Persistent Homology for High-dimensional Data Based on Spectral Methods

Sebastian Damrich[†] **Philipp Berens**[†‡] **Dmitry Kobak**[†§]
[†] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
[‡] Tübingen AI Center, Germany
[§] IWR, Heidelberg University, Germany
{sebastian.damrich,philipp.berens,dmitry.kobak}@uni-tuebingen.de

Abstract

Persistent homology is a popular computational tool for analyzing the topology of point clouds, such as the presence of loops or voids. However, many real-world datasets with low intrinsic dimensionality reside in an ambient space of much higher dimensionality. We show that in this case traditional persistent homology becomes very sensitive to noise and fails to detect the correct topology. The same holds true for existing refinements of persistent homology. As a remedy, we find that spectral distances on the k-nearest-neighbor graph of the data, such as diffusion distance and effective resistance, allow to detect the correct topology even in the presence of high-dimensional noise. Moreover, we derive a novel closed-form formula for effective resistance, and describe its relation to diffusion distances. Finally, we apply these methods to high-dimensional single-cell RNA-sequencing data and show that spectral distances allow robust detection of cell cycle loops.

1 Introduction

Algebraic topology can describe the shape of a continuous manifold. In particular, it can detect if a manifold has holes, using its so-called homology groups [39]. For example, a cup has a single one-dimensional hole, or *loop* (its handle), whereas a football has a single two-dimensional hole, or *void* (its hollow interior). These global topological properties are often helpful for understanding an object's overall structure. However, real-world datasets are typically given as point clouds, a discrete set of points sampled from an underlying manifold. In this setting, true homologies are trivial, as there is one connected component per point and no holes whatsoever; instead, *persistent homology* can be used to find holes in point clouds and to assign an importance score called *persistence* to each [25, 90]. Holes with high persistence are indicative of holes in the underlying manifold. Persistence homology has been successfully applied in machine learning pipelines, for instance for gait recognition [48], instance segmentation [44], and protein binding [84], as well as for neural network analysis [67].

Persistent homology works well for low-dimensional data [78] but we find that it has difficulties in high dimensionality. If data points are sampled from a low-dimensional manifold embedded in a high-dimensional ambient space (*manifold hypothesis*), then the measurement noise typically affects all ambient dimensions. In this setting, traditional persistent homology is not robust against even low levels of noise. On a dataset as simple as a circle in $\mathbb{R}^{50}$, persistent homology based on the Euclidean distance between noisy points can fail to identify the correct loop as a clear outlier in the persistence diagram (Figure 1). The aim of our work is to find alternatives to traditional persistent homology that can robustly detect the correct topology despite high-dimensional noise.

We were inspired by visualization methods t-SNE [79] and UMAP [53] that are able to depict the loop in the same noisy dataset (Figure 1d,e). They approximate the data manifold by the k-nearest-neighbor (kNN) graph [76, 70, 5, 41, 56]. Therefore, we suggest to use persistent homology with spectral

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1328

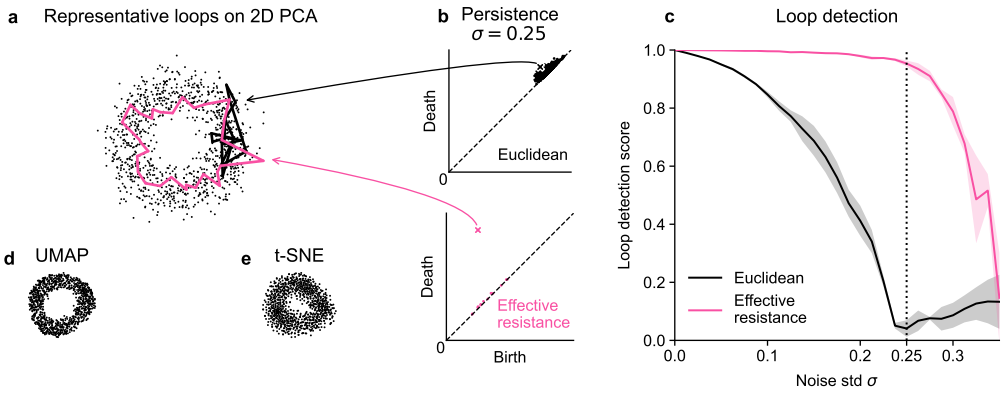

Figure 1: **a.** 2D PCA of a noisy circle ($\sigma = 0.25$, radius 1) in $\mathbb{R}^{50}$. Overlaid are representative cycles of the most persistent loops. **b.** Persistence diagrams using Euclidean distance and the effective resistance. **c.** Loop detection scores of persistent homology using effective resistance and Euclidean distance. **d, e.** UMAP and t-SNE embeddings of the same data, showing the loop structure in 2D.

distances on this kNN graph, such as the effective resistance [24] and the diffusion distance [20]. Effective resistance successfully identified the correct loop in the above toy example (Figure 1). We also found spectral distances to outperform other distances in detecting the correct topology of several synthetic datasets as well as finding the cell cycles in single-cell RNA-sequencing data.

Our contributions are:

1. an analysis of the failure modes of persistent homology for noisy high-dimensional data;
2. a closed-form expression for effective resistance, explaining its relation to diffusion distances;
3. a synthetic benchmark, with spectral distances outperforming state-of-the-art alternatives;
4. an application to a range of single-cell RNA-sequencing datasets with ground-truth cycles.

Our code is available at `https://github.com/berenslab/eff-ph/tree/neurips2024`.

2 Related work

Persistent homology has long been known to be sensitive to outliers [15] and several extensions have been proposed to make it more robust. One recurring idea is to replace the Euclidean distance with a different distance matrix, before running persistent homology. Bendich et al. [6] suggested to use diffusion distances [20], but their empirical validation was limited to a single dataset in 2D. Anai et al. [2] suggested to use the distance-to-measure (DTM) [15] and Fernández et al. [26] proposed to use Fermat distances [34]. Vishwanath et al. [80] introduced persistent homology based on robust kernel density estimation, an approach that itself becomes challenging in high dimensionality. All of these works focused on low-dimensional datasets (<10D, mostly 2D or 3D), while our work specifically addresses the challenges of persistent homology in high dimensionality.

The concurrent work of Hiraoka et al. [42] is the most relevant related work. Their treatment of the curse of dimensionality of persistent homology is mostly theoretical, while ours has an empirical focus. The two works are thus complementary to each other. Hiraoka et al.'s theoretical description of the curse of dimensionality is similar to ours (Appendix B) in that it analyses how distance concentration with high-dimensional noise impairs persistent homology, but is more general. Practically, Hiraoka et al. propose normalized PCA to mitigate the curse of dimensionality. However, this approach assumes the true dimensionality of the data to be known, which is not realistic in real-world applications, and performs worse than our suggestions (Appendix C).

Below, we recommend using effective resistance and diffusion distances for persistent homology in high-dimensional spaces. Both of these distances, as well as the shortest path distance, have been used in combination with persistent homology to analyze the topology of graph data [62, 37, 1, 77, 11, 55, 23]. Shortest paths on the kNN graph were also used by Naitzat et al. [60] and Fernández et al. [26]. Motivated by the performance of UMAP [53] for dimensionality reduction, Gardner et al. [31] and Hermansen et al. [40] used UMAP affinities to define distances for persistent homology.

https://doi.org/10.52202/079017-1328

Effective resistance is a well-established graph distance [24, 29]. A correction, more appropriate for large graphs, was suggested by von Luxburg et al. [81, 83]. When speaking of *effective resistance*, we mean this corrected version, if not otherwise stated. It has not yet been combined with persistent homology. Conceptually similar diffusion distances [20] have been used in single-cell RNA-sequencing data analysis, for dimensionality reduction [56], trajectory inference [35], feature extraction [16], and hierarchical clustering, similar to 0D persistent homology [9, 46].

Persistent homology has been applied to single-cell RNA-sequencing data, but only the concurrent work of Flores-Bautista and Thomson [27] applies it directly to the high-dimensional data. Wang et al. [85] used a Witness complex on a PCA of the data. Other works applied persistent homology to a derived graph, e.g., a gene regulator network [52] or a Mapper graph [73, 68]. In other biological contexts, persistent homology has also been applied to a low-dimensional representation of the data: 3D projection of cytometry data [58], 6D PCA of hippocampal spiking data [31], and 3D PHATE embedding of calcium signaling [57]. Several recent applications of persistent homology only computed 0D features (i.e. clusters) [37, 45, 61], which amounts to doing single linkage clustering [33]. Here we only investigate the detection of higher-dimensional (1D and 2D) holes with persistent homology. The dimensionality of the data itself, however, is typically much higher.

3 Background: persistent homology

Persistent homology computes topological invariants of a space at different scales. For point clouds, the different scales are typically given by growing a ball around each point (Figure 2a), and letting the radius τ grow from 0 to infinity. For each value of τ, homology groups of the union of all balls are computed to find the holes, and holes that *persist* for longer time periods are considered more prominent. At $\tau \approx 0$, there are no holes as the balls are non-overlapping, while at $\tau \to \infty$ there are no holes as all the balls merge together.

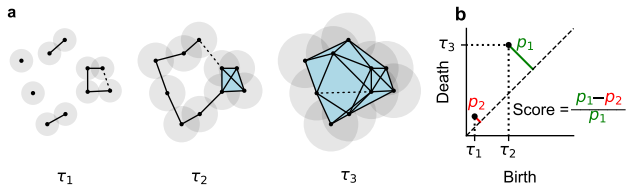

Figure 2: **a.** Persistent homology applied to a noisy circle ($n = 10$) in 2D tracks appearing and disappearing holes as balls grow around each datapoint. Dotted lines show the graph edges that lead to the birth / death of two loops (Section 3). **b.** The corresponding persistence diagram with two detected 1D holes (loops). Our *hole detection score* measures the gap in persistence between the first and the second detected holes (Section 7).

To keep the computation tractable, instead of the union of growing balls, persistent homology operates on a so-called *filtered simplicial complex* (Figure 2a). A simplicial complex is a hypergraph containing points as nodes, edges between nodes, triangles bounded by edges, and so forth. These building blocks are called *simplices*. At each time τ, the complex encodes all intersections between the balls and suffices to find the holes. The complexes at smaller τ values are nested within the complexes at larger τ values, and together form a filtered simplicial complex, with τ being the filtration time. In this work, we only use the Vietoris–Rips complex, which includes an n-simplex $(v_0, v_1, \ldots, v_n)$ at filtration time τ if the distances between all pairs v_i, v_j are at most τ. Therefore, to build a Vietoris–Rips complex, it suffices to provide pairwise distances between all pairs of points. We compute persistent homology via the `ripser` package [4] to which we pass a distance matrix.

Persistent homology consists of a set of holes for each dimension. We limit ourselves to loops and voids. Each hole has associated birth and death times (τ_b, τ_d), i.e., the first and last filtration value τ at which that hole exists. Their difference $p = \tau_d - \tau_b$ is called the *persistence* of the hole and quantifies its prominence. The birth and death times can be visualized as a scatter plot (Figure 2b), known as the *persistence diagram*. Points far from the diagonal have high persistence. This process is illustrated in Figure 2 for a noisy sample of $n = 10$ points from a circle $S^1 \subset \mathbb{R}^2$. At τ_1, a small spurious loop is formed thanks to the inclusion of the dotted edge, but it dies soon afterwards. The ground-truth loop is formed at τ_2 and dies at τ_3, once the hole is completely filled in by triangles. Both loops (one-dimensional holes) found in this dataset are shown in the persistence diagram.

https://doi.org/10.52202/079017-1328

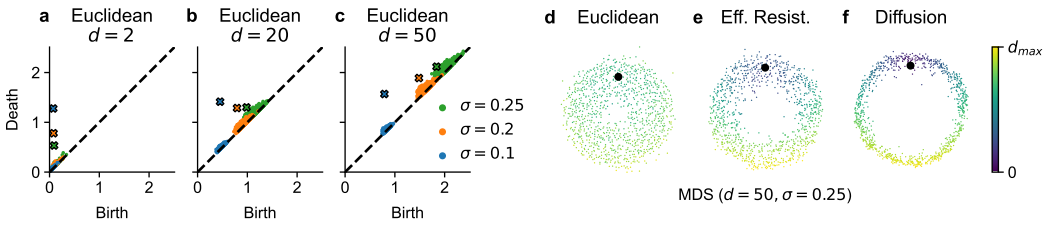

Figure 3: **a – c.** Persistence diagrams of a noisy circle in different ambient dimensionality and with different amount of noise. Ideally, there should be one feature (point) with high persistence, corresponding to the circle. But for high noise and dimensionality that feature vanishes into the noise cloud near the diagonal. **d – f.** Multidimensional scaling of Euclidean, effective resistance, and diffusion distances for a noisy circle in $\mathbb{R}^{50}$. Color indicates the distance to the highlighted point.

4 The curse of dimensionality for persistent homology

While persistent homology is robust to small changes in point positions [19], the curse of dimensionality can still severely hurt its performance. To illustrate, we consider the same toy setting as in Figure 1: we sample points from $S^1 \subset \mathbb{R}^d$, and add Gaussian noise of standard deviation σ to each ambient coordinate. When $d = 2$, higher noise does not affect the birth times but leads to lower death times (Figure 3a), because some points get distorted to the middle of the circle and the hole fills up at earlier τ. When we increase the ambient dimensionality to $d = 20$, higher noise leads to later birth times (Figure 3b) because in higher dimensionality distances get dominated by the noise dimensions rather than by the circular structure. Indeed, in Corollary B.5 we prove that for any two points $x_i, x_j \in \mathbb{R}^d$ and two isotropic, multivariate normal noise vectors $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ the ratio $\|\varepsilon_1 - \varepsilon_2\| / \|x_i + \varepsilon_1 - (x_2 + \varepsilon_2)\| \to 1$ in probability as $d \to \infty$. Finally, for $d = 50$ both the birth *and* the death times increase with σ (Figure 3c, Corollary B.7) and the ground-truth hole disappears in the cloud of spurious holes. Applying MDS to the Euclidean distances obtained with $d = 50$ and $\sigma = 0.25$ yields a 2D layout with almost no visible hole, because all distances have become similar (Figure 3d). See the concurrent work of Hiraoka et al. [42] for a more detailed treatment.

Therefore, the failure modes of persistent homology differ between low- and high-dimensional spaces. While in low dimensions, persistent homology is susceptible to outlier points in the middle of the circle, in high dimensions, there are no points in the middle of the circle; instead, all distances become too similar, hiding the true loops. See Appendix N for more details on the effect of outliers.

5 Spectral distances are more robust

Many modern manifold learning and dimensionality reduction methods rely on the k-nearest-neighbor (kNN) graph of the data. This works well because, although distances become increasingly similar in high-dimensional spaces, nearest neighbors still carry information about the data manifold. To make persistent homology overcome high-dimensional noise, we therefore suggest to rely on the symmetric kNN graph, which contains edge ij if node i is among the k nearest neighbors of j or vice versa. A natural choice is to use its geodesics, but, as we show below, this does not work well, likely because a single graph edge across a circle can destroy the corresponding feature too early. Instead, we propose to use spectral methods, such as the effective resistance or diffusion distance. Both methods rely on random walks and thus incorporate information from all edges.

For a connected graph G with n nodes, e.g., the symmetric kNN graph, let A be its symmetric, $n \times n$ adjacency matrix with elements $a_{ij} = 1$ if edge ij exits in G and $a_{ij} = 0$ otherwise. The degree matrix D is defined by $D = \operatorname{diag}\{d_i\}$, where $d_i = \sum_{j=1}^{n} a_{ij}$ are the node degrees. We define $\operatorname{vol}(G) = \sum_{i=1}^{n} d_i$. Let H_{ij} be the *hitting time* from node i to j, i.e., the average number of edges it takes a random walker, that starts at node i randomly moving along edges, to reach node j. The naive effective resistance is defined as $\tilde{d}_{ij}^{\text{eff}} = (H_{ij} + H_{ji})/\operatorname{vol}(G)$. This version is known to be unsuitable for large graphs (Figure S20) because it reduces to $\tilde{d}_{ij}^{\text{eff}} \approx 1/d_i + 1/d_j$ [81]. Therefore, we used von Luxburg et al. [81]'s corrected version

$$d_{ij}^{\text{eff}} = \tilde{d}_{ij}^{\text{eff}} - 1/d_i - 1/d_j + 2a_{ij}/(d_i d_j) - a_{ii}/d_i^2 - a_{jj}/d_j^2. \tag{1}$$

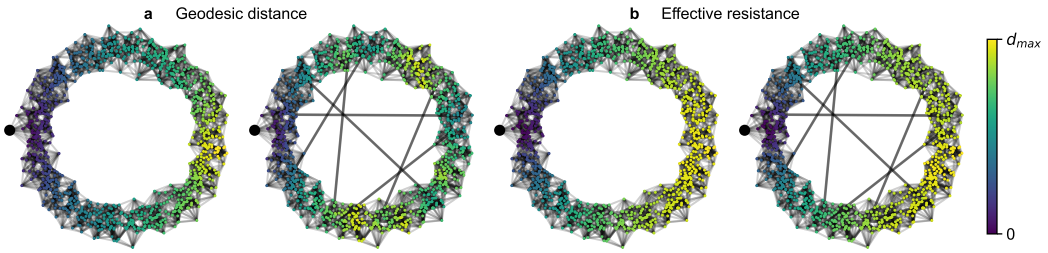

Figure 4: Robustness of effective resistance. We sampled $n = 1\,000$ points from a noisy circle in 2D with Gaussian noise of standard deviation $\sigma = 0.1$, constructed the unweighted symmetric 15-NN graph, and optionally added 10 random edges (thick lines). Node colors indicate the graph distance from the fat black dot. **a.** The geodesic distance is severely affected by the random edges. **b.** The effective resistance distance is robust to them.

Diffusion distances also rely on random walks. The random walk transition matrix is given by $P = D^{-1}A$. Then $P^t_{i,:}$, the i-th row of P^t, holds the probability distribution over nodes after t steps of a random walker starting at node i. The diffusion distance is then defined as

$$d_{ij}(t) = \sqrt{\text{vol}(G)}\|(P^t_{i,:} - P^t_{j,:})D^{-\frac{1}{2}}\|. \tag{2}$$

There are many possible random walks between nodes i and j if they both reside in the same densely connected region of the graph, while it is unlikely for a random walker to cross between sparsely connected regions. As a result, both effective resistance and diffusion distance are small between parts of the graph that are densely connected and are robust against single stray edges (Figure 4). This makes spectral distances on the kNN graph the ideal input to persistent homology for detecting the topology of data in high-dimensional spaces. Indeed, the MDS embedding of the effective resistance and of the diffusion distance of the circle in ambient $\mathbb{R}^{50}$ both clearly show the circular structure (Figure 3e,f).

6 Relation between spectral distances

We show in Section 7 that spectral methods excel as input distances to persistent homology for high-dimensional data. But first, we explain the relationships between them. Laplacian Eigenmaps distance and diffusion distance can be written as Euclidean distances in data representations given by appropriately scaled eigenvectors of the graph Laplacian. In this section, we derive a similar closed-form formula for effective resistance and show that effective resistance aggregates all but the most local diffusion distances.

Let $A^{\text{sym}} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ and $L^{\text{sym}} = I - A^{\text{sym}}$ be the symmetrically normalized adjacency and Laplacian matrix. We denote the eigenvectors of L^{sym} by $u_1, \ldots, u_n$ and their eigenvalues by $\mu_1, \ldots, \mu_n$ in increasing order. For a connected graph, $\mu_1 = 0$ and $u_1 = D^{\frac{1}{2}}(1, \ldots, 1)^\top / \sqrt{\text{vol}(G)}$.

