

Supplementary Methods for Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models

Yuchi Qiu¹ and Guo-Wei Wei^{*1,2,3}

¹*Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA*

²*Department of Biochemistry and Molecular Biology, Michigan State University, MI, 48824, USA*

³*Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA*

1 Mathematical theory of topological data analysis (TDA)

1.1 Simplicial complex and chain complex

Graph is a representation for a point cloud consisting of vertices and edges for modeling pairwise interactions, such as atoms and bonds in molecules. Simplicial complex, the generalization of graph, constructs more enriched shapes to include high dimensional objects. A simplicial complex is composed of simplexes up to certain dimensions. A k -simplex, σ^k , is a convex hull of $k+1$ affinely independent points $v_0, v_1, v_2, \dots, v_k$:

$$\sigma^k := [v_0, v_1, v_2, \dots, v_k] = \left\{ \sum_{i=0}^k \lambda_i v_i \mid \sum_{i=0}^k \lambda_i = 1; \lambda_i \in [0, 1], \forall i \right\}. \quad (1)$$

In Euclidean space, 0-simplex is a point, 1-simplex is an edge, 2-simplex is a triangle, and 3-simplex is a tetrahedron. The k -simplex can describe abstract simplex for $k > 3$.

A subset of the $k+1$ vertices of a k -simplex, σ^k , with $m+1$ vertices forming a convex hull in a lower dimension and is called an m -face of the k -simplex σ^k , denoted as $\sigma^m \subset \sigma^k$. A simplicial complex K is a finite collection of simplexes satisfying two conditions:

- 1) Any face of a simplex in K is also in K .
- 2) The intersection of any two simplexes in K is either empty or a shared face.

The interactions between two simplexes can be described by adjacency. For example, in graph theory, two vertices (0-simplexes) are adjacent if they share a common edge (1-simplex). Adjacency for k -simplexes with $k > 0$ includes both upper and lower adjacency. Two distinct k -simplexes, σ_1 and σ_2 , in K are upper adjacent, denoted $\sigma_1 \sim_U \sigma_2$, if both are faces of a $(k+1)$ -simplex in K , called a common upper simplex. Two distinct k -simplexes, σ_1 and σ_2 , in K are lower adjacent, denoted $\sigma_1 \sim_L \sigma_2$, if they share a common $(k-1)$ -simplex as their face, called a common lower simplex. Either common upper simplex or common lower simplex is unique for two upper or lower

*Corresponding author: weig@msu.edu

adjacent simplexes. The upper degree of a k -simplex, $\deg_U(\sigma^k)$, is the number of $(k+1)$ -simplexes in K of which σ^k is a face; the lower degree of a k -simplex, $\deg_L(\sigma^k)$, is the number of nonempty $(k-1)$ -simplexes in K that are faces of σ^k , which is always $k+1$. The degree of k -simplex ($k > 0$) is defined as the sum of its upper and lower degree

$$\deg(\sigma^k) = \deg_U(\sigma^k) + \deg_L(\sigma^k) = \deg_U(\sigma^k) + k + 1. \quad (2)$$

For $k = 0$, the degree of a vertex is:

$$\deg(\sigma^0) = \deg_U(\sigma^0). \quad (3)$$

A simplex has orientation determined by the ordering of its vertices, except 0-simplex. For example, clockwise and anticlockwise orderings of three vertices determine the two orientation of a triangle. Two simplexes, σ_1 and σ_2 , defined on the same vertices are similarly oriented if their orderings of vertices differ from an even number of permutations, otherwise, they are dissimilarly oriented.

Algebraic topology provides a tool to calculate simplicial complex. A k -chain is a formal sum of oriented k -simplexes in K with coefficients on \mathbb{Z} . The set of all k -chains of simplicial complex K together with the addition operation on \mathbb{Z} constructs a free Abelian group $C_k(K)$, called chain group. To link chain groups from different dimensions, the k -boundary operator, $\partial_k : C_k(K) \rightarrow C_{k-1}(K)$, maps a k -chain in the form of a linear combination of k -simplexes to the same linear combination of the boundaries of the k -simplexes. For a simple example where the k -chain has one oriented k -simplex spanned by $k+1$ vertices as defined in Eq. (1), its boundary operator is defined as the formal sum of its all $(k-1)$ -faces:

$$\partial_k \sigma^k = \sum_{i=0}^k (-1)^i \sigma_i^{k-1} = \sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k], \quad (4)$$

where $\sigma_i^{k-1} = [v_0, \dots, \hat{v}_i, \dots, v_k]$ is the $(k-1)$ -simplex with its vertex v_i being removed. The most important topological property is that a boundary has no boundary: $\partial_{k-1} \partial_k = \emptyset$.

