolecular and intermolecular interactions exist within molecular structures, characterized by different forces such as covalent bonds, van der Waals forces, electrostatic interactions, hydrophobic interactions, and hydrophilic interactions. To this end, we follow the element interaction characterization for pairwise atom groups [19] and use persistent Mayer homology to analyze these element-specific topological data structures. A cutoff distance of 12 Å is applied to extract the protein atoms around the ligand, considering that intermolecular interactions predominantly occur in the binding pocket region.

Figure 2.16b displays the PMH ($N = 2$) barcodes for C-C and O-C atom groups in the protein-ligand complex (PDBID: 1A94), with the simplicial complex constructed using the alpha complex. The persistence and variance of the β_0, β_1, and β_2 information are revealed. The ligand has more carbon atoms than oxygen atoms, leading to the faster decay of the β_0 value during filtration for C-C atom groups. Persistent attributes associated with β_1 and β_2 are also distinguishable in the characterization of C-C and O-C atom groups. The Betti curves of different dimensions are for these two atom groups as shown in Figure 2.16c and Figure 2.16d, respectively. The changes in $\beta_{n,q}$ values from PMH with $N = 3$ and $N = 5$ for C-C groups are shown in Figure 2.16e and Figure 2.16g. The changes for O-C groups are exhibited in Figure 2.16f and Figure 2.16h. Unlike the PMH characterization for 2D point clouds, which shows overlapping curves, there are distinct $\beta_{0,q}$ or $\beta_{1,q}$ curves in Figure 2.16g and Figure 2.16h for $N = 5$. These PMH ($N = 3$ or $N = 5$) Betti changes for these atom groups tend to plateau when the filtration parameter reaches 10 Å, or even as early as 5 Å. Therefore, it is sufficient to collect the Betti information with the filtration parameter

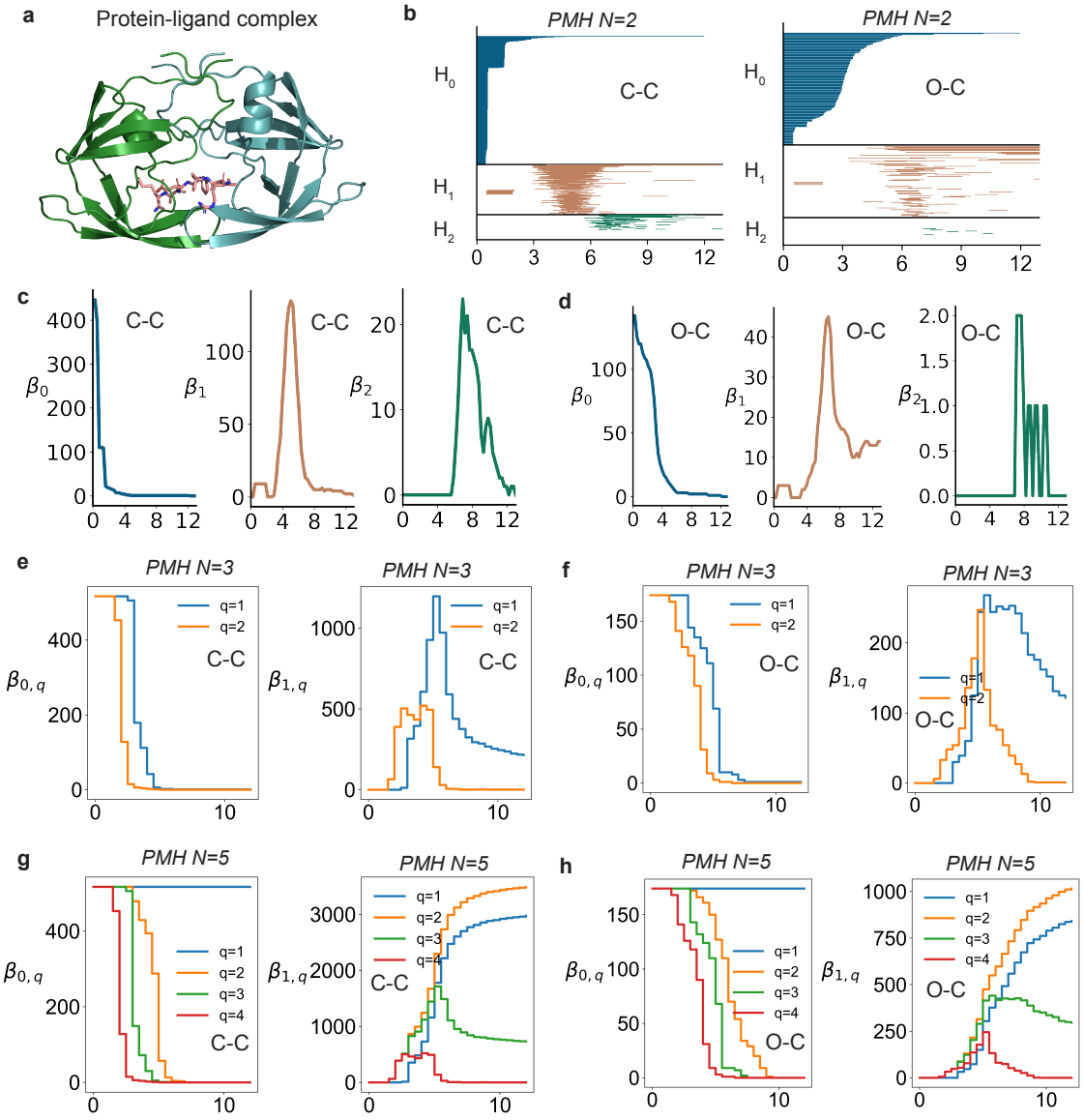


Figure 2.16 Persistent Mayer homology characterization for a protein-ligand complex (PDBID: 1A94) on alpha complex. a: The 3D structure of protein 1A94. b: The barcodes of different dimensions for a pair of atom sets in protein 1A94 with PMH (N=2). The first letter in C-C or O-C stands from atom group from protein and the second one indicates atom group from the ligand. c: The Betti curves of different dimensions for the C-C atom group in b. d: The Betti curves of different dimensions for the O-C atom group in b. e: The $\beta_{0,q}$ and $\beta_{1,q}$ curves for the C-C atom groups in b using PMH with N=3 (q=1,2). f: The $\beta_{0,q}$ and $\beta_{1,q}$ curves for the O-C atom groups in b using PMH with N=3 and (q=1,2). g: $\beta_{0,q}$ and $\beta_{1,q}$ curves for the C-C atom group in b using PMH with N=5 (q=1, 2,3,4). h: The $\beta_{0,q}$ and $\beta_{1,q}$ curves for the O-C atom group in b using PMH with N=5 (q=1, 2,3,4).

ranging from 0 Å to 10 Å. For PMH ($N = 2$) or traditional persistent homology characterization of the protein-ligand complex, persistent attributes analysis extends to an upper filtration parameter

of 12 Å.

It is observed that the $\beta_{0,1}$ and $\beta_{0,2}$ curves in Figure 2.16e resemble the $\beta_{0,3}$ and $\beta_{0,4}$ curves in Figure 2.16g. A similar pattern is seen between Figure 2.16f and Figure 2.16h. However, there are subtle numerical differences along the filtration. The $\beta_{0,1}$ and $\beta_{0,2}$ curves, along with the distinct $\beta_{1,q}$ curves, still differentiate PMH ($N = 5$) from PMH ($N = 3$).

A multiscale molecular representation can be obtained either by directly using PMH Betti numbers or by extracting useful statistical information from barcodes. Persistence bars represent the persistence of topological invariants in nested simplicial complexes, from which PMH Betti numbers can be directly read. Molecular features can be designed by collecting the Betti numbers at a set of filtration parameters. However, the inconsistent number of atoms across atom groups or molecules makes barcodes not directly suitable for scalable representation learning. Various stable learning strategies for topological data analysis have been proposed, such as persistent landscapes [20] and persistent images [21]. The bin-spaced statistical functions [3], incorporating the maximum, minimum, average, and standard deviation of barcodes, provide a reliable and effective vector representation. This approach offers competitive descriptive capacity and the advantage of scalable modeling. We utilize both the Betti numbers from PMH and barcodes to design molecular features.

To address computational efficiency, simplicial complexes using alpha complexes are primarily considered for PMH with $N > 2$. For PMH with $N = 2$, both VR complexes and alpha complexes can be utilized. When VR complexes are used, we incorporate physical properties in addition to the original molecular structure data to ensure that sufficient molecular interactions are captured. Technically, the filtration process and persistent Mayer homology are induced using either the Euclidean distance metric in space or a kernel function-defined correlation matrix for a group of atomic coordinates. Collectively, these methods enhance our PMH theory-based molecular representation learning. We provide more details about our PMH features in the following section.

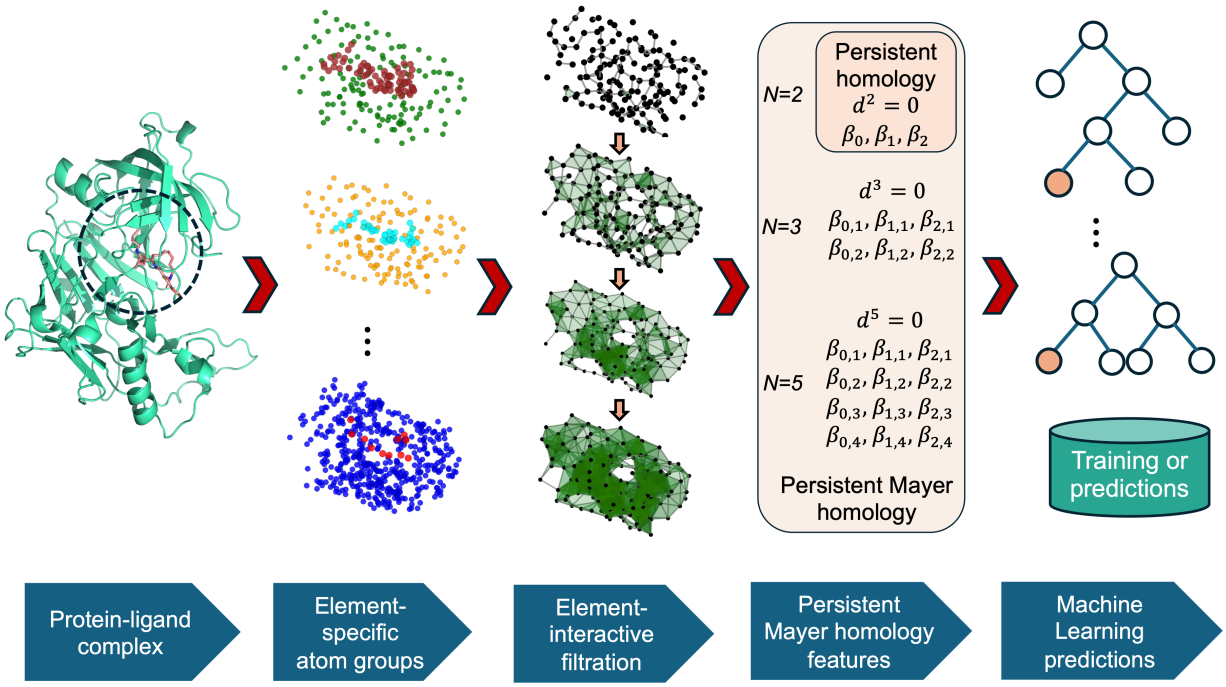


Figure 2.17 The illustration of persistent Mayer homology feature extraction for a protein-ligand complex (PDBID: 1A94) and the subsequent machine learning model development.

2.3.2 PMH learning models for drug design

2.3.3 PMH-based multiscale molecular vectorization

We utilize element-interactive PMH representation learning for biomolecular data, as discussed above. This strategy captures crucial biological information and enhances characterization capacity, as validated by extensive modeling work [3, 22, 23]. Specifically, for a protein-ligand complex, the types of elements considered for proteins are $S_P = \{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}\}$, and for ligands, they are $S_L = \{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}, \mathrm{P}, \mathrm{F}, \mathrm{Cl}, \mathrm{Br}, \mathrm{I}\}$. Therefore, we can have up to 36 element combinations and design interactive PMH features accordingly. The interactions between all the ligand atoms and protein atoms near the binding pocket can also be characterized by PMH.

We denote S^c_{X-Y} as the set of atoms consisting of X types of atoms in the protein and Y types of atoms in the ligand, where the distance between any pair of atoms in these two groups is within a cutoff c:

$$S^c_{X-Y} = \{a | a \in X, \min_{b \in Y} \mathrm{dis}(a, b) \leq c\} \cup \{b | b \in Y\}, \tag{2.3.1}$$

where a and b denote atoms. We also consider all heavy atoms in the ligand together with all heavy atoms in the protein that are within the cutoff distance c from the ligand molecule, and denote this set as S^c_{all}. Similarly, we denote the set of all heavy atoms in the protein that are within the cutoff distance c from the ligand molecule as S^c_{pro}.

Both the correlation matrix and the Euclidean distance matrix are used for the VR complex-induced persistent homology (PMH) ($N = 2$). We use $A(i)$ to indicate the affiliation of an atom with index i in a group of atoms from either the protein or the ligand. We define four types of matrices as follows.

- F$RI^{agst}_{\tau,\nu}$:

$$d(i,j) = \begin{cases} 1 - e^{-(r_{ij}/\eta_{ij})^{\kappa}}, & A(i) \neq A(j) \\ d_{\infty}, & A(i) = A(j) \end{cases} \tag{2.3.2}$$

- F$RI_{\tau,\nu}$:

$$d(i,j) = 1 - e^{-(r_{ij}/\eta_{ij})^{\kappa}} \tag{2.3.3}$$

- EUC^{agst}:

$$d(i,j) = \begin{cases} r_{ij}, & A(i) \neq A(j) \\ d_{\infty}, & A(i) = A(j) \end{cases} \tag{2.3.4}$$

- EUC:

$$d(i,j) = r_{ij}. \tag{2.3.5}$$

Equation 2.3.2 is inspired by the development of the flexibility-rigidity index (FRI) theory [24], which utilizes a decaying radial basis function to effectively quantify atomic interactions. The parameter r_{ij} represents the Euclidean distance between atoms with indices i and j, and $\eta_{ij} = \tau \cdot (r_i + r_j)$, where k and τ are positive adjustable parameters that control the decay rate of the exponential kernel, allowing us to model interactions with different strengths. Here, η_{ij} is the characteristic distance between the ith and jth atoms and is typically set as the sum of the van der Waals radii of the two atoms. The exponential kernel function is non-negative and strictly

monotonically decreasing with respect to the Euclidean distance between a pair of atoms. When the Euclidean distance between two atoms is close to 0, their correlation distance $d(i, j)$ approaches 1. Conversely, when the atoms are far apart, $d(i, j)$ approaches 0. This ensures that the correlation matrix is well-defined. We use the superscript $agst$ to distinguish correlations between atoms from the same or different affiliations. When both atoms are within the same molecule, their correlation distance is set to infinity. This approach excludes intramolecular interactions and highlights the intermolecular interactions between proteins and ligands, which are then represented in the construction of VR simplices and ultimately aid in characterizing these interactions through persistent Mayer homology (PMH).

In contrast, the correlation matrix defined by Equation 2.3.3 captures both physical and chemical information from intramolecular and intermolecular interactions. Furthermore, Equation 2.3.4 and Equation 2.3.5, which are based on the Euclidean distance metric, provide a better characterization of molecular 3D structures. The EUC^{agst} metric places greater emphasis on the shape derived from intermolecular 3D data and is used in conjunction with alpha complexes for our PMH analysis. We primarily use PMH(N=2) and PMH(N=5) to extract molecular features, employing five different feature extraction strategies as shown in Table 2.6. Consequently, for each protein-ligand complex, we generate five feature vectors: the first four are derived from PMH(N=2), while the final vector is based on PMH(N=5).

2.3.4 PMH learning models for binding affinity prediction

We demonstrate the learning capacity of the proposed PMH through protein-ligand binding affinity prediction, a critical problem in drug discovery. We consider three well-established PDBbind datasets [25], including PDBbind-v2007, PDBbind-v2013, and PDBbind-v2016. These datasets contain a collection of 3D structures for protein-ligand complexes and their experimental binding affinities and have been widely used to test new methods [26, 27, 28]. Detailed information about the data size for the three datasets and the related training-test splits can be found in Table 2.7. Based on the 3D structures, each protein-ligand complex is represented by five sets of molecular vectors according to Table 2.6. In our implementation, feature sets I-IV are concatenated into a

I	$\mathrm{PMH2}(P^{12}_{\mathrm{ep-el}}, \mathrm{FRI}^{agst}, \mathrm{VR})$ $\mathrm{ep} \in S_P$, $\mathrm{el} \in S_L$	Length sum of all Betti-0 bars.
II	$\mathrm{PMH2}(P^{6}_{all}, \mathrm{FRI}, \mathrm{VR})$ $\mathrm{PMH2}(P^{6}_{pro}, \mathrm{FRI}, \mathrm{VR})$	Length sum and birth sum of Betti-0, Betti-1, and Betti-2 bars for protein, complex, as well as the sum differences between protein and complex.
III	$\mathrm{PMH2}(P^{12}_{\mathrm{ep-el}}, \mathrm{EUC}^{agst}, \mathrm{VR})$ $\mathrm{ep} \in S_P$, $\mathrm{el} \in S_L$	Counts of Betti-0 bars with 'death' values within: $[0, 2.5]$, $[2.5, 3]$, $[3, 3.5]$, $[3.5, 4.5]$, $[4.5, 6]$, $[6, 12]$.
IV	$\mathrm{PMH2}(P^{9}_{all}, \mathrm{EUC}, \mathrm{Alpha})$ $\mathrm{PMH2}(P^{9}_{pro}, \mathrm{EUC}, \mathrm{Alpha})$	Length sum of Betti-1 and Betti-2 bars with 'birth' values within each interval: $[0, 2]$, $[2, 3]$, $[3, 4]$, $[4, 5]$, $[5, 6]$, $[6, 9]$. The sum differences between complex and protein are also considered.
V	$\mathrm{PMH5}(P^{12}_{\mathrm{ep-el}}, \mathrm{EUC}, \mathrm{Alpha})$ $\mathrm{ep} \in S_P \setminus \{C\}$, $\mathrm{el} \in S_L \cup \{H\}$ $\mathrm{ep} \in \{C\}$, $\mathrm{el} \in S_L \cup \{H\}$	$\beta_{n,q}$ (n=0,1, q=1,2,$\cdots$,4) over filtration parameter range from 0 to 10 with stepsize of 0.2. $\beta_{n,q}$ (n=0,1, q=1,2,$\cdots$,4) over filtration parameter range from 0 to 8 with stepsize of 0.5.

Table 2.6 Molecular feature extraction with PMH. PMH2 and PMH5 indicates the PMH on 2-chain and 5-chain complex, respectively. The first argument in PMH2 or PMH5 specifies the group of molecular coordinate data, while the second argument denotes the correlation or Euclidean distance matrix. The third argument indicates the type of complex used to construct simplical complex.

long vector representation, while feature set V is used as a separate vector representation. These two vectors are combined with the gradient boosting decision tree (GBDT) algorithm to build regression models, resulting in model-PMH2 and model-PMH5. The GBDT hyperparameters used for modeling are listed in Table 2.8. A general workflow of our PMH featurization and the resulting machine learning modeling is provided in Figure 2.17.

The final PMH modeling prediction is determined by the consensus of the predictions from the two models. We build models twenty times with different random seeds and use two evaluation metrics: Pearson correlation coefficient (R) and root mean square error (RMSE). The average R values of the PMH machine learning models for the three datasets are 0.824, 0.787, and 0.834, respectively, as shown in Table 2.9. These high R values validate the effectiveness and reliability of our PMH molecular representation. We also obtain low RMSE values (in units of kcal/mol), which compare the predicted binding energies with the experimental values. The binding energy is calculated from the given pK_d in the original data by multiplying it by a constant of 1.3633.

To enhance the predictive performance of our PMH machine learning models, we incorporate natural language processing (NLP)-based molecular features and develop an additional set of machine learning models. The pretrained NLP models generate molecular features using molecular sequences as input. Specifically, we utilize molecular features from transformer-based pretrained models for proteins [29] and small molecules [30]. These features are then integrated with the GBDT algorithm to create a new predictive model, referred to as model-seq. The modeling performance of this approach is presented in the third column of Table 2.9. The average R value of the PMH model exceeds that of the transformer-based machine learning model. Additionally, we create a consensus model by combining the strengths of the three models—model-PMH2, model-PMH5, and model-seq—by averaging their predictions to determine the final predicted binding affinity. The last column of Table 2.9 shows the performance of the consensus model. The consensus model significantly boosts the performance of the PMH model, with an average R value of 0.832.

A series of advanced mathematical theories from algebraic topology and graph theory were employed to design molecular descriptors [22, 23, 31, 3], leading to reliable machine learning models. Their success significantly relies on molecular characterization through topological invariants. Our machine learning model is comparable to these competitive models and demonstrates superior performance compared to a wide range of other published models. The Betti numbers from PMH include crucial topological invariants and provide additional mathematical analysis of molecular data. This significantly enhances the descriptive and predictive power of our molecular features.

Dataset	**Total**	**Training set**	**Test set**
PDBbind-v2007 [32]	1300	1105	195
PDBbind-v2013 [33]	2959	2764	195
PDBbind-v2016 [34]	4057	3767	290

Table 2.7 Details of the datasets utilized for benchmark tests in this study.

We compare the performance of our consensus model with various models from the literature. Figure 2.18 depicts these comparisons across the three PDBbind datasets. Our model outperforms a wide range of models and represents the state of the art. The second column in Figure 2.18 shows

No. of estimators	Max depth	Min. sample split	Learning rate
20000	7	5	0.002
Max features	Subsample size	Repetition	
Square root	0.8	20 times	

Table 2.8 Hyperparameters used for build gradient boosting regression models.

Dataset	**PMH**	**Transformer**	**PMH+Transformer**
PDBbind-v2007	0.824(1.95)	0.795(2.006)	**0.837(1.907)**
PDBbind-v2013	0.787(2.036)	0.791(1.977)	**0.807(1.982)**
PDBbind-v2016	0.834(1.755)	0.836(1.716)	**0.851(1.701)**
Average	0.815 (1.914)	0.807 (1.9)	**0.832 (1.863)**

Table 2.9 Modeling performance of different strategies on the test sets of PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. Pearson correlation coefficient and root mean square error (unit, kcal/mol) are the two evaluation metrics.

the comparison between experimental energy and predictions from our final consensus model. The high consistency between the two sets of binding energies validates the accuracy and reliability of our machine learning model. Deep neural networks have advanced the development of the scientific community. Integrating our PMH molecular descriptors with deep neural networks has the potential to offer even more accurate predictive models.

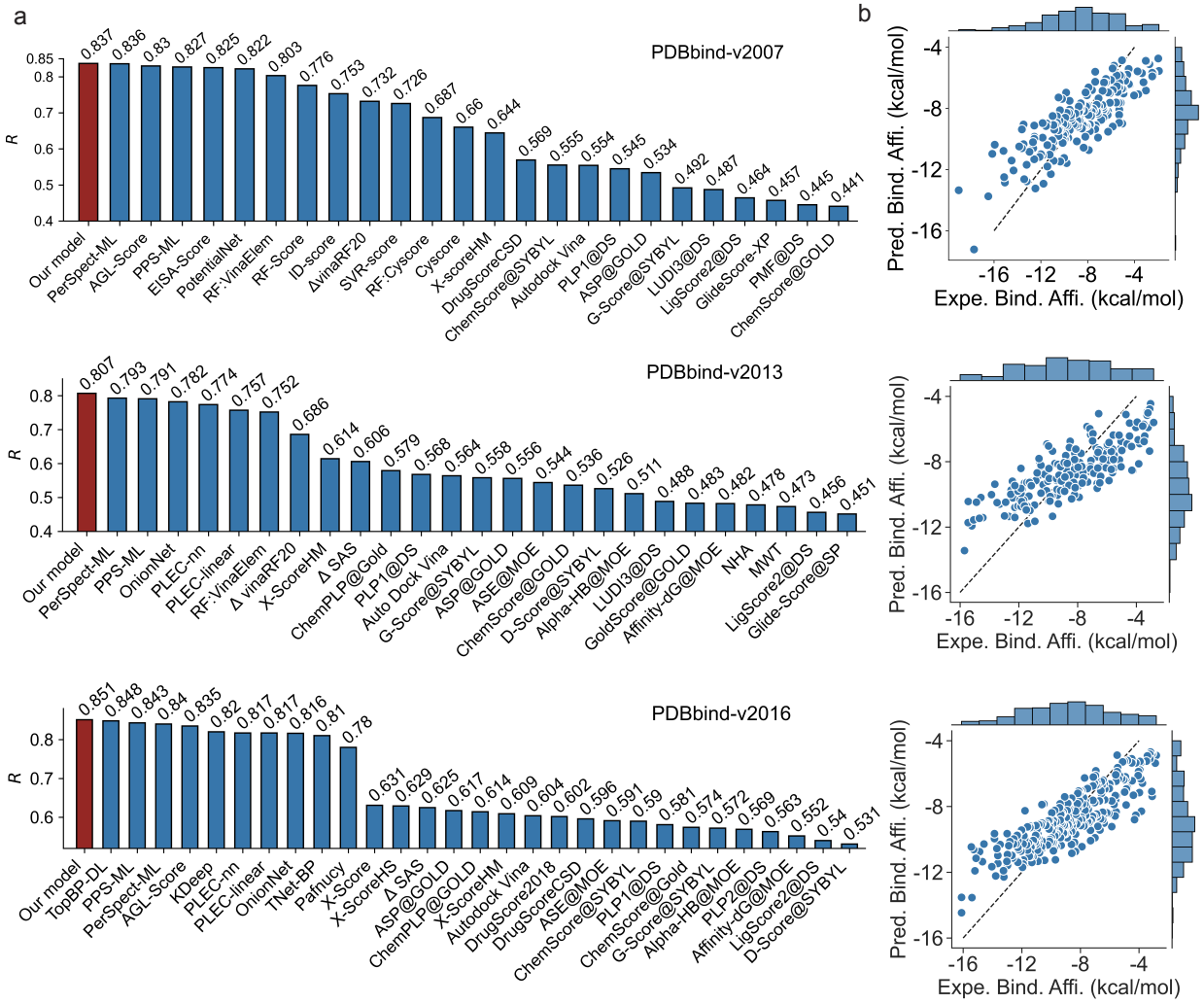


Figure 2.18 The prediction performance of my final machine learning model for three well-established protein-ligand binding affinity datasets including PDBbind-v2007, PDBbind-v2013, and PDBbind-v2016. The comparison of the experimental and predicted binding affinities for the three datasets are exhibited in the right column.

CHAPTER 3

COMPUTATIONAL GEOMETRIC TOPOLOGY IN BIOLOGICAL STUDIES

3.1 Knot theory

To introduce Khovanov homology and establish notations, we review some fundamental concepts of knot theory in this section, including Reidemeister moves, knot invariants, Gauss code, Kauffman brackets, Jones polynomials, and Khovanov homology. We aim to present these topics in a self-contained manner. For readers interested in a more detailed study of knot theory, we recommend the references [35, 36].

3.1.1 Knot invariant

A *knot* is an embedding of the circle S^1 into three-dimensional Euclidean space $\mathbb{R}^3$ or into the 3D sphere S^3. Sometimes, the knot is required to be piecewise smooth and to have a non-vanishing derivative on each closed interval.

Two embeddings $f, g : N \to M$ of manifolds are called *ambient isotopy* if there is a continuous map $F : M \times [0, 1] \to M$ such that if F_0 is the identity map, each $F_t : M \to M$ is a homeomorphism, and $F_1 \circ f = g$.

Two knots are *equivalent* if there is an ambient isotopy between them. It is one of the pivotal challenges in knot theory to study the equivalence classes of knots. This equivalence allows us to systematically study the properties and characteristics of knots without considering their specific shapes or spatial positions. Based on this, researchers have developed various knot invariants and established the topology of knots.

A knot in $\mathbb{R}^3$ (resp. S^3) can be projected into the Euclidean plane $\mathbb{R}^2$ (resp. S^2). From now on, unless specifically stated otherwise, we will focus on knots in $\mathbb{R}^3$. For knots in S^3, we can provide analogous descriptions.

A projection $p : K \to \mathbb{R}^2$ of a knot K is *regular* if it is injective everywhere, except at a finite number of crossing points. These crossing points are the projections of double points of the knot, and should occur only where lines intersect. Moreover, the crossing points contain the information

of overcrossings and undercrossings. Such a projection is commonly referred to as a *knot diagram*.

It is worth noting that a knot can have different regular projections. Consequently, for a given knot, we can obtain different knot diagrams. Indeed, the knot diagram is independent of the choice of projection up to equivalence. Before proceeding, let us recall the Reidemeister moves.

The *Reidemeister moves* are the following three operations on a small region of the diagram :

(R1) Twist and untwist in either direction;

(R2) Move one loop completely over or under another; and

(R3) Move a string completely over or under a crossing.

Figure 3.1(a) provides a graphical representation of the Reidemeister moves.

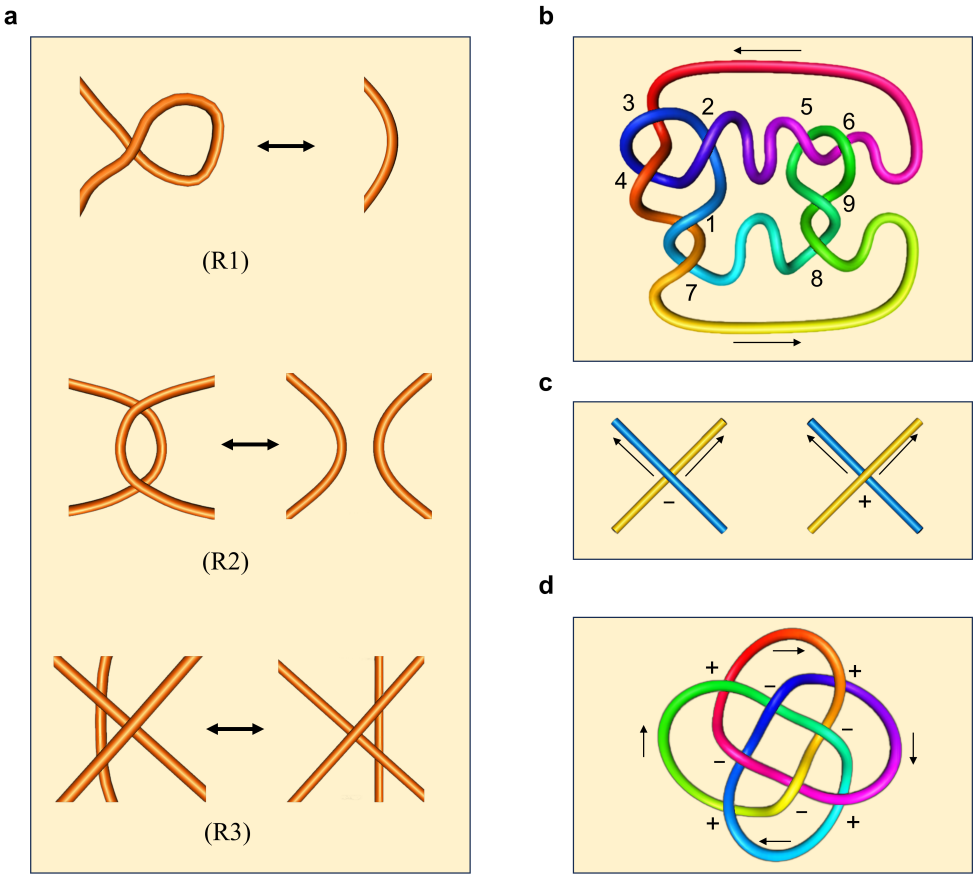


Figure 3.1 **(a)** The three types of Reidemeister moves; **(b)** The marked diagram of a knot can be used to obtain the Gauss code; **(c)** The left is the left-handed crossing, and the right is the right-handed crossing; **(d)** The knot with crossings marked by + or −. The corresponding writhe number is $w(L) = 4 - 4 = 0$.

