Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

TOPOLOGY MEETS BIOLOGY: PERSISTENT HOMOLOGY IN LASSO PROTEINS DETECTION

Tesi di Laurea in Topologia Computazionale

Relatore:
Prof.ssa Alessia Cattabriga
Correlatore:
Dott. Paolo Cavicchioli

Presentata da:
Beatrice Lucia Iasi

Anno Accademico 2023-2024

"Puoi cambiare camicia se ne hai voglia
e se hai fiducia puoi cambiare scarpe
con scarpe nuove puoi cambiare strada
e cambiando strada puoi cambiare idee
e con le idee si cambia il mondo
ma il mondo non cambia spesso
allora la tua vera rivoluzione
sarà cambiare te stesso"
A. Mannarino, *Vivere la vita.*

Introduction

Over the past few decades, persistent homology (PH) has emerged as a crucial area of modern algebraic topology [6, 23, 24], especially within the field of topological data analysis (TDA). TDA integrates mathematics with data analysis, algebraic topology, computational geometry, computer science, statistics, and related fields [19]. The foundational concept of persistent homology was introduced by Frosini and Landi [7] and later developed in its general form by Robins [21], Edelsbrunner [5], and Zomorodian and Carlsson [25]. Persistent homology builds upon traditional homology by examining how topological features evolve across different scales, which enhances its ability to characterize complex geometric structures and uncover patterns in data that may be obscured at a single scale. Its applications are broad, including image analysis [2], structural and computational biology [9, 24], and complex networks [10, 20].

The appeal of this method lies in the fact that it is based on algebraic topology, which offers a well-established theoretical framework for analyzing the qualitative features of complex data. It is computable using linear algebra and is both stable and robust against small perturbations in input data. Moreover, compared to ordinary homology, PH introduces an additional dimension—the filtration parameter—which allows embedding crucial geometric or quantitative information into topological invariants.

This work aims to introduce persistent homology and explore its application in identifying lasso proteins—a unique class of proteins distinguished by a structural motif where a segment of the protein chain forms a loop, which is pierced by another piece of the same chain. This structure provides significant stability and resistance to degradation, making lasso proteins a field of interest for drug design and molecular engineering, including peptide-based therapies and bioengineering. We will specifically reference the algorithm presented in [9], which is still an ongoing work.

Chapter 1 will cover preliminary concepts of simplices and simplicial complexes, essential for defining (simplicial) homology. Since homology computation requires a simplicial complex as input, and the objects of interest are often not provided in this form, we will

conclude the chapter with methods for constructing simplicial complexes, such as Čech and Vietoris-Rips complexes. These constructions enable the association of a simplicial complex with a point cloud—a finite collection of points in a metric space representing a sample of our shape.

Chapter 2 will explore persistent homology, beginning with the concept of filtration. We will present the formal definition of PH and discuss the two primary visualization methods: barcodes and persistence diagrams. Additionally, we will address the robustness and stability of the method, defining the bottleneck distance metric for comparing persistence diagrams.

Finally, Chapter 3 will focus on applying persistent homology to identify lasso proteins. We will provide a brief overview of the biological context of proteins and lasso structures before presenting the theoretical framework that underlies the algorithm used to detect these structures. A simple 2-dimensional example will illustrate the application of this framework to a small set of points, followed by a brief description of the algorithm pipeline as cited in [9].

Contents

Introduction **ii**

1 Homology groups and simplicial complexes **1**

1.1 Simplicial complexes 1

1.2 Homology groups 7

1.3 Constructions of simplicial complexes 14

2 Persistent Homology **19**

2.1 Filtrations 19

2.2 Definition and Visualization 22

2.3 Stability 26

3 Identifying Lasso Proteins with Persistent Homology **29**

3.1 Proteins: Structure Fundamentals 29

3.2 Protein topology: the lasso proteins 33

3.3 Detecting lasso proteins 35

List of Figures

1.1 Fundamental simplices; *from* [6], *p. 62.* 2
1.2 A labeled geometric complex; *by author* 5
1.3 The torus; *from* [23], *p. 40, adapted.* 6
1.4 The chain complex, *from* [6], *p. 96.* 9
1.5 A geometric representation of K, coordinates free; *by author.* 10
1.6 Example of geometric realization of a nerve; *from* [23], *p. 63, adapted.* . . 15
1.7 Two examples of Čech complexes; *from* [23], *p. 63.* 16

2.1 A filtration of simplicial complexes and its correspondent 0-dimensional barcode; *from* [23], *p. 122.* 20
2.2 Example of persistent homology for a point cloud; *from* [19], *p. 3.* 21
2.3 A filtration and their corresponding zero-dimensional barcode and persistence diagram; *from* [23], *p. 127.* 25
2.4 0-dimensional persistence diagram; *from* [23], *p. 127.* 25
2.5 Matching diagrams and bottleneck distances; *from* [23], *p. 146, adapted.* 28

3.1 Peptide bond; from `https://en.wikipedia.org/wiki/Peptide_bond`. . 30
3.2 Fundamental levels of organization of the proteins; *from* [1], *Chapter 2 of the e-book version, adapted.* 31
3.3 The secondary structures; *from* [1], *Chapter 2 of the e-book version*, adapted. 32
3.4 Various types of complex lasso motifs; *from* [16], *p. 3.* 34
3.5 An example of supercoiling lasso complex in a real protein; *from* [18], *p. 2.* 34
3.6 Rips filtration for the decagon without (i) and with (ii) the center; *by author.* 39
3.7 Barcodes of persistent homology, *by author.* 40
3.8 Persistence diagrams, *by author.* 40
3.9 Bottleneck distance; *by author.* 41
3.10 Lasso complex detection in glutamate-like receptor GLR3.2 ligand-binding domain, *from* `https://lassoprot.cent.uw.edu.pl`. 42

Chapter 1

Homology groups and simplicial complexes

In this chapter, we begin by introducing simplicial complexes. This allows us to define (simplicial) homology[1], whose treatment is essential for discussing persistent homology. Finally, we explain how to construct simplicial complexes. Unless otherwise noted, references are made to [4, 6, 8, 23], particularly in Sections 1.1 and 1.3, while for more details on homology theory, see [11, 14].

1.1 Simplicial complexes

Simplicial complexes are intuitively sets composed of points, line segments, triangles, and their higher-dimensional counterparts. They provide a convenient combinatorial description of *certain* metric spaces[2]. Thus, it is common practice to replace the original spaces with simplicial complexes for concrete computations. Before presenting their formal definition, we need to clarify some general concepts.

Definition 1.1. Let $u_0, \ldots, u_k \in \mathbb{R}^d$ and $\lambda_0, \ldots, \lambda_k \in \mathbb{R}$, with $\sum_{i=0}^{k} \lambda_i = 1$. The point $p = \sum_{i=0}^{k} \lambda_i u_i$ is called an *affine combination* of $u_0, \ldots, u_k$, with coefficients $\lambda_0, \ldots, \lambda_k \in \mathbb{R}$. The set of all affine combinations of the points $u_0, \ldots, u_k \in \mathbb{R}^d$ is called the *affine hull* of $u_0, \ldots, u_k$.

Definition 1.2. The points $u_0, \ldots, u_k \in \mathbb{R}^d$ are called *affinely independent* if, given coefficients $\lambda_0, \ldots, \lambda_k$ and $\mu_0, \ldots, \mu_k$ in $\mathbb{R}$ such that $\sum_i \lambda_i = \sum_i \mu_i = 1$, the equality $\sum_i \lambda_i u_i = \sum_i \mu_i u_i$ holds *iff* $\lambda_i = \mu_i$ for all i.

[1] From now on, simplicial homology will simply be referred to as homology.

[2] This is not true for arbitrary metric spaces, but for "nice" ones. See [11] for further details.

One can show that $u_0, \ldots, u_k \in \mathbb{R}^d$ are affinely independent *iff* the vectors $\{u_1 - u_0, \ldots, u_k - u_0\}$ are linearly independent.

Definition 1.3. Each affine combination $x = \sum_{i=0}^{k} \lambda_i u_i$ with non-negative coefficients is called a *convex combination* of $u_0, \ldots, u_k$. Given $U \subseteq \mathbb{R}^d$, the set of all convex combinations of point of U defined as $\mathrm{Conv}(U)$ is called the *convex hull* of U.

If $U = \{u_0, \ldots, u_k\}$, we will often write $\langle u_0, \ldots, u_k \rangle$ in place of $\mathrm{Conv}(U)$. One can prove that $\langle u_0, \ldots, u_k \rangle$ is the smallest convex set containing $u_0, \ldots, u_k$.

Definition 1.4. Let $k, d \in \mathbb{N}$ with $k \leq d$. A *geometric k-simplex* σ in $\mathbb{R}^d$ is the convex hull of an affinely independent family $U = \{u_0, \ldots, u_k\} \subset \mathbb{R}^d$, *i.e.*, $\sigma = \langle u_0, \ldots, u_k \rangle$. The number k is called the dimension of the geometric simplex σ.

We use special names for the smallest simplices: *vertex* for 0-simplex, *edge* for 1-simplex, *triangle* for 2-simplex and *tetrahedron* for 3-simplex as in Figure 1.1. We also say that the empty set is the unique (-1)-simplex.

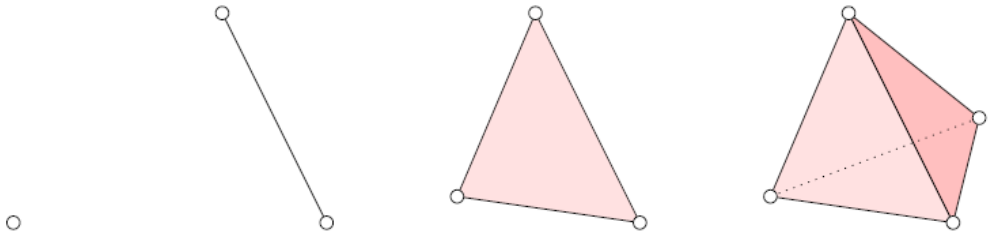


Figure 1.1: From left to right: a vertex, an edge, a triangle, and a tetrahedron.

Definition 1.5. If $\sigma = \langle u_0, \ldots, u_k \rangle$ is a simplex, the convex hull of any subset of $\{u_0, \ldots, u_k\}$ is called a *face* of σ. If τ is a face of σ, we write $\tau \leq \sigma$ (or $\sigma \geq \tau$). If $\tau \leq \sigma$ and $\tau \neq \sigma$, we say that τ is a *proper face* of σ and write $\tau < \sigma$ (or $\sigma > \tau$). If τ is a (proper) face of σ, we say that σ is a (proper) *coface* of τ. We also define the *boundary* $\mathrm{bd}\,\sigma$ of a simplex σ as the union of all proper faces of σ. The *interior* $\mathrm{int}\,\sigma$ of σ is the set $\sigma \setminus \mathrm{bd}\,\sigma$.

Remark 1.6. Let $\sigma = \langle u_0, \ldots, u_k \rangle$. By considering the empty set as a subset of $\{u_0, \ldots, u_k\}$, the (-1)-simplex is a face of σ. Thus, the vertices u_i, with $i = 0$ have empty boundary, *i.e.*, $\mathrm{int}\, u_i = u_i$, for $i = 0, \ldots, k$.

Geometric simplicial complexes

Now, we are ready to give the actual definition of a (geometric) simplicial complex in $\mathbb{R}^d$.

Definition 1.7. A (finite) *geometric simplicial complex* $K \neq \{\emptyset\}$ is a set of simplices such that

1. if $\sigma \in K$ and $\tau \leq \sigma$, then $\tau \in K$;
2. if $\sigma_1, \sigma_2 \in K$, then $\sigma_1 \cap \sigma_2 \leq \sigma_1$ and $\sigma_1 \cap \sigma_2 \leq \sigma_2$.

The *dimension* of a geometric simplicial complex is the maximum dimension of its simplices. The *body* $|K|$ of a geometric simplicial complex K is the union of all simplices in K, endowed with the topology induced by the Euclidean topology in $\mathbb{R}^d$. When K is finite, the body $|K|$ is compact, as a finite union of compact subspaces σ. A *polyhedron* is the body of a geometric simplicial complex.

Formally speaking, a geometric simplicial complex K in $\mathbb{R}^d$ is a collection of simplices, while the related body $|K|$ is a subset of $\mathbb{R}^d$. However, in geometric discussions, we will often identify these two objects. Indeed, from this point forward, we will visualize simplicial complexes by illustrating their body while implicitly assuming the underlying simplicial structure.

We are now ready to describe the relationship between a metric subspace of $\mathbb{R}^d$ and its combinatorial representation.

Definition 1.8. A *triangulation* of a topological space X is a homeomorphism from X to a polyhedron. If a topological space X admits a triangulation, we say that X is *triangulable*.

Example 1.9. Let K be a geometric simplicial complex. The identity map $\mathrm{id} : |K| \to |K|$ is a triangulation of the body of K.

We conclude this subsection with a useful definition.

Definition 1.10. Let K be a geometric simplicial complex. Every geometric simplicial complex L such that $L \subseteq K$ is called a *subcomplexes* of K. The subcomplex $K^{(j)} := \{\sigma \in K : \dim \sigma \leq j\}$ is called the *j-skeleton* of K. $K^{(0)}$ is also referred as *vertex set* of K and denoted by the symbol $\mathrm{Vert}\, K$.

Abstract simplicial complex

There is another way to describe simplicial complex. Rather than listing all its simplices by providing the coordinates of their vertices, a more efficient approach is to represent them abstractly. This method allows for a coordinate-free description, simplifying the process of defining how the simplices fit together within the complex. For this reason, we introduce abstract simplicial complexes.

Definition 1.11. A (finite) *abstract simplicial complex* is a finite family $\mathcal{A} \neq \{\emptyset\}$ of sets such that if $\alpha \in \mathcal{A}$ and $\beta \subseteq \alpha$, then $\beta \in \mathcal{A}$. Each element of an abstract simplicial complex is called an *abstract simplex* (and of course is finite).

While geometric simplicial complexes capture the geometric aspects of a space such as sizes, lengths, and shapes, in contrast, abstract simplicial complexes focus on topological properties, representing the structure of the space up to homeomorphism.

The *dimension* of an abstract simplex α in $\mathcal{A}$ is $|\alpha| - 1$, while the dimension of an abstract simplicial complex $\mathcal{A}$ is the maximum dimension of any simplex in the complex. Any (proper) subset of $\alpha \in \mathcal{A}$ is called a (proper) *face* of α. The union of the abstract simplices of $\mathcal{A}$ is called the *vertex set* of $\mathcal{A}$, and denoted by the symbol $\operatorname{Vert} \mathcal{A}$.

Every abstract simplicial complex $\mathcal{B}$ such that $\mathcal{B} \subseteq \mathcal{A}$ is called a *subcomplex* of $\mathcal{A}$.

Remark 1.12. The empty set is always included as an abstract simplex of dimension -1.

Definition 1.13. Two abstract simplicial complexes $\mathcal{A}, \mathcal{B}$ are *isomorphic* if there exists a bijection $\phi : \operatorname{Vert} \mathcal{A} \to \operatorname{Vert} \mathcal{B}$ such that $\alpha \in \mathcal{A}$ implies that $\phi(\alpha) = \{\phi(a) : a \in \alpha\} \in \mathcal{B}$ and $\beta \in \mathcal{B}$ implies that $\phi^{-1}(\beta) = \{\phi^{-1}(b) : b \in \beta\} \in \mathcal{A}$. The map $\bar{\phi} : \mathcal{A} \to \mathcal{B}$ induced by the map ϕ is called an *isomorphism* between the abstract simplicial complexes $\mathcal{A}, \mathcal{B}$.

