Types of Topology Data Analysis

Introduction of TDA

Advanced mathematics models, including algebraic topology, differential geometry and algebraic graph theory, have been proposed for **topology data analysis (TDA)** in recent years. Several invariants, such as topological invariant (Betti numbers), geometric invariant (curvatures) and algebraic graph invariant (eigenvalues) are considered in TDA . The combination of these invariants with learning models has achieved great successes in various aspects. For instance, persistent homology can be applied in distinguishing the morphological differences of objects and in drug design, including protein-ligand binding affinity prediction, toxicity prediction, drug discovery.

Topological networks are a framework for machine learning. Topological networks represent data by gathering similar data points into nodes, with an edge connecting the nodes if the corresponding set has a common data point.

Advantages

(1) A new method to extract data features and classifying big data.

(2) As an emerging mathematical method, it can be applied to many fields.

Challenges

(1) Geometric information is difficult to quantify.

(2) Non-geometric or non-topological information is neglected.

Prospects

(1) Embedding other information in topological invariants.

(2) Combine more geometry methods and topological means together.

1. Persistent homology

Introduction

Persistent homology, a method for studying topological features over changing scales, has received tremendous attention in the past decade. The basic idea is to measure the life cycle of topological features within a filtration, i.e., a nested family of abstract simplicial complexes, such as Vietoris-Rips complexes, alpha complexes. Thus, long-lived topological characteristics, which are often the intrinsic invariants of the underlying system, can be extracted.

Application fields

As an efficient tool to unveil topological invariants, persistent homology has been applied to various fields, such as image analysis, complex network, data analysis, geometric processing, and computational biology.

Advantages

It is a new method to extract data features and classifying big data. As an emerging mathematical method , it can be applied to many fields.

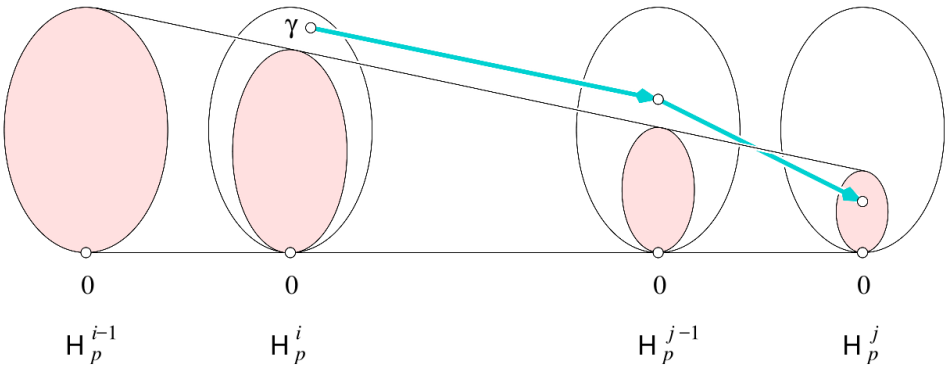
Disadvantages

It doesn't do a good job of describing the situation where the points interact with each other. Moreover, this approach loses a lot of information about the point cloud data besides the topology information.

Basic principles

Given a point cloud data, i.e., a point set $S \subset R^n$ without additional structure, a filtration is constructed that grows a solid ball centered at each point with an ever-increasing radius r. As the value of r increases, the solid balls will grow and simplices can be defined through the overlaps among the set of balls. There are various ways of constructing abstract simplicial complexes from the intersection patterns of the set of expanding balls, such as Vietoris-Rips complex and alpha complex. Persistent homoloy is introduced to describe toplogical features of these complex by using topological invariants such as Betti numbers.

For instance, the topological characteristics of $3D$ objects typically include connected components, tunnels or rings, and cavities or voids, which are invariant under the non-degenerate deformation of the structure. Homology characterizes such structures as groups, whose generators can be considered independent components, tunnels, cavities, etc. Their times of “birth” and “death” can be measured by a function associated with the filtration, calculated with ever more efficient computational procedures, and further visualized through barcodes, a series of horizontal line segments with the x-axis representing the changing scale of filtration and the y-axis representing the index of the homology generators.