The $\tilde{d}$-dimensional Laplacian Eigenmaps embedding is given by the first $\tilde{d}$ nontrivial eigenvectors:

$$d^{\text{LE}}_{ij}(\tilde{d}) = \|e^{\text{LE}}_i(\tilde{d}) - e^{\text{LE}}_j(\tilde{d})\|, \text{ where } e^{\text{LE}}_i(\tilde{d}) = (u_{2,i}, \ldots, u_{(\tilde{d}+1),i}). \tag{3}$$

The diffusion distance after t diffusion steps is given by [20]

$$d^{\text{diff}}_{ij}(t) = \sqrt{\text{vol}(G)}\|e^{\text{diff}}_i(t) - e^{\text{diff}}_j(t)\|, \text{ where } e^{\text{diff}}_i(t) = \frac{\big((1-\mu_2)^t u_{2,i}, \ldots, (1-\mu_n)^t u_{n,i}\big)}{\sqrt{d_i}}. \tag{4}$$

The original uncorrected version of effective resistance is given by [51]

$$\tilde{d}^{\text{eff}}_{ij} = \|\tilde{e}^{\text{eff}}_i - \tilde{e}^{\text{eff}}_j\|^2, \text{ where } \tilde{e}^{\text{eff}}_i = \big(u_{2,i}/\sqrt{\mu_2}, \ldots, u_{n,i}/\sqrt{\mu_n}\big)/\sqrt{d_i}. \tag{5}$$

In Appendix F we prove that the corrected effective resistance [81] can also be written in this form:

Proposition 6.1. *The corrected effective resistance distance can be computed by*

$$d^{\text{eff}}_{ij} = \|e^{\text{eff}}_i - e^{\text{eff}}_j\|^2, \text{ where } e^{\text{eff}}_i = \left(\frac{1-\mu_2}{\sqrt{\mu_2}}u_{2,i}, \ldots, \frac{1-\mu_n}{\sqrt{\mu_n}}u_{n,i}\right) \Big/ \sqrt{d_i}. \tag{6}$$

https://doi.org/10.52202/079017-1328

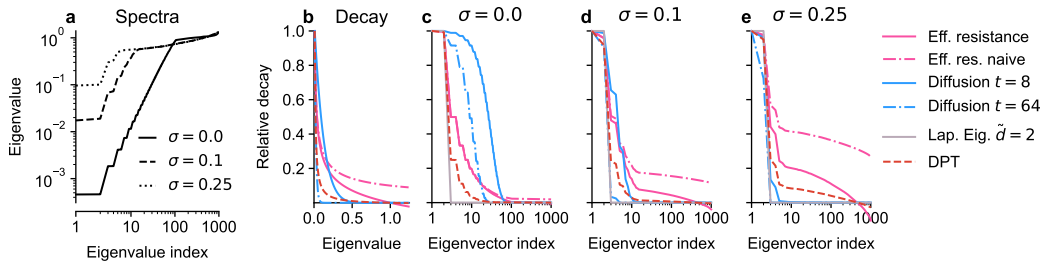

Figure 5: **a.** Eigenvalue spectra of the kNN graph Laplacian for the noisy circle in ambient $\mathbb{R}^{50}$ for noise levels $\sigma = \{0.0, 0.1, 0.25\}$. **b.** Decay of eigenvector contribution based on the eigenvalue for effective resistance, diffusion distances and DPT. **c – e.** Relative contribution of each eigenvector for eff. resistance, diffusion distance, Laplacian Eigenmaps, and DPT for various noise levels (Section 6).

It has been known [54] that the uncorrected effective resistance can be written in terms of diffusion distances as $\tilde{d}_{ij}^{\text{eff}} = \sum_{t=0}^{\infty} d_{ij}^{\text{diff}}(t/2)^2/\text{vol}(G)$, see Proposition G.1. Here, based on Proposition 6.1, we derive a similar result for the corrected effective resistance (proof in Appendix G):

Corollary 6.2. *If G is connected and not bipartite, we have*

$$d_{ij}^{\text{eff}} = \sum_{t=2}^{\infty} d_{ij}^{\text{diff}}(t/2)^2/\text{vol}(G) \quad \textit{and hence} \quad \tilde{d}_{ij}^{\text{eff}} - d_{ij}^{\text{eff}} = \left(d_{ij}^{\text{diff}}(0)^2 + d_{ij}^{\text{diff}}(1/2)^2\right)/\text{vol}(G). \quad (7)$$

In words, the corrected effective resistance combines all diffusion distances, save for those with the shortest diffusion time. These most local diffusion distances form exactly the correction from naive to corrected effective resistance. While the effective resistance is a *squared* Euclidean distance, omitting the square amounts to taking the square root of all birth and death times, maintaining the loop detection performance of effective resistance (Figure S20). Therefore, the main difference between the spectral methods is in to how they decay eigenvectors based on the corresponding eigenvalues.

The naive effective resistance decays the eigenvectors with $1/\sqrt{\mu_i}$, which is much slower than diffusion distances' $(1-\mu_i)^t$ for $t \in [8, 64]$. Corrected effective resistance shows intermediate behavior (Figure 5b). When represented as a sum over diffusion distances, it contains all diffusion distances with $t \geq 1$, making it decay slower than diffusion distances with $t = 8$ or 64, but does not contain the non-decaying $t = 0$ term, so it decays faster than its naive version. The correction matters little for $S^1 \subset \mathbb{R}^{50}$ in the absence of noise, when the first eigenvalues are much smaller than the rest and dominate the embedding (Figure 5a,c) but becomes important as the noise and consequently the low eigenvalues increase (Figure 5a,d,e). As the noise increases, the decay for diffusion distances gets closer to a step function preserving only the first two non-constant eigenvectors, sufficient for the circular structure. In contrast, Laplacian Eigenmaps needs the number of components as input (Figure 5c – e).[1]

7 Spectral distances find holes in high-dimensional spaces

High-dimensional data is ubiquitous, but traditional persistent homology can fail to detect its topology. Here, we benchmark the performance of various distances as input to persistent homology.

Distance measures We examined twelve distances as input to persistent homology, beyond the Euclidean distance. Full definitions are given in Appendix I. First, there are some state-of-the-art approaches for persistent homology in the presence of noise and outliers. Fermat distances [26] aim to exaggerate large over small distances to incorporate the density of the data. Distance-to-measure (DTM) [2] aims for outlier robustness by combining the Euclidean distance with the distances from each point to its k nearest neighbors, which are high for outliers. Similarly, the core distance used in the HDBSCAN algorithm [10, 21] raises each Euclidean distance at least to the distance between

[1] Diffusion pseudotime (DPT) [36] has a very similar expression as corrected effective resistance, using the scaling $(1-\mu_i)/\mu_i$, see Appendix H. This means that DPT decays eigenvalues faster than both versions of effective resistance (Figure 5). We prove an analogous statement to Corollary 6.2 for DPT in Proposition H.1.

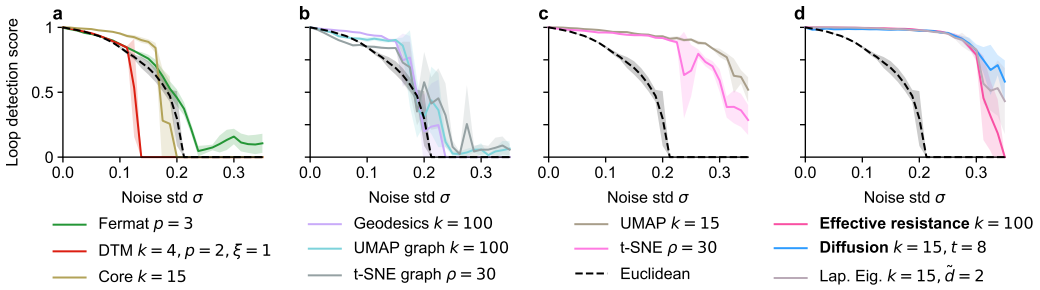

Figure 6: Loop detection score for persistent homology with various distances on a noisy circle in $\mathbb{R}^{50}$. The best hyperparameter setting for each distance is shown. Methods are grouped into panels for visual clarity. Recommended methods in **bold**.

incident points and their k-th nearest neighbors. We evaluate these methods here with respect to Gaussian noise in high-dimensional ambient space, a different noise model than the one for which these methods were designed. Second, we consider some non-spectral graph distances. The geodesic distance on the kNN graph was popularized by Isomap [76] and used for persistent homology by Naitzat et al. [60]. Following Gardner et al. [31] we used distances based on UMAP affinities, and also experimented with t-SNE affinities. Third, we computed t-SNE and UMAP embeddings and used distances in the 2D embedding space. Finally, we explored methods using the spectral decomposition of the kNN graph Laplacian, see Section 6: effective resistance, diffusion distance, and the distance in Laplacian Eigenmaps' embedding space.

All methods come with hyperparameters. We report the results for the best hyperparameter setting on each dataset (Appendix K) but found spectral methods to be robust to these choices (Appendix L).

Performance score The output of persistent homology is a persistence diagram showing birth and death times for all detected holes. It may be difficult to decide whether this procedure has actually detected a hole in the data, or not. Ideally, for a dataset with m ground-truth holes, the persistence diagram should have m points with high persistence while all other points should have low persistence and lie close to the diagonal. Therefore, for m ground-truth features, our *hole detection score* $s_m \in [0, 1]$ is the relative gap between the persistences p_m and p_{m+1} of the m-th and $(m + 1)$-th most persistent features: $s_m = (p_m - p_{m+1})/p_m$. This corresponds to the visual gap between them in the persistence diagram (Figure 2b). Rieck and Leitte [66] as well as Smith and Kurlin [74] used similar quantities to find important features. We prove a continuity property of s_m in Appendix D and consider alternative scores in Appendix E.

In addition, we set $s_m = 0$ if all features in the persistence diagram have very low death-to-birth ratios $\tau_d/\tau_b < 1.25$. This handles situations with very few detected holes that die very quickly after being born, which otherwise can have spuriously high s_m values. This was done everywhere apart from the qualitative Figures 1, 9 and in Figure S25. We call this heuristic *thresholding*.

Note that the number of ground-truth topological features was used only for evaluation. We report the mean over three random seeds; shading and error bars indicate the standard deviation.

7.1 Synthetic benchmark

Benchmark setup In our synthetic benchmark, we evaluated the performance of various distance measures in conjunction with persistent homology on five manifolds: a circle, a pair of linked circles, the eyeglasses dataset (a circle squeezed nearly to a figure eight) [26], the sphere, and the torus. The radii of the circles, the sphere, and the torus' tube were set to 1, the bottleneck of the eyeglasses was 0.7, and the torus' tube followed a circle of radius 2. In each case, we uniformly sampled $n = 1\,000$ points from the manifold, mapped them isometrically to $\mathbb{R}^d$ for $d \in [2, 50]$, and then added isotropic Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$ for $\sigma \in [0, 0.35]$. More details can be found in Appendix J. For each resulting dataset, we computed persistent homology for loops and, for the sphere and the torus, also for voids. We never computed holes of dimension 3 or higher.

https://doi.org/10.52202/079017-1328

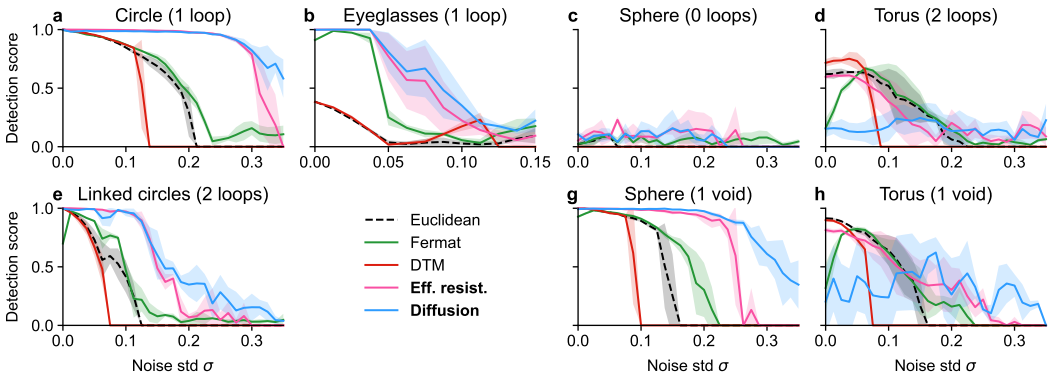

Figure 7: Loop detection score for selected methods on synthetic datasets in ambient $\mathbb{R}^{50}$. More experimental results can be found in Figures S23 – S33. Recommended methods in **bold**.

Results on synthetic data On the circle dataset in $\mathbb{R}^{50}$, persistent homology with all distance metrics found the correct hole when the noise level σ was very low (Figure 6). However, as the amount of noise increased, the performance of Euclidean distance quickly deteriorated, reaching zero score at $\sigma \approx 0.2$. Most other distances outperformed the Euclidean distance, at least in the low noise regime. Fermat distance did not have any effect, and neither did DTM distance, which collapsed at $\sigma \approx 0.15$ due to our thresholding (Figure 6a). Geodesics, UMAP/t-SNE graph, and core distance offered only a modest improvement over Euclidean (Figure 6b) highlighting that many kNN-graph-based distances cannot handle high-dimensional noise. In contrast, embedding-based distances performed very well on the circle (Figure 6c), but have obvious limitations: for example, a 2D embedding cannot possibly have a void. UMAP with higher embedding dimension struggled with loop detection on surfaces and the torus' void (Appendix O). Finally, all spectral methods (effective resistance, diffusion, and Laplacian Eigenmaps) showed similarly excellent performance (Figure 6d).

In line with these results, spectral methods outperformed other methods across most synthetic datasets in $\mathbb{R}^{50}$ (Figure 7). DTM collapsed earlier than Euclidean but detected loops on the torus for low noise levels best by a small margin. Fermat distance typically had little effect and provided a benefit over Euclidean only on the eyeglasses and the sphere. Spectral distances outperformed all other methods on all datasets apart from the torus, where effective resistance was on par with Euclidean but diffusion performed poorly. On a more densely sampled torus, all methods performed better and the spectral methods again outperformed the others (Figure S31). On all other datasets diffusion distance slightly outperformed effective resistance for large σ. Reassuringly, all methods passed the negative control and did not find any persistent loops on the sphere (Figure 7c).

As discussed in Section 4, persistent homology with Euclidean distances deteriorates with increasing ambient dimensionality. Using the circle data in $\mathbb{R}^d$, we found that if the noise level was fixed at $\sigma = 0.25$, no persistent loop was found using Euclidean distances for $d \gtrsim 30$ (Figure 8). In the same setting, DTM deteriorated even more quickly than Euclidean distances. In contrast, effective resistance and diffusion distance were robust against both the high noise level and the large ambient dimension (Figure 8a,c – e). See Figure S1 for an extended analysis.

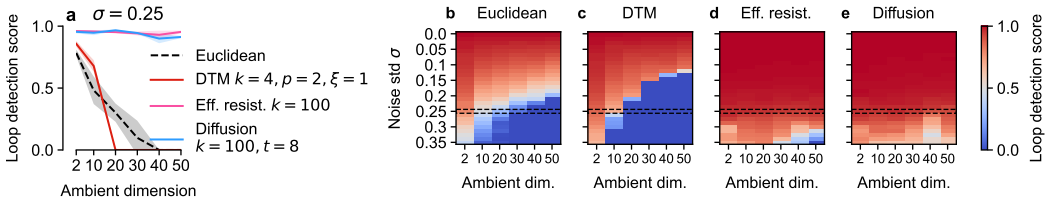

Figure 8: **a.** Loop detection score of various methods on a noisy circle depending on the ambient dimensionality. Noise $\sigma = 0.25$. **b – e.** Heat maps for $\sigma \in [0, 0.35]$ and $d \in [2, 50]$.

7.2 Detecting cycles in single-cell data

We applied our methods to six single-cell RNA-sequencing datasets: Malaria [43], Neurosphere and Hippocampus from [89], HeLa2 [72], Neural IPCs [8], and Pancreas [3]. Single-cell RNA-sequencing data consists of expression levels for thousands of genes in individual cells, so the data is high-dimensional and notoriously noisy. Importantly, all selected datasets are known to contain circular structures, usually corresponding to the cell division cycle during which gene expression levels cyclically change. As a result, we know how many loops to expect in each dataset and can therefore use them as a real-world benchmark of various distances for persistent homology. In each case, we followed preprocessing pipelines from prior publications leading to representations with 10 to 5 156 dimensions. We downsampled datasets with more than $4\,000$ cells to $n = 1\,000$ (Appendix J).

The Malaria dataset is expected to contain two cycles: the parasite replication cycle in red blood cells, and the parasite transmission cycle between human and mosquito hosts. Following Howick et al. [43], we based all computations for this dataset (and all derived distances) on the correlation distance instead of the Euclidean distance. Persistent homology based on the correlation distance itself failed to correctly identify the two ground-truth cycles and DTM produced representatives that only roughly approximate the two ground truth cycles (Figure 9a,b). Both effective resistance and diffusion distance successfully uncovered both cycles with $s_2 > 0.9$ (Figure 9c,d).

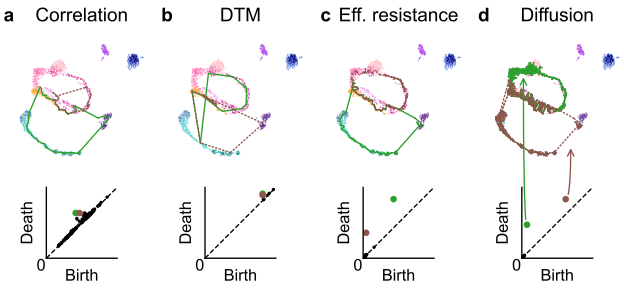

Figure 9: Malaria dataset. **a – d.** Representatives of the two most persistent loops overlaid on UMAP embedding (top) and persistence diagrams (bottom) using four methods. Biology dictates that there should be two loops (in warm colors and in cold colors) connected as in a figure eight.

Across all six datasets, the detection scores were higher for spectral methods than for their competitors (Figure 10). Furthermore, we manually investigated representative loops for all considered methods on all datasets and found several cases where the most persistent loop(s) was/were likely not correct (hatched bars in Figure 10). Overall, we found that the spectral methods, and in particular effective resistance, could reliably find the correct loops with high detection score. Persistent homology based on the t-SNE and UMAP embeddings worked on average better than traditional persistent homology, Fermat distances, and DTM, but worse than the spectral methods.

8 Limitations and future work

In the real-world applications, it was important to look at representatives of detected holes as some holes were persistent, but arguably incorrect. That said, each homology class has many different

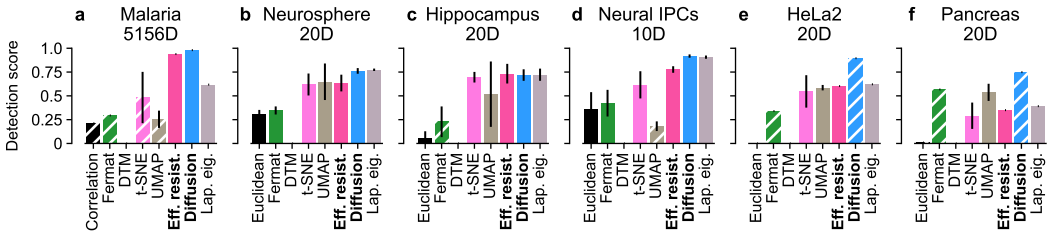

Figure 10: Loop detection scores on six high-dimensional scRNA-seq datasets. Hatched bars indicate implausible representatives. See Figure S34 for detection scores for different hyperparameter values. Recommended methods in **bold**.

https://doi.org/10.52202/079017-1328

representative cycles, making interpretation difficult. Given ground-truth cycles, an automatic procedure for evaluating cycle correctness remains an interesting research question.

Persistent homology can only detect topology, which is often a useful global level of abstraction. However, it may therefore fail to distinguish some non-isomorphic point clouds [75]. There exist dedicated measures for detecting isometry [7, 87, 47].

Dimensionality reduction methods are designed to handle high-dimensional data. t-SNE and UMAP indeed performed well on many datasets, but on average worse than spectral distances on the real data. Moreover, UMAP struggled with the surface of the 3D toy datasets and the torus' void (Appendix O). Finally, they require the choice of an embedding dimension and are known to produce artifacts [49, 12, 85], e.g., leading to poor scores in the noiseless setting in Figures S17, S22. In contrast, spectral distances on the symmetric kNN graph worked well without a low-dimensional embedding (Section 7).

Using effective resistance or diffusion distances is easy in practice as their computation time $O(n^3)$ is dwarfed by that of the persistent homology (Table S5), which scales as $\mathcal{O}(n^{3(\delta+1)})$ for n points and topological holes of dimension δ [59]. This high complexity of persistent homology aggravates other problems of high-dimensional datasets as dense sampling in high-dimensional space would require a prohibitively large sample size (recall that spectral methods needed a high sampling density for good performance on some of our datasets such as the torus). Combining persistent homology with non-Euclidean distance measures could mitigate this problem via the approach of Bendich et al. [6], who performed subsampling after computation of the distance matrix. This is a particularly attractive avenue for future research.

