

EFFICIENT DECOMPOSITION OF FORMAN-RICCI CURVATURE ON VIETORIS-RIPS COMPLEXES AND DATA APPLICATIONS *

DANILLO BARROS DE SOUZA[†], JONATAS TEODOMIRO[‡], FERNANDO A. N. SANTOS[§],
MENGJUN DING[¶], WEIQIANG SUN^{||}, MATHIEU DESROCHES[#], JÜRGEN JOST^{††}, AND
SERAFIM RODRIGUES^{‡‡}

Abstract. Discrete Forman-Ricci curvature (FRC) is an efficient tool that characterizes essential geometrical features and associated transitions of real-world networks, extending seamlessly to higher-dimensional computations in simplicial complexes. In this article, we provide two major advancements: First, we give a decomposition for FRC, enabling local computations of FRC. Second, we construct a set-theoretical proof enabling an efficient algorithm for the local computation of FRC in Vietoris-Rips (VR) complexes. Strikingly, this approach reveals critical information and geometric insights often overlooked by conventional classification techniques. Our findings open new avenues for geometric computations in VR complexes and highlight an essential yet under-explored aspect of data classification: the geometry underpinning statistical patterns.

Key words. Forman-Ricci curvature, discrete geometry, set theory, optimization, complex systems, higher-order networks, data science.

AMS subject classifications. 05C85, 52C99, 90C35, 62R40, 68T09.

1. Introduction. Network-based analysis [31] provides a versatile and effective framework for data mining [40] and thus applies to a wide range of applications, particularly for data with complex relationships and hierarchies, such as community detection [17]. Recently, it has been instrumental in advancing machine learning techniques [21]. Beyond the dyadic network framework, Topological data analysis (TDA) has recently emerged as the leading approach to studying higher-order relationships and hierarchies in data, as well as enhancing the signal-to-noise ratio of data features [43]. As a consequence, TDA has permeated across several fields, including neuroscience [36]. Due to the intimate relationship between topology (global property) and geometry (local property), as stated for instance in the Gauss-Bonnet theorem [22],

*This manuscript is for review purposes only.

Funding: This research is supported by the grant PID2023-146683OB-100 funded by MICIU/AEI /10.13039/501100011033 and by ERDF, EU. Additionally, it is supported by Ikerbasque Foundation and the Basque Government through the BERC 2022-2025 program and by the Ministry of Science and Innovation: BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIU / AEI / 10.13039/501100011033. Moreover, the authors acknowledge the financial support received from BCAM-IKUR, funded by the Basque Government by the IKUR Strategy and the European Union NextGenerationEU/PRTR. We also acknowledge the support of ONBODY no. KK-2023/00070 funded by the Basque Government through ELKARTEK Programme. Weiqiang Sun and Mengjun Ding are supported by the National Key Research and Development Project of China under Grant 2024YFB2908301 and by the National Natural Science Foundation of China (NSFC) under Grant 62331017

[†]Basque Center for Applied Mathematics (danillo.dbs16@gmail.com, dbarros@bcamath.org)

[‡]Universidade Federal de Pernambuco (jonatas.teodomiro@ufpe.br)

[§]Dutch Institute for Emergent Phenomenal (f.a.nobregasantos@uva.nl)

[¶]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China (mengjun.ding@sjtu.edu.cn)

^{||}School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China (sunwq@sjtu.edu.cn)

[#]Inria, Montpellier, France (mathieu.desroches@inria.fr)

^{††}Max Planck Institute for Mathematics in the Sciences, Leipzig, and Center for Scalable Data Analytics and Artificial Intelligence, Leipzig University, Germany, and Santa Fe Institute, New Mexico, USA (jost@mis.mpg.de).

^{‡‡}Basque Center for Applied Mathematics (srodrigues@bcamath.org)

TDA has also advanced with the study of geometric properties of data (Topological and Geometrical Data Analysis - TGDA). In this context, discrete Ricci curvatures have been shown to be a powerful geometric descriptor of networks, enabling, for example, efficient detection of network clusters [16, 26, 37]. Its application is now widespread in several fields, such as stock market fragility [38], neuroscience [4], and epidemiology [8], to name a few. A step further is given by the discrete Forman-Ricci curvature (FRC), which provides a powerful geometric descriptor of complex networks, including higher-order networks [16].

Despite the relevance of geometric descriptors, topological approaches, such as those based on persistent homology [12, 44] prioritize topological descriptors of data while largely disregarding the geometric perspective. This is partly explained due to the computational complexity of geometric descriptors, particularly in the context of constructing higher-order networks. Moreover, there is no local geometric descriptor that naturally extends to a global computation, as for example expressed by the Gauss-Bonnet theorem [23].

To overcome these challenges, we leverage set theory [14, 20, 32, 41], which has proven effective in optimizing discrete computational algorithmic methods. For instance, our recent works unveiled the construction of efficient set-theoretical approaches for computing geometric invariants [6, 9]. Building on these results, we explore geometric computations in greater detail and identify computational patterns in the construction of Vietoris-Rips (VR) complexes that reduce the complexity of the FRC computation. Strikingly, we find a decomposition strategy that enables the local computations of FRC from local network neighbourhoods. This permits the geometric local update to be possible with minor numerical increments and efficient computational complexity.

To clarify the rationale behind our approach, we give an overview of our computational strategy. In traditional homology barcode computations, the topological invariant is recomputed as a function of the cutoff distances of VR complexes, once the global structure needs to be recomputed from scratch. For example, in order to recompute the Betti numbers [2], the clique's neighbourhood needs to be recomputed, as well as the boundary operators. From a geometric point of view, the local neighbourhood changes as the new structures are added as a function of the distance, and may imply re-computations of local geometry as well. To reduce the computational complexity, we develop an alternative set-theoretical approach that updates the numerical computation of these curvatures instead of recomputing the FRC for the updated network. As a proof-of-concept, we test our novel computational method on various datasets with the aim of exploring and understanding the effect of geometric approaches in VR complexes. This also includes noise sensitivity studies. Here, we denote our methodological procedure *data geometrization*. Our studies validate our approach and, crucially, introduce a novel descriptor suited for noisy high-dimensional datasets. This geometric descriptor, facilitated by our algorithm, complements topological and statistical information, enhancing data analysis.

The rest of this article is organised as follows. In [section 2](#), we give a brief review of the theoretical underpinnings of networks and VR complexes, define the FRC and examine associated optimizations methods in past works. Then in [section 3](#), we benchmark our novel algorithm to compute FRC in synthetic and real datasets. Finally, in [section 4](#), we discuss the details of our findings and give a few perspectives on this topic.

2. Theoretical background. In this section, we provide a brief background of graphs and the VR complexes. Subsequently, we define the discrete FRC curvature. We refer the reader to classic literature for more details on these concepts [13, 16, 42, 44].

2.1. Vietoris-Rips complexes from networks. A simple weighted undirected graph is a pair $G = (V, E)$, together with a weight function $w : E \rightarrow \mathbb{R}$, where $V := V(G)$ is the (finite) set of nodes of G and $E := E(G)$ is the set of edges connecting nodes in G , such that the following equation is satisfied:

$$(2.1) \quad E \subseteq \{\{x, y\} \mid x, y \in V, x \neq y\}.$$

The neighbours of a node $x \in V$ will be denoted by π_x and are defined by

$$(2.2) \quad \pi_x = \{y \in V \mid \{x, y\} \in E\}.$$

Similarly, the neighbours of a set of nodes $\alpha = \{x_1, \dots, x_n\}$ are defined by

$$(2.3) \quad \pi_\alpha = \bigcap_{x \in \alpha} \pi_x.$$

A *clique complex* is the set of all complete subgraphs of G . We say that these subgraphs are its simplices, and, when such a subgraph has $d+1$ nodes, we call it a d -face. We define the set of all d -faces by C_d . We say that a d -face has dimension d and that the dimension of the clique complex is the highest dimension among its faces. We can define this clique complex as the union $C = \bigcup_d C_d$. We see that $V(G)$, together with C , is an *abstract simplicial complex*, i.e., they satisfy the following conditions:

1. For each $v \in V(G)$, $\{v\} \in C$;
2. If $\gamma \subseteq \alpha$ and $\alpha \in C$, then $\gamma \in C$.

We define the *Vietoris-Rips complex* constructed from C as a function of the radius distance ε by

$$(2.4) \quad \text{VR}_C(\varepsilon) := \{\sigma \subseteq C \mid \text{diam } \sigma \leq \varepsilon\},$$

where $\text{diam } \sigma = \max\{w_e \mid e \in \sigma\}$, $w_e = w(e)$, and ε is the *cutoff* distance for generating the simplicial complex. Finally, let $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k$ be the diameters of the simplices in the simplicial complex such that $\varepsilon_1 \leq \varepsilon_2 \leq \dots \leq \varepsilon_k$. Then it is clear that

$$(2.5) \quad \emptyset \subseteq \text{VR}(\varepsilon_1) \subseteq \dots \subseteq \text{VR}(\varepsilon_k) = C.$$

Figure 1 (top) elucidates the construction of a VR complex with an intuitive example. Next, we define the FRC and summarise the key results from our previous work, which will serve as the foundation for proposing a novel, efficient formulation.

2.2. Discrete Forman-Ricci curvature. The *boundary* of a d -face α is the set of all $(d-1)$ -faces in α , and it is denoted by $\partial(\alpha)$. When $\gamma \in C_{d-1}$ is contained in the boundary of α , we denote it by $\gamma < \alpha$; alternatively, α containing γ in its boundary will be denoted by $\alpha > \gamma$. Two faces $\alpha_1, \alpha_2 \in C_d$ are said to be *neighbours* if at least one of the following conditions is satisfied:

1. there exists a $(d-1)$ -face γ such that $\gamma < \alpha_1, \alpha_2$,
2. there exists a $(d+1)$ -face β such that $\alpha_1, \alpha_2 < \beta$,

and we define by N_α the set of all neighbours of α . We say that α_1 and α_2 are *parallel neighbours* if condition 1. is satisfied, but not 2.. If both conditions are satisfied simultaneously, then α_1 and α_2 are said *transverse neighbours*.

The original formulation for higher-order FRC, defined for CW-complexes, originates from [16]. In the special case of simplicial complexes, we can express it as proposed in [6] as follows

$$(2.6) \quad F(\alpha) = |H_\alpha| + (d+1) - |P_\alpha|,$$

where H_α is the set of $(d+1)$ -faces containing α in its boundary, and P_α is the set of parallel neighbours of α . We therefore define the d -th Forman-Ricci curvature (or the d -FRC) of a non-empty simplicial complex as follows:

$$(2.7) \quad F_d(C) = \frac{1}{|C_d|} \sum_{\alpha \in C_d} F(\alpha),$$

assuming that $C_d \neq \emptyset$. For convenience of numerical computations, we define $F_d(\emptyset) := 0$. This will facilitate the formalism of our approach in this article. Finally, we define $F(\alpha)$ using a key result from our previous work [6] as follows:

$$(2.8) \quad F(\alpha) = (d+2) \cdot \left| \bigcap_{\gamma \in \partial \alpha} \pi_\gamma \right| + 2 \cdot (d+1) - \sum_{\gamma \in \partial \alpha} |\pi_\gamma|,$$

where π_γ is defined as in (2.3). In the case of geometrical simplicial complexes, formally, we do not define curvature if the simplicial complex is degenerate.

3. Results. We subsequently outline our main theoretical enhancements alongside their computational implications. Finally, to demonstrate its validity, we apply it to synthetic data.

3.1. Alternative formulation for local FRC computations. Despite the FRC formulation being originally defined from local neighbourhood interactions as in (2.6) and (2.7), the current classic formulation (that uses the average local *FRC* curvature of neighbouring nodes) is unable to recover the geometry from local to global scale. This is usually possible in topological approaches, e.g., the Euler characteristic computation obtained from the Knill curvature [24]. To tackle this issue, we investigated alternative ways to consider local computations that extend the geometric information to a global viewpoint. To this end, we constructed an alternative local computation of FRC inspired by the Gauss-Bonnet theorem for simplicial complexes [22, 25]. More precisely, we define the *local Forman-Ricci curvature* to the nodes $x \in V$ (for x in d -faces) as follows:

$$(3.1) \quad f_d(x) = \frac{1}{(d+1) \cdot |C_d|} \sum_{\alpha \ni x} F_d(\alpha).$$

Thus, the global FRC can be recovered from the local computations from (3.1), specifically, by the formula

$$(3.2) \quad F_d(C) = \sum_{x \in V} f_d(x).$$

The derivation of (3.2) is detailed in A.3. Figure 9 elucidates the computation of FRC.