Reidemeister et al. have shown that two knot diagrams belonging to the same knot can be transformed into each other by a sequence of the three Reidemeister moves up to ambient isotopy [37, 38]. Moreover, two knots are equivalent if and only if all their projections are equivalent [36]. This suggests that the equivalence relation of knots can be established using Reidemeister moves, which are more user-friendly compared to ambient isotopy. They also facilitate proving whether a quantity is a knot invariant.

A *knot invariant* is a quantity defined on knots that remains unchanged under knot equivalence. The most common knot invariants include tricoloring [39], crossing number [35], bridge number [40], and the Jones polynomial [41]. However, these knot invariants cannot determine the equivalent class of knots; indeed, it is even difficult to determine if a knot is the trivial knot. This underscores the inadequacy of current knot invariants, prompting ongoing efforts to seek new ones. Among these knot invariants, the Jones polynomial stands out as one of the most successful. It encapsulates critical information regarding knot topology and structure, including symmetry, crossing distribution, and complexity. Furthermore, its profound links to fields such as topological quantum field theory and quantum braid theory in physics underscore its importance in understanding topological phase transitions and quantum states.

3.1.2 Gauss code

The Gauss code represents a knot diagram using a sequence of integer numbers [42]. This digital representation facilitates recording and understanding of the knot diagram. Moreover, we can reconstruct the original knot diagram from its Gauss code. This implies that Gauss code holds significant importance in classifying knots and computing knot invariants.

Given a knot diagram K, one can obtain a Gauss code $G(K)$ as follows:

1) Choose a crossing as the starting point and select a direction to begin from the starting point;

2) Assign the starting crossing a value of 1, and then assign values of 2, 3, and so on to each subsequent unlabeled crossing along the chosen direction;

3) For each crossing, we assign a sign. If the crossing is an overcrossing, the sign is positive;

otherwise, it is negative.

The integer sequence written down following the aforementioned procedure is what we refer to as the Gauss code. For example, see Figure 3.1(b). Starting from 1 and proceeding to 2, we obtain a sequence of numbers, denoted as $1, 2, 3, 4, 2, 5, 6, 3, 4, 1, 7, 8, 9, 6, 5, 9, 8, 7$. By assigning a sign to each number based on the type of crossing, we get a new sequence of numbers:

$$+1, -2, +3, -4, +2, +5, -6, -3, +4, -1, +7, +8, -9, +6, -5, +9, -8, -7.$$

This sequence is the Gauss code for the knot in Figure 3.1(b).

For a Gauss code C, we can reconstruct a knot diagram $D(C)$. So, the natural question arises: for a knot diagram K, is the knot diagram $D(G(K))$ equivalent to K? In general, this is not entirely correct. To address this issue, people have introduced extended Gauss code. The construction of the extended Gauss code is similar to the Gauss code, with one key difference in how the signs of the integers are assigned. When the crossing is right-handed, the integer is assigned a positive value, and when it is left-handed, the integer is assigned a negative value. For Figure 3.1(b), by considering the right-handed or left-handed nature of each crossing, we obtain the extended Gauss code:

$$+1L, -2R, +3R, -4R, +2R, +5L, -6L, -3R, +4R, -1L,$$
$$+7L, +8R, -9R, +6L, -5L, +9R, -8R, -7L.$$

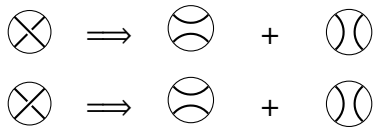
In theory, Gauss code helps us examine and understand information about knots, which allows us to study their properties. In computation, Gauss code can be utilized to calculate various knot invariants, such as the Jones polynomial, Alexander polynomial, and others. Furthermore, from an algorithmic perspective, digitizing and processing knot data through the Gauss code are invaluable for computer-assisted knot research and computation.

3.1.3 Kauffman bracket and Jones polynomial

In the previous section, we concluded that to study the invariants of knots, it is sufficient to explore the invariance of knot diagrams under Reidemeister moves. From now on, our attention will

directed toward knot diagrams as we revisit the Kauffman bracket and Jones polynomial associated with them.

For a crossing, there is a 0-smoothing and a 1-smoothing  . The process of smoothing can be understood as untangling a crossing, as illustrated below.



A *link* is a collection of knots that do not intersect but may be linked (or knotted) together. In particular, a knot is a link with only one component. If not explicitly stated, the links discussed in this paper are assumed to be orientable.

Given a knot K and a crossing x of K, we can create links by replacing the crossing x with the 0-smoothing and the 1-smoothing, respectively. Let **Knot** denote the set of knots, and let **Link** denote the set of links. Given a link L, let $\mathcal{X}(L)$ denote the set of crossings of L. For each $x \in \mathcal{X}(L)$, the smoothing operators at x lead to the 0-smoothing and the 1-smoothing maps $\rho_0, \rho_1 : \textbf{Link} \to \textbf{Link}$ as $L \mapsto \rho_0(L, x)$ and $L \mapsto \rho_1(L, x)$, respectively. In the following construction of the Kauffman bracket, for an unoriented knot, the smoothing is always performed on the undercrossing  .

The *Kauffman bracket* is a bracket function $\langle - \rangle : \textbf{Link} \to \mathbb{Z}[a, a^{-1}]$ satisfying:

(a) $\langle \bigcirc \rangle = 1$;

(b) $\langle \bigcirc \cup L \rangle = (-a^2 - a^{-2})\langle L \rangle$;

(c) $\langle L \rangle = a\langle \rho_0(L, x) \rangle + a^{-1}\langle \rho_1(L, x) \rangle$ for any $x \in \mathcal{X}(L)$.

Here, $\bigcirc$ denotes the trivial knot.

The Kauffman bracket does always exist, and it is uniquely determined in $\mathbb{Z}[a, a^{-1}]$. Now, let $n = |\mathcal{X}(L)|$ be the number of crossings of L. For each crossing, we have the options of performing 0-smoothing and 1-smoothing. Thus, we can obtain a total of 2^n different smoothing links. Each of these smoothing links is referred to as a *state* of the link L. All the states together form a state

cube. Another description of the Kauffman bracket is given in terms of the state cube of a link [39]. For a state s of L, let $\alpha(s)$ and $\beta(s)$ denote the number of 0-smoothings and 1-smoothings of crossings in state s, respectively. The Kauffman bracket is

$$\langle L\rangle = \sum_{s}(-1)^{\alpha(s)-\beta(s)}(-a^2-a^{-2})^{\gamma(s)-1}. \tag{3.1.1}$$

Here, s runs through all the states of L, and $\gamma(s)$ is the number of circles of L in the state s.

It is worth noting that the Kauffman bracket is invariant under the Reidemeister moves (R2) and (R3). However, the Kauffman bracket is not a knot invariant, as it is not invariant under (R1). To define a knot invariant, we first introduce the concept of the writhe number. Consider an oriented diagram of a link L. Let us define $w(L)$ as follows: with each crossing of L, we associate $+1$ if it is a right-handed crossing, and -1 if it is a left-handed crossing. For an example, see Figures 3.1(c) and (d). By summing these numbers at all crossings, we obtain the writhe number $w(L)$.

The *Kauffman polynomial* (or normalized Kauman bracket) of a link L is the polynomial defined as follows

$$X_L(a) = (-a)^{-3w(L)}\langle L\rangle. \tag{3.1.2}$$

The Kauffman polynomial is a knot invariant [43]. By substituting a in $X_L(t)$ with $t^{-\frac{1}{4}}$, we obtain the *Jones polynomial*

$$V_L(t) = X_L(t^{-\frac{1}{4}}). \tag{3.1.3}$$

The Jones polynomial is a famous knot invariant introduced by Jones [41].

Remark 3.1.1. With the previous notations, if we set $q = -a^{-2}$, then the Kauffman bracket can be described by the conditions

(a') $\langle \bigcirc\rangle = q + q^{-1}$;

(b') $\langle \bigcirc \cup L\rangle = (q + q^{-1})\langle L\rangle$;

(c') $\langle L\rangle = \langle \rho_0(L,x)\rangle - q\langle \rho_1(L,x)\rangle$ for any $x \in \mathcal{X}(L)$.

Let n_+ be the number of right-handed crossings in $\mathcal{X}(L)$, and let n_- be the number of left-handed crossings in $\mathcal{X}(L)$. The unnormalized Jones polynomial is defined by

$$\hat{J}(L) = (-1)^{n_-} q^{n_+ - 2n_-} \langle L \rangle. \tag{3.1.4}$$

Then, the Jones polynomial of L is defined as $J(L) = \hat{J}(L)/(q + q^{-1})$. This definition is more convenient for categorifying the Jones polynomial, as specifically detailed in the literature [44].

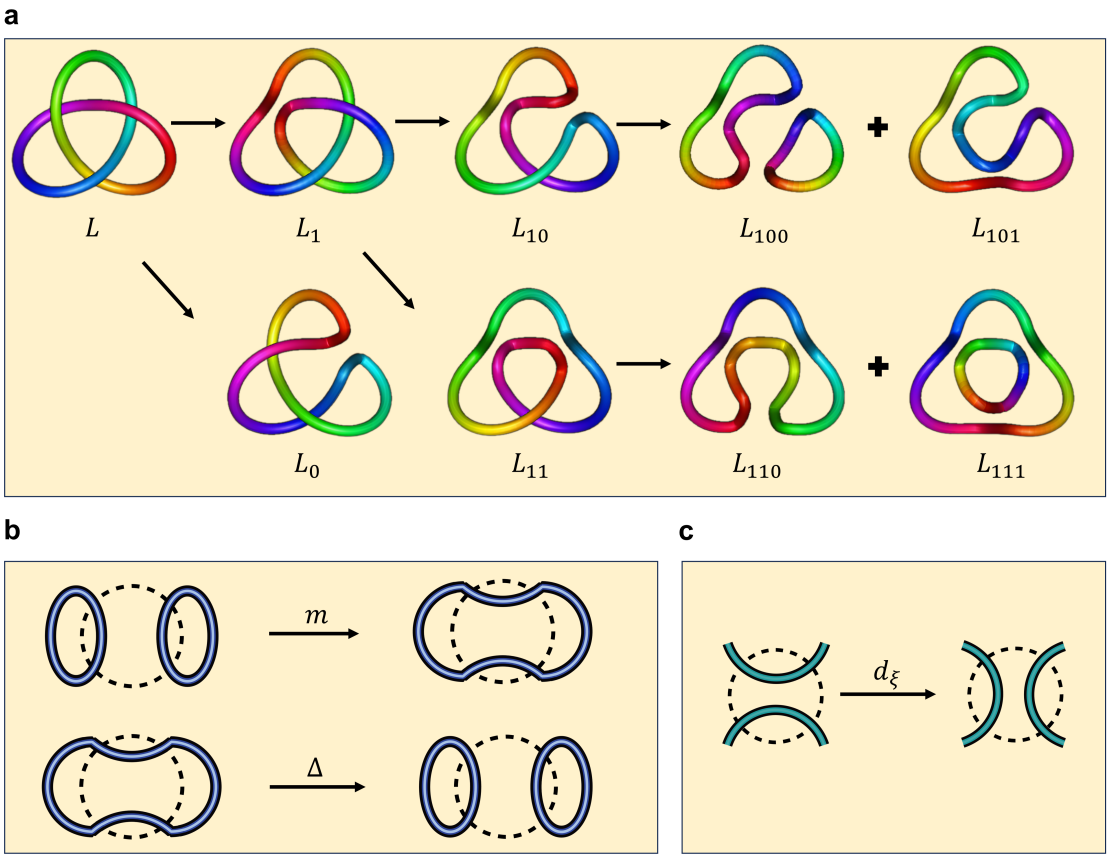


Figure 3.2 (a) The links by conducting 0-smoothings and 1-smoothings of the undercrossings of a left-handed trefoil; (b) Two circles merging into one, or one circle splitting into two; (c) An illustration of the differential.

Example 3.1.2. Let L be a left-handed trefoil. Consider the smoothing of L shown in Figure 3.2(a). For example, the link L_{100} represents the original link after performing one 1-smoothing, followed by two 0-smoothings. Note that

$$\langle L_{100}\rangle = \langle \bigcirc \cup \bigcirc \rangle = (q+q^{-1})^2,$$
$$\langle L_{101}\rangle = \langle \bigcirc \rangle = (q+q^{-1}),$$
$$\langle L_{110}\rangle = \langle \bigcirc \rangle = (q+q^{-1}),$$
$$\langle L_{111}\rangle = \langle \bigcirc \cup \bigcirc \rangle = (q+q^{-1})^2.$$

It follows that

$$\langle L_{10}\rangle = \langle L_{100}\rangle - q\langle L_{101}\rangle = q^{-1}(q+q^{-1}),$$
$$\langle L_{11}\rangle = \langle L_{110}\rangle - q\langle L_{111}\rangle = -q^2(q+q^{-1}).$$

Thus, we have $\langle L_1\rangle = \langle L_{10}\rangle - q\langle L_{11}\rangle = (q^{-1}+q^3)(q+q^{-1})$. By a similar calculation, we can obtain $\langle L_0\rangle = q^{-2}(q+q^{-1})$. Hence, we obtain

$$\langle L\rangle = \langle L_0\rangle - q\langle L_1\rangle = (q^{-2}-1-q^4)(q+q^{-1}).$$

Thus the unnormalized Jones polynomial of L is

$$\hat{J}(L) = (-1)^3 q^{-6}\langle L\rangle = q^{-1}+q^{-3}+q^{-5}-q^{-9},$$

and the Jones polynomial of L is $q^{-2}+q^{-6}-q^{-8}$.

3.1.4 Khovanov homology

Khovanov homology, introduced by Khovanov around year 2000, is regarded as a categorification of the Jones polynomial, providing a topological interpretation of the Jones polynomial [45, 43]. Specifically, the graded Euler characteristic of Khovanov homology corresponds to the Jones polynomial. Compared to the Jones polynomial, Khovanov homology contains more information. Notably, Khovanov homology can detect the unknot [46].

Graded dimension: Let $V = \sum_{k\in\mathbb{Z}} V_k$ be a graded vector space. The graded dimension of V is the power series

$$\mathrm{qdim}\, V = \sum_{k\in\mathbb{Z}} q^k \dim V_k.$$

For example, if V is generated by three elements v_{-1}, v_0, v_1 with the grading -1, 0, 1, respectively, then the graded dimension of V is $q^{-1} + 1 + q$.

Degree shift: The degree shift on a graded vector space $V = \sum_{k\in\mathbb{Z}} V_k$ is an operation $\cdot\{l\}$ such that $W\{l\}_k = W_{k-l}$. By definition, one has that

$$\mathrm{qdim}\, V\{l\} = q^l \mathrm{qdim}\, V.$$

Height shift: Let C denote the cochain complex $\cdots \to C^n \xrightarrow{d^n} C^{n+1} \to \cdots$. The height shift of C^* is the operation $\cdot[m]$ such that $C[m]$ is a cochain complex with $C[m]^n = C^{n-m}$ and $d[m]^n = d^{n-m} : C^{n-m} \to C^{n-m+1}$.

Recall that for a link, we have a state cube $\{0,1\}^{\mathcal{X}(L)}$. Each state s in $\{0,1\}^{\mathcal{X}(L)}$ can be represented as $(s_1, s_2, \ldots, s_n)$, where $n = |\mathcal{X}(L)|$. Now, let $\mathbb{K}$ be the ground field, and let V be a graded vector space with two generators v_-, v_+. Then, $\mathrm{qdim}\, V = q^{-1} + q$. For each state $s \in \{0,1\}^{\mathcal{X}(L)}$, we have a space $V_s(L) = V^{\otimes c(s)}\{\ell(s)\}$, where $c(s)$ is the number of circles in the smoothing of L at state s, and $\ell(s) = \sum_{i=1}^{n} s_i$ is the number of ones in the representation of s. The k-th chain group of L is defined as

$$[[L]]^k := \bigoplus_{s:\ell(s)=k} V_{c(s)}(L). \tag{3.1.5}$$

Then, $[[L]]$ is a graded vector space. Furthermore, we can obtain a cochain complex $[[L]]\{n_+ - 2n_-\}$. The Khovanov chain group of L is defined by

$$C(L) := [[L]][-n_-]\{n_+ - 2n_-\}. \tag{3.1.6}$$

More precisely, we have

$$C^k(L) = \bigoplus_{\ell(s)=k+n_-} V^{\otimes c(s)}\{\ell(s) + n_+ - 2n_-\}. \tag{3.1.7}$$

Note that $C^k(L)$ itself is a graded vector space. Thus there is a natural graded structure on $C^k(L)$. To obtain a cochain complex, we will endow $C(L)$ with a differential as follows. Consider the state cube $\{0,1\}^{\mathcal{X}(L)}$ with $n \cdot 2^{n-1}$ edges. Each of the edges is of the form

$$(s_1, s_2, \ldots, s_{i-1}, 0, s_{i+1}, \ldots, s_n) \to (s_1, s_2, \ldots, s_{i-1}, 1, s_{i+1}, \ldots, s_n).$$

We denote the edge by $\xi = (\xi_1, \xi_2, \ldots, \xi_{i-1}, \star, \xi_{i+1}, \ldots, \xi_n)$. Let $\mathrm{sgn}(\xi) = (-1)^{\xi_1+\cdots+\xi_{i-1}}$, and let $|\xi| = \sum_{t\neq i} \xi_t$. The differential $d^k : C^k(L) \to C^{k+1}(L)$ is defined by $d = \sum_{|\xi|=k} \mathrm{sgn}(\xi) \cdot d_\xi$. Now, we will review the construction of d_ξ. Note that an edge of the state cube connects two adjacent states. The two states differ by just one crossing's smoothing, which implies that the diagrams corresponding to these two states differ by just one circle. Geometrically, this is manifested as two circles merging into one, or one circle splitting into two, see Figures 3.2(b) and (c).

Algebraically, the above process can be understood as $V \otimes V \to V$ or $V \to V \otimes V$, because the word length of the term $V^{\otimes c(s)}\{\ell(s) + n_+ - 2n_-\}$ is equal to the number of circles. The map $d_\xi : C^k(L) \to C^{k+1}(L)$ is defined as:

$$m : V \otimes V \to V, \quad m : \begin{cases} v_+ \otimes v_+ \mapsto v_+, \quad v_- \otimes v_+ \mapsto v_-, \\ v_+ \otimes v_- \mapsto v_-, \quad v_- \otimes v_- \mapsto 0 \end{cases} \tag{3.1.8}$$

on the components involved in merging,

$$\Delta : V \to V \otimes V, \quad \Delta : \begin{cases} v_+ \mapsto v_+ \otimes v_- + v_- \otimes v_+, \\ v_- \mapsto v_- \otimes v_- \end{cases} \tag{3.1.9}$$

on the components involved in splitting, and the identity at other components. It can be verified that the above construction indeed provides a differential structure on $C(L)$. Therefore, $C(L)$ is a cochain complex, called the *Khovanov complex*. The *Khovanov (co)homology* of L is defined by

$$H^k(L) := H^k(C(L)), \quad k \geq 1.$$

As a well-known knot invariant, Khovanov homology can decode the Jones polynomial. We call the rank of $H^k(L)$ the *k-th Betti polynomial* of L, denoted by $\beta_k(q)$.

The *graded Poincaré polynomial* of $C(L)$ is defined by

$$Kh(L) = \sum_k \mathrm{qdim}\, H^k(L) \cdot t^k. \tag{3.1.10}$$

By taking $t = -1$, we have the *graded Euler characteristic* of L given by

$$\chi_q(L) = \sum_k (-1)^k \mathrm{qdim}\, H^k(L). \tag{3.1.11}$$

It is worth noting that $\mathcal{X}_q(L) = \sum_k (-1)^k \mathrm{qdim} C^k(L)$. A famous result asserts that the graded Euler characteristic of L equals the unnormalized Jones polynomial of L.

Theorem 3.1.1. Let L be a link. We have $\mathcal{X}_q(L) = \hat{J}(L)$.

The above result demonstrates that Khovanov homology provides a categorical interpretation of the Jones polynomial, thereby establishing the significant role of Khovanov homology in knot theory. In this work, our focus lies in applying the features of Khovanov homology to analyze and study knots with spatial twists. Persistence is the core principle in analyzing the spatial geometric structure of knots. This prompts us to investigate evolutionary Khovanov homology in subsequent sections.

Example 3.1.3. Let L be the left-handed trefoil. All the crossings are left-handed. Then, we have the Khovanov cochain complex of L given by

$$0 \longrightarrow C^{-3}(L) \xrightarrow{d^{-3}} C^{-2}(L) \xrightarrow{d^{-2}} C^{-1}(L) \xrightarrow{d^{-1}} C^{0}(L) \longrightarrow 0.$$

Here, the space $C^k(L)$ is obtained by the circles of states listed as follows:

$$\begin{aligned}
C^{-3}(L) &= \overbrace{V \otimes V \otimes V}^{(0,0,0)}, \\
C^{-2}(L) &= \overbrace{V \otimes V}^{(1,0,0)} \oplus \overbrace{V \otimes V}^{(0,1,0)} \oplus \overbrace{V \otimes V}^{(0,0,1)}, \\
C^{-1}(L) &= \overbrace{V}^{(1,1,0)} \oplus \overbrace{V}^{(1,0,1)} \oplus \overbrace{V}^{(0,1,1)}, \\
C^{0}(L) &= \overbrace{V \otimes V}^{(1,1,1)}.
\end{aligned}$$

Recall that V has two generators v_+ and v_-. Thus, the space $C^{-3}(L)$ has the basis

$$v_+ \otimes v_+ \otimes v_+, v_+ \otimes v_+ \otimes v_-, v_+ \otimes v_- \otimes v_+, v_- \otimes v_+ \otimes v_+,$$

$$v_+ \otimes v_- \otimes v_-, v_- \otimes v_+ \otimes v_-, v_- \otimes v_- \otimes v_+, v_- \otimes v_- \otimes v_-,$$

the space $C^{-2}(L)$ has the basis

$$(v_+ \otimes v_+, 0, 0), (v_+ \otimes v_-, 0, 0), (v_- \otimes v_+, 0, 0), (v_- \otimes v_-, 0, 0),$$
$$(0, v_+ \otimes v_+, 0), (0, v_+ \otimes v_-, 0), (0, v_- \otimes v_+, 0), (0, v_- \otimes v_-, 0),$$
$$(0, 0, v_+ \otimes v_+), (0, 0, v_+ \otimes v_-), (0, 0, v_- \otimes v_+), (0, 0, v_- \otimes v_-),$$

the space $C^{-1}(L)$ is generated by

$$(v_+, 0, 0), (v_-, 0, 0), (0, v_+, 0), (0, v_-, 0), (0, 0, v_+), (0, 0, v_-),$$

and the space $C^0(L)$ has the basis

$$v_+ \otimes v_+, v_+ \otimes v_-, v_- \otimes v_+, v_- \otimes v_-.$$

We represent the basis of the corresponding space $C^k(L)$ using column vectors. The left representation matrix B_{-1} for the differential d^{-1} is then given as follows:

$$d^{-1}\begin{pmatrix} (v_+, 0, 0) \\ (v_-, 0, 0) \\ (0, v_+, 0) \\ (0, v_-, 0) \\ (0, 0, v_+) \\ (0, 0, v_-) \end{pmatrix} = B_{-1}\begin{pmatrix} v_+ \otimes v_+ \\ v_+ \otimes v_- \\ v_- \otimes v_+ \\ v_- \otimes v_- \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} v_+ \otimes v_+ \\ v_+ \otimes v_- \\ v_- \otimes v_+ \\ v_- \otimes v_- \end{pmatrix}.$$

Similarly, the left representation matrices of the differentials d^{-3} and d^{-2} with respect to the chosen

basis are given by

$$B_{-3} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, B_{-2} = \begin{pmatrix} -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

By step-by-step calculation, we can obtain the corresponding Khovanov homology presented in Table 3.1.

$H^{k,l}(L)$	$k=0$	$k=-1$	$k=-2$	$k=-3$
$l=-1$	$[v_+ \otimes v_+]$	0	0	0
$l=-2$	0	0	0	0
$l=-3$	$[v_+ \otimes v_-]$	0	0	0
$l=-4$	0	0	0	0
$l=-5$	0	0	$[v_+ \otimes v_- - v_- \otimes v_+]$	0
$l=-6$	0	0	0	0
$l=-7$	0	0	$[v_- \otimes v_-]_2$	0
$l=-8$	0	0	0	0
$l=-9$	0	0	0	$[v_- \otimes v_- \otimes v_-]$

Table 3.1 The Khovanov homology $H^{k,l}(L)$ of L.

Here, k is the height and l is the degree of the homology generators. The generator $[v_- \otimes v_-]_2$ exhibits a torsion of 2, meaning that $2[v_- \otimes v_-]_2 = 0$. The remaining generators are free. Thus, we

have

$$
\begin{aligned}
H^{-3}(L) &\cong \mathbb{K}, \\
H^{-2}(L) &\cong \begin{cases} \mathbb{K} \oplus \mathbb{K}, & \mathbb{K} \text{ is the field of characteristic 2;} \\ \mathbb{K}, & \text{otherwise.} \end{cases} \\
H^{-1}(L) &= 0, \\
H^{0}(L) &\cong \mathbb{K} \oplus \mathbb{K}.
\end{aligned}
$$

Consider the case that 2 is invertible in $\mathbb{K}$. The corresponding unnormalized Jones polynomial is given by

$$\hat{J}(L) = \mathcal{X}_q(L) = \sum_k (-1)^k \mathrm{qdim}\, H^k(L) = q^{-1} + q^{-3} + q^{-5} - q^{-9}.$$

This coincides with the result shown in Example 3.1.2.

3.2 Knot data analysis using multiscale Guass linking integral

Knots are ubiquitous in nature, from animal nests, interlocked tree branches, vines, tendrils, chromosome chains, to DNA double helices. Humans have been intrigued by knot tying due to their practical functions, aesthetic appeal, and spiritual symbolism since prehistoric times. Mathematical theory of knots dated back to 1771 by Alexandre-Théophile Vandermonde. Knot theory is one of the most active areas of mathematical studies, concerning the embeddings of a closed circle S^1 into the three-dimensional (3D) Euclidean space, their classification, equivalence after continuous deformations, or ambient isotopy [35]. Some of the most important knot invariants, which differentiate knots, include knot crossing number, knot group [47], knot polynomials [35], knot Floer homology [48], Khovanov homology [44], etc.

Knot theory has been applied to various fields such as physics [49], biochemistry [50], and biology [51, 52, 53], with limited success. Most real-world objects might not be a closed circle. In applications, ambient isotopy typically has major different properties, while keeping the global knot information unchanged. For instance, the realization of many object functions, such as the molecular recognition of DNA, depends on local structures. Therefore, it is imperative to develop knot theory-based tools that are robust and effective for applications.

Several attempts have been made to address the aforementioned challenge. Jamroz et al. proposed the protein topology database KnotProt to study knot and slipknot type of proteins [54]. Dabrowski-Tumanski et al. extend the database to include links and spatial graphs, and also enable the calculation of topological polynomials invariant of those structures [55]. Recently, Panagiotou and Kauffman have proposed new invariants for open curves in 3-space [56]. In addition, Baldwin et al. [57] attempted to localize knot information by intercepting some specific intervals in the linear structure of an open curve. Nevertheless, these approaches are still global topological in nature.

Multiscale analysis can offer a viable localization scheme for knot data analysis, given its remarkable success in diverse areas such as wavelet theory and topological data analysis (TDA). Persistent homology, as a prominent technique in TDA, combines concepts from algebraic topology, geometry, and multiscale analysis to analyze complex datasets [2, 58]. It uncovers the complex topological invariants and patterns of data at various scales, which are not easily discernible with traditional geometric and statistical techniques. Topological features facilitate valuable representation learning, and their efficacy is demonstrated through integration with deep learning models, specifically in the context of topological deep learning (TDL) coined by us 2017 [3]. Compelling applications which consistently demonstrate the relevant advantages of TDL over existing methods are the victories of TDL in the D3R Grand Challenges, a worldwide annual competition series in computer-aided drug, [5], the discovery of SARS-CoV-2 evolution mechanisms [59], and the successful forecasting of SARS-CoV-2 variants BA.2 [60], and BA.5 [18] about two months in advance.

Mathematically, linking number is a knot invariant that measures the extent of linkage between two closed curves in 3D space, representing the number of times that each curve winds around the other. The Gauss linking integral [61], also known as Gauss's integral, gives an explicit formulation for the linking number. It serves as a fundamental tool for studying knots, links, and other topological structures within 3D space. This tool holds significance in various fields, including knot theory, geometric topology, differential geometry, and quantum field theory. For

example, for idealized Dirac-string center vortices, the Chern-Simons number, can be given by the Gauss link integral [62]. High-order link integrals were proposed [63]. However, these approaches are typically global and qualitative.

The objective of this work is to introduce knot data analysis (KDA) as a new paradigm for data science. To this end, we propose a new framework called multiscale Gauss linking integral (mGLI) by integrating multiscale analysis with classical knot and knot-related theories. The proposed mGLI can capture both local and global information of knots, curves, and other curve-like objects by admitting a family of open balls around each segment on the objects. We define a metric to describe the degree of the local entanglement within each ball. By increasing the ball radius, the metric will incorporate additional local information in objects and finally reveal the global properties of the original structure such as knots and entangled links. The proposed mGLI effectively captures intrinsic structures and patterns in complex data, offers valuable low-dimensional embeddings of the data. To assess the performance of mGLI, we consider 13 benchmark datasets across various domains, including protein flexibility analysis, protein-ligand binding affinity prediction, human Ether-à-go-go-Related Gene (hERG) blockade classification, and quantitative toxicity predictions. The performance of mGLI is compared with that of other state-of-art approaches, including TDA, unlocking geometric topology's potential.

In contrast to the previous qualitative and descriptive knot theory approaches, the mGLI is a quantitative and predictive strategy. It offers an unprecedented tool in knot theory analysis and opens a new area in data analysis and knot learning.

3.2.1 Overview of mGLI in knot data analysis(KDA) platform

Figure 3.3 outlines the proposed KDA platform. Like TDA, KDA utilizes a multiscale strategy to capture local structural information of data at various scales and represent the information in a knot invariant, the Gauss link integral or Gauss link number. While globally the Gauss link number quantifies the linking or entanglement between two curves or loops in 3D space, our mGLI further measures local entanglements at each pair of link or curve segments. As shown in Figure 3.3**a**, such local information are systematically collected across scales and assembled over all segments,

giving rise to a vectorization of the original structure.