Both effective resistance and diffusion distances require the choice of hyperparameters. However, effective resistance only needs a single hyperparameter: the number of kNN neighbors. For this reason and due to its greater outlier resistance (Appendix N), we tend to recommend effective resistance over diffusion distances, but a principled criterion when to use which of the two is still missing.

Moreover, we do not have a theoretical proof that spectral distances mitigate the curse of dimensionality. Such a proof may be achieved in the future taking inspiration from the stability results in [11, 42, 77] and, more generally, spectral perturbation theory.

Our empirical results focus on benchmarking which distances identify the correct topology in the presence of high-dimensional noise. Therefore, we only considered datasets with known ground-truth topology. The next step will be to use spectral distances to detect non-trivial topology in real-world exploratory contexts.

High-dimensional data and thus application areas for our improved topology detection pipeline are becoming ubiquitous. Within biology, we see possible applications for our method in other single-cell omics modalities, population genomics, or neural activity data [31, 40]. Beyond biology, we believe that our approach can improve the topological analysis of artificial neural network activations [60], and in general be used to detect topology of any high-dimensional data, e.g. in the climate sciences, in astronomical measurements, or wearable sensor data.

9 Conclusion

In this work we asked how to use persistent homology on high-dimensional noisy datasets which are very common in real-world applications even if the intrinsic data dimensionality is low. We found spectral methods to be the optimal approach. We demonstrated that, as the dimensionality of the data increases, the main problem for persistent homology shifts from handling outliers to handling noise dimensions (Section 4). We used a synthetic benchmark to show that traditional persistent homology and many of its existing extensions struggle to find the correct topology in this setting. Our main finding is that spectral methods based on the kNN graph, such as the effective resistance and diffusion distances, still work well (Section 7.1). Furthermore, we view it as an advantage that we found existing methods that are able to handle the important problem of high-dimensional noise. We derived an expression for effective resistance based on the eigendecomposition of the graph Laplacian, and demonstrated that it combines all but the most local diffusion distances (Section 6). Finally, we showed that spectral distances outperform all competitors on single-cell data (Section 7.2).

Acknowledgments and Disclosure of Funding

We thank Enrique Fita Sanmartin and Ulrike von Luxburg for productive discussions on effective resistance and persistent homology, and Benjamin Dunn, Erik Hermansen, and David Klindt for helpful discussions on combining persistent homology with other dissimilarities than Euclidean distance. Moreover, we thank Sten Linnarsson and Miri Danan Gotthold for sharing the scVI representation of their pallium data. Final thanks go to Bastian Rieck for feedback on the writing.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via Germany's Excellence Strategy (Excellence cluster 2064 "Machine Learning — New Perspectives for Science", EXC 390727645; Excellence cluster 2181 "STRUCTURES", EXC 390900948), the German Ministry of Science and Education (BMBF) via the Tübingen AI Center (01IS18039A), the Gemeinnützige Hertie-Stiftung, and the National Institutes of Health (UM1MH130981). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] M. E. Aktas, E. Akbas, and A. E. Fatmaoui. Persistence homology of networks: methods and applications. *Applied Network Science*, 4(1):1–28, 2019.

[2] H. Anai, F. Chazal, M. Glisse, Y. Ike, H. Inakoshi, R. Tinarrage, and Y. Umeda. DTM-based filtrations. In *Topological Data Analysis: The Abel Symposium 2018*, pages 33–66. Springer, 2020.

[3] A. Bastidas-Ponce, S. Tritschler, L. Dony, K. Scheibner, M. Tarquis-Medina, C. Salinno, S. Schirge, I. Burtscher, A. Böttcher, F. J. Theis, et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12): dev173849, 2019.

[4] U. Bauer. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3), 2021.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, pages 585–591, 2002.

[6] P. Bendich, T. Galkovskyi, and J. Harer. Improving homology estimates with random walks. *Inverse Problems*, 27(12):124002, 2011.

[7] M. Boutin and G. Kemper. On reconstructing n-point configurations from the distribution of distances or areas. *Advances in Applied Mathematics*, 32(4):709–735, 2004.

[8] E. Braun, M. Danan-Gotthold, L. E. Borm, K. W. Lee, E. Vinsland, P. Lönnerberg, L. Hu, X. Li, X. He, Ž. Andrusivová, et al. Comprehensive cell atlas of the first-trimester developing human brain. *Science*, 382(6667):eadf1226, 2023.

[9] N. Brugnone, A. Gonopolskiy, M. W. Moyle, M. Kuchroo, D. van Dijk, K. R. Moon, D. Colon-Ramos, G. Wolf, M. J. Hirn, and S. Krishnaswamy. Coarse graining of data via inhomogeneous diffusion condensation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2624–2633. IEEE, 2019.

[10] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–51, 2015.

[11] M. Carrière, F. Chazal, Y. Ike, T. Lacombe, M. Royer, and Y. Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR, 2020.

[12] T. Chari and L. Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, 2023.

 https://doi.org/10.52202/079017-1328

[13] B. Charlier, J. Feydy, J. A. Glaunès, F.-D. Collin, and G. Durif. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

[14] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:108, 2021.

[15] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for measures based on distance functions. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

[16] J. Chew, H. Steach, S. Viswanath, H.-T. Wu, M. Hirn, D. Needell, M. D. Vesely, S. Krishnaswamy, and M. Perlmutter. The manifold scattering transform for high-dimensional point cloud data. In *Topological, Algebraic and Geometric Learning Workshops 2022*, pages 67–78. PMLR, 2022.

[17] F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[18] L. Clarté, A. Vandenbroucque, G. Dalle, B. Loureiro, F. Krzakala, and L. Zdeborová. Analysis of bootstrap and subsampling in high-dimensional regularized regression. *arXiv preprint arXiv:2402.13622*, 2024.

[19] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 263–271, 2005.

[20] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[21] D. Damm. Core-distance-weighted persistent homology and its behavior on tree-shaped data. Master's thesis, Heidelberg University, 2022.

[22] S. Damrich and F. A. Hamprecht. On UMAP's true loss function. In *Advances in Neural Information Processing Systems*, volume 34, pages 5798–5809, 2021.

[23] T. Davies, Z. Wan, and R. J. Sanchez-Garcia. The persistent Laplacian for data science: Evaluating higher-order persistent spectral representations of data. In *International Conference on Machine Learning*, pages 7249–7263. PMLR, 2023.

[24] P. G. Doyle and J. L. Snell. *Random Walks and Electric Networks*, volume 22. American Mathematical Soc., 1984.

[25] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

[26] X. Fernández, E. Borghini, G. Mindlin, and P. Groisman. Intrinsic persistent homology via density-based metric learning. *Journal of Machine Learning Research*, 24(75):1–42, 2023.

[27] E. Flores-Bautista and M. Thomson. Unraveling cell differentiation mechanisms through topological exploration of single-cell developmental trajectories. *bioRxiv*, pages 2023–07, 2023.

[28] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on knowledge and data engineering*, 19(3):355–369, 2007.

[29] F. Fouss, M. Saerens, and M. Shimbo. *Algorithms and Models for Network Data and Link Analysis*. Cambridge University Press, 2016.

[30] I. García-Redondo, A. Monod, and A. Song. Fast topological signal identification and persistent cohomological cycle matching. *Journal of Applied and Computational Topology*, pages 1–32, 2024.

[31] R. J. Gardner, E. Hermansen, M. Pachitariu, Y. Burak, N. A. Baas, B. A. Dunn, M.-B. Moser, and E. I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895): 123–128, 2022.

[32] F. Göbel and A. Jagers. Random walks on graphs. *Stochastic processes and their applications*, 2(4):311–336, 1974.

[33] J. C. Gower and G. J. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 18(1):54–64, 1969.

[34] P. Groisman, M. Jonckheere, and F. Sapienza. Nonhomogeneous Euclidean first-passage percolation and distance learning. *Bernoulli*, 28(1):255–276, 2022.

[35] L. Haghverdi, F. Buettner, and F. J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.

[36] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016.

[37] M. Hajij, B. Wang, C. Scheidegger, and P. Rosen. Visual detection of structural changes in time-varying graphs using persistent homology. In *2018 IEEE Pacific Visualization Symposium (pacificvis)*, pages 125–134. IEEE, 2018.

[38] P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3): 427–444, 2005.

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

[40] E. Hermansen, D. A. Klindt, and B. A. Dunn. Uncovering 2-d toroidal representations in grid cell ensemble activity during 1-d behavior. *Nature Communications*, 15(1):5429, 2024.

[41] G. E. Hinton and S. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 857–864, 2002.

[42] Y. Hiraoka, Y. Imoto, S. Kanazawa, and E. Liu. Curse of dimensionality on persistence diagrams. *arXiv preprint arXiv:2404.18194*, 2024.

[43] V. M. Howick, A. J. Russell, T. Andrews, H. Heaton, A. J. Reid, K. Natarajan, H. Butungi, T. Metcalf, L. H. Verzier, J. C. Rayner, et al. The malaria cell atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science*, 365(6455):eaaw2619, 2019.

[44] X. Hu, F. Li, D. Samaras, and C. Chen. Topology-preserving deep image segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[45] J. Jia and L. Chen. Single-cell RNA sequencing data analysis based on non-uniform ε-neighborhood network. *Bioinformatics*, 38(9):2459–2465, 2022.

[46] M. Kuchroo, M. DiStasio, E. Song, E. Calapkulu, L. Zhang, M. Ige, A. H. Sheth, A. Majdoubi, M. Menon, A. Tong, et al. Single-cell analysis reveals inflammatory interactions driving macular degeneration. *Nature Communications*, 14(1):2589, 2023.

[47] V. Kurlin. Polynomial-time algorithms for continuous metrics on atomic clouds of unordered points. *MATCH Commun. Math. Comput. Chem.*, 91:79–108, 2024.

[48] J. Lamar-León, E. B. Garcia-Reyes, and R. Gonzalez-Diaz. Human gait identification using persistent homology. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17*, pages 244–251. Springer, 2012.

[49] P. S. Landweber, E. A. Lazar, and N. Patel. On fiber diameters of continuous maps. *The American Mathematical Monthly*, 123(4):392–397, 2016.

[50] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[51] L. Lovász. Random walks on graphs. *Combinatorics, Paul Erdős is Eighty*, 2(1-46):4, 1993.

 https://doi.org/10.52202/079017-1328

[52] H. Masoomy, B. Askari, S. Tajik, A. K. Rizi, and G. R. Jafari. Topological analysis of interaction patterns in cancer-specific gene regulatory network: Persistent homology approach. *Scientific Reports*, 11(1):16414, 2021.

[53] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[54] F. Mémoli. personal communication, 2023.

[55] F. Mémoli, Z. Wan, and Y. Wang. Persistent Laplacians: properties, algorithms and implications. *SIAM Journal on Mathematics of Data Science*, 4(2):858–884, 2022.

[56] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.

[57] J. L. Moore, D. Bhaskar, F. Gao, C. Matte-Martone, S. Du, E. Lathrop, S. Ganesan, L. Shao, R. Norris, N. Campamà Sanz, et al. Cell cycle controls long-range calcium signaling in the regenerating epidermis. *Journal of Cell Biology*, 222(7):e202302095, 2023.

[58] S. Mukherjee, D. Wethington, T. K. Dey, and J. Das. Determining clinically relevant features in cytometry data using persistent homology. *PLoS Computational Biology*, 18(3):e1009931, 2022.

[59] A. D. Myers, M. M. Chumley, F. A. Khasawneh, and E. Munch. Persistent homology of coarse-grained state-space networks. *Physical Review E*, 107(3):034303, 2023.

[60] G. Naitzat, A. Zhitnikov, and L.-H. Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(1):7503–7542, 2020.

[61] A. Petenkaya, F. Manuchehrfar, C. Chronis, and J. Liang. Identifying transient cells during reprogramming via persistent homology. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2920–2923. IEEE, 2022.

[62] G. Petri, M. Scolamiero, I. Donato, and F. Vaccarino. Networks and cycles: a persistent homology approach to complex networks. In *Proceedings of the European Conference on Complex Systems 2012*, pages 93–99. Springer, 2013.

[63] P. G. Poličar, M. Stražar, and B. Zupan. openTSNE: A modular python library for t-sne dimensionality reduction and embedding. *Journal of Statistical Software*, 109:1–30, 2024.

[64] Y. Reani and O. Bobrowski. Cycle registration in persistent homology with applications in topological bootstrap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5579–5593, 2022.

[65] A. J. Reid, A. M. Talman, H. M. Bennett, A. R. Gomes, M. J. Sanders, C. J. Illingworth, O. Billker, M. Berriman, and M. K. Lawniczak. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *eLife*, 7:e33105, 2018.

[66] B. Rieck and H. Leitte. Agreement analysis of quality measures for dimensionality reduction. In *Topological Methods in Data Analysis and Visualization IV: Theory, Algorithms, and Applications VI*, pages 103–117. Springer, 2017.

[67] B. Rieck, M. Togninalli, C. Bock, M. Moor, M. Horn, T. Gumbsch, and K. Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *International Conference on Learning Representations*, 2018.

[68] A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and R. Rabadan. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35(6):551–560, 2017.

[69] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1–9, 2010.

[70] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[71] B. Roycraft, J. Krebs, and W. Polonik. Bootstrapping persistent Betti numbers and other stabilizing statistics. *The Annals of Statistics*, 51(4):1484–1509, 2023.

[72] D. Schwabe, S. Formichetti, J. P. Junker, M. Falcke, and N. Rajewsky. The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology*, 16(11):e9946, 2020.

[73] G. Singh, F. Mémoli, G. E. Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3D object recognition. *PBG@ Eurographics*, 2:091–100, 2007.

[74] P. Smith and V. Kurlin. Skeletonisation algorithms with theoretical guarantees for unorganised point clouds with high levels of noise. *Pattern Recognition*, 115:107902, 2021.

[75] P. Smith and V. Kurlin. Generic families of finite metric spaces with identical or trivial 1-dimensional persistence. *Journal of Applied and Computational Topology*, pages 1–17, 2024.

[76] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[77] Q. H. Tran, Y. Hasegawa, et al. Scale-variant topological information for characterizing the structure of complex networks. *Physical Review E*, 100(3):032308, 2019.

[78] R. Turkes, G. F. Montufar, and N. Otter. On the effectiveness of persistent homology. *Advances in Neural Information Processing Systems*, 35:35432–35448, 2022.

[79] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[80] S. Vishwanath, K. Fukumizu, S. Kuriki, and B. K. Sriperumbudur. Robust persistence diagrams using reproducing kernels. *Advances in Neural Information Processing Systems*, 33:21900–21911, 2020.

[81] U. von Luxburg, A. Radl, and M. Hein. Getting lost in space: Large sample analysis of the resistance distance. *Advances in Neural Information Processing Systems*, 23, 2010.

[82] U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large graphs are often misleading. *arXiv preprint arXiv:1003.1266*, 2010. version 1 from Mar 05 2010.

[83] U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15(1):1751–1798, 2014.

[84] M. Wang, Z. Cang, and G.-W. Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2): 116–123, 2020.

[85] S. Wang, E. D. Sontag, and D. A. Lauffenburger. What cannot be seen correctly in 2D visualizations of single-cell 'omics data? *Cell Systems*, 14(9):723–731, 2023.

[86] L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[87] D. Widdowson and V. Kurlin. Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1275–1284, 2023.

[88] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20:1–9, 2019.

[89] S. C. Zheng, G. Stein-O'Brien, J. J. Augustin, J. Slosberg, G. A. Carosso, B. Winer, G. Shin, H. T. Bjornsson, L. A. Goff, and K. D. Hansen. Universal prediction of cell-cycle position using transfer learning. *Genome Biology*, 23(1):1–27, 2022.

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

 https://doi.org/10.52202/079017-1328

A Broader impact

The goal of our paper is to advance the field of machine learning. Our comprehensive benchmark required a lot of compute. However, we expect that the lessons learned will save compute for researchers applying persistent homology to their high-dimensional data. Beyond this, we do not see any potential societal consequences of our work that would be worth specifically highlighting here.

B Noise in high-dimensional spaces eventually dominates structure

In this section, we rigorously prove the intuitive result that the Euclidean distances — and thus the traditional persistence diagrams — eventually get dominated by noise as the number of ambient dimensions grows (assuming homoscedastic noise). The main results are Corollary B.5 and Corollary B.7. Our analysis is similar to the concurrent work of Hiraoka et al. [42]. Their analysis extends ours in that they consider an arbitrary noise distribution, the Čech complex, and make more precise statements on how persistent homology gets impaired by high-dimensional noise.

We begin with some helpful lemmata.

Lemma B.1. *Let $\{\varepsilon_{1,d}\}$ and $\{\varepsilon_{2,d}\}$ be two sequences of d-dimensional multivariate normally distributed random variables $\varepsilon_{i,d} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. For each fixed d, we have $\mathbb{E}(\|\varepsilon_{1,d} - \varepsilon_{2,d}\|^2) = 2\sigma^2 d$. Moreover, the sequence $\{\|\varepsilon_{1,d} - \varepsilon_{2,d}\|/\sqrt{d}\}$ converges to $\sqrt{2}\sigma$ almost surely.*

Proof. All entries of $\varepsilon_{1,d} - \varepsilon_{2,d}$ are i.i.d. Gaussian random variables with zero mean and variance $2\sigma^2$. The squared Euclidean distance is the sum of the squared entries of the vector, which implies the first statement. By the strong law of large numbers, $\|\varepsilon_{1,d} - \varepsilon_{2,d}\|^2/d$ converges almost surely to $2\sigma^2$. By the continuous mapping theorem, this implies almost sure convergence of $\|\varepsilon_{1,d} - \varepsilon_{2,d}\|/\sqrt{d}$ to $\sqrt{2}\sigma$. □

Lemma B.2. *Let $\{X_d\}$ be a sequence of random variables with finite means and variances. If $\mathbb{E}(X_d)$ converges to a constant c and $\mathrm{Var}(X_d) \to 0$ as $d \to \infty$, then $\{X_d\}$ converges to c in squared mean, i.e., $\mathbb{E}\big((X_d - c)^2\big) \to 0$.*

Proof. We have

$$\begin{aligned}\mathbb{E}\big((X_d - c)^2\big) &= \mathbb{E}(X_d^2) - 2c\mathbb{E}(X_d) + c^2 \\ &= \mathbb{E}(X_d^2) - \mathbb{E}(X_d)^2 + \mathbb{E}(X_d)^2 - 2c\mathbb{E}(X_d) + c^2 \\ &= \mathrm{Var}(X_d) + \mathbb{E}(X_d)^2 - 2c\mathbb{E}(X_d) + c^2 \\ &\to 0 + c^2 - 2c^2 + c^2 = 0. \end{aligned} \tag{8}$$

□

Lemma B.3. *Let $\{X_d\}$ be a sequence of random variables which converges in squared mean to a random variable X. Then it also converges to X in probability.*

Proof. This is an application of Markov's inequality. Let $\varepsilon > 0$. Then

$$\mathbb{P}\big(|X_d - X| > \varepsilon\big) = \mathbb{P}\big((X_d - X)^2 > \varepsilon^2\big) \leq \mathbb{E}\big((X_d - X)^2\big)/\varepsilon^2 \to 0$$

for $d \to \infty$ by convergence in squared mean. □

Now we show that for sufficiently many noise dimensions the distance between any two points gets dominated by the noise. This follows from the Pythagorean theorem, because most noise dimensions are orthogonal to the difference of the noise-free points. A similar result can be found in Hall et al. [38].