2. Weighted persistent homology

Weighted persistent homology (WPH) models have been proposed to incorporate physical, chemical and biological or other properties into topological modeling. They can also be designed to characterize local topological information and certain special interaction patterns. WPH models contain localized persistent homology (LPH) and interactive persistent homology (IPH). This method assigns a value to the distance between two points, taking into account the interactions between points, , which improves the effectiveness of topological data analysis.

(1) Localized persistent homology

In LPH, the structure is decomposed into a series of local domains or regions, that may overlap with each other, and persistent homology analysis is then systematically applied on part (or all) of these local domains ore regions.

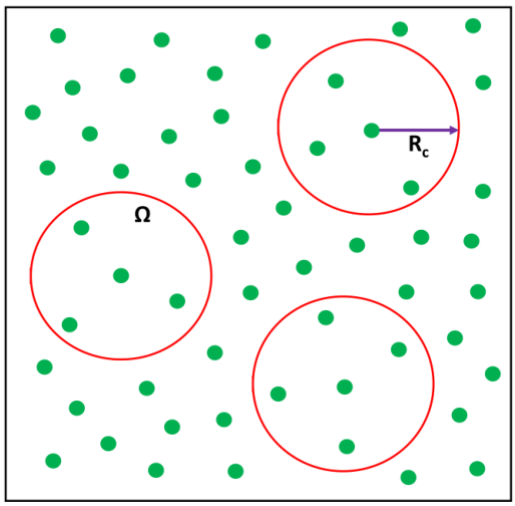


Figure 1: A schematic representation of the localized persistent homology (LPH) analysis. The red circle (a sphere in 3D) specifies the local region. Persistent homology is carried out on all the molecules within the local region.

(2) Interactive persistent homology

Interactive persistent homology (IPH) is proposed to study the topological invariant of the interaction networks formed between points. For instance, for a protein-ligand complex, an interaction matrix can be built with its element as the Euclidean distance between two atoms. However, if two atoms come from the some molecule (either protein or ligand), its distance is set to infinity, meaning they will never interact in IPH.

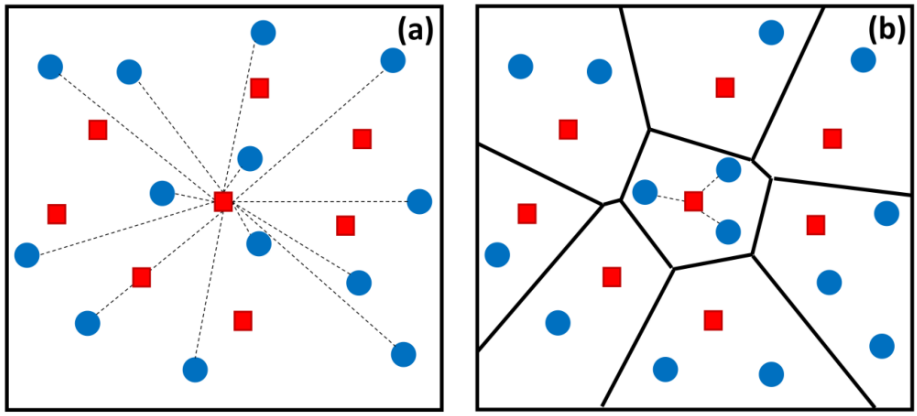


Figure 2. Illustration of interactions in global-scale and local-scale interactive persistent homology (IPH). In global-scale IPH model, an osmolyte molecule (red rectangle) interacts with all the water molecules (blue dots). In local-scale model, an osmolyte molecule (red rectangle) interact only with water molecules (blue dots) in its Voronoi cell.

figure 3. The left is a LPH model, the right is a IPH model describing the pairing of genes.

3. Persistent spectra graph

Mathematically, graph, simplicial complex and hypergraph and three topological models for structure characterization. Based on them, spectral graph theory, spectral simplicial complex and spectral hypergraph are proposed.