Example 1.14. Let K be a geometric simplicial complex as in Figure 1.2. As a geometric simplicial complex, K contains specific geometric simplices described by the coordinates of their vertices. We can construct a corresponding abstract simplicial complex L. Labelling the vertices as in figure, then:

$$L = \{\{a, c, d\}, \{a, b\}, \{b, c\}, \{c, d\}, \{d, a\}, \{a, c\}, \{a\}, \{b\}, \{c\}, \{d\}\}.$$

In this description, no coordinates are involved.

There is a connection between abstract simplicial complexes and geometric ones, which implies that for any given geometric simplicial complex, an abstract simplicial

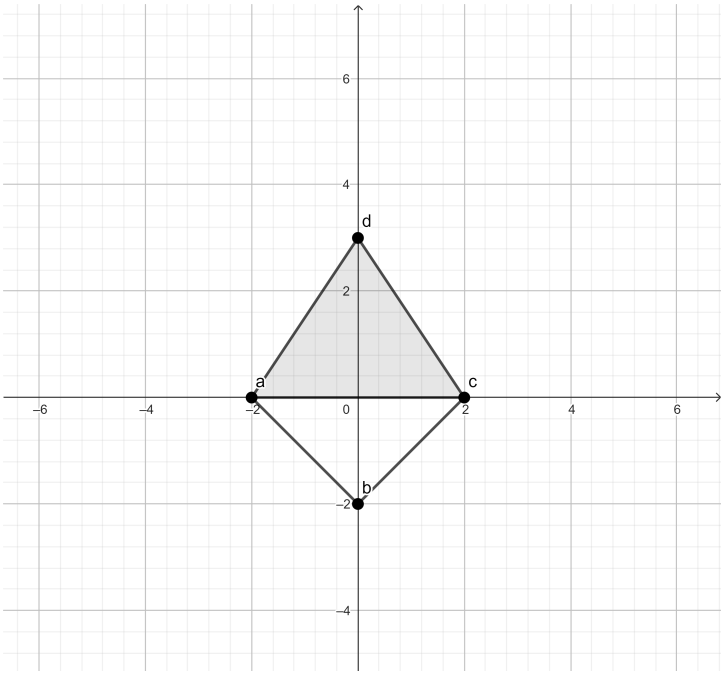


Figure 1.2: A geometric complex labeled to be represented abstractly

complex can always be constructed, and vice versa. The former is a trivial consequence of the definitions above.

Proposition 1.15. Let K be a geometric simplicial complex. Let $\mathrm{Ab}\,(K)$ be the set whose elements are the subsets $\{u_0, \ldots, u_k\}$ of $\mathrm{Vert}\,K$ such that $\langle u_0, \ldots, u_k\rangle \in K$. Then $\mathrm{Ab}\,(K)$ is an abstract simplicial complex, and is called the *vertex scheme* of K. If an abstract simplicial complex $\mathcal{B}$ is isomorphic to $\mathrm{Ab}\,(K)$, then K is called a *geometric realization* of $\mathcal{B}$.

The reverse direction is not as straightforward as the previous proposition, and proving it is beyond our current scope. For a complete proof see for instance [6].

Theorem 1.16 (Geometric Realization Theorem)**.** *Every abstract simplicial complex $\mathcal{A}$ of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$.*

We conclude this subsection with a standard example, to clarify the previous concepts.

Example 1.17. The torus T is a topological subspace of $\mathbb{R}^3$. One can provide a triangulation of that through the standard square model, as in Figure 1.2. By triangulating the square and respecting the identifications, we obtain a structure of an abstract simplicial complex L. Taking the geometric representation of L, $\tilde{L}$ and identifying the correspondent simplices, we obtain an homeomorphism between the body of $\tilde{L}$ and T, our desired triangulation.

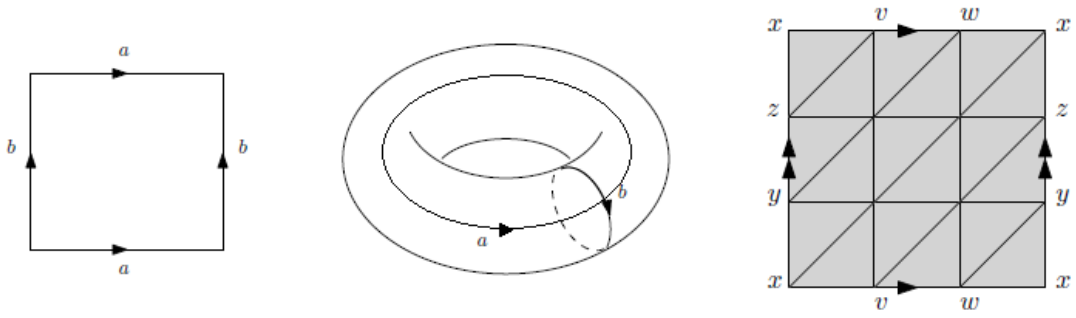


Figure 1.3: On the left: the torus arising from a square. On the right: a triangulation of a torus in terms of an abstract simplicial complex.

Simplicial maps

In the same way as simplicial complexes offer a useful combinatorial framework for representing metric spaces, there exists a natural combinatoric analog of continuous map, the simplicial maps, which we now introduce. Let us begin with a useful proposition.

Proposition 1.18. Let K be a geometric simplicial complex. Every point $x \in |K|$ belongs to the interior of exactly one simplex of K.

Proof. **Existence** $|K| = \bigcup_{\sigma \in K} \sigma$, x belongs to at least one simplex of K. Let τ be the simplex that contains x with minimal dimension. If $x \in \partial\tau$, then x belongs to a face of τ against the minimality of it. Therefore, $x \in \text{int}\tau$.

Uniqueness Let us suppose by absurd that $x \in \operatorname{int}\sigma_1 \cap \operatorname{int}\sigma_2$. Then $x \in \sigma_1 \cap \sigma_2$. Since K is a geometric simplicial complex, τ is a face of both σ_1 and σ_2. If τ were a proper face of σ_i, x would belong to $\partial\sigma_i$, which contradicts the condition $x \in \operatorname{int}\sigma_i$, for $i = 1, 2$. Therefore, $\sigma_1 = \sigma_2$.

□

Using this result, we can introduce the concept of barycentric coordinates. In particular, let K be a geometric simplicial complex and $\{u_0, u_1, \ldots, u_n\} = \operatorname{Vert} K$. Take $x \in |K|$. By the Proposition 1.18 and Definition 1.2, there exists a unique $\sigma = \langle u_0, u_1, \ldots, u_k \rangle$ and unique $\lambda_0, \ldots, \lambda_k$ with $\sum_{i=0}^{k} \lambda_i = 1$ and $\lambda_i \geq 0$ for all i, such that $x = \sum_{i=0}^{k} \lambda_i u_i$.

By setting $b_i(x) = \lambda_i$ for $0 \leq i \leq k$ and $b_i(x) = 0$ for $k+1 \leq i \leq n$, we can write $x = \sum_{i=0}^{n} b_i(x) u_i$, and we refer to the $b_i(x)$ as the *barycentric coordinates* of x in K.

We use these coordinates to construct a piecewise linear, continuous map between (geometric) simplicial complexes (as they are linear and continuous on each simplex).

Definition 1.19. Suppose K and L are geometric simplicial complexes. A map $f : K \to L$ is a *simplicial map* if:

1. For each vertex u of K, its image $f(u)$ is a vertex of L (we call the restricted map $f_{|\text{Vert K}}$ *vertex map*).

2. The induced map between the corresponding abstract simplicial complexes is simplicial, *i.e.*, if $\{u_0, u_1, \ldots, u_k\}$ span a geometric simplex in K, then $\{f(u_0), f(u_1), \ldots, f(u_k)\}$ span a geometric simplex in L.

3. The map f is linear on simplices (in terms of barycentric coordinates), *i.e.*,

$$\forall \lambda_i \in [0,1] : \sum_{i=0}^{k} \lambda_i = 1, \text{ and } \forall u_i \in K^{(0)}, \quad f\left(\sum_{i=0}^{k} \lambda_i u_i\right) = \sum_{i=0}^{k} \lambda_i f(u_i).$$

1.2 Homology groups

Now that we have presented the combinatorial and algebraic prerequisites, we are ready to define homology, our tool for detecting basic topological features such as the number of components, holes, and voids in a simplicial complex (or in triangulable spaces).

Chains and boundary

Definition 1.20. Let K be a simplicial complex, either abstract or geometric, p be any given dimension, and $\mathbb{F}$ a chosen field. A *p-chain* is a linear combination of p-simplices in K, expressed as $c = \sum_{i=1}^{k} a_i \sigma_i$, where the σ_i are the p-simplices in K and the a_i are their respective coefficients in $\mathbb{F}$.

The field $\mathbb{F}$ could be, for instance, $\mathbb{Z}$, $\mathbb{R}$, or $\mathbb{Z}_n$, where n is a prime. In computational topology, it is common to work with coefficients a_i in $\mathbb{Z}_2$, so we will fix $\mathbb{F} = \mathbb{Z}_2$. Under this condition, a chain can be interpreted as the set of p-simplices for which $a_i = 1$.

Two p-chains are added component-wise, like polynomials, and the p-chains together with the addition operation form a free Abelian group called *chain group* and denoted as $(\mathsf{C}_p(K, \mathbb{F}), +)$, or simply $\mathsf{C}_p = \mathsf{C}_p(K)$ if the operation is understood. Associativity follows from associativity of addition modulo 2. The neutral element is $0 = \sum 0\sigma_i$. The inverse of c is $-c = c$ since $c + c = 0$. Moreover, C_p is Abelian because addition modulo 2 is Abelian. For p less than zero and greater than the dimension of K, this group is trivial, consisting only of the neutral element.

Remark 1.21. It is straightforward to see that the chain group $\mathsf{C}_p(K)$ is also a *vector space* over $\mathbb{F}$, where scalar multiplication is defined componentwise.

Thinking of p-simplices of K as an abstract collection of linearly independent vectors, the resulting linear space over $\mathbb{F}$ spanned by them is the chain group. By setting $n_p(K) := \dim \mathsf{C}_p(K)$, then $\mathsf{C}_p(K)$ is isomorphic to $\mathbb{F}^{n_p(K)}$.

Remark 1.22. If $p < 0$ or $p > \dim K$, we set $\mathsf{C}_p(K) := 0$, *i.e.*, the trivial vector space over $\mathbb{F}$.

Definition 1.23. Let $\sigma = \langle u_0, u_1, \ldots, u_p \rangle$ be the p-simplex spanned by the listed vertices. The *boundary* of σ is the sum of its $(p-1)$-dimensional faces, which is

$$\partial_p \sigma = \sum_{j=0}^{p} \langle u_0, \ldots, \hat{u}_j, \ldots, u_p \rangle,$$

where the hat indicates that u_j is omitted[3]. For a p-chain $c = \sum a_i \sigma_i$, the boundary is the sum of the boundaries of its simplices, i.e. $\partial_p c = \sum a_i \partial_p \sigma_i$. Notice also that taking the boundary commutes with addition, that is, $\partial_p(c + c') = \partial_p c + \partial_p c'$. Hence, the boundary defines an homomorphism $\partial_p : \mathsf{C}_p \to \mathsf{C}_{p-1}$ which maps a p-chain to a $(p-1)$-chain. We will therefore refer to ∂_p as the *boundary map for chains*.

Definition 1.24. A *chain complex* $\mathcal{C}$ is a sequence of vector space V_p over a field $\mathbb{F}$ and homomorphisms $d_p : V_p \to V_{p-1}$ indexed by the integer numbers, such that $d_{p-1} \circ d_p$ is the null homomorphism for any $p \in \mathbb{Z}$. Each homomorphism d_p is called a p-boundary map. The elements of V_p, $\ker d_p$, $\operatorname{Im} d_{p+1}$ are respectively called *p-chains*, *p-cycles* and *p-boundaries*. Sometimes, we will use the symbols $Z_p(\mathcal{C})$ to denote the p-cycles and $B_p(\mathcal{C})$ to denote the p-boundaries.

Proposition 1.25. Let K be a geometric simplicial complex, $p \in \mathbb{Z}$. The sequence $\mathcal{C}(K) := (\mathsf{C}_p, \partial_p)_{p \in \mathbb{Z}}$ is a chain complex.

Proof. We only need to prove that $\partial_{p-1} \circ \partial_p(\sigma)$ is the null chain for any $p \in \mathbb{Z}$. The statement is trivial for $p \leq 0$ and for $p > \dim K$, because by Remark 1.22 $\mathsf{C}_p = 0$ for $p < 0$ or $p > \dim K$, thus ∂_p is the trivial map. Therefore, we can assume that $0 < p \leq \dim K$. Take a p-simplex $\sigma = \langle u_0, \ldots, u_p \rangle$ of K. Let us define the symbol σ_{ij} by setting

[3] If the coefficients are in a generic field, the boundary is defined as $\sum_{j=0}^{p} (-1)^i \langle u_0, \ldots, \hat{u}_j, \ldots, u_p \rangle$.

$$\sigma_{ij} := \begin{cases} \langle u_0, \ldots, \hat{u}_i, \ldots, \hat{u}_j, \ldots, u_p \rangle, & \text{if } i < j \\ \text{null chain in } \mathsf{C}_{p-2}(K), & \text{if } i \geq j. \end{cases}$$

We have that

$$\partial_{p-1} \circ \partial_p(\sigma) = \partial_{p-1} \left(\sum_{i=0}^{p} \langle u_0, \ldots, \hat{u}_i, \ldots, u_p \rangle \right) = \sum_{i=0}^{p} \partial_{p-1} \left(\langle u_0, \ldots, \hat{u}_i, \ldots, u_p \rangle \right)$$

$$= \sum_{i=0}^{p} \left(\sum_{j=0}^{i-1} \langle u_0, \ldots, \hat{u}_j, \ldots, \hat{u}_i, \ldots, u_p \rangle + \sum_{j=i+1}^{p} \langle u_0, \ldots, \hat{u}_i, \ldots, \hat{u}_j, \ldots, u_p \rangle \right)$$

$$= \sum_{i=0}^{p} \left(\sum_{j=0}^{i-1} \sigma_{ji} + \sum_{j=i+1}^{p} \sigma_{ij} \right) = \sum_{i=0}^{p} \left(\sum_{j=0}^{p} \sigma_{ji} + \sum_{j=0}^{p} \sigma_{ij} \right)$$

$$= \sum_{i=0}^{p} \sum_{j=0}^{p} \sigma_{ji} + \sum_{i=0}^{p} \sum_{j=0}^{p} \sigma_{ij} = \sum_{i=0}^{p} \sum_{j=0}^{p} \sigma_{ij} + \sum_{i=0}^{p} \sum_{j=0}^{p} \sigma_{ij} = 0,$$

where the last equality follows from the fact that the coefficients are in $\mathbb{Z}_2$. □

We will denote the chain complex $\mathcal{C}(K)$ related to a chain group $\mathsf{C}_p = \mathsf{C}_p(K)$ as

$$\cdots \xrightarrow{\partial_{p+2}} \mathsf{C}_{p+1} \xrightarrow{\partial_{p+1}} \mathsf{C}_p \xrightarrow{\partial_p} \mathsf{C}_{p-1} \xrightarrow{\partial_{p-1}} \cdots$$

It will often be convenient to drop the index from the boundary homomorphism when the dimension of the chain it applies to is clear. See Figure 1.4 for an intuitive representation of a chain complex.