A sequence of chain groups connected by boundary operators defines the chain complex:

$$\dots \xrightarrow{\partial_{n+1}} C_n(K) \xrightarrow{\partial_n} C_{n-1}(K) \xrightarrow{\partial_{n-1}} \dots \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} \emptyset. \quad (5)$$

When n exceeds the dimension of K , $C_n(K)$ is an empty vector space and the corresponding boundary operator is a zero map.

1.2 Filtration for multiscale chain complexes

Filtration is a process that constructs a nested sequence of simplicial complex allowing a multiscale analysis of the point cloud. It creates a family of simplicial complexes ordered by inclusion (Figure 2c):

$$\emptyset = K^{t_0} \subseteq K^{t_1} \subseteq \dots \subseteq K^{t_n} = K. \quad (6)$$

where K is the largest simplicial complex can be obtained from the point cloud.

The filtration induces a sequence of chain complexes

$$\begin{array}{ccccccc}
\cdots & \xrightarrow[\partial_{k+2}^{t_1*}]{\partial_{k+2}^{t_1}} & C_{k+1}^{t_1} & \xrightarrow[\partial_{k+1}^{t_1*}]{\partial_{k+1}^{t_1}} & C_k^{t_1} & \xrightarrow[\partial_k^{t_1*}]{\partial_k^{t_1}} & \cdots \xrightarrow[\partial_1^{t_1*}]{\partial_1^{t_1}} C_0^{t_1} \xrightarrow[\partial_0^{t_1*}]{\partial_0^{t_1}} \emptyset \\
& & \cap & & \cap & & \cap \\
\cdots & \xrightarrow[\partial_{k+2}^{t_2*}]{\partial_{k+2}^{t_2}} & C_{k+1}^{t_2} & \xrightarrow[\partial_{k+1}^{t_2*}]{\partial_{k+1}^{t_2}} & C_k^{t_2} & \xrightarrow[\partial_k^{t_2*}]{\partial_k^{t_2}} & \cdots \xrightarrow[\partial_1^{t_2*}]{\partial_1^{t_2}} C_0^{t_2} \xrightarrow[\partial_0^{t_2*}]{\partial_0^{t_2}} \emptyset \\
& & \cap & & \cap & & \cap \\
& & \vdots & & \vdots & & \vdots \\
& & \cap & & \cap & & \cap \\
\cdots & \xrightarrow[\partial_{k+2}^{t_n*}]{\partial_{k+2}^{t_n}} & C_{k+1}^{t_n} & \xrightarrow[\partial_{k+1}^{t_n*}]{\partial_{k+1}^{t_n}} & C_k^{t_n} & \xrightarrow[\partial_k^{t_n*}]{\partial_k^{t_n}} & \cdots \xrightarrow[\partial_1^{t_n*}]{\partial_1^{t_n}} C_0^{t_n} \xrightarrow[\partial_0^{t_n*}]{\partial_0^{t_n}} \emptyset
\end{array} \tag{7}$$

where $C_k^t = C_k(K^t)$ is the chain group for subcomplex K^t , and its k -boundary operator is $\partial_k^t : C_k(K^t) \rightarrow C_{k-1}(K^t)$. ∂_k^t is the co-boundary operator. Associated with the k -boundary operator, its adjoint operator is the k -adjoint boundary operator, $\partial_k^{t*} : C_{k-1}(K^t) \rightarrow C_k(K^t)$.