In **spectral graph**, Laplacian matrixes are proposed as the algebraic description of graphs. The eigen spectral information of the Laplacian matrix can then be used in characterization of graph properties.

Application fields

Chemistry and biomolecular research, drug design.

Advantages

The interaction between two points makes the data analysis more reasonable.

Basic principles

Graph structure encodes inter-dependencies among constituents and provides low-dimensional representations of high-dimensional datasets. One of the representations frequently used in spectral graph theory (SGT) is to associate graphs with matrices, such as the Laplacian matrix and adjacency matrix. Analyzing the spectra from such matrices leads to the understanding of the topological and spectral properties of the graph.
Let V be the vertex set, and E be the edge set. For a given simple graph $G(V, E)$ (A simple graph can be either connected or disconnected), the degree of the vertex $v \in V$ is the number of edges that are adjacent to v, denoted $\deg(v)$. The adjacency matrix $\mathcal{A}$ is defined by

$$\mathcal{A}(G) = \begin{cases} 1 & \text{if } v_i \text{ and } v_j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

and the Laplacian matrix $\mathcal{L}$ is given by

$$\mathcal{L}(G) = \begin{cases} \deg(v_i) & \text{if } v_i = v_j, \\ -1 & \text{if } v_j \text{ and } v_j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

figure 4

4. Persistent spectra simplicial complex

In **spectral simplicial complex**, combinatorial Laplacians or Hodge Laplacians can be defined from boundary matrixes, which characterize the topological connection between low-dimensional simplexes and high-dimensional simplexes. Essentially, graph Laplacian describes the relation between 1-simplexes (edges) and 0-simplexes (vertices), while combinatorial Laplacians are the generalization of the relation to higher-dimensional simplexes, such as, 2-simplexes (triangles), 3-simplexes (tetrahedrons), etc.

Application fields

Analyzing the stability of chemical molecular structure.

Advantages

It can be used to describe the interaction of high dimensional point cloud data.