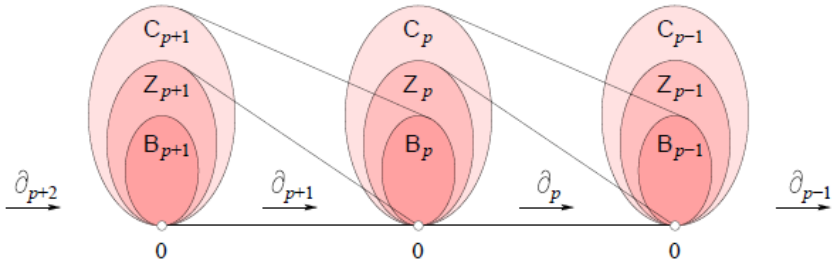


Figure 1.4: The chain complex consisting of a linear sequence of chain, cycle, and boundary groups connected by boundary homomorphisms.

Boundary matrices

Let K be a geometric simplicial complex. For computational purposes, the boundary maps related to $\mathcal{C}(K)$ are typically represented as matrices with entries in $\mathbb{Z}_2$. Take $p \in \mathbb{Z}$ and fix some bases of $\mathsf{C}_p(K)$ and $\mathsf{C}_{p-1}(k)$[4]. The matrix $M_p(K)$ corresponding to ∂_p related to these bases is obtained as follows:

- The p-simplices of K are represented by columns.
- The $(p-1)$-simplices of K are represented by rows.
- The entry at position (i, j) equals 1 if the i-th $(p-1)$ simplex is a face of the j-th p-simplex, otherwise is equal to 0.

In particular, the boundary $\partial_p c$ of a chain $c \in \mathsf{C}_p(K)$ is obtained by multiplying the boundary matrix with the natural representation of c in the chosen basis.

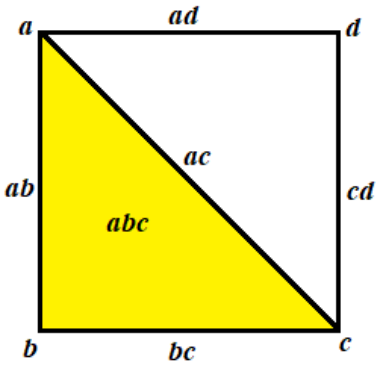


Figure 1.5: Geometric representation of the simplicial complex K, without the Cartesian coordinate system.

Example 1.26. Consider the (abstract) simplicial complex

$$K = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{c, d\}, \{a, b, c\}\}.$$

See Figure 1.5 for a geometric representation. For brevity, denote with abc the basis vector that corresponds to the simplex $\{a, b, c\} = \mathsf{C}_2(K)$. Similarly, we use ab, ac, ad, bc and cd to denote the basis vectors of $\mathsf{C}_1(K)$ and a, b, c, d to denote the basis vectors related to $\mathsf{C}_0(K)$. We order the bases of the vector spaces using lexicographic order. We then have:

[4] As vector spaces over $\mathbb{Z}_2$, you can take the natural ones.

$$M_2(K) = \begin{matrix} abc \\ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \end{matrix} \begin{matrix} \\ ab \\ ac \\ ad \\ bc \\ cd \end{matrix}$$

and

$$M_1(K) = \begin{matrix} \begin{matrix} ab & ac & ad & bc & cd \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \end{matrix} \begin{matrix} a \\ b \\ c \\ d \end{matrix}$$

Homology

We are now finally ready to define homology. For that reason, we first make the following preliminary observation.

Remark 1.27. Take a chain complex $\mathcal{C} = (V_p, d_p)_{p\in\mathbb{Z}}$. The fact that $d_p \circ d_{p+1} \equiv 0$ immediately implies that $\operatorname{Im} d_{p+1} \subseteq \ker d_p$, and hence the quotient vector space $\ker d_p / \operatorname{Im} d_{p+1} = Z_p(\mathcal{C})/B_p(\mathcal{C})$ is well-defined.

Definition 1.28. Given a chain complex $\mathcal{C} = (V_p, d_p)_{p\in\mathbb{Z}}$, we define

$$H_p(\mathcal{C}) = Z_p(\mathcal{C})/B_p(\mathcal{C})$$

and call it the p-th *homology group of* $\mathcal{C}$. Thus, for any given geometric simplicial complex K and every $p \in \mathbb{Z}$ we will simply write $Z_p(K), B_p(K)$ and $H_p(K)$ to denote $Z_p(\mathcal{C}(K)), B_p(\mathcal{C}(K))$ and $H_p(\mathcal{C}(K))$. The vector space $H_p(K)$ is called the *p-th homology group of K with coefficients in $\mathbb{Z}_2$*. The dimension of $H_p(K)$ as a $\mathbb{Z}_2$ vector space is called p-*Betti number*, and it is denoted $\beta_p(K) := \dim H_p(K)$[5]. For convenience, we will also denote $b_p(K) := \dim B_p(K)$, $z_p(K) := \dim Z_p(K)$.

Proposition 1.29. If K is a geometric simplicial complex, then $\beta_p(K) = n_p(K) - b_p(K) - b_{p-1}(K)$ for any $p \in \mathbb{Z}$.

Proof. Consider the dimensional equations for the linear maps $\partial_p : \mathsf{C}_p(K) \to \mathsf{C}_{p-1}(K)$ and $\pi_p : Z_p(K) \to H_p(K) = Z_p(K)/B_p(K)$, where π_p is the quotient projection map. These equations state that $n_p(K) - z_p(K) = b_{p-1}(K)$ and $z_p(K) = b_p(K) + \beta_p(K)$. Our thesis immediately follows from these two equalities. □

[5]Intuitively, it represents the number of p-dimensional holes.

Remark 1.30. Let $b_p(K) := B_p(K)$. Since $B_p(K) \subseteq \mathsf{C}_p(K)$ for any index p, and $\mathsf{C}_p(K)$ is the null space for every $p < 0$ and every $p > \dim K$, it follows that $b_p(K) = 0$ for every $p < 0$ and every $p > \dim K$. Moreover, fixing $q = \dim K$, since $\mathsf{C}_{q+1}(K)$ is the null space, $b_q(K) := \dim \partial_{q+1}(\mathsf{C}_{q+1}(K)) = 0$. Therefore, $b_p(K) = 0$ for every $p < 0$ and every $p \geq \dim K$.

Remark 1.31. Let K be a simplicial complex. When representing the boundary map ∂_p as a matrix $M_p(K)$, the rank of this matrix is $b_p(K)$.

More generally, since each p-th homology group associated with a geometric complex is a finitely generated vector space over $\mathbb{Z}_2$, it is isomorphic to $\mathbb{Z}_2^{\beta_p}$. We will denote this isomorphism as $H_p(K) \cong \mathbb{F}^{\beta_p}$.

Example 1.32. Let us revisit Example 1.26. We can compute the homology of K: firstly, $b_0(K) = 3$ and $b_1(K) = 1$. Moreover, $b_p(K) = 0$ for every $p < 0$ and every $p \geq 2$, as stated in Remark 1.30. Hence, by Proposition 1.29, we have:

- $\beta_0(K) = n_0(K) - b_0(K) - b_{-1}(K) = 4 - 3 - 0 = 1$;
- $\beta_1(K) = n_1(K) - b_1(K) - b_0(K) = 5 - 1 - 3 = 1$;
- $\beta_2(K) = n_2(K) - b_2(K) - b_1(K) = 1 - 0 - 1 = 0$;
- $\beta_p(K) = n_p(K) - b_p(K) - b_{p-1}(K) = 0 - 0 - 0 = 0$ for $p \neq 0, 1, 2$.

It follows that $H_0(K) \cong H_1(K) \cong \mathbb{Z}_2$, while $H_p(K) \cong 0$ for $p \neq 0, 1$.

Induced maps

Let us consider two geometric simplicial complexes, K and L, and a simplicial map $f : K \to L$ between them. Recall that concretely this means that f maps each simplex of K linearly to a simplex of L. This induces a map from the chain groups of K to the chain groups of the same dimension in L, which we will call the *induced map*. Induced maps are particularly useful because they provide the property of functoriality with respect to homology groups, which will become important in the next chapter when discussing persistent homology. Let us formalize these concepts.

Definition 1.33. Let $\mathcal{C} = (\mathsf{C}_p, \partial_p)_{p \in \mathbb{Z}}$ and $\mathcal{C}' = (\mathsf{C}'_p, \partial'_p)_{p \in \mathbb{Z}}$ be two chain complexes. An indexed family of homomorphisms $\varphi = (\varphi_p : \mathsf{C}_p \to \mathsf{C}'_p)_{p \in \mathbb{Z}}$ is called a *chain map* from

$\mathcal{C}$ to $\mathcal{C}'$ if $\partial'_p \circ \varphi_p = \varphi_{p-1} \circ \partial_p$ for every index p. Equivalently, if the following diagram commutes:

$$\begin{array}{ccc} \mathsf{C}_p & \xrightarrow{\partial_p} & \mathsf{C}_{p-1} \\ {\scriptstyle \varphi_p}\downarrow & & \downarrow{\scriptstyle \varphi_{p-1}} \\ \mathsf{C}'_p & \xrightarrow{\partial'_p} & \mathsf{C}_{p-1} \end{array}$$

If each homomorphism φ_p is an isomorphism, we say that φ is an *isomorphism between chain complexes*.

This concept is important for the next proposition.

Proposition 1.34. Let $\varphi = (\varphi_p : \mathsf{C}_p \to \mathsf{C}'_p)_{p\in\mathbb{Z}}$ as above. Then, each map $\varphi_{p,*} : H_p(\mathcal{C}) \to H_p(\mathcal{C}')$, defined by $\varphi_{p,*}([z]) := [\varphi_p(z)]$, is a well-defined homomorphism for every $p \in \mathbb{Z}$. If φ is an isomorphism between chain complexes, then each $\varphi_{p,*}$ is also an isomorphism.

Proof. Take an element $[z]$ in $H_p(\mathcal{C})$, where z is a p-cycle. Since φ is a chain map, we have $\partial'_p(\varphi_p(z)) = \varphi_{p-1}(\partial_p(z)) = \varphi_{p-1}(0) = 0$. This shows $\varphi_p(z)$ is a p-cycle in $\mathcal{C}'$, which means that $[\varphi_p(z)]$ is an element of $H_p(\mathcal{C}')$.

If two elements $[z'] = [z]$ in $H_p(\mathcal{C})$ are equivalent, then $z' - z$ is a p-boundary. That means there exists a $(p+1)$-chain $c \in \mathsf{C}_{p+1}$ such that $z' - z = \partial_{p+1}(c)$. Applying φ_p, we get $\varphi_p(z') - \varphi_p(z) = \varphi_p(z' - z) = \varphi_p(\partial_{p+1}(c)) = \partial'_{p+1}(\varphi_{p+1}(c))$. Thus, $\varphi_p(z') - \varphi_p(z)$ is a p-boundary in $\mathcal{C}'$, showing $[\varphi_p(z')] = [\varphi_p(z)]$ in $H_p(\mathcal{C}')$.

Therefore, $\varphi_{p,*}$ is well-defined.

The linearity of $\varphi_{p,*}$ follows from the linearity of φ_p. If φ is an isomorphism, there exists an inverse map $\varphi^{-1} = (\varphi_p^{-1} : \mathsf{C}'_p \to \mathsf{C}_p)_{p\in\mathbb{Z}}$ which is also a chain map from $\mathcal{C}'$ to $\mathcal{C}$. It is an isomorphism because $(\varphi_p^{-1})_{,*} \circ \varphi_{p,*}([z]) = [\varphi_p^{-1}(\varphi_p(z))] = [z]$. This confirms that $\varphi_{p,*}$ is an isomorphism for each $p \in \mathbb{Z}$.

□

Definition 1.35. Let $\varphi = (\varphi_p : \mathsf{C}_p \to \mathsf{C}'_p)_{p\in\mathbb{Z}}$ be a chain map from $\mathcal{C}$ to $\mathcal{C}'$. The indexed family of homomorphisms $\varphi_* = (\varphi_{p,*} : H_p(\mathcal{C}) \to H_p(\mathcal{C}'))_{p\in\mathbb{Z}}$ defined above is called an *induced map* from $H(\mathcal{C}) := (H_p(\mathcal{C}))_{p\in\mathbb{Z}}$ to $H(\mathcal{C}') := (H_p(\mathcal{C}'))_{p\in\mathbb{Z}}$.

The next result is not hard to prove (for a sketch of the proof, see [6]) and, by applying Proposition 1.34, it provides the definition of induced map related to two (geometric) simplicial complexes.

Proposition 1.36. Take $f : K \to L$ a simplicial map. Consider $f_{p\#} : \mathsf{C}_p(K) \to \mathsf{C}_p(L)$ the map that sends p-chain $c = \sum a_i\sigma_i$ in $\mathsf{C}_p(K)$ to $f_{p\#}(c) = \sum a_i\tau_i$, where

$$\tau_i := \begin{cases} f(\sigma_i), & \text{if } \dim f(\sigma_i) = p \\ 0 \text{ as a chain in } \mathsf{C}_p(L), & \text{if } \dim f(\sigma_i) < p. \end{cases}$$

For $p < 0$ and $p > \dim K$, we set $f_{p\#} : \mathsf{C}_p(K) \to \mathsf{C}_p(L)$ equal to the null map. The collection of maps $f_\# = (f_{p\#} : \mathsf{C}_p(K) \to \mathsf{C}_p(L))_{p\in\mathbb{Z}}$ is a chain map from $\mathcal{C}(K)$ to $\mathcal{C}(L)$.

The fact that the induced map commutes with the boundary map implies that $f_\#$ takes cycles to cycles, $f_\#(Z_p(K)) \subseteq Z_p(L)$, and boundaries to boundaries, $f_\#(B_p(K)) \subseteq B_p(L)$. Therefore, it defines a map on the quotient spaces, which we refer to as the induced map on homology, denoted by $f_* : H_p(K) \to H_p(L)$.

We conclude this section with the following theorem (its proof follows straightforwardly from the definition of the chain map induced by a simplicial map), which clarifies what we mean by the property of functoriality.

Theorem 1.37. *The map F_p taking each geometric simplicial complex K to $H_p(K)$ and each simplicial map $f : K \to L$ to the map $f_* : H_p(K) \to H_p(L)$ (induced by the chain map $f_{p\#}$ is a covariant functor for every $p \in \mathbb{Z}$, i.e.,*

$$(g \circ f)_* = g_* \circ f_*.$$

1.3 Constructions of simplicial complexes

Since homology computations for simplicial complexes can be carried out algorithmically, it is often advantageous to construct simplicial complexes that either compute the homology of an underlying space X or are closely related to it. To ensure that a simplicial complex accurately computes the homology of X, a rigorous approach is to establish a *homotopy equivalence* between X and the simplicial complex K, or between a space homotopy equivalent to X and K. A *homotopy equivalence* is a map $g : X \to K$ such that there exists a map $f : K \to X$ with $f \circ g$ homotopic to the identity map on X and $g \circ f$ homotopic to the identity map on K. In this case, K and X are said to be *homotopy equivalent*. If two spaces X and Y are homotopy equivalent, then their homology groups $H_p(X)$ and $H_p(Y)$ are isomorphic for all p (see [11] for a complete proof).

Various simplicial complexes can be derived from X. We introduce two of the most used simplicial complexes in computational topology, Čech and Rips complexes.

Čech complexes

Čech complexes are a special case of a general topological construction called nerve.