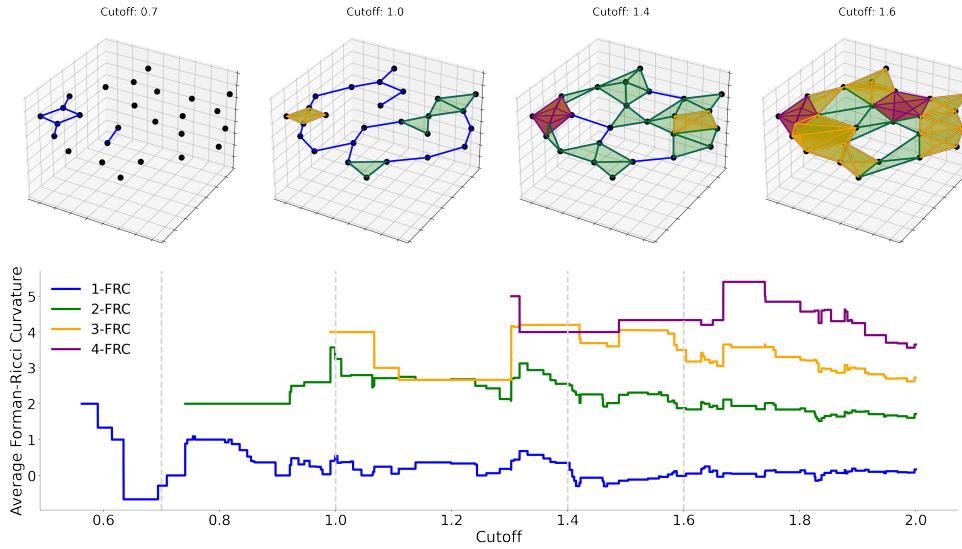


FIG. 1. *FRC computation on a VR complex as a function of the cutoff distance. The dashed-grey lines correspond to the cutoff values 0.7, 1.0, 1.4 and 1.6. The blue, green, yellow and red curves correspond to the average FRC for edges ($d = 1$), triangles ($d = 2$), tetrahedra ($d = 3$) and pentahedra ($d = 4$), respectively.*

3.2. A novel algorithm to efficiently compute FRC in a VR complex.

Computations of higher-order geometric invariants in complex networks are not new. However, the computations on VR complexes are challenging, since the additional computation of several sequential networks may be necessary. Figure 1 provides an example of FRC computation in a VR complex. Subsequently, we define such computations and explicitly point out the challenges that they entail. Then, we will propose a novel set-theoretical approach for efficiently computing higher-order FRC in VR complexes filtrations.

Let $C_d = \{\alpha_1, \dots, \alpha_k\}$ be a finite set (or sequence) of d -faces, with a positive weight function $\omega : C_d \rightarrow \mathbb{R}$ such that $\omega(\alpha_k) = w_{\alpha_k}$. We also assume that faces' weights are sorted in ascending order (i.e., $w_{\alpha_1} \leq w_{\alpha_2} \leq \dots \leq w_{\alpha_k}$). We define $C_d^i = C_d^{i-1} \sqcup \{\alpha_i\}$, for $i > 1$, with $C_d^0 = \emptyset$ and $C_d^1 = \{\alpha_1\}$. It is clear that the sequence of simplicial complexes $(C_d^i)_i$ are such that

$$(3.3) \quad \emptyset \subseteq C_d^1 \subseteq C_d^2 \subseteq \dots \subseteq C_d^k = C_d.$$

Therefore, (3.3) defines a VR complex. A natural and straightforward FRC computation in a VR complex can be accomplished by simply computing the sequence

$$(3.4) \quad 0 =: F(\emptyset), F(C_d^1), F(C_d^2), \dots, F(C_d^k) = F(C_d).$$

However, this would imply an algorithm that requires exhaustive recomputation of the FRC for each simplicial complex C_d^i . Instead, we find that linking equations (3.3) and (2.5) enables the construction of an optimal algorithm for computing the FRC of higher-order faces across a filtration. This observation leads us to propose an optimal algorithm which takes as input the sorted simplicial complex C and the maximum face dimension d_{max} . Such an algorithm will be detailed in this section.

Observation	Feature 1	Feature 2	...	Feature n
x_1	x_{11}	x_{12}	...	x_{1n}
x_2	x_{21}	x_{22}	...	x_{2n}
...
x_m	x_{m1}	x_{m2}	...	x_{mn}

TABLE 1

Classic Tabular Input Data: The input table is given as an $m \times n$ matrix (i.e., m points in a n -dimensional space). More precisely, each observation point is in the shape $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ for $i \in \{1, 2, \dots, m\}$.

Henceforth, for all $\alpha \in C_d^i$, we will denote $N_\alpha^i := N_\alpha(C_d^i)$, $T_\alpha^i := T_\alpha(C_d^i)$, $P_\alpha^i := P_\alpha(C_d^i)$ and $H_\alpha^i := H_\alpha(C_d^i)$ the set of neighbours, transverse neighbours, parallel neighbours and higher-order faces containing α with regards to C_d^i , respectively. In the special case where $\alpha = \alpha_i$, we will denote $N_{\alpha_i}^i$ by N^i and $P_{\alpha_i}^i$ by P^i . We also define $F^i(\alpha)$ by the computation of FRC to α at the i -th step, i.e., when $\alpha \in C_d^i$, $\alpha \neq \alpha_i$. More precisely, we define: $F^i(\alpha) = |H_\alpha^i| + (d+1) - |P_\alpha^i|, \forall \alpha \in C_d^{i-1}$.

From Proposition 1 (see section subsection A.3), if α^i is the new neighbour of some $\alpha \in C_d^i$, then it can be either a parallel or transverse neighbour of α . If α_i is a new parallel neighbour of α , then we have $F^i(\alpha) = F^{i-1}(\alpha) - 1$. Otherwise, we have α as a new transverse neighbour and thus $F^i(\alpha) = F^{i-1}(\alpha) + (d+1)$. For illustration, Figure 9 provides an example for edges ($d=1$) and triangles ($d=2$).

This finding, together with the new formulation provided in (2.8) from [6] motivated us to derive and design an algorithm for computing FRC in a filtration of a VR complex. In order to facilitate the implementation of our algorithm, we define $\Delta F(\alpha) := F^i(\alpha) - F^{i-1}(\alpha)$ as an auxiliary function to be applied over the neighbours $\alpha \in N_{\alpha_i}^i$, which can be re-written as follows:

$$(3.5) \quad \Delta F(\alpha) = \begin{cases} -1, & \alpha \in P^i \\ (d+1), & \text{otherwise} \end{cases}.$$

Using our results from [8] (see also subsection A.1), this formula induces an equivalent set-theoretical formulation as a function of the nodes $x \in \partial(\alpha)$ as follows:

$$(3.6) \quad \delta(x) = \begin{cases} -1, & x \notin \pi_{\alpha_i} \\ (d+1), & \text{otherwise} \end{cases}.$$

The core idea behind formulas (3.5) and (3.6) is to examine the neighbourhood of the new face at step i and locally iterate through the neighbouring faces to assess the contribution of the new face added to the curvature. This is achieved simply by comparing the curvature before and after adding the new face. In particular, equation (3.6) allows this verification by identifying the neighbourhood of the current face through a set-theoretical representation, as performed in our previous work in [6]. The above formulation, particularly through (3.6), enables the implementation of an efficient algorithm for computing the FRC on VR complexes, using a decision tree based on the neighbourhood of each newly added face α_i . This leads us to our algorithm (A.2) (see subsection A.2). This novel derivation not only enables dynamic computation of the FRC as a function of the cutoff distance, but it also allows for revealing underlying structural changes and patterns in simplicial complexes from a geometric perspective.

Observation	ε_1	ε_2	\dots	ε_k
$f_d(x_1)$	f_d^{11}	f_d^{12}	\dots	f_d^{1k}
$f_d(x_2)$	f_d^{21}	f_d^{22}	\dots	f_d^{2k}
\dots	\dots	\dots	\dots	\dots
$f_d(x_m)$	f_d^{m1}	f_d^{m2}	\dots	f_d^{mk}

TABLE 2

Geometrized Tabular Data: The new input table is provided from the geometric information (FRC) computed on each observation originally provided by [Table 1](#). More precisely, each geometrized observation is a vector $f_d(x_i) = (f_d^{i1}, f_d^{i2}, \dots, f_d^{ik})$, for all $i \in \{1, 2, \dots, k\}$, where f_d^{ij} is the computation of the d -th local FRC for the node x_i restricted to $VR(\varepsilon_j)$. The global FRC is recovered from [\(3.2\)](#), and the geometric output is similar to the content in [Figure 1](#).

3.3. Application of Data Geometrization. Data science approaches aim at determining descriptive statistical summaries that synthesizes data, enabling inferences and predictions. Typically, this data is provided as in [Table 1](#). In our approach, we perform the FRC per observation (i.e., points in the aforementioned assumed coordinate system defined by the data features) by using equation [\(3.1\)](#) as a function of each distance ε (in this case, the Euclidean distance), which leads to [Table 2](#). The data geometrization concept has been applied to manyfold learning approaches, as in [\[39\]](#). To validate our proposed approach, we consider both synthetic point cloud data and breast cancer data from public repositories [\[10, 1\]](#). Our algorithm and all data processing were implemented in the Python language [\[35\]](#). To facilitate understanding, we also provide the pseudo-code to compute both local and global FRC in [subsection A.2](#). To generate the faces of the simplicial complexes, we use the package Gudhi [\[3, 34\]](#). Following our approach, the computed curvature values (as a function of the cutoff distance) can then be used as features for classification. In particular, we employ the Uniform Manifold Approximation and Projection algorithm (UMAP [\[29, 30\]](#)) to classify the geometric outputs from the considered data.

3.3.1. FRC computations on Synthetic point cloud data. We examine two sub-categories of synthetic point clouds, namely Random Geometric Graphs and Replicas of the Datasaurus datasets. The corresponding generated datasets are publicly available via [\[7\]](#). These two sub-cases are detailed below:

1. Random geometric graphs [\[5, 33\]](#) with $n = 100$ nodes and different box dimensions ($\dim \in \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$) and network densities ($\rho \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$). To generate the random geometric graphs, we used the package NetworkX [\[19\]](#). Noteworthy, we tested the sensitivity of the box dimension across the network density;
2. Replicas of Datasaurus datasets, (a synthetic point-cloud data repository) that generates different geometries with the same basic statistics [\[18\]](#)(see [Figure 12](#)). This is inspired by an alternative randomization algorithm that shuffles the points coordinates while preserving the average, the standard deviation and Pearson correlation of points coordinates [\[28\]](#). For more details, see [subsection A.2](#). As above, we tested the sensitivity of the data recognition across datasets.