A specific application of mGLI to a protein-ligand complex is given in Figure 3.3**b**. An element-specific mGLI strategy is introduced to elucidate physical and chemical interactions (Figure 3.3**c**) and to ensure the scalability across different complexes via statistics (Figure 3.3**d**). In the case of protein-ligand complex characterization, chemical and biological information, such as hydrogen bonds, electrostatics, hydrophilicity, and hydrophobicity can be delineated by element-specific mGLI strategy. The intrinsic molecular properties in the 3D structures are properly decoded into low-dimensional topological representations, which are suitable for downstream molecular property analysis and prediction. Theoretical details are provide in Methods section.

The proposed mGLI method captures stereochemical information that is crucial for molecular interactions. In complement, pretrained deep language models are able to access evolutionary and constitutional information of the problem under study. Specifically, we use a transformer-based pretrained model for protein embedding [29], while transformer and autoencoder-based pretrained models are utilized for small molecule embedding[64, 30] as indicated in Figure 3.3**e**. These embeddings are paired with mGLIs for downstream prediction tasks as shown in Figure 3.3**f**.

Multiscale Gauss linking integral (mGLI)

It is intrinsic to describe real-world data by mathematical objects, such as knots, knotoids, lassos, links, linkoids, cysteine knots, etc. (see Figure 3.7**a**). The mGLI involves partitioning knots and other curved objects into segments and conducting a multiscale analysis at each segment. Upon curve segmentation, Gauss link integrals are defined at various scales to quantitatively capture structure, connectivity, and entanglement. The global topological invariant properties are ultimately recovered when a sufficiently large scale is reached. Below, we give some essential formulations of the proposed mGLI method.

Definition 3.2.1 (Gauss linking integral)**.** Given two disjoint open or closed curves l_1 and l_2, parametrized as $\gamma_1(s)$ and $\gamma_2(t)$, respectively, the following double integral gives the the Gauss linking integral that characterizes the degree of interlinking between l_1 and l_2 [65]:

$$L(l_1, l_2) = \frac{1}{4\pi} \int_{[0,1]} \int_{[0,1]} \frac{\det(\dot{\gamma}_1(s), \dot{\gamma}_2(t), \gamma_1(s) - \gamma_2(t))}{|\gamma_1(s) - \gamma_2(t)|^3} \, ds \, dt, \tag{3.2.1}$$

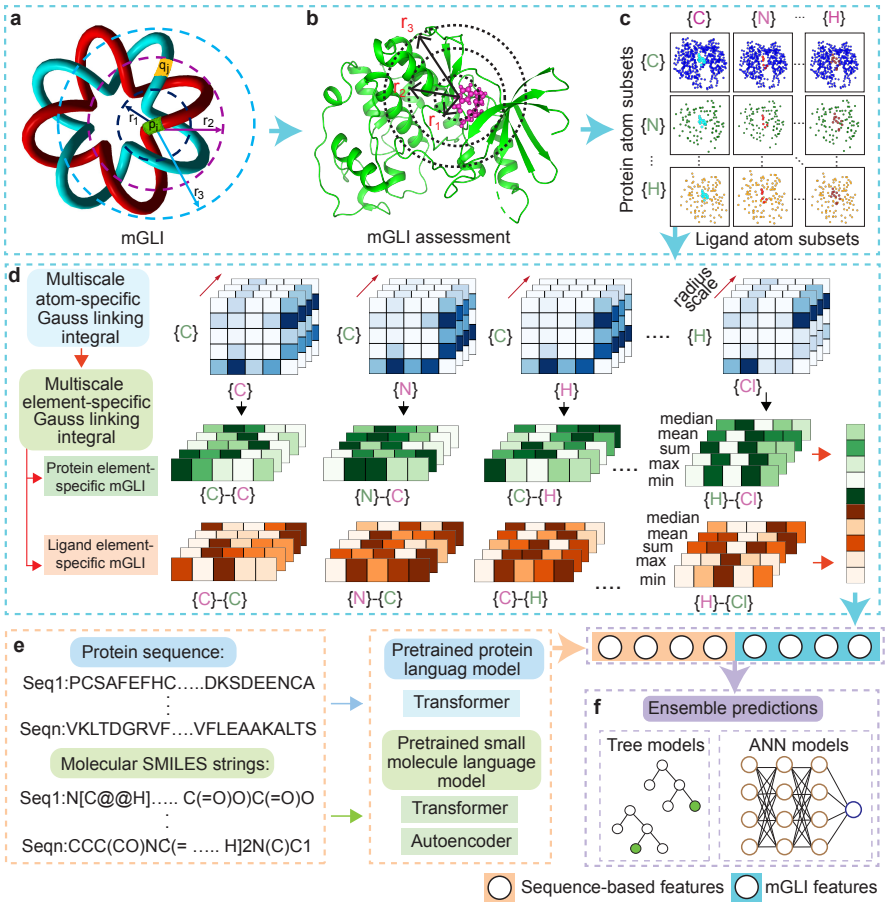


Figure 3.3 The conceptual diagram of the knot data analysis (KDA) platform for biological data learning. **a**. An illustration of multiscale Gauss linking integral-based KDA on a (2, 8) torus. **b** mGLI is applied to the assessment of biomolecular 3D structures with multiple radius scales applied around each atom. **c**. An element-specific mGLI strategy is introduced to embed physical and chemical interactions. **d**. Atom-specific mGLI features are extracted to characterize atomic interactions in the protein-ligand complex. Statistics is used to ensure the scalability across different complexes. **e**. Sequence-based features are generated for the amino acid sequence and the SMILES string, respectively, using pretrained natural language processing models. **f**. The mGLI features and sequence-based features are paired for downstream predictions and analysis using gradient boosting decision tree models or deep neural networks. Colors of frames and large arrows indicate the workflows in different modules: **(a, b, c**, and **d)** denote a structure-based module (blue), **e** highlights a sequence-based module (orange), and **f** represents a prediction module (purple).

where $\dot{\gamma}_1(s)$ and $\dot{\gamma}_2(t)$ are derivative of $\gamma_1(s)$ and $\gamma_2(t)$, respectively.

Definition 3.2.2 (Segmentation of Gauss linking integral)**.** Given finite curve segments P_n and Q_m for disjoint open or closed curves $l_1$and l_2, respectively, the segmentation of Gauss linking integral

induced by the curve segments is defined as the following $n \times m$ segmentation matrix:

$$G = \begin{pmatrix} L(p_1, q_1) & L(p_1, q_2) & \cdots & L(p_1, q_m) \\ L(p_2, q_1) & L(p_2, q_2) & \cdots & L(p_2, q_m) \\ \vdots & \vdots & \ddots & \vdots \\ L(p_n, q_1) & L(p_n, q_2) & \cdots & L(p_n, q_m) \end{pmatrix}, \tag{3.2.2}$$

where $p_i \in P_n$ and $q_j \in Q_m$ are curve segments of l_1 and l_2, respectively. Examples on segmentation of Gauss linking integral for Hopf link are offered in subsection A in the Appendix file.

Remark 3.2.3. The segmentation of the Gauss linking integral serves as the basis for our multiscale modeling. Since the objects in the segmentation of Gauss linking integral are curve segments, we define the distance of curve segments $d(p_i, q_j)$ with Euclidean distance.

Definition 3.2.4 (Scaled Gauss linking integral)**.** Given a finite set of real numbers $R = \{r_0, r_1, r_2, r_3, \cdots, r_k\}$ where $0 = r_0 < r_1 < r_2 < \cdots < r_k$, the Gauss linking integral at scale $[r_t, r_{t+1}]$ is defined as (3.2.3) and (3.2.4).

$$G^{r_t, r_{t+1}} = \begin{pmatrix} \chi_{[r_t, r_{t+1}]}(d(p_1, q_1))L(p_1, q_1) & \cdots & \cdots & \chi_{[r_t, r_{t+1}]}(d(p_1, q_m))L(p_1, q_m) \\ \chi_{[r_t, r_{t+1}]}(d(p_2, q_1))L(p_2, q_1) & \cdots & \cdots & \chi_{[r_t, r_{t+1}]}(d(p_2, q_m))L(p_2, q_m) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_{[r_t, r_{t+1}]}(d(p_n, q_1))L(p_n, q_1) & \cdots & \cdots & \chi_{[r_t, r_{t+1}]}(d(p_n, q_m))L(p_n, q_m) \end{pmatrix}, \tag{3.2.3}$$

where

$$\chi_{[r_t, r_{t+1}]}(x) = \begin{cases} 1, \text{if } x \in [r_t, r_{t+1}] \\ 0, \text{else} \end{cases} \tag{3.2.4}$$

Remark 3.2.5. The scaled Gauss linking integral is used to extract appropriate linking integral within the scale. As shown in the curve segmentation for a (2, 8) torus of Figure 3.3a, each torus has a collection of segments. We have $G_{ij}^{0,r_1} = 0$, $G_{ij}^{r_1,r_2} = L(p_i, q_j)$, and $G_{ij}^{r_2,r_3} = 0$. The scaled integral provides a way to capture local interactions between segments for a given scales. Cumulative integrals across expanding scales offer additional local structural insights, gradually

unveiling broader global characteristics and relationships. Accordingly, multiscale Gauss linking integral features can be designed for various system (see Methods).

Definition 3.2.6 (Localized scaled Gauss linking integral)**.** For given scale $[r_t, r_{t+1}]$, we can define the localized scaled Gauss linking integral at p_i or q_j by the followings:

$$J^{r_t,r_{t+1}}(p_i) = \sum_{s=1}^{m} G_{is}^{r_t,r_{t+1}}, \tag{3.2.5}$$

$$J^{r_t,r_{t+1}}(q_j) = \sum_{s=1}^{n} G_{sj}^{r_t,r_{t+1}} \tag{3.2.6}$$

Remark 3.2.7. By examining Gauss linking integrals at different scales, we obtain multiscale representation. The localized scaled Gauss linking integral gives rise to a measurement for each curve segment in the curve. By considering different scales, the localized scaled Gauss linking integral provides a featurization of each curve segment u:

$$Feature(u) = (J^{r_1,r_2}(u), J^{r_2,r_3}(u), \cdots, J^{r_{k-1},r_k}(u)). \tag{3.2.7}$$

In the case of biomolecular data characterization, curve segmentation is centered at atoms. Consequently, a scaled Gauss linking integral is tailored in an atom-specific or element-specific manner. Localized scaled Gauss linking integrals characterize atomic interactions across various scales, facilitating molecular multiscale analysis.

KDA of biological data

Biological systems are intricately complex and pose grand challenges. We evaluate the performance of mGLI with 13 benchmark datasets in four classes of biological systems, including protein flexibility analysis, protein-ligand binding affinity prediction, the classification of hEGR channel blockers, and quantitative toxicity prediction. To develop predictive machine learning models, we incorporate mGLI features with linear regression algorithm, gradient boosting decision trees (GBDT), deep neural networks (DNN), and multi-task deep neural networks (MTDNN). Extensive comparison with the state-of-the-art is carried to demonstrate utility, reliability, and robustness of the proposed mGLI-based KDA platform.

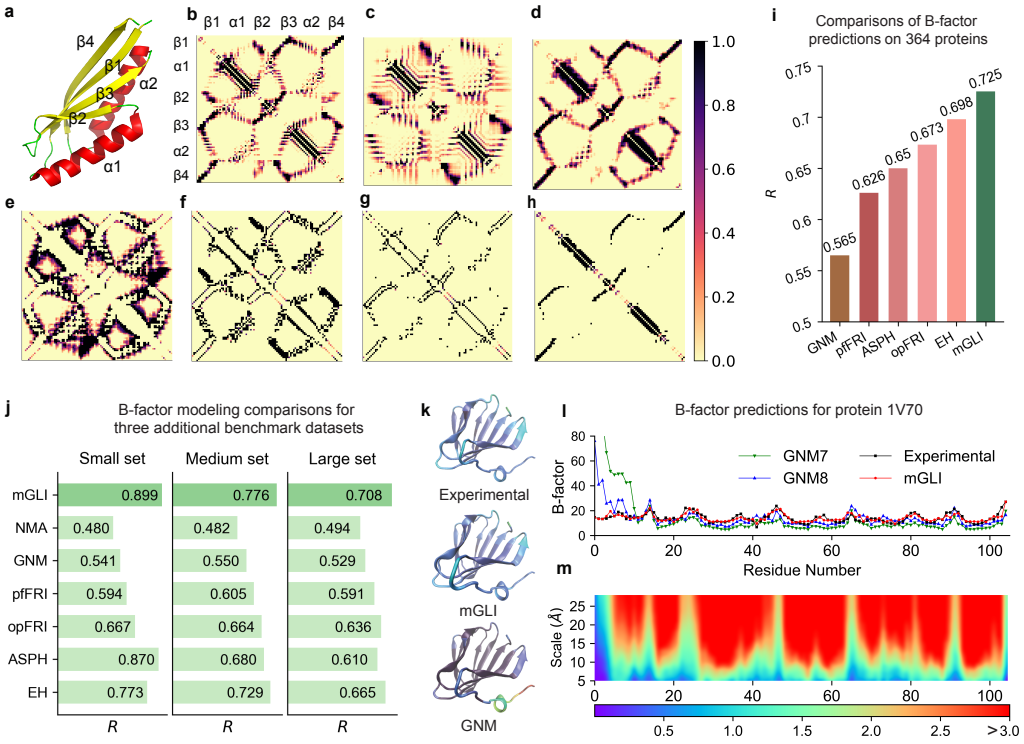


Figure 3.4 An illustration of mBLI analysis for protein B-factor predictions. **a**. The 3D structure of protein 1J27 consisting of two α-helices and four β-sheets. **b**. The segmentation of the Gauss linking integral of protein 1J27. **c**. The absolute value of Gauss linking integral matrix of protein 1J27. **d**. The absolute Gauss linking integral matrix of protein 1J27. **e**-**h**. Absolute Gauss linking integral matrices of protein 1J27 at different scales. **i**. The comparison of B-factor predictions between our mGLI method and other literature approaches on a benchmark dataset of 364 proteins. **j**. The comparison of B-factor predictions on three additional benchmark datasets between our mGLI method and other literature approaches (refer to Table S2 for detailed information). **k**. The visualization of protein 1J27 B-factors obtained from experiments, mGLI, and GNM [66]. **l**. Comparison of protein 1J27 B-factors obtained from experiments, mGLI, and GNM [66]. Here GNM7 and GNM8 indicate the cutoff value at 7 Å and 8 Å for the GNM. The x-axis represents the residue number, and the y-axis represents the B-factor value. **m**. The visualization of mGLI features with the maximal cutoff at 30Å. The x-axis represents the residue number and the y-axis represents the scale range. Note that all values exceed 3.0 are labeled as red.

Protein flexibility analysis

Proteins are inherently flexible and undergo various motions to maintain their functions. Protein flexibility is often experimentally measured with B-factors, also known as temperature factors or atomic displacement parameters. High B-factors indicate increased atomic mobility, suggesting the location of the protein that is flexible or involves conformational changes. Low B-factors, on the other hand, indicate rigid regions with limited atomic motion. We assess the effectiveness of the

proposed mGLI-base features in predicting protein B-factors (see Methods). The mGLI features are integrated with linear regression algorithm. It has been a tradition in B-factor predictions for all methods to utilize the same simple machine learning algorithm, thereby ensuring a fair comparison of various approaches.

Typically, the B-factor prediction focuses on C_α atoms in a protein as shown in Figure 3.4**a** for protein (PDBID: 1J27). We segment the protein polymer chain structure into C_α atoms to facilitate Gauss linking integral calculations of atomic interactions among C_α atoms. The resulting atom-wise mGLI matrix is depicted in Figure 3.4**b** with reference to the secondary structure. It is noteworthy that the Gauss linking integral depends on the orientations of segments or curves. Eliminating this orientation factor may lead to a more insightful analysis for specific tasks, regardless of curve orientation. To completely disregard orientation impact, we consider the absolute Gauss linking integral as

$$\bar{L}(l_1, l_2) = \frac{1}{4\pi} \int_{[0,1]} \int_{[0,1]} \left| \frac{\det(\dot{\gamma_1}(s), \dot{\gamma_2}(t), \gamma_1(s) - \gamma_2(t))}{|\gamma_1(s) - \gamma_2(t)|^3} \right| \, ds \, dt, \tag{3.2.8}$$

along with its corresponding integral segmentation matrix. The absolute Gauss linking integral of Figure 3.4**b** is given in Figure 3.4**c**. In the rest of this work, we use absolute Gauss linking integral in our computations.

Figures 3.4**c-h** show the absolute mGLIs at various scales from large to small. At the smallest scale (Figure 3.4**h**), only the nearest neighbor interactions are recorded in Gauss linking integral. This multiscale analysis characterize each C_α atom's local environment and interactions.

Numerous computational methods have been developed for B-factor predictions, such as Gauss network model (GNM) [66], anisotropic network model (ANM) [67], normal mode analysis (NMA) [68]. However, Park et al. [69] demonstrated that both GNM and NMA were ineffective in analyzing a wide range of protein structures. Their findings revealed that, on average, the correlation coefficients for GNM and NMA, across three protein sets categorized by size (small, medium, and large), were consistently below 0.6 and 0.5, respectively. Recently, advanced methods have emerged to address this challenge, including flexibility rigidity index-based approaches such as pfFRI [70] and opFRI [70], as well as topology-based methods like atom-specific persistent homology (ASPH)

[71] and evolutionary homology (EH) [72].

To evaluate the performance of the proposed multiscale Gauss linking integral (mGLI) for protein flexibility analysis, we employed a dataset consisting of 364 protein structures, sourced from [70]. This dataset served as a benchmark for comparing mGLI against established methods, specifically opFRI [70], pfFRI [70], and GNM [69].

In Table S11, we present the comparative results of mGLI with previous methods for each protein in the dataset. Remarkably, mGLI outperformed previous methods in 320 out of 364 proteins. On average, mGLI achieved the highest correlation coefficient of 0.725, surpassing the values of 0.673 for opFRI, 0.626 for pfFRI, and 0.565 for GNM, as illustrated in Figure 3.4**i**. This represents a significant improvement of 7.7%, 15.8%, and 28.3%, respectively.

In addition, to validate the effectiveness of mGLI for predicting C_α atom B-factors in proteins of different sizes, we compared our method with previous approaches including EH [72], ASPH [71], opFRI [70], pfFRI [70], GNM [69], and NMA [69] on three protein sets, as shown in Figure 3.4**j**. mGLI achieved average correlation coefficients of 0.899, 0.776, and 0.708 for the small, medium, and large protein sets, respectively. Our results on the three datasets significantly outperformed the previous methods, demonstrating improvements of 16.3%, 6.4%, and 6.5% on the small, medium, and large protein sets, respectively, compared to the previous state-of-the-art method EH [72].

To understand mGLI's performance, we present a case study with a potential antibiotic synthesis protein (PDBID: 1V70) 105 residues. Figure 3.4**k** shows the protein colored with B-factor values. Apparently, mGLI-predicted B-factor values are very close to those of the experimental ones, whereas, GNM predicted values are unmatched. Figure 3.4**l** presents detailed comparison. GNM methods have large errors around residues 1-10, which can also be seen in Figure 3.4**k**. In contrast, mGLI gives accurate B-factor prediction for these residues. The mGLI features are presented in Figure 3.4**m**. For each scale, we calculate the cumulative absolute Gauss linking integral, represented by a colored bar along with its accumulated value below. We designate the values exceeding a specific threshold (3.0 in this case) as red. Consequently, it becomes evident that the pattern of mGLI values in Figure 3.4**m** matches the experimental B-factors in Figure 3.4**l** directly.

This observation holds true in a broader sense and is further validated in Figure S6 and Figure S7.

Protein-ligand binding affinity predictions

Protein-ligand binding affinity describes the interaction strength between a potential drug molecule and its target protein or receptor, and its prediction plays a crucial role in drug design and discovery [73, 74]. The development of machine learning models for protein-ligand binding affinity prediction represents a pivotal advancement in computational biology [75]. We explore the utility of mGLI for machine learning predictive models. The PDBbind database [76] offers a comprehensive repository of protein-ligand complex structures along with their corresponding binding affinity data [74]. In our study, we have included two of the most commonly utilized protein-ligand databases, namely, PDBbind-v2013 and PDBbind-v2016 [25]. It is challenging to improve performance on these datasets as they have been studied by numerous researchers. The detailed information for the two datasets and related rigorous training-test splittings can be found in Table S1.

In Methods, we propose two mGLI featurization approaches on two distinct scale intervals $[r_t, r_{t+1}]$ or $[0, r_{t+1}]$, on which localized scaled Gauss linking integral is given. We use notations mGLI-bin and mGLI-all to indicate the protein-ligand complex features and mGLI-lig-bin and mGLI-lig-all to indicate two sets of ligand features. The mGLI-lig-all features can be used as additional features for protein-ligand interactions. We also utilize pretrained natural language processing (NLP) models, i.e., transformer features (TF), to complement mGLI features (see details in the Methods). Gradient boosting decision algorithm is used for the predictions. Given a training dataset, models are built 20 times with different random seeds to address initialization-related errors. The median of Pearson correlation coefficient (R) values from the 20 experiments are reported below.

Figure 3.5**a** illustrates the comparison of Pearson correlation coefficients (R) obtained from our model and the literature ones. Our mGLI-assisted model outperforms existing models for the two PDBbind datasets. The R values of 0.819 and 0.862, are achieved by our models in modeling PDBbind-2013 and PDBbind-2016, respectively, and are the highest values ever reported

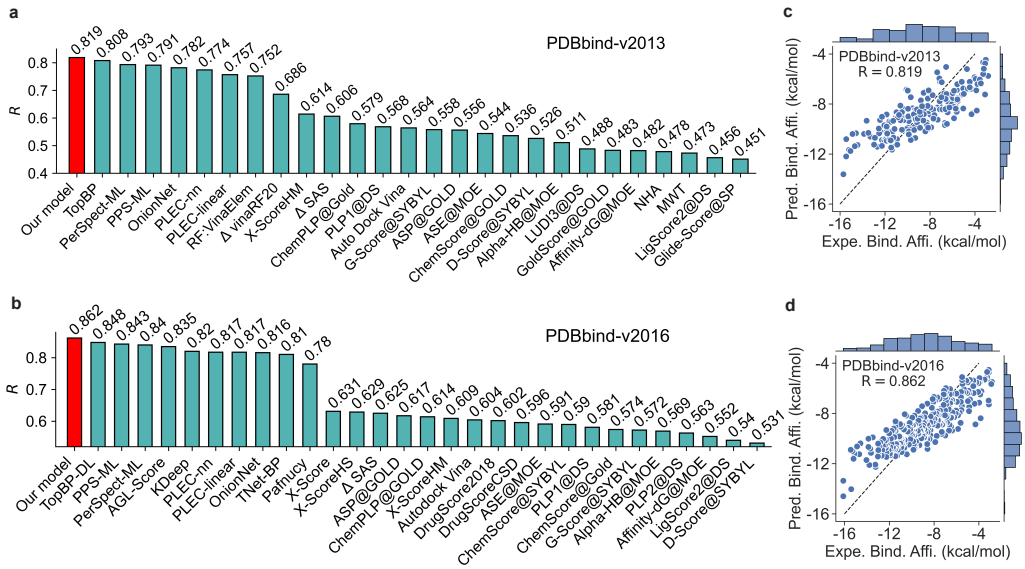


Figure 3.5 The performance summary of our mGLI-assisted machine learning predictions for two PDBbind datasets. **a-b**: The Pearson correlation coefficient (R) comparison for the binding affinity predictions of PDBbind-v2013 and PDBbind-v2016 core sets. Our models outperform other state-of-art methods (refer to Table S4 for detailed information). **c-d**: The comparison between the experimental binding affinity (BA) and the predicted BA from our best models across the two PDBbind datasets.

in the literature. This highlights our model's superiority and establishes it as a new state-of-the-art protein-ligand binding affinity prediction model. Notably, our model demonstrates a significant improvement in R values in modeling the PDBbind-v2013 and PDBbind-v2016 datasets compared to others. The PDBbind-v2013 and PDBbind-v2016 datasets contain 2764 and 3767 complexes, respectively.

Persistent homology [77] and persistent spectral theories [4, 23, 31] give rise to competitive molecular representation and are widely utilized for molecular properties predictions. For example, TopBP [77], PerSpect-ML [23], and PPS-ML [31] rank among the top-performing models in binding affinity prediction, as demonstrated in Figure 3.5**a**. The efficacy of these models can be further augmented when additional physical information is integrated. For instance, the average R value of PerSpect-ML [23] across the two datasets increased from 0.806 to 0.817, while that of PPS-ML [31] increased from 0.804 to 0.817. Our mGLI-assisted models, which are based on mGLI-all&mGLI-lig-all or mGLI-bin&mGLI-lig-all features, provide accurate predictions across the two PDBbind datasets, as shown in Table S3. The symbol '&' denotes feature concatenation.

The average R values of the two mGLI-based models across the two PDBbind dataset are 0.814 and 0.818. The best consensus models, formed by averaging predictions from mGLI-all&mGLI-lig-all or mGLI-bin&mGLI-lig-all feature-based models along with the transformer feature-based models further enhance the modeling performance, achieving an average R value of 0.838 and 0.841 across the two PDBbind datasets. This exceeds the average R of 0.835 obtained from persistent homology [77], as well as the averages of 0.817 from PerSpect-ML [23] and 0.817 from PPS-ML [31].

Figure 3.5**b** offer visualization comparison between the experimental and predicted binding affinities generated by our best models for the two PDBbind datasets. The details of our models are provided in Table S3.

hERG blockade classification predictions

Ligand-based virtual screening plays a significant role in drug discovery. Appropriate molecular descriptors are of vital importance for predictive accuracy. We investigate the performance of our mGLI molecular features in several ligand-based virtual screening prediction tasks. Predictions for hERG blockage are critically important in drug discovery due to the potential cardiac safety risks associated with drugs that inhibit the hERG potassium channel [78].

Several machine learning predictive models are available in the literature [79, 80, 78, 81, 82], and we benchmark our mGLI-based models against them. Among these models, the persistent Laplacian theory [4, 78] was used in conjunction with several NLP molecular embeddings [83, 30] to build predictive models, yielding the best hERG blockade prediction model. The persistent Laplacian approach, rooted in spectral graph theory, can be regarded as an extension of persistent homology theory. It preserves the topological persistence as persistent homology, while revealing additional geometric insights from those non-harmonic portions of the spectrum. We provided the detailed discussion of these two theories in section 7 in the appendix file. Here, we employ mGLI theory alongside several other molecular descriptors, including the same two NLP embeddings as in [78], and algebraic graph (AG)-based molecular features [22]. The NLP embeddings are paired with artificial neural network algorithms, while mGLI and AG features are used with gradient boosting decision tree (GBDT) algorithms. Our final prediction model is obtained with the consensus

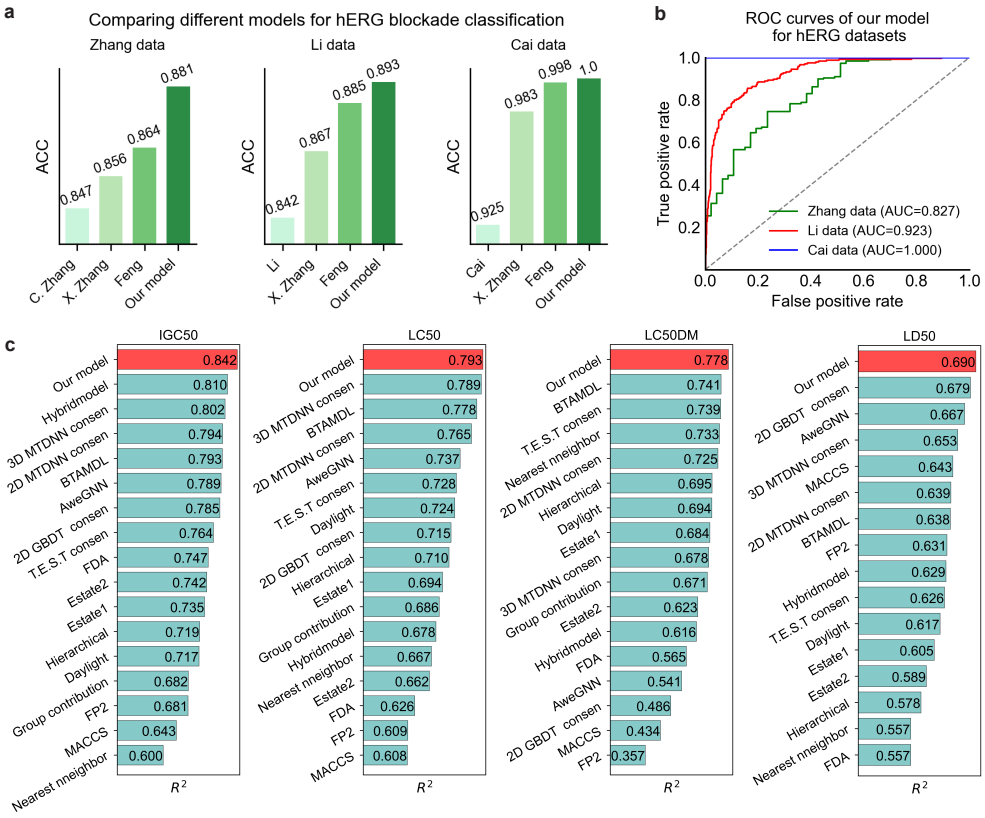


Figure 3.6 The performance summary of our machine learning models for hERG blockade classification and drug toxicity predictions. **a**. Accuracy (ACC) comparisons of our mGLI-assisted consensus model with literature models. These comparisons indicate that our model represents the state-of-the-art machine learning predictive tool. **b**. ROC curves of our model for four hERG blockade classification tasks. **c.** Prediction comparisons of our model with literature models for the four toxicity datasets in terms of the squared Pearson correlation coefficient (R^2) (Refer to Table S7 for detailed comparative information.)

prediction of these four models.

Three hERG blockade datasets with binary classification labels from the literature were used to investigate the performance of our models. Details of these datasets and five utilized evaluation metrics including AUC, ACC, MCC, sensitivity, and specificity are included in Table S1 and section 1 in the Appendix file. Among these metrics, ACC gives the percentage of the correctly predicted blockers and non-blockers. Given a training dataset, each individual model was built ten times with different random seeds. In the comparison with other literature models, the highest ACC scores, along with corresponding metrics evaluations from the ten prediction results, are reported in Table S5. Our models yield state-of-the-art predictions. Figure 3.6a displays the ACC score comparisons

across the three datasets, while the comparison in terms of AUC and MCC is displayed in Figure S12. Figure 3.6**b** exhibits the ROC curves of our model in predicting the test sets of the three datasets.