Definition 1.38. The *nerve* Nrv$\mathcal{F}$ of a finite collection $\mathcal{F}$ of subsets of a nonempty set X is the abstract simplicial complex, whose simplices are all the subcollections $\mathcal{F}'$ of $\mathcal{F}$ such that

$$\bigcap_{Y \in \mathcal{F}'} Y \neq \emptyset.$$

We can realize geometrically the nerve in some Euclidean space, as stated by Theorem 1.16, in order to talk about its topology and homotopy type (see Figure 1.6). One of

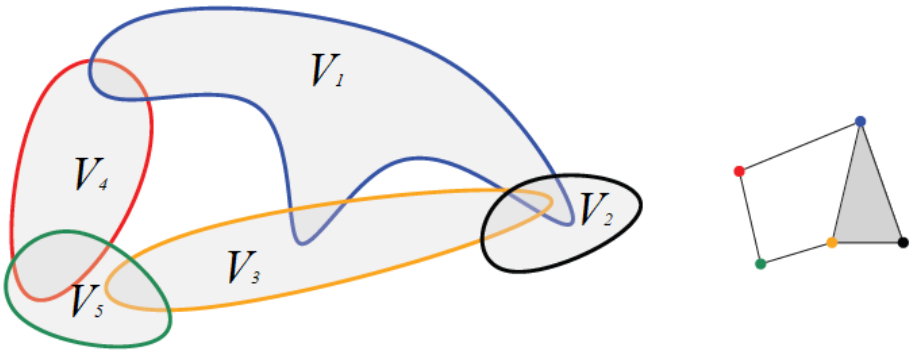


Figure 1.6: A family F of sets and a geometric realization of its nerve

the advantages of nerve complexes is that, in some cases, their homotopy type [6] is equal to the union of the elements of the given collection. This is formalized by the following theorem, which is proved for example in [23].

Theorem 1.39 (Nerve Theorem)**.** *Let $F = \{V_1, \ldots, V_k\}$ be a finite collection of closed convex subsets of $\mathbb{R}^d$. Then Nrv$\mathcal{F}$ and the set $\bigcup_{i=1}^{k} V_i$ have the same homotopy type, i.e. $\bigcup_{i=1}^{k} V_i \simeq$ Nrv$\mathcal{F}$.*

It follows that taking a topological space X and an open, finite covering $\mathcal{F} = \{U_1, \ldots, U_n\}$ of convex sets, then Nrv$\overline{\mathcal{F}}$ is homotopy equivalent to X, where $\overline{\mathcal{F}} = \{\overline{U_1}, \ldots, \overline{U_n}\}$[7]. The requirement on the sets can be relaxed without affecting the conclusion. Specifically, if $X = \bigcup_{F \in \mathcal{F}} F$ is triangulable, all sets in $\mathcal{F}$ are closed, and all non-empty common intersections are contractible[8], then Nrv$\mathcal{F} \simeq \bigcup_{F \in \mathcal{F}} F$ (see [6] for

[6] The homotopy type of a space X is the class of topological space homotopic equivalent to X.

[7] Given a set U, the symbol $\overline{U}$ denotes the closure of that set.

[8] A space is contractible if it has the homotopy type of a point.

details).

Definition 1.40. Let S be a finite set of points in $\mathbb{R}^d$. The Čech complex of S with radius $r \geq 0$ is the abstract simplicial complex defined as follows:

$$\check{C}(S,r) = \check{C}(r) = \left\{ \sigma \subseteq S \mid \bigcap_{x \in \sigma} B_x(r) \neq \emptyset \right\}.$$

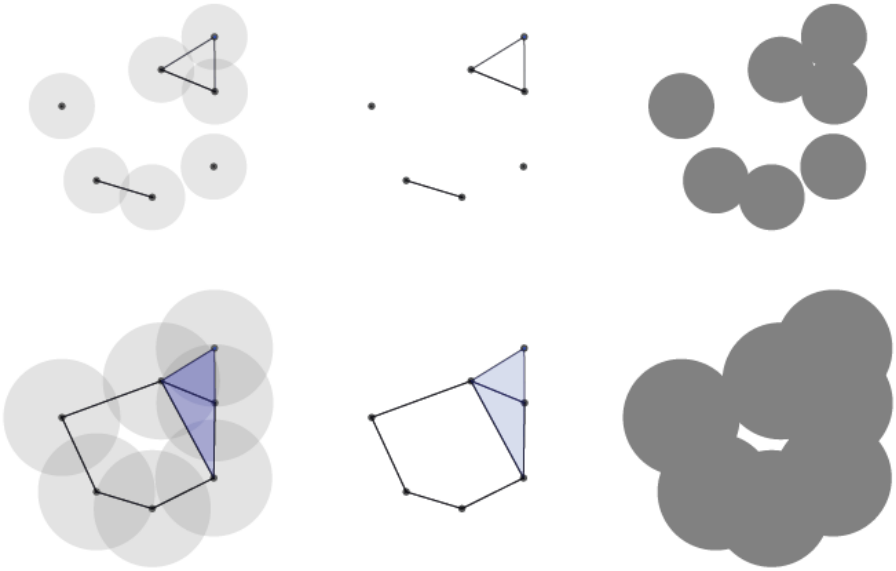


Figure 1.7: Two examples of Čech complexes.

It is easy to see that $\check{C}(S,r)$ is isomorphic to $\mathrm{Nrv}\mathcal{F}$, with $\mathcal{F} = \{B(s,r)\}_{s \in S}$. Thus, by the *Nerve Theorem* 1.39, we have $\check{C}(S,r) \simeq \bigcup_{F \in \mathcal{F}} F$. Furthermore, this observation can be used to prove reconstruction results: given a closed, connected surface X in Euclidean space, for each sufficiently small scale parameter $r \geq 0$ and for each sufficiently dense finite subset $S \subseteq X$, we have $X \simeq \check{C}(S,r)$, i.e., the homotopy type of the space X can be reconstructed using Čech complexes.

Vietoris-Rips Complexes

From a computational point of view, Čech complexes are expensive to construct because one has to check for large numbers of intersections. To circumvent this issue, one can instead consider the Vietoris–Rips (VR) complex, which approximates the Čech complex.

Definition 1.41. Let S be a finite set of points in $\mathbb{R}^d$. Then the Vietoris-Rips complex for S with radius r, denoted by $\mathrm{VR}(S, r)$ or simply $\mathrm{VR}(r)$, will be the abstract simplicial complexes whose simplexes are all the subsets of S with diameter[9] at most $2r$. In symbols:

$$\mathrm{VR}(r) = \{\sigma \subseteq S \mid \operatorname{diam} \sigma \leq 2r\}.$$

Remark 1.42. The condition $\operatorname{diam} \sigma \leq 2r$ means that the distance between any two vertices of σ is at most $2r$. Thus, it immediately implies that if $0 \leq r_1 \leq r_2$, then $\check{C}(S, r_1) \subseteq \check{C}(S, r_2) \subseteq \mathrm{VR}(S, r_2)$.

The containment relation can be reversed if we are willing to increase the radius of the Čech complex by a multiplicative constant. Before stating this result, let us see some preliminary notions.

Definition 1.43. For each non-empty compact set $K \subseteq \mathbb{R}^d$, the unique closed ball containing K and having minimum radius is called *miniball* of K.

Proposition 1.44. Let $\{p_1, \ldots, p_n\} \subset \mathbb{R}^d$. If $F = \{\bar{B}_1, \ldots, \bar{B}_n\}$ is the set of all *closed* balls $\bar{B}_i$ of center p_i and radius $r \geq 0$, for $1 \leq i \leq n$, then these three properties are equivalent:

1. $\bigcap_{i=1}^n \bar{B}_i \neq \emptyset$;
2. There exists a point $z \in \mathbb{R}^d$, such that the closed ball of center z and radius r contains $\{p_1, \ldots, p_n\}$;
3. The miniball of $\{p_1, \ldots, p_n\}$ has radius less than or equal to r.

Proof. 1) $\implies$ 2) Just take a point $z \in \bigcap_{i=1}^n \bar{B}_i$. By symmetry, since $z \in \bar{B}_i$, $p_i \in \bar{B}(z, r)$[10], for $1 \leq i \leq n$.

2) $\implies$ 3) It follows from the minimality of the radius of the miniball containing $\{p_1, \ldots, p_n\}$.

3) $\implies$ 1) It is sufficient to observe that the center of the miniball containing $\{p_1, \ldots, p_n\}$ belongs to $\bigcap_{i=1}^n \bar{B}_i$.

□

[9] The diameter of a subset $A \subseteq X$ of a metric space X is defined as $\operatorname{diam}(A) = \sup_{x,y \in A} d(x, y)$. If A is finite, it coincides with the max.

[10] It is the closed ball centered in z with radius r.

Lemma 1.45 (Vietoris-Rips Lemma)**.** Let S be a finite set of points in some Euclidean space and $r \geq 0$, then $\mathrm{VR}(r) \subseteq \check{C}(\sqrt{2}r)$.

Sketch of the proof. Let $M(\sigma)$ be the miniball of a k-simplex $\sigma = \{p_{i_0}, \ldots, p_{i_k}\} \in \mathrm{VR}(r)$. Call z the center of $M(\sigma)$ and ρ its radius. It is sufficient to prove that

$$\sqrt{2}\rho \leq \operatorname{diam} \sigma \leq 2\rho.$$

Since this statement trivially holds for $k = 0$, let us assume that $k \geq 1$ (and hence $\rho > 0$). For a proof that $\sqrt{2}\rho \leq \operatorname{diam} \sigma$, see [8]. The inequality $\operatorname{diam} \sigma \leq 2\rho$ follows from the fact that $\sigma \subseteq M(\sigma)$ and $\operatorname{diam} M(\sigma) = 2\rho$.

Since $\sigma \in \mathrm{VR}(r)$, we know that $\operatorname{diam} \sigma \leq 2r$. It follows that $\sqrt{2}\rho \leq 2r$, and hence $\rho \leq \sqrt{2}r$. Therefore, the definition of Čech complex and Proposition 1.44 imply that the collection σ' of balls of radius $\sqrt{2}r$, whose centers belong to σ, correspond (their center) to a simplex in $\check{C}(\sqrt{2}r)$. It follows that $\mathrm{VR}(S, r) \subseteq \check{C}(S, \sqrt{2}r)$. □

Chapter 2

Persistent Homology

In this chapter, we provide an overview of persistent homology, a natural extension of traditional homology that measures how homological elements, such as components, holes, and other features, persist, *i.e.*, remain non-trivial, through the steps of a *filtration*. To that end, we first introduce the concept of filtration, then provide a formal definition of persistent homology, and finally present two ways to visualize it. We conclude this chapter with a brief section on the stability of this method. Unless otherwise noted, we refer to [6, 23, 4] for definitions and to [19, 23] for representations.

2.1 Filtrations

We begin by formally introducing the concept of a filtration of simplicial complexes, which consists of a nested sequence of increasingly larger complexes that represent the evolution of a growing simplicial complex. Filtrations can be either discrete or continuous. We present both approaches: the discrete setting allows for a more straightforward definition and visualization of persistent homology, while the stability of persistent homology relies on the continuous variation of a parameter[1].

Discrete filtration

Definition 2.1. Let K be a simplicial complex. A *(discrete) simplicial filtration* $\mathcal{F} = \mathcal{F}(K)$ of K is a nested sequence of subcomplexes

$$\mathcal{F} : \emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K.$$

[1]This will be discussed in detail later. In essence, continuous variation ensures the continuity of persistent homology.

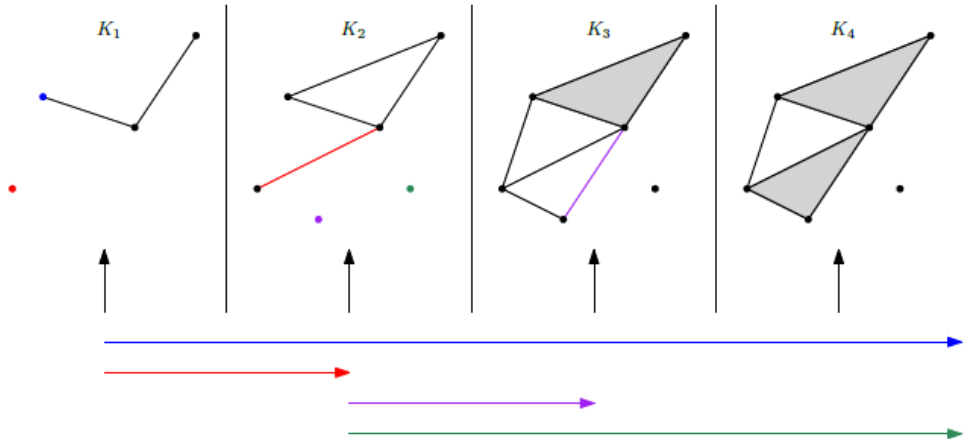


Figure 2.1: A filtration of simplicial complexes and its correspondent 0-dimensional barcode. The left endpoint of each bar corresponds to the birth complex of a component. The right endpoint of each bar corresponds to the terminal complex of a component. The color of each bar also appears on one vertex (the representative of the component) and potentially on one edge (the edge, that terminates the component).

which can also be expressed as a sequence of natural inclusion maps[2] denoted by

$$\mathcal{F} : \emptyset = K_0 \xhookrightarrow{i_{0,1}} K_1 \xhookrightarrow{i_{1,2}} K_2 \xhookrightarrow{i_{2,3}} \ldots \xhookrightarrow{i_{m-1,m}} K_m = K.$$

Example 2.2. An example of a simplicial filtration is illustrated in Figure 2.1. The nested simplicial complexes $K_1 \subseteq K_2 \subseteq K_3 \subseteq K_4$ are separated by vertical lines. K_0 is omitted for simplicity. The horizontal arrows below, referred to as "bars", form a barcode, which visually represents the persistence of specific features. In this case, they show the persistence of zero-dimensional homology classes, namely the components. We will provide a more precise definition of a barcode later in this chapter.

Continuous Filtrations

Definition 2.3. A *continuous filtration* of a finite simplicial complex K is a collection of subcomplexes $\{K_r\}_{r\geq 0}$ of K such that:

$$\forall r < q : K_r \subseteq K_q \subseteq K.$$

Definition 2.4. Given a simplicial complex K, let f be a *filtration function*[3], *i.e.*, an

[2] In general, $i_{s,t} : K_s \hookrightarrow K_t$

[3] It can be referred also as annotation function, monotonic function or simplex-wise monotone function, see for instance [6] or [4]

annotation of each of the simplices of K by a real number such that $\sigma \leq \tau \Rightarrow f(\sigma) \leq f(\tau)$. The *sublevel filtration* associated to f is a continuous filtration consisting of sublevel complexes $K_r = f^{-1}((-\infty, r]) \subseteq K$ for $r \in \mathbb{R}$.

Remark 2.5. The property $\sigma \leq \tau \Rightarrow f(\sigma) \leq f(\tau)$ ensures that the sublevel sets $f^{-1}((-\infty, a])$ are subcomplexes of K for every $a \in \mathbb{R}$. Additionally, we need to set a value a_0 to get $K_{a_0} = K_0 = \emptyset$, such as $a_0 = -\infty$.

The Rips and Čech filtrations, as defined in Chapter 1, are continuous filtrations of this sort.

Example 2.6 (Čech Filtration)**.** Let X be a metric space $S \subseteq X$ be a finite subset. The *Čech filtration* of S is the collection of abstract simplicial complexes $\{\check{C}(S, r)\}_{r \geq 0}$ along with inclusions $i_{r_1, r_2} : \check{C}(S, r_1) \hookrightarrow \check{C}(S, r_2)$ for all $r_1 \leq r_2$.