Basic principles

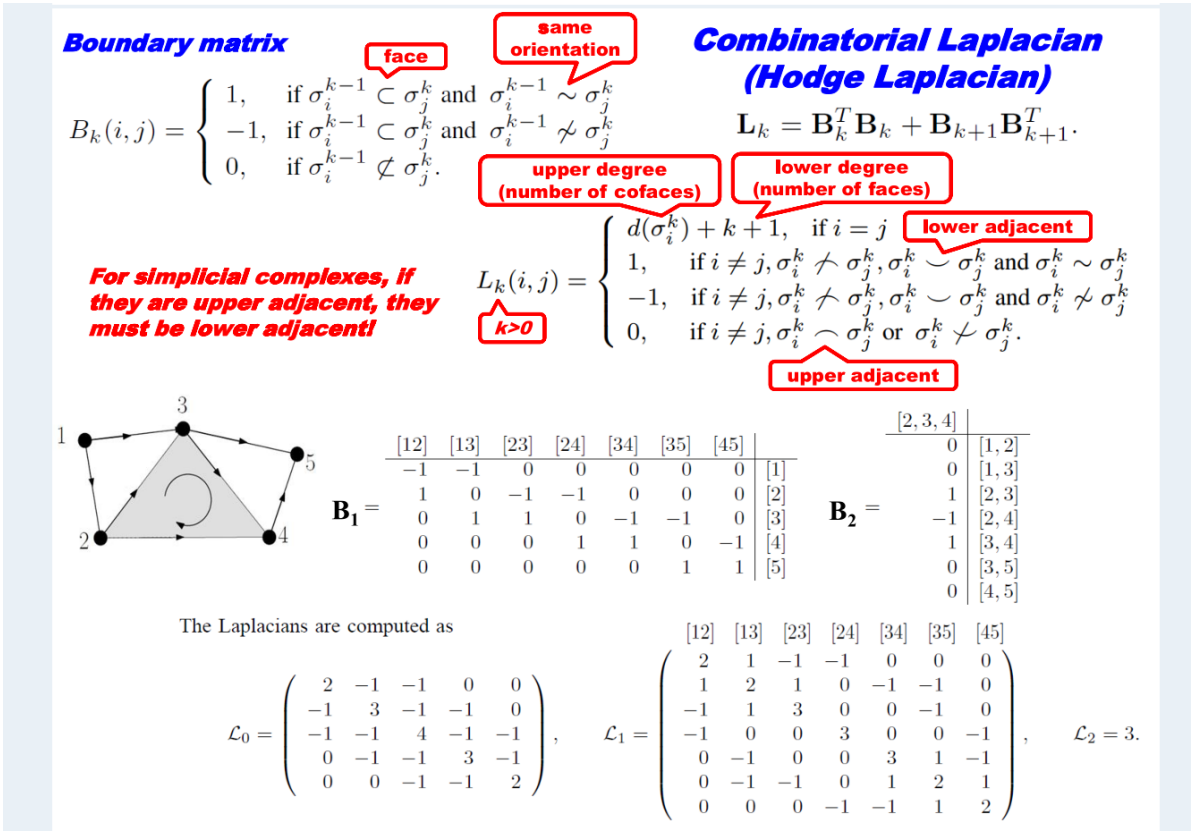


figure 5

5. Persistent spectra hypergraph

A hypergraph is a generalization of a graph in which an edge can join any number of vertices. Formally, **an undirected hypergraph** H is a pair $H = (X, E)$ where X is a set of elements called nodes or vertices, and E is a set of non-empty subsets of X called hyperedges or edges. The size of the vertex set is called the order of the hypergraph, and the size of edges set is the size of the hypergraph. **A directed hypergraph** differs in that its hyperedges are not sets, but an ordered pair of subsets of X, constituting the tail and head of the hyperedge.

In spectral hypergraph, boundary matrix or incident matrix can be defined between hyperedges and vertices. Hypergraph Laplacian matrix can then be constructed from the incident matrix. Hypergraph Laplacians can also be defined as combinatorial Laplacians of the Clique complex, which is induced from the hypergraph.