Starting with the case of random geometric graph, in [Figure 2](#), we show classification of geometric graphs with fixed density or $\rho = 0.25$ and different box dimensions. In particular, we compute the global d -FRC for $d \in \{1, 2, 3\}$ (top panels) and parameterize UMAP with these geometric figures. The UMAP classification is depicted in the bottom panels. Notably, a similar classification pattern is observed for different

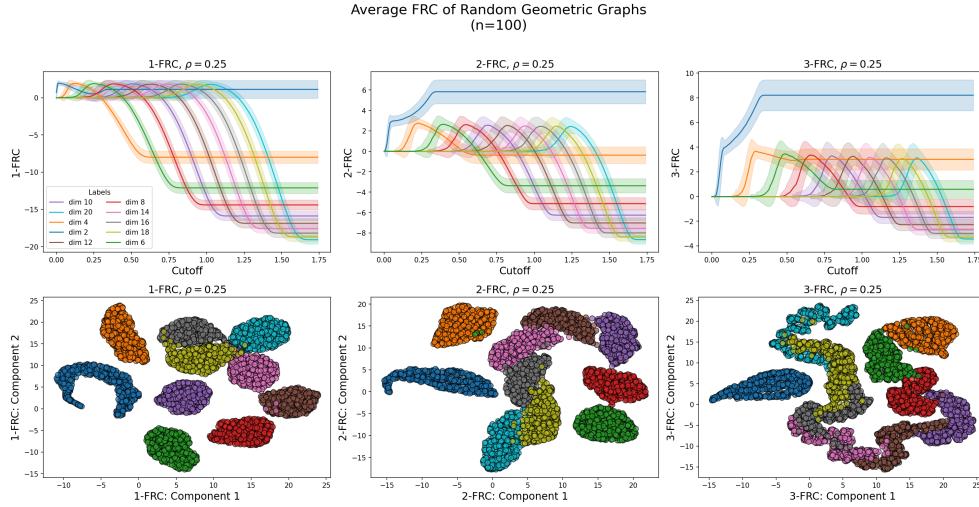


FIG. 2. Computation of d -FRC for $d = 1$ (first column), $d = 2$ (middle column) and $d = 3$ (third column), and fixed edge density $\rho = 0.25$ in random geometric graphs with different box dimensions (see labels). The UMAP classification (bottom figures) used the Euclidean metric and the minimum distance of points 1.0 (the highest allowed). We used this parameter to split away the points as much as possible and test the sensitivity to noise.

d . This is in part because the higher orders ($d = 2, d = 3$) information builds upon pairwise interactions ($d = 1$). Specifically, the distance-based cliques are computed from pairwise information. In other words, the tetrahedra are dependent on the preceding existence of triangles, which only exist from the coupling of three edges. The above results are robust against other network density values (see subsection A.5 for further examples). Our findings highlight the importance of methods for detecting network dimensionality and extend recent works in the field [15]. Moreover, it can be used as an alternative to the state-of-the-art methods used for pattern recognition in networks [27].

Moving to the case of DataSaurus datasets, we also test the sensitivity to noise added from the randomization iteration process. Figure 3 illustrates the sensitivity of the global FRC over the data randomization process. In Figure 4, we provide the comparison between the geometry from Datasaurus datasets randomization and its UMAP classification. Noteworthy, we show that the global FRC provides a structural classification even in the presence of noise after the randomization steps. Similar to the results on random networks, higher-order curvatures do not provide additional information since they are built upon pairwise information (i.e., due to the lack of independence of face generation in distance-based approaches). In subsection A.4, we test the sensitivity to higher levels of noise. Crucially, our approach provides a geometric classification of the Datasaurus that generalizes the results obtained in [11], where the authors used homology to identify statistical differences in the data.

3.3.2. FRC computations on breast cancer datasets. We now test our approach on two sets of breast cancer data, namely:

1. Breast cancer diagnosis from Wisconsin dataset [10] with two classification labels (benign and malignant);
2. Breast cancer classification from the Molecular Taxonomy of Breast Cancer

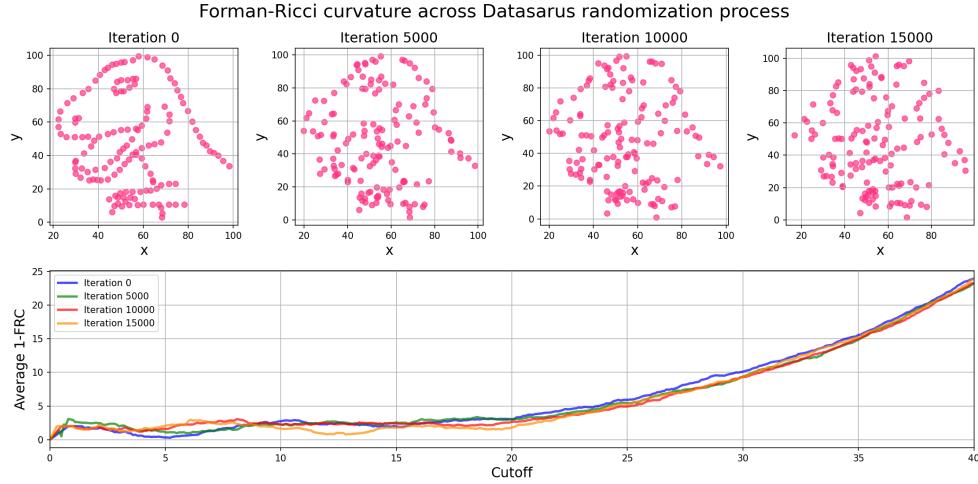


FIG. 3. *FRC computation for different steps of Datasaurus dataset randomization. In the example, the algorithm A.3 was performed to disturb the original data (iteration 0) in a total of 9000 iterations, which totaled 17469 effective iterations. From these, we show the dataset changes for iterations 5000, 10000 and 15000. Despite the randomization process that generates distinct geometry from VR complexes, the FRC presents robustness for low noise levels.*

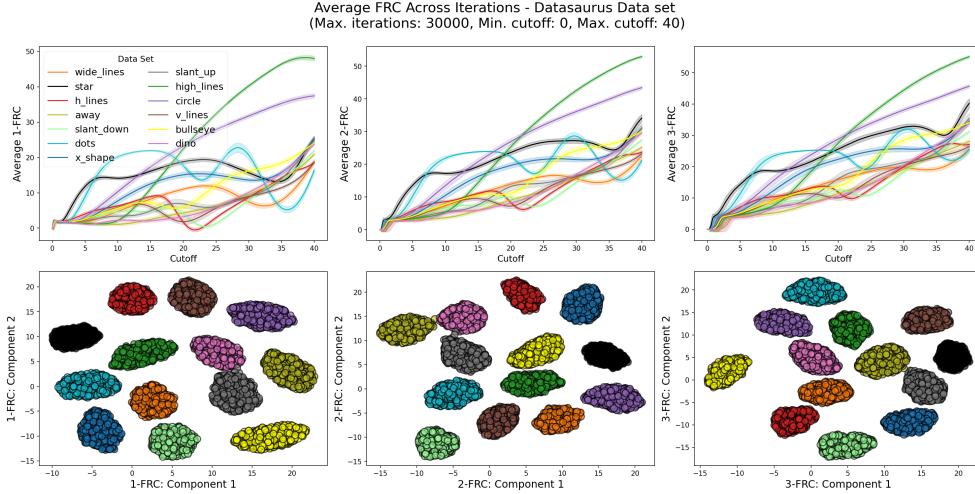


FIG. 4. *Computation of d-FRC, for $d = 1$ (first column), $d = 2$ (middle column) and $d = 3$ third column of the Datasaurus dataset randomization process, for a maximum of 30000 randomization steps. The algorithm A.3 was used for generating the randomized datasets.*

International Consortium (METABRIC) [1]. We used the patient's vital status (alive, died of the disease, and died of other causes) for the classification labels.

In both datasets, we restricted the dataset to the numerical data features and only considered patients without missing information. We apply the geometrization process as described in Table 2. For the METABRIC dataset, we additionally included a

feature representing the duration of time the patient lived with the tumor, calculated as the difference between the date of death and the treatment start date. Due to computational limitations, we restricted the computation of the local FRC to edges (1-FRC).

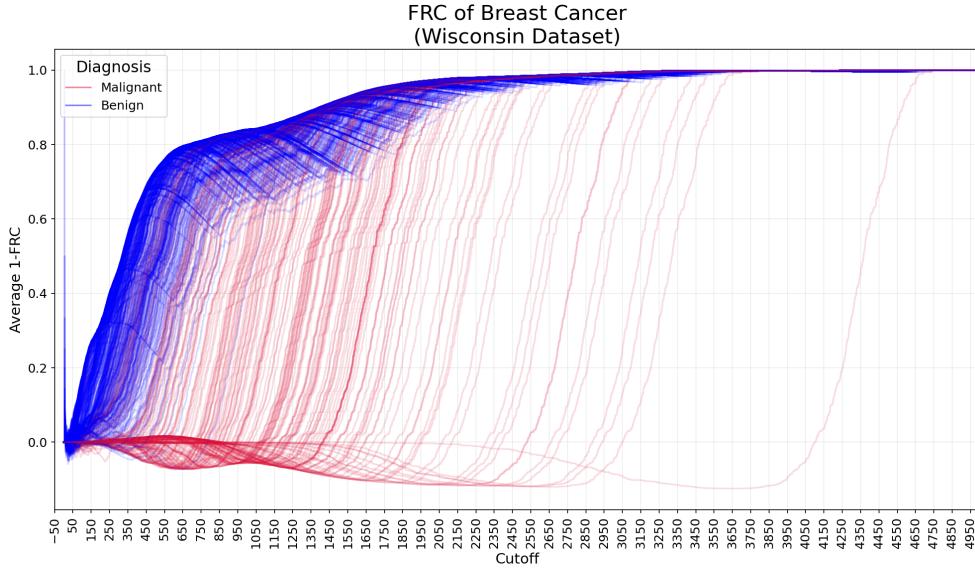


FIG. 5. Average 1-FRC per patient from Wisconsin breast cancer database. The benign and malignant diagnoses are represented in blue and red, respectively. Notoriously, there is a geometric separation between the two groups from the statistical features.

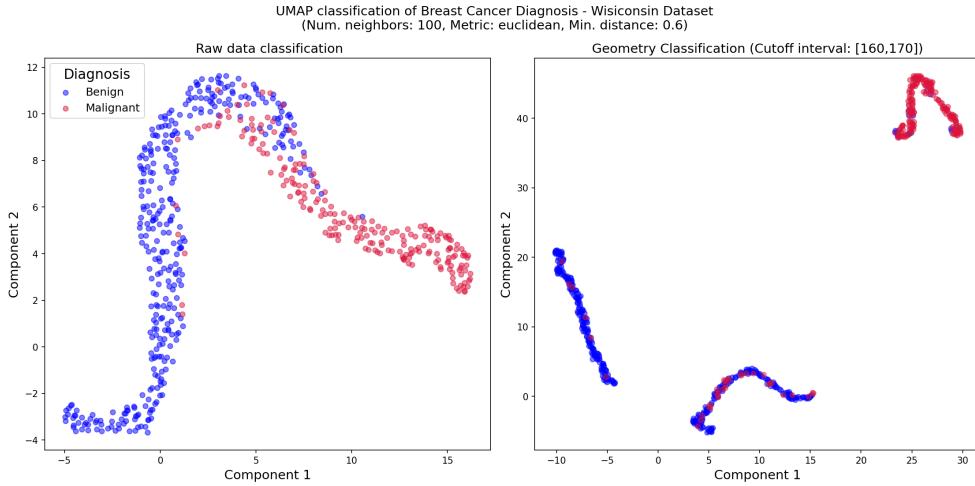


FIG. 6. Comparison between the UMAP classification of breast cancer (Wisconsin database) by using raw data (left panel) and geometrized data (right panel).

In Figures 5 and 7, we show the local 1-FRC as a function of the cutoff distance for each patient from the Wisconsin database and METABRIC database, respectively. The resulting classification is represented via different colorings of the FRC curves. In the Wisconsin data, we used red for the malignant tumors and blue for the benign ones, respectively. For the METABRIC data, we used green, blue and red to distinguish the groups of alive, and died of other causes and died of the diseases, respectively.

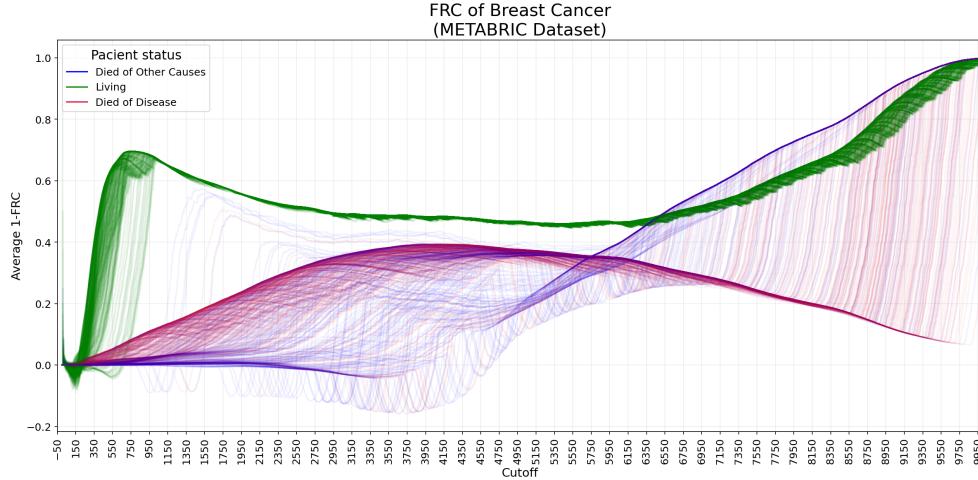


FIG. 7. Average 1-FRC per patient from METABRIC dataset. Green, blue and red curves represent the patients who survived, died of other causes and died of the disease, respectively.