Proposition B.4. *Let x_1 and x_2 be two points in $\mathbb{R}^{d'}$. Let $\{\iota_d\}$ be a sequence of isometries $\iota_d : \mathbb{R}^{d'} \to \mathbb{R}^d$ for $d \geq d'$. Let $\{\varepsilon_{1,d}\}$ and $\{\varepsilon_{2,d}\}$ be two sequences of d-dimensional multivariate normally distributed random variables $\varepsilon_{i,d} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ starting at $d = d'$. Let $\{\Delta_d\}$ be the sequence of Euclidean distances between $\iota_d(x_1) + \varepsilon_{1,d}$ and $\iota_d(x_2) + \varepsilon_{2,d}$. Then $\{\Delta_d^2/d\}$ converges to $2\sigma^2$ in squared mean.*

Proof. Denote the distance of the embedded points by $\delta = \|x_1 - x_2\| = \|\iota_d(x_1) - \iota_d(x_2)\|$. We have

$$\Delta_d = \|\iota_d(x_1) + \varepsilon_{1,d} - \big(\iota_d(x_2) + \varepsilon_{2,d}\big)\| = \|\big(\iota_d(x_1) + \varepsilon_{1,d} - \varepsilon_{2,d}\big) - \iota_d(x_2)\|.$$

So instead of the setting where both points get noised, we can just consider the case where only the first point gets noised by a d-dimensional multivariate normally distributed variable of double variance $\varepsilon_d \sim \mathcal{N}(\mathbf{0}, 2\sigma^2\mathbf{I}_d)$. We denote $\|\varepsilon_d\|$ by l. We will drop the use of ι_d now and just write $x_1, x_2 \in \mathbb{R}^d$ in slight abuse of notation.

Let α be the angle that ε_d makes with the vector $x_2 - x_1$. Then, by the law of cosines,

$$\Delta_d^2 = l^2 + \delta^2 - 2\delta\cos(\alpha)l. \tag{9}$$

Since ε_d is isometrically distributed, we can assume that $x_2 - x_1$ is parallel to $(1, 0, \ldots, 0)$ without loss of generality. Then $\cos(\alpha)l = \varepsilon'$, where $\varepsilon' \sim \mathcal{N}(0, 2\sigma^2)$ is the first entry of ε_d. Thus, $\mathbb{E}(\Delta_d^2) = 2\sigma^2 d + \delta^2$ by Lemma B.1 and $\mathbb{E}(\Delta_d^2/d) \to 2\sigma^2$ for $d \to \infty$.

For the variance, we obtain

$$\mathrm{Var}(\Delta_d^2) = \mathrm{Var}(l^2 + \delta^2 - 2\delta\varepsilon') = (d-1)\mathrm{Var}(\varepsilon'^2) + \mathrm{Var}(\varepsilon'^2 - 2\delta\varepsilon').$$

The first term contains the variances of all the directions orthogonal to $x_2 - x_1$. Since ε_d is isotropic, all directions are independent and we can simply add up the individual variances. The second term contains the variance in direction $x_2 - x_1$. We have

$$\begin{aligned}
\mathrm{Var}(\varepsilon'^2) &= 3 \cdot 4\sigma^4 - 4\sigma^4 \\
&= 8\sigma^4 \qquad (10)\\
\mathrm{Var}(\varepsilon'^2 - 2\delta\varepsilon') &= \mathbb{E}(\varepsilon'^4 - 4\delta\varepsilon'^3 + 4\delta^2\varepsilon'^2) - \mathbb{E}(\varepsilon'^2 - 2\delta\varepsilon')^2 \\
&= 12\sigma^4 + 0 + 8\delta^2\sigma^2 - 4\sigma^2 + 0 \\
&= 8\sigma^4 + 8\delta^2\sigma^2. \qquad (11)
\end{aligned}$$

Together, we have

$$\begin{aligned}
\mathrm{Var}(\Delta_d^2/d) &= \big((d-1)8\sigma^4 + 8\sigma^4 + 8\delta^2\sigma^2\big)/d^2 \\
&= 8\sigma^2(d\sigma^2 + \delta^2)/d^2 \\
&\to 0 \qquad (12)
\end{aligned}$$

as $d \to \infty$. By Lemma B.2, we conclude that $\{\Delta_d^2/d\}$ converges to $2\sigma^2$ in squared mean. □

The next corollary generalizes Proposition B.4 to an arbitrary arrangement of n points. Independent of their structure, noise will dominate all distances for sufficiently high dimensionality.

Corollary B.5. *Let $x_1, \ldots, x_n$ be n pairwise distinct points in $\mathbb{R}^{d'}$. Let $\{\iota_d\}$ be a sequence of isometries $\iota_d : \mathbb{R}^{d'} \to \mathbb{R}^d$ for $d \geq d'$. Let further $\{\varepsilon_{1,d}\}, \ldots, \{\varepsilon_{n,d}\}$ be n sequences of multivariate normally distributed random variables $\varepsilon_{i,d} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$ in d dimensions. Let $\Delta_{i,j,d}$ be the sequence of random variables of Euclidean distances between $\iota_d(x_i) + \varepsilon_{i,d}$ and $\iota_d(x_j) + \varepsilon_{j,d}$. Then each sequence $\{\Delta_{i,j,d}/\sqrt{d}\}$ converges to $\sqrt{2}\sigma$ in probability as $p \to \infty$. The sequence $\{\|\varepsilon_{i,d} - \varepsilon_{j,d}\|/\Delta_{i,j,d}\}$ converges to 1 in probability and thus the joint vector $(\|\varepsilon_{i,d} - \varepsilon_{j,d}\|/\Delta_{i,j,d})_{i,j}$ with entries for each pair $i \neq j$ converges to the vector of all ones in probability.*

Proof. Proposition B.4, Lemma B.3, and the continuous mapping theorem imply that $\Delta_{i,j,d}/\sqrt{d}$ converges to $\sqrt{2}\sigma$ in probability as $p \to \infty$ and that $\sqrt{d}/\Delta_{i,j,d}$, which is well-defined up to a set of measure zero, converges to $(\sqrt{2}\sigma)^{-1}$ in probability. By Lemma B.1 and since almost sure convergence implies convergence in probability, we have that $\|\varepsilon_{i,d} - \varepsilon_{j,d}\|/\sqrt{d}$ converges to $\sqrt{2}\sigma$ in probability as well. Since convergence in probability is preserved under multiplication, we obtain that $\|\varepsilon_{i,d} - \varepsilon_{j,d}\|/\Delta_{i,j,d}$ converges to 1 in probability. Finally, convergence in probability of two sequences implies joint convergence in probability. □

https://doi.org/10.52202/079017-1328

Corollary B.6. *Consider the setting of Corollary B.5 and choose some $\varepsilon > 0$. Then we have that*

$$\mathbb{P}(\min_{i,j} \Delta_{i,j,d}/\sqrt{d} > \sqrt{2}\sigma - \varepsilon), \tag{13}$$

$$\mathbb{P}(\min_{i,j} \Delta_{i,j,d}/\sqrt{d} < \sqrt{2}\sigma + \varepsilon), \tag{14}$$

$$\mathbb{P}(\max_{i,j} \Delta_{i,j,d}/\sqrt{d} > \sqrt{2}\sigma - \varepsilon), \text{ and} \tag{15}$$

$$\mathbb{P}(\max_{i,j} \Delta_{i,j,d}/\sqrt{d} < \sqrt{2}\sigma + \varepsilon) \tag{16}$$

all go to one as $d \to \infty$.

Proof. Since $\max$ and $\min$ are continuous maps, the continuous mapping theorem and Corollary B.5 imply convergence of the sequences $\{\max_{i,j} \Delta_{i,j,d}/\sqrt{d}\}$ and $\{\min_{i,j} \Delta_{i,j,d}/\sqrt{d}\}$ to $\sqrt{2}\sigma$ in probability. The claim follows since the terms (13), (14) and (15), (16) are lower bounded by

$$\mathbb{P}(|\min_{i,j} \Delta_{i,j,d}/\sqrt{d} - \sqrt{2}\sigma| > \varepsilon) \text{ and } \mathbb{P}(|\max_{i,j} \Delta_{i,j,d}/\sqrt{d} - \sqrt{2}\sigma| > \varepsilon), \tag{17}$$

respectively. □

The next corollary shows that for sufficiently high ambient dimension d, increasing the noise level will increase birth and death times and thus drive persistent homology with high probability.

Corollary B.7. *Consider the setting of Corollary B.5 but for different noise levels $\sigma > \sigma' > 0$. Let $\beta_{d,\sigma}$ and $\delta_{d,\sigma}$ be the minimal birth and death times among all at least one-dimensional homological features at that noise level and ambient dimensionality. Similarly, let $B_{d,\sigma}$ and $D_{d,\sigma}$ be the maximal birth and death times among all at least one-dimensional homological features. Then $\mathbb{P}(\beta_{d,\sigma} > D_{d,\sigma'})$ and hence also $\mathbb{P}(\beta_{d,\sigma} > B_{d,\sigma'}), \mathbb{P}(\delta_{d,\sigma} > D_{d,\sigma'})$ converge to one as $d \to \infty$.*

Proof. Denote the distance between the noised d-dimensional points i, j with noise level σ by $\Delta_{i,j,d,\sigma}$. The main idea is that $\beta_{d,\sigma} \geq \min_{i,j} \Delta_{i,j,d,\sigma}$ and $D_{d,\sigma'} \leq \max_{i,j} \Delta_{i,j,d,\sigma}$. We prove $\mathbb{P}(\beta_{d,\sigma} > D_{d,\sigma'}) \to 1$. The other statements follow from $\beta_{d,\sigma} < B_{d,\sigma}$ and $\delta_{d,\sigma} > \beta_{d,\sigma}$.

Choose any $0 < \varepsilon < (\sigma - \sigma')/\sqrt{2}$. Then $\sqrt{2}\sigma - \varepsilon > \sqrt{2}\sigma' + \varepsilon$. Furthermore, we have

$$\mathbb{P}(b_{d,\sigma} > D_{d,\sigma'}) \geq \mathbb{P}\Big(b_{d,\sigma} > (\sqrt{2}\sigma - \varepsilon)\sqrt{d} \text{ and } (\sqrt{2}\sigma' + \varepsilon)\sqrt{d} > D_{d,\sigma'}\Big).$$

By Corollary B.6, we have for $d \to \infty$ that

$$\mathbb{P}\big(b_{d,\sigma} > (\sqrt{2}\sigma - \varepsilon)\sqrt{d}\big) \geq \mathbb{P}\big(\min_{i,j} \Delta_{i,j,d,\sigma} > (\sqrt{2}\sigma - \varepsilon)\sqrt{d}\big) \to 1.$$

Similarly,

$$\mathbb{P}\big(D_{d,\sigma} < (\sqrt{2}\sigma' + \varepsilon)\sqrt{d}\big) \geq \mathbb{P}\big(\max_{i,j} \Delta_{i,j,d,\sigma} < (\sqrt{2}\sigma' + \varepsilon)\sqrt{d}\big) \to 1.$$

As a result,

$$1 \geq \mathbb{P}\big(b_{d,\sigma} > (\sqrt{2}\sigma - \varepsilon)\sqrt{d} \text{ and } (\sqrt{2}\sigma' + \varepsilon)\sqrt{d} > D_{d,\sigma'}\big) \to 1,$$

and hence $\mathbb{P}(b_{d,\sigma} > D_{d,\sigma'}) \to 1$ for $d \to \infty$. □

These statements show that noise will eventually dominate the Euclidean distances and drive traditional persistent homology if the ambient dimension is large enough. In Figure 8 we saw empirically that tens of ambient dimensions already severely impact traditional persistent homology, while spectral distance offer more noise robustness. This also holds true in hundreds to thousands of ambient dimensions. Compared to the Euclidean distance, spectral distance can detect the correct topology in more than an order of magnitude higher dimensionality, before they eventually fail too (Figure S1).

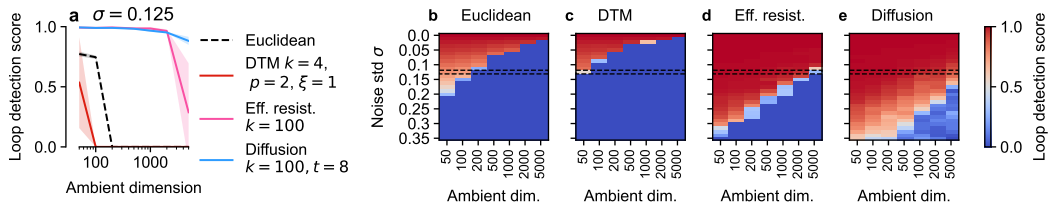

Figure S1: Extension of Figure 8 to higher ambient dimensionalities. **a.** Loop detection scores of various methods on a noisy circle depending on the ambient dimensionality. Due to the higher dimensionalities, we use here the noise with standard deviation $\sigma = 0.125$, one half compared to Figure 8a. **b – e.** Heat maps for $\sigma \in [0, 0.35]$ and $d \in [50, 5\,000]$. Spectral methods are much more noise robust, but eventually also fail to detect the correct topology.

C Persistent homology with PCA

Hiraoka et al. [42] recommend combating the curse of dimensionality by performing a normalized PCA before applying persistent homology. We explore this approach here in our setting.

Let $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times d}$ be the centered data matrix with each data point as a row. Let $X = USV^T$ be the singular value decomposition of X with U and V orthogonal matrices and $S \in \mathbb{R}^{n \times d}$ diagonal with decreasing non-negative values on the diagonal. PCA reduces the dimensionality to $k \leq d$ dimensions by considering the first k columns of US, while normalized PCA instead considers the first k columns of U.

We ran experiments on the 1D datasets for both versions with varying number of principal components (PCs) (Figure S2). Normalization provided no consistent benefit. The performance was worse for higher number of PCs as more noise remained after PCA. Using fewer PCs than the dataset's true dimensionality (Figure S2f, j) also led to poor performance. Knowing the true dimensionality of the data is therefore crucial for this approach, but true dimensionality is typically not available in real-world applications. Effective resistance outperformed PCA preprocessing on the linked circles and eyeglasses dataset and showed similar performance on the circle.

D Continuity of the hole detection score

In this section we prove that our hole detection score is a continuous map, as long as the persistence diagrams have sufficiently many points. This means that small changes in the persistence diagram result in small changes of the hole detection score, a desirable property for a performance measure.

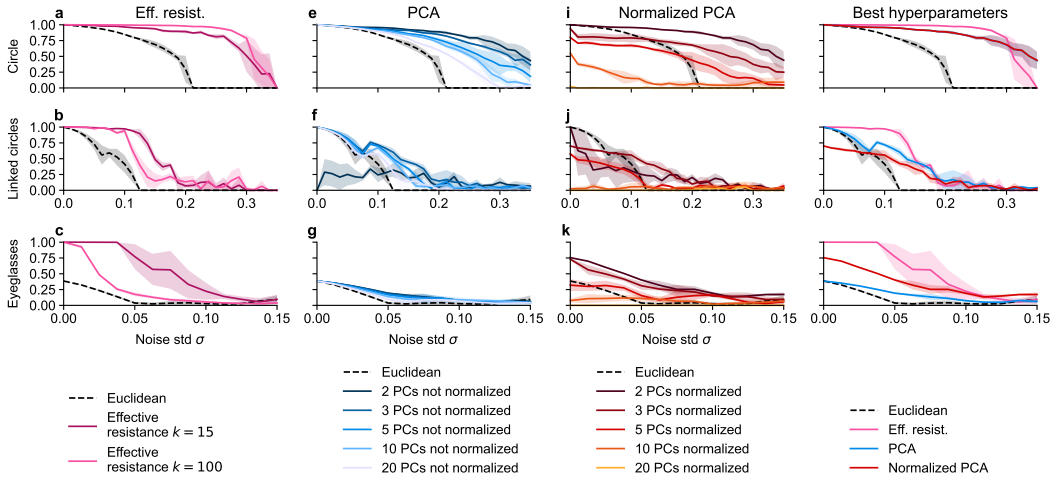

Figure S2: Loop detection score for effective resistance and PCA on three 1D toy datasets.

https://doi.org/10.52202/079017-1328

To state our result formally, we recall some definitions from Cohen-Steiner et al. [19]:

Definition D.1 (Multiset)**.** A multiset is a set in which each element has a multiplicity in $\mathbb{N}_{>0} \cup \infty$.

Definition D.2 (Persistence diagram)**.** A persistence diagram is a multiset with elements from $\{(b, d) \in (\mathbb{R} \cup \{-\infty, \infty\})^2 \mid b < d\}$ counted with multiplicities, together with the diagonal $\{(b, b) \in (\mathbb{R} \cup \{-\infty, \infty\})^2\}$ counted with infinite multiplicity.

Since we build persistence diagrams only for distances-based Vietoris–Rips complexes and do not consider 0-dimensional homologies in this paper, no birth times will be $-\infty$. Moreover, since we do not consider infinite distances (Appendix I), no death times will be infinite either. Therefore, in the following, we will treat persistence diagrams as multisets with elements in $\mathbb{R}^2$.

Definition D.3 (Bottleneck distance)**.** Let D and D' be persistence diagrams. Then the bottleneck distance between them is given by

$$d_B(D, D') = \inf_{\eta: D \to D'} \sup_{x \in D} \|x - \eta(x)\|_\infty,$$

where the infimum is over the bijections between the multisets D and D'.

A bijection η realizing the bottleneck distance between persistence diagrams has three types of paired points: Pairs with both points off the diagonal, pairs with exactly one point off the diagonal, and pairs with both points equal and on the diagonal. Intuitively, the bottleneck distance matches the off-diagonal points between each other. However, the second type of pairs is needed in case the persistence diagrams do not have the same number of points off the diagonal. This is why the full diagonal with infinite multiplicity is considered to belong to each persistence diagram. The third type of pairs does not contribute to the bottleneck distance.

We can now proceed with our proofs. Let $\mathcal{D}$ be the space of all persistence diagrams with finitely many points off the diagonal, endowed with the topology induced by the bottleneck distance d_B.

Lemma D.4. *For any $m \in \mathbb{N}$, the map $p_m : \mathcal{D} \to \mathbb{R}_{\geq 0}$ assigning to a diagram its m-th largest persistence value is continuous.*

Proof. For a persistence diagram $D \in \mathcal{D}$ and a feature $f = (\tau_b, \tau_d) \in D$ we denote by $p(f) = \tau_d - \tau_b$ the persistence of f. Note that for two features f, f' in a persistence diagram, we have $|p(f) - p(f')| \leq \sqrt{2}\|f - f'\|_\infty$.

We need to show that for any $\varepsilon > 0$ and any $D \in \mathcal{D}$, we can choose some $\delta > 0$ such that $|p_m(D') - p_m(D)| < \varepsilon$ for all D' whose bottleneck distance from D is at most δ, i.e., $D' \in B_\delta(D)$. Recall that in this setting, by definition of the bottleneck distance, there is a bijection (of multisets) $\eta : D \to D'$ such that for all $f \in D$ we have $\|f - \eta(f)\|_\infty < \delta$.

Continuity of p_m is easier to demonstrate at diagrams $D \in \mathcal{D}$ for which all off-diagonal points have different persistences. We discuss this case first. Assume D has l off-diagonal points. Denote by f_k the feature with k-th largest persistence.

Let $\tilde{\delta} = \min\{p(f_k) - p(f_{k+1}) \mid k \leq l\}$ and $\delta = \min\left(\tilde{\delta}/(2\sqrt{2}), \varepsilon/(2\sqrt{2})\right)$. Consider $D' \in B_\delta(D)$. By the choice of δ, we have that $|p(\eta(f_k)) - p(f_k)| < \tilde{\delta}/2$, so that the values $p(\eta(f_1)), \ldots, p(\eta(f_l)), 0$ are all distinct and strictly decreasing. As a result, for all $m \leq l$ we have $p_m(D') = p(\eta(f_m))$ and thus

$$|p_m(D') - p_m(D)| = |p(\eta(f_m)) - p(f_m)| < \sqrt{2}\delta < \varepsilon.$$

For $m > l$, we have $p_m(D) = 0$ and any $f' \in D'$ with $p_m(D') = p(f')$ is paired with a point on the diagonal of D under η. Hence

$$|p_m(D') - p_m(D)| = |p(f') - 0| < \sqrt{2}\delta < \varepsilon.$$

The case where D might contain off-diagonal points with identical persistence follows the same overall argument, but requires more careful bookkeeping. Assume D has l off-diagonal points. Let $\tilde{\delta} = \min\left(\{p_k(D) - p_{k+1}(D) \mid k \leq l\} \backslash \{0\}\right)$. Let

$$\delta = \min\left(\tilde{\delta}/(2\sqrt{2}), \varepsilon/(2\sqrt{2})\right).$$

Set

$$F_k = \{f \in D \mid p(f) = p_k(D)\}$$

for $k \leq l$ and F_{l+1} equal to the diagonal. These multisets all contain at least one element. We have $p_k(D) = p_{k+1}(D)$ if and only if $F_k = F_{k+1}$.