C. Zhang et al. [79] investigated their model performance with a hERG dataset containing 1163 compounds. Different training and test sets were partitioned from the 1163 compounds. Various thresholds defined by IC_{50} values were used to discriminate hERG blockers from non-blockers. Their SVM model had the best ACC scores of 0.848 on the test set with threshold of 30 μM. X. Zhang et al.'s model [80] had an boosted prediction ACC score of 0.856. Feng et al.'s model [78] achieved much higher improvement in many metric. Our model has significantly higher predictive power than Feng et al.'s model [78] with ACC scores increased from 0.864 to 0.881, and MCC results boosted from 0.518 to 0.587, respectively, while it also achieved high sensitivity and specificity scores.

Li et al. [81] constructed two consensus models based on their dataset composed of 3721 compounds with a threshold of IC_{50} equals to 1 μM classifying blockers and non-blockers. Their best consensus results on a test set of 1092 compounds achieved an ACC score of 0.842. Feng et al.'s model [78] improved the results of Li et al. [81] and X. Zhang et al.[80]. The AUC, ACC, and MCC scores of our mGLI-assisted model are 0.924, 0.893 and 0.661, which are even higher than the corresponding scores of 0.917, 0.885, and 0.629 in Feng et al.'s model [78].

Cai et al. [82] developed a multitask deep neural network-based model and had their best predictive power on a hERG dataset with blockade threshold value of 80 μM. The reported AUC and ACC scores achieved 0.967 and 0.925. Feng et al.'s [78] model had boosted performance. Our model accomplished perfect scores of 1.000 in all the five evaluation metrics. The detailed performance of our individual models is provided in Table S6 or Figure S13. The mGLI models outperform or achieve comparable results. This indicates the critical impact of mGLI modeling on the resulting consensus predictions. Our model consistently exhibits outstanding predictive performance, placing it among the top-tier machine learning models for hERG blocker/non-blocker classification.

Quantitative toxicity predictions

Toxicity in drug discovery refers to the potential harmful effects or adverse reactions that a drug or chemical compound may have on living organisms [84]. Assessing drug toxicity is essential in drug discovery. We assess the performance of our mGLI-assisted predictive models on four toxicity datasets, including IGC50, LC50, LC50DM, and LD50. Information about the toxicity datasets is provided in Table S1 and subsection B in the Appendix file.

In addition to mGLI, we also employ transformer (TF) [83] and autoencoder (AE) models [30] to enhance the modeling performance. We pair GBDT with mGLI features to model the four datasets. Due to the similarity of the toxicity datasets, a multitask deep neural network (MTDNN) was employed to enhance modeling performance [85, 84, 64]. We employed TF and AE features to build two MTDNN models, resulting in two additional sets of predictions. Our final predictive model is obtained by averaging these three sets of predictions. Given a training dataset, models are built 10 times with random seeds.

Table S7 presents the detailed comparison in terms of squared Pearson correlation coefficients (R^2) and root mean squared error (RMSE). The comparisons in terms of R^2 are depicted in Figure 3.6**b**. Our model stands out in toxicity predictions, achieving the higher R^2 values of 0.842, 0.793, 0.778, and 0.690 for the IGC50, LC50, LC50DM, and LD50 datasets, respectively. Figure S16 presents a comparison between the experimental toxicity and our predicted toxicity values for the four datasets. The high consistency underscores the effectiveness of our machine learning models.

Two competitive models were proposed by Gao et al. [85], namely the 2D-GBDT and 2D-MTDNN consensus models, which utilize traditional 2D molecular fingerprints along with various machine learning algorithms. Their multitask learning consensus model achieved R^2 values of 0.794, 0.765, 0.725, and 0.639 for the IGC50, LC50, LC50DM, and LD50 datasets, respectively. They surpassed many other models in the literature, including those from Toxicity Estimation Software Tool (T.E.S.T) and related approaches, such as hierarchical, FDA, nearest neighbor, and T.E.S.T consensus [86]. Wu et al. [84] introduced molecular fingerprints using persistent

homology theory and developed a consensus multitask learning model. Additional molecular descriptors based on physical attributes, including energy, surface energy, and electric charge, were incorporated into their consensus model, significantly enhancing predictive performance. Their model achieved R^2 values of 0.802, 0.789, 0.678, and 0.653 for the aforementioned datasets. Our model outperforms these exceptional models. Several other models have recently been developed based on traditional molecular fingerprints such as estate1, estate2, daylight MACCS, or other advanced strategies. However, our model outperforms them by a significant margin, as observed in Figure 3.6, and detailed comparisons are provided in Table S7. This demonstrates that our mGLI-based knot theory provides an effective approach for molecular representation learning.

In addition, Table S7 or Figure S15 displays the detailed performance results of our GBDT and MTDNN models. We compared the mGLI-based GBDT model with GBDT models based on TF or AE features. The mGLI-GBDT model is competitive across the four prediction tasks, outperforming the TF-GBDT model in all tasks except for LC50DM. The inferior performance for the LC50DM task can be primarily attributed to overfitting issues. The large number of features in the mGLI model makes it less suitable for the LC50DM dataset, whose training set only has 283 molecules. The comparisons indicate that mGLI provides valuable 3D structure-based features for small molecule representations compared to NLP molecular features and is competitive in modeling individual tasks.

Discussion

Generalization to other topological objects and real-world structures

It is intriguing to consider the range of data to which the present KDA can be applied. Mathematically, the multiscale Gauss link integral theory proposed in this work can naturally extend to a wide variety of other topological objects, such as knotoids [87], links, linkoids [88], lassos [89], and cysteine knots [90] in Figure 3.7**a**, as well as curve segments in Figure 3.7**b-c**, tangles, and braids. These types of curved structures are ubiquitous in real-world objects, ranging from ropes, shoelaces, highways, and powerline networks to polymers, DNA, RNA, nucleosomes, chromosomes, and the trajectories of space vehicles and interceptor missiles. In a comparative

analysis, our KDA deals with curved data, whereas TDA handles point cloud data defined on simplicial complexes, graphs, hypergraphs, etc. Additionally, our earlier persistent Hodge Laplacian is defined on manifolds and addresses volumetric data [9].

Curve segment size and multiscale granularity

In principle, our method allows for the arbitrary combination of curve segmentation with any multiscale schemes. However, in practical applications, the performance of mGLI is highly dependent on their selection. First and foremost, the values of the Gauss linking integral of a local curve segment depend not only on their spatial alignment but also on their relative lengths compared to the global curve. When the length of a curve segment approaches zero, the corresponding Gauss linking integral approximates to 0. Similarly, as curve segments expand to cover the global curve, the Gauss linking integral returns global information. In both cases, the Gauss linking integral fails to extract useful spatial information regarding local alignments. The choice of segmentation depends on the specific application. For example, in dealing with molecular properties, atomic segments are needed. In modeling a crowded highway, the segment of individual car size is a natural choice. Secondly, the selection of the multiscale range impacts the featurization of the Gauss linking integral. Ideally, different scales should capture distinct spatial structure information, and the choice of scales should reflect important interactions in the data. If the information between different scales is negligible, it can result in a large number of identical or trivial features. Conversely, if the scale is too coarse, it may lead to information loss.

The superiority of mGLI for biomolecular data

Proteins, DNA, and RNA are polymers and are naturally modeled as curved structures at certain scales. The proposed multiscale Gauss linking integral proves to be a superior tool for biomolecular data analysis compared to previous methods. The analysis of biomolecular structures using mGLI can lead to insights. To demonstrate this, we conducted a structural analysis of protein 1J27 by segmenting the absolute multiscale Gauss linking integral and compared it with the previous transient probability matrix (TPM) [91]. The structural information that was previously obscured becomes considerably more evident and clear when using mGLI, as depicted in Figure 3.7**e**. For

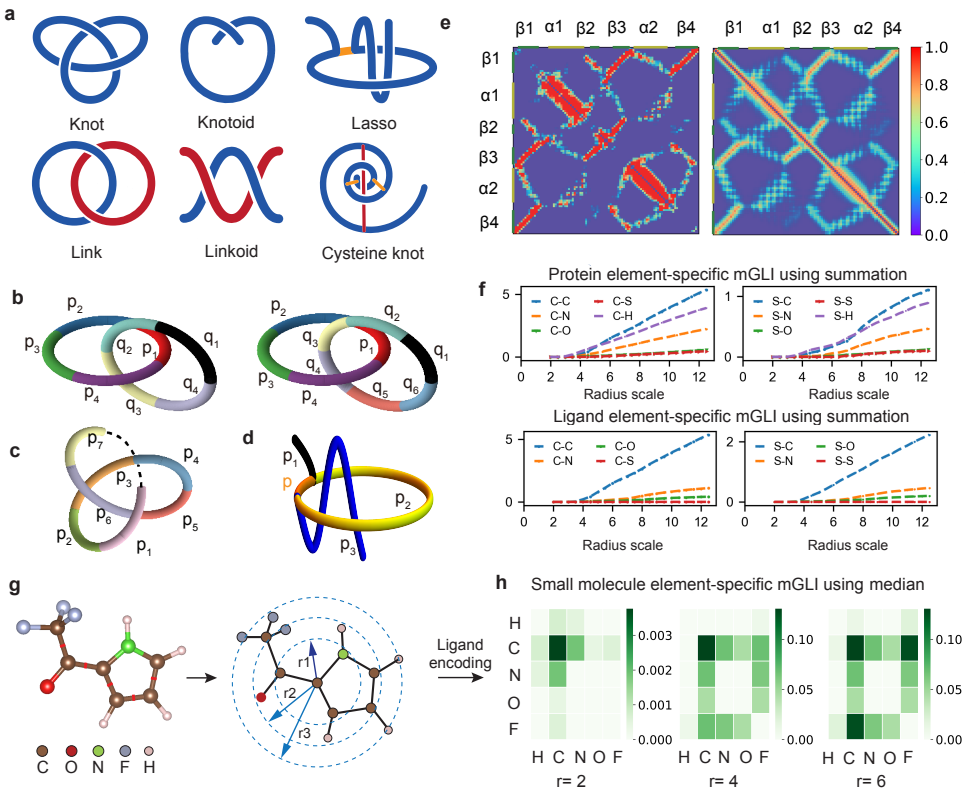


Figure 3.7 **a**. Examples of topological objects which can be studied by the multiscale Gauss linking integral. **b**. Hopf link with two types of curve segmentations. **c**. Slipknot with seven curve segments. **d**. Lasso with four curve segments. **e**. Left is the absolute Gauss linking integral matrix for protein 1J27. Right is the transient probability matrix (TPM) for protein 1J27. Points in top row and left column are colored green or yellow, denoting β-sheet or α-helix of 1J27. **f**. The protein or ligand element-specific mGLI features based on summation statistics for protein 1PXO, as formulated in Equation 3.2.16. Additional features for other element-specific cases are offered in Figure S2, while features based on median statistics are provided in Figure S3. **g**. The curve segmentation illustration of molecule 2-Trifluroacetyl along with radius scales centered at each atom. **h**. The feature of element-specific mGLI under three scales for the molecule using median statistics, as formulated in Equation 3.2.16. The magnitude of feature values increases as the scales increase. Features with alternative statistics measures for element-specific mGLI features are presented in Figure S4 and Figure S5.

instance, in the TPM, interactions such as α1-α1 and α2-α2 are represented as slightly thicker yellow blocks along the diagonal. In contrast, mGLI portrays these interactions as larger, more expressive, and prominently red blocks. This enhanced visualization enables a more precise distinction between the self-interaction of the alpha chain and other structural elements, such as the self-interaction between the β2 and β3 regions. Furthermore, the contrast between different values within each block is more pronounced in mGLI compared to TPM. This distinction is particularly

noticeable in blocks representing interactions like α1-α2, β1-α2, β1-β2, and so forth.

Topological data analysis vs knot data analysis

Recent years have witnessed the rapid growth of TDA in data science, driving its success in various domains, particularly in computational biology [5, 59, 60]. However, the major tool of TDA, persistent homology, has many drawbacks [18], including its qualitative and global nature, as well as the lack of localization. It is imperative to develop new mathematical/topological methods that overcome the drawbacks of TDA and potentially impact various domains of data science.

The proposed mGLI is a local method but recovers global topological properties at sufficiently large scales. Therefore, mGLI-based KDA models can outperform TDA models, as shown in this work. A direct comparison of TDA and KDA in protein B-factor prediction shows that KDA has a 17.2% improvement over TDA as sown in Figure 3.4**i** (ASPH vs mGLI). Besides, our mGLI models demonstrate superiority over TDA models [23, 31] for predicting protein-ligand binding affinity. Our model, based on mGLI features, achieves an average R value of 0.818 across the two PDBbind datasets. This surpasses the R values of 0.806 from PerSpect-ML [23] and 0.804 from PPS-ML [31] as well. The proposed KDA is computationally efficient, as it takes only a few minutes on a personal computer to generate mGLI features for a moderately sized dataset. Recently, a new KDA tool, persistent Khovanov homology, has also been reported [11]. Given the tremendous success of TDA, we expect that KDA will become a powerful new topological learning tool for a wide variety of problems in data science.

Methods

Multiscale Gauss linking integral

We introduced several essential definitions related to the Gauss linking integral in the Results section. Additional important proposition or theorems are presented below.

Proposition 3.2.1. The Gauss linking integral in Equation 3.2.1 is identical to the average of half the algebraic sum of inter-crossings in the projection of the two curves in any possible projection direction for both open and closed curves.

Theorem 3.2.2 (Panagiotou et al.[56])**.** For closed curves, the Gauss linking integral is an integer and a topological invariant. For open curves, the Gauss linking integral is a real number and a continuous function of curve coordinates.

Theorem 3.2.3 (The grand sum of the segmentation matrix)**.** The grand sum of the segmentation matrix of two curves equals the Gauss linking integral of the original curves:

$$L(l_1, l_2) = \sum_i \sum_j L(p_i, p_j). \tag{3.2.9}$$

Remark 3.2.8 (Generalization of Gauss linking integral)**.** Vassiliev measure, a generalization of Gauss linking integral, can be applied to open and closed curves in 3-space [88]. Similarly, the proposed mGLI obtained by combining the Gauss linking integral and multiscale process can naturally be applied to links, linkoids, open and closed curves, and other segmentable objects as shown in Figure 3.7**b**. It can be noticed that any element in the segmentation of the Gauss linking integral is defined on local curve segments. This indicates that one can define a generalized form of the multiscale Gauss linking integral if the segmentation of the Gauss linking integral is well-defined on local curve segments. In fact, for any topological or geometric structure that can be segmented into curve segments P_n, Q_m, we can define the following segmentation matrix:

$$\bar{G} = \begin{pmatrix} g(p_1, q_1) & g(p_1, q_2) & \cdots & g(p_1, q_m) \\ g(p_2, q_1) & g(p_2, q_2) & \cdots & g(p_2, q_m) \\ \vdots & \vdots & \ddots & \vdots \\ g(p_n, q_1) & g(p_n, q_2) & \cdots & g(p_n, q_m) \end{pmatrix}, \tag{3.2.10}$$

where

$$g(p_i, q_j) = \begin{cases} L(p_i, q_j) & \text{if } p_i \cap q_j \text{ is a null-set,} \\ 0 & \text{else.} \end{cases} \tag{3.2.11}$$

In the above definition, unlike in Equation 3.2.2, the curve segments in P_n and Q_m are allowed to intersect or even be equal. Thus, the mGLI can be applied in multiple topological/geometric structures as long as they can locally be represented as curve segments. Featurization can be similarly derived as in Equation 3.2.7.

mGLI featurization for B-factor prediction

We consider a protein as an open curve, acknowledging that the polypeptide chain of a protein molecule can be seen as an open polygon l whose vertices are corresponding to the C_α atoms, while the edges represent the pseudobonds that connect a C_α atom to another one in an adjacent amino acid residue. We propose a curve segmentation induced by C_α atoms:

$$p_i = \{x \in l_1 | f(x, c_i) = \inf_{c \in C} f(x, c)\}, 1 \leq i \leq n, \tag{3.2.12}$$

where $f(a, b)$ is the distance of points a and b along l, c_i is the 3D coordinates of a C_α atom, and C is the set of C_α atoms. Then, the $d(p_i, q_j)$ assumed in Equation 3.2.3 can be defined:

$$d(p_i, q_j) = d_E(c_i, c_j), \tag{3.2.13}$$

where d_E is the Euclidean distance in the 3D space.

Then, according to the generalized multiscale Guass linking integral, the segmentation of Gauss linking integral that investigates the inter-crossings between segments of the protein can be given:

$$\begin{aligned} G &= \begin{pmatrix} L(p_1, p_1) & L(p_1, p_2) & \cdots & L(p_1, p_n) \\ L(p_2, p_1) & L(p_2, p_2) & \cdots & L(p_2, p_n) \\ \vdots & \vdots & \ddots & \vdots \\ L(p_n, p_1) & L(p_n, p_2) & \cdots & L(p_n, p_n) \end{pmatrix} \\ &= \begin{pmatrix} 0 & L(p_1, p_2) & \cdots & L(p_1, p_n) \\ L(p_2, p_1) & 0 & \cdots & L(p_2, p_n) \\ \vdots & \vdots & \ddots & \vdots \\ L(p_n, p_1) & L(p_n, p_2) & \cdots & 0 \end{pmatrix}. \end{aligned}$$

The localized scaled Gauss linking integral, detailed in Remark Remark 3.2.7, is a natural way to characterize each C_α atom in B-factor predictions. We naturally choose a segment that precisely covers a single C_α atom along the polymer chain. Additionally, in our study, the multiscale scheme is selected to start from 5Å and extend up to 17Å, with each scale interval set at 1Å. This choice is based on the fact that the average distance between C_α atoms is approximately 3.8Å.

Such a selection of the multiscale scheme results in a powerful featurization method that provides abundant representations of local protein structures.

Traditional B-factor analysis methods predominantly concentrate on individual atoms and their spatial positions in three-dimensional space, accounting for the thermal motion and disorder of atoms within a protein structure. However, the incorporation of bonding interactions between atoms, which indirectly impacts the observed B-factor values, is rarely employed in B-factor analysis. Through the incorporation of mGLI, our method introduces the notion of pseudobonds between protein atoms, effectively capturing the influence of bonding interactions. The integration of knot theory with the multiscale procedure enables the precise localization of measurements, capitalizing on the spatial positions and atomic environments. The synergy between multiscale analysis and knot theory culminates in a robust method for predicting protein B-factors, showcasing the potential of multiscale approaches in effectively pinpointing measurements derived from knot theory.

mGLI featurization for protein-ligand complex

Localized scaled Gauss linking integral is also utilized to characterize protein-ligand interactions. This approach defines distinct curve segments and computes integrals with other segments across various scales. For molecular structures, we adopt atom-specific curve segmentation. Each atom c_i in a protein or ligand molecule is linked by multiple covalent bonds to neighboring atoms, determining the curve segmentation specific to c_i. These segments originate from the central atom and extend to the midpoint of associated covalent bonds, resulting in atom-specific curve segmentation.

$$p_i = \{x \in l | f(x, c_i) \leq \frac{1}{2} f(c, c_i), c \in C\}, \tag{3.2.14}$$

Here, C represents the set of adjacent atoms connected to atom c_i by covalent bonds, and l denotes the straight line along each covalent bond.

We focus on the binding core region where protein-ligand interactions primarily occur, extracting protein atoms within a 12 Å cutoff distance from the ligand. We can obtain atom-specific

curve segmentations for both the protein and ligand. Using these segmentations (p_i in protein and q_j in ligand), we compute atom-by-atom Gauss linking integrals (a-GLI) $L(p_i, q_j)$. Multiple segment pairs between the two atoms may exist, resulting in numerous Gauss linking integral between a segment pair. We consider the absolute Gauss linking integrals to mitigate curve orientation effects. Due to the multiple integrals between pairs, we utilize statistical analysis, specifically median and standard deviation, to define $L(p_i, q_j)$.

Element-specific approach is used in designing mGLI protein-ligand features. Specifically, we primarily focus on the protein atom groups of four elements (C, N, O, and S) within the protein, while considering atom groups of ten elements (C, N, O, H, S, P, F, Cl, Br, and I) within the ligand. We extract these atom groups in the core binding region, and then apply mGLI to characterize pairwise interactions between these atom groups from the protein and ligand.

Let P_n^C and Q_m^N represent collections of carbon (C) atom-specific curve segmentations in the protein and nitrogen (N) atom-specific curve segmentations in the ligand, respectively, given by $P_n^C = \{p_i^C | i = 1, 2, \cdots, n\}$ and $Q_m^N = \{q_j^N | j = 1, 2, \cdots, m\}$. We use the two groups to illustrate element-specific mGLI for protein-ligand featurization. The atomic coordinates in the two groups are labeled as $\{\mathbf{r}_i^C | i = 1, 2, \cdots, n\}$ and $\{\mathbf{r}_j^N | j = 1, 2, \cdots, m\}$. With the atom-by-atom Gauss linking integral $L(p_i^C, q_j^N)$ defined, we further determine the multiscale element-by-element Gauss linking integral. Assuming a scale $R = \{r_0, r_1, r_2, r_3, \cdots, r_k\}$ where $0 = r_0 < r_1 < r_2 < \cdots < r_k$, the distance between p_i^C and q_j^N is denoted as $d(p_i^C, q_j^N) = d_E(\mathbf{r}_i^C, \mathbf{r}_j^N)$ (in Å), where $d_E(\cdot, \cdot)$ indicates the Euclidean distance. The scaled Gauss linking integral $G^{r_t, r_{t+1}}$ in Equation 3.2.3 for curve segments generalizes to atom-by-atom Gauss linking integral. Atom-specific localized scaled Gauss linking integrals between two atom groups can be similarly derived as in Equation 3.2.5 and Equation 3.2.6:

$$J^{r_t, r_{t+1}}(p_i^C, Q_m^N) = \sum_{s=1}^{m} G_{is}^{r_t, r_{t+1}},$$
$$J^{r_t, r_{t+1}}(q_j^N, P_n^C) = \sum_{s=1}^{n} G_{sj}^{r_t, r_{t+1}}$$

where the second variable in $J^{r_t, r_{t+1}}$ indicate linking atom sets with the specified atom in the first variable. These expressions quantify the inter-crossing between a C atom-specific segmentation

p_i^C in the protein and a set of C atom-specific segmentations in the ligand within a given scale from r_t to r_{t+1}, or between a N atom-specific segmentation q_i^N in the ligand and a set of C atom-specific segmentations in the protein within a given scale.

To provide a scalable description of atomic interactions between two atom groups, we compute all atom-specific localized scaled Gauss linking integrals $J^{r_t,r_{t+1}}(p_i^C, Q_m^N)$ for $i = 1, 2, \cdots, n$, and $J^{r_t,r_{t+1}}(q_j^N, P_n^C)$ for $j = 1, 2, \cdots, m$. Statistical measures are then used to determine the multiscale element-specific Gauss linking integral (e-GLI) through the following formulations:

$$\begin{aligned}
&J^{r_t,r_{t+1}}(P_n^C, Q_m^N) = \text{statistics of} \\
&\{J^{r_t,r_{t+1}}(p_1^C, Q_m^N), J^{r_t,r_{t+1}}(p_2^C, Q_m^N), \cdots, J^{r_t,r_{t+1}}(p_n^C, Q_m^N)\}, \\
&J^{r_t,r_{t+1}}(Q_m^N, P_n^C) = \text{statistics of} \\
&\{J^{r_t,r_{t+1}}(q_1^N, P_n^C), J^{r_t,r_{t+1}}(q_2^N, P_n^C), \cdots, J^{r_t,r_{t+1}}(q_m^N, P_n^C)\}
\end{aligned} \tag{3.2.15}$$

We employ various statistical measures such as sum, minimum, maximum, mean, and median in Equation 3.2.15, which depict the atomic interactions between C atom-specific segmentations in the protein and N atom-specific segmentations in the ligand within the scale $[r_t, r_{t+1}]$. We consider the two formulations in Equation 3.2.15 as protein and ligand element-specific Gauss linking integral, respectively.

We can extend starting point of the scale interval to 0, giving rise to following formulation:

$$\begin{aligned}
&J^{0,r_{t+1}}(P_n^C, Q_m^N) = \text{statistics of} \\
&\{J^{0,r_{t+1}}(p_1^C, Q_m^N), J^{0,r_{t+1}}(p_2^C, Q_m^N), \cdots, J^{0,r_{t+1}}(p_n^C, Q_m^N)\}, \\
&J^{0,r_{t+1}}(Q_m^N, P_n^C) = \text{statistics of} \\
&\{J^{0,r_{t+1}}(q_1^N, P_n^C), J^{0,r_{t+1}}(q_2^N, P_n^C), \cdots, J^{0,r_{t+1}}(q_m^N, P_n^C)\}
\end{aligned} \tag{3.2.16}$$

We refer to the first and second approaches as mGLI-bin and mGLI-all featurization, respectively. In characterizing protein-ligand complexes, we define the scale radius set as $R = \{0, 2, 3, \cdots, 11, 12\}$ (in Å). Each of these featurization approaches results in an mGLI feature vector with a length of 40 (number of element combinations) × 2 (e-GLI fro two formulations in Equation 3.2.15) × 11 (scale number) × 5 (statistics for e-GLI) × 2 (statistics for a-GLI) = 8800. Figure 3.7**e-f** give an illustration of protein and ligand element-specific mGLI features.

Figure 3.7**f** illustrates a few cases of protein or ligand element-specific mGLI over the radius scales based on statistics of summation for two formulations in Equation 3.2.16. Additional cases are provided in Figure S2 and Figure S3.

We investigate the potential improvements in modeling performance resulting from employing statistical measures for mGLI features. Figure S8, Figure S9 and Figure S10 demonstrate the effectiveness of utilizing various statistical measures. Comparative analysis in subsection B in Appendix file validates the enhancement induced by incorporating additional statistical measures.

Adjusting the upper scale of protein-specific mGLI features could lead to an improvement in modeling performance. Figure S11 presents the resulting performance comparisons across various upper scales r_k, ranging from 12 to 20. Despite the increase in upper scales, the modeling performance remains consistent, indicating that an upper scale of 12 Å is adequate for ensuring optimal mGLI feature performance. The scale range and equal partitioning with an increment of 1 Å are appropriate for capturing local atomic interactions and recovering global molecular interactions.

mGLI featurization for small molecules

The mGLI featurization for small molecules can utilize the same approach based on the aforementioned 10 atom groups. Two mGLI feature strategies for ligands are available: mGLI-bin-lig and mGLI-all-lig, depending on local integral scale ranges. For a ligand with atom-specific curve segmentations p_i and q_j, the atom-by-atom Gauss linking integral $L(p_i, q_j)$ is determined using median statistics, adhering to the element-specific strategy to capture more atomic interactions. For atom-specific curve segmentations p_i^C $(i = 1, 2, \cdots, n)$ and q_j^N $(j = 1, 2, \cdots, m)$, statistics including summation, minimum, maximum, mean, and median are applied to the multiscale element-specific Gauss linking integral in equations such as Equation 3.2.15, or Equation 3.2.16. The scale values are defined as $R = \{0, 2.0, 2.44, 2.98, 3.63, 4.43, 5.41, 6.59, 8.05, 10\}$ for characterizing small molecules. Both mGLI-bin-lig and mGLI-all-lig features have a length of 2475. The upper scale of 10 Å is reasonable based on the 3D structure size of general small molecules as analyzed for hERG blockade molecules

in Figure S14.

An illustration of the multiscale element-specific Gauss linking integral for a molecule is depicted in Figure 3.7**g-h**, with corresponding additional feature analysis provided in Figure S4 and Figure S5.

Additional molecular descriptors and machine learning algorithms

In this work, transformer and autoencoder-based natural language processing (NLP) molecular descriptors are employed to enhance mGLI knot learning for various predictive tasks. Details about these descriptors are provided in subsection C in the Appendix file. Additionally, the integration of various molecular descriptors with machine learning and deep learning algorithms is discussed in the Appendix file.

3.3 Evolutionary Khovanov homology

We encounter challenges in establishing a filtration process for links, to the extent that we lack even the concept of sublinks. In fact, morphisms in the category of links are provided by cobordisms, and cobordism constructions are geometric in nature. This presents a challenge in the application of links. Thus directly studying the filtration process on the category of links is not a favorable approach. Therefore, in order to obtain a persistent process for link versions, we consider establishing filtration from the perspective of Khovanov cochain complexes of links.

3.3.1 Smoothing link

Let L be a link diagram. Let $x \in \mathcal{X}(L)$ be a crossing of L. At crossing x, there are two smoothing options: the 0-smoothing denoted as $\rho_0(L, x)$ and the 1-smoothing denoted as $\rho_1(L, x)$. It is worth noting that $2^{\mathcal{X}(L)} = 2^{\mathcal{X}(\rho_0(L,x))} \sqcup 2^{\mathcal{X}(\rho_1(L,x))}$. Thus the Khovanov chain groups of $\rho_0(L, x)$ and $\rho_1(L, x)$ are subspaces of the Khovanov chain group of L without considering the gradings. Moreover, even when we consider gradings, the Khovanov complex $C(\rho_0(L, x))$ or $C(\rho_1(L, x))$ can still be a subcomplex of $C(L)$ in certain cases.

When x is a left-handed crossing, assume that $n = |\mathcal{X}(L)|$ is the number of crossing of L. Each crossing in $\mathcal{X}(L)$ can be written of the form $(s_1, s_2, \ldots, s_n)$. Let λ be the index of the crossing x

in $\mathcal{X}(L)$. We have a map $j_0 : 2^{\mathcal{X}(\rho_0(L,x))} \to 2^{\mathcal{X}(L)}$ given by

$$(s_1, s_2, \ldots, s_{n-1}) \to (s_1, \ldots, s_{\lambda-1}, 1, s_\lambda, \ldots, s_{n-1}).$$

Let $n_{-,0}$ be the number of left-handed crossings in $\mathcal{X}(\rho_0(L,x))$, and let $n_{+,0}$ be the number of right-handed crossings in $\mathcal{X}(\rho_0(L,x))$. It follows that

$$c(s) = c(j_0(s)), \quad n_{-,0} = n_- - 1, \quad n_{+,0} = n_+, \quad \ell(s) = \ell(j_0(s)) - 1.$$

Then, we have an isomorphism of vector spaces

$$V^{\otimes c(s)}\{\ell(s) + n_+ - 2n_-\} \cong V^{\otimes c(j_0(s))}\{\ell(j_0(s)) + n_{+,0} - 2n_{-,0}\},$$

which is given by the degree shift. The degree difference is

$$\ell(j_0(s)) + n_{+,0} - 2n_{-,0} - \ell(s) - n_+ + 2n_- = 1.$$

The height of both side are equal: $\ell(s) - n_- = \ell(j_0(s)) - n_{-,0}$. Thus the induced map

$$i_0 : C(\rho_0(L,x)) \to C(L)$$

is an inclusion of degree -1 shift from the Khovanov complex $C(\rho_0(L,x))$ to the Khovanov complex $C(L)$. Moreover, one can verify $i_0 d = d i_0$ step by step by confirming $i_0 d_\xi = d_\xi i_0$ for each ξ. Hence, $C(\rho_0(L,x))$ is the subcomplex of $C(L)$.