Example 2.7 (Rips Filtration)**.** Let X be a metric space and let $S \subseteq X$ be a finite subset. The *Rips filtration* on S is the collection of abstract simplicial complexes $\{\mathrm{VR}(S, r)\}_{r \geq 0}$ along with inclusions $i_{r_1, r_2} : \mathrm{VR}(S, r_1) \hookrightarrow \mathrm{VR}(S, r_2)$ for all $r_1 \leq r_2$. See the top part of Figure 2.2 for an example.

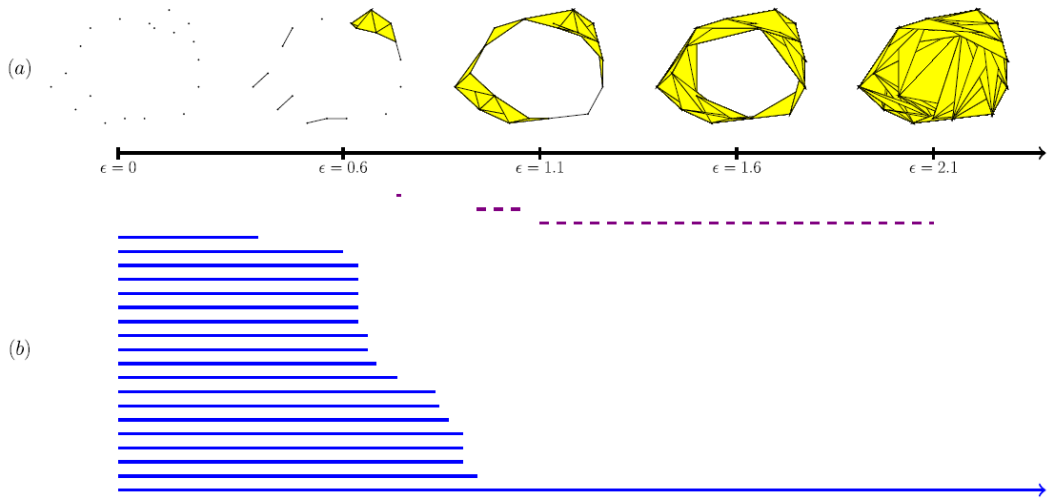


Figure 2.2: **(a)** A finite set of points in $\mathbb{R}^2$ (for $\epsilon = 0$) and the Vietoris-Rips filtration constructed from it (for ϵ ranging from 0 to 2.1). **(b)** The barcode corresponding to the nested sequence of spaces shown in (a), where solid lines represent the lifetimes of connected components and dashed lines represent the lifetimes of holes.

Given a discrete filtration, there is a straightforward extension of it as the sublevel filtration of the annotation function. Conversely, for a continuous sublevel filtration $\{K_r\}_{r\geq 0}$ associated with a filtration function f, the most effective way to generate a discrete filtration is to consider the index i as the index of the critical scale[4] of the continuous filtration. Formally, since K is finite by definition, Im f has finite cardinality. Therefore, we can define the critical scales $r_1 < r_2 < \cdots < r_k$ as the enumeration of the image Im $f = \{r_1, r_2, \ldots, r_k\}$, and define K_i as the corresponding sublevel sets. This finite filtration retains all the information about the changes in the original continuous filtration.

Continuous filtrations conveniently model the geometric setup of standard filtrations. On the other hand, discrete filtrations are a convenient finite description on which we may develop algorithmic approaches.

2.2 Definition and Visualization

We are now ready to introduce the formal definition of persistent homology. Note that, from this point forward, we will focus on discrete filtrations. However, the definition of persistent homology can also be applied to continuous filtrations, with the appropriate adjustments to the indices of the groups.

Definition 2.8. Let K be a simplicial complex, $\mathbb{F} = \mathbb{Z}_2$[5], and $p \in \{0, 1, 2, \ldots\}$. Given a filtration

$$\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K$$

of K, the corresponding p-dimensional *persistent homology groups* with coefficients in $\mathbb{F}$ are the images of the maps $(i_{s,t})_* : H_p(K_s; \mathbb{F}) \to H_p(K_t; \mathbb{F})$ for all $0 \leq s \leq t \leq m$. In symbols:

$$H_p^{s,t} = \text{Im}(i_{s,t})_* \qquad \text{for } 0 \leq s \leq t \leq m.$$

The corresponding ranks $\beta_{s,t}^p = \text{rank}(i_{s,t})_*$ are called *persistent Betti numbers*[6].

Remark 2.9. By the functoriality of homology, *i.e.* $(i_{u,t})_* \circ (i_{s,u})_* = (i_{s,t})_*$, we obtain a sequence of homology groups

$$\emptyset = H_p(K_0; \mathbb{F}) \xhookrightarrow{(i_{0,1})_*} H_p(K_1; \mathbb{F}) \xhookrightarrow{(i_{1,2})_*} \ldots \xhookrightarrow{(i_{m-1,m})_*} H_p(K_m; \mathbb{F}) = H_p(K; \mathbb{F}).$$

[4] A scale r of a continuous filtration is considered critical if at least one simplex appears at r.

[5] In this work we focus on $\mathbb{Z}_2$, but the definition is true for an arbitrary field.

[6] Note that $\beta_{s,t}^p$ is a non-increasing function in t and a non-decreasing function in s.

As is the case with ordinary homology, each persistent homology group is determined up to isomorphism by its Betti number. As we go from K_{i-1} to K_i, we gain new homology classes and we lose some when they become trivial or merge with each other. Intuitively, the persistent homology groups consist of the homology classes of K_s that are "still alive" in K_t. In the next section, we will see more precisely what does it mean to be still alive and how to visualize this idea.

Barcodes

Let us fix a filtration $K_1 \subseteq K_2 \subseteq \ldots \subseteq K_m = K$ of the simplex complex K, $\mathbb{F} = \mathbb{Z}_2$ and $p \in \{0, 1, \ldots\}$. As introduced in the previous section, the persistent Betti number $\beta^p_{s,t}$ represents the dimension of the subspace of homology elements in K_t that have a representative in K_s. More rigorously, it is the dimension of the collection of homology elements in K_s that remain non-trivial in K_t, defined as $\beta^p_{s,t} = \dim H_p(K_s)/\ker(i_{s,t})_*$[7]. Barcodes, as mentioned above, provide a more intuitive representation of specific information: the *lifetime* of homology classes. A bar $[s, t)$ represents a homology element that is born at s and terminates at t. Formally, we define:

1. $\beta_{s,t}$ represents the number of bars that begin at s and persist through t.

2. Homology *born* at s is defined as $H_p(K_s)/(\text{Im } i_{s-1,s})_*$, where we quotient the homology classes that already have a representative in K_{s-1}. For formal reasons, $(i_{0,t})_*$ is defined as the trivial map. The dimension of this homology is $\beta_{s,s} - \beta_{s-1,s}$, which represents the number of bars that begin at s.

3. Homology *terminating* at t is defined as $\ker(i_{t-1,t})_*$. Its dimension is $\beta_{t-1,t-1} - \beta_{t-1,t}$, since $\dim \ker(i_{t-1,t})_* = \dim H_p(K_{t-1})/\text{Im}(i_{t-1,t})_*$. This value corresponds to the number of bars that terminate at t.

4. The quantity $\beta_{s,t} - \beta_{s-1,t}$ measures the dimension of homology born at s that is still alive at t. Specifically, $\beta_{s,t} - \beta_{s-1,t} = \dim((\text{Im } i_{s,t})_*/\text{Im}(i_{s-1,t})_*)$, *i.e.*, the dimension of homology classes in $H_p(K_t)$ that have representatives in K_s, modulo those already present in K_{s-1}. This represents the number of bars starting at s and continuing through t.

5. The quantity $n_{s,t} = \beta_{s,t-1} - \beta_{s-1,t-1} - (\beta_{s,t} - \beta_{s-1,t})$ indicates[8] the dimension of

[7] Throughout the rest of this section, we will omit the superscript p indicating the fixed dimension.

[8] This can be interpreted as (the dimension of homology born at s and still alive at $t-1$) minus (the dimension of homology born at s and still alive at t).

homology born at s that terminates at t, representing the number of bars that begin at s and end at t.

6. Finally, $n_{s,\infty} = \beta_{s,m} - \beta_{s-1,m}$ represents the dimension of homology born at s that persists through to the end of the filtration.

The p-dimensional barcode consists of intervals of the form:

i. $[s, t)$ for $1 \leq s < t \leq m$, and

ii. $[s, \infty)$ for $1 \leq s < m$.

Each interval can have different multiplicity: the number of intervals $[s, t)$ is denoted by $n_{s,t}$, with $1 \leq s < t \leq \infty$.

Example 2.10. We again turn our attention to the filtration in Figure 2.1. From the Table 2.1 below, we can deduce that $n_{2,3} = 1 - 1 - (2 - 3) = 1$, and as a result, there is 1 bar of the form $[2, 3)$, as displayed in the figure. Similarly, we compute $n_{1,2} = n_{1,\infty} = n_{2,\infty} = 1$ and $n_{1,3} = n_{1,4} = n_{2,4} = n_{3,4} = n_{3,\infty} = n_{4,\infty} = 0$

Persistence diagrams

Another well-established method for visualizing persistent homology is through *persistence diagrams*, which are defined as follows. Given a barcode, as previously described, we can represent each interval $[s, t)$ as a pair of numbers and visualize it as a point (s, t) in $\mathbb{R}^2$. Note that a point of the form (s, ∞) cannot be directly represented on a plane, so we select a y-coordinate above a certain value k, such as $k + 1$, to approximate ∞.

Each point (s, t) on a persistence diagram is assigned a multiplicity $n_{s,t}$, which indicates the number of intervals of the form $[s, t)$.

The resulting collection of weighted points in the plane is known as a *persistence diagram*. An example of this can be seen in Figure 2.3. A barcode encodes the same information as a persistence diagram, but while the *persistence* of a bar is measured by its length, the *persistence* of a point on a persistence diagram is measured by its distance from the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$. All points on a persistence diagram lie above this diagonal. In theory, if bars of length zero $[s, s)$ existed, they would correspond to points on the diagonal (s, s).

Persistence diagrams are often preferred for visualizing persistent homology, especially when the number of points and bars is large, as their distribution tends to be

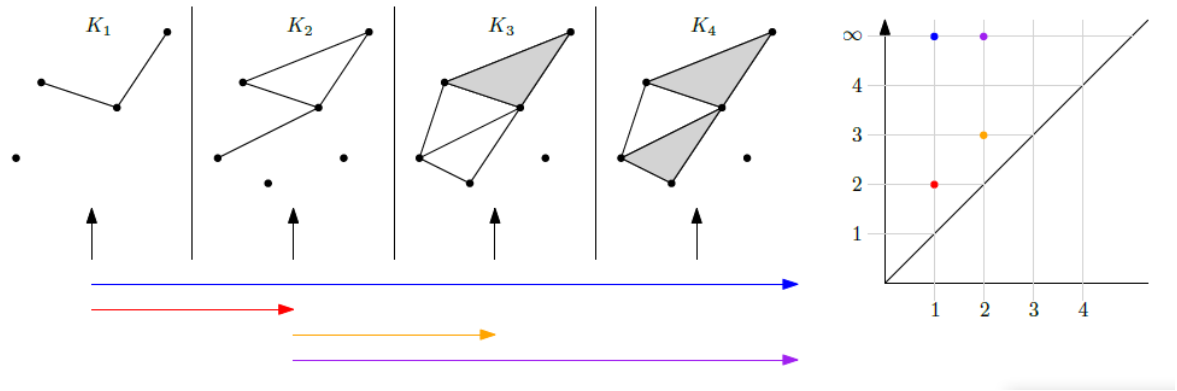


Figure 2.3: A filtration along with the corresponding zero-dimensional barcode and persistence diagram. The colors of bars match the colors of the corresponding points in the persistence diagram.

s \ t	1	2	3	4
1	2	1	1	1
2	/	3	2	2
3	/	/	2	2
4	/	/	/	2

Table 2.1: Table of zero-dimensional persistent Betti-numbers $\beta^0_{s,t}$.

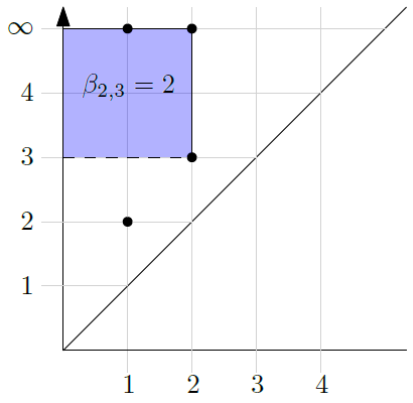


Figure 2.4: The sum of multiplicities of points in the blue quadrant with apex $(2, 3)$ is $\beta_{2,3}$ by Lemma 2.11.

well-represented in this form. Conversely, when the number of points and bars is small, barcodes are often more descriptive.
As the multiplicities $n_{s,t}$ are defined using persistent Betti numbers $\beta_{s,t}$, it turns out that a reverse expression also exists.

Lemma 2.11 (Fundamental Lemma of Persistent Homology)**.**

$$b_{s,t} = \sum_{s' \leq s,\, t' > t} n_{s,t},$$

with t' that could also take the value ∞.

The formula in the lemma can be verified explicitly. However, the statement is apparent from the definitions, as $\beta_{s,t}$ represents the homology born at s or before and terminating after t, while $n_{s,t}$ represents the homology born precisely at s and terminating precisely at t.

Remark 2.12. Lemma 2.11 has a geometric interpretation in the context of persistence diagrams (see Figure 2.4). It essentially states that $\beta_{s,t}$ is the sum of all multiplicities of points in a persistence diagram that lie in the upper-left quadrant $[0, s] \times (t, \infty]$. In the context of this interpretation, the formula for multiplicity

$$n_{s,t} = b_{s,t-1} - b_{s-1,t-1} - b_{s,t} + b_{s-1,t}$$

is the expression of the square $(s-1, s] \times [t-1, t)$ in terms of such quadrants.

Remark 2.13. The Lemma 2.11 also implies that the information encoded in a barcode or in a persistence diagram is precisely the same as the information encoded by persistent Betti numbers.

2.3 Stability

Persistence is a measure-theoretic concept built upon algebraic structures, with its most significant property being stability under data perturbations, which means that small changes in the data result in only minor changes in the persistence. We can formalize this statement using a simple tool, the bottleneck distance, that formalize the concepts of similarity between persistence diagrams.

Bottleneck distance

Suppose $\mathcal{A} = (a_1, a_2, \ldots, a_m)$ and $\mathcal{B} = (b_1, b_2, \ldots, b_n)$ are persistence diagrams, *i.e.,* a_i and b_J are point above the diagonal in the first quadrant in the plane, with their multiplicity, for all $i = 1, \ldots m$ and $j = 1, \ldots, n$,

For a point $v = (v_1, v_2) \in \mathbb{R}^2$, let $\bar{v} = \left(\frac{v_1+v_2}{2}, \frac{v_1+v_2}{2}\right) \in \mathbb{R}^2$ be the point on the diagonal $\Delta = \{(z, z) \mid z \in \mathbb{R}\}$ which is the closest to v in d_∞ (and also in d_2) metric[9].