There are various simplicial complex that can be used to construct the filtration, such as Rips complex, Čech complex, and Alpha complex. For example, the Rips complex of K with radius t consists of all simplexes with diameter at most $2t$:

$$V(t) = \{\sigma \subseteq K \mid \text{diam}(\sigma) \leq 2t\}. \tag{8}$$

1.3 Homology group and persistent homology

With the chain complex defined in Eq. (5), the k -cycle and k -boundary groups are defined as:

$$\begin{aligned}
Z_k &= \ker \partial_k = \{c \in C_k \mid \partial_k c = 0\} \\
B_k &= \text{im } \partial_{k+1} = \{\partial_{k+1} c \mid c \in C_{k+1}\}
\end{aligned} \tag{9}$$

Then the k -th homology group H_k is defined as

$$H_k = Z_k / B_k. \tag{10}$$

The k -th Betti number, β_k , is defined by the rank of k -th homology group H_k which counts k -dimensional holes. For example, $\beta_0 = \text{rank}(H_0)$ reflects the number of connected components, $\beta_1 = \text{rank}(H_1)$ reflects the number of loops, and $\beta_2 = \text{rank}(H_2)$ reveals the number of voids or cavities.

Persistent homology is devised to track the multiscale topological information along the filtration [1]. The inclusion map $K_i \subseteq K_j$ induces a homomorphism $f_k^{i,j}$ between homology groups $H_k(K_{t_i}) \rightarrow H_k(K_{t_j})$ for each dimension k . The p -persistent k -th homology group of K_t is defined by

$$H_k^{t,p} = Z_k^t / (B_k^{t+p} \cap Z_k^t), \tag{11}$$

where $Z_k^t = \ker \partial_k^t$ and $B_k^{t+p} = \text{im } \partial_{k+1}^{t+p}$. Intuitively, this homology group records the k -dimensional homology classes of K_t that are persistent at least until K_{t+p} . The birth and death of homology classes can be represented by a barcode, a set of intervals (Figure 2d).

1.4 Combinatorial Laplacian.

For k -boundary operator $\partial_k : C_k \rightarrow C_{k-1}$ in K , let \mathcal{B}_k be the matrix representation of this operator relative to the standard bases for C_k and C_{k-1} in K . $\mathcal{B}_k \in \mathbb{Z}^{M \times N}$ is the matrix representation of boundary operator under the standard bases $\{\sigma_i^k\}_{i=1}^N$ and $\{\sigma_j^{k-1}\}_{j=1}^M$ of C_k and C_{k-1} . Associated with the boundary operator ∂_k , the adjoint boundary operator is $\partial_k^* : C_{k-1} \rightarrow C_k$, where its matrix representation is the transpose of the matrix, \mathcal{B}^T , with respect to the same ordered bases to the boundary operator.

The k -combinatorial Laplacian, a topological Laplacian, is a linear operator $\Delta_k : C_k(K) \rightarrow C_k(K)$

$$\Delta_k := \partial_{k+1} \partial_{k+1}^* + \partial_k^* \partial_k, \quad (12)$$

and its matrix representation, L_k , is given by

$$L_k = \mathcal{B}_{k+1} \mathcal{B}_{k+1}^T + \mathcal{B}_k^T \mathcal{B}_k. \quad (13)$$

In particular, the 0-combinatorial Laplacian (i.e. graph Laplacian) is given as follows since ∂_0 is an zero map:

$$L_0 = \mathcal{B}_1 \mathcal{B}_1^T. \quad (14)$$

The elements of k -combinatorial Laplacian matrices are

$$(L_k)_{i,j} = \begin{cases} \deg(\sigma_i^k), & \text{if } i = j \\ 1, & \text{if } i \neq j, \sigma_i^k \sim_U \sigma_j^k \text{ and } \sigma_i^k \sim_L \sigma_j^k \text{ with similar orientation} \\ -1, & \text{if } i \neq j, \sigma_i^k \sim_U \sigma_j^k \text{ and } \sigma_i^k \sim_L \sigma_j^k \text{ with dissimilar orientation} \\ 0, & \text{if } i \neq j, \text{ either } \sigma_i^k \sim_U \sigma_j^k \text{ or } \sigma_i^k \sim_L \sigma_j^k. \end{cases} \quad (15)$$

For $k = 0$, the graph Laplacian matrix L_0 is

$$(L_0)_{i,j} = \begin{cases} \deg(\sigma_i^0), & \text{if } i = j \\ -1, & \text{if } i \neq j, \sigma_i^0 \sim_U \sigma_j^0 \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

The multiplicity of zero spectra of L_k gives the Betti- k number, according to combinatorial Hodge theorem [2]:

$$\beta_k = \dim(L_k) - \text{rank}(L_k) = \text{null}(L_k). \quad (17)$$

The Betti numbers describe topological invariants. Specifically, β_0 , β_1 , and β_2 may be regarded as the numbers of independent components, rings, and cavities, respectively.