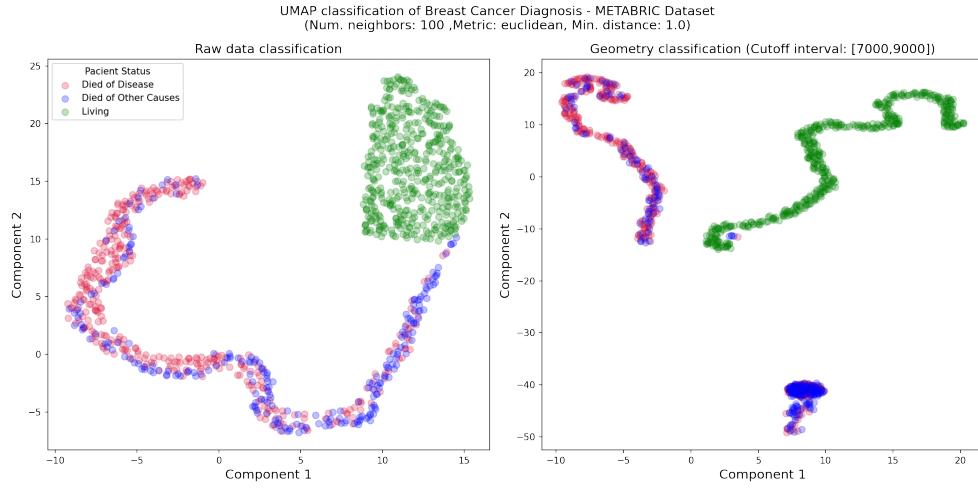


FIG. 8. Comparison between the UMAP classification of breast cancer (METABRIC database) by using raw data (left panel) and geometrized data (right panel). Notably, there is a more evident geometric separation between the groups.

The FRC effectively reveals meaningful patterns embedded in the dataset's intrinsic geometry. However, it is important to emphasize that data geometrization should be viewed as a toolkit—a method that guides analysis rather than a one-time procedure. It should be approached as an iterative process, informed by continuous visualization of the resulting classifications. To illustrate this, consider a dataset composed of statistical features. Applying data geometrization to these features helps identify optimal cutoff intervals where statistical differences between data points become apparent, and it highlights statistical outliers. This insight can be directly leveraged

to inform clustering decisions and improve the interpretability of the results. For example, in [Figure 5](#), the FRC becomes less distinguishable for cutoff values above 2500, as most curvature values converge toward similar magnitudes. This convergence suggests that the corresponding data points are not statistically separable, indicating that cutoff values greater than 2500 are suboptimal for distinguishing between malignant and benign tumors. Conversely, cutoff intervals between 50 and 2000 reveal clear statistical divergence between the two groups, making them suitable for classification purposes. Similarly, in [Figure 7](#), the choice of cutoff interval plays a critical role in balancing outlier detection, redundancy reduction, and the identification of statistically meaningful differences—factors essential for optimal data classification.

To enhance the result, we use UMAP to compare the usual data classification (from the data features) and the FRC as a function of the cutoff interval (geometrized data). The results are shown in [Figures 6](#) and [8](#). Notably, the FRC offers a significant enhancement in data classification. A key advantage of our approach is the ability to select a cutoff interval that reveals the greatest statistical divergence between groups, which is something that is not achievable with traditional classification methods. In [subsection A.7](#) and [A.8](#) we test the sensitivity between other cutoff intervals and different UMAP input parameters, e.g., the metric, the minimum distance between points, and the number of neighbours.

4. Conclusion. In this work, we introduced an alternative set-theoretical algorithm for computing higher-order Forman-Ricci curvature (FRC) in Vietoris-Rips complexes, based on the numerical increments of local curvature values as a function of cutoff distances. We also provided an alternative theoretical formulation for global geometric computation from local curvature definitions, drawing parallels with topological approaches like the Gauss-Bonnet theorem in complex networks. To benchmark our approach, we applied both global and local FRC to synthetic and real-world datasets, producing what we refer to as “geometrized data.” We then classified the data using the UMAP algorithm, examining its sensitivity to noise and the parameter space.

In contrast to state-of-the-art data classification methods, our data geometrization approach demonstrates robustness against noise and enhances the results of existing dimensionality reduction techniques, such as UMAP. This improvement stems from the fact that existing state-of-the-art geometry-based dimensionality reduction algorithms assume data uniformity for accurate classification. Since real-world data typically lacks such uniformity, our data geometrization process offers an alternative way to find the optimal balance between signal and noise, leading to improved accuracy. This is achieved by selecting appropriate cutoff distance intervals within the Vietoris-Rips (VR) complex.

However, despite these improvements, a visual inspection of the FRC is still necessary, which can become exhaustive when dealing with large datasets containing multiple labels. Additionally, the geometric computations involved remain time-consuming, limiting their application to low-dimensional structures and requiring significant computational resources and efficient code implementation. In our tests, the FRC filtration was computed using Euclidean distance (due to a limitation in the Gudhi algorithm), but the approach can be extended to other metrics, requiring alternative clique algorithms capable of constructing VR complexes in different metric spaces.

The results show that the FRC significantly enhances the statistical relevance of clustering groups, acting as a geometric classifier for datasets. These findings not only offer an alternative to barcode homology representations of datasets but also improve

classification outcomes compared to traditional methods. In conclusion, our approach facilitates the statistical analysis of large datasets through a geometric lens.

REFERENCES

- [1] R. ALHARBI, *Breast Cancer Gene Expression Profiles (METABRIC)*, 2020, <https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric>. Accessed: 2024-11-13.
- [2] S. BOCHNER, *Curvature and Betti numbers. II*, Annals of Mathematics, 50 (1949), pp. 77–93.
- [3] J.-D. BOISSONNAT AND C. MARIA, *The GUDHI Library: Simplicial Complexes and Persistent Homology*, ACM Communications in Computer Algebra, 47 (2014), pp. 85–87.
- [4] T. CHATTERJEE, R. ALBERT, S. THAPLIYAL, N. AZARHOOSHANG, AND B. DASGUPTA, *Detecting network anomalies using Forman–Ricci curvature and a case study for human brain networks*, Scientific Reports, 11 (2021), p. 8121.
- [5] J. DALL AND M. CHRISTENSEN, *Random geometric graphs*, Physical Review E, 66 (2002), p. 016121.
- [6] D. DE SOUZA, J. DA CUNHA, F. SANTOS, J. JOST, AND S. RODRIGUES, *An efficient set-theoretic algorithm for high-order Forman–Ricci curvature*. Proceedings of the Royal Society A (in press), 2025.
- [7] D. DE SOUZA, *F.R.C. on V.R. filtrations of random graphs*. <https://kaggle.com/datasets/fa3926660ecbe1ced3e2de6012a32ba9e3bdf8988f1e6ec9250387aafc73214>, 2025. Accessed: 2025-04-24.
- [8] D. B. DE SOUZA, J. T. DA CUNHA, E. F. DOS SANTOS, J. B. CORREIA, H. P. DA SILVA, J. L. DE LIMA FILHO, J. ALBUQUERQUE, AND F. A. SANTOS, *Using discrete Ricci curvatures to infer COVID-19 epidemic network fragility and systemic risk*, Journal of Statistical Mechanics: Theory and Experiment, 2021 (2021), p. 053501.
- [9] D. B. DE SOUZA, J. TEODOMIRO, F. A. N. SANTOS, M. DESROCHES, AND S. RODRIGUES, *Alternative set-theoretical algorithms for efficient computations of cliques in vietoris-rips complexes*, 2025, <https://arxiv.org/abs/2502.14593>, <https://arxiv.org/abs/2502.14593>.
- [10] D. DHEERU AND E. KARRA TANISKIDOU, *Breast Cancer Wisconsin (Diagnostic) Data Set*, 2017, [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)). Accessed: 2024-11-13.
- [11] P. DLOTKO AND S. RUDKIN, *Persistence Norms and the Datasaurus*. arXiv e-print 2309.13479, <https://arxiv.org/abs/2309.13479>, 2023.
- [12] H. EDELSBRUNNER, J. HARER, ET AL., *Persistent Homology – a Survey*, Contemporary Mathematics, 453 (2008), pp. 257–282.
- [13] H. EDELSBRUNNER AND J. L. HARER, *Computational Topology: An introduction*, American Mathematical Society, 2022.
- [14] H. B. ENDERTON, *Elements of set theory*, Academic press, 1977.
- [15] V. ERBA, S. ARIOSTO, M. GHERARDI, AND P. ROTONDO, *Random geometric graphs in high dimension*, Physical Review E, 102 (2020), p. 012306.
- [16] R. FORMAN, *Bochner's method for cell complexes and combinatorial Ricci curvature*, Discrete and Computational Geometry, 29 (2003), pp. 323–374.
- [17] S. FORTUNATO, *Community detection in graphs*, Physics Reports, 486 (2010), pp. 75–174.
- [18] C. GILLESPIE, S. LOCKE, R. DAVIES, AND L. D'AGOSTINO McGOWAN, *datasauRus: Datasets from the Datasaurus Dozen*, 2024, <https://github.com/jumpingrivers/datasauRus>. R package version 0.1.8, <https://jumpingrivers.github.io/datasauRus/>.
- [19] A. A. HAGBERG, D. A. SCHULT, AND P. J. SWART, *NetworkX: Network Analysis in Python*. Accessed: 2023-11-13, 2008. <https://networkx.org>.
- [20] T. JECH, *Set theory: The third millennium edition, revised and expanded*, Springer, 2003.
- [21] D. JIN, Z. YU, P. JIAO, S. PAN, D. HE, J. WU, S. Y. PHILIP, AND W. ZHANG, *A survey of community detection approaches: From statistical modeling to deep learning*, IEEE Transactions on Knowledge and Data Engineering, 35 (2021), pp. 1149–1170.
- [22] O. KNILL, *A discrete gauss-bonnet type theorem*, Elemente der Mathematik, 67 (2012), pp. 1–17.
- [23] O. KNILL, *An index formula for simple graphs*. arXiv e-print 1205.0306, <https://arxiv.org/abs/1205.0306>, 2012.
- [24] O. KNILL, *On index expectation curvature for manifolds*. arXiv e-print 2001.06925, <https://arxiv.org/abs/2001.06925>, 2020.
- [25] O. KNILL, *Gauss-Bonnet for Form Curvatures*. arXiv e-print 2409.01425, <https://arxiv.org/abs/2409.01425>, 2024.