Definition 2.14. Let $\mathcal{A}' \subseteq \mathcal{A}$ and $\mathcal{B}' \subseteq \mathcal{B}$. A *partial matching* between $\mathcal{A}$ and $\mathcal{B}$ is a bijective map $\varphi : \mathcal{A}' \to \mathcal{B}'$. The *matching distance* of such φ is defined as

$$d_M(\varphi) = \max\left\{\max_{v\in\mathcal{A}'} d_\infty(v, \varphi(v)), \max_{v\in\mathcal{A}\setminus\mathcal{A}'} d_\infty(v, \bar{v}), \max_{v\in\mathcal{B}\setminus\mathcal{B}'} d_\infty(v, \bar{v})\right\}.$$

Definition 2.15. Let $\mu(\mathcal{A}, \mathcal{B})$ denote the collection of all partial matchings between $\mathcal{A}$ and $\mathcal{B}$. The *bottleneck distance* between persistence diagrams $\mathcal{A}$ and $\mathcal{B}$ is the minimal matching distance between them, *i.e.*,

$$d_B(\mathcal{A}, \mathcal{B}) = \min_{\varphi\in\mu(A,B)} d_M(\varphi).$$

Remark 2.16. Clearly, $d_B(\mathcal{A}, \mathcal{B}) = 0$ if and only if $\mathcal{A} = \mathcal{B}$. Furthermore, $d_B(\mathcal{A}, \mathcal{B}) = d_B(\mathcal{B}, \mathcal{A})$, and $d_B(\mathcal{A}, \mathcal{C}) \leq d_B(\mathcal{A}, \mathcal{B}) + d_B(\mathcal{B}, \mathcal{C})$ for persistence diagrams $\mathcal{A}, \mathcal{B}, \mathcal{C}$, thereby justifying its classification as a distance.

Examples of partial matchings are given in Figure 2.5. The unmatched points are connected to the closest point on the diagonal. Note that the $d_\infty(a, b)$ distance between points a and b can be thought of as representing one half of the side length of the square centered at a which has b on its boundary, as shown in the second line of the cited figure. According to this, it is easy to see that the pair with the smallest matching distance is the second from the left, and this quantity is the actual bottleneck distance d_B.

Stability theorem

Now, we are ready to present one of the main theorem for stability of persistence diagrams related to filtrations (for the proof, see [3]). More specific result, related to particular set of functions along with stability theorem related to other concepts of distance between persistence diagrams, can be found in [6] or [23].

[9] Recall that these two distances in $\mathbb{R}^d$ are defined as follows:

$d_\infty(v, u) = \max_{i=1,2}\{|v_i - u_i|\}$, where $v = (v_1, v_2)$ and $u = (u_1, u_2)$.

$d_2(v, u) = \sqrt{|v_1 - u_1|^2 + |v_2 - u_2|^2}$ (euclidean distance).

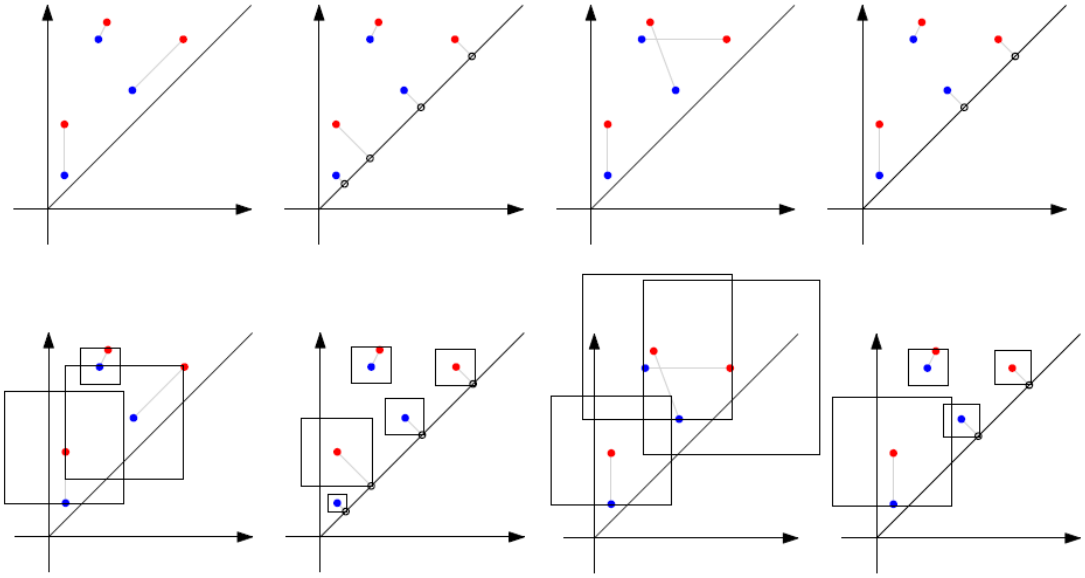


Figure 2.5: In the first line, examples of partial matchings between the red and the blue persistence diagrams, with points unmatched by φ being matched to the closest diagonal point; in the second line the distances between pairs are demonstrated by the squares arising as the balls of the d_∞ metric.

Theorem 2.17. *Let K be a simplicial complex and $f, g : K \to \mathbb{R}$ be two monotonic functions. Let $Dg_p(f)$ and $Dg_p(g)$ denote the persistence diagrams related to the sublevel filtration associated to f and g, respectively. For each dimension p, the bottleneck distance between the diagrams is bounded from above by the d_∞-distance between the functions, $d_B(Dg_p(f), Dg_p(g)) \leq \|f - g\|_\infty$*[10].

The essence of the theorem is that small perturbations of the input lead to small changes in persistence diagrams. This property has a crucial impact in computer science and practical applications: it ensures that minor variations in the positions of points in a data cloud do not significantly alter the persistence diagram, thereby providing a robust topological analysis even with imprecise data.

[10]Recall that $\|f - g\|_\infty = d_\infty(f, g) = \sup_{x \in K} |f(x) - g(x)|$.

Chapter 3

Identifying Lasso Proteins with Persistent Homology

In this chapter, we see how persistent homology can be applied to biology, by identifying lasso structures in proteins. We start by introducing key concepts related to protein structure and topology, focusing on lasso proteins. We then discuss how persistent homology can be used to identify these lasso structures, in the last section. General information on proteins is based on [1, 15], while the discussion on lasso proteins draws from [18, 17, 16], and the final section references to [9].

3.1 Proteins: Structure Fundamentals

Proteins are one of the three main biological macromolecules, along with DNA and RNA. These molecules are closely interconnected: DNA is transcribed into RNA, which is then translated into proteins. Proteins carry out essential functions, including catalyzing reactions, transporting molecules, coordinating cell processes, and providing structural support. Studying protein structure is crucial, as their functions depend on their three-dimensional shapes, which to a large extent is determined by their specific amino acid sequences[1].

[1]Predicting a protein's tertiary structure from its sequence alone is difficult because of the complexities involved in protein folding. However, if the structure of a related protein (from the same family) is known, it is possible to accurately predict the tertiary structure through a computational technique called homology modeling.

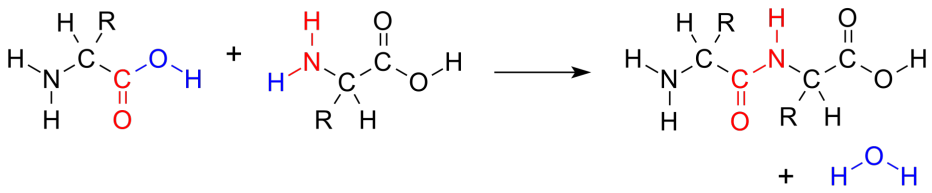


Figure 3.1: The dehydration condensation of two amino acids to form a peptide bond (red) with expulsion of water (blue)

Amino Acids: The Building Blocks of Proteins

While DNA and RNA are composed of four different nucleotides, proteins are composed of 20 different amino acids. These basic units share a common structure, consisting of a central carbon atom (the α-carbon C_α) bonded to an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom, and an R-group or side chain. The R-group varies between amino acids and determines their chemical properties[2].

Amino acids are linked together by peptide bonds, which are formed through a dehydration synthesis reaction between the carboxyl group of one amino acid and the amino group of another. This bond formation results in a *polypeptide chain* (see Figure 3.1), which then folds into a specific three-dimensional structure to form a functional protein.

Levels of Organization

The multiplicity of functions carried out by proteins arises from the vast number of different shapes they can adopt. The three-dimensional structure of a protein is typically described at four different levels of organization: primary, secondary, tertiary, and quaternary (see Figure 3.2).

Primary Structure The sequence of amino acids in a polypeptide chain is referred to as its *primary structure*. This sequence is determined by the gene encoding of the protein and is unique for each protein. It can vary greatly in length. Generally, when a polypeptide consists of only a few amino acids, it is referred to as an oligopeptide (or simply peptide). The average size of proteins is on the order of hundreds of amino acid residues, although proteins composed of many thousands of amino acids are also known.

[2]There are four main categories of amino acids based on their side chains: aromatic, polar, non-polar, and charged. See [15] for details.

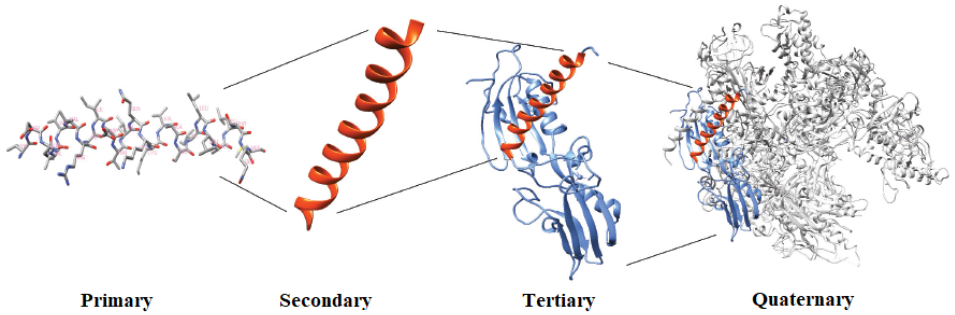


Figure 3.2: Increasing levels of organization in protein structure. The figure shows how a short segment of 25 amino acids in the RNA polymerase II of S. cerevisiae folds into an α-helix, and subsequently integrates within a subunit of the complete protein, which is composed of ten different subunits.

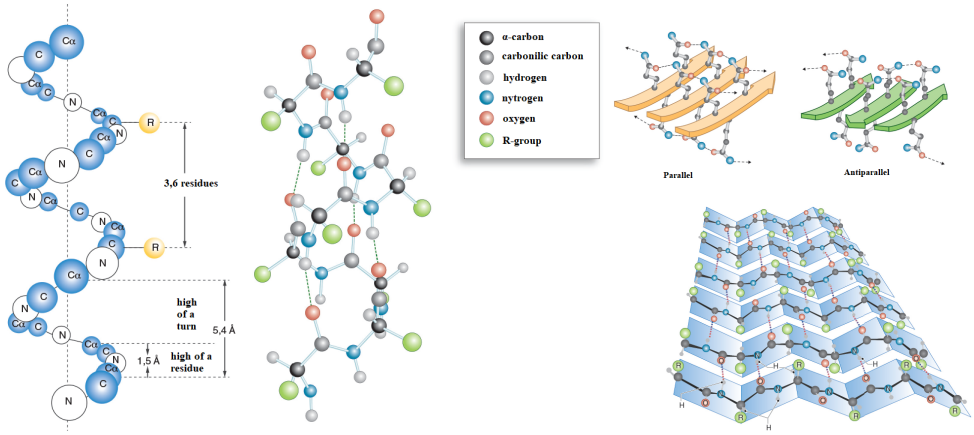
The primary structure dictates the higher levels of protein structure and ultimately its function.

Secondary Structure The secondary structure guides the *local* folding of the polypeptide chain through interactions between backbone[3] atoms. The most common secondary structures are the α-helix and the β-sheet. In both cases, hydrogen bonds[4] are crucial for the formation and stabilization of these structures. Furthermore, since secondary structures are localized, a single protein molecule can contain various regions with different secondary structures, increasing the complexity of the molecule. For this reason, secondary structures are usually represented in a simplified manner through the *ribbon representation*, where the backbone is shown as a strip to highlight the arrangement of secondary structural elements, such as alpha-helices and beta-sheets. In particular, α-helices are represented by coiled ribbons or thick tubes, β-sheets by arrows, and non-repetitive coils or loops by lines or thin tubes. This visualization method simplifies complex three-dimensional structures, making it easier to understand the overall folding and organization of the protein (example in Figure 3.5).

Tertiary Structure The tertiary structure is the overall three-dimensional shape of a single polypeptide chain, resulting from interactions between the R-groups of the amino acids. These interactions can be weak, such as hydrogen bonding and Van der Waals

[3]The backbone is the main chain of a protein molecule. It consists in a repeated pattern: N-C_α-C'-N-C_α-C'-N-C_α-C', and so on, where each unit corresponds to one amino acid. N is the nitrogen of the amide group, C_α is the alpha carbon, C' is the carbonyl carbon.

[4]A hydrogen bond is a weak chemical bond formed when a hydrogen atom covalently bonded to a highly electronegative atom (such as nitrogen, oxygen, or fluorine) interacts with another electronegative atom with a lone pair of electrons. Electronegativity is the tendency of an atom to attract electrons.



(a) α-helix structure **(b)** β-sheet structure

Figure 3.3: The secondary structures.

forces[5], or strong, such as ionic interactions and *disulfide bonds*. Disulfide bonds are covalent linkages formed between the sulfur atoms of two cysteine residues within a protein. Because of this, they are also known as SS-bonds, disulfide bridges, or cysteine bonds. They are essential for the formation of lasso motifs, as we will see in the next section. However, these interactions, in addition to being generally stronger than the hydrogen bonds that stabilize an α-helix or a β-sheet, are also geometrically more variable. As a result, the tertiary structures are far more complex and diverse than secondary structures in proteins.

The tertiary structure is crucial for a protein's functionality, as it determines the spatial arrangement of the active sites and other functional regions of the protein.

Quaternary Structure Some proteins are composed of more than one polypeptide chain. The quaternary structure refers to the arrangement and interaction of multiple polypeptide chains (subunits) in a protein. Hemoglobin, for example, is a protein with quaternary structure, composed of four subunits. The quaternary structure is stabilized by the same types of interactions that stabilize tertiary structure.

[5]Van der Waals forces are weak, non-covalent interactions that arise from transient dipole moments caused by the random movement of electrons, which induce dipoles in neighboring molecules. See [15] for details.

3.2 Protein topology: the lasso proteins

In recent years, it has become clear that the conventional framework of folding (with the subdivision between primary, secondary and tertiary structure in general) is not sufficient to describe at least 6% of the proteins in the Protein Data Bank (PDB)[6] [22]. These proteins exhibit complex entanglements, forming structures such as knots[7], slipknots[8] and non-trivial lassos. In this case, the nontrivial topology of the protein chain occurs when disulfide bonds (introduced in Section 3.1) or other kinds of bridges are positioned in such a way that a portion of the protein forms a closed loop (called covalent or cysteine loop [16]), through which another segment threads. In many cases, the threading occurs during the protein's folding process and remains locked for stability reasons.

Classification of complex lassos

Complex lasso proteins can be classified based on the number of piercings through the minimal surface[9] spanned by the covalent loop (classification introduced in [16]; see Figure 3.4). Specifically, four distinct classes of complex lasso proteins have been identified:

- L_n *class* (simple lasso): where the same tail pierces the surface n times;
- LS_n *class* (supercoiling lasso [18]: where one tail pierces the surface n times and winds the protein chain comprising the loop;
- $LL_{i,j}$ *class* (double lasso): where both tails pierce the surface i and j times, respectively;
- $LSL_{i,j}$ *class*: where one tail pierces the surface i times in a supercoiling manner, while the second tail pierces the surface in a simple manner.