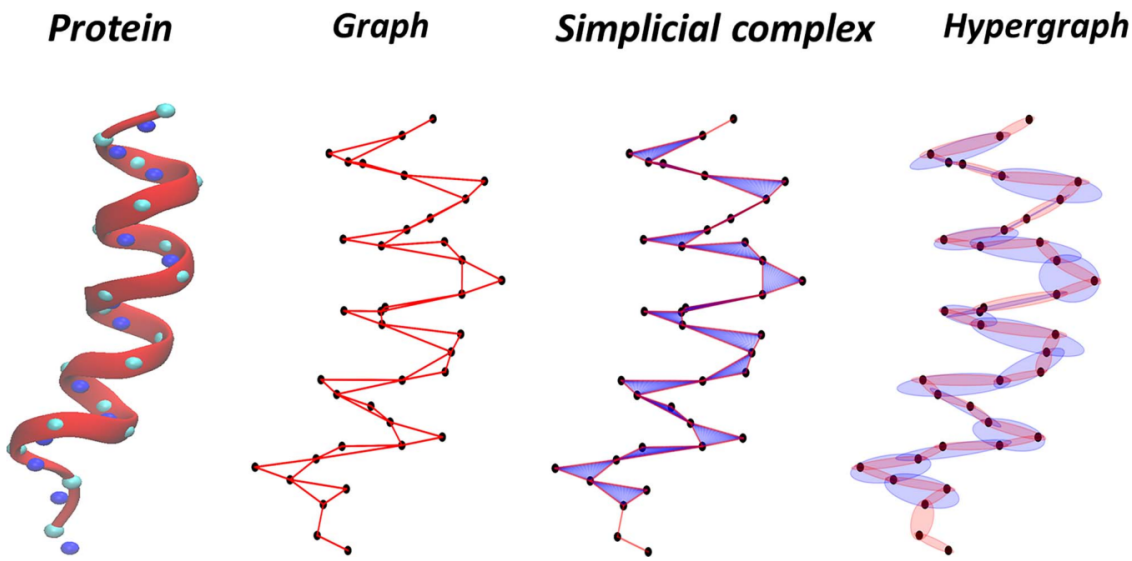


figure 6. The graph shows three different topological representations of a protein molecule: graph, simplex, and hypergraph. Graphs consist only of 0-simplex (vertices) and 1-simplex (edges); Simplex complexes can have higher dimensional simplex, such as 2-simplex (blue triangle); Hypergraphs consist of hyperedges, which are nonempty subsets of vertices and can be viewed as extensions of edges and simplex. 1-hyperedges and 2-hyperedges are represented by red and blue ellipses, respectively.

Application fields

Protein-ligand binding prediction, Interpersonal social network analysis.

Advantages

Describing the interaction between different types of point cloud data , loosening the conditions of complex generated by point cloud data.