- [26] Y. LIN, L. LU, AND S.-T. YAU, *Ricci curvature of graphs*, Tohoku Mathematical Journal, 63 (2011), pp. 605–627, <https://doi.org/10.2748/tmj/1325886283>, <http://projecteuclid.org/euclid.tmj/1325886283>.
- [27] D. J. MARCHETTE, *Random Graphs for Statistical Pattern Recognition*, John Wiley & Sons, 2005.
- [28] J. MATEJKÁ AND G. FITZMAURICE, *Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing*, in Proceedings of the 2017 CHI conference on human factors in computing systems, 2017, pp. 1290–1294.
- [29] L. MCINNES, J. HEALY, AND J. MELVILLE, *UMAP: Uniform Manifold Approximation and Projection*, 2018. <https://umap-learn.readthedocs.io>.
- [30] L. MCINNES, J. HEALY, AND J. MELVILLE, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv e-print 1802.03426, <https://arxiv.org/abs/1802.03426>, 2018, <https://arxiv.org/abs/1802.03426>.
- [31] M. E. J. NEWMAN, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167–256.
- [32] Z. PAWLAK, *Rough set theory and its applications*, Journal of Telecommunications and Information Technology, (2002), pp. 7–10.
- [33] M. PENROSE, *Random Geometric Graphs*, vol. 5, Oxford University Press, 2003.
- [34] T. G. PROJECT, *GUDHI: Geometry Understanding in Higher Dimensions*, 2023. Version 3.6.0, <https://gudhi.inria.fr>.
- [35] PYTHON SOFTWARE FOUNDATION, *Python Language Reference, version 3.13*, 2023. <https://docs.python.org/3/>.
- [36] M. SAGGAR, O. SPORNS, J. GONZALEZ-CASTILLO, P. A. BANDETTINI, G. CARLSSON, G. GLOVER, AND A. L. REISS, *Towards a new approach to reveal dynamical organization of the brain using topological data analysis*, Nature Communications, 9 (2018), pp. 1–14.
- [37] A. SAMAL, R. SREEJITH, J. GU, S. LIU, E. SAUCAN, AND J. JOST, *Comparative analysis of two discretizations of Ricci curvature for complex networks*, Scientific Reports, 8 (2018), pp. 1–16.
- [38] R. S. SANDHU, T. T. GEORGIOU, AND A. R. TANNENBAUM, *Ricci curvature: An economic indicator for market fragility and systemic risk*, Science Advances, 2 (2016), p. e1501495.
- [39] Z. XU, S. WEN, J. WANG, G. LIU, L. WANG, Z. YANG, L. DING, Y. ZHANG, D. ZHANG, J. XU, ET AL., *Amcad: adaptive mixed-curvature representation based advertisement retrieval system*, in 2022 IEEE 38th International Conference on Data Engineering (ICDE), IEEE, 2022, pp. 3439–3452.
- [40] J. ZHOU, G. CUI, S. HU, Z. ZHANG, C. YANG, Z. LIU, L. WANG, C. LI, AND M. SUN, *Graph neural networks: A review of methods and applications*, AI open, 1 (2020), pp. 57–81.
- [41] H.-J. ZIMMERMANN, *Applications of fuzzy set theory to mathematical programming*, in Fuzzy Sets for Intelligent Systems, D. Dubois, H. Prade, and R. Yager, eds., Elsevier, 1993, pp. 795–809.
- [42] A. ZOMORODIAN, *Fast construction of the Vietoris–Rips complex*, Computers & Graphics, 34 (2010), pp. 263–271.
- [43] A. ZOMORODIAN, *Topological data analysis*, Advances in Applied and Computational Topology, 70 (2012), pp. 1–39.
- [44] A. J. ZOMORODIAN, *Topology for Computing*, vol. 16, Cambridge University Press, 2005.

Appendix A. Supplementary Material. Herein, we provide the theoretical results used to derive the set-theoretical formulations that extend the computation of FRC to VR complexes. We also provide the algorithms and the proofs for our theoretical findings. We also include a few additional figures.

A.1. Theoretical Results. Below, we provide a set-theoretical representation of simplicial complexes, and we give an alternative formulation to the FRC. This derivation used the disjoint union $N_\alpha = P_\alpha \sqcup T_\alpha$ and the result $|T_\alpha| = (d+1)|H_\alpha|$, where N_α , T_α and P_α are the sets of neighbours, transverse neighbours and parallel neighbours to α , respectively. Using elementary algebra, we simplify the computation for FRC to

$$(A.1) \quad F(\alpha) = (d+2) \cdot |H_\alpha| + 2 \cdot (d+1) - |N_\alpha|,$$

where N_α is the set of neighbors of α . We also obtain that

$$(A.2) \quad N_\alpha = \bigsqcup_{\substack{\gamma \in \partial(\alpha) \\ x \in \pi_\gamma \neq \emptyset \\ x \notin \alpha}} \{\gamma \cup \{x\}\},$$

and hence, we can rewrite $|N_\alpha|$ as

$$(A.3) \quad |N_\alpha| = \sum_{\gamma \in \partial\alpha} |\pi_\gamma| - (d+1).$$

Finally we derive the number of $(d+1)$ -cells containing α as

$$(A.4) \quad |H_\alpha| = \left| \bigcap_{\gamma \in \partial\alpha} \pi_\gamma \right|.$$

Equations (A.3) and (A.4) together with (A.1) provide a new formulation for FRC computation, namely:

$$(A.5) \quad F(\alpha) = (d+2) \cdot \left| \bigcap_{\gamma \in \partial\alpha} \pi_\gamma \right| + 2 \cdot (d+1) - \sum_{\gamma \in \partial\alpha} |\pi_\gamma|.$$

This last formulation is crucial in the implementation of our novel FRC algorithm as a function of cutoff distance.

A.2. Algorithms. We now give details on the algorithms used in our approach. The [Algorithm A.1](#) is a simplified (basic) version of the proposed algorithm for computing FRC in VR complexes. The [Algorithm A.2](#) computes FRC in the function of the distance for VR complexes, and provides a more detailed version, which includes variable declarations and interpolation in the post-processing data. The [Al-](#)

Algorithm A.1 Compute Average local and global Forman-Ricci Curvature

```

Input:  $C, d_{\max} \geq 1$ 
Set  $F_d(x) := 0, \forall x \in V$ 
Set  $c_d := 0, \forall d \in \{1, \dots, d_{\max}\}$ 
Set  $w = \infty$ 

for  $\alpha \in C$  do
    Set  $d := |\alpha| - 1$ 
     $c_d := c_d + 1$ 
    Update the neighborhood of each  $\gamma \in \partial(\alpha)$ 
    Compute  $w_\alpha$  the diameter of  $\alpha$ 
    Compute  $F_d(\alpha)$  and  $F_d(x)$  according to (2.8) and the help of (3.6), i.e.:
    Set the number of neighbors of  $\alpha$ ,  $n := 0$ 
    Compute  $H := \bigcap_{\gamma \in \partial(\alpha)} \pi_\gamma$ 
    for  $\gamma \in \partial(\alpha)$  do
        update  $n := n + |\pi_\gamma|$ 
    end for

    Set  $f_d := F_d(\alpha) = (d + 2)|H| + 2 \cdot (d + 1) - n$ 

    for  $x \in V$  do
        Update the contribution of  $F(\alpha)$  to the node  $x$ , i.e.,  $F_d(x) := F_d(x) + f_d / (d + 1)$ 
    end for
    for  $\gamma \in \partial(\alpha)$  do
        for  $x \in \pi_\gamma \setminus \alpha$  do
             $\delta := \delta(x)$ 
            for  $y \in \partial(\alpha)$  do
                Update local total FRC for the nodes in the boundary:  $F_d(y) := F_d(y) + \delta / (d + 1)$ 
            end for
            Update the total FRC:  $f_d := f_d + \delta$ 
        end for
    end for
    end for
    if  $w \neq w_\alpha$  then
        if  $c_d \neq 0$  then
            print  $(w, \frac{f_d}{c_d}, (\frac{F_d(x)}{c_d}, \forall x \in V))$ 
        else
            print  $(w, 0, (0, \forall x \in V))$ 
        end if
    end if
    Update  $w := w_\alpha$ 
end for

```

gorithm A.3 was used to randomize n -dimensional point cloud data so that the original statistics of the points are maintained (with an error of 0.1). In particular, it was used to randomize the Datasaurus dataset (see Figure 12), where $n = 2$, scale = 0.5 and temp = 1. It is worth noting that the algorithm performs several steps of point randomization in its iterations, however, they are not always effective as the randomization is not performed when the statistical conditions are not reached in that specific iteration. Therefore, a high number of total steps must be performed in order to effectively randomize the data. For instance, in Figure 3, a total of 90000 iterations were performed to obtain that 17469 sequential iterations provide effective data randomization.

Algorithm A.2 Compute Local and Global Forman-Ricci Curvature (FRC)

1: **Input:** Distance matrix D , maximum dimension d_{\max} , maximum distance max_dist, precision p
 2: **Output:** Average local FRC and total FRC for nodes V
 3:
 4: Initialize node set V with $|V|$ nodes (if not provided, assume $V = \{0, 1, \dots, |D|\}$)
 5: Set cutoff step size $\delta \leftarrow 10^{-p}$ and compute cutoffs $\leftarrow [0, \text{max_dist}]$ with step δ
 6: Initialize:
 7: $C_d \leftarrow 0$, $F_d \leftarrow 0$ (global curvature), $nF_d(x) \leftarrow 0 \forall x \in V$
 8: Neighborhood $N_d \leftarrow \{\}$ for $d \in \{1, \dots, d_{\max}\}$
 9: **if** $d_{\max} = 1$:
 10: Generate edge list E and corresponding distances w_{ij} sorted by increasing order
 11: **else:**
 12: Use `cliques_gudhi` to generate cliques α with $w_\alpha \leq \text{max_dist}$
 13:
 14: **for** each clique α with weight w_α **do**
 15: $d \leftarrow |\alpha| - 1$ {Dimension of the clique}
 16: $C_d \leftarrow C_d + 1$ {Count the clique}
 17: Compute boundary $\partial(\alpha) \leftarrow \{B \mid B \subset \alpha, |B| = d\}$
 18: **for** each boundary $B \in \partial(\alpha)$ **do**
 19: Update Neighborhood: $N_d[B] \leftarrow N_d[B] \cup (\alpha \setminus B)$
 20: **end for**
 21: Compute $H \leftarrow \bigcap_{B \in \partial(\alpha)} N_d[B]$ {Intersection of neighbors of boundaries}
 22: Compute total neighbors $n \leftarrow \sum_{B \in \partial(\alpha)} |N_d[B]|$
 23: Compute curvature: $f_d \leftarrow (d + 2) \cdot |H| + 2 \cdot (d + 1) - n$
 24: **for** each node $x \in \alpha$ **do**
 25: $nF_d(x) \leftarrow nF_d(x) + f_d / (d + 1)$
 26: **end for**
 27: **for** each boundary $B \in \partial(\alpha)$ **do**
 28: **for** each node $x \in N_d[B] \setminus \alpha$ **do**
 29: Compute $\delta \leftarrow \delta(x \in H, d)$ {Transverse or parallel neighbor}
 30: **for** each node $y \in B$ **do**
 31: $nF_d(y) \leftarrow nF_d(y) + \delta / (d + 1)$
 32: **end for**
 33: $nF_d(x) \leftarrow nF_d(x) + \delta / (d + 1)$
 34: Update total curvature: $f_d \leftarrow f_d + \delta$
 35: **end for**
 36: **end for**
 37: Update global curvature: $F_d \leftarrow F_d + f_d$
 38: **if** cutoff w_α changes from previous w **then**
 39: **if** $C_d > 0$ **then**
 40: Output: $(w, C_d / \text{combs}_d, F_d / C_d, \{nF_d(x) / C_d, \forall x \in V\})$
 41: **else**
 42: Output: $(w, 0, 0, \{0, \forall x \in V\})$
 43: **end if**
 44: Update $w \leftarrow w_\alpha$
 45: **end if**
 46: **end for**
 47: Fill in missing cutoff results by interpolating last valid curvature values
 48: Return final FRC values: Average local FRC and total FRC for nodes

Algorithm A.3 Dataset Perturbation

```

1: Input: Initial dataset initial_ds, Number of iterations iterations, Temperature
   temp , Perturbation scale.
2: Output: Final dataset current_ds with preserved statistical properties
3:
4: Set current_ds ← initial_ds
5:
6: Function FIT(ds):
7:   Return the sum of distances from the origin for all points in ds
8:
9: Function ISERROROK(test_ds, initial_ds):
10:   Calculate the mean and standard deviation for each coordinate in initial_ds
    and test_ds
11:   Calculate the correlation matrices for initial_ds and test_ds
12:   Round all values to 3 decimal places
13:   Return True if means, standard deviations, and correlations match to 3 dec-
    imal places; otherwise, return False
14:
15: Function MOVERANDOMPOINTS(ds):
16:   Create a copy of ds as test_ds
17:   Select a random index idx from ds
18:   Generate a random movement vector from a normal distribution with small-
    scale
19:   Move the point at test_ds[idx] by the movement vector
20:   Return test_ds
21:
22: Function RANDOM(scale):
23:   Return a random value between 0 and 1 (with standard deviation = scale)
24:
25: Function PERTURB(ds, temp):
26:   Loop until a valid perturbation is found:
27:     test ← MOVERANDOMPOINTS(ds)
28:     If FIT(test) > FIT(ds) or temp > RANDOM():
29:       Return test
30:
31: Main Function SIMULATEDANNEALING(initial_ds, iterations, temp):
32:   Set current_ds ← initial_ds
33:   For each iteration in 1 to iterations:
34:     test_ds ← PERTURB(current_ds, temp)
35:     If ISERROROK(test_ds, initial_ds):
36:       Update current_ds ← test_ds
37:   Return current_ds

```

A.3. Main results and Demonstrations. In this section, we develop the theoretical assumptions of our work.