1.5 Persistent spectral graph (PSG)

The homotopic shape changes with a small increment of filtration parameter may be subject to noise from the data. The persistence may be considered to enhance the robustness when calculating the Laplacian. First, we define the p -persistent chain group $\mathbb{C}_k^{t,p} \subseteq C_k^{t+p}$ whose boundary is in C_{k-1}^t :

$$\mathbb{C}_k^{t,p} = \left\{ \alpha \in C_k^{t+p} \mid \partial_k^{t+p}(\alpha) \in C_{k-1}^t \right\}, \quad (18)$$

where $\partial_k^{t+p} : C_k^{t+p} \rightarrow C_{k-1}^{t+p}$ is the k -boundary operator for chain group C_k^{t+p} . Then we can define a p -persistent boundary operator, $\partial_k^{t,p}$, as the restriction of ∂_k^{t+p} on the p -persistent chain group $C_k^{t,p}$:

$$\partial_k^{t,p} = \partial_k^{t+p}|_{C_k^{t,p}} : C_k^{t,p} \rightarrow C_{k-1}^t. \quad (19)$$

Then PSG defines a family of p -persistent k -combinatorial Laplacian operators $\Delta_k^{t,p} : C_k(K_t) \rightarrow C_k(K_t)$ [3, 4] which is defined as

$$\Delta_k^{t,p} = \partial_{k+1}^{t,p} \left(\partial_{k+1}^{t,p} \right)^* + \left(\partial_k^t \right)^* \partial_k^t. \quad (20)$$

We denote $\mathcal{B}_{k+1}^{t,p}$ and \mathcal{B}_k^t as the matrix representations for boundary operators $\partial_{k+1}^{t,p}$ and ∂_k^t , respectively. Then the Laplacian matrix for $\Delta_k^{t,p}$ is

$$\mathcal{L}_k^{t,p} = \mathcal{B}_{k+1}^{t,p} \left(\mathcal{B}_{k+1}^{t,p} \right)^T + \left(\mathcal{B}_k^t \right)^T \mathcal{B}_k^t. \quad (21)$$

Since the Laplacian matrix, $\mathcal{L}_k^{t,p}$, is positive-semidefinite, its spectra are all real and non-negative

$$S_k^{t,p} = \text{Spectra}(\mathcal{L}_k^{t,p}) = \{(\lambda_1)_k^{t,p}, (\lambda_2)_k^{t,p}, \dots, (\lambda_N)_k^{t,p}\}, \quad (22)$$

where N is the dimension of a standard basis for C_k^t , and $\mathcal{L}_k^{t,p}$ has dimension $N \times N$. The k -persistent Betti number $\beta_k^{t,p}$ can be obtained from the multiplicity of harmonic spectra of $\mathcal{L}_k^{t,p}$:

$$\beta_k^{t,p} = \dim(\mathcal{L}_k^{t,p}) - \text{rank}(\mathcal{L}_k^{t,p}) = \text{null}(\mathcal{L}_k^{t,p}) = \#\{i | (\lambda_i)_k^{t,p} \in S_k^{t,p}, \text{ and } (\lambda_i)_k^{t,p} = 0\}. \quad (23)$$

In addition, the rest of the spectra, i.e., the non-harmonic part, capture additional geometric information. The family of spectra of the persistent Laplacians reveals the homotopic shape evolution [5].

References

- [1] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.
- [2] Beno Eckmann. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255, 1944.
- [3] Rui Wang, Duc Duy Nguyen, and Guo-Wei Wei. Persistent spectral graph. *International journal for numerical methods in biomedical engineering*, 36(9):e3376, 2020.
- [4] Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Persistent laplacians: Properties, algorithms and implications. *SIAM Journal on Mathematics of Data Science*, 4(2):858–884, 2022.
- [5] Yuchi Qiu and Guo-Wei Wei. Persistent spectral theory-guided protein engineering. *Nature Computational Science*, pages 1–15, 2023.