Basic principles

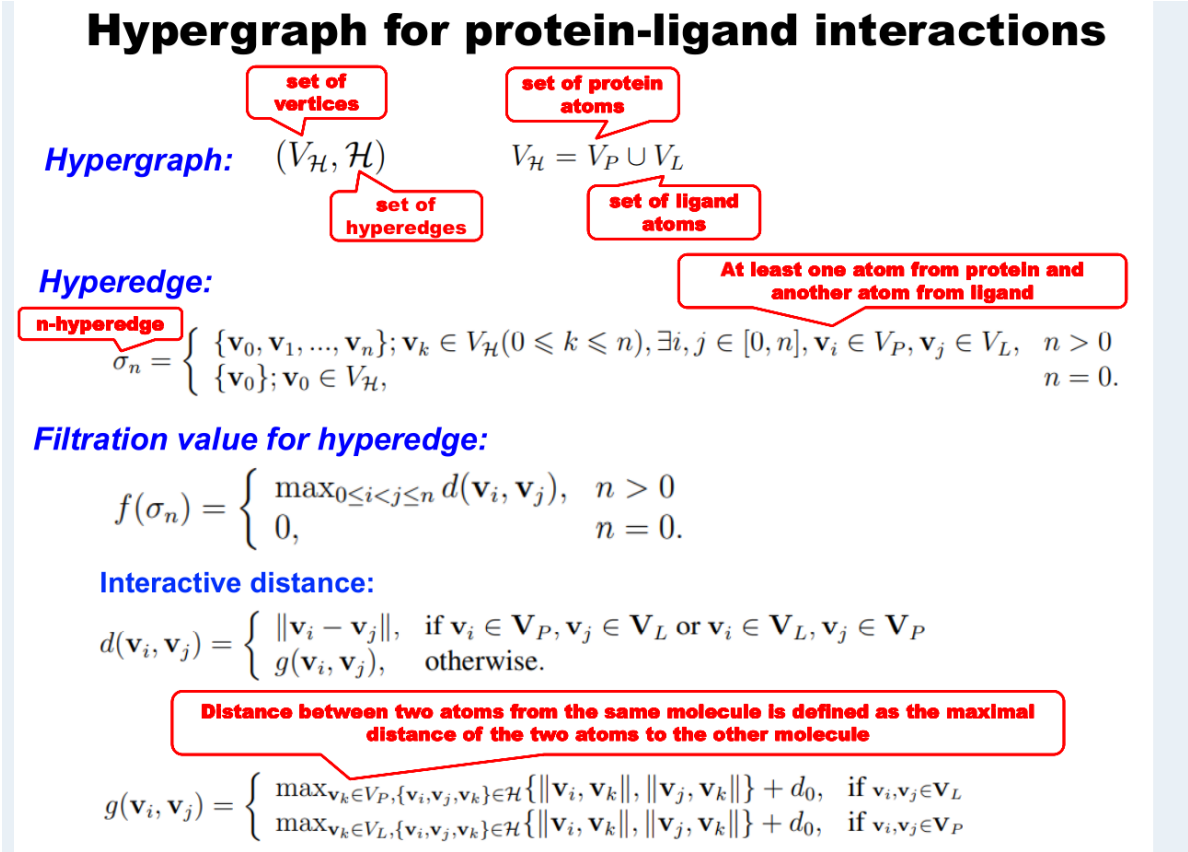


figure 7

6. Persistent cohomology

Persistent cohomology theory has also been applied to topological data analysis. Usually, when analyzing datasets with persistent homology, the geometric information is built into topological invariants while nongeometric information is usually neglected. for instance, during the topological abstraction of biomolecular datasets, some physical, chemical, and biological information (such as atom types, partial charges, and pairwise interaction strengths in a dataset) is neglected.

Persistent cohomology encodes non-geometric properties into functions fully or partially defined on simplicial complexes locally associated with the cohomology generators, thus it can embed other important information in a dataset into topological invariants generated from the geometric information of the dataset. We seek a formulation that organizes geometric information into a simplicial complex while encoding chemical, physical, and biological properties into functions fully or partially defined on simplicial complexes locally associated with the cohomology generators.

Application fields

Prediction of protein-ligand binding affinities from large datasets.

Advantages

Persistent cohomology enrichs barcodes by embedding multicomponent nongeometric information into topological representations of the geometric information.

Basic principles

Given a simplicial complex X of dimension n and a function $f\colon X_k \to R\ (0 \le k \le n)$, we seek a method to embed the information of f on the persistence barcodes obtained with a given filtration of X. In other words, we seek a representation of f on cohomology generators. But some representative cocycles in persistent cohomology may not reflect the overall location and structure associated with their cohomology generators.

To better embed the additional information in the data into cohomology generators, we look for a smoothed representative cocycle in each cohomology class. The smoothness of functions can be measured by using a Laplacian. We then construct smoothed representative cocycles with this Laplacian. There can be many choices for the Laplacian operator such as the discrete Hodge Laplacian for manifold-like complexes and the graph Laplacian or its higher-order generalizations for graphs.

Examples

Annuluses. Consider a point cloud sampled from two adjacent annuluses with radius 1 and centered at $(0, 0)$ and $(2, 2)$ as shown in Figure 8. The persistent cohomology computation was carried out using a Vietoris-Rips complex based filtration with the Euclidean distance. There are two persistent H_1 bars associated to the two significant circles whose smoothed cocycles show the contribution of simplices to the bars .