As a result, for all $k \leq l$, we have $|\bigcup_{i=1}^{k} F_i| = k$ if and only if $F_k \neq F_{k+1}$ which in turn holds if and only if $p_k(D) \neq p_{k+1}(D)$.

For any $f, \tilde{f} \in D$ with $p(f) > p(\tilde{f})$ we have

$$\begin{aligned} p(\eta(f)) - p(\eta(\tilde{f})) &= p(\eta(f)) - p(f) + p(f) - p(\tilde{f}) + p(\tilde{f}) - p(\eta(\tilde{f})) \\ &\geq p(f) - p(\tilde{f}) - |p(\eta(f)) - p(f)| - |p(\tilde{f}) - p(\eta(\tilde{f}))| \\ &> p(f) - p(\tilde{f}) - \tilde{\delta} \\ &\geq 0, \end{aligned} \tag{18}$$

so $p(\eta(f)) > p(\eta(\tilde{f}))$. We will now show that if $m \leq l$ and $f' \in D'$ such that $p_m(D') = p(f')$, then $\eta^{-1}(f') \in F_m$.

Let a be the highest number below m such that $F_a \neq F_m$. Then

$$\left|\bigcup_{i=1}^{a} \eta(F_i)\right| = \left|\bigcup_{i=1}^{a} F_i\right| = a.$$

and $\bigcup_{i=1}^{a} \eta(F_i)$ contains the $a < m$ features of D' with smallest persistence. In particular, $\eta^{-1}(f') \notin \bigcup_{i=1}^{a} F_i$.

Let b be the largest number such that $F_h = F_m$. By a similar argument, $\bigcup_{i=1}^{b} \eta(F_i)$ contains the $b \geq m$ features in D' with largest persistence, including f'. Thus, $\eta^{-1}(f') \in \bigcup_{i=a+1}^{b} F_i$. But by choice of a and b, we have $F_{a+1} = \cdots = F_m = \cdots = F_b$, so $\eta^{-1}(f') \in F_m$.

Thus,

$$|p_m(D') - p_m(D)| = |p(f') - p(\eta^{-1}(f'))| \leq \sqrt{2}\|f' - \eta^{-1}(f')\|_\infty < \varepsilon.$$

For $m > l$, we have $p_m(D) = 0$ and F_m is the set of points on the diagonal of D. Moreover,

$$\left|\bigcup_{i=1}^{l} F_i\right| = l = \left|\bigcup_{i=1}^{l} \eta(F_i)\right|,$$

so that any $f' \in D'$ with $p_m(D') = p(f')$ must be in the image of the diagonal under η and we can proceed as in the case of unique positive persistences. □

Our hole detection score is a continuous score if there are at least m points in the persistence diagram, and not continuous otherwise:

Proposition D.5. *Let $\mathcal{D}$ be the space of all persistence diagrams with finitely many points off the diagonal, endowed with the topology induced by the bottleneck distance d_B. Let $m \in \mathbb{N}$ and define $\mathcal{D}'$ as the subset of $\mathcal{D}$ with at least m points off the diagonal. Then the map $s_m : \mathcal{D} \to \mathbb{R}_{\geq 0}$ is continuous at all diagrams in $\mathcal{D}'$ but is discontinuous at all diagrams in $\mathcal{D} \backslash \mathcal{D}'$.*

Proof. The discontinuity of s_m at diagrams with less than m points happens when adding the m-th point to a diagram. Let D be a persistence diagram with $l < m$ points off the diagonal. Let $\varepsilon > 0$ and define D' as the persistence diagram obtained by adding $m - l$ copies of $(0, \varepsilon/2)$. Then $s_m(D) = 0$ and $s_m(D') = 1$, because D' has exactly m off-diagonal points, but $d_B(D, D') < \varepsilon$. Thus, s_m is not continuous at D.

For the continuity part, we write s_m as the composition of two continuous maps. The map $q : \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $(x, y) \mapsto (x - y)/x$ is continuous. Furthermore, since each diagram in $\mathcal{D}'$ has at least m points off the diagonal, $p_m(D) \in \mathbb{R}_{>0}$ for all $D \in \mathcal{D}'$. By lemma D.4 the map $p_m \times p_{m+1} : \mathcal{D}' \to \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ is continuous. Finally, $s_m : \mathcal{D}' \to \mathbb{R}_{\geq 0}$ factors as $q \circ (p_m \times p_{m+1})$, showing its continuity. □

https://doi.org/10.52202/079017-1328

In other words, on persistence diagrams with sufficiently many points our hole detection score is continuous. This makes our score robust against small perturbations in the persistence diagram, a property expected for a reliable score.

In the typical case, where the persistence diagram contains a noise cloud of many points close to the diagonal and, possibly, some outliers far away from the diagonal, the requirement on the number of points is satisfied. Only in the case where we expect a large gap after the m-th feature, but the diagram does not even have m features, can there be a sudden jump in our hole detection score.

E Alternative performance scores

While our hole detection score corresponds to an intuitive visual assessment of persistence diagrams and is continuous (Appendix D), in this section we discuss two alternative performance scores. We argue that both of them are less suited for our benchmark.

E.1 Widest gap score

A natural way to interpret a persistence diagram is to deem all features above the widest gap in persistence values as true features and the rest as noise, i.e., infer that a diagram has a true features if $a = \text{argmax}_\alpha(p_\alpha - p_{\alpha+1})$.

We considered the binary score which is equal to 1 if the number of features above the widest gap equals the number of ground-truth features and 0 otherwise. We call it the *widest gap score*.

Note that our hole detection score can be lower than 1 even if the persistence diagram clearly shows the correct number of outlier features (e.g. $s_m \approx 0.5$ for $\sigma = 0.2$ in Figure 3b). In contrast, the widest gap score is equal to 1 in this case.

On the other hand, the binary nature of the widest gap score leads to instability in case of nearly equisized gaps. For instance in Figure 9d, a small perturbation could make the gap after the most persistent feature the widest, dropping the widest gap score from 1 to 0. In contrast, our $s_2 \approx 0.9$ correctly reflects the two clear outlier features.

We evaluated several distances on the three 1D datasets using the widest gap score in Figure S3. The binary nature of the widest gap score led to high-variance results even though we increased the number of random seeds from 3 to 10. Spectral methods still performed better than Euclidean, Fermat, and DTM distances. However, the widest gap score led to some false positives. For example, for the Fermat distance and $\sigma = 0.3$ for the circle dataset, in 9 of 10 trials the widest gap appeared after the first feature. However, all 9 persistence diagrams looked like noise clouds and in 8 cases the most persistent feature's representative cycle was clearly a noise feature (Figure S3d – e). Conversely, for the effective resistance there was always a clear outlier in the diagram and the representative did follow the ground-truth loop (Figure S3f – g). Although the Fermat distance failed in this setting and the effective resistance performed well, their widest gap scores were both high.

For these reasons, we prefer our hole detection score to the widest gap score.

E.2 Cycle matching

The need to distinguish between true signal and noise features in a persistence diagram motivated Reani and Bobrowski [64] to develop *cycle matching*, a technique that quantifies the correspondence of features in two persistence diagrams by a prevalence score in $[0, 1]$. True signal corresponds to features that reappear with high prevalence in variations of the data obtained, e.g., via resampling.

This approach to detecting the true topology of high-dimensional data is orthogonal to our exploration of different input distances. Indeed, cycle matching can be combined with any distance.

An important downside to cycle matching is the need for resampling. In an exploratory context, the true data distribution is not known. Instead, Reani and Bobrowski [64] suggest to either bootstrap the existing finite dataset or sample from a kernel density estimate. Unfortunately, for high-dimensional data both approaches are problematic. Bootstrapping high-dimensional data is error-prone [18] and persistent homology of bootstraps is biased because repeated points effectively decrease the sample

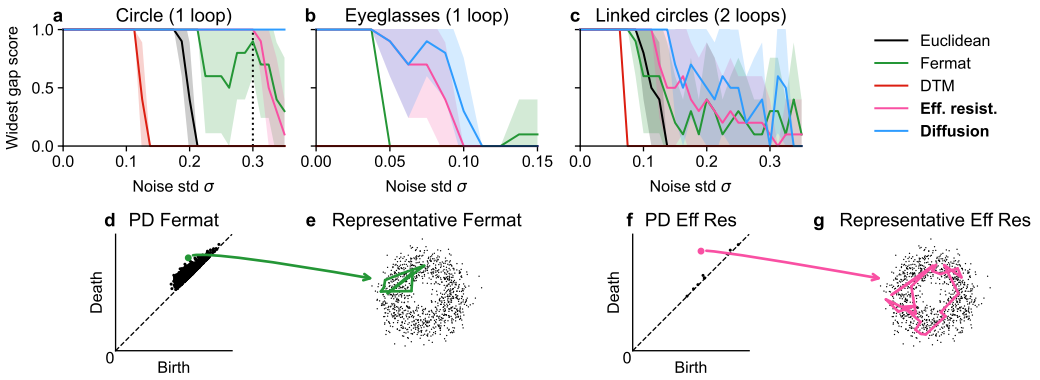

Figure S3: Top row: Widest gap score on the three 1D toy datasets. We used 10 random seeds here. Spectral methods outperformed Fermat, DTM, and Euclidean in this score. Bottom row: Example persistence diagrams and representatives for the toy circle with $\sigma = 0.3$ with Fermat distance and effective resistance illustrate that despite the high score for both, Fermat distance actually failed.

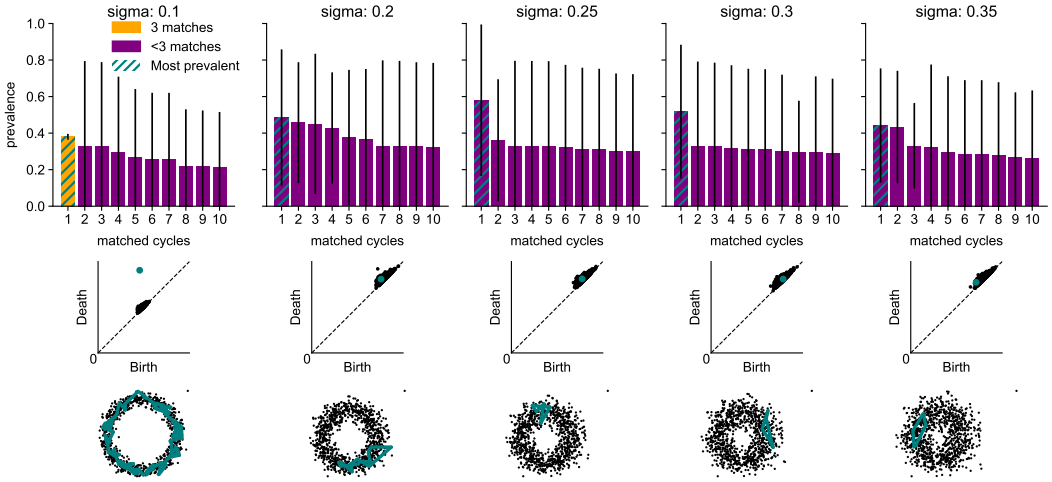

Figure S4: Results for cycle matching with the Euclidean distance on a noisy circle in $\mathbb{R}^{50}$ with noise level σ. Top row: Prevalences of the 10 most prevalent cycles. Means and standard deviation over three seeds. Color indicates whether the cycle was matched for all three random seeds or not. Second row: Persistence diagrams with the most prevalent features highlighted. Third row: Representative of the most prevalent feature overlaid on a 2D PCA of the data.

size [71]. Kernel density estimation requires a prohibitively large sample size in high dimensions [86, chapter 6.5].

Nevertheless, as a proof of concept, we explored cycle matching with the relatively fast `ripser`-based implementation of García-Redondo et al. [30]. The authors report a runtime of about 90 minutes for a dataset of sample size $n = 1000$ (their Table 3) and in our experiments we experienced even longer run times. As the result, this implementation of cycle matching was about 450 times slower than our approach (Table S5). For this reason, we only used three resamples.

García-Redondo et al. [30]'s implementation accepts datasets as point clouds and implicitly assumes the Euclidean distance. Therefore, in this experiment we only used Euclidean distances and diffusion distances, which can be realized as Euclidean distances (Eq. 4). We computed prevalences of matched cycles on the noisy circle with $n = 1000$ points in $\mathbb{R}^{50}$ for Gaussian noise of standard deviations $\sigma \in [0.1, 0.2, 0.25, 0.3, 0.35]$.

https://doi.org/10.52202/079017-1328

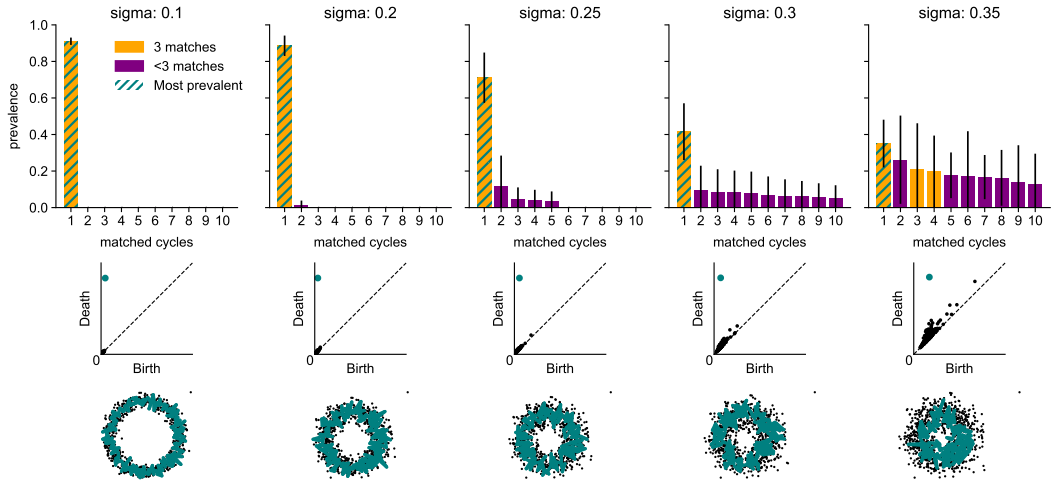

Figure S5: Results for cycle matching with the diffusion distance ($k = 15, t = 8$) on a noisy circle in $\mathbb{R}^{50}$ with noise level σ. Top row: Prevalences of the 10 most prevalent cycles. Means and standard deviation over three seeds. Color indicates whether the cycle was matched for all three random seeds or not. Second row: Persistence diagrams with the most prevalent features highlighted. Third row: Representative of the most prevalent feature overlaid on a 2D PCA of the data.

Cycle matching with the Euclidean distance did not reliably identify the ground-truth feature for $sigma > 0.1$ (Figure S4). In all experiments with the Euclidean distance and $\sigma > 0.1$, the feature with the highest prevalence did not correspond to the ground-truth loop.

For $\sigma = 0.2, 0.25$, we also checked the prevalences of the ground truth cycle. For $\sigma = 0.2$ the ground truth cycle was not matched for any of the three resamples, leading to a prevalence of 0. For $\sigma = 0.25$ the ground-truth cycle had the average prevalence of 0.1 ± 0.1, much lower than the cycle with the highest prevalence.

One might hope that the cycles that do get matched for all resamples are true topological features, even if their prevalence values are low. Indeed, for $\sigma = 0.1$ only the correct cycle was matched for all three resamples. However this heuristic did not work either because for $\sigma \in \{0.2, 0.25, 0.3\}$ no cycle was matched for all three resamples, while for $\sigma = 0.35$ there was a cycle matched for all three resamples (not shown in Figure S4 because it was not among the 10 most prevalent ones), but it did not encode the ground-truth feature (and its prevalence 0.1 ± 0.1 was very low).

The fact that most cycles were not matched for all resamples led to high standard deviations of the prevalences. The reason for the high mean prevalences for $\sigma > 0.1$ was that some noise cycles got matched with very high prevalence (sometimes > 0.9) for one or two of the resamples. For the same reason, the prevalence of the ground truth cycle for $\sigma = 0.1$ was not much higher than the second highest prevalence. Increasing the number of resamples may help, but would make the procedure even more computationally expensive.

Overall, we conclude that cycle matching is not only very slow, but also does not alleviate the curse of dimensionality for persistent homology. In contrast, our spectral methods produced a near-perfect loop detection score until $\sigma = 0.3$ (Figure 6).

That said, cycle matching can serve as an alternative performance score. We applied cycle matching to the diffusion distance (Figure S5) and found that

1. the maximal prevalences were larger than for the Euclidean distance for $\sigma < 0.3$,
2. the cycle with maximal prevalence always represented the ground truth loop,
3. the prevalence of the correct cycle was a clear outlier among all prevalences for $\sigma \leq 0.3$,
4. the cycle representing the ground truth feature was always matched for all three resamples and was the only one matched for all three resamples for $\sigma \leq 0.3$.

This confirms the superior performance of spectral methods for detecting topology in high-dimensional settings.

F Effective resistance as a spectral method

In this section, we derive an explicit spectral embedding realizing the square root of von Luxburg et al. [81]'s *corrected* effective resistance. We also show that it is not necessarily a proper metric.

Let us recall the notation from Sections 5 and 6, extended to weighted graphs. Let G be a weighted connected graph with n nodes. We denote by $A = (a_{ij})_{i,j=1,\dots,n}$ the weighted adjacency matrix whose entries $a_{ij} = a_{ji}$ equal the edge weight of edge ij if edge ij is part of the graph and zero otherwise. We further denote by $D = \operatorname{diag}(d_i)$ the degree matrix, where $d_i = \sum_j a_{ij}$ are the node degrees. We define $\operatorname{vol}(G) = \sum_i d_i$. Let further $A^{\text{sym}} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ and $L^{\text{sym}} = I - A^{\text{sym}}$ be the symmetrically normalized adjacency matrix and the symmetrically normalized graph Laplacian. Denote the eigenvectors of L^{sym} by $u_1, \dots, u_n$ and their eigenvalues by $\mu_1, \dots, \mu_n$ in increasing order. The eigenvectors and eigenvalues of A^{sym} are $u_1, \dots, u_n$ and $1 - \mu_1, \dots, 1 - \mu_n$.

Definition F.1. The naive effective resistance distance between nodes i and j is defined as

$$\tilde{d}_{ij}^{\text{eff}} = \frac{1}{\operatorname{vol}(G)}(H_{ij} + H_{ji}), \tag{19}$$

where H_{ij} is the hitting time from i to j, i.e., the expected number of steps that a random walker starting at node i takes to reach node j for the first time.

The corrected effective resistance distance between nodes i and j is defined as

$$d_{ij}^{\text{eff}} = \tilde{d}_{ij}^{\text{eff}} - \frac{1}{d_i} - \frac{1}{d_j} + 2\frac{a_{ij}}{d_i d_j} - \frac{a_{ii}}{d_i^2} - \frac{a_{jj}}{d_j^2}. \tag{20}$$

The following proposition refines Proposition 4 in [81].

Proposition F.2. *The corrected effective resistance distance d_{ij}^{eff} between nodes i and j can be computed by $d_{ij}^{\text{eff}} = \|e_i^{\text{eff}} - e_j^{\text{eff}}\|^2$, where*

$$e_i^{\text{eff}} = \frac{1}{\sqrt{d_i}}\left(\frac{1-\mu_2}{\sqrt{\mu_2}}u_{2,i}, \dots, \frac{1-\mu_n}{\sqrt{\mu_n}}u_{n,i}\right).$$

Proof. By the fourth step of the large equation in the proof of Proposition 2 in [82], we have

$$\frac{1}{\operatorname{vol}(G)}H_{ij} = \frac{1}{d_j} + \langle b_j, A^{\text{sym}}(b_j - b_i)\rangle + \sum_{r=2}^{n}\frac{1}{\mu_r}\langle A^{\text{sym}}b_j, u_r u_r^\top A^{\text{sym}}(b_j - b_i)\rangle, \tag{21}$$

where $b_i = \frac{1}{\sqrt{d_i}}\hat{e}_i = D^{-\frac{1}{2}}\hat{e}_i$ with $\hat{e}_i$ being the i-th standard basis vector.