When x is a right-handed crossing, we can verify that $C(\rho_1(L,x))$ is a subcomplex of $C(L)$ using a similar approach as described above. Consider the map $j_1 : 2^{\mathcal{X}(\rho_1(L,x))} \to 2^{\mathcal{X}(L)}$ given by

$$(s_1, s_2, \ldots, s_{n-1}) \to (s_1, \ldots, s_{\lambda-1}, 0, s_\lambda, \ldots, s_{n-1}).$$

We can obtain an injection $i_1 : C(\rho_1(L,x)) \to C(L)$ of degree 1 shift from the Khovanov complex $C(\rho_1(L,x))$ to the Khovanov complex $C(L)$. Thus, we have the following proposition.

Proposition 3.3.1. Let L be a link, and let x be a crossing of L. If x is a left-handed crossing, $C(\rho_0(L,x))$ is a subcomplex of $C(L)$. If x is a right-handed crossing, $C(\rho_1(L,x))$ is a subcomplex of $C(L)$.

The construction described above is called the *smoothing link*, denoted by $\rho_x L$. Note that $\rho_x L = \rho_0(L, x)$ if x is left-handed, and $\rho_x L = \rho_1(L, x)$ if x is right-handed. By construction, we have the following result.

Lemma 3.3.2. Let L be a link, and let x, y be crossings of L. Then, we have $\rho_x \rho_y L = \rho_y \rho_x L$.

In view of Lemma 3.3.2, for a subset S of $\mathcal{X}(L)$, we obtain a link $\rho_S L$ by applying the smoothing link step by step to crossings in S. Obviously, $C(\rho_S(L, x))$ is the subcomplex of $C(L)$.

3.3.2 Evolutionary Khovanov homology

A *weighted link* is a link L equipped with a function $f : \mathcal{X}(L) \to \mathbb{R}$ on the set of crossings of L. We arrange the crossings in $\mathcal{X}(L)$ in ascending order of their assigned values, denoted as $x_1, x_2, \ldots, x_n$. Then, we have a filtration of links

$$L, \rho_{x_1} L, \rho_{x_2} \rho_{x_1} L, \ldots, \rho_{x_n} \cdots \rho_{x_2} \rho_{x_1} L.$$

Note that the link $\rho_{x_n} \cdots \rho_{x_2} \rho_{x_1} L$ is unknotted, comprising a collection of disjoint circles. The filtration of links characterizes the process by which a complex link is gradually untangled, crossing by crossing, through smoothing. This process can be understood as the evolution of a link from complexity to simplicity.

For any real number a, we have the subset $\mathcal{X}(L, a)$ of $\mathcal{X}(L)$ consists of crossings x such that $f(x) \leq a$. Then we have a link $\rho_{\mathcal{X}(L,a)} L$, which is called the *$a$-indexed link*.

Let $(\mathbb{R}, \leq)$ the category with real numbers as objects and pairs of form $a \leq b$ as morphisms.

Theorem 3.3.3. The construction $C(\rho_{\mathcal{X}(L,-)} L)$ is a functor from the category $(\mathbb{R}, \leq)^{\mathrm{op}}$ to the category of cochain complexes.

Proof. For any $a \leq b$, let $x_{t_1}, \ldots, x_{t_u}$ be the crossings in $\mathcal{X}(L, b) \backslash \mathcal{X}(L, a)$. By Proposition 3.3.1 and Lemma 3.3.2, the cochain complex $C(\rho_{\mathcal{X}(L,b)} L) = C(\rho_{t_1} \cdots \rho_{t_u} \rho_{\mathcal{X}(L,a)} L)$ is the subcomplex of $C(\rho_{\mathcal{X}(L,a)} L)$. Let us denote $\theta_{a,b} : C(\rho_{\mathcal{X}(L,b)} L) \to C(\rho_{\mathcal{X}(L,a)} L)$. For real numbers $a \leq b \leq c$,

we have the following commutative diagram.

$$\begin{array}{ccc} C(\rho_{\mathcal{X}(L,c)}L) & \xrightarrow{\theta_{b,c}} & C(\rho_{\mathcal{X}(L,b)}L) \\ & {\scriptstyle\theta_{a,c}}\searrow \quad \swarrow{\scriptstyle\theta_{a,b}} & \\ & C(\rho_{\mathcal{X}(L,a)}L) & \end{array}$$

It follows that $\theta_{a,b}\theta_{b,c} = \theta_{a,c}$. Note that $\theta_{a,a} = \mathrm{id}|_{C(\rho_{\mathcal{X}(L,a)}L)}$ for any real number a. The desired result follows. □

For real numbers $a \leq b$, we have links $\rho_{\mathcal{X}(L,a)}L$ and $\rho_{\mathcal{X}(L,b)}L$. Note that there is an inclusion of Khovanov cochain complexes

$$C(\rho_{\mathcal{X}(L,b)}L) \hookrightarrow C(\rho_{\mathcal{X}(L,a)}L).$$

This induces the morphism of Khovanov homology

$$\lambda_{a,b} : H(\rho_{\mathcal{X}(L,b)}L) \to H(\rho_{\mathcal{X}(L,a)}L).$$

The (a, b)-*evolutionary Khovanov homology* of the weighted link (L, f) is defined by

$$H^k_{a,b}(L, f) := \mathrm{im}\,(H^k(\rho_{\mathcal{X}(L,b)}L) \to H^k(\rho_{\mathcal{X}(L,a)}L)), \quad k \geq 0.$$

Remark 3.3.1. For a weighted link (L, f) with crossings $x_1, x_2, \ldots, x_n$ of ascending weights, one can also obtain a filtration of links

$$L, \rho_{x_n}L, \rho_{x_{n-1}}\rho_{x_n}L, \ldots, \rho_{x_1}\cdots\rho_{x_{n-1}}\rho_{x_n}L.$$

For any real number a, let $\mathcal{X}_a(L)$ be the set of crossing with weight $f(x) \geq a$. Then, the construction $C(\rho_{\mathcal{X}_-(L)}L)$ is a functor from the category $(\mathbb{R}, \leq)$ to the category of cochain complexes. For real numbers $a \leq b$, we define the (a, b)-*evolutionary Khovanov homology* of the weighted link (L, f) as

$$H^k_{a,b}(L, f) := \mathrm{im}\,(H^k(\rho_{\mathcal{X}_a(L)}L) \to H^k(\rho_{\mathcal{X}_b(L)}L)), \quad k \geq 0.$$

This definition shares the same fundamental idea as the previous definition.

The rank of $H^k_{a,b}(L, f)$ is called the (a, b)-*evolutionary Betti number*, denoted by $\beta_{a,b}(L, f)$, which is the crucial feature for us to conduct data analysis. In particular, if we take $a = b$, we have that $H^k_{a,b}(L, f) = H^k(\rho_{\chi(L,a)}L)$. Furthermore, we can define the (a, b)-*evolutionary unnormalized Jones polynomial* as

$$\hat{J}_{a,b}(L) = \sum_k (-1)^k \mathrm{qdim}\, H^k_{a,b}(L).$$

As a direct corollary of Proposition 3.3.3, we have the following result, which shows that the evolutionary Khovanov homology is a (co)persistence module [92].

Theorem 3.3.4. The evolutionary Khovanov homology $H : (\mathbb{R}, \leq)^{\mathrm{op}} \to \mathrm{Vec}_{\mathbb{K}}$ is a functor from the category $(\mathbb{R}, \leq)^{\mathrm{op}}$ to the category of $\mathbb{K}$-module.

Evolutionary Khovanov homology tracks how the generators of Khovanov homology evolve with changes in parameter filtration. This concept shares a remarkable similarity with persistent homology. Yet, there are fundamental distinctions between the evolution process of evolutionary Khovanov homology and the persistence process of persistent homology: the former relies on smoothing the link, while the latter is established through the Vietoris-Rips complex, ensuring a continuous persistence.

Example 3.3.2. Consider the link L in Figure 3.8. Link L has four crossings, labeled x_1, x_2, x_3, and x_4 in the figure. We consider the weighted functions $f, g : \mathcal{X}(L) \to \mathbb{R}$ defined by

$$f(x_1) = 1, f(x_2) = 2, f(x_3) = 3, f(x_4) = 5,$$

and

$$g(x_1) = 1, g(x_2) = 3, g(x_3) = 2, g(x_4) = 4.$$

This gives us the following filtrations of links:

$$L, \rho_{x_1}L, \rho_{x_2}\rho_{x_1}L, \rho_{x_3}\rho_{x_2}\rho_{x_1}L, \rho_{x_4}\rho_{x_3}\rho_{x_2}\rho_{x_1}L,$$

and

$$L, \rho_{x_1}L, \rho_{x_3}\rho_{x_1}L, \rho_{x_2}\rho_{x_3}\rho_{x_1}L, \rho_{x_4}\rho_{x_2}\rho_{x_3}\rho_{x_1}L.$$

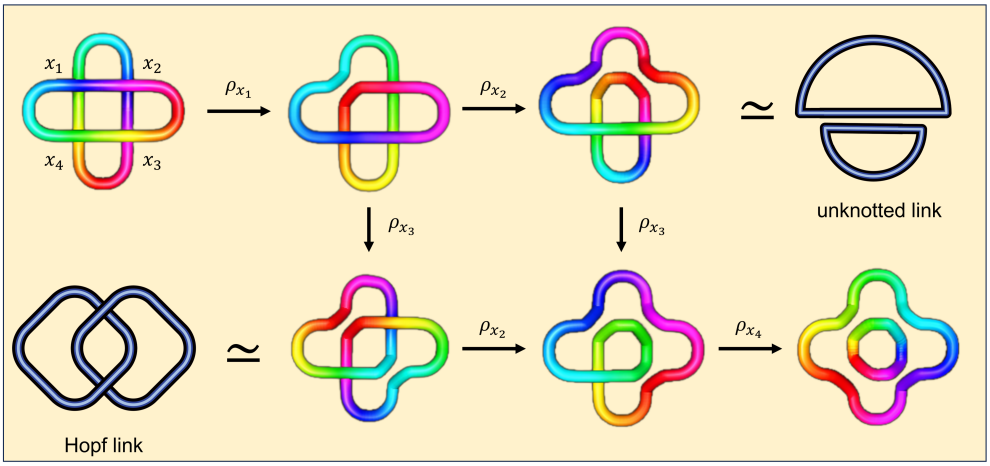


Figure 3.8 Link L produces different filtrations of links when processed through the crossings x_1, x_2, x_3 and through the crossings x_1, x_3, x_2.

Note that link L is unknotted, so its Khovanov homology is trivial. The links in the filtration given by the weighted function f are all unknotted links, hence their corresponding evolutionary Khovanov homologies are also trivial. On the other hand, note that the link $\rho_{x_3}\rho_{x_1}L$ is a Hopf link. Its Khovanov homology has four generators, and the Khovanov homology is given by

$$H^{-2}(\rho_{x_3}\rho_{x_1}L) \cong \mathbb{K} \oplus \mathbb{K},$$

$$H^{-1}(\rho_{x_3}\rho_{x_1}L) = 0,$$

$$H^{0}(\rho_{x_3}\rho_{x_1}L) \cong \mathbb{K} \oplus \mathbb{K}.$$

The evolutionary Khovanov homology $H^*_{2,2}(L, g)$ is non-trivial. This example illustrates that even if an unknotted link has trivial Khovanov homology, its evolutionary Khovanov homology may not be trivial. Moreover, different choices of weighting functions can produce different filtrations of links, leading to variations in their evolutionary Khovanov homology.

3.3.3 Representations of evolutionary features

In the previous section, we proved that evolutionary Khovanov homology is a functor. Consequently, evolutionary Khovanov homology also has representations similar to the barcode and persistence diagram in persistent homology theory.

Given a weighted link (L, f), since the links we consider have a finite number of crossings, we can arrange the crossings of the link L in ascending order of their weights as $x_1, x_2, \ldots, x_n$.

For any integers $1 \leq i \leq j \leq n$, we obtain an evolutionary Khovanov homology $H^k_{f(x_i),f(x_j)}(L,f)$. Let $\mathbf{H} = \bigoplus_i H_{f(x_i)}(L,f)$, and let $t : \mathbf{H} \to \mathbf{H}$ be given by the map $\lambda_{f(x_i),f(x_{i+1})} : H_{f(x_{i+1})}(L,f) \to H_{f(x_i)}(L,f)$. Then, for any element g in the polynomial ring $\mathbb{K}[t]$, we obtain a map

$$g : \mathbf{H} \to \mathbf{H}.$$

This implies that $\mathbf{H}$ is a finitely generated $\mathbb{K}[t]$-module. By the decomposition theorem for finitely generated modules over a principal ideal domain, we have:

Theorem 3.3.5. Let (L,f) be a weighted link. We have a decomposition of the evolutionary Khovanov homolog of (L,f) given by

$$\mathbf{H} \cong \bigoplus_k t^{b_k} \cdot \mathbb{K}[t] \oplus \left(\bigoplus_l t^{c_l} \cdot \frac{\mathbb{K}[t]}{t^{d_l} \cdot \mathbb{K}[t]} \right). \tag{3.3.1}$$

In the decomposition mentioned above, the $\mathbb{K}[t]$-module $\mathbf{H}$ has two components: the free part and the torsion part. For the free part, b_k represents a generator of the evolutionary Khovanov homology, which has weight 1 until smoothing at crossing x_{b_k} and becomes weight 0 after smoothing at crossing x_{b_k}. For the torsion part, c_l represents a generator that, after smoothing at crossing x_{c_l}, its weight becomes 0. Before smoothing at crossing x_{c_l}, this generator has weight 1 after smoothing at crossing $x_{c_l-d_l}$ and weight 0 before smoothing at crossing $x_{c_l-d_l}$.

Evolutionary Khovanov homology reflects the changes in homological generators of a link as it undergoes smoothing. This provides a more nuanced characterization of the topological features of the link. It also implies that the characteristic representation of evolutionary Khovanov homology is highly valuable in application. Common representations include barcode and persistence diagrams. Considering the decomposition of evolutionary Khovanov homology, each generator's information can be represented using intervals. For the decomposition (3.3.1), the generators of the free part can be represented by intervals $(-\infty, b_k]$, while for the torsion part, their generators can be represented by intervals $[c_l - d_l, c_l]$. This collection of intervals provides the barcode of evolutionary Khovanov homology. Another well-known representation is the persistence diagram. For the generators of the free part, they are represented by pairs of the form $(-\infty, b_k)$, while for the torsion part, pairs

of the form $(c_l - d_l, c_l)$ are used. These pairs correspond to points on the plane $\mathbb{R}^2$, and these discrete points provide the persistence diagram representation of evolutionary Khovanov homology. Other tools such as Betti curves and persistence landscapes are commonly used for representing and analyzing topological features. We demonstrate these representations in examples and applications.

Example 3.3.3. Consider the weighted trefoil knot (L, f) with $f : \mathcal{X}(L) \to \mathbb{R}$ defined as $f(x_1) = 1$, $f_{x_2} = 2$, and $f_{x_3} = 3$. Then, we have a filtration of links $L, \rho_{x_1}L, \rho_{x_2}\rho_{x_1}L, \rho_{x_3}\rho_{x_2}\rho_{x_1}L$, shown in Figure 3.9(a). This filtration illustrates the process of untangling a crossing of a trefoil by smoothing.

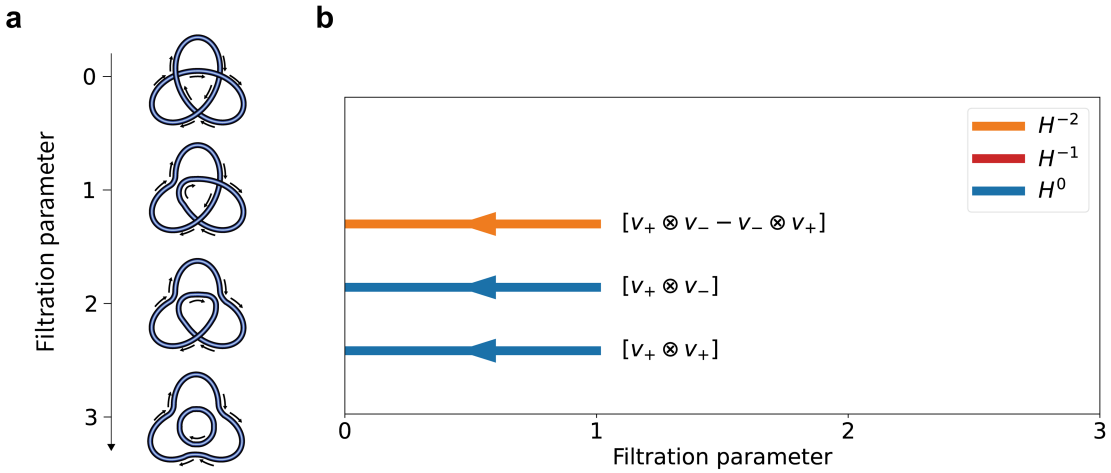


Figure 3.9 **(a)** The filtration of smoothing links of the weighted trefoil link (L, f); **(b)** The barcode of the evolutionary Khovanov homology of (L, f).

Note that the last two links are both unknotted, so they have trivial Khovanov homology. Now, let us first examine the Khovanov complex of the link $\rho_{x_1}L$. Note that the map $i_0 : 2^{\mathcal{X}(\rho_{x_2}L)} \to 2^{\mathcal{X}(L)}$ is given by $(s_1, s_2) \to (1, s_1, s_2)$. Hence, we can verify the commutative diagram between the Khovanov complex of $\rho_{x_1}L$ and the Khovanov complex of L.

$$\begin{array}{ccccccccccc}
0 & \longrightarrow & V \otimes V & \xrightarrow{d^{-2}} & V \oplus V & \xrightarrow{d^{-1}} & V \otimes V & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & V \otimes V \otimes V & \xrightarrow{d^{-3}} & \bigoplus_{i=1}^{3} V \otimes V & \xrightarrow{d^{-2}} & V \oplus V \oplus V & \xrightarrow{d^{-1}} & V \otimes V & \longrightarrow & 0
\end{array}$$

We select the basis of $V \otimes V$ as $v_+ \otimes v_+$, $v_+ \otimes v_+$, $v_+ \otimes v_+$, $v_+ \otimes v_+$, and for $V \oplus V$, the basis is chosen as $(v_+, 0)$, $(v_-, 0)$, $(0, v_+)$, $(0, v_-)$. Then, the left representation matrices of the differentials

d^{-2} and d^{-1} in the Khovanov complex $C^*(\rho_{x_1}L)$ are as follows:

$$B_{-2} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad B_{-1} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

From matrix calculations, we can obtain the generators of the Khovanov homology of $\rho_{x_1}L$ as in Table 3.2.

$H^{k,l}(\rho_{x_1}L)$	$k=0$	$k=-1$	$k=-2$
$l=0$	$[v_+ \otimes v_+]$	0	0
$l=-1$	0	0	0
$l=-2$	$[v_+ \otimes v_-]$	0	0
$l=-3$	0	0	0
$l=-4$	0	0	$[v_+ \otimes v_- - v_- \otimes v_+]$
$l=-5$	0	0	0
$l=-6$	0	0	$[v_- \otimes v_-]$

Table 3.2 The Khovanov homology $H^{k,l}(\rho_{x_1}L)$ of $\rho_{x_1}L$.

Therefore, the Khovanov homology of $\rho_{x_1}L$ is given by

$$H^{-2}(\rho_{x_1}L) \cong \mathbb{K} \oplus \mathbb{K},$$
$$H^{-1}(\rho_{x_1}L) = 0,$$
$$H^{0}(\rho_{x_1}L) \cong \mathbb{K} \oplus \mathbb{K}.$$

The corresponding unnormalized Jones polynomial is given by

$$\hat{J}(L) = \mathcal{X}_q(L) = \sum_k (-1)^k \mathrm{qdim}\, H^k(L) = 1 + q^{-2} + q^{-4} + q^{-6}.$$

Comparing Tables 3.1 and 3.2, we observe that the homology generators $[v_+ \otimes v_+]$, $[v_+ \otimes v_-]$, and $[v_+ \otimes v_- - v_- \otimes v_+]$ of $H^*(\rho_{x_1}L)$ are mapped to generators in $H^*(L)$. The generator $[v_- \otimes v_-]$ maps to the torsion part in $H^*(L)$. Assuming that 2 is invertible in $\mathbb{K}$, we can conclude that the generator $[v_- \otimes v_-]$ vanishes in $H^*(L)$. The corresponding barcode of the evolutionary Khovanov homology is shown in Figure 3.9(b). There are three bars, representing the generators $[v_+ \otimes v_+]$,

$[v_+ \otimes v_-]$, and $[v_+ \otimes v_- - v_- \otimes v_+]$. The arrows indicate that the cohomology generators emerge from later moments and persist toward earlier moments. These generators can be represented by intervals as $[0, 1]$, $[0, 1]$, and $[0, 1]$, respectively, each with degrees -1, -3, and -5. Besides, the $(0, 1)$-evolutionary unnormalized Jones polynomial of (L, f) is

$$\hat{J}_{0,1}(L) = \sum_k (-1)^k \mathrm{qdim} H^k_{0,1}(L) = q^{-1} + q^{-3} + q^{-5}.$$

3.3.4 Distance-based filtration of links

Traditional approaches to studying knots or links primarily focus on their topological properties. However, considering knots and links as objects within a metric space, their geometric properties are equally significant. In this section, we study the geometric information and topological characteristics of links by exploring distance-based filtration. This method allows us to extract richer and more effective information about links.

Consider a link L with crossings projected into a space $\mathbb{R}^2$. Let $\mathcal{X}(L)$ be the set of crossings. We have a function $f : \mathcal{X}(L) \to \mathbb{R}$ defined as follows: For a crossing $x \in \mathbb{R}^2$, we can construct a disk $D(x, r)$ with center x and radius r. Then, $f(x)$ is defined as the maximal real number r such that there are no other crossings within the interior of $D(x, r)$ apart from x. Mathematically, we have

$$f(x) = \max\{r | d(x, y) \geq r \text{ for any crossing } y \neq x \text{ in } \mathcal{X}(L)\}. \tag{3.3.2}$$

Geometrically, we connect points that are within a distance r. When $r < f(x)$, the point x remains isolated. Based on this construction, we obtain a weighted link (L, f). Using the method described in 3.3.1, we can obtain a filtration of links, which we refer to as the *distance-based filtration of links*. In the above construction, we can metaphorically say that we smooth out the isolated crossings first, gradually breaking down the entire knot step by step.

Now, for real numbers $a \leq b$, the (a, b)-evolutionary Khovanov homology of the link L is

$$H^k_{a,b}(L) := \mathrm{im}\,(H^k(\rho_{\mathcal{X}_a(L)}L) \to H^k(\rho_{\mathcal{X}_b(L)}L)), \quad k \geq 0.$$

Specifically, when a and b are sufficiently large, $H^k_{a,b}(L) = H^k(L)$. Conversely, when a and b are sufficiently small, we have $H^k_{a,b}(L) = 0$. We will illustrate this method with an example.

Example 3.3.4. Consider the link L embedded in $\mathbb{R}^3$ shown in Figure 3.10(a). This is a knot of 7_6 type.

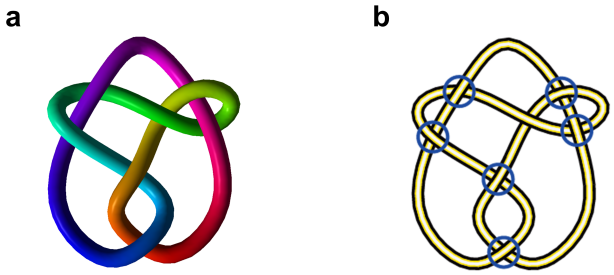


Figure 3.10 **(a)** A knot L of type 7_6 in 3-dimensional space; **(b)** The corresponding knot diagram of L.

The coordinates of these crossings are given below:

$$(-3.68122, 2.1618, 0.520849), (-2.31313, 4.52637, -0.526226),$$

$$(-0.291898, -0.0329635, 0.5289), (-0.000160251, -3.82999, -0.657526),$$

$$(1.29451, 3.02755, -0.309725), (2.99467, 4.45183, 0.450002),$$

$$(3.79753, 2.50471, -0.482759).$$

We project the knot onto the xy-plane, obtaining a knot diagram as shown in Figure 3.10(b).

Through the construction of the weighted function in Eq (3.3.2), we can obtain a weighted link (L, f). Figure 3.11(b) depicts the process of assigning weights to crossings. Subsequently, we can derive a filtration of links as illustrated in Figure 3.11(b). The variations in Figure 3.11(a) correspond to eight different cases, each yielding a distinct result. In Table 3.3, we describe the different critical distances corresponding to the changes in Figure 3.11(a), along with their respective link types. Here, 7_6 and 3_1 represent types in the knot table. Specifically, 3_1 denotes the trefoil. The links 5^2_1 and 2^2_1 are representations in Rolfsen's Table of Links, where 5^2_1 is the Whitehead link and 2^2_1 is the Hopf link. Additionally, $n\bigcirc$ denotes n separate unknots $\bigcirc$.

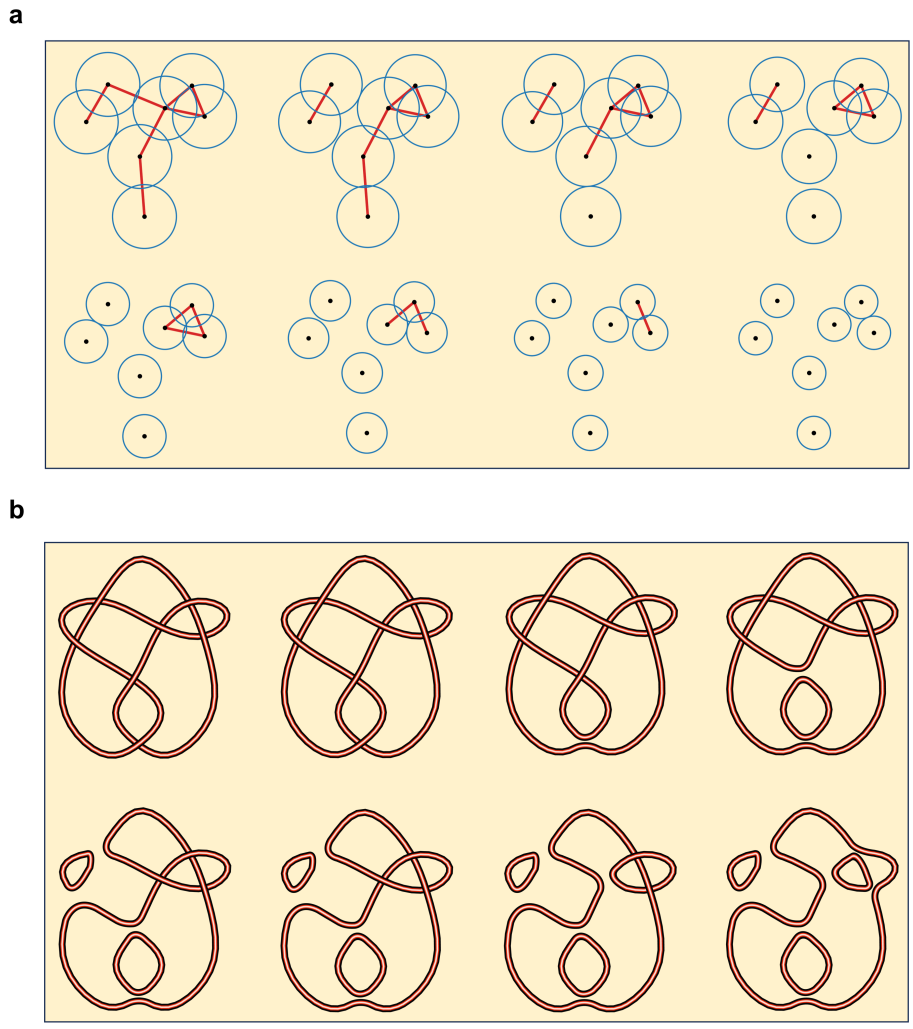


Figure 3.11 **(a)** As the distance decreases, isolated crossing points undergo gradual smoothing; **(b)** The filtration of links provided by the distance-based weighted function.

Filtration	1	2	3	4	5	6	7	8
Critical distance	2.019	1.953	1.904	1.724	1.366	1.279	1.109	1.053
Type of links	7_6	5_1^2	5_1^2+◯	5_1^2+ ◯	3_1+2◯	3_1+2◯	2_1^2+2◯	4 ◯

Table 3.3 The link types of the filtration of links.

Furthermore, for each filtration distance, we can obtain the corresponding Khovanov homology. Figure 3.12 illustrates the evolution of the graded Poincaré polynomial of Khovanov homology. The x-axis represents the filtration distance, while the y-axis denotes the Euler characteristic $\chi_1 = \chi_1(L_r)$ for the link L_r at distance r. Each subfigure in Figure 3.12 represents the surface of the graded Poincaré polynomial of the Khovanov homology $H^*(L_r)$.

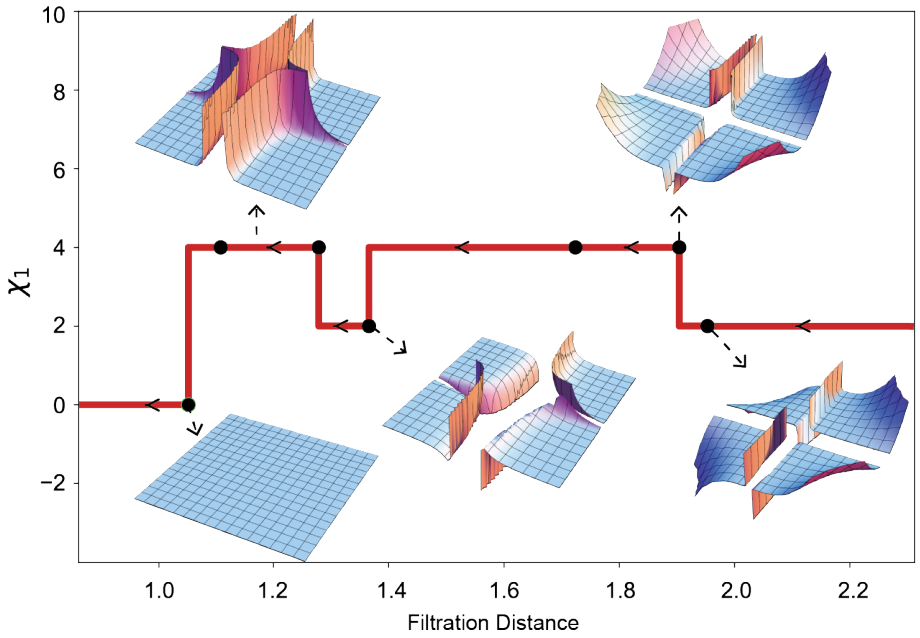


Figure 3.12 The representation of evolutionary Khovanov homology. Each subfigure represents the surface of the graded Poincaré polynomial of the Khovanov homology at the corresponding distance parameter. The y-axis denotes the value of Euler characteristic χ_q for the case $q = 1$.

The graded dimensions of the Khovanov homology of the links are the graded Betti numbers parameterized by q. When we set $q = 1$, it reduces to the usual Betti numbers, representing the number of generators. In persistent homology theory, for a given dimension k and distance r, the Betti number β_k is a real number. In evolutionary Khovanov homology, for a given dimension k and distance r, the graded Betti number $\beta_k(q)$ is a polynomial in q. In other words, the graded Betti number not only includes information about the number of generators but also about the degree of each generator. In Table 3.4, we observe the evolution of the graded Betti numbers in evolutionary Khovanov homology for different values of k.