PROPOSITION 1. Let $\alpha_i \in C_d$, $\alpha \in N^i$. Let $F : C_d \rightarrow \mathbb{Z}$ and F^i be the FRC function as defined in 2.6. If α is parallel to α_i , then $F^i(\alpha) = F^{i-1}(\alpha) - 1$. Otherwise, we have $F^i(\alpha) = F^{i-1}(\alpha) + (d + 1)$.

Proof. We have that $F^{i-1}(\alpha) = |H_\alpha^{i-1}| + (d+1) - |P_\alpha^{i-1}|$. Suppose that α is parallel to α_i . Thus, $P_\alpha^i = P_\alpha^{i-1} \sqcup \{\alpha_i\}$, which implies that $|P_\alpha^i| = |P_\alpha^{i-1}| + 1$ and that $H_\alpha^i = H_\alpha^{i-1}$. It follows that $F^i(\alpha) = |H_\alpha^i| + (d+1) - |P_\alpha^i| = |H_\alpha^{i-1}| + (d+1) - |P_\alpha^{i-1}| - 1 = F^{i-1}(\alpha) - 1$. Suppose that $\alpha \notin P^{i+1}$. Then, $\alpha \in T^{i+1}$, which implies that $|H_\alpha^i| = |H_\alpha^{i-1}| + 1$ and $P_\alpha^i = P_\alpha^{i-1}$. It follows that $F_\alpha^i = |H_\alpha^i| + (d+1) - |P_\alpha^i| = (d+1) \cdot (|H_\alpha^{i-1}| + 1) + (d+1) - |P_\alpha^{i-1}| = F^{i-1}(\alpha) + (d+1)$. \square

THEOREM A.1 (Geometric Gauss-Bonnet Theorem). *Let C_d be a non-empty set of d -faces in a simplicial complex C generated from an undirected graph $G = (V, E)$. Let F_d and f_d be the global and local FRC definitions, as defined in (2.7) and (3.1), respectively. Then, the following equality holds:*

$$(A.6) \quad F_d(C) = \sum_{x \in V} f_d(x).$$

Proof. It is sufficient to notice that

$$(A.7) \quad \sum_{x \in V} \sum_{\substack{\alpha \subset C_d \\ x \in \alpha}} F_d(\alpha) = (d+1) \sum_{\alpha \subset C_d} f_d(\alpha),$$

once that for each $\alpha \in C_d$, the value $f_d(\alpha)$ is counted exactly $(d+1)$ times in the sum over all the nodes $x \in V$. The result follows by multiplying both sides of the equation above by $\frac{1}{(d+1) \cdot |C_d|}$. \square

A.4. Additional figures. We now provide all the supplementary figures of our work, which include examples, illustrations and results. [Figure 9](#) provides an example of FRC value change as a function of the local neighbourhood. [Figure 12](#) shows the 2D plots of the Datasaurus dataset.

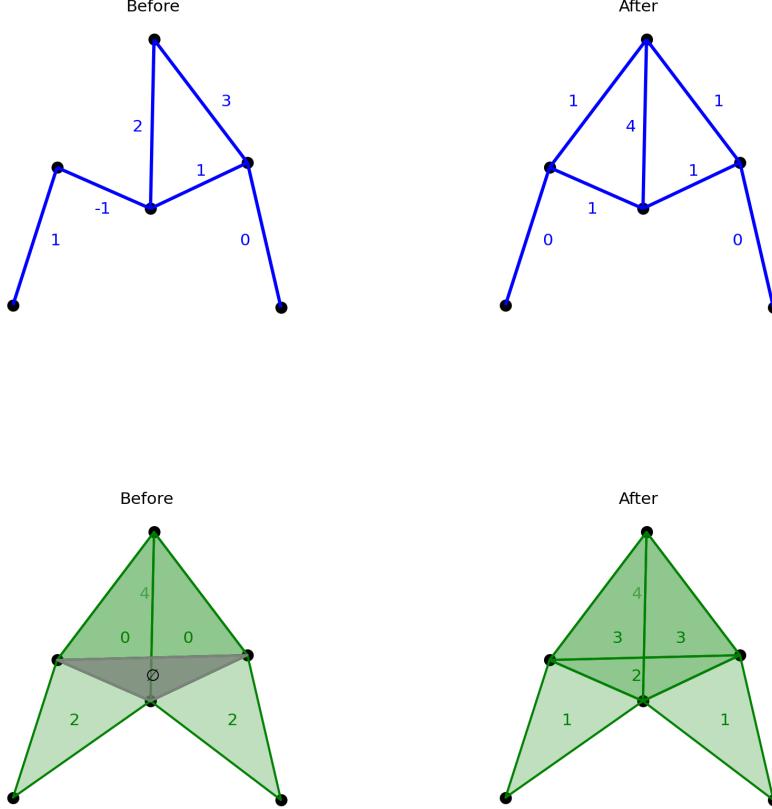


FIG. 9. Example of an evolving simplicial complex (by adding a new face) and how it influences the FRC (the numerical values given in the figure) for edges (blue) and triangles (green). The new edge has 4 neighbours, of which two are transverse (the neighbours sharing the same triangle) and 2 parallel (The ones outside the triangle). The 1-FRC of the new edge is the number of triangles containing the new edge plus the length of the boundary minus the number of parallel neighbours, i.e., $1 + 2 - 2 = 1$. When a new edge is added, the FRC increases by 2 units when a new triangle is created and decreases by 1 unit for the new neighbours outside the new triangle. Similarly, the new triangle has 3 neighbours, where 1 shares a tetrahedron and 2 are parallel neighbours, therefore, the 2-FRC of the new triangle is $1 + 3 - 2 = 2$. Similarly, for triangle faces, the FRC increases by 3 units when a new tetrahedron is created in the neighbourhood and decreases by 1 unit otherwise.

A.5. Random geometric graphs. Figures 10 and 11 provide the FRC computations on random geometric graphs for different edge densities and box dimensions, as well as the comparison with the geometric classification performed by UMAP. In

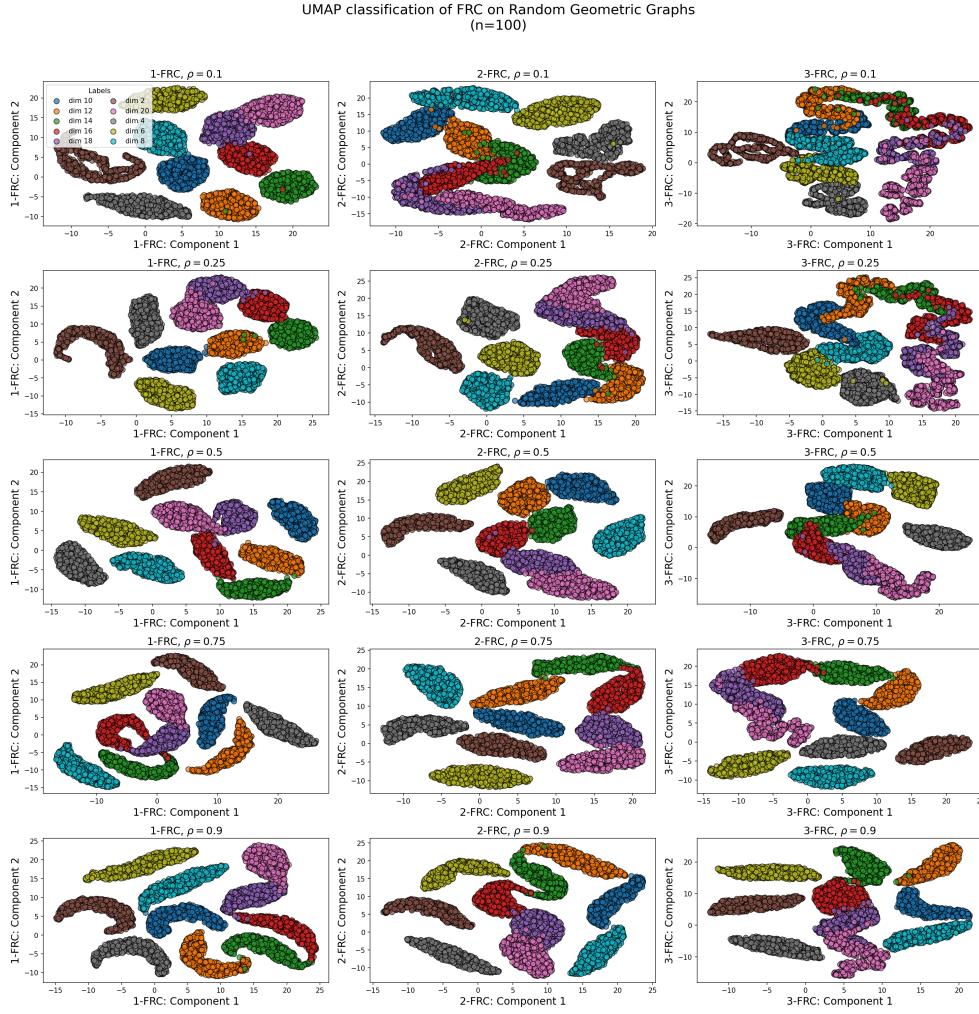


FIG. 10. Computations of the average d -FRC for $d = 1$ (left column), $d = 2$ (middle column) and $d = 3$ (right column) on Random geometric graphs with $n = 100$ nodes and different box dimensions and densities (plot lines). The solid-coloured lines are the average, while the error bands are computed from the standard deviation. The box dimension classification can be better visualised in Figure 11.

all tests, we used the Euclidean distance and minimum distance of 1.0 as UMAP parameters.

A.6. Datasaurus dataset. Here, we provide the result of the data randomization process of the Datasaurus datasets performed by Algorithm A.3, as well as its FRC computation for data classification with UMAP and sensitivity to noise. The original dataset is shown in Figure 12. In Figures 13 and 14 we provide the comparison of the FRC performance and classification in the presence of different levels of noise.

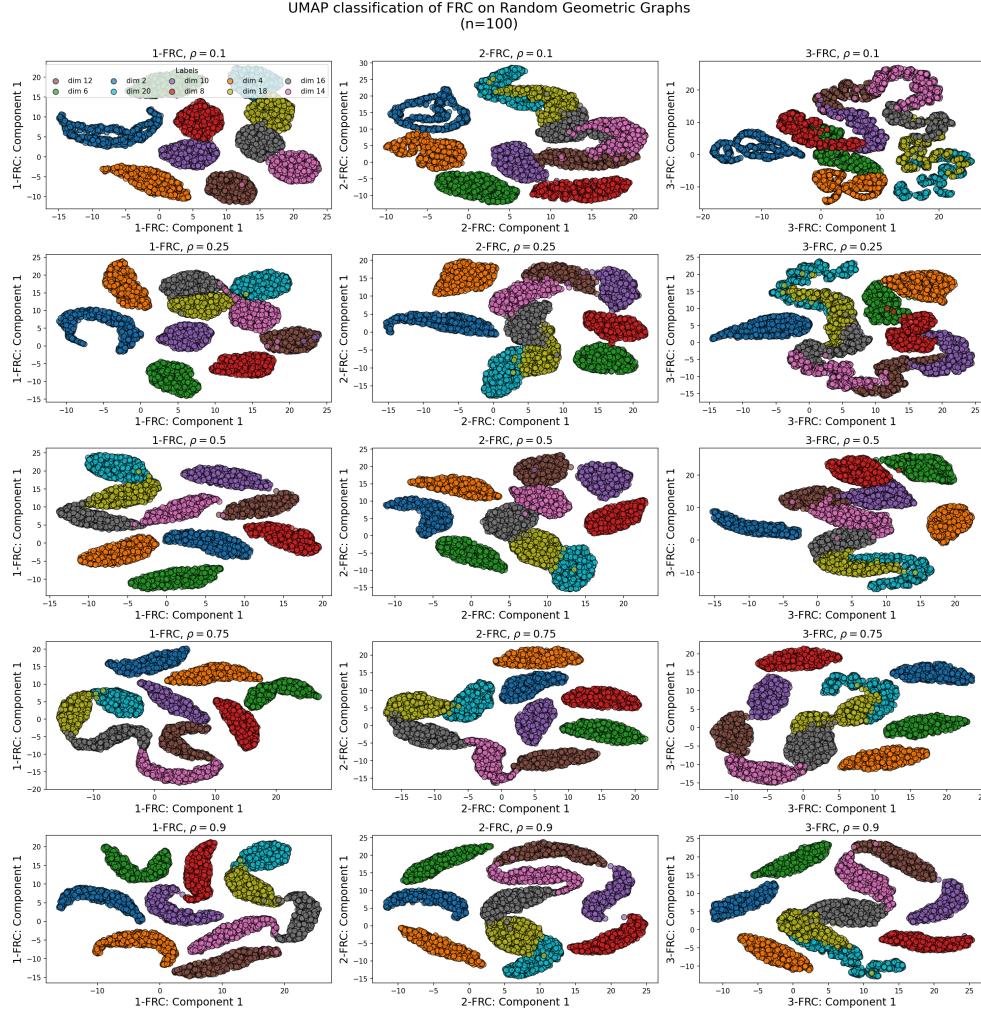


FIG. 11. UMAP classification from the computations of the average d -FRC for $d = 1$ (left column), $d = 2$ (middle column) and $d = 3$ (right column) on random geometric graphs with $n = 100$ nodes and different box dimensions and densities; see also [Figure 10](#)

A.7. Breast cancer Wisconsin database. Here, we provide the FRC computations on the Wisconsin Breast Cancer database. [Figures 15 to 18](#) show FRC computations and their sensitivity to cutoff intervals and UMAP parameters.