Adding this expression for ij and ji and using the definition of $\tilde{d}_{ij}^{\text{eff}}$, we get

$$\begin{aligned}
\tilde{d}_{ij}^{\text{eff}} - \frac{1}{d_i} - \frac{1}{d_j} &= \frac{1}{\operatorname{vol}(G)}(H_{ij} + H_{ji}) - \frac{1}{d_i} - \frac{1}{d_j} && (22)\\
&= \langle b_j, A^{\text{sym}}(b_j - b_i)\rangle + \langle b_i, A^{\text{sym}}(b_i - b_j)\rangle && (23)\\
&\quad + \sum_{r=2}^{n}\frac{1}{\mu_r}\langle A^{\text{sym}}b_j, u_r u_r^\top A^{\text{sym}}(b_j - b_i)\rangle && (24)\\
&\quad + \sum_{r=2}^{n}\frac{1}{\mu_r}\langle A^{\text{sym}}b_i, u_r u_r^\top A^{\text{sym}}(b_i - b_j)\rangle && (25)\\
&= \langle b_j - b_i, A^{\text{sym}}(b_j - b_i)\rangle && (26)\\
&\quad + \sum_{r=2}^{n}\frac{1}{\mu_r}\langle A^{\text{sym}}(b_j - b_i), u_r u_r^\top A^{\text{sym}}(b_j - b_i)\rangle && (27)
\end{aligned}$$

https://doi.org/10.52202/079017-1328

For ease of exposition, we treat the two terms separately. By unpacking the definitions and using the symmetry of D, we get

$$\langle b_j - b_i, A^{\text{sym}}(b_j - b_i)\rangle = \langle D^{-\frac{1}{2}}(\hat{e}_j - \hat{e}_i), A^{\text{sym}} D^{-\frac{1}{2}}(\hat{e}_j - \hat{e}_i)\rangle \tag{28}$$

$$= \langle (\hat{e}_j - \hat{e}_i), D^{-1} A D^{-1}(\hat{e}_j - \hat{e}_i)\rangle \tag{29}$$

$$= \frac{a_{jj}}{d_j^2} - 2\frac{a_{ij}}{d_i d_j} + \frac{a_{ii}}{d_i^2} \tag{30}$$

Since the u_r are eigenvectors of A^{sym} with eigenvalue $1 - \mu_r$ and A^{sym} is symmetric, we also get

$$\sum_{r=2}^{n} \frac{1}{\mu_r} \langle A^{\text{sym}}(b_j - b_i), u_r u_r^\top A^{\text{sym}}(b_j - b_i)\rangle \tag{31}$$

$$= \sum_{r=2}^{n} \frac{1}{\mu_r} \langle b_j - b_i, (A^{\text{sym}} u_r)(A^{\text{sym}} u_r)^\top (b_j - b_i))\rangle \tag{32}$$

$$= \sum_{r=2}^{n} \frac{1}{\mu_r} \langle b_j - b_i, (1-\mu_r) u_r \big((1-\mu_r) u_r\big)^\top (b_j - b_i)\rangle \tag{33}$$

$$= \sum_{r=2}^{n} \left(\frac{1-\mu_r}{\sqrt{\mu_r}} u_r^\top D^{-\frac{1}{2}}(\hat{e}_j - \hat{e}_i) \right)^2 \tag{34}$$

$$= \left\| \frac{1}{\sqrt{d_j}} \left(\frac{1-\mu_r}{\sqrt{\mu_r}} u_{r,j} \right)_{r=2,\dots,n} - \frac{1}{\sqrt{d_i}} \left(\frac{1-\mu_r}{\sqrt{\mu_r}} u_{r,i} \right)_{r=2,\dots,n} \right\|^2 \tag{35}$$

$$= \|e_j - e_i\|^2 \tag{36}$$

Putting everything together yields the result

$$d_{ij}^{\text{eff}} = \tilde{d}_{ij}^{\text{eff}} - \frac{1}{d_i} - \frac{1}{d_j} + 2\frac{a_{ij}}{d_i d_j} - \frac{a_{ii}}{d_i^2} - \frac{a_{jj}}{d_j^2} \tag{37}$$

$$= \frac{a_{jj}}{d_j^2} - 2\frac{a_{ij}}{d_i d_j} + \frac{a_{ii}}{d_i^2} + \|e_j - e_i\|^2 + 2\frac{a_{ij}}{d_i d_j} - \frac{a_{ii}}{d_i^2} - \frac{a_{jj}}{d_j^2} \tag{38}$$

$$= \|e_j^{\text{eff}} - e_i^{\text{eff}}\|^2. \tag{39}$$

□

The corrected version of effective resistance and diffusion distances are, in general, not proper metrics, unlike the naive effective resistance [32, Corollary 2.4.]. We show this by giving concrete examples of graphs where the metric axioms for these distances do not hold.

Proposition F.3. *Neither corrected effective resistance nor diffusion distances between distinct points are necessarily positive. Moreover, corrected effective resistance does not always satisfy the triangle inequality.*

Proof. Consider the unweighted chain graph with three nodes (Figure S6, left). The uncorrected effective resistance between the first and the last node is 2. So the corrected effective resistance between these distinct nodes is $2 - 1 - 1 = 0$. On the same graph, the random walker is necessarily at node 1 after one step, independent whether it started at node 0 or 2, so the diffusion distance with $t = 1$ between these two distinct nodes is zero.

Consider now an unweighted graph with 5 nodes, the first three of which form a triangle and the other two a chain connected to the triangle (Figure S6, right). Then

$$\tilde{d}_{04}^{\text{eff}} = \frac{8}{3}, \quad \tilde{d}_{02}^{\text{eff}} = \frac{2}{3}, \quad \tilde{d}_{24}^{\text{eff}} = 2; \quad d_{04}^{\text{eff}} = \frac{7}{6}, \quad d_{02}^{\text{eff}} = \frac{1}{6}, \quad d_{24}^{\text{eff}} = \frac{2}{3}. \tag{40}$$

We see that the triangle $0, 2, 4$ violates the triangle inequality

$$d_{04}^{\text{eff}} = \frac{7}{6} > \frac{2}{3} + \frac{1}{6} = d_{02}^{\text{eff}} + d_{24}^{\text{eff}}.$$

□

Figure S6: Counterexample graphs for Proposition F.3

The square root of corrected effective resistance does satisfy the triangle inequality and can also be used as input to persistent homology. This amounts to taking the square root of all birth and death times found with the corrected effective resistance as computing persistent homology commutes with strictly monotonic maps, and hence does not strongly affect the loop detection performance of corrected effective resistance (Figure S20). We therefore believe that it is not a problem for our persistent homology application that corrected effective resistance fails to satisfy the triangle inequality.

G Effective resistance integrates diffusion distances

First, we describe the connection between diffusion distances and naive effective resistance communicated to us by Mémoli [54]. To the best of our abilities, we could not find another reference. We use the same notation as in Section 6 and Appendix F.

Below, we speak of diffusion distances at half-integer time points, although the random walk analogy does not extend to half-steps. What we mean is inserting a half-integer value of t in Eq. (4).

Proposition G.1. *The sum of the squared diffusion distances over all time points $t/2$ for $t = 0, 1, 2, \ldots$ equals the naive effective resistance. Formally, if G is a connected and non-bipartite graph, we have*

$$\frac{1}{\operatorname{vol}(G)} \sum_{t=0}^{\infty} d_{ij}^{\text{diff}}(t/2)^2 = \tilde{d}_{ij}^{\text{eff}}. \tag{41}$$

Proof. This is an application of the geometric series. Since G is not bipartite, the largest eigenvalue μ_n of its Laplacian L^{sym} is smaller than 2 [17, Lemma 1.7]. Recall that by Eq. (4) the diffusion distance for t diffusion steps is given by

$$d_{ij}^{\text{diff}}(t) = \sqrt{\operatorname{vol}(G)} \|e_i^{\text{diff}}(t) - e_j^{\text{diff}}(t)\|, \text{ where } e_i^{\text{diff}}(t) = \frac{1}{\sqrt{d_i}} \left((1-\mu_2)^t u_{2,i}, \ldots, (1-\mu_n)^t u_{n,i} \right). \tag{42}$$

and that the naive effective resistance is given by

$$\tilde{d}_{ij}^{\text{eff}} = \|\tilde{e}_i^{\text{eff}} - \tilde{e}_j^{\text{eff}}\|^2, \text{ where } \tilde{e}_i^{\text{eff}} = \frac{1}{\sqrt{d_i}} \left(\frac{1}{\sqrt{\mu_2}} u_{2,i}, \ldots, \frac{1}{\sqrt{\mu_n}} u_{n,i} \right). \tag{43}$$

We compute

$$\begin{aligned}
\frac{1}{\operatorname{vol}(G)} \sum_{t=0}^{\infty} d_{ij}^{\text{diff}}(t/2)^2 &= \sum_{t=0}^{\infty} \sum_{l=2}^{n} (1-\mu_l)^{2\cdot t/2} \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 && (44)\\
&= \sum_{l=2}^{n} \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 \sum_{t=0}^{\infty} (1-\mu_l)^t && (45)\\
&= \sum_{l=2}^{n} \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 \frac{1}{1-(1-\mu_l)} && (46)\\
&= \sum_{l=2}^{n} \frac{1}{\mu_l} \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 && (47)\\
&= \|\tilde{e}_i^{\text{eff}} - \tilde{e}_j^{\text{eff}}\|^2 && (48)\\
&= \tilde{d}_{ij}^{\text{eff}}. && (49)
\end{aligned}$$

https://doi.org/10.52202/079017-1328

The application of the geometric series is justified as $\mu_l \in (0,2)$ for $l = 2, \dots, n$ by assumption on G, so that $|1 - \mu_l| < 1$. □

Similarly, we show the corresponding result for the corrected version of effective resistance.

Corollary G.2. *The sum of the squared diffusion distances over all time points $t/2$ for $t = 2, 3, \dots$ equals the corrected effective resistance. Formally, if G is a connected non-bipartite graph, we have*

$$\frac{1}{\mathrm{vol}(G)} \sum_{t=2}^{\infty} d_{ij}^{\mathrm{diff}}(t/2)^2 = d_{ij}^{\mathrm{eff}} \tag{50}$$

and

$$\frac{d_{ij}^{\mathrm{diff}}(0)^2}{\mathrm{vol}(G)} = \frac{1}{d_i} + \frac{1}{d_j}, \qquad \frac{d_{ij}^{\mathrm{diff}}(1/2)^2}{\mathrm{vol}(G)} = \frac{a_{ii}}{d_i^2} + \frac{a_{jj}}{d_j^2} - 2\frac{a_{ij}}{d_i d_j} \tag{51}$$

Proof. For $x \in (-1, 1)$, starting a geometric series at the m-th term yields $\sum_{t=m}^{\infty} x^t = \frac{x^m}{1-x}$. By Proposition F.2, we have

$$d_{ij}^{\mathrm{eff}} = \sum_{l=2}^{n} \frac{(1-\mu_l)^2}{\mu_l} \left(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \right)^2 = \sum_{l=2}^{n} \frac{(1-\mu_l)^2}{1-(1-\mu_l)} \left(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \right)^2. \tag{52}$$

By the same argument as in the proof of Proposition G.1, this shows the first part of the Corollary.

The statement for diffusion distance after $t = 0$ steps is an easy computation from the original definition, Eq. (2):

$$\begin{aligned} d_{ij}^{\mathrm{diff}}(0) &= \sqrt{\mathrm{vol}(G)} \|(P^0_{i,:} - P^0_{j,:}) D^{-0.5}\| \\ &= \sqrt{\mathrm{vol}(G)} \|(\hat{e}_i^\top - \hat{e}_j^\top) D^{-0.5}\| \\ &= \sqrt{\mathrm{vol}(G) \left(\frac{1}{d_i} + \frac{1}{d_j} \right)}, \end{aligned} \tag{53}$$

where $\hat{e}_i$ is the i-th standard basis vector. The last part follows from the definition of the corrected effective resistance as

$$d_{ij}^{\mathrm{eff}} = \tilde{d}_{ij}^{\mathrm{eff}} - 1/d_i - 1/d_j + 2a_{ij}/(d_i d_j) - a_{ii}/d_i^2 - a_{jj}/d_j^2. \tag{54}$$

and the expression of the naive effective resistance as full geometric series of squared diffusion distances in Proposition G.1. □

H Diffusion pseudotime

Diffusion pseudotime (DPT), d^{dpt}, [36] also considers diffusion processes of arbitrary length. There are several variants of diffusion pseudotime [36, 88]. They are all computed as $d_{ij}^{\mathrm{dpt}} = \|e_i^{\mathrm{dpt}} - e_j^{\mathrm{dpt}}\|$ for

$$e_i^{\mathrm{dpt}} = \left(\frac{1-\mu_2}{\mu_2} v_{2,i}, \dots, \frac{1-\mu_n}{\mu_n} v_{n,i} \right). \tag{55}$$

The difference is the $v_1, \dots, v_n$'s. In the original publication [36], they were the normalized eigenvectors of the random walker graph Laplacian $D^{-\frac{1}{2}} L^{\mathrm{sym}} D^{\frac{1}{2}}$. These are given by $v_l = D^{-\frac{1}{2}} u_l / \|D^{-\frac{1}{2}} u_l\|$. Wolf et al. [88] introduced a version using the eigenvectors of the symmetric graph Laplacian L^{sym}, so that $v_l = u_l$. Closest to the corrected resistance distance is the case where $v_l = D^{-\frac{1}{2}} u_l$ are non-normalized. We refer to these three versions as "rw", "sym" and "symd".

The only difference between the "symd" version of DPT and corrected effective resistance is that the former decays eigenvalues with $(1-\mu_l)/\mu_l$, while the latter decays it with $(1-\mu_l)/\sqrt{\mu_l}$, so that DPT decays large eigenvalues more strongly than corrected effective resistance (Figure 5).

Similar to effective resistance, one can also write diffusion pseudotime in terms of diffusion distances. But for diffusion pseudotime the diffusion distances corresponding to higher diffusion times contribute more.

Proposition H.1. *Let G be a connected, non-bipartite graph. We can write the "symd" version of DPT as*

$$d_{ij}^{\text{dpt}} = \sqrt{\frac{1}{\text{vol}(G)} \sum_{t=1}^{\infty} (t-1) d_{ij}^{\text{diff}}(t/2)^2}. \tag{56}$$

Proof. We will use that for a real number x with $|x| < 1$ the derivative of the geometric series is

$$\sum_{t=1}^{\infty} t x^{t-1} = \frac{1}{(1-x)^2}.$$

Multiplying with x^2, we get

$$\frac{x^2}{(1-x)^2} = \sum_{t=1}^{\infty} t x^{t+1} = \sum_{t=2}^{\infty} (t-1) x^t. \tag{57}$$

By assumption on G, for all eigenvalues μ_l we have $|1-\mu_l| < 1$ for $l = 2, \dots, n$, so that the above equation holds for all $x = 1 - \mu_l$ with $l = 2, \dots, n$. Together with the "symd" expression for diffusion pseudotime, we compute

$$\begin{aligned}
\left(d_{ij}^{\text{dpt}}\right)^2 &= \sum_{l=2}^{n} \frac{(1-\mu_l)^2}{\mu_l^2} \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 \\
&= \sum_{l=2}^{n} \sum_{t=2}^{\infty} (t-1)(1-\mu_l)^t \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 \\
&= \sum_{t=2}^{\infty} (t-1) \sum_{l=2}^{n} (1-\mu_l)^{2\cdot(t/2)} \Big(\frac{u_{l,i}}{\sqrt{d_i}} - \frac{u_{l,j}}{\sqrt{d_j}} \Big)^2 \\
&= \sum_{t=2}^{\infty} (t-1) d_{ij}^{\text{diff}}(t/2)^2.
\end{aligned} \tag{58}$$

□

H.1 Performance of diffusion pseudotime and potential distance

Here we compare the spectral methods of our main benchmark to the three versions of DPT and potential distance.

For DPT, we used all three versions defined above and used $k \in \{15, 100\}$ nearest neighbors.

The potential distance underlies the visualization method PHATE [56] and is closely related to the diffusion distance. It is defined by

$$d_t^{\text{pot}}(i,j) = \| \log(P_{i,:}^t) - \log(P_{j,:}^t) \|,$$

where the logarithm is applied element-wise. We used $k \in \{15, 100\}$ and $t \in \{8, 64\}$.

Overall, all spectral methods performed very similarly on the 1D datasets circle, linked circles, and eyeglasses (Figures S7, S8), as well as on the single-cell datasets (Figure S9).

I Details on the distances used in our benchmark

Let $x_1, \dots, x_n \in \mathbb{R}^d$. We denote pairwise Euclidean distances by $d_{ij} = \|x_i - x_j\|$, the k nearest neighbors of x_i in increasing distance by $x_{i_1}, \dots, x_{i_k}$, and the set containing them by N_i. Many distances rely on the symmetric k-nearest-neighbor (skNN) graph. This graph contains edge ij if x_i is among the k nearest neighbors of x_j or vice versa.

https://doi.org/10.52202/079017-1328

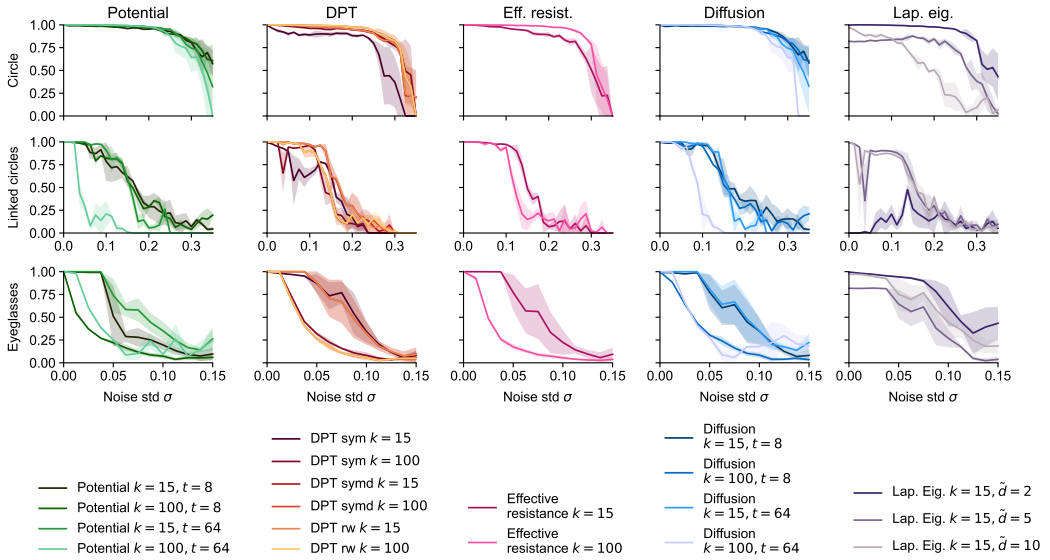

Figure S7: Comparison of all spectral methods on the noised versions of the circle, the linked circles, and the eyeglassed dataset in $\mathbb{R}^{50}$.

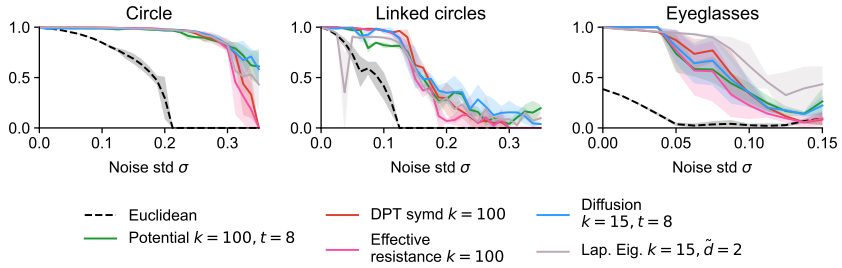

Figure S8: Best hyperparameter choices for the methods in Figure S7. All spectral methods reach very similar performance.

Fermat distances For $p \geq 1$, the Fermat distance is defined as

$$d_p^F(ij) = \inf_{\pi} \left(\sum_{(uv)\in\pi} d_{uv}^p \right), \tag{59}$$

where the infimum is taken over all finite paths from x_i to x_j in the complete graph with edge weights d_{ij}^p. As a speed-up, Fernández et al. [26] suggested to compute the shortest paths only on the kNN graph, but for our sample sizes we could perform the calculation on the complete graph. For $p = 1$ this reduces to normal Euclidean distances due to the triangle inequality. We used $p \in \{2, 3, 5, 7\}$.