	Distance				
Degree	0–1.053	1.053–1.109	1.109–1.366	1.366–1.953	1.953–2.019
$k \geq 1$	0	0	0	$q^4 + 1$	$q^3 + q + q^{-1}$
$k = 0$	0	$1 + q^{-2}$	$q^{-1} + q^{-3}$	$2 + 2q^{-2}$	$2q^{-1} + 2q^{-3}$
$k = -1$	0	0	0	q^{-2}	$2q^{-3} + q^{-5}$
$k = -2$	0	$q^{-4} + q^{-6}$	q^{-5}	$q^{-4} + q^{-6}$	$2q^{-5} + 2q^{-7}$
$k \leq -3$	0	0	q^{-9}	q^{-8}	$q^{-7} + 3q^{-9} + q^{-11} + q^{-3}$

Table 3.4 The graded Betti of the filtration of links.

3.3.5 Unzipping filtration of links

The unzipping filtration of links presents another innovative method for extracting geometric and topological information from link diagrams. Starting from a given initial point and direction, this technique involves progressively smoothing out each crossing along the link until none remain, simplifying the complex links into simple circles. This process preserves crucial geometric and topological characteristics, allowing for enhanced insight and detailed analysis at each stage of simplification. By systematically reducing visual complexity, unzipping filtration uncovers hidden structural features and enables systematic featurization of links, making it a valuable evolutionary technique compared to traditional knot theory techniques.

Given a link L, we can assign it a Gauss code representation. In this Gauss code, each crossing x of L is assigned a number $G(x)$ and its sign. We define a function $f : \mathcal{X}(L) \to \mathbb{Z}$ by $f(x) = G(x)$, resulting in a weighted link (L, f). This process involves starting at an initial crossing and progressively unwrapping the link in a specified direction, akin to unzipping a zipper. The links obtained in this evolutionary process form what is known as the unzipping filtration of links.

For real numbers $a \leq b$, the (a, b)-evolutionary Khovanov homology of the link L is given by

$$H^k_{a,b}(L) := \operatorname{im}(H^k(\rho_{\chi_b(L)}L) \to H^k(\rho_{\chi_a(L)}L)), \quad k \geq 0.$$

Unzipping filtration offers a distinctive alternative to distance-based filtration, with several unique attributes. First, it is less sensitive to local disturbances, making it more resistant to noise. Second, it has a strong connection to the Gauss code of a link diagram, directly relating the filtration process to the link's combinatorial properties. Third, unzipping filtration is less influenced by the spatial distribution of crossings. While distance-based methods may struggle in isolating crossings in complex local regions, unzipping filtration can sequentially separate and resolve individual crossings, providing a robust method for link analysis. This makes unzipping filtration a valuable complement to distance-based filtration as an effective evolutionary technique, offering an alternative perspective in the study of EKH.

Example 3.3.5. In this example, we employed evolutionary Khovanov homology of a unzipping filtration to investigate the knot structure of the SARS-CoV-2 frameshifting pseudoknot (PDB ID: 7LYJ). The knot structure was generated with the following process. Initially, we simplified the molecular structure by representing each RNA residue solely by its phosphorus atom, and connecting these atoms with linear segments to form a continuous backbone, directed from the 5' to 3' end, see Figure 3.13(a). This abstraction was followed by transforming the linear RNA chain into a closed loop, ensuring continuity by connecting the terminal phosphorus atoms. Such closure is essential for applying knot theory, as it converts the molecular structure into a topologically relevant form as in Figure 3.13(b). Lastly, to facilitate the analysis of the RNA's topological properties, we projected the closed-loop structure onto the xz-plane, generating a knot diagram. Along the numbering of crossings, the value of the weight function corresponds to the number assigned to each crossing. Consequently, we obtain a filtration of links, as shown in Figure 3.14.

Figure 3.13 **(a)** The representation of the SARS-CoV-2 frameshifting pseudoknot with the 5' and 3' ends; **(b)** The corresponding abstract knot of the SARS-CoV-2 frameshifting pseudoknot formed by connecting the two ends.

Using the method described in Section 3.3, we computed the evolutionary Khovanov homology of the corresponding knot diagram of the SARS-CoV-2 frameshifting pseudoknot. We obtained the corresponding barcode information, as shown in Figure 3.15. Note that the knot in Figure 3.13(b) is unknotted, and its Khovanov homology is trivial. However, Figure 3.15 shows that its evolutionary Khovanov homology is non-trivial, with four bars. Here, since the dimensions of generators remain unchanged during the evolution, but their degrees change, we use the vertical axis to represent the degree. We use polyline segments to indicate the changes in the degrees of these generators.

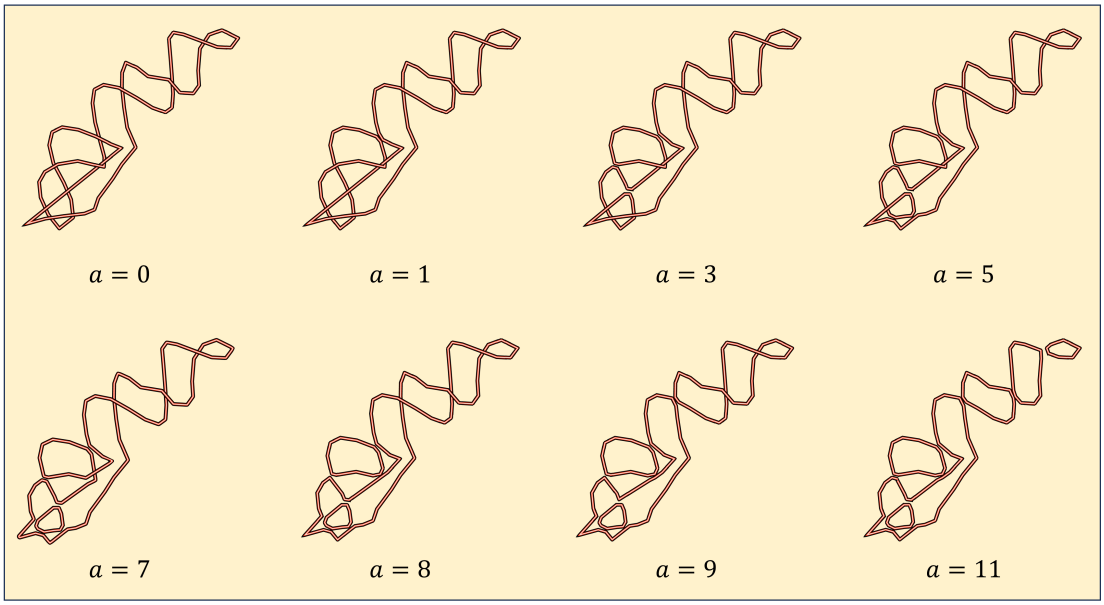


Figure 3.14 The filtration of smoothing links of the corresponding knot diagram of the SARS-CoV-2 frameshifting pseudoknot.

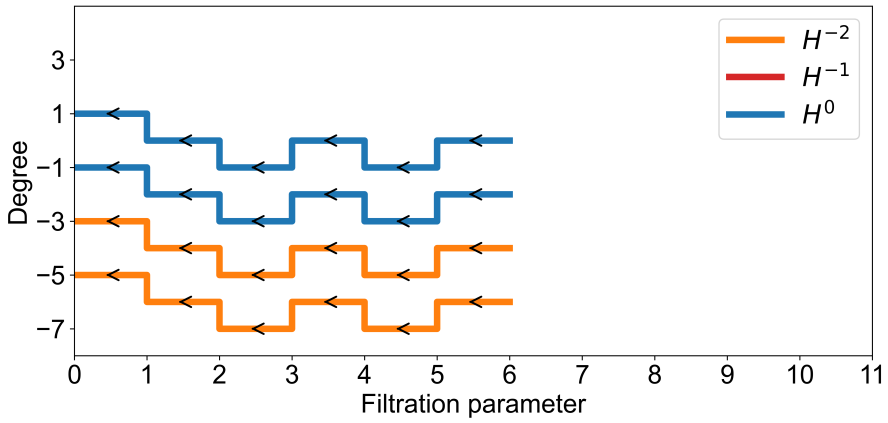


Figure 3.15 The barcode of the evolutionary Khovanov homology of the corresponding knot diagram of the SARS-CoV-2 frameshifting pseudoknot.

3.4 Persistent Khovanov homology of tangle

3.4.1 Tangle and Khovanov homology

In this section, we review the fundamental concepts and results related to tangles. We refer to [93] for basic concepts related to tangles. For the classical theory of Khovanov homology of tangles, we refer to [94] and [44]. Additionally, [95] explores the homology of $(1,1)$-tangles. Our approach in this work builds upon the relevant theory of the Khovanov homology of tangles as presented in [94].

3.4.1.1 Tangle

A *tangle* is an embedding of finitely many arcs and circles into $\mathbb{R}^2 \times [0,1]$. More precisely, a tangle T is defined as a 1-dimensional compact oriented piecewise smooth submanifold of $\mathbb{R}^3$ lying between two horizontal planes, with every boundary point of T lying on both the top and bottom planes. Another way to describe a tangle is as an embedding of finitely many arcs and circles into a 3-dimensional ball B^3, with the ends of the arcs required to lie on the boundary ∂B^3 of B^3. From now on, we will consider tangles embedded in the 3-dimensional ball B^3.

Figure 3.16 The tangle representations of a tangle in $\mathbb{R}^2 \times [0,1]$ and B^3.

Two tangles T and T' are *isotopic* if there exists a continuous map $H : B^3 \times [0,1] \to B^3$ such that $H(-,0)$ is the identity map, $H(-,1)$ maps T to T', and each map $H(-,t)$ is a homeomorphism that restricts to the identity map on ∂B^3.

A *tangle diagram* is a projection $T \to B^2$ of a tangle onto a maximal disk B^2 in B^3 such that it is injective everywhere except at a finite number of crossing points, which are the projections of only two points of the tangle. A tangle diagram can be seen as a generalization of the concepts of knot diagrams and link diagrams. Two tangle diagrams are equivalent if they are related by a series of Reidemeister moves.

From now on, unless otherwise specified, the tangles considered will always refer to tangle

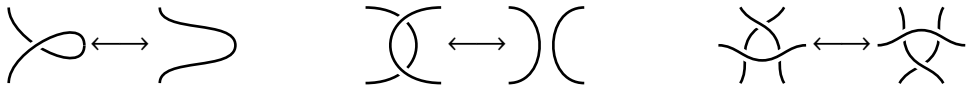


Figure 3.17 The three types of Reidemeister moves.

diagrams. For a tangle T, we denote the set of crossings of T by $\mathcal{X}(T)$. A crossing of the form  is called an overcrossing, while a crossing of the form is an undercrossing. Each crossing has a smoothing resolution:  $\Rightarrow$ + or $\Rightarrow$ + . Here, , called the 0-smoothing, is the tangle obtained by locally changing a crossing into two opposing arcs, one above the other. Similarly, , called the 1-smoothing, is obtained by locally changing a crossing into two opposing arcs, one to the left and one to the right. In this work, the 0-smoothings and 1-smoothings are always conducted on the undercrossing . Let $n = |\mathcal{X}(T)|$ be the number of crossings of T. Then there are 2^n states of the smoothing resolution of T. The 2^n states form a state cube $\{0,1\}^n$. Each vertex represents a state of the smoothing resolution and can be described by a sequence $(s_i)_{0\leq i\leq n} \in \{0,1\}^n$ of 0s and 1s of length n. Each edge represents two state sequences that differ in exactly one position. For an oriented tangle, we have the right-handed crossing and the left-handed crossing . We always assign the symbol + to the right-handed crossing and the symbol $-$ to the left-handed crossing. Let n_+ denote the number of right-handed crossings, and let n_- denote the number of left-handed crossings.

The study of the category of tangles involves the 2-category structure of tangles, which has been developed in [96, 97, 98]. Roughly speaking, this category has the boundaries of tangles as objects, tangles as 1-morphisms, and cobordisms connecting tangles as 2-morphisms. The 2-morphisms, depicted by movies, are generated by a family of moves as detailed in [99, 100]. In particular, the edges of the state cube can be characterized by a cobordism between the smoothings of a tangle.

3.4.1.2 Cobordism and bracket complex

Let M and N be two compact manifolds without boundary. A *cobordism* Σ between M and N is a compact manifold with boundary such that its boundary is the disjoint of M and N, $\partial\Sigma = M \sqcup N$.

Given a tangle T, recall that we can obtain a state cube $\{0,1\}^n$. Each vertex of the cube represents a tangle with the boundary ∂T. Moreover, there is a cobordism connecting the tangles corresponding

to the end vertices of an edge of the state cube. Considering such tangles corresponding to some smoothing of T as objects, and the cobordisms between these tangles as morphisms, we obtain a category $Cube(T)$. Generally, for a finite set of points B on a circle, we have a category $Cob^3(B)$, whose objects are the tangles corresponding to some smoothing of a tangle, and whose morphisms are the cobordisms between such tangles. For a fixed tangle T, the category $Cube(T)$ is a subcategory of $Cob^3(\partial T)$.

Let $\mathbf{k}$ be a commutative ring with a unit. One can extend $Cob^3(B)$ to a pre-additive category $\mathbf{k}Cob^3(B)$ as follows. The objects in $\mathbf{k}Cob^3(B)$ are the same as the objects in $Cob^3(B)$, and the morphisms in $\mathbf{k}Cob^3(B)$ are linear combinations of morphisms in $Cob^3(B)$. That is, the set $\mathrm{Hom}_{\mathbf{k}Cob^3(B)}(T, T')$ is a $\mathbf{k}$-module generated by the morphisms in the set $\mathrm{Hom}_{Cob^3(B)}(T, T')$ of morphisms from T to T' for any objects T and T' in $Cob^3(B)$.

Definition 3.4.1. For a pre-additive category C, we can define a category $\mathrm{Mat}_{\mathbf{k}}(C)$ with:

- Objects of the form $O = \bigoplus_{i=1}^{m} O_i$ for $O_i \in C$.
- Morphisms that are matrices of the form $f = (f_{ij})_{i,j} : \bigoplus_{i=1}^{m} O_i \to \bigoplus_{j=1}^{k} O'_j$, where $f_{ij} : O_i \to O'_j$ are morphisms in C for $1 \leq i \leq m$ and $1 \leq j \leq k$.
- Composition of morphisms given by matrix multiplication.

The construction $\mathrm{Mat}_{\mathbf{k}}(C)$ is an additive category, which is the additive closure of the category C. Furthermore, one can define a cochain complex in an additive category.

Definition 3.4.2. Let C be an additive category. The category $\mathbf{Ch}^{\bullet}(C)$ of cochain complexes over C is defined as follows. Its objects are of the form

$$\cdots \longrightarrow \Omega^{r-1} \xrightarrow{d^{r-1}} \Omega^r \xrightarrow{d^r} \Omega^{r+1} \longrightarrow \cdots$$

such that $d^{r+1} \circ d^r = 0$ for any r, and its morphisms are of the form $f^r : (\Omega^r_a, d_a) \to (\Omega^r_b, d_b)$ such that $f^{r-1} \circ d_a = d_b \circ f^r$ for any r.

Let T be a tangle with n crossings. The state cube associated with T has vertices indexed by states $s = (s_i)_{0 \leq i \leq n} \in \{0, 1\}^n$, where each s_i represents a smoothing choice at the i-th crossing of the

tangle. For a given state s, we denote $\ell(s) = \sum_{i=1}^{n} s_i$. Next, for the smoothing tangle T_s corresponding to state s, we assign a height function $h(s) = \ell(s) - n_-$, where n_- is the number of left-handed crossings in the original tangle T. This height measures the relative position of each smoothing state in the cube. Recall that the category $Cube(T)$ is a subcategory of $Cob^3(\partial T)$. We have a graded object in $\mathrm{Mat}(\mathbf{k}Cob^3(B))$ given by

$$\cdots \longrightarrow [[T]]^{k-1} \xrightarrow{d^{k-1}} [[T]]^k \xrightarrow{d^k} [[T]]^{k+1} \xrightarrow{d^{k+1}} \cdots,$$

where each graded piece $[[T]]^k = \bigoplus_{h(s)=k} T_s$ is a direct sum over all smoothing tangles T_s whose height $h(s) = k$. The morphism d^k is given by

$$d^k = \sum_{\xi} (-1)^{\mathrm{sgn}(\xi)} d_\xi : [[T]]^k \to [[T]]^{k+1},$$

where the sum is over all edges $\xi = (\xi_1, \ldots, \xi_{i-1}, \star, \xi_{i+1}, \ldots, \xi_{|\chi(T)|}) \in \{0, 1, \star\}^{|\chi(T)|}$ in the state cube that connect a state s with a neighboring state s' that differs by one position. Here, $\xi_j \in \{0, 1\}$ for $j \neq i$ and $\star$ indicates an edge connecting 0 to 1. The map d_ξ denotes the cobordism morphism between the smoothing tangles T_s and $T_{s'}$. The sign $\mathrm{sgn}(\xi)$ is determined by the number of 1s in ξ that appear before the first $\star$.

Note that the cube $Cube(T)$ is anti-commutative. This means that for each face of the cube, represented by the following diagram:

$$\begin{array}{ccc} T_s & \xrightarrow{d_\xi} & T_{s'} \\ {\scriptstyle d_\eta}\downarrow & & \downarrow{\scriptstyle d_{\eta'}} \\ T_{\tilde{s}} & \xrightarrow{d_{\tilde{\xi}}} & T_{\tilde{s}'} \end{array}$$

we have the anti-commutativity relation $d_{\tilde{\xi}} \circ d_\eta = -d_{\eta'} \circ d_\xi$. This condition ensures that the composition of differentials along the edges of each face of the state cube satisfies the appropriate signs, maintaining the structure of a cochain complex.

Proposition 3.4.1 ([94, Proposition 3.4])**.** The construction $([[T]]^*, d^*)$ above is a cochain complex over $\mathrm{Mat}(\mathbf{k}Cob^3(\partial T))$.

The cochain complex $([[T]]^*, d^*)$ is called the *bracket complex* of T. However, the bracket complex $([[T]]^*, d^*)$ is not a tangle invariant in the category $\mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3(\partial T)))$ of cochain complexes over $\mathrm{Mat}(\mathbf{k}Cob^3(\partial T))$. In [94], Bar-Natan obtains a new category from $\mathrm{Mat}(\mathbf{k}Cob^3(\partial T))$ by modding out some equivalence relations. In this new category, he proves that the bracket complex is a tangle invariant up to chain homotopy.

Let B be a finite set of points on a circle. The category $\mathbf{k}Cob^3_{/l}(B)$ is a localization of the category $\mathbf{k}Cob^3(B)$ defined as follows. The objects are the same as the objects in $\mathbf{k}Cob^3(B)$. The morphisms are those of $\mathbf{k}Cob^3(B)$ under the following equivalence relations:

- (S) $C + S^2 = 0$ for any cobordism C in $\mathbf{k}Cob^3(B)$. Here, S^2 is the cobordism of the 2-dimensional sphere.
- (T) $C + T^2 = 2C$ for any cobordism C in $\mathbf{k}Cob^3(B)$. Here, T^2 is the cobordism corresponding to the torus.
- $(4Tu)$ $C_{12} + C_{34} = C_{13} + C_{24}$. Here, C is a cobordism whose intersection with a ball is the union of four disks D_i, $i = 1, 2, 3, 4$, and C_{ij} is the cobordism obtained by removing D_i and D_j from C and replacing them with a tube that has the same boundary.

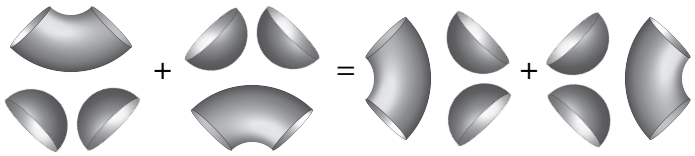


Figure 3.18 The cobordism representation of the $(4Tu)$ relation.

Since $\mathbf{k}Cob^3(B)$ is a pre-additive category, so is $\mathbf{k}Cob^3_{/l}(B)$. Moreover, one has an additive category $\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B))$.

Theorem 3.4.2 ([94, Theorem 1])**.** The construction $([[T]]^*, d^*)$ is a tangle invariant up to chain homotopy in the category $\mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(\partial T)))$ of cochain complexes over $\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(\partial T))$.

The above theorem says that the bracket complex of T in the category of cochain complexes over $\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(\partial T))$ is an invariant under Reidemeister moves up to chain homotopy.

Definition 3.4.3. Let T be a tangle. The *Khovanov complex* of T is the cochain complex $(Kh^*(T), d_T^*)$ given by $Kh^p(T) = [[T]]^{p+n_+-n_-}$ and $d_T^p = d^{p+n_+-n_-}$.

The Khovanov complex and the bracket complex differ by a height shift. Specifically, when the tangle T is a knot or link, the corresponding Khovanov complex is consistent with the Khovanov complex of the knot or link. Similarly, if two tangles T_1 and T_2 differ by some Reidemeister moves, there exists a chain homotopy equivalence $Kh(T_1) \simeq Kh(T_2)$.

Let $B \subseteq S^1$ be a finite set of points. Let $Cob^4(B)$ be the category whose objects are tangles in a disk D with boundary B, and whose morphisms are 2-dimensional cobordisms between these tangles in $D \times [-\epsilon, \epsilon] \times [0, 1]$ with boundary $B \times [-\epsilon, \epsilon] \times [0, 1]$. The construction Kh gives a functor $Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B)))$ from the category $Cob^4(B)$ of tangles with boundary B to the category of cochain complexes over $\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B))$.

Theorem 3.4.3. The functor $Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B)))$ maps the equivalence classes of isotopy of tangles to the equivalence classes of chain homotopy of cochain complexes.

It is worth noting that Bar-Natan's construction directly forms cochain complexes in the category $\mathbf{k}Cob^3_{/l}(B)$, which provides a more fundamental approach compared to the Khovanov complex constructed within the framework of topological quantum field theory (TQFT). However, this more intrinsic construction comes with a significant limitation: we cannot directly define Khovanov homology because the category $\mathbf{k}Cob^3_{/l}(B)$ is not an abelian category.

3.4.1.3 Khovanov homology of tangles

Let $\mathcal{A}b$ be an abelian category. Note that any functor $\mathcal{F} : \mathbf{k}Cob^3_{/l}(B) \to \mathcal{A}b$ can extend to a functor $\mathcal{F} : \mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B)) \to \mathcal{A}b$. Thus one can obtain a functor $\mathcal{F}^\bullet : \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B))) \to \mathbf{Ch}^\bullet(\mathcal{A}b)$ given by $\mathcal{F}^\bullet(\Omega^*, d^*) = (\mathcal{F}\Omega^*, \mathcal{F}d^*)$. Recall that the homology is a functor $H : \mathbf{Ch}^\bullet(\mathcal{A}b) \to \mathcal{A}b$ from the category of cochain complexes to an abelian category. We have the definition of Khovanov homology of tangles as follows.

Definition 3.4.4. Let B be a finite set of points on a circle. Let $\mathcal{F} : \mathbf{k}Cob^3_{/l}(B) \to \mathcal{A}b$ be a functor into an abelian category. The *Khovanov homology of tangles* with respect to $\mathcal{F}$ is the composition

of functors

$$Cob^4(B) \xrightarrow{Kh_B} \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B)) \xrightarrow{\mathcal{F}^\bullet} \mathbf{Ch}^\bullet(\mathcal{A}b) \xrightarrow{H} \mathcal{A}b.$$

It can be verified that $H\mathcal{F}^\bullet Kh_B$ is an isotopy invariant of tangles with boundary B. The definition of Khovanov homology mentioned above relies on the functor $\mathcal{F}$. Recall that the category $\mathcal{M}od_{\mathbf{k}}$ of modules is an abelian category. In TQFT, there is a standard construction of the functor $\mathcal{F} : Cob^3(\emptyset) \to \mathcal{M}od_{\mathbf{k}}$, which yields the usual definition of Khovanov homology of links.

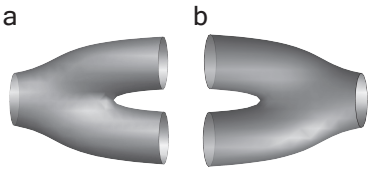


Figure 3.19 The cobordisms corresponding to the maps $\wedge$ and $\vee$.

Consider the functor $\mathcal{F} : Cob^3(\emptyset) \to \mathcal{M}od_{\mathbf{k}}$ constructed as follows. Let V be a $\mathbf{k}$-module generated by the elements v_+ and v_-. For a link L, the $\mathbf{k}$-module $\mathcal{F}(L)$ is the tensor product $\underbrace{V \otimes_{\mathbf{k}} V \otimes_{\mathbf{k}} \cdots \otimes_{\mathbf{k}} V}_{r(L)}$, where $r(L)$ is the number of circles of L. Note that the morphisms in $Cob^3(\emptyset)$ are compositions of those represented by cap and cup cobordisms, along with the morphisms $\wedge$ and $\vee$, corresponding to saddle cobordisms. The functor $\mathcal{F}$ is given as follows:

$$\mathcal{F}(\bigcap) = \epsilon : \mathbf{k} \to V, \quad 1 \mapsto v_+,$$

where $\bigcap : \emptyset \to \bigcirc$ denotes the morphism corresponding to the cap cobordism shown in Figure 3.20a,

$$\mathcal{F}(\bigcup) = \eta : V \to \mathbf{k}, \quad v_+ \mapsto 0, v_- \mapsto 1$$

where $\bigcup : \bigcirc \to \emptyset$ denotes the morphism corresponding to the cup cobordism depicted in Figure 3.20b.

$$\mathcal{F}(\wedge) = \Delta : V \to V \otimes_{\mathbf{k}} V, \quad \Delta : \begin{cases} v_+ \mapsto v_+ \otimes v_- + v_- \otimes v_+, \\ v_- \mapsto v_- \otimes v_-, \end{cases}$$

where $\wedge$ denotes the splitting of a circle into two circles, as shown in Figure 3.19a.

$$\mathcal{F}(\vee) = m : V \otimes_{\mathbf{k}} V \to V, \quad m : \begin{cases} v_+ \otimes v_+ \mapsto v_+, & v_- \otimes v_+ \mapsto v_-, \\ v_+ \otimes v_- \mapsto v_-, & v_- \otimes v_- \mapsto 0, \end{cases}$$

where $\vee$ denotes the merging of circles into a circle, as shown in Figure 3.19b. One can verify that the construction above can extend to a functor $\mathcal{F} : \mathbf{k}Cob^3_{/l}(\emptyset) \to \mathcal{M}od_{\mathbf{k}}$. Fix a link L, the construction $\mathcal{F}^\bullet Kh_\emptyset(L)$ is a cochain complex of $\mathbf{k}$-modules. The differential is the $\mathbf{k}$-module homomorphism $d^k = \sum_\xi (-1)^{\mathrm{sgn}(\xi)} \mathcal{F}(d_\xi)$, where d_ξ is the map given by $\mathcal{F}(\vee)$ or $\mathcal{F}(\wedge)$ on the components involved in merging or splitting, and the identity on other components. In this case, the homology $H(\mathcal{F}^\bullet Kh_\emptyset(L))$ coincides with the classical definition of Khovanov homology for links. Additionally, each element x in the cochain complex $\mathcal{F}^\bullet Kh_\emptyset(L)$ has a quantum grading given by $\Phi(x) = p(x) + n_+ - n_- + \theta(x)$, where $p(x)$ is the height of x in the cochain complex, and $\theta(x)$ is obtained by taking $\theta(v_+) = 1$ and $\theta(v_-) = -1$.

In the remainder of this paper, we will denote $\mathcal{K}_B = \mathcal{F}^\bullet Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ and $H(-;\mathcal{F}) = H\mathcal{K}_B : Cob^4(B) \to \mathcal{M}od_{\mathbf{k}}$ for simplicity. Unless otherwise specified, the notation $\mathcal{K}_\emptyset = \mathcal{F}^\bullet Kh_\emptyset : Cob^4(\emptyset) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ will always be based on the construction of $\mathcal{F}$ given in Section 3.4.1.3.

Now, consider the case where $\mathbf{k}$ is a field. Let $M = \bigoplus_{i\in\mathbb{Z}} M_i$ be a graded $\mathbf{k}$-linear space. The *graded dimension* of M is defined as the polynomial $\mathrm{qdim}\, M = \sum_{i\in\mathbb{Z}} q^i \dim M_i$ in the variable q. For the above construction of V, let $\deg v_+ = 1$ and $\deg v_- = -1$. Then we have $\mathrm{qdim}\, V = q + q^{-1}$. The *graded Euler characteristic* of a cochain complex C^* of $\mathbf{k}$-linear spaces is defined by $\mathcal{X}_q = \sum_k (-1)^k \mathrm{qdim}\, C^k$. Let T be a tangle. Then the Jones polynomial of T can be expressed as

$$J_q(T) = \sum_k (-1)^k \mathrm{qdim}\, H^k(T, \mathcal{F}).$$

When $\partial T = \emptyset$, $J_q(T)$ corresponds to the classical unnormalized Jones polynomial.

3.4.2 Persistent Khovanov homology of tangles

Tangles are common research objects across various disciplines, such as curve-like data which locally appear as tangles, and they have significant application potential. Studying the persistent

Khovanov homology of tangles is a natural idea, and it offers a new tool for understanding complex entangled structures. In this section, we introduce the concept of persistent Khovanov homology of tangles. Moreover, to ensure the computability of tangle homology, we provide a construction from the category $Cob^3(B)$ of tangles to the category of $\mathbf{k}$-modules.

3.4.2.1 Persistent Khovanov homology

Let B be a finite set of points on the circle S^1. Suppose $(X, \leq)$ is a poset with the partial order $\leq$. Then $(X, \leq)$ can be regarded as a category whose objects are the elements in X, and whose morphisms are the pairs $x \leq x'$ with $x, x' \in X$.

Definition 3.4.5. A *persistence tangle* with boundary B is a functor $\mathcal{P} : (X, \leq) \to Cob^4(B)$ into the category of tangles with boundary B.

Example 3.4.6. A movie of a tangle cobordism Σ is the intersection of the tangle cobordism in $D \times [-\epsilon, \epsilon] \times [0, 1]$ with cylinder spaces $D \times [-\epsilon, \epsilon] \times \{t\}$. This movie is called the *movie representation* of the tangle cobordism Σ. For each $t \in [0, 1]$, the intersection corresponds to a tangle. The movie representation of a tangle cobordism can be understood as depicting each frame of the movie.

A movie representation of the tangle cobordism in the category $Cob^4(B)$ can equivalently be described as a persistence tangle. Given a tangle cobordism Σ with boundary $B \times [-\epsilon, \epsilon] \times [0, 1]$, the functor $\mathcal{P} : ([0, 1], \leq) \to Cob^4(B)$ given by $\mathcal{P}(t) = \Sigma \cap (D \times [-\epsilon, \epsilon] \times \{t\})$ is a persistence tangle. The persistence tangle $\mathcal{P}(t)$ is also a movie representation of the tangle cobordism Σ.