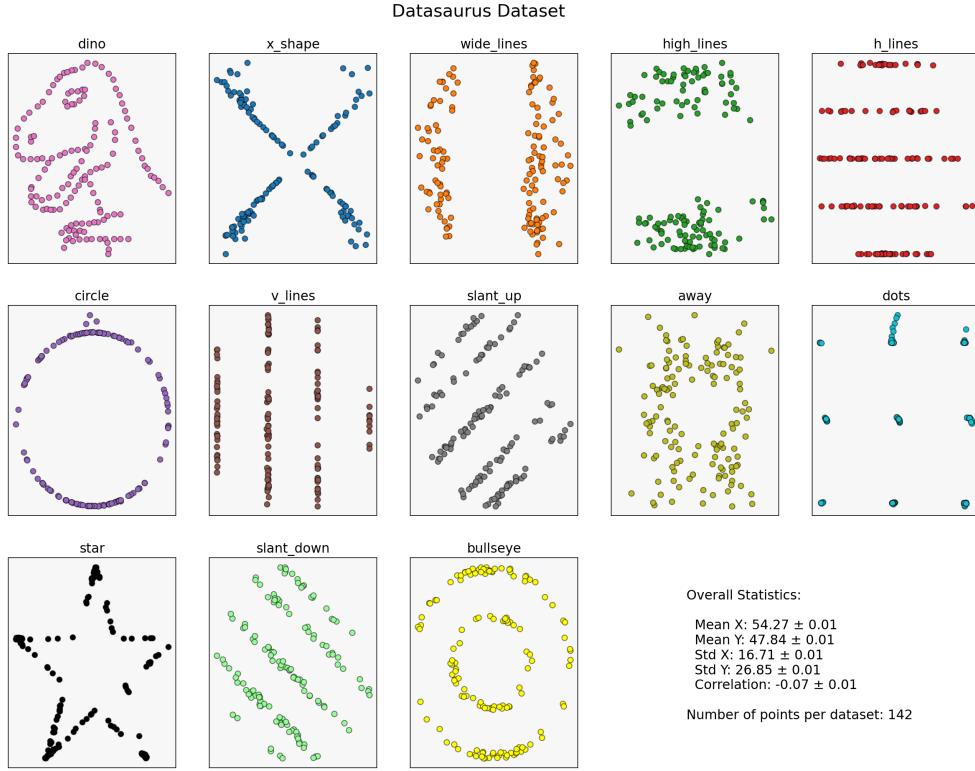


FIG. 12. The Datasaurus dataset and its overall statistics. The purpose of this dataset was to generate 2D point cloud data that provides different geometries with (proximally) the same basic statistics.

A.8. Breast cancer METABRIC dataset. Here, we provide the FRC computations on the breast cancer METABRIC database. Figures 19 to 22 display FRC computations and its sensitivity to cutoff intervals and UMAP parameters.

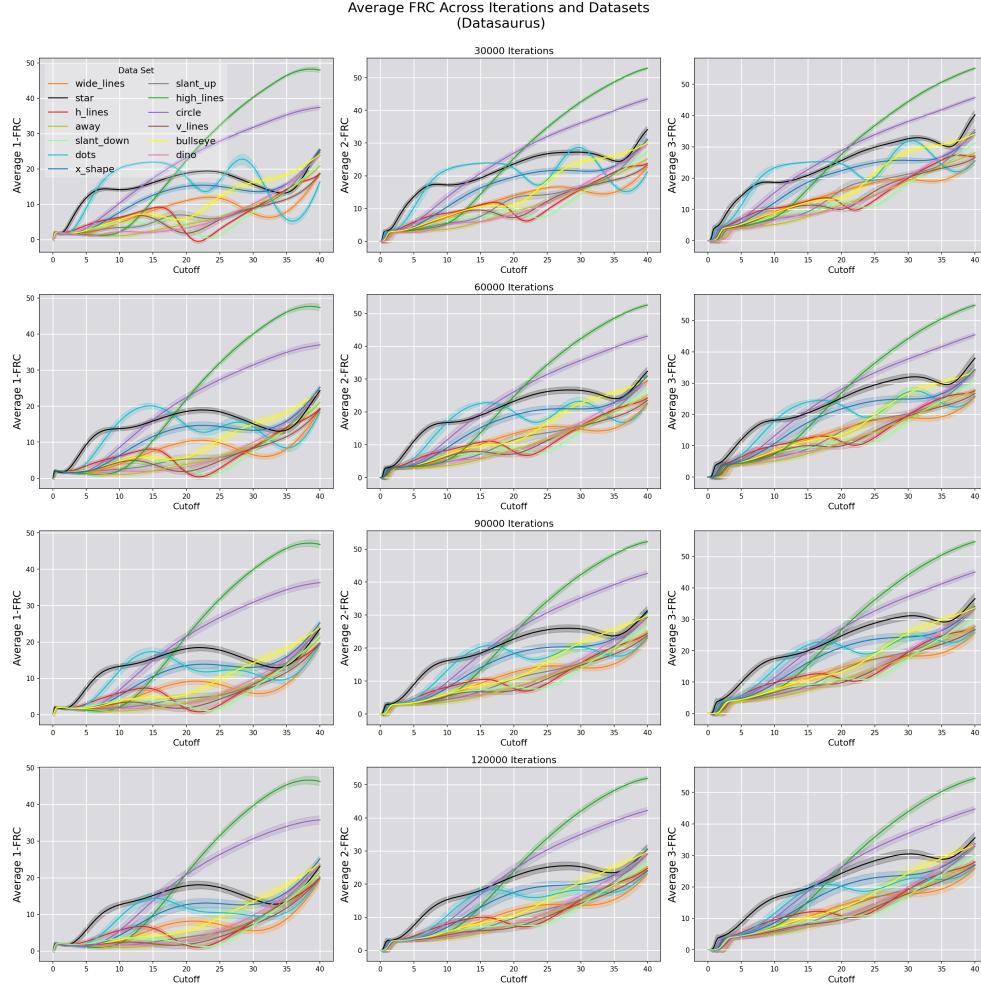


FIG. 13. Average k -FRC statistical computation on Datasaurus datasets, for $d = 1$ (left column), $d = 2$ (middle column) and $d = 3$ (right column), and different number of maximum randomization steps (lines). The result is the mean curvatures (central coloured lines) and the error bands (computed from the standard deviation). This result can be compared with the UMAP classification in Figure 14.

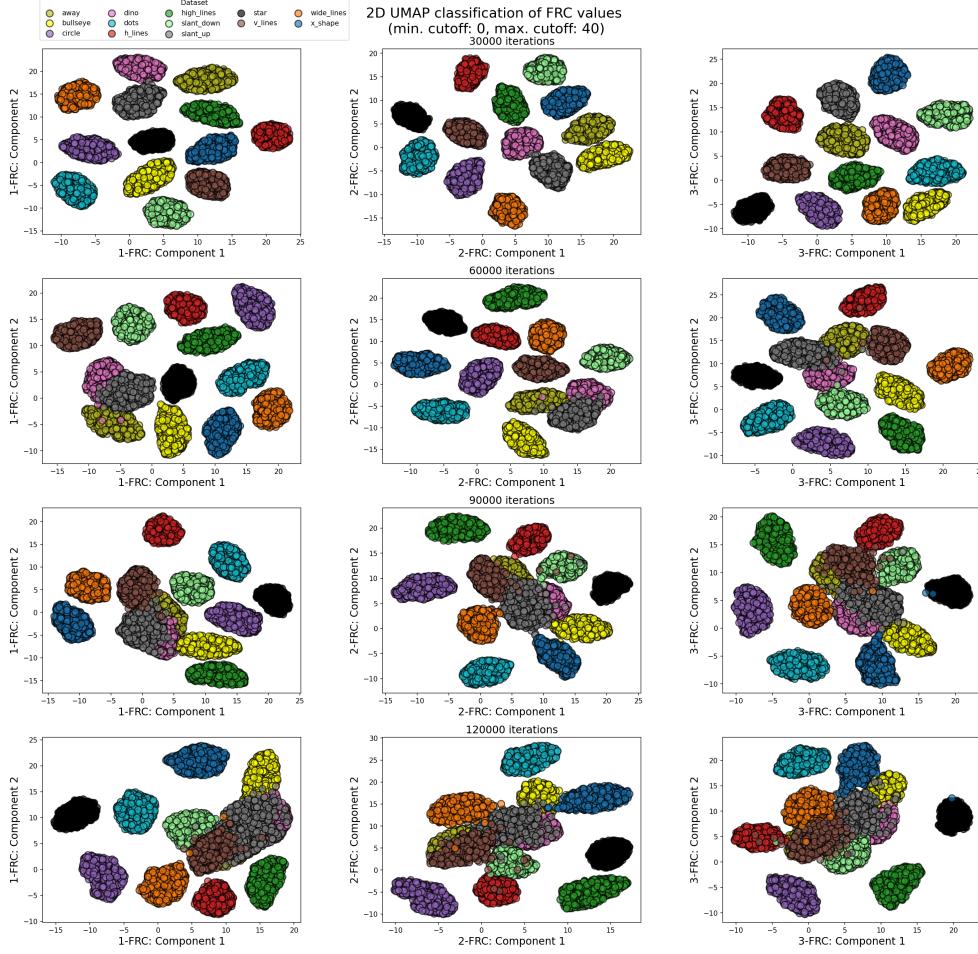


FIG. 14. UMAP classification of the average k -FRC statistical computation on Datasaurus datasets, for $d = 1$ (left column), $d = 2$ (middle column) and $d = 3$ (right column), and different number of maximum randomization steps (lines). Compare with

[Figure 13.](#)

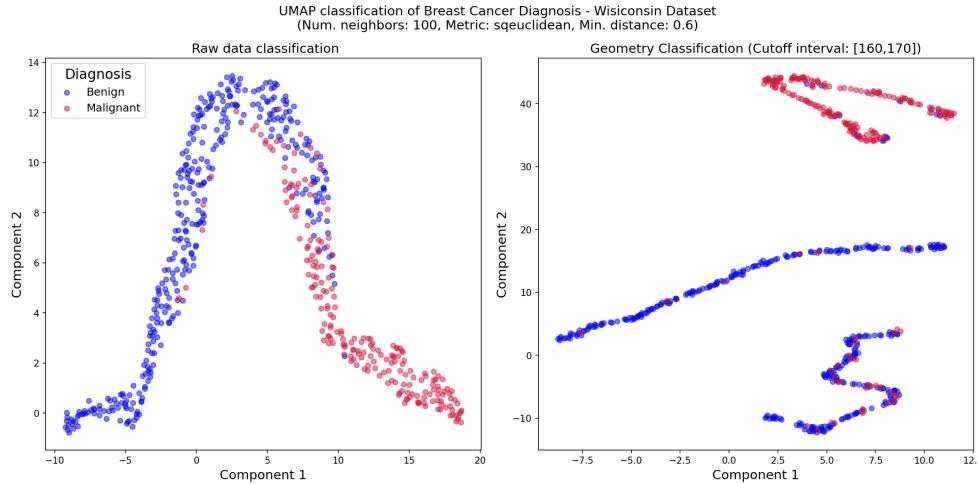


FIG. 15. *Breast cancer diagnosis comparison between UMAP classification and raw data input from Wisconsin dataset (left) and its geometrized version(right).*