DTM distances The DTM distances depend on three hyperparameters: the number of nearest neighbors k, one hyperparameter controlling the distance to measure p, and finally a hyperparameter ξ controlling the combination of DTM and Euclidean distance. The DTM value for each point is given by

$$\mathrm{dtm}_i = \begin{cases} \sqrt[p]{\sum_{\kappa=1}^{k} \|x_i - x_{i_\kappa}\|^p / k} & \text{if } p < \infty \\ \|x_i - x_{i_k}\| & \text{else.} \end{cases} \tag{60}$$

These values are combined with pairwise Euclidean distances to give pairwise DTM distances:

$$d_{k,p,\xi}^{\mathrm{DTM}}(ij) = \begin{cases} \max(\mathrm{dtm}_i, \mathrm{dtm}_j) & \text{if } \|x_i - x_j\| \leq \sqrt[\xi]{|\mathrm{dtm}_i^\xi - \mathrm{dtm}_j^\xi|} \\ \theta & \text{else,} \end{cases} \tag{61}$$

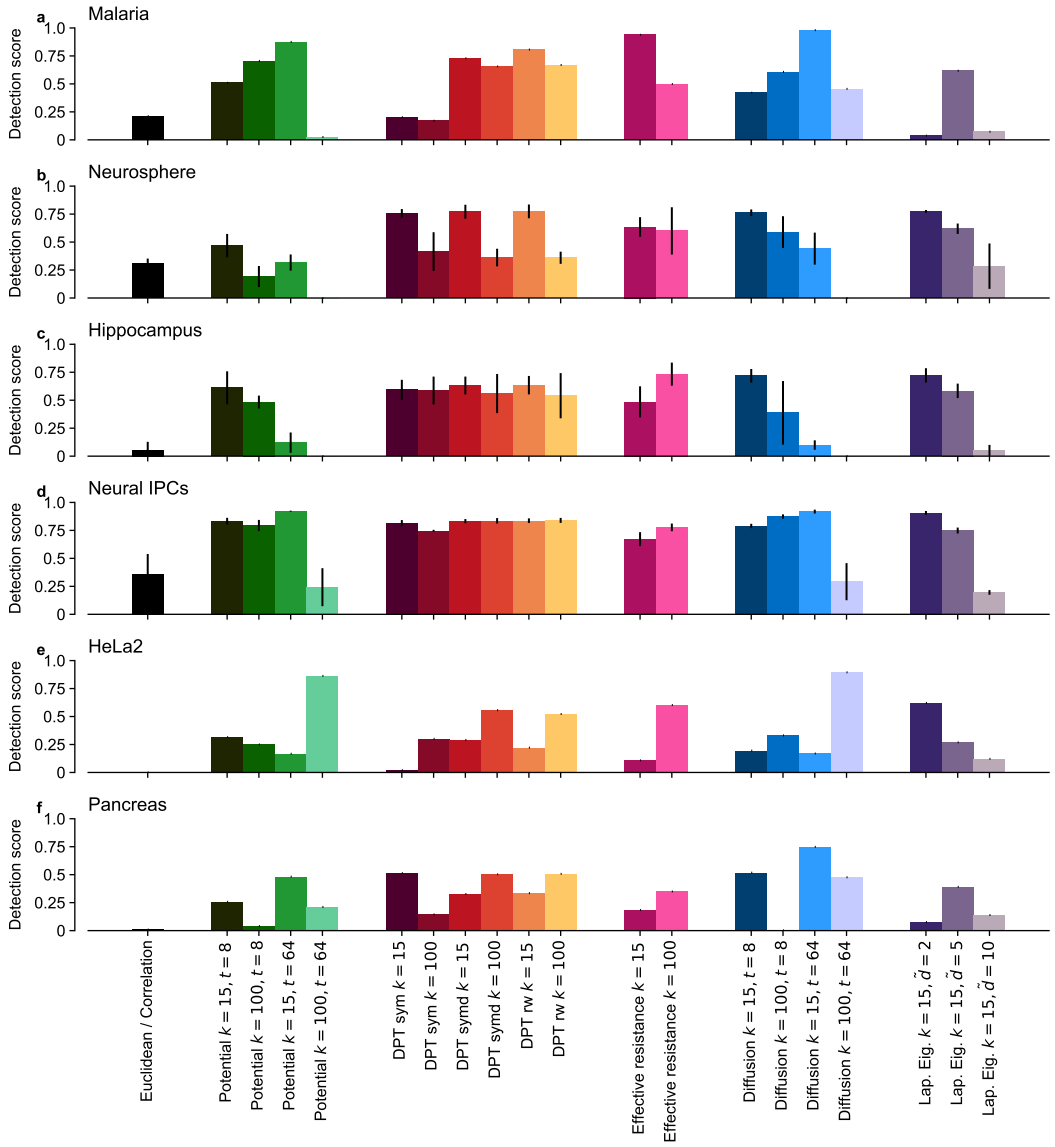

Figure S9: Comparison of all spectral methods on the single cell datasets. They all achieve similar performance.

https://doi.org/10.52202/079017-1328

where θ is the only positive root of $\sqrt[\xi]{\theta^\xi - \text{dtm}_i^\xi} + \sqrt[\xi]{\theta^\xi - \text{dtm}_j^\xi} = d_{ij}$. We only considered the values $\xi \in \{1, 2, \infty\}$, for which the there are closed-form solutions:

$$\theta = \begin{cases} (\text{dtm}_i + \text{dtm}_j + d_{ij})/2 & \text{if } \xi = 1 \\ \sqrt{\big((\text{dtm}_i + \text{dtm}_j)^2 + d_{ij}^2\big) \cdot \big((\text{dtm}_i - \text{dtm}_j)^2 + d_{ij}^2\big)} \Big/ (2d_{ij}) & \text{if } \xi = 2 \\ \max(\text{dtm}_i, \text{dtm}_j, d_{ij}/2) & \text{if } \xi = \infty. \end{cases} \tag{62}$$

We used $k \in \{4, 15, 100\}$, $p \in \{2, \infty\}$, and $\xi \in \{1, 2, \infty\}$.

The original exposition of DTM-based filtrations [2] only considered the setting $p = 2$, while DTM has been defined for arbitrary $p \geq 1$ [14]. We explore an additional value, $p = \infty$, in order to possibly strengthen DTM. Indeed, in several experiments it outperformed the $p = 2$ setting.

Moreover, Anai et al. [2] actually used a small variant of the Vietoris-Rips complex on the above distance $d_{ij}^{\text{DTM}}(k, p, \xi)$: They only included point x_i in the filtered complex once the filtration value exceeds dtm_i. This, however, only affects the 0-th homology, which we do not consider in our experiments.

Core distance The core distance is similar to the DTM distance with $\xi = \infty$ and $p = \infty$ and is given by

$$d_k^{\text{core}}(ij) = \max(d_{ij}, \|x_i - x_{i_k}\|, \|x_j - x_{j_k}\|). \tag{63}$$

We used $k \in \{15, 100\}$.

t-SNE graph affinities The t-SNE affinities are given by

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}, \quad p_{j|i} = \frac{\nu_{j|i}}{\sum_{k \neq i} \nu_{k|i}}, \quad \nu_{j|i} = \begin{cases} \exp\big(\|x_i - x_j\|^2/(2\sigma_i^2)\big) & \text{if } x_j \in N_i \\ 0 & \text{else,} \end{cases} \tag{64}$$

where σ_i is selected such that the distribution $p_{j|i}$ has pre-specified perplexity ρ. Standard implementations of t-SNE use $k = 3\rho$. We transformed t-SNE affinities into pairwise distances by taking the negative logarithm. Pairs x_i and x_j with $p_{ij} = 0$ (i.e. not in the kNN graph) get distance ∞. We used $\rho \in \{30, 200, 333\}$.

UMAP graph affinities The UMAP affinities are given by

$$\mu_{ij} = \mu_{i|j} + \mu_{j|i} - \mu_{i|j}\mu_{j|i}, \quad \mu_{j|i} = \begin{cases} \exp\big(-(d_{ij} - \mu_i)/\sigma_i\big) & \text{for } j \in \{i_1, \ldots, i_k\} \\ 0 & \text{else,} \end{cases} \tag{65}$$

where $\mu_i = \|x_i - x_{i_1}\|$ is the distance between x_i and its nearest non-identical neighbor. The scale parameter σ_i is selected such that

$$\sum_{\kappa=1}^{k} \exp\Big(-\big(d(x_i, x_{i_\kappa}) - \mu_i\big)/\sigma_i\Big) = \log_2(k). \tag{66}$$

As above, to convert these affinities into distances, we take the negative logarithm and handle zero similarities as for the t-SNE case. We used $k \in \{100, 999\}$; $k = 15$ resulted in memory overflow on one of the void-containing datasets.

Gardner et al. [31] and Hermansen et al. [40] first used these distances, but omitted μ_i, which we included to completely reproduce UMAP's affinities.

Note that distances derived from UMAP and t-SNE affinities are not guaranteed to obey the triangle inequality.

Geodesic distances We computed the shortest path distances between all pairs of nodes in the skNN graph with edges weighted by their Euclidean distances. We used the Python function `scipy.sparse.csgraph.shortest_path`. We used $k = \{15, 100\}$.

UMAP embedding We computed the UMAP embeddings in 2 embedding dimensions using 750 optimization epochs, `min_dist` of 0.1, exactly computed k nearest neighbors, and PCA initialization. Then we used Euclidean distances between the embedding points. We used UMAP commit `a7606f2`. We used $k \in \{15, 100, 999\}$.

t-SNE embedding We computed the t-SNE embeddings in 2 embedding dimensions using openTSNE [63] with default parameters, but providing manually computed affinities. For that we used standard Gaussian affinities on the skNN graph with $k = 3\rho$. Then we used the Euclidean distances between the embedding points. We used perplexity $\rho \in \{8, 30, 333\}$.

For UMAP and t-SNE affinities as well as for UMAP and t-SNE embeddings we computed the skNN graph with PyKeOps [13] instead of using the default approximate methods. The UMAP and t-SNE affinities (without negative logarithm) were used by the corresponding embedding methods.

Effective resistance We computed the effective resistance on the skNN graph. Following the analogy with resistances in an electric circuit, if the skNN graph is disconnected, we computed the effective resistance separately in each connected component and set resistances between components to ∞. The uncorrected resistances were computed via the pseudoinverse of the unnormalized graph Laplacian [28]

$$\tilde{d}_{ij}^{\text{eff}} = l_{ii}^{\dagger} - 2l_{ij}^{\dagger} + l_{jj}^{\dagger}, \tag{67}$$

where $l_{ij}^{\dagger}$ is the ij-th entry of the pseudoinverse $L^{\dagger}$ of the unnormalized skNN graph Laplacian $L = D - A$. The pseudoinverse inverts all non-zero eigenvalues. Denote the eigenvalue decomposition of L by $L = V\Lambda V^T$, where $V = (v_1, \dots, v_n)^T$ is the matrix of eigenvectors of L and Λ is the diagonal matrix of their eigenvalues $\lambda_1, \dots, \lambda_n$. Then $L^{\dagger} = V\Lambda^{\dagger}V^T$, where $\Lambda^{\dagger}$ is the diagonal matrix with entry $1/\lambda_i$ if $\lambda_i > 0$ and 0 otherwise. We can use this to derive a similar coordinate expression as in Eq. 5, but based on the unnormalized graph Laplacian ([28, 4.B]). Let e_i be the i-th standard basis vector

$$\begin{aligned}
\tilde{d}_{ij}^{\text{eff}} &= l_{ii}^{\dagger} - 2l_{ij}^{\dagger} + l_{jj}^{\dagger} \\
&= (e_i - e_j)^T L^{\dagger}(e_i - e_j) \\
&= (e_i - e_j)^T V\Lambda^{\dagger}V^T(e_i - e_j) \\
&= \left(\sqrt{\Lambda^{\dagger}}V^T e_i - \sqrt{\Lambda^{\dagger}}V^T e_j\right)^T \left(\sqrt{\Lambda^{\dagger}}V^T e_i - \sqrt{\Lambda^{\dagger}}V^T e_j\right) \\
&= \|\hat{e}_i^{\text{eff}} - \hat{e}_j^{\text{eff}}\|^2 \text{ where} \\
\hat{e}_i^{\text{eff}} &= \left(\frac{v_{2,i}}{\sqrt{\lambda_i}}, \dots, \frac{v_{n,i}}{\sqrt{\lambda_n}}\right).
\end{aligned} \tag{68}$$

For the corrected version, we used

$$d_{ij}^{\text{eff}} = \tilde{d}_{ij}^{\text{eff}} - \frac{1}{d_i} - \frac{1}{d_j} + 2\frac{a_{ij}}{d_i d_j} - \frac{a_{ii}}{d_i^2} - \frac{a_{jj}}{d_j^2}. \tag{69}$$

For the weighted version of effective resistance, each edge in the skNN graph was weighted by the inverse of the Euclidean distance. We experimented with the weighted and unweighted versions, but only reported the unweighted version in the paper as the difference was always minor. We also experimented with the unweighted and uncorrected version and saw that correcting is crucial for high noise levels (Figure S20). We used $k \in \{15, 100\}$.

Both forms of the effective resistance can be written as squared distances between certain embedding points (5), (6). Nevertheless, the uncorrected effective resistance is a proper metric [32, Corollary 2.4.]. The corrected version in general is not a proper metric (Proposition F.3).

Diffusion distance We computed the diffusion distances on the unweighted skNN graph directly by equation (2), i.e.,

$$d_t^{\text{diff}}(i, j) = \sqrt{\text{vol}(G)}\|(P_{i,:}^t - P_{j,:}^t)D^{-\frac{1}{2}}\|. \tag{70}$$

Note that our skNN graphs do not contain self-loops. We used $k \in \{15, 100\}$ and $t \in \{8, 64\}$.

It is clear from the above definition of the diffusion distance that it satisfies the triangle inequality, but it can fail to be positive on distinct points (Proposition F.3).

Laplacian Eigenmaps For an skNN graph with K connected components, we computed the $K + \tilde{d}$ eigenvectors $u_1, \dots, u_{K+\tilde{d}}$ of the normalized graph Laplacian L^{sym} of the skNN graph and discarded

https://doi.org/10.52202/079017-1328

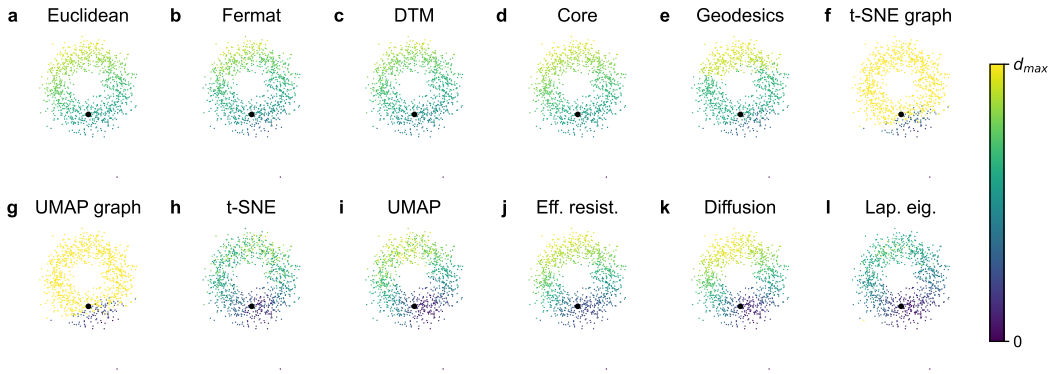

Figure S10: Visualization of all distances on the noisy circle in $\mathbb{R}^{50}$ with $\sigma = 0.25$. All scatter plots are the 2D PCA of the 50D dataset. The colors indicate the distance to the highlighted point.

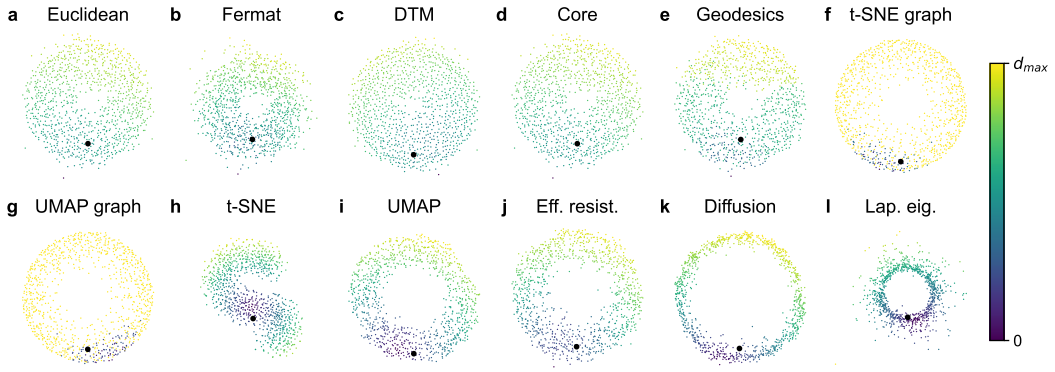

Figure S11: Visualization of all distances on the noisy circle in $\mathbb{R}^{50}$ with $\sigma = 0.25$. The scatter plots are 2D multidimensional scaling embeddings using the respective distances. The colors indicate the distance to the highlighted point. For this random seed the t-SNE embedding tore the circle apart.

the first K eigenvectors $u_1, \dots, u_K$, which are coding for the connected components. Then we computed the Euclidean distances between the embedding vectors $e_i^{\text{LE}} = (u_{K+1,i}, \dots, u_{(K+\tilde{d}),i})$. We used $k = 15$ and embedding dimensions $\tilde{d} \in \{2, 5, 10\}$.

Alternatively, one can compute Laplacian Eigenmaps using the un-normalized graph Laplacian L. We tried this normalization for $\tilde{d} = 2$ but obtained very similar embeddings.

Diffusion pseudotime Diffusion pseudotime [36] also integrates the different time scales of the diffusion distance, but on the level of the transition matrices, rather than on the level of the distances themselves (Proposition G.2). There are multiple variants of diffusion pseudotime. They are all computed as $d_{ij}^{\text{dpt}} = \|e_i^{\text{dpt}} - e_j^{\text{dpt}}\|$ for

$$e_i^{\text{dpt}} = \left(\frac{1-\mu_2}{\mu_2} v_{2,i}, \dots, \frac{1-\mu_n}{\mu_n} v_{n,i} \right). \tag{71}$$

The difference is the $v_1, \dots, v_n$'s. In the original publication [36], they were the normalized eigenvectors of the random walker graph Laplacian $D^{-\frac{1}{2}} L^{\text{sym}} D^{\frac{1}{2}}$. These are given by $v_i = D^{-\frac{1}{2}} u_i / \|D^{-\frac{1}{2}} u_i\|$. Wolf et al. [88] introduced a version using the eigenvectors of the symmetric graph Laplacian L^{sym}, so that $v_i = u_i$. Closest to the corrected resistance distance is the case where $v_i = D^{-\frac{1}{2}} u_i$ are non-normalized. We tested all three versions, to which we refer as "rw", "sym" and "symd" version. We used $k \in \{15, 100\}$ nearest neighbors.

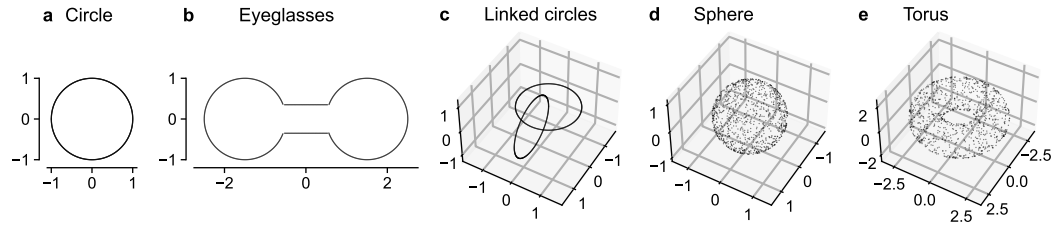

Figure S12: Synthetic, noiseless datasets with $n = 1\,000$ points each.