Definition 3.4.7. Let $\mathcal{P} : (X, \leq) \to Cob^4(B)$ be a persistence tangle. The *persistent homology of* $\mathcal{P}$ is the composition of functors

$$(X, \leq) \xrightarrow{\mathcal{P}} Cob^4(B) \xrightarrow{H(-;\mathcal{F})} \mathcal{M}od_{\mathbf{k}}.$$

Here, $H(-; \mathcal{F}) : Cob^4(B) \to \mathcal{M}od_{\mathbf{k}}$ is the homology of tangles.

Specifically, for any $a \leq b$ in X, the (a, b)-persistent Khovanov homology of the persistence tangle $\mathcal{P} : (X, \leq) \to Cob^4(B)$ is given by

$$H^p_{a,b}(\mathcal{P}, B) = \text{im}\,(H^p(\mathcal{P}(a), B) \to H^p(\mathcal{P}(b), B)), \quad p \in \mathbb{Z}.$$

The graded dimension of $H^p_{a,b}(\mathcal{P}, B)$ is the Betti polynomial $\beta^p_{a,b}(q) = \sum_{\omega \in H^p_{a,b}(\mathcal{P},B)} q^{\Phi(\omega)}$, where $\Phi(\omega)$ is the quantum grading of ω.

Specifically, let $(X, \leq) = (\mathbb{Z}, \leq)$. Let $\mathbf{H} = \bigoplus_{a \in \mathbb{Z}} H^*(\mathcal{P}(a), B)$. For any $a \in \mathbb{Z}$, we have a map

$$z : H^*(\mathcal{P}(a), B) \to H^*(\mathcal{P}(a+1), B),$$

which induces a map $z : \mathbf{H} \to \mathbf{H}$. Thus, $\mathbf{H}$ is a $\mathbf{k}[z]$-module. This implies that the persistent Khovanov homology of tangles is also a persistence module. Under certain conditions, persistent Khovanov homology exhibits the structure theorem of persistence modules, the fundamental characterization of the corresponding barcodes, as well as the stability theorem for persistence modules. We shall not expend further in elaborating on these analogous results.

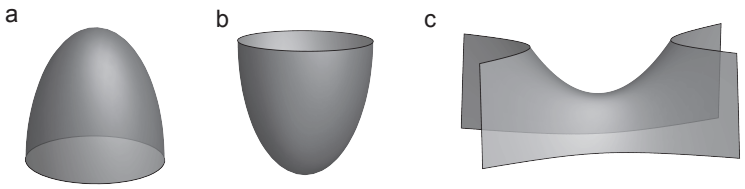


Figure 3.20 The subfigures a, b, and c represent the cap cobordism, cup cobordism, and saddle cobordism, respectively.

Let  be the morphism representing the saddle cobordism (see Figure 3.20c). It is known that the morphisms in the category $Cob^4(\emptyset)$ are generated by the three Reidemeister moves and the morphisms $\bigcap$, $\bigcup$, and .

Let $\bigcap : T \to T \amalg \bigcirc$ be the morphism of tangles that produces a circle. Here, $T \amalg \bigcirc$ denotes the disjoint union of the tangle T and the circle $\bigcirc$. Note that the cochain complex $\mathcal{K}_\emptyset(T \amalg \bigcirc) = \mathcal{K}_\emptyset(T) \otimes_{\mathbf{k}} V$. Thus, the morphism $\mathcal{K}_\emptyset(\bigcap) : \mathcal{K}_\emptyset(T) \to \mathcal{K}_\emptyset(T) \otimes_{\mathbf{k}} V$ is given by $\mathcal{K}_\emptyset(\bigcap)(x) = x \otimes v_+$. Therefore, the corresponding persistent Khovanov homology of $\bigcap$ is

$$\text{im}\, H^*(\bigcap; \mathcal{F}) = H^*(T; \mathcal{F}) \otimes v_+.$$

Let $\bigcup : T \amalg \bigcirc \to T$ be the morphism of tangles corresponding to the cup cobordism. The morphism $\mathcal{K}_\emptyset(\bigcup) : \mathcal{K}_\emptyset(T) \otimes_{\mathbf{k}} V \to \mathcal{K}_\emptyset(T)$ is given by $\mathcal{K}_\emptyset(\bigcup)(x \otimes v_+) = 0$ and $\mathcal{K}_\emptyset(\bigcup)(x \otimes v_-) = x$. Thus, the persistent Khovanov homology of $\bigcup$ is

$$\mathrm{im}\, H^*(\bigcup; \mathcal{F}) = H^*(T; \mathcal{F}).$$

Let  $: T \to T'$ be the morphism of tangles with a local saddle cobordism. We have a morphism $\mathcal{K}_\emptyset($  $) : \mathcal{K}_\emptyset(T) \to \mathcal{K}_\emptyset(T')$ of cochain complexes. Let $\widetilde{T}$ be the tangle obtained by changing  of T into . By [94], one obtains a cochain complex

$$\mathcal{K}_\emptyset(\widetilde{T}) = \mathcal{K}_\emptyset(T)[-1] \oplus \mathcal{K}_\emptyset(T')$$

with the differential given by $\widetilde{d}(z, z') = (-dz, \mathcal{K}_\emptyset($  $)(z) + d'z')$, where $\mathcal{K}_\emptyset(T)[-1]$ is the height shift of $\mathcal{K}_\emptyset(T)$ given by $\mathcal{K}_\emptyset(T)[-1]^p = \mathcal{K}_\emptyset(T)^{p+1}$. Here, $z \in \mathcal{K}_\emptyset(T)[-1]$, $z' \in \mathcal{K}_\emptyset(T')$, and d, d' are the differentials of $\mathcal{K}_\emptyset(T)[-1]$ and $\mathcal{K}_\emptyset(T')$, respectively. Thus, the morphism $\mathcal{K}_\emptyset($  $) : \mathcal{K}_\emptyset(T) \to \mathcal{K}_\emptyset(T')$ is given by $\mathcal{K}_\emptyset($  $)(z) = p_1\widetilde{d}z$. Here, $p_1 : \mathcal{K}_\emptyset(\widetilde{T}) \to \mathcal{K}_\emptyset(T')$ is the projection onto the component $\mathcal{K}_\emptyset(T')$. Therefore, one has a $\mathbf{k}$-module homomorphism $(p_1\widetilde{d})^* : H^*(T; \mathcal{F}) \to H^*(T'; \mathcal{F})$ given by $(p_1\widetilde{d})^*([z]) = [p_1\widetilde{d}z]$ for any cohomology class $[z] \in H^*(T; \mathcal{F})$. It follows that the persistent Khovanov homology of the saddle morphism is

$$\mathrm{im}\, H^*(\,$ $\,; \mathcal{F}) = \mathrm{im}\,(p_1\widetilde{d})^*.$$

Besides, the morphisms of the Khovanov cochain complexes induced by the three Reidemeister moves are chain homotopy equivalences, and the corresponding morphisms of the Khovanov homology are isomorphisms. Therefore, for any persistence tangle with boundary B, the persistent Khovanov homology is a composition of a sequence of the three types of morphisms mentioned above and can be computed step by step.

3.4.2.2 The construction of functors on tangles

In [44], Khovanov provides a construction of a functor from the category of $(1, 1)$-tangles to the category of modules. In [95], he assigns graded bimodules to tangle smoothings by considering all

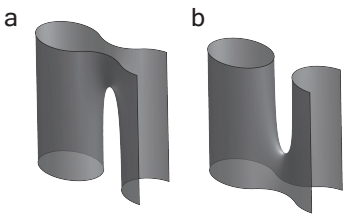


Figure 3.21 The tangle cobordisms corresponding to the saddle maps in our construction.

closures of tangles. However, these constructions have limitations for our application to persistent homology. In this section, we will present a different construction of $\mathbf{k}$-modules on tangles.

Let us define the functor $\mathcal{G} : Cob^3(B) \to \mathcal{M}od_{\mathbf{k}}$ as follows. Let V be the $\mathbf{k}$-module generated by the elements v_+ and v_-, and let W be the $\mathbf{k}$-module generated by an element w. For a tangle T in $Cob^3(B)$, the $\mathbf{k}$-module $\mathcal{G}(T)$ is the tensor product $\underbrace{W \otimes_{\mathbf{k}} \cdots \otimes_{\mathbf{k}} W}_{t(T)} \otimes \underbrace{V \otimes_{\mathbf{k}} \cdots \otimes_{\mathbf{k}} V}_{r(T)}$, where $r(T)$ and $t(T)$ represent the number of circles and arcs in T, respectively. Here, $V = \mathbf{k}\{v_+, v_-\}$ and $W = \mathbf{k}\{w\}$ are free $\mathbf{k}$-modules.

The functor $\mathcal{G}$ is defined as follows:

$\mathcal{G}($  $) : W \otimes W \to W \otimes W, \quad w \otimes w \mapsto 0,$

$\mathcal{G}($  $) : W \to W \otimes V, \quad w \mapsto w \otimes v_-,$

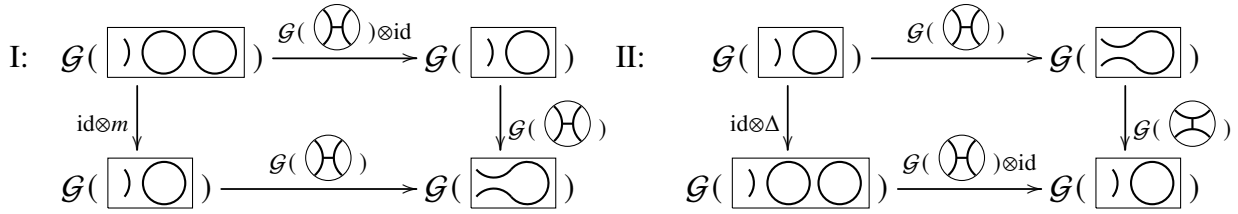
$\mathcal{G}($  $) : W \otimes V \to W, \quad \begin{cases} w \otimes v_+ \mapsto w, \\ w \otimes v_- \mapsto 0, \end{cases}$

and $\mathcal{G}$ coincides with $\mathcal{F}$ on the maps corresponding to the operations on the components of circles described in Section 3.4.1.3. Here, the square boxes indicate that the arcs or circles within them are independent components in the tangle, in contrast to the arcs in the round boxes in  which represent local arcs within the tangle. Besides, in the above construction, the degree of w is set to be -1.

Proposition 3.4.4. The construction $\mathcal{G} : Cob^3(B) \to \mathcal{M}od_{\mathbf{k}}$ is functorial.

Proof. Recall that the construction $\mathcal{G}$ coincides with $\mathcal{F}$ on circles and the map between circles. We will focus on mappings that include arcs. To show $\mathcal{G}$ is a functor, the nontrivial steps are to verify

the following diagrams commute.



Indeed, a step-by-step calculation shows that

$$\begin{array}{ccccc} \mathcal{G}(\,\cdot\,\bigcirc\bigcirc) & \xrightarrow{\mathcal{G}(\,\cdot\otimes\mathrm{id})} & \mathcal{G}(\,\cdot\,\bigcirc) & \xrightarrow{\mathcal{G}(\,\cdot\,)} & \mathcal{G}(\,\cdot\,\sim) \\ w\otimes v_+\otimes v_+ & \longmapsto & w\otimes v_+ & \longmapsto & w \\ w\otimes v_+\otimes v_- & \longmapsto & w\otimes v_- & \longmapsto & 0 \\ w\otimes v_-\otimes v_+ & \longmapsto & 0 & \longmapsto & 0 \\ w\otimes v_-\otimes v_- & \longmapsto & 0 & \longmapsto & 0 \end{array}$$

and

$$\begin{array}{ccccc} \mathcal{G}(\,\cdot\,\bigcirc\bigcirc) & \xrightarrow{\mathrm{id}\otimes m} & \mathcal{G}(\,\cdot\,\bigcirc) & \xrightarrow{\mathcal{G}(\,\cdot\,)} & \mathcal{G}(\,\cdot\,\sim) \\ w\otimes v_+\otimes v_+ & \longmapsto & w\otimes v_+ & \longmapsto & w \\ w\otimes v_+\otimes v_- & \longmapsto & w\otimes v_- & \longmapsto & 0 \\ w\otimes v_-\otimes v_+ & \longmapsto & w\otimes v_- & \longmapsto & 0 \\ w\otimes v_-\otimes v_- & \longmapsto & 0 & \longmapsto & 0. \end{array}$$

For the second diagram, we have that

$$\begin{array}{ccccc} \mathcal{G}(\,\cdot\,\bigcirc) & \xrightarrow{\mathcal{G}(\,\cdot\,)} & \mathcal{G}(\,\cdot\,\sim) & \xrightarrow{\mathcal{G}(\,\cdot\,)} & \mathcal{G}(\,\cdot\,\bigcirc) \\ w\otimes v_+ & \longmapsto & w & \longmapsto & w\otimes v_- \\ w\otimes v_- & \longmapsto & 0 & \longmapsto & 0 \end{array}$$

and

$$\begin{array}{ccccc} \mathcal{G}(\,\cdot\,\bigcirc) & \xrightarrow{\mathrm{id}\otimes\Delta} & \mathcal{G}(\,\cdot\,\bigcirc\bigcirc) & \xrightarrow{\mathcal{G}(\,\cdot\otimes\mathrm{id})} & \mathcal{G}(\,\cdot\,\bigcirc) \\ w\otimes v_+ & \longmapsto & w\otimes v_+\otimes v_- + w\otimes v_-\otimes v_+ & \longmapsto & w\otimes v_- \\ w\otimes v_- & \longmapsto & w\otimes v_-\otimes v_- & \longmapsto & 0. \end{array}$$

The desired result follows. The remaining verifications are straightforward. □

Note that the relations (S), (T), and $(4Tu)$ occur at the components of cobordism between closed curves. Therefore, the functor $\mathcal{G} : Cob^3(B) \to \mathcal{M}od_{\mathbf{k}}$ can descend to a functor $\mathcal{G} : \mathbf{k}Cob^3_{/l}(B) \to \mathcal{M}od_{\mathbf{k}}$ from the additive category $\mathbf{k}Cob^3_{/l}(B)$ to the abelian category $\mathcal{M}od_{\mathbf{k}}$. Thus we can obtain a functor $\mathcal{G}^\bullet : \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B))) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ between the category of cochain complexes. Now, we will give the detailed construction of the cochain complex of $\mathbf{k}$-module derived from $\mathcal{G}$. For a tangle T, let $\mathcal{G}[[T]]^k = \bigoplus_{h(s)=k} \mathcal{G}(T_s)$. And the map $\mathcal{G}(d)^k = \mathcal{G}(d^k) : \mathcal{G}[[T]]^k \to \mathcal{G}[[T]]^{k+1}$ is given by $\mathcal{G}(d^k) = \sum_{\xi}(-1)^{\mathrm{sgn}(\xi)}\mathcal{G}(d_\xi)$. Since $d^{k+1} \circ d^k = 0$, we have $\mathcal{G}(d^{k+1}) \circ \mathcal{G}(d^k) = \mathcal{G}(d^{k+1} \circ d^k) = 0$. Hence, the construction $(\mathcal{G}[[T]]^*, \mathcal{G}(d)^*)$ is a cochain complex. Let $\mathcal{G}Kh^p(T) = \mathcal{G}[[T]]^{p+n_+-n_-}$ and $\mathcal{G}(d_T)^p = \mathcal{G}(d)^{p+n_+-n_-}$. Thus, we have the following result.

Proposition 3.4.5. The construction $(\mathcal{G}Kh^*(T), \mathcal{G}(d_T)^*)$ is a cochain complex.

For any element x in the cochain complex $\mathcal{G}Kh^p(T)$, we define the quantum grading of x by $\Phi(x) = p + n_+ - n_- + \theta(x)$, where $\theta(x)$ is obtained by taking $\theta(v_+) = 1$, $\theta(v_-) = -1$, and $\theta(w) = -1$.

Lemma 3.4.6 ([101])**.** Let $\mathcal{A}$ and $\mathcal{B}$ be additive categories. Any additive functor $F : \mathcal{A} \to \mathcal{B}$ induces an additive functor $F^\bullet : \mathbf{Ch}^\bullet(\mathcal{A}) \to \mathbf{Ch}^\bullet(\mathcal{B})$ that preserves homotopy equivalences.

Recall that we have the functor $Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B)))$. By composing it with $\mathcal{G}^\bullet : \mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B))) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$, we obtain the functor $\mathcal{G}^\bullet Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$, which maps the category of tangles with boundary B to the category of cochain complexes of $\mathbf{k}$-modules.

Theorem 3.4.7. The functor $\mathcal{G}^\bullet Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ maps isotopy classes of tangles to homotopy classes of cochain complexes.

Proof. Note that the functor $\mathcal{G} : \mathbf{k}Cob^3_{/l}(B) \to \mathcal{M}od_{\mathbf{k}}$ is additive, and it extends to an additive functor $\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(B)) \to \mathcal{M}od_{\mathbf{k}}$. The desired result follows from a variant of [94, Theorem 4] and Lemma 3.4.6. □

3.4.3 Planar tangles and persistent Khovanov homology

In the study of persistent Khovanov homology for tangles, it is often the case that the boundary of the tangle does not remain fixed as the tangle evolves over persistence parameter. This presents a challenge for the application of persistence tangles. A natural approach is to consider that as the persistence parameter increases, the tangle at earlier times can be viewed as an interior part of the tangle at later times. The relationship between these two tangles can be described using operations induced by input planar tangles.

3.4.3.1 Input planar tangle

A *d-input planar tangle* consists of a large output disk equipped with d input disks, along with a collection of disjoint embedded arcs that are either closed or have endpoints on the boundary. These input disks are sequentially numbered from 1 to d, and both the input disks and the output disk are marked with $*$ as base points.

Let $\mathcal{T}(k)$ be the collection of all the classes of tangles with k endpoints up to Reidemeister moves. Suppose D is a d-input planar tangle such that there are k_r endpoints of arcs on the r-th input disk in D for $R = 1, 2, \ldots, d$. Then one has an operation

$$D : \mathcal{T}(k_1) \times \cdots \times \mathcal{T}(k_d) \to \mathcal{T}(k),$$

which embeds d tangles, each with $k_1, \ldots, k_d$ endpoints respectively, into the d-input planar tangle D by connecting their endpoints, resulting in a new tangle. Let $\mathcal{P}(k)$ be the vector space generated by the elements in $\mathcal{T}(k)$. Then the collection $\{\mathcal{P}(k)\}_{k\geq 0}$, equipped with the operation D, forms a planar algebra. For more details on planar algebras, refer to [102].

Now, let D be a 1-input planar tangle. We can obtain an operation

$$D : \mathcal{T}(k_1) \to \mathcal{T}(k)$$

by embedding a tangle T into D, resulting in a larger tangle $D(T)$, with T as a part of $D(T)$, as shown in Figure 3.22.

Our goal in this work is to establish the distance-based persistent Khovanov homology of tangles. A natural idea is to determine whether we can obtain a morphism $Kh(\mathcal{T}(k_1)) \to Kh(\mathcal{T}(k))$

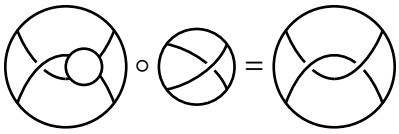


Figure 3.22 An example of the operation of a 1-input planar tangle.

of cochain complexes. Unfortunately, it is challenging to construct such a morphism of cochain complexes. Even with the constructions from TQFT, we have not been able to establish a morphism $\mathcal{F}^{\bullet}Kh(\mathcal{T}(k_1)) \to \mathcal{F}^{\bullet}Kh(\mathcal{T}(k))$ of cochain complexes.

3.4.3.2 The category $\mathcal{P}la$

Consider the category $\mathcal{P}la$ of tangles, where the objects are tangles and the morphisms are given by maps $T \to T' = D(T)$ for some 1-input planar tangle D. In this setting, any morphism in $\mathcal{P}la$ can be viewed as an inclusion of 1-dimensional manifolds, where arcs are mapped to either arcs or circles, and circles are mapped to circles.

For any morphism $T \to T'$, we can associate cochain complexes $(\mathcal{G}Kh^*(T), \mathcal{G}(d_T)^*)$ and $(\mathcal{G}Kh^*(T'), \mathcal{G}(d_{T'})^*)$. Our goal is to construct a map $\Psi : (\mathcal{G}Kh^*(T), \mathcal{G}(d_T)^*) \to (\mathcal{G}Kh^*(T'), \mathcal{G}(d_{T'})^*)$. Recall that each direct summand of $Kh^*(T)$ consists of a collection of disjoint arcs and circles. The map Ψ is a $\mathbf{k}$-module homomorphism defined as follows:

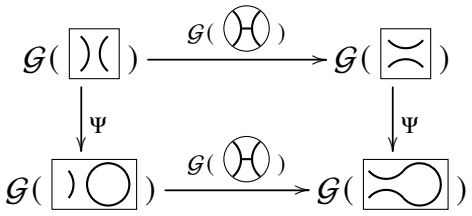
$$\Psi : \mathcal{G}(\boxed{}) \to \mathcal{G}(\boxed{}), \quad w \mapsto v_-,$$

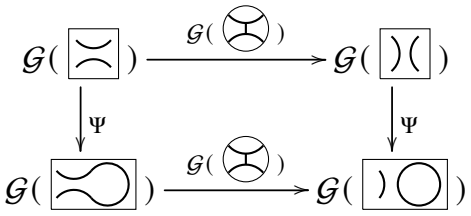
and Ψ acts as the identity on the identity maps $\boxed{} \to \boxed{}$ and $\boxed{} \to \boxed{}$ of independent components. In other words, Ψ maps arcs to arcs and circles to circles wherever the structure of the tangle is preserved, and performs the specified homomorphism on components that transition between arcs and circles.

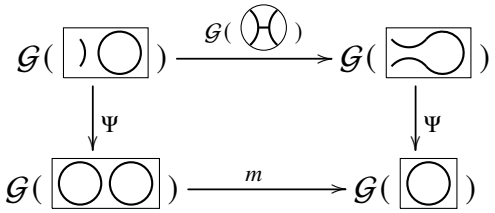
Theorem 3.4.8. The map $\Psi : \mathcal{G}Kh^*(T) \to \mathcal{G}Kh^*(T')$ is a morphism of cochain complexes.

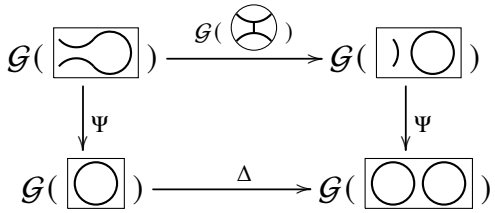
Proof. To prove Ψ is a morphism of cochain complexes, it suffices to show $\mathcal{G}(d_\xi) \circ \Psi = \Psi \circ \mathcal{G}(d_\xi)$. Here, $d_\xi : T_s \to T_{s'}$ is a saddle map given by the edge $\xi = (\xi_1, \ldots, \xi_{i-1}, \star, \xi_{i+1}, \ldots, \xi_{|\chi(T)|}) \in \{0, 1, \star\}^{|\chi(T)|}$ in the state cube that connect a state s with a neighboring state s' that differs by one position. Here, $\xi_j \in \{0, 1\}$ for $j \neq i$ and $\star$ indicates an edge connecting 0 to 1. Hence, we need to

prove that the following four diagrams are commutative.

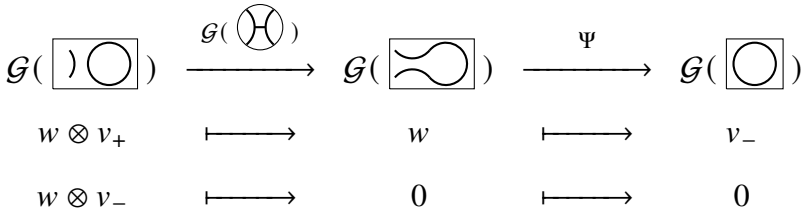
I: 

II: 

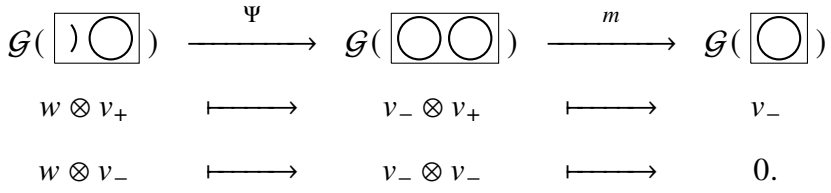
III: 

IV: 

We will only verify the third diagram. A straightforward calculation shows that

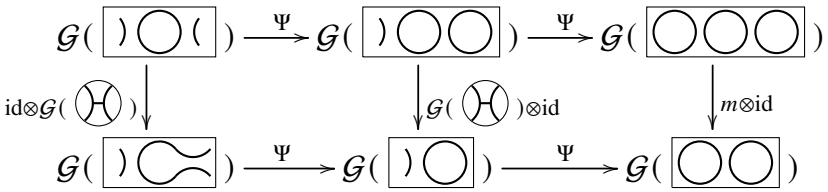


and



Thus, Diagram III commutes. The commutativity of the other diagrams can be verified similarly, following analogous calculations. □

Example 3.4.8. Now, we will give an example of tangles with more independent components. Consider the following diagram.



It is worth noting that $\mathcal{G}($  $) \otimes \mathrm{id} = \mathrm{id} \otimes m$ and $m \otimes \mathrm{id} = \mathrm{id} \otimes m$. The corresponding element mappings and their associated diagrams are listed as follows.

$$\begin{array}{ccccc} w \otimes v_+ \otimes w & \longmapsto & v_- \otimes v_+ \otimes w & \longmapsto & v_- \otimes v_+ \otimes v_- \\ \downarrow & & \downarrow & & \downarrow \\ w \otimes w & \longmapsto & v_- \otimes w & \longmapsto & v_- \otimes v_- \end{array}$$

$$\begin{array}{ccccc} w \otimes v_- \otimes w & \longmapsto & w \otimes v_- \otimes v_- & \longrightarrow & v_- \otimes v_- \otimes v_- \\ \downarrow & & \downarrow & & \downarrow \\ 0 & \longmapsto & 0 & \longmapsto & 0 \end{array}$$

The calculation shows that the above diagram of $\mathbf{k}$-modules is commutative.

Theorem 3.4.9. The construction $\mathcal{G}^\bullet Kh : \mathcal{P}la \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ is functorial.

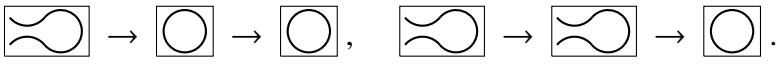
Proof. Let $T \xrightarrow{D} T' \xrightarrow{D'} T''$ be morphisms of tangles in the category $\mathcal{P}la$. We need to prove that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{G}^\bullet Kh(T) & \xrightarrow{\quad\Psi_D\quad} & \mathcal{G}^\bullet Kh(T') \\ & {\scriptstyle \Psi_{D'\circ D}} \searrow \quad \swarrow {\scriptstyle \Psi_{D'}} & \\ & \mathcal{G}^\bullet Kh(T''). & \end{array}$$

Here, $D' \circ D$ is the composition of morphisms in the category $\mathcal{P}la$. In other words, we need to prove that $\Psi_{D'\circ D} = \Psi_{D'} \circ \Psi_D$. It suffices to prove the diagram

$$\begin{array}{ccc} \mathcal{G}(T) & \xrightarrow{\quad\mathcal{G}(T\to T')\quad} & \mathcal{G}(T') \\ & {\scriptstyle \mathcal{G}(T\to T'')} \searrow \quad \swarrow {\scriptstyle \Psi_{\mathcal{G}(T'\to T'')}} & \\ & \mathcal{G}(T'') & \end{array}$$

is commutative. We only need to verify the commutativity for the two cases of morphisms $T \to T' \to T''$:



This follows from a straightforward step-by-step computation. The remaining part of the proof can be checked directly. □

It is worth noting that, at present, there is no definition of isotopy for tangles with different boundaries. Consequently, we do not have a result stating that the functor $\mathcal{G}^\bullet Kh : \mathcal{P}la \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ maps isotopy classes of tangles to homotopy classes of cochain complexes.

3.4.3.3 Homology functors for tangles

In the previous sections, we introduced a new construction $\mathcal{G} : Cob^3(B) \to \mathcal{M}od_{\mathbf{k}}$ for tangles. This construction is functorial and leads to two functors: $\mathcal{G}^\bullet Kh_B : Cob^4(B) \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$ and $\mathcal{G}^\bullet Kh : \mathcal{P}la \to \mathbf{Ch}^\bullet(\mathcal{M}od_{\mathbf{k}})$. The functor $\mathcal{G}^\bullet Kh_B$ is a tangle invariant up to homotopy equivalence, but it has limitations for applications because it requires the boundaries of tangles to be fixed. In contrast, although functor $\mathcal{G}^\bullet Kh$ does not capture tangle invariants, it has greater potential for application.

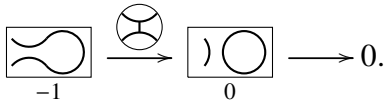
For a given tangle T, the constructions $\mathcal{G}^\bullet Kh_{\partial T}(T)$ and $\mathcal{G}^\bullet Kh(T)$ produce the same cochain complex. Thus, although $\mathcal{G}^\bullet Kh_{\partial T}$ and $\mathcal{G}^\bullet Kh$ are different functors, this does not impact the computation of the Khovanov homology of tangles. For practical purposes, we will use the homology functor associated with $\mathcal{G}^\bullet Kh$.

Definition 3.4.9. Let T be a tangle. The *Khovanov homology of T associated with $\mathcal{G}$* is defined by

$$H^p(T;\mathcal{G}) = H^p(\mathcal{G}^\bullet Kh(T)), \quad p \in \mathbb{Z}.$$

The Khovanov homology associated with $\mathcal{G}$ is a functor $H^p(-;\mathcal{G}) : \mathcal{P}la \to \mathcal{M}od_{\mathbf{k}}$. Moreover, if $\partial T = \emptyset$, the Khovanov homology associated with $\mathcal{G}$ reduces to the Khovanov homology of links, that is, $H^p(T;\mathcal{G}) = H^p(T;\mathcal{F})$ for any p. The Khovanov homology of tangles associated with $\mathcal{G}$ can be explicitly computed. The following computation provides a detailed example.