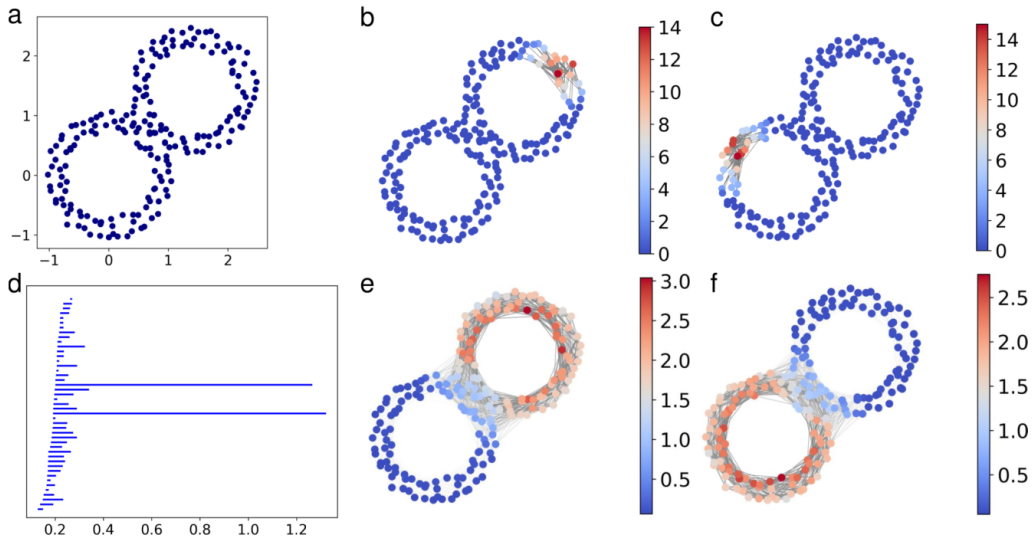


Figure 8. a: A point cloud sampled from two adjacent annulus. b,c: Two representative cocycles corresponding to the two long bars in the H_1 barcode. d: The H_1 persistence barcode of the point cloud with Vietoris-Rips filtration. e,f: The smoothed cocycles. In b, c, e, and f, the node color shows the weight induced by the smoothed cocycle projected to the nodes and the opaqueness of edges shows the weight induced by the smoothed cocycle.

Futhermore, if given datasets with similar geometry but different nongeometric information, values on the nodes, we can use enriched barcodes to distinguish between them as shown in Figure 9. Here, D1 and D2 have the same geometry, and thus their curve is more on the left side, which means there is a smaller distance between their persistent homology barcodes. On the

other hand, D3 has a similar value assignment on the points as that of D2, so their curve is on the bottom, which means there is a smaller distance in the nongeometric information.

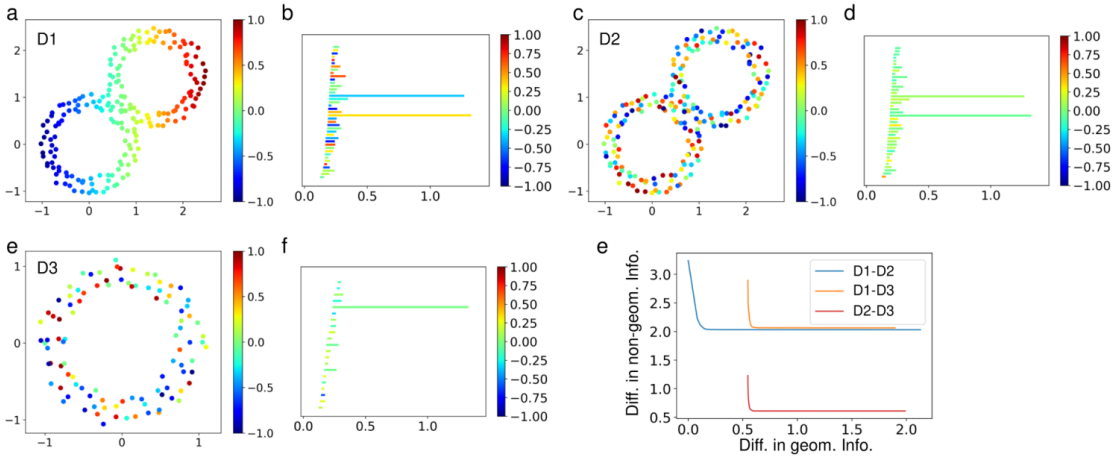


Figure 9. a-f: Three datasets with nongeometric information on the nodes and their H_1 enriched barcodes. e: The Wasserstein characteristics curves among these three datasets. The computation is done on a finite set of γ values, from 0 to 1 with a step size of 0.005.

References

[1] Cang Z , Wei G W . Persistent cohomology for data with multicomponent heterogeneous information[J]. 2018.

[2] Xia K , Anand D V , Saxena S , et al. Persistent homology analysis of osmolyte molecular aggregation and their hydrogen-bonding networks[J]. Physical Chemistry Chemical Physics, 2019, 21.

[3] Meng Z , Xia K . Persistent spectral based machine learning (PerSpect ML) for drug design[J]. 2020.

[4] Wang R , Nguyen D D , Wei G . Persistent spectral graph[J]. International Journal for Numerical Methods in Biomedical Engineering, 2020, 36(9).

[5] Rote G , Vegter G . Computational Topology: An Introduction[J]. Springer Berlin Heidelberg, 2006.

[6] Anand D V , Xia K , Mu Y . Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis[J]. Scientific Reports.