Certain lasso motifs are associated with specific functions: L_1 is common in binding, antimicrobial, viral, and immune-related proteins, while L_2 is often found in signaling proteins. L_3 motifs frequently appear in transport proteins. Although supercoiling and LL structures are less common, they are frequently found in adhesion proteins (see Figure 3.5 for an example of supercoiling lasso).

[6]The Protein Data Bank (PDB) is a database for the three-dimensional structural data of large biological molecules such as proteins and nucleic acids.

[7]A knot is defined as embedding of the circle $\mathbb{S}^1$ in the 3-dimensional sphere $\mathbb{S}^3$

[8]Proteins that have knotted sub-chains despite their overall backbone chain being unknotted [12].

[9]A minimal surface is a surface with local minimal area (see [13] for details). This condition of a minimal area removes the ambiguity in a definition of such a surface

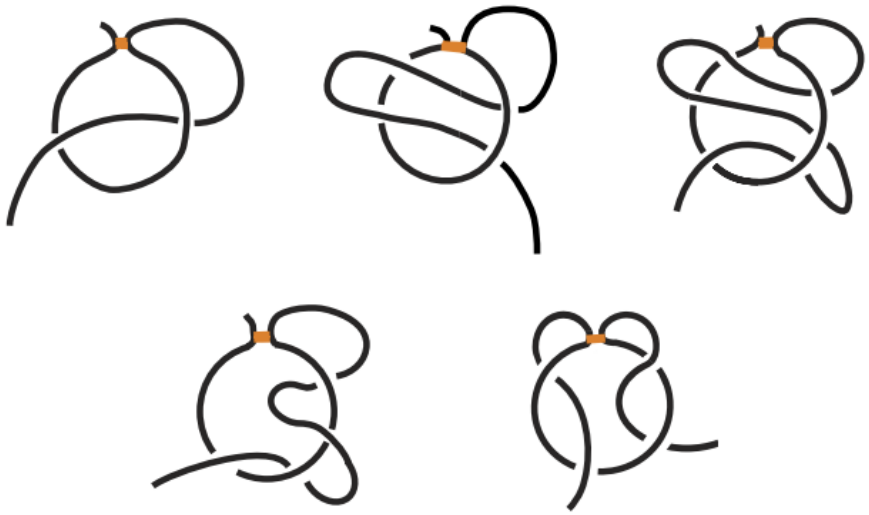


Figure 3.4: Various types of complex lasso motifs. Top row, left to right: L_1 (single lasso), L_2 (double lasso), L_3 (triple lasso); bottom row, left to right: L_S (supercoiling) and $LL_{1,1}$ (two-sided lasso).

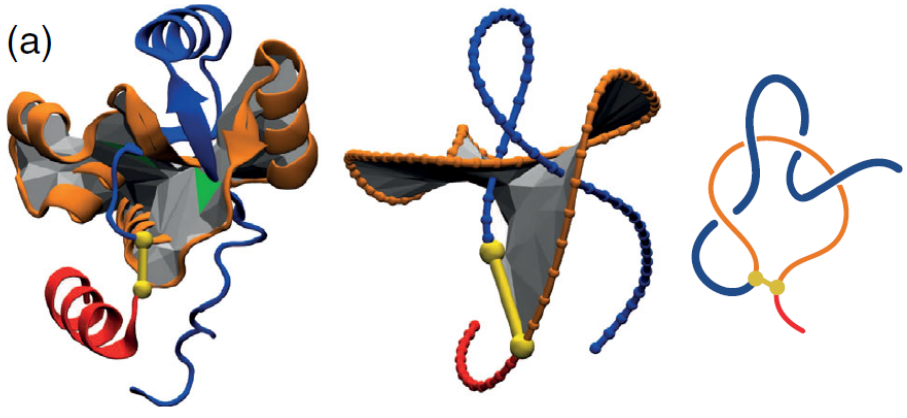


Figure 3.5: Protein 1zd0 in ribbon representation (left), with smoothed backbone (middle), and in schematic representation (right). The gray portion spanned by the protein's backbone in orange is closed by a disulfide bond (yellow) and is pierced twice in the same direction by the N terminal

Biological role and application

Around 18% of proteins with disulfide bridges have complex lasso [18], and they have been identified across various organisms, suggesting their evolutionary significance and potential functional diversity. Many bacterial lasso peptides, for example, have been identified as antibiotics due to their ability to inhibit enzymes or disrupt cellular processes in pathogens. Their tightly knotted structure is believed to contribute to their stability and resistance to proteolytic degradation, enhancing their longevity and effectiveness in hostile environments [18]. Similarly, lasso proteins found in eukaryotic systems have been implicated in signaling, transport, and regulatory roles, where their mechanical stability plays a critical role in maintaining functional integrity [16].

Identifying and studying lasso complexes is essential because their unique topology can have a big impact on a protein's stability, folding mechanisms, and biological functions. Their remarkable thermal and mechanical stability, along with their resistance to degradation, makes them great candidates for drug design, especially in developing peptide-based therapies that need to stay active over extended periods. Additionally, due to their distinct topology, lasso proteins are also being explored as scaffolds for molecular design, where their structure can be modified or functionalized for specific applications in nanotechnology and bioengineering.

3.3 Detecting lasso proteins

Persistent homology has demonstrated a wide range of applications, from biological systems [2, 24] to social sciences [20] and neuroscience [2, 10]. Recently, it has proven particularly effective for detecting lasso topologies in proteins, thanks to the introduction of a valuable algorithm [9] that is still being refined.

This algorithm offers a robust framework for identifying lasso topologies in proteins and presents several advantages over commonly used methods such as minimal surfaces analysis [17, 18]. To begin with, persistent homology is notably robust to noise, which is crucial given the variability and imperfections often present in protein structure data. It is also more computationally efficient compared to minimal surface methods and can identify topological features at multiple scales, making it well-suited for analyzing the complex, multi-scale nature of protein structures.

However, there are challenges as well. While the algorithm is generally faster and more efficient for simple and well-behaved structures compared to other methods, calculating persistent homology for large and complex protein structures can still be compu-

tationally intensive. Furthermore, interpreting results can be challenging, as comparing persistence diagrams is not always straightforward and may require careful interpretation and supplementary analysis.

A Theoretical Perspective on the Framework

Detecting lasso topologies in proteins using persistent homology involves several steps:

1. **Representation of the Protein Structure**. The first step is to represent the protein's three-dimensional structure as a point cloud, where each point corresponds to the position of an atom in the protein. This can be done by representing only the backbone atoms or including the side chains as well.

2. **Filtration Process**. From the point cloud, we construct a nested sequence of simplicial complexes, such as Vietoris-Rips or Čech complexes (see Chapter 2). The choice of filtration method significantly affects the detection of lasso structures, as it influences how the protein's topological features are revealed across different scales.

3. **Persistent Homology Computation**. Persistent homology is computed by tracking the birth and death of homological features as the filtration progresses. For lasso detection, the focus is typically on identifying persistent zero- or one-dimensional homology classes.

4. **Interpretation of Results**. The persistence diagram or barcode generated from the analysis is then interpreted. This is a crucial step, which will be explained in more detail in the next section. Intuitively, if the analyzed protein contains a lasso motif, an intersection occurs between the disk spanned by the covalent loop and one terminus of the protein. By sequentially adding atoms from the termini and comparing the persistence diagrams of the isolated covalent loop with the covalent loop plus the added atom, a *significant* difference between the diagrams will appear when the intersection occurs.

   What does it mean for persistence diagrams to be *significantly different*? Several criteria can be applied, such as comparing the lifetimes of one-dimensional homology classes or summing the lifetimes of each one-dimensional class along the filtration. However, the most effective method has been proven to be the bottleneck distance (see Section 2.3).

A simplified model: the decagon

To understand how the algorithm identifies lasso motifs, we analyze its behavior in a simplified two-dimensional model. We consider a set of points arranged as the vertices of a regular decagon, representing the atoms in a covalent loop. Specifically, we examine two cases: a decagon without a center, consisting only of the 10 vertices $\{A_1, \ldots, A_{10}\} = S$, and a decagon with an additional central point, which we denote by C (see Figures 3.6i.a and 3.6ii.a; both decagons are constructed from a circle with a radius of 4). The central point represents the intersection between the disk bounded by the covalent loop and one of the two ends of the protein, giving rise to the lasso motif. Note that, for simplicity, this point was placed at the center, but it could have been positioned anywhere inside the decagon.

From the point cloud, we constructed a nested sequence of Vietoris-Rips simplicial complexes by increasing the radius r of the balls centered at the vertices.[10]

Given the nature of the model, the persistent homology analysis, which tracks the life of homology classes (both 0-dimensional and 1-dimensional), was carried out intuitively (without matrix computations, which are required for more complex models). This analysis produced the barcodes and persistence diagrams shown in Figures 3.7 and 3.8. Let us now analyze the persistent homology of the two cases, referring to the snapshots in Figures 3.6i and 3.6ii, which show the process of constructing the Rips filtration.

Zero-Dimensional Persistent Homology

Decagon without the center (S) Initially, we have 10 disconnected vertices, each representing a separate connected component (Figure 3.6i.a). As the radius r increases, edges form between the vertices when the balls centered at each vertex intersect. The first edges appear at $r = \frac{2.47}{2}$ (see Figure 3.6i.c), which corresponds to the minimum distance between adjacent vertices on the decagon. All vertices eventually merge into a single connected component. The barcode for zero-dimensional homology (Figure 3.7.a) shows this merging process, with the single connected component persisting indefinitely (represented as a bar extending to infinity).

Decagon with the center ($S \cup \{C\}$) When the central point C is added, the number of connected components at $r = 0$ increases to 11 (see Figure 3.6ii.a). As before, edges between the vertices emerge at $r = \frac{2.47}{2}$, and the number of components reduces to two

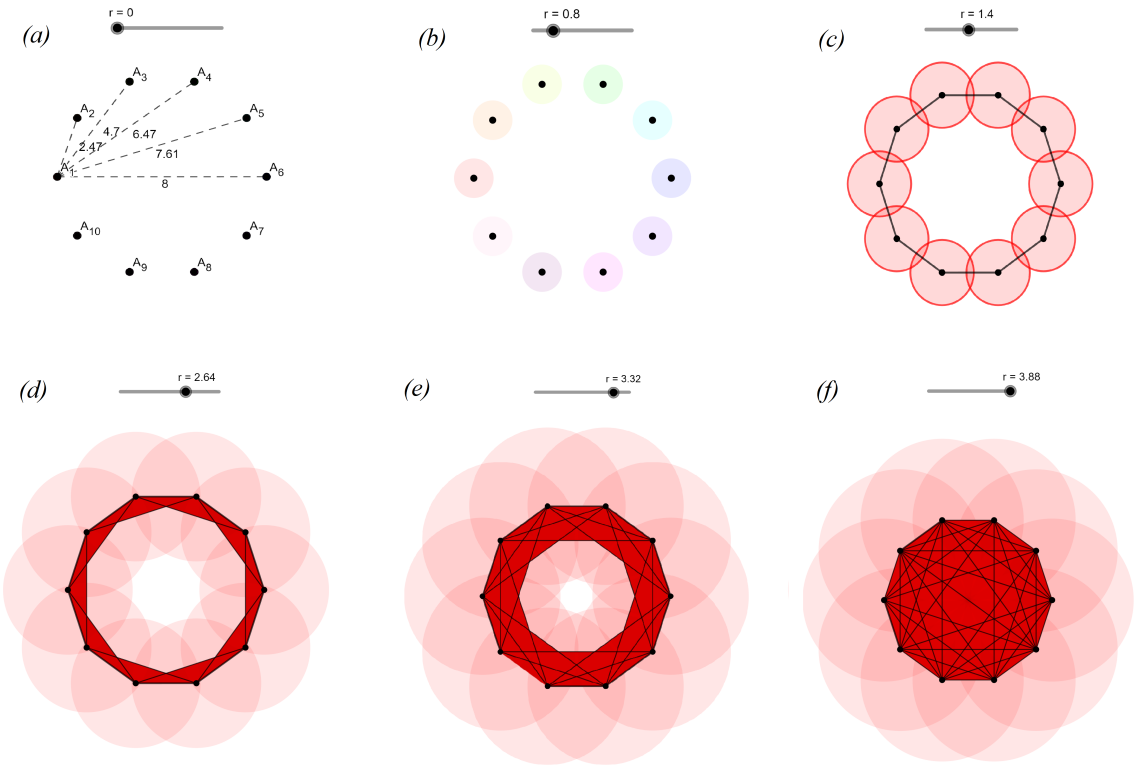
[10] Remember that, for a fixed $r \in [0, 4]$, a simplex $\sigma \in \mathrm{VR}(S, r)$ *iff* $\operatorname{diam}\sigma \leq 2r$ *iff* $\mathrm{d}_2(x, y) \leq 2r \quad \forall x, y \in \sigma$ (see Definition 1.41). Intuitively, a new edge emerges when two balls intersect.

(Figure 3.6ii.c). For $r \geq 2$, the balls around the vertices intersect with the ball around C, leading to the formation of new simplices and the eventual merging of the two remaining components (Figures 3.6ii.d and 3.6ii.e). The barcode for zero-dimensional homology (Figure 3.7.b) captures this progression.

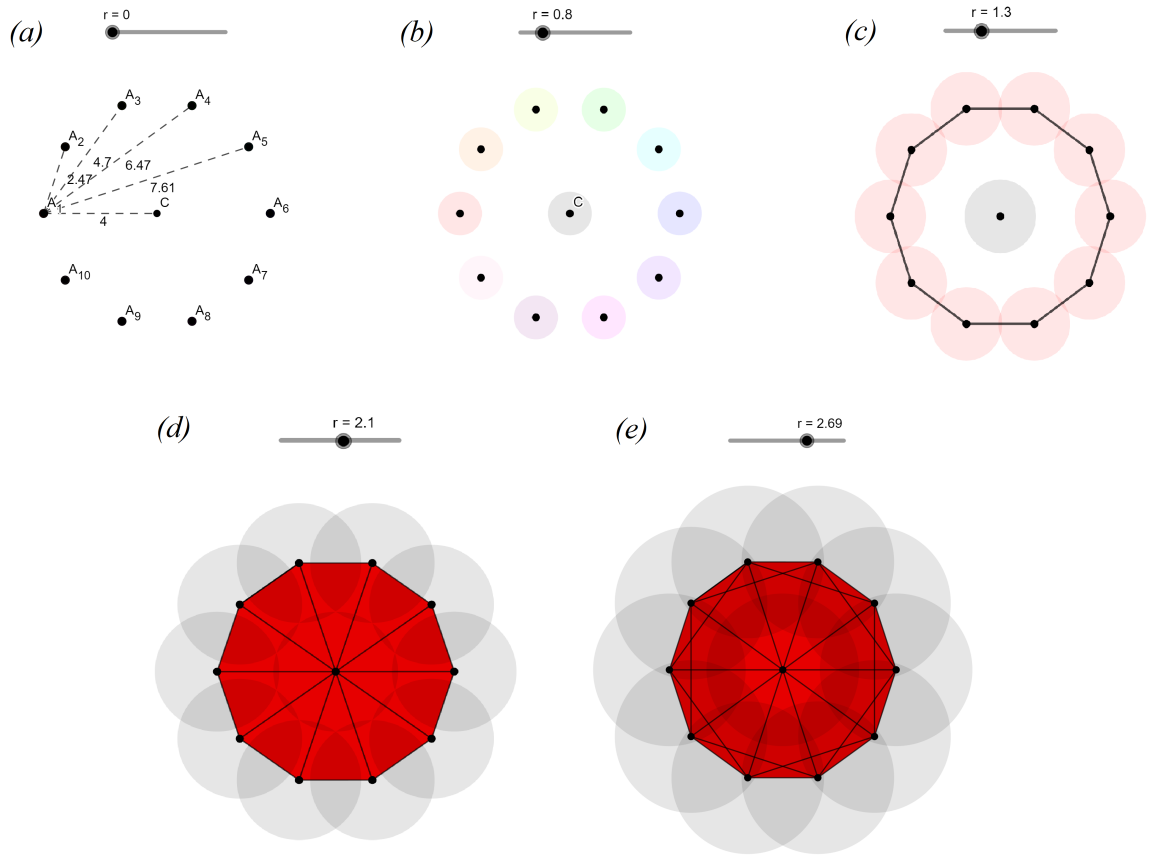
One-Dimensional Persistent Homology

Decagon without the center (S) As the edges between vertices emerge at $r = \frac{2.47}{2}$ (Figure 3.6i.c), a non-trivial loop forms. This loop persists until $r = \frac{7.61}{2}$, when the loop is "filled in" by newly added simplices (Figure 3.6i.f), at which point the one-dimensional homology class dies.