Example 3.4.10. Consider the tangle $T =$  . The corresponding cochain complex $Kh(T)$ of T in $\mathbf{Ch}^\bullet(\mathrm{Mat}(\mathbf{k}Cob^3_{/l}(\partial T))$ is described as follows:



The cochain complex $Kh^*(T)$ collapses at heights -1 and 0. The only nontrivial differential is $d^{-1} =$  $: Kh^{-1}(T) \to Kh^0(T)$. Applying to the functor $\mathcal{G}$, we have a cochain complex of

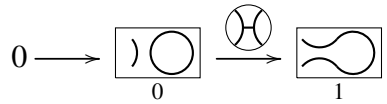
k-modules

$$0 \longrightarrow W \xrightarrow{d} W \otimes V \longrightarrow 0.$$

Here, $dw = w \otimes v_-$. A straightforward calculation shows that

$$H^p(T;\mathcal{G}) = \begin{cases} \mathbf{k}\{w \otimes v_+\}, & p = 0; \\ 0, & \text{otherwise.} \end{cases}$$

Recall that $\deg w = -1$. Then the quantum grading of $w \otimes v_+$ is given by -1. Now, consider the tangle $T' =$  . Then the cochain complex $Kh(T')$ is described as follows:



The differential at dimension 0 is given by $d^0 =$  $: Kh^0(T) \to Kh^1(T)$. Thus, we have a cochain complex of **k**-modules

$$0 \longrightarrow W \otimes V \xrightarrow{d} W \longrightarrow 0,$$

where $d(w \otimes v_+) = w$ and $d(w \otimes v_-) = 0$. The corresponding Khovanov homology is

$$H^p(T';\mathcal{G}) = \begin{cases} \mathbf{k}\{w \otimes v_-\}, & p = 0; \\ 0, & \text{otherwise.} \end{cases}$$

The quantum grading of $w \otimes v_-$ is -1. Now, consider the tangle T'' consisting of a single arc. It is clear that the Khovanov homology is

$$H^p(T'';\mathcal{G}) = \begin{cases} \mathbf{k}\{w\}, & p = 0; \\ 0, & \text{otherwise.} \end{cases}$$

The quantum grading of w here is also -1. In this example, T, T', and T'' are equivalent up to Reidemeister moves. Their corresponding Khovanov homology groups are also identical, with even the quantum gradings of the homology generators being equal.

3.4.3.4 Application

In Section 3.4.2.1, we defined persistent Khovanov homology of tangles within the category of tangles with fixed boundaries. However, in practical applications, it is uncommon to encounter the filtration of tangles with fixed boundaries. In this section, we present an application that describes how, with a given tangle in a metric space, one can construct persistent tangles within the category $\mathcal{P}la$, thereby obtaining the persistent Khovanov homology of tangles.

Let $(X, \leq)$ be a poset. Then $(X, \leq)$ can be regarded as a category, where the objects are the elements of X, and the morphisms are the pairs (x, x') such that $x \leq x'$ for $x, x' \in X$.

Definition 3.4.11. A *persistence tangle in category* $\mathcal{P}la$ is a functor $\mathcal{P} : (X, \leq) \to \mathcal{P}la$.

Definition 3.4.12. Let $\mathcal{P} : (X, \leq) \to \mathcal{P}la$ be a persistence tangle. The *persistent Khovanov homology of tangles* is the composition of functors

$$(X, \leq) \xrightarrow{\mathcal{P}} \mathcal{P}la \xrightarrow{H(-;\mathcal{G})} \mathcal{M}od_{\mathbf{k}}.$$

For any $a \leq b$ in X, the (a, b)-persistent Khovanov homology of tangles $\mathcal{P} : (X, \leq) \to \mathcal{P}la$ is given by

$$H^p_{a,b}(\mathcal{P}; \mathcal{G}) = \text{im}\left(H^p(\mathcal{P}(a); \mathcal{G}) \to H^p(\mathcal{P}(b); \mathcal{G})\right), \quad p \in \mathbb{Z}.$$

Example 3.4.13. Consider a tangle T in a Euclidean plane. Fix a point P as the center, and let D_ε denote a disk centered at P with radius ε. For each ε, define the tangle $T_\varepsilon = T \cap D_\varepsilon$, which may be empty. It is evident that the functor $\mathcal{P} : (\mathbb{R}, \leq) \to \mathcal{P}la$ defined by $\mathcal{P}(\varepsilon) = T_\varepsilon$ is a persistence tangle. For any real numbers $a \leq b$, we have the corresponding (a, b)-persistent Khovanov homology of tangles $H^*_{a,b}(\mathcal{P}; \mathcal{G})$.

Example 3.4.14. Let C be a finite collection of curves in 3-dimensional Euclidean space, and let $q : C \to \mathbb{R}^2$ be a projection such that there are finitely many crossings, each of which is required to be a double point. Let $\{D_\varepsilon\}_{\varepsilon \in \mathbb{R}}$ be a family of disks in $\mathbb{R}^2$ with the same center. Then the intersection $T_\varepsilon = q(C) \cap D_\varepsilon$ is a tangle (or the empty set) for any $\varepsilon > 0$. This defines a persistent

tangle $T_\varepsilon : (\mathbb{R}, \leq) \to \mathcal{P}la$, which can be used to compute the persistent Khovanov homology of tangles and extract topological features.

In practical applications, persistent tangles can be derived from one-dimensional manifolds embedded in three-dimensional space, or even from collections of non-smooth curves. By computing the persistent Khovanov homology of tangles, one can extract multi-scale topological features, which can then be used to analyze curve-type data. This highlights the significant potential of persistent Khovanov homology of tangles across various application domains in data science.

CHAPTER 4

THESIS CONTRIBUTION

The main contributions of this dissertation are listed as follows:

- In chapter 2.1, we introduce a new construction of N-chain complexes on simplicial complexes and develop the associated Mayer homology, persistent Mayer homology, and persistent Mayer Laplacians.
- In chapter 2.2, we perform the application of using Mayer homology to study protein-ligand binding affinities.
- In chapter 3.1, we review essential knot–theoretic foundations required for computational geometric topology in biology.
- In chapter 3.2, we propose the multiscale Gauss linking integral (mGLI) and illustrate its power for knot data analysis of biomolecules.
- In chapter 3.3, we study evolutionary Khovanov homology, providing a multiscale refinement that captures topological transitions of knots and links.
- In chapter 3.4, we develop persistent Khovanov homology of tangles, extending multiscale analysis of knot-type data beyond closed curves to open tangles.

The contents of this dissertation are mostly adopted from the following publications and preprints:

- Li Shen, Jian Liu, and Guo-Wei Wei. "Persistent Mayer Homology and Persistent Mayer Laplacian." *Foundations of Data Science* **6** (2024): 584–612. doi:10.3934/fods.2024032.
- Hongsong Feng, Li Shen, Jian Liu, and GuoWei Wei. "MayerHomology Learning Prediction of Protein–Ligand Binding Affinities." *Journal of Computational Biophysics and Chemistry* **24** (2) (2025): 253–266. doi:10.1142/S2737416524500613.
- Li Shen, Jian Liu, and Guo-Wei Wei. "Evolutionary Khovanov Homology." *AIMS Mathematics* **9** (9) (2024): 26139–26165. doi:10.3934/math.20241277.

- Li Shen, Hongsong Feng, Fengling Li, Fengchun Lei, Jie Wu, and Guo-Wei Wei. "Knot Data Analysis Using Multiscale Gauss Link Integral." *Proceedings of the National Academy of Sciences* (2024). doi:10.1073/pnas.2408431121.
- Jian Liu, Li Shen, and Guo-Wei Wei. "Persistent Khovanov Homology of Tangle." *arXiv preprint* (2024). Available at `https://arxiv.org/abs/2409.18312`.

CHAPTER 5

FUTURE WORK

Many future directions are available, including:

- Design and implement scalable algorithms—potentially leveraging parallelization, finite-field arithmetic —to accelerate Mayer homology and persistent Mayer Laplacian computations on large simplicial complexes.
- The Mayer framework extends the classical differential d to an N-differential with $d^N = 0$. Developing analogous N-operator extensions for other homology theories (e.g., Hochschild, quantum, or interaction homology) could open new algebraic and computational avenues.
- Apply the multiscale Gauss linking integral to problems beyond knot entanglement, such as protein mutation analysis, neuronal arbor geometry, and other biology domains involving highly segmented or filamentous structures.
- Generalize evolutionary Khovanov homology and persistent Khovanov homology to spatial graphs that admit singular vertices, enabling topological analysis of complex knot–type data with branching or junction points.
- Generalize evolutionary Khovanov homology and persistent Khovanov homology produce invariants indexed by quantum degrees; developing task-specific featurization or embedding strategies for these quantum-graded signatures will be crucial for downstream machine-learning applications.

BIBLIOGRAPHY

[1] G. Carlsson, G. Singh, and A. Zomorodian. Computing multidimensional persistence. In *Algorithms and Computation: 20th International Symposium, ISAAC 2009, Honolulu, Hawaii, USA, December 16–18, 2009. Proceedings 20*, pages 730–739. Springer, 2009.

[2] H. Edelsbrunner and J. Harer. Persistent homology–a survey. *Contemp. Math.*, 453:257–282, 2008.

[3] Z. Cang and G.-W. Wei. Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Computational Biology*, 13(7):e1005690, 2017.

[4] R. Wang, D. D. Nguyen, and G.-W. Wei. Persistent spectral graph. *International Journal for Numerical Methods in Biomedical Engineering*, 36(9):e3376, 2020.

[5] Duc Duy Nguyen, Zixuan Cang, Kedi Wu, Menglun Wang, Yin Cao, and Guo-Wei Wei. 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.

[6] W. Mayer. A new homology theory. *Annals of Mathematics*, pages 370–380, 1942.

[7] E. H. Spanier. The mayer homology theory. *Bulletin of the American Mathematical Society*, 55(2):102–112, 1949.

[8] M. Dubois-Violette. Generalized differential spaces with $d^n = 0$ and the q-differential calculus. *Czechoslovak Journal of Physics*, 46(12):1227–1233, 1996.

[9] J. Chen, R. Zhao, Y. Tong, and G.-W. Wei. Evolutionary de rham-hodge method. *Discrete and Continuous Dynamical Systems. Series B*, 26(7):3785, 2021.

[10] Li Shen, Hongsong Feng, Fengling Li, Fengchun Lei, Jie Wu, and Guo-Wei Wei. Knot data analysis using multiscale gauss link integral. *arXiv preprint arXiv:2311.12834*, 2023.

[11] Li Shen, Jian Liu, and Guo-Wei Wei. Evolutionary khovanov homology. *arXiv preprint arXiv:2406.02821*, 2024.

[12] M. Dubois-Violette. $d^n = 0$: Generalized homology. *K-theory*, 14(4):371–404, 1998.

[13] D. Chen, J. Liu, J. Wu, and G.-W. Wei. Persistent hyperdigraph homology and persistent hyperdigraph laplacians, 2023.

[14] J. Liu, J. Li, and J. Wu. The algebraic stability for persistent laplacians, 2023.

[15] P. Bubenik and J. A. Scott. Categorification of persistent homology. *Discrete &*

Computational Geometry, 51(3):600–627, 2014.

[16] U. Bauer and M. Lesnick. Persistence diagrams as diagrams: A categorification of the stability theorem. In *Topological Data Analysis: The Abel Symposium 2018*, pages 67–96. Springer, 2020.

[17] U. Bauer and M. Lesnick. Induced matchings and the algebraic stability of persistence barcodes, 2013.

[18] J. Chen, Y. Qiu, R. Wang, and G.-W. Wei. Persistent laplacian projected omicron ba.4 and ba.5 to become new dominating variants. *Computers in Biology and Medicine*, 151:106262, 2022.

[19] Zixuan Cang and Guo-Wei Wei. 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):e2914, 2018.

[20] Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, 2015.

[21] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[22] Duc Duy Nguyen and Guo-Wei Wei. Agl-score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling*, 59(7):3291–3304, 2019.

[23] Zhenyu Meng and Kelin Xia. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science advances*, 7(19):eabc5329, 2021.

[24] Kelin Xia, Kristopher Opron, and Guo-Wei Wei. Multiscale multiphysics and multidomain models—flexibility and rigidity. *The Journal of chemical physics*, 139(19):11B614_1, 2013.

[25] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

[26] Md Masud Rana and Duc Duy Nguyen. Geometric graph learning with extended atom-types features for protein-ligand binding affinity prediction. *Computers in Biology and Medicine*, 164:107250, 2023.

[27] Md Masud Rana and Duc Duy Nguyen. Eisa-score: Element interactive surface area score for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*,

62(18):4329–4341, 2022.

[28] Xiang Liu, Huitao Feng, Jie Wu, and Kelin Xia. Dowker complex based machine learning (dcml) models for protein-ligand binding affinity prediction. *PLoS computational biology*, 18(4):e1009943, 2022.

[29] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[30] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

[31] Ran Liu, Xiang Liu, and Jie Wu. Persistent path-spectral (pps) based machine learning for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 63(3):1066–1075, 2023.

[32] Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling*, 49(4):1079–1093, 2009.

[33] Yan Li, Zhihai Liu, Jie Li, Li Han, Jie Liu, Zhixiong Zhao, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *Journal of chemical information and modeling*, 54(6):1700–1716, 2014.

[34] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

[35] C. C. Adams. *The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots*. American Mathematical Society, 1994.

[36] G. Burde and H. Zieschang. *Knots*. De Gruyter, 2002.

[37] J. W. Alexander and G. B. Briggs. On types of knotted curves. *Ann. Math.*, 28:562–586, 1926.

[38] K. Reidemeister. Elementare begründung der knotentheorie. *Abh. Math. Semin. Univ. Hambg.*, 5:24–32, 1927.

[39] L. H. Kauffman. State models and the jones polynomial. *Topology*, 26:395–407, 1987.

[40] H. Schubert. Über eine numerische knoteninvariante. *Math. Z.*, 61:245–288, 1954.

[41] V. F. R. Jones. A polynomial invariant for knots via von neumann algebras. In *Fields Medallists' Lectures*, pages 448–458. World Scientific, Singapore, 1997.

[42] A. Gibson. Homotopy invariants of gauss words. *Math. Ann.*, 349:871–887, 2011.

[43] V. O. Manturov. *Knot Theory*. CRC Press, Boca Raton, 2 edition, 2018.

[44] Mikhail Khovanov. A categorification of the Jones polynomial. *Duke Mathematical Journal*, 101(3):359 – 426, 2000.

[45] Dror Bar-Natan. On khovanov's categorification of the Jones polynomial. *Algebraic & Geometric Topology*, 2(1):337–370, 2002.

[46] P. B. Kronheimer and T. S. Mrowka. Khovanov homology is an unknot-detector. *Publ. Math. IHES*, 113:97–208, 2011.

[47] Richard H Crowell and Ralph Hartzler Fox. *Introduction to knot theory*, volume 57. Springer Science & Business Media, 2012.

[48] Ciprian Manolescu. An introduction to knot floer homology. *Physics and mathematics of link homology*, 680:99–135, 2014.

[49] Tomotada Ohtsuki. *Quantum invariants: A study of knots, 3-manifolds, and their sets*, volume 29. World Scientific, 2002.

[50] Chengzhi Liang and Kurt Mislow. Knots in proteins. *Journal of the American Chemical Society*, 116(24):11189–11190, 1994.

[51] DW Sumners. The role of knot theory in dna research. In *Geometry and Topology*, pages 297–318. CRC Press, 2020.

[52] Tamar Schlick, Qiyao Zhu, Abhishek Dey, Swati Jain, Shuting Yan, and Alain Laederach. To knot or not to knot: multiple conformations of the sars-cov-2 frameshifting rna element. *Journal of the American Chemical Society*, 143(30):11404–11422, 2021.

[53] Kenneth C Millett, Eric J Rawdon, Andrzej Stasiak, and Joanna I Sułkowska. Identifying knots in proteins. *Biochemical Society Transactions*, 41(2):533–537, 2013.

[54] M. Jamroz, W. Niemyska, E. J. Rawdon, A. Stasiak, K. C. Millett, and P. et al. Sułkowski. Knotprot: A database of proteins with knots and slipknots. *Nucleic Acids Res.*, 43:D306–D314, 2015.

[55] Pawel Dabrowski-Tumanski, Pawel Rubach, Wanda Niemyska, Bartosz Ambrozy Gren, and Joanna Ida Sulkowska. Topoly: Python package to analyze topology of polymers. *Briefings in Bioinformatics*, 22(3):bbaa196, 2021.

[56] Eleni Panagiotou and Louis H Kauffman. Knot polynomials of open and closed curves. *Proceedings of the Royal Society A*, 476(2240):20200124, 2020.

[57] Quenisha Baldwin, Bobby Sumpter, and Eleni Panagiotou. The local topological free energy of the sars-cov-2 spike protein. *Polymers*, 14(15):3014, 2022.

[58] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 347–356, 2004.

[59] Rui Wang, Jiahui Chen, and Guo-Wei Wei. 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.

[60] Jiahui Chen and Guo-Wei Wei. 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.

[61] Carl Friedrich Gauss. Integral formula for linking number. *In Zur mathematischen theorie der electrodynamische wirkungen*, 5:605, 1833.

[62] John M Cornwall and Noah Graham. Sphalerons, knots, and dynamical compactification in yang-mills-chern-simons theories. *Physical Review D*, 66(6):065012, 2002.

[63] Mitchell A Berger. Third-order link integrals. *Journal of Physics A: Mathematical and General*, 23(13):2787, 1990.

[64] Dong Chen, Kaifu Gao, Duc Duy Nguyen, Xin Chen, Yi Jiang, Guo-Wei Wei, and Feng Pan. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nature communications*, 12(1):3521, 2021.

[65] Renzo L Ricca and Bernardo Nipoti. Gauss'linking number revisited. *Journal of Knot Theory and Its Ramifications*, 20(10):1325–1343, 2011.

[66] AJ Rader, Chakra Chennubhotla, Lee-Wei Yang, and Ivet Bahar. The gaussian network model: Theory and applications. In *Normal mode analysis*, pages 65–88. Chapman and Hall/CRC, 2005.

[67] Eran Eyal, Lee-Wei Yang, and Ivet Bahar. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, 22(21):2619–2627, 2006.

[68] Ivet Bahar and AJ Rader. Coarse-grained normal mode analysis in structural biology. *Current opinion in structural biology*, 15(5):586–592, 2005.

[69] Jun-Koo Park, Robert Jernigan, and Zhijun Wu. Coarse grained normal mode analysis vs. refined gaussian network model for protein residue-level structural fluctuations. *Bulletin of*

mathematical biology, 75:124–160, 2013.

[70] Kristopher Opron, Kelin Xia, and Guo-Wei Wei. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *The Journal of chemical physics*, 140(23):06B617_1, 2014.

[71] David Bramer and Guo-Wei Wei. Atom-specific persistent homology and its application to protein flexibility analysis. *Computational and mathematical biophysics*, 8(1):1–35, 2020.

[72] Zixuan Cang, Elizabeth Munch, and Guo-Wei Wei. Evolutionary homology on coupled dynamical systems with applications to protein flexibility analysis. *Journal of applied and computational topology*, 4:481–507, 2020.

[73] Heng Cai, Chao Shen, Tianye Jian, Xujun Zhang, Tong Chen, Xiaoqi Han, Zhuo Yang, Wei Dang, Chang-Yu Hsieh, Yu Kang, et al. Carsidock: a deep learning paradigm for accurate protein–ligand docking and screening based on large-scale pre-training. *Chemical Science*, 15(4):1449–1471, 2024.

[74] Qurrat Ul Ain, Antoniya Aleksandrova, Florian D Roessler, and Pedro J Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(6):405–424, 2015.

[75] Xiaolin Pan, Hao Wang, Yueqing Zhang, Xingyu Wang, Cuiyu Li, Changge Ji, and John ZH Zhang. Aa-score: a new scoring function based on amino acid-specific interaction for molecular docking. *Journal of Chemical Information and Modeling*, 62(10):2499–2509, 2022.

[76] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

[77] Zixuan Cang, Lin Mu, and Guo-Wei Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology*, 14(1):e1005929, 2018.

[78] Hongsong Feng and Guo-Wei Wei. Virtual screening of drugbank database for herg blockers using topological laplacian-assisted ai models. *Computers in biology and medicine*, 153:106491, 2023.

[79] Chen Zhang, Yuan Zhou, Shikai Gu, Zengrui Wu, Wenjie Wu, Changming Liu, Kaidong Wang, Guixia Liu, Weihua Li, Philip W Lee, et al. In silico prediction of herg potassium channel blockage by chemical category approaches. *Toxicology research*, 5(2):570–582, 2016.

[80] Xudong Zhang, Jun Mao, Min Wei, Yifei Qi, and John ZH Zhang. Hergspred: Accurate classification of herg blockers/nonblockers with machine-learning models. *Journal of Chemical Information and Modeling*, 62(8):1830–1839, 2022.

[81] Xiao Li, Yuan Zhang, Huanhuan Li, and Yong Zhao. Modeling of the herg k+ channel blockage using online chemical database and modeling environment (ochem). *Molecular Informatics*, 36(12):1700074, 2017.

[82] Chuipu Cai, Pengfei Guo, Yadi Zhou, Jingwei Zhou, Qi Wang, Fengxue Zhang, Jiansong Fang, and Feixiong Cheng. Deep learning-based prediction of drug-induced cardiotoxicity. *Journal of chemical information and modeling*, 59(3):1073–1084, 2019.

[83] Dong Chen, Jiaxin Zheng, Guo-Wei Wei, and Feng Pan. Extracting predictive representations from hundreds of millions of molecules. *The journal of physical chemistry letters*, 12(44):10793–10801, 2021.

[84] Kedi Wu and Guo-Wei Wei. Quantitative toxicity prediction using topology based multitask deep neural networks. *Journal of chemical information and modeling*, 58(2):520–531, 2018.

[85] Kaifu Gao, Duc Duy Nguyen, Vishnu Sresht, Alan M Mathiowetz, Meihua Tu, and Guo-Wei Wei. Are 2d fingerprints still valuable for drug discovery? *Physical chemistry chemical physics*, 22(16):8373–8390, 2020.

[86] T Martin. User's guide for test (version 4.2)(toxicity estimation software tool) a program to estimate toxicity from molecular structure. us epa office of research and development, washington, dc. Technical report, EPA/600/R-16/058, 2016.

[87] Neslihan Gügümcü and Louis H Kauffman. New invariants of knotoids. *European Journal of Combinatorics*, 65:186–229, 2017.

[88] Eleni Panagiotou and Louis H Kauffman. Vassiliev measures of complexity of open and closed curves in 3-space. *Proceedings of the Royal Society A*, 477(2254):20210440, 2021.

[89] Wanda Niemyska, Pawel Dabrowski-Tumanski, Michal Kadlof, Ellinor Haglund, Piotr Sułkowski, and Joanna I Sulkowska. Complex lasso: new entangled motifs in proteins. *Scientific reports*, 6(1):36895, 2016.

[90] Pawel Dabrowski-Tumanski, Pawel Rubach, Dimos Goundaroulis, Julien Dorier, Piotr Sułkowski, Kenneth C Millett, Eric J Rawdon, Andrzej Stasiak, and Joanna I Sulkowska. Knotprot 2.0: a database of proteins with knots and other entangled structures. *Nucleic acids research*, 47(D1):D367–D375, 2019.

[91] Kristopher Opron, Kelin Xia, and Guo-Wei Wei. Communication: Capturing protein multiscale thermal fluctuations. *The Journal of chemical physics*, 142(21):06B401_1, 2015.

[92] W. Bi, J. Li, J. Liu, and J. Wu. On the cayley-persistence algebra, 2022.

[93] Tu Quoc Thang Le and Jun Murakami. Representation of the category of tangles by kontsevich's iterated integral. *Communications in mathematical physics*, 168:535–562, 1995.

[94] Dror Bar-Natan. Khovanov's homology for tangles and cobordisms. *Geometry & Topology*, 9(3):1443–1499, 2005.

[95] Mikhail Khovanov. A functor-valued invariant of tangles. *Algebraic & Geometric Topology*, 2(2):665–741, 2002.

[96] John C Baez and Laurel Langford. Higher-dimensional algebra iv: 2-tangles. *Advances in Mathematics*, 180(2):705–764, 2003.

[97] John E Fischer Jr. 2-categories and 2-knots. *Duke Math. J.*, 76(1):493–526, 1994.

[98] Laurel Tamara Fearnley Langford. *2-tangles as a free braided monoidal 2-category with duals*. University of California, Riverside, 1997.

[99] J Scott Carter and Masahico Saito. Reidemeister moves for surface isotopies and their interpretation as moves to movies. *Journal of Knot Theory and its Ramifications*, 2(03):251–284, 1993.

[100] Dennis Roseman. Reidemeister-type moves for surfaces in four-dimensional space. *Banach Center Publications*, 42(1):347–380, 1998.

[101] Charles A Weibel. *An introduction to homological algebra*. Number 38. Cambridge university press, 1994.

[102] Vaughan Jones. Planar algebras. *New Zealand Journal of Mathematics*, 52:1–107, 2021.

[103] V. Abramov. On a graded q-differential algebra. *Journal of Nonlinear Mathematical Physics*, 13(sup1):1–8, 2006.

[104] S. Bressan, J. Li, S. Ren, and J. Wu. The embedded homology of hypergraphs and applications, 2016.

[105] G. Carlsson and V. De Silva. Zigzag persistence. *Foundations of Computational Mathematics*, 10:367–405, 2010.

[106] G. Carlsson and A. Zomorodian. The theory of multidimensional persistence. In *Proceedings of the twenty-third annual symposium on Computational geometry*, pages 184–193, 2007.

[107] C. Kassel and M. Wambst. Algébre homologique des n-complexes et homologie de

hochschild aux racines de l'unité. *Publications of the Research Institute for Mathematical Sciences*, 34(2):91–114, 1998.

[108] X. Liu, H. Feng, J. Wu, and K. Xia. Persistent spectral hypergraph based machine learning (psh-ml) for protein-ligand binding affinity prediction. *Briefings in Bioinformatics*, 22(5):bbab127, 2021.

[109] B. Lu and Z. Di. Gorenstein cohomology of n-complexes. *Journal of Algebra and Its Applications*, 19(09):2050174, 2020.

[110] B. Lu, Z. Di, and Y. Liu. Cartan-eilenberg n-complexes with respect to self-orthogonal subcategories. *Frontiers of Mathematics in China*, 15:351–365, 2020.

[111] A. Sitarz. On the tensor product construction for q-differential algebras. *Letters in Mathematical Physics*, 44(1):17–21, 1998.

[112] R. Wang and G.-W. Wei. Persistent path laplacian. *Foundations of Data Science (Springfield, Mo.)*, 5(1):26, 2023.

[113] X. Wei and G.-W. Wei. Persistent sheaf laplacians, 2021.

[114] Louis H Kauffman. An introduction to Khovanov homology. In *Knot theory and its applications*, pages 105–139, 2016.

[115] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[116] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & computational geometry*, 28:511–533, 2002.

[117] M. F. Atiyah. *The Geometry and Physics of Knots*. Cambridge University Press, 1990.

[118] O. Lukin and F. Vögtle. Knotting and threading of molecules: chemistry and chirality of molecular knots and their assemblies. *Angew. Chem. Int. Edit.*, 44:1456–1477, 2005.

[119] K. Murasugi. *Knot Theory and Its Applications*. Birkhauser, 1996.

[120] D. Endy. Foundations for engineering biology. *Nature*, 438:449–453, 2005.

[121] Y. Pommier, E. Leo, H. L. Zhang, and C. Marchand. Dna topoisomerases and their poisoning by anticancer and antibacterial drugs. *Chem. Biol.*, 17:421–433, 2010.

[122] D. Goundaroulis, N. Gügümcü, S. Lambropoulou, J. Dorier, A. Stasiak, and L. Kauffman. Topological models for open-knotted protein chains using the concepts of knotoids and bonded knotoids. *Polymers*, 9:444, 2017.

[123] N. C. H. Lim and S. E. Jackson. Molecular knots in biology and chemistry. *J. Phys.: Condens. Matter*, 27:354101, 2015.

[124] P. Dabrowski-Tumanski and J. I. Sulkowska. Topological knots and links in proteins. *P. Natl. A. Sci.*, 114:3415–3420, 2017.

[125] X. Q. Wei and G.-W. Wei. Persistent topological laplacians–a survey, 2023.

[126] J. Liu, D. Chen, and G.-W. Wei. Persistent interaction topology in data analysis, 2024.

[127] Greg Kuperberg. From the mahler conjecture to gauss linking integrals. *Geometric And Functional Analysis*, 18(3):870–892, 2008.

[128] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30(8):814–844, 2014.

[129] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[130] Danijela Horak and Jürgen Jost. Spectra of combinatorial laplace operators on simplicial complexes. *Advances in Mathematics*, 244:303–336, 2013.

[131] Mark Anthony Armstrong. *Basic topology*. Springer Science & Business Media, 2013.

[132] Yiwei Wang, Lei Huang, Siwen Jiang, Yifei Wang, Jun Zou, Hongguang Fu, and Shengyong Yang. Capsule networks showed excellent performance in the classification of herg blockers/nonblockers. *Frontiers in pharmacology*, 10:1631, 2020.

[133] Munikumar R Doddareddy, Elisabeth C Klaasse, Adriaan P IJzerman, and Andreas Bender. Prospective validation of a comprehensive in silico herg model and its applications to commercial compound and drug databases. *ChemMedChem*, 5(5):716–729, 2010.

[134] Kevin S Akers, Glendon D Sinks, and T Wayne Schultz. Structure–toxicity relationships for selected halogenated aliphatic chemicals. *Environmental toxicology and pharmacology*, 7(1):33–39, 1999.

[135] Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, and Igor V Tetko. Combinatorial qsar modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of chemical information and modeling*, 48(4):766–784, 2008.

[136] Li Shen, Hongsong Feng, Yuchi Qiu, and Guo-Wei Wei. Svsbi: sequence-based virtual screening of biomolecular interactions. *Communications Biology*, 6(1):536, 2023.

[137] Hongsong Feng, Jian Jiang, and Guo-Wei Wei. Machine-learning repurposing of drugbank

compounds for opioid use disorder. *Computers in biology and medicine*, 160:106921, 2023.

[138] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[139] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[140] Herbert Edelsbrunner and John L Harer. *Computational topology: an introduction*. American Mathematical Society, 2022.

[141] Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14):15956–15965, 2019.

[142] Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. Potentialnet for molecular property prediction. *ACS central science*, 4(11):1520–1530, 2018.

[143] Jonas Dittrich, Denis Schmidt, Christopher Pfleger, and Holger Gohlke. Converging a knowledge-based scoring function: Drugscore2018. *Journal of chemical information and modeling*, 59(1):509–521, 2018.

[144] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

[145] Hongjian Li, Gang Lu, Kam-Heung Sze, Xianwei Su, Wai-Yee Chan, and Kwong-Sak Leung. Machine-learning scoring functions trained on complexes dissimilar to the test set already outperform classical counterparts on a blind benchmark. *Briefings in bioinformatics*, 22(6):bbab225, 2021.

[146] Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[147] Timothy Szocinski, Duc Duy Nguyen, and Guo-Wei Wei. Awegnn: Auto-parametrized weighted element-specific graph neural networks for molecules. *Computers in biology and medicine*, 134:104460, 2021.

[148] Edison Mucllari, Vasily Zadorozhnyy, Qiang Ye, and Duc Duy Nguyen. Novel molecular representations using neumann-cayley orthogonal gated recurrent unit. *Journal of Chemical Information and Modeling*, 63(9):2656–2666, 2023.

[149] Jian Jiang, Rui Wang, Menglun Wang, Kaifu Gao, Duc Duy Nguyen, and Guo-Wei Wei. Boosting tree-assisted multitask deep learning for small scientific datasets. *Journal of chemical information and modeling*, 60(3):1235–1244, 2020.

[150] S Jannicke Moe, Anders L Madsen, Kristin A Connors, Jane M Rawlings, Scott E Belanger, Wayne G Landis, Raoul Wolf, and Adam D Lillicrap. Development of a hybrid bayesian network model for predicting acute fish toxicity using multiple lines of evidence. *Environmental modelling & software*, 126:104655, 2020.