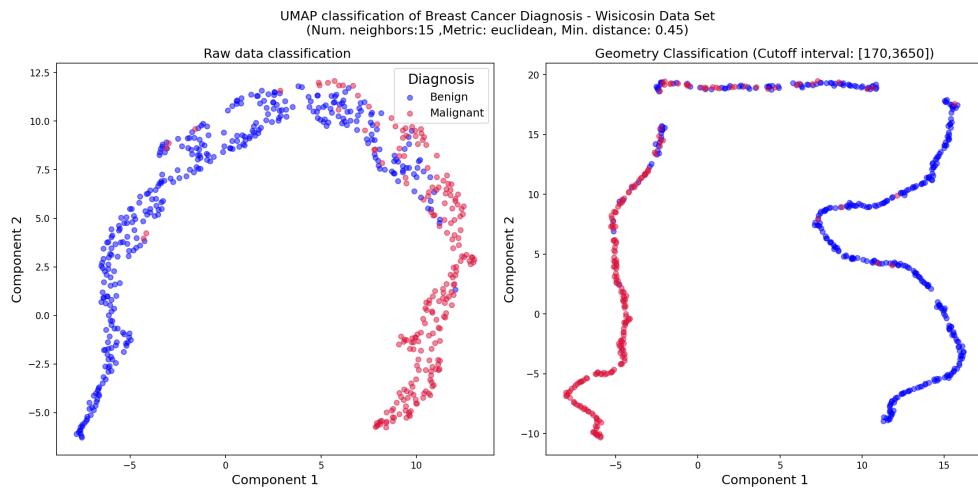


FIG. 16. *Breast cancer diagnosis comparison between UMAP classification and raw data input from Wisconsin dataset (left) and its geometrized version(right).*

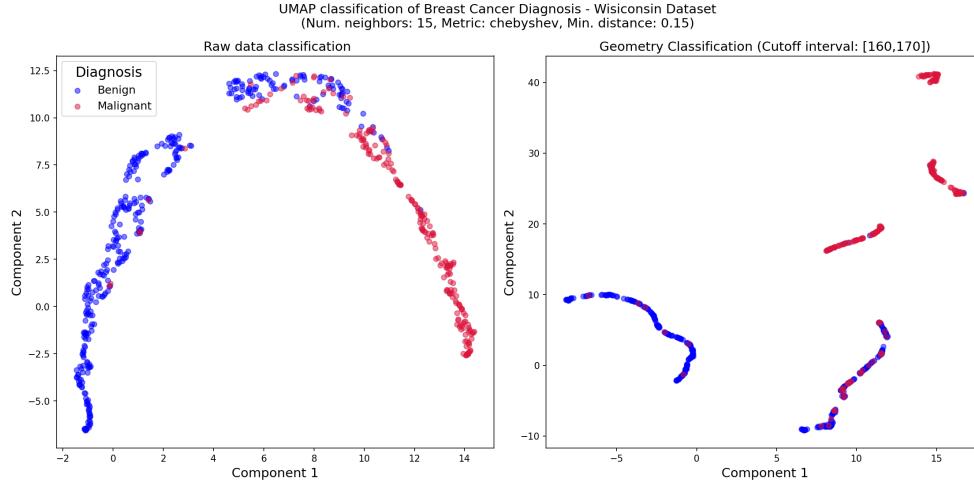


FIG. 17. *Breast cancer diagnosis comparison between UMAP classification and raw data input from Wisconsin dataset (left) and its geometrized version(right).*

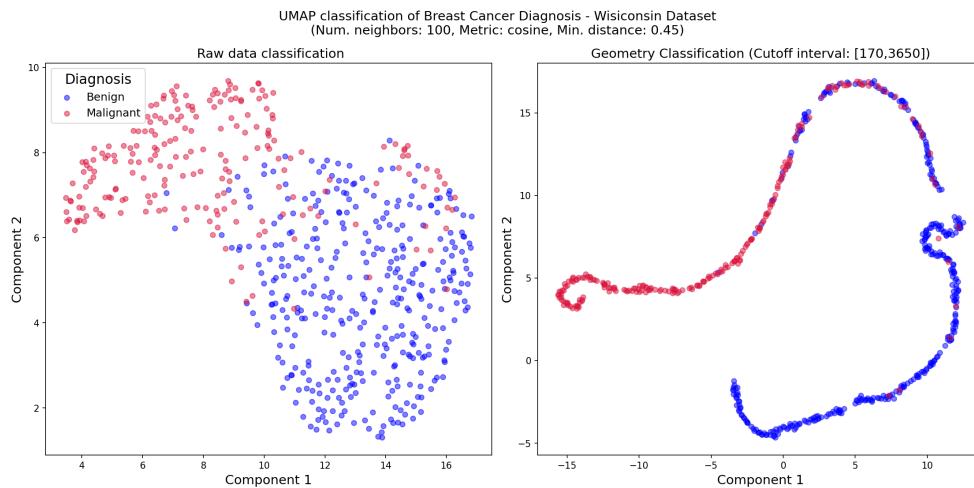


FIG. 18. *Breast cancer diagnosis comparison between UMAP classification and raw data input from Wisconsin dataset (left) and its geometrized version(right).*

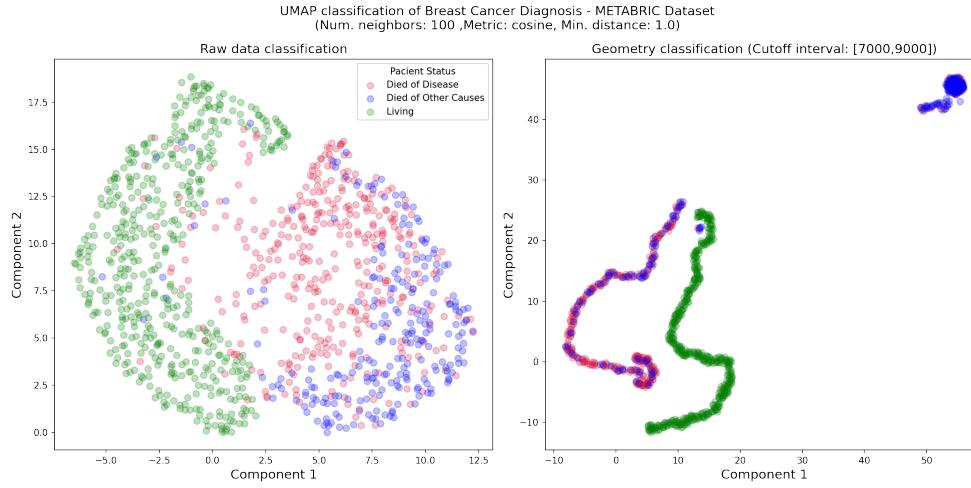


FIG. 19. *Breast cancer diagnosis comparison between UMAP classification and raw data input from METABRIC dataset (left) and its geometrized version(right).*

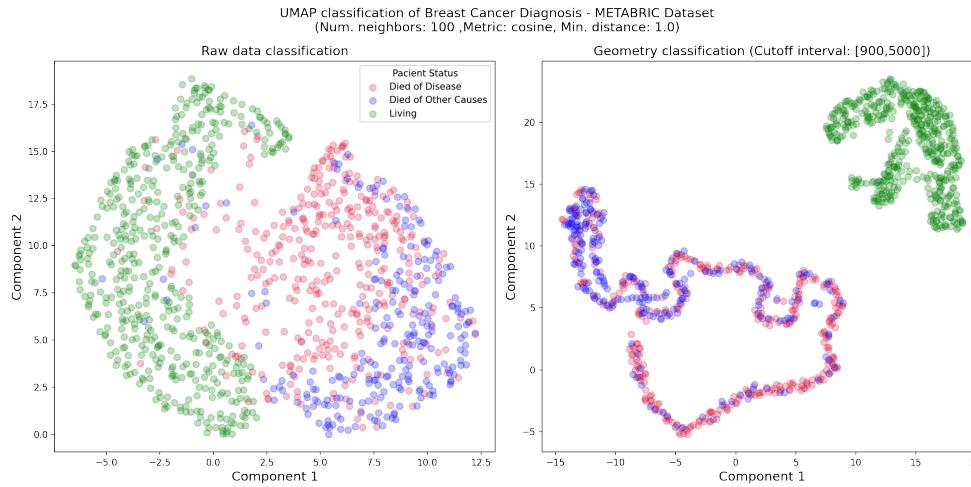


FIG. 20. *Breast cancer diagnosis comparison between UMAP classification and raw data input from METABRIC dataset (left) and its geometrized version(right).*

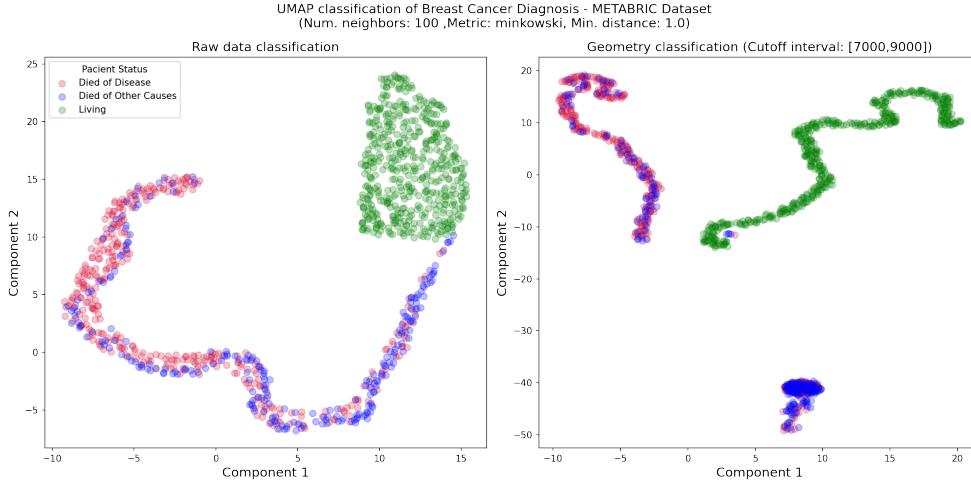


FIG. 21. *Breast cancer diagnosis comparison between UMAP classification and raw data input from METABRIC dataset (left) and its geometrized version(right).*

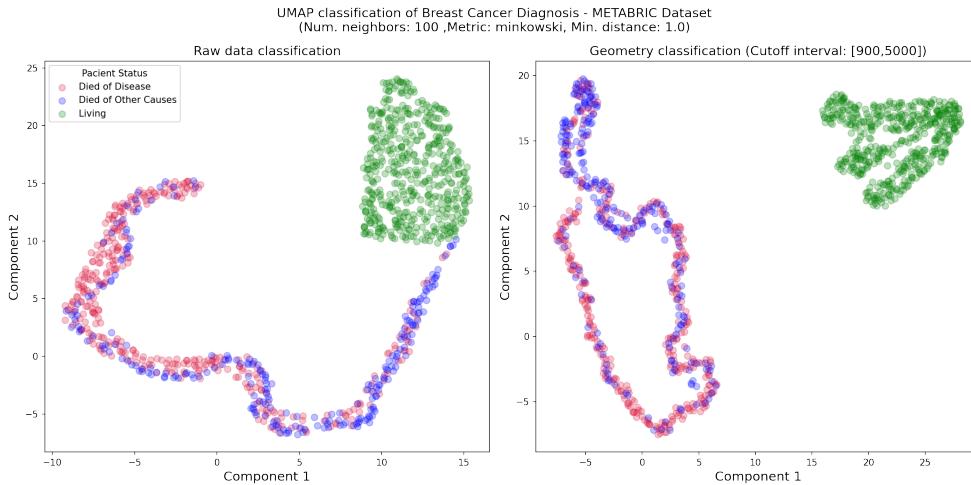


FIG. 22. *Breast cancer diagnosis comparison between UMAP classification and raw data input from METABRIC dataset (left) and its geometrized version(right).*