Decagon with the center ($S \cup \{C\}$) In this case, the central point influences the formation and lifespan of the one-dimensional homology classes. A loop forms at $r = \frac{2.47}{2}$ (Figure 3.6ii.c), but due to the central point, this loop is filled in much earlier (at $r = 2$), as simplices form between the vertices and the center (Figure 3.6ii.d). Thus, the one-



(i) (a) Points with their mutual distances, which serve as a reference for determining the values of r at which new complexes are added. (b) VR$(S, 0.8)$ (c) VR$(S, 1.3)$, where the edges of the decagon become apparent as $r \geq \frac{2.47}{2}$. (d) and (e) VR$(S, 2.64)$ and VR$(S, 3.32)$, where new simplices emerge, but the homology remains unchanged. (f) VR$(S, 3.38)$, where the union of the simplices forms a single convex connected component.



(ii) (a) Points with their mutual distances, as before. (b) Simplicial complex $\mathrm{VR}(S \cup \{C\}, 0.8)$, where each connected component is still separate. (c) $\mathrm{VR}(S \cup \{C\}, 1.3)$, where the edges of the decagon become apparent as $r \geq \frac{2.47}{2}$. (d) $\mathrm{VR}(S \cup \{C\}, 2.1)$, where the balls centered at the vertices and the center intersect, covering the central hole with new simplices. (e) $\mathrm{VR}(S \cup \{C\}, 2.69)$, where new simplices emerge, but they do not alter the homology of the complex.

Figure 3.6: Rips filtration for the decagon without (i) and with (ii) the center.

dimensional homology class dies earlier than in the case without the center, as reflected in the barcode (Figure 3.7.b).

In summary, the key differences between the two cases are as follows:

- *Zero-dimensional homology*: In both cases, all vertices eventually merge into a single connected component, but the case with the center initially has one additional component, affecting the merging process. This difference is visible in both the barcodes (Figure 3.7) and the persistence diagrams (Figures 3.8).

- *One-dimensional homology*: The main distinction lies in the lifespan of the non-trivial loop in one-dimensional homology. In the decagon without the center, the loop persists longer (until $r = \frac{7.61}{2}$, giving it a lifespan of 2.57). In contrast, in the

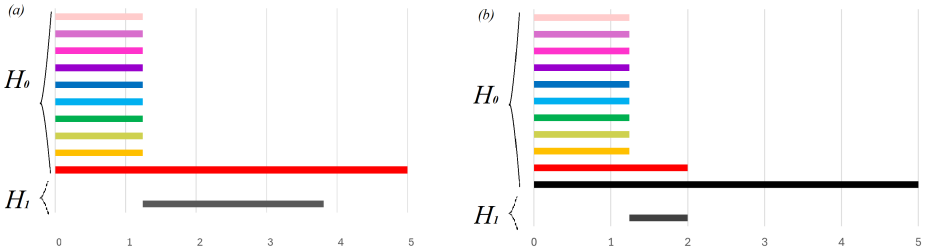


Figure 3.7: Barcodes of persistent homology. (a) Barcode related to zero-dimensional persistent homology. (b) Barcode related to one-dimensional persistent homology.

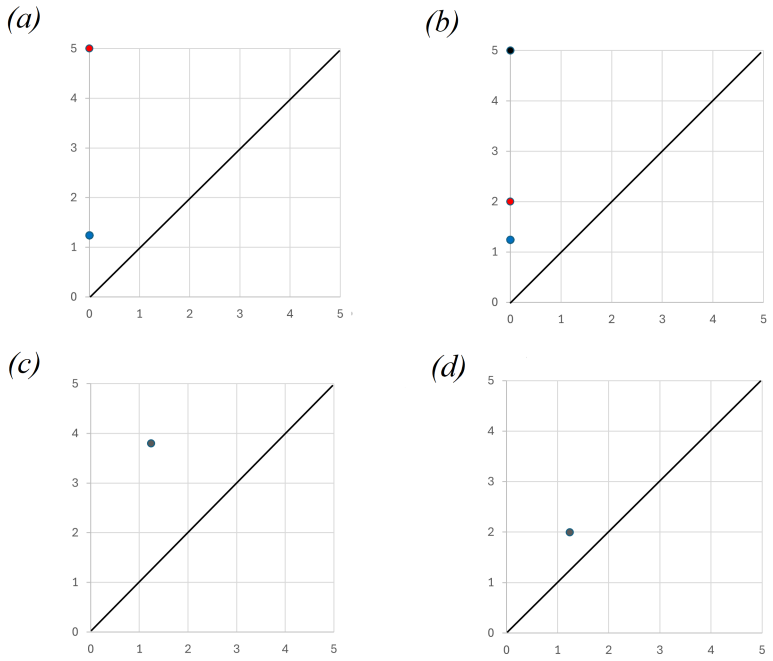


Figure 3.8: Persistence diagrams. (a) and (b) are the zero-dimensional persistence diagrams for the decagon with and without the center, respectively. (c) and (d) are the one-dimensional persistence diagrams for the decagon with and without the center, respectively.

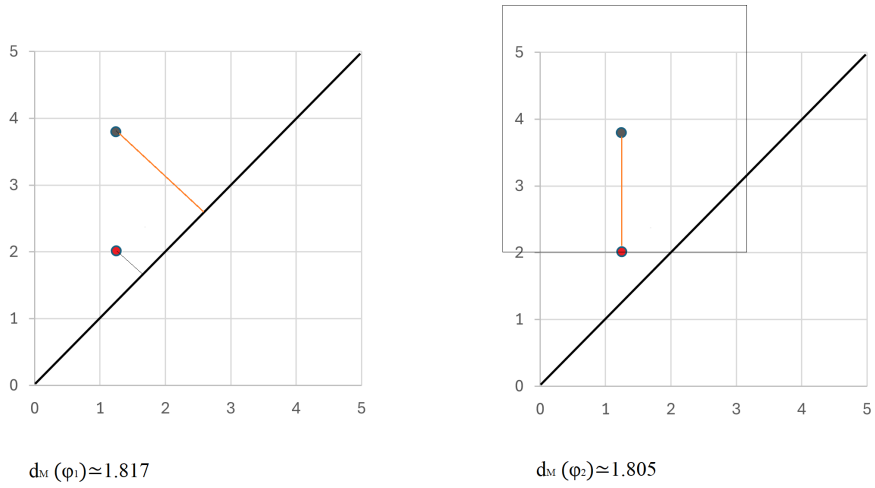


Figure 3.9: Bottleneck distance between persistence diagrams.

case with the center, the loop is filled in much earlier (at $r = 2$, giving it a lifespan of 0.76), due to the connections between the vertices and the central point.

To further highlight the differences, we can calculate the bottleneck distance between the persistence diagrams. Figure 3.9 shows an example of partial matchings between the persistence diagrams related to one-dimensional homology. According to the definition, the bottleneck distance between the two is represented by the diagram on the right.

Obviously, the framework we have just discussed is applied to much larger and more complex datasets, in 3 dimensions, by the algorithm we mentioned. See Figure 3.10, which shows a lasso protein identified through minimal surface analysis, to understand the complexity of the dataset typically used.

The algorithm

We conclude this section by showing the pipeline followed by the algorithm to detect lasso structures. More details regarding the implementation, optimization, and efficiency of this algorithm can be found in [9].

- The algorithm starts by loading the (x, y, z) coordinates of each atom in the protein from a provided file. It is possible to choose whether to include all atoms, including the side chains, or just the backbone of the protein.

- Based on information from the *Lassoprot* database[11], the protein is divided into

[11] A server and database dedicated to proteins with lasso structures. It allows users to analyze new

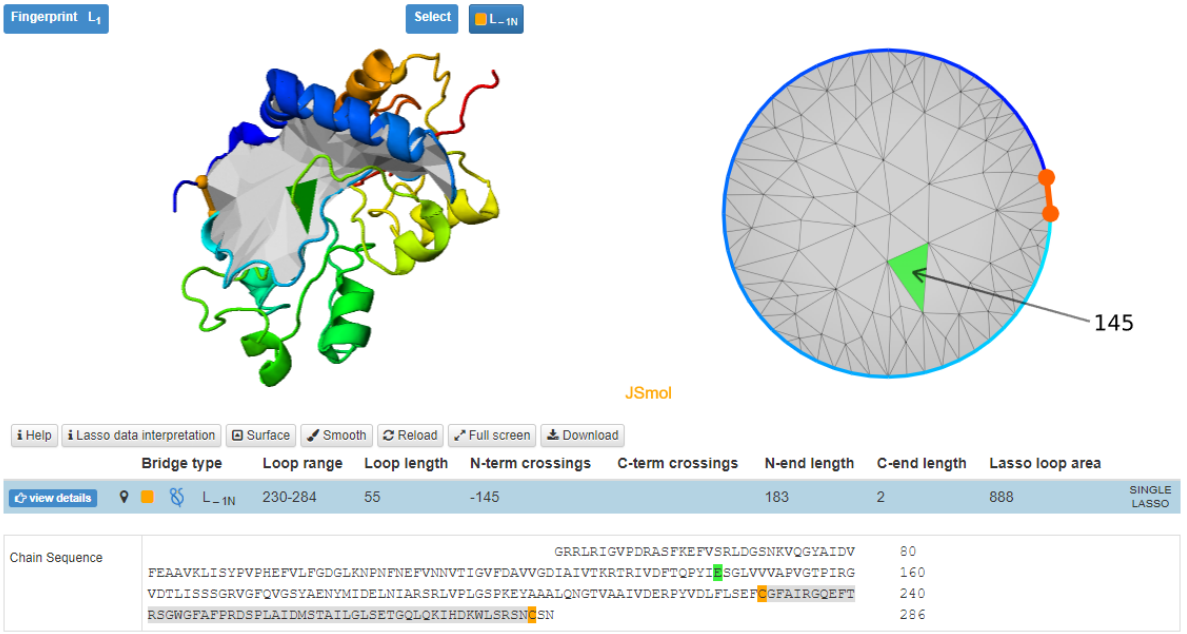


Figure 3.10: Structure of the glutamate-like receptor GLR3.2 ligand-binding domain in complex with glycine, on the left. Minimal surface spanned by the covalent loop and its triangulation, on the right. At the bottom, the chain sequence. The cysteines in orange delimit the covalent loop.

three distinct sections: the loop, the head, and the tail. This segmentation is crucial for isolating the loop for further analysis.

- To increase the resolution of the method, additional atoms (typically 1 or 2) are inserted between the original atoms in the loop. This step refines the structure and enhances the precision of the persistent homology calculations, thereby reducing the likelihood of false negatives.

- The persistent homology of the loop is computed independently.

- Each atom in the head and tail is considered individually. For each atom, the persistent homology of the combined set of the loop and that atom is calculated. This comparison allows for the detection of interactions between the head or tail and the loop.

- The bottleneck distance is then measured between the persistent homology of the loop alone and that of the loop combined with each individual atom from the head

protein structures and includes a comprehensive database with detailed information about proteins with lasso structures previously analyzed. The resource is accessible at the following link `https://lassoprot.cent.uw.edu.pl/`.

or tail. This distance quantifies the difference in topological features between the two scenarios.

- A graph of all these distances is constructed. The graph is then smoothed to remove noise and highlight significant changes.
- The maxima of the smoothed graph are identified, as these correspond to the points where the head or tail intersects with the loop.
- Finally, any maxima below a predefined threshold are discarded, ensuring that only the most significant intersection points are retained.

Bibliography

[1] G. Capranico. *Biologia molecolare*. Edises, 2016.

[2] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[3] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. In *Proceedings of the Twenty-First Annual Symposium on Computational Geometry*, SCG '05, page 263–271, New York, NY, USA, 2005. Association for Computing Machinery.

[4] T.K. Dey and Y. Wang. *Computational Topology for Data Analysis*. Computational Topology for Data Analysis. Cambridge University Press, 2022.

[5] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & computational geometry*, 28:511–533, 2002.

[6] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied Mathematics. American Mathematical Society, 2010.

[7] Patrizio Frosini. Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, volume 1607, pages 122–133. SPIE, 1992.

[8] Patrizio Frosini. Lecture notes in computational topology, April 2024.

[9] Boštjan Gabrovšek, Paolo Cavicchioli, Bartosz Greń, Žiga Virk, and Joanna Sułkowska. Topological analysis of protein lasso structures using persistent homology. *in development*, 2024.

[10] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.

[11] A. Hatcher. *Algebraic Topology*. Algebraic Topology. Cambridge University Press, 2002.

[12] Michal Jamroz, Wanda Niemyska, Eric J. Rawdon, Andrzej Stasiak, Kenneth C. Millett, Piotr Sułkowski, and Joanna I. Sulkowska. KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Research*, 43(D1):D306–D314, 10 2014.

[13] W. Meeks and J. Pérez. *A Survey on Classical Minimal Surface Theory*. University lecture series. American Mathematical Society, 2012.

[14] J.R. Munkres and J.W. Munkres. *Elements Of Algebraic Topology*. CRC Press, 2018.

[15] D.L. Nelson and M.M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 2017.

[16] Wanda Niemyska, Pawel Dabrowski-Tumanski, Michal Kadlof, Ellinor Haglund, Piotr Sułkowski, and Joanna Sulkowska. Complex lasso: New entangled motifs in proteins. *Scientific Reports*, 6:36895, 11 2016.

[17] Wanda Niemyska, Kenneth C Millett, and Joanna I Sulkowska. Gln: a method to reveal unique properties of lasso type topology in proteins. *Scientific reports*, 10(1):15186, 2020.

[18] Szymon Niewieczerzał and Joanna I Sulkowska. Supercoiling in a protein increases its stability. *Physical review letters*, 123(13):138102, 2019.

[19] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1), August 2017.

[20] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Topological strata of weighted complex networks. *PLOS ONE*, 8(6):1–8, 06 2013.

[21] V. Robin. Towards computing homology from finite approximations. In *Topology Proceedings*, volume 24, page 503–532, Summer 1999.

[22] Joanna Ida Sulkowska. On folding of entangled proteins: knots, lassos, links and θ-curves. *Current Opinion in Structural Biology*, 60:131–141, 2020. Folding and Binding ∘ Proteins.

[23] Ž. Virk. *Introduction to Persistent Homology*. Založba UL FRI, 2022.

[24] 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.

[25] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 347–356, 2004.