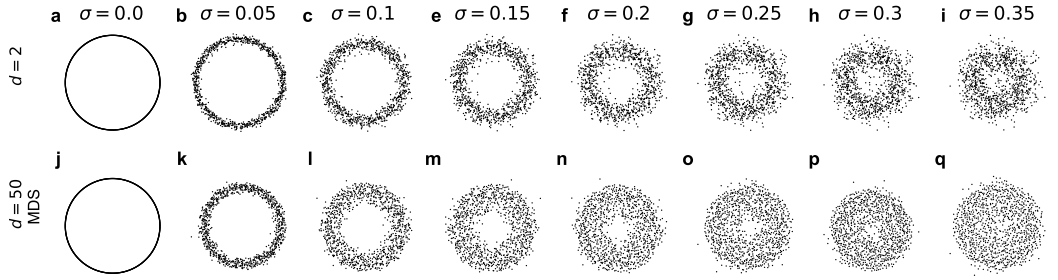

Figure S13: Circle with Gaussian noise of different standard deviation σ. **a – i.** Original data in ambient dimension $d = 2$. **j – q.** Multidimensional scaling of the Euclidean distance of the data in ambient dimension $d = 50$.

Potential distance The potential distance underlies the visualization method PHATE [56] and is closely related to the diffusion distance. It is defined by

$$d_t^{\text{pot}}(i, j) = \| \log(P_{i,:}^t) - \log(P_{j,:}^t) \|,$$

where the logarithm is applied element-wise. We used $k \in \{15, 100\}$ and $t \in \{8, 64\}$.

For all methods, we replaced infinite distances with twice the maximal finite distance to be able to compute our hole detection scores.

We illustrate the distances used in the main benchmark in Figures S10 and S11.

J Datasets

J.1 Synthetic datasets

The synthetic, noiseless datasets with $n = 1\,000$ points each are depicted in Figure S12. Noised versions of the circle for ambient dimensions $d = 2, 50$ are depicted in Figure S13.

Circle The circle dataset consists of n points equidistantly spaced along a circle of radius $r = 1$.

Linked circles The linked circles dataset consists of two circle datasets of $n/2$ points each, arranged such that each circle perpendicularly intersects the plane spanned by the other and goes through the other's center.

Eyeglasses The eyeglasses dataset consists of four parts: Two circle segments of arclength $\pi + 2.4$ and radius $r = 1$, centered 3 units apart with the gaps facing each other. The third and fourth part are two straight line segments of length 1.06, separated by 0.7 units linking up the two circle segments. The circle segments consist of $0.425n$ equidistantly distributed points each and the line segments consist of $0.075n$ equispaced points each. As the length scale of this dataset is dominated by the bottleneck between the two line segments, we only considered noise levels $\sigma \in [0, 0.15]$ for this dataset, as at this point the bottleneck essentially merges in $\mathbb{R}^2$.

https://doi.org/10.52202/079017-1328

Sphere The sphere dataset consists of n points sampled uniformly from a sphere S^2 with radius $r = 1$.

Torus The torus dataset consists of n points sampled uniformly from a torus. The radius of the torus' tube was $r = 1$ and the radius of the center of the tube was $R = 2$. Note that we do not sample the points to have uniform angle distribution along the tube's and the tube center's circle, but uniform on the surface of the torus.

High-dimensional noise We mapped each dataset to $\mathbb{R}^d$ for $d \in [2, 50]$ using a random matrix $\mathbf{V}$ of size $d \times 2$ or $d \times 3$ with orthonormal columns, and then added isotropic Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ for $\sigma \in [0, 0.35]$.

The orthogonal embedding in $\mathbb{R}^d$ does not change the shape of the data. The procedure is equivalent to adding $d - 2$ or $d - 3$ zero dimensions and then randomly rotating the resulting dataset in $\mathbb{R}^d$.

J.2 Single-cell datasets

We depict 2D embeddings of all single-cell datasets in Figure S14.

Malaria The Malaria dataset [43] consists of gene expression measurement of $5\,156$ genes obtained with the modified SmartSeq2 approach of Reid et al. [65] in $n = 1\,787$ cells from the entire life cycle of *Plasmodium berghei*. The resulting transcripts were pre-processed with the trimmed mean of M-values method [69]. We obtained the pre-processed data from `https://github.com/vhowick/MalariaCellAtlas/raw/v1.0/Expression_Matrices/Smartseq2/SS2_tmmlogcounts.csv.zip`. The data is licensed under the GNU GPLv3 licence.

The UMAP embedding shown in Figure 9 follows the authors' setup and uses correlation distance as input metric, $k = 10$ nearest neighbors, and a `min_dist` of 1 and `spread` of 2. Note that when computing persistent homology with UMAP-related distances, we used our normal UMAP hyperparameters and never changed `min_dist` or `spread`.

Neural IPCs The Neural IPC dataset [8] consists of gene expressions of $n = 26\,625$ neural IPCs from the developing human cortex. scVI [50] was used to integrate cells with different ages and donors based on the 700 most highly variable genes, resulting in a $d = 10$ dimensional embedding. Braun et al. [8] shared this representation with us for a superset of $297\,927$ telencephalic exitatory cells and allowed us to share it with this paper (MIT License). We limited our analysis to the neural IPCs because they formed a particularly prominent cell cycle.

Neurosphere The Neurosphere dataset [89] consists of gene expressions for $n = 12\,805$ cells from the mouse neurosphere. After quality control, the data was library size normalized and $\log_2$ transformed. Seurat was used to integrate different samples based on the first 30 PCs of the top $2\,000$ highly variable genes, resulting in a $12\,805 \times 2\,000$ matrix of $\log_2$ transformed expressions. These were subsetted to the genes in the gene ontology (GO) term cell cycle (GO:0007049). The 500 most highly variable genes were selected and a PCA was computed to $d = 20$. The GO PCA representation was downloaded from `https://zenodo.org/record/5519841/files/neurosphere.qs`. It is licensed under CC BY 4.0.

Hippocampus The Hippocampus dataset [89] consists of gene expressions for $n = 9\,188$ mouse hippocampal NPCs. The pre-processing was the same as for the Neurosphere dataset. The GO PCA representation was downloaded from `https://zenodo.org/record/5519841/files/hipp.qs`. It is licensed under CC BY 4.0.

HeLa2 The HeLa2 dataset [72, 89] consists of gene expressions for $2\,463$ cells from a human cell line derived from cervical cancer. After quality control, the data was library size normalized and $\log_2$ transformed. From here the GO PCA computation was the same as for the neurosphere dataset. The GO PCA representation was downloaded from `https://zenodo.org/record/5519841/files/HeLa2.qs`. It is licensed under CC BY 4.0.

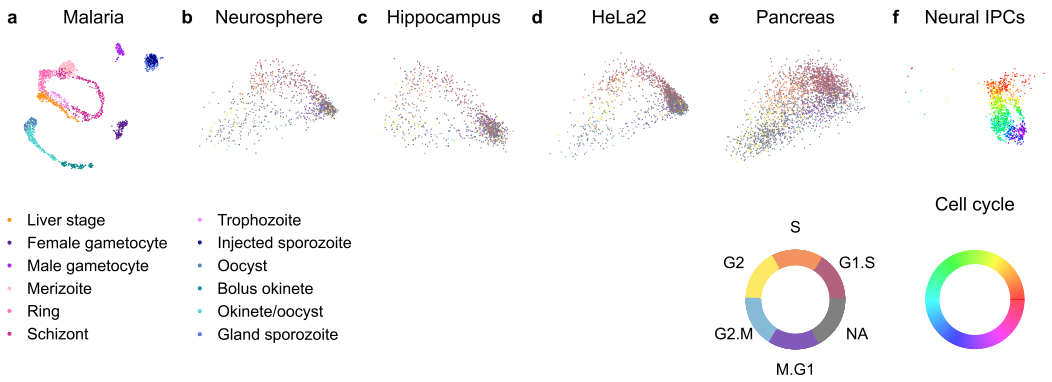

Figure S14: 2D embeddings of all six single-cell datasets. **a, f.** UMAP embeddings of the Malaria [43] and the Neural IPC datasets [8]. We recomputed the embedding for the Malaria dataset using UMAP hyperparameters provided in the original publication, and subsetted an author-provided UMAP of a superset of telencephalic exitatory cells to the Neural IPC. The text legend refers to Malaria cell types. **b – e.** 2D linear projection constructed to bring out the cell cycle ('tricycle embedding') [89] of the Neurosphere, Hippocampus, HeLa2, and Pancreas datasets. We used the projection coordinates provided by Zheng et al. [89].

Pancreas The Pancreas dataset [3, 89] consists of gene expressions for 3 559 cells from the mouse endocrine pancreas. After quality control, the data was library size normalized and $\log_2$ transformed. From here the GO PCA computation was the same as for the neurosphere dataset. The GO PCA representation was downloaded from `https://zenodo.org/record/5519841/files/endo.qs`. It is licensed under CC BY 4.0.

K Hyperparameter selection

For each of the datasets and hole dimensions, we showed the result with the best hyperparameter setting. For the synthetic experiments, this meant the highest area under the hole detection score curve, while for the single-cell datasets it meant the highest loop detection score. Here, we give details of the selected hyperparamters.

For Figure 1 we used effective resistance with $k = 100$ as in Figure 6.

For Figure 6 we specified the selected hyperparameters directly in the figure. For the density-based methods, they were $p = 3$ for Fermat distances, $k = 4, p = 2, \xi = \infty$ for DTM, and $k = 15$ for the core distance. For the graph-based methods, they were $k = 100$ for the geodesics, $k = 100$ for the UMAP graph affinities, and $\rho = 30$ for t-SNE graph affinities. The embedding-based methods used $k = 15$ for UMAP and $\rho = 30$ for t-SNE. Finally, as spectral methods, we selected effective resistance with $k = 100$, diffusion distance with $k = 100, t = 8$ and Laplacian Eigenmaps with $k = 15, d = 2$.

The hyperparameters for Figure 7 are given in Table S1.

In Figure 8 we specified the hyperparameters used. They were the same as for Figure 6 save for the diffusion distance, for which we used $k = 100, t = 8$. This setting had better performance for $d = 2$ and only marginally lower performance than $k = 15, t = 8$ in higher dimensionalities.

For Figure 9, we selected DTM with $k = 15, p = \infty, \xi = \infty$, effective resistance with $k = 15$ and diffusion distance with $k = 15, t = 64$. They are the same for the Malaria dataset in Figure 10.

The selected hyperparameters for Figure 10 can be found in Table S2.

The hyperparameters for Figures S10 and S11 are the same as those used in Figure 6. The hyperparameters for Figure S3 are given in Table S3 and those for Figure S16 in Table S4. All other supplementary figures either do not depend on hyperparameters or detail them directly in the figure or the caption.

https://doi.org/10.52202/079017-1328

Table S1: The optimal hyperparameters that were selected in Figure 7. For torus and sphere, we consider the case of loop detection (H_1) and void detection (H_2) separately.

Dataset	Fermat	DTM	Eff. res.	Diffusion
Circle	$p = 3$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 15, t = 8$
Eyeglasses	$p = 7$	$k = 100, p = 2, \xi = 1$	$k = 15$	$k = 15, t = 64$
Linked circles	$p = 7$	$k = 15, p = \infty, \xi = 1$	$k = 15$	$k = 15, t = 8$
Torus H_1	$p = 2$	$k = 4, p = 2, \xi = \infty$	$k = 100$	$k = 15, t = 8$
Sphere H_1	$p = 2$	$k = 100, p = 2, \xi = 1$	$k = 15$	$k = 100, t = 64$
Torus H_2	$p = 2$	$k = 4, p = 2, \xi = \infty$	$k = 100$	$k = 15, t = 8$
Sphere H_2	$p = 2$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 100, t = 8$

Table S2: The optimal hyperparameters that were selected in Figure 10. For DTM we report the best setting without thresholding (because none of the DTM runs passed our birth/death thresholding, so all s_m scores for all parameter combinations are zero).

Dataset	Fermat	DTM	t-SNE	UMAP	Eff. res.	Diffusion	Lap. Eig.
Malaria	$p = 2$	$k = 15$ $p = \infty$ $\xi = \infty$	$\rho = 8$	$k = 15$	$k = 15$	$k = 15$ $t = 64$	$\tilde{d} = 5$
Neurosphere	$p = 2$	$k = 100$ $p = 2$ $\xi = 2$	$\rho = 30$	$k = 999$	$k = 15$	$k = 15$ $t = 8$	$\tilde{d} = 2$
Hippocampus	$p = 7$	$k = 100$ $p = 2$ $\xi = 2$	$\rho = 8$	$k = 15$	$k = 100$	$k = 15$ $t = 8$	$\tilde{d} = 2$
Neural IPC	$p = 2$	$k = 4$ $p = \infty$ $\xi = \infty$	$\rho = 30$	$k = 15$	$k = 100$	$k = 15$ $t = 64$	$\tilde{d} = 2$
HeLa2	$p = 3$	$k = 4$ $p = 2$ $\xi = \infty$	$\rho = 30$	$k = 100$	$k = 100$	$k = 100$ $t = 64$	$\tilde{d} = 2$
Pancreas	$p = 7$	$k = 100$ $p = 2$ $\xi = \infty$	$\rho = 8$	$k = 100$	$k = 4$	$k = 15$ $t = 64$	$\tilde{d} = 5$

L Hyperparameter sensitivity

While all distances other than the Euclidean distance have hyperparameters and we show results for the best hyperparameter setting in most figures, hyperparameter selection does not pose a serious problem for diffusion distance and effective resistance. We only tuned hyperparameters very mildly for these methods (4 settings for diffusion distances, only 2 for effective resistance, as opposed to 24 for the competing method DTM). Moreover, we conduced a more fine-grained sensitivity analysis and found the performance of diffusion distances and effective resistances to be robust to the exact value of their hyperparameters. For this sensitivity analysis we computed the area under the noise-level / detection score curves for the circle, the interlinked circles and the eyeglasses dataset in 50 ambient dimensions with $k = 4, 15, 30, 45, 60, 75, 90, 105, 120, 135, 150$ nearest neighbors and $t = 2, 4, 8, 16, 32, 64, 128, 256$ diffusion steps. For all but the most extreme hyperparameter values both effective resistance and diffusion distances strongly outperformed the Euclidean distance (Figure S15). We observed a trade-off between the optimal number of neighbors k and diffusion steps t for the diffusion distance, since both control how fast the diffusion can progress across the dataset.

Table S3: The optimal hyperparameters that were selected in Figure S3.

Dataset	Fermat	DTM	Eff. res.	Diffusion
Circle	$p = 2$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 15, t = 8$
Eyeglasses	$p = 7$	$k = 4, p = 2, \xi = 1$	$k = 15$	$k = 15, t = 8$
Linked circles	$p = 5$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 15, t = 8$

Table S4: The optimal hyperparameters that were selected in Figure S16.

Ambient dimension	Number of outliers	Fermat	DTM	Effective resistance	Diffusion
2	0	$p = 2$	$k = 100, p = \infty, \xi = 2$	$k = 100$	$k = 100, t = 64$
2	50	$p = 7$	$k = 100, p = \infty, \xi = 2$	$k = 100$	$k = 100, t = 64$
2	100	$p = 7$	$k = 100, p = \infty, \xi = 2$	$k = 100$	$k = 100, t = 64$
2	200	$p = 7$	$k = 100, p = \infty, \xi = 2$	$k = 100$	$k = 100, t = 64$
50	0	$p = 3$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 100, t = 8$
50	50	$p = 2$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 15, t = 8$
50	100	$p = 2$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 15, t = 8$
50	200	$p = 2$	$k = 4, p = 2, \xi = 1$	$k = 100$	$k = 15, t = 64$

M Implementation details

We computed persistent homology using the `ripser` [4] project's `representative-cycles` branch at commit `140670f` to compute persistent homologies and representative cycles. We used coefficients in $\mathbb{Z}/2\mathbb{Z}$. To compute kNN graphs, we used the PyKeops package [13]. The rest of our implementation is in Python. Our code is available at `https://github.com/berenslab/eff-ph/tree/neurips2024`.

Our experiments were run on a machine with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz with 64 kernels, 377GB memory, and an NVIDIA RTX A6000 GPU. The persistent homology computations only ever used a single kernel.

Our benchmark consisted of many individual experiments. We explored 47 hyperparameter settings across all distances, computed results for 3 random seeds and 29 noise levels σ. In the synthetic

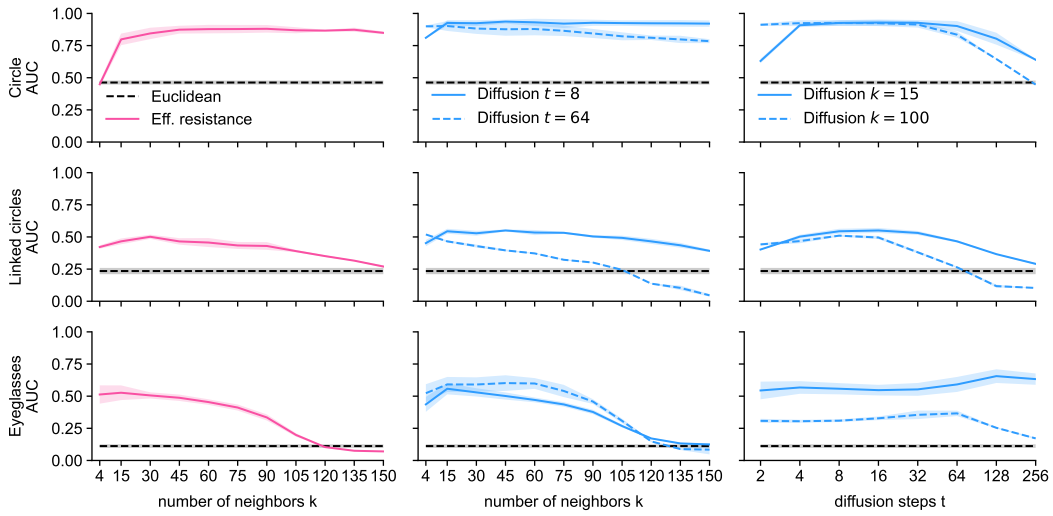

Figure S15: Diffusion distances and effective resistance are robust to their hyperparameters and outperform the Euclidean distance for most choices. We depict the area under the noise-level / detection score curve for the three 1D datasets circle, linked circles and eyeglasses in $\mathbb{R}^{50}$.

https://doi.org/10.52202/079017-1328

Table S5: Exemplary run times in seconds.

Dataset	n	σ	Distance	Feature dim	Time distance [s]	Time PH [s]
Circle	1 000	0.0	Euclidean	1	0.013 ± 0.002	12.3 ± 0.4
Circle	1 000	0.0	Eff. res $k = 100$	1	0.17 ± 0.04	12.0 ± 0.2
Circle	2 000	0.0	Euclidean	1	0.09 ± 0.04	117 ± 9
Sphere	1 000	0.0	Euclidean	1	0.012 ± 0.001	1.31 ± 0.06
Sphere	1 000	0.0	Euclidean	2	0.017 ± 0.002	4687 ± 2501
Circle	1 000	0.35	Euclidean	1	0.016 ± 0.001	5 ± 2
Sphere	1 000	0.35	Euclidean	2	0.03 ± 0.02	258 ± 18

benchmark, we computed only 1D persistent homology for 3 datasets and both 1D and 2D persistent homology of 2 more datasets. So the synthetic benchmark with ambient dimension $d = 50$ alone consisted of 12 267 computations of 1D persistent homology and 8 178 computations of both 1D and 2D persistent homology.

The run time of persistent homology vastly dominated the time taken by the distance computation. The persistent homology run time depended most strongly on the sample size n, the dataset, and on the highest dimensionality of holes. The difference between distances was usually small. However, we observed that there were some outliers, depending on the noise level and the random seed, that had much longer run time. Overall, we found that methods that produce many pairwise distances of the same value (e.g., because of infinite distance in the graph affinities or maximum operations like for DTM with $p = \infty, \xi = \infty$) often had a much longer run time than other settings. We presume this was because equal distances led to many simplices being added to the complex at the same time. We give exemplary run times in Table S5.

As a rough estimate for the total run time, we extrapolated the run times for the circle to all 1D persistent homology experiments for ambient dimension $d = 50$ and the times for the sphere to all 2D experiments. In both cases we took the mean between the noiseless ($\sigma = 0$) and highest noise ($\sigma = 0.35$) setting in Table S5. This way, we estimated a total sequential run time of about 60 days, but we parallelized the runs.

N Effect of outliers

Persistent homology with the Euclidean distance is known to be sensitive to outliers. Methods such as DTM were introduced to handle this issue. Here we show that spectral methods can also handle outliers well. Moreover, in high ambient dimensionality outliers are distributed over the large volume and hence are very sparse, making them less of a problem.

We experimented with the noisy circle with $n = 1\,000$ points in ambient $\mathbb{R}^d$ for $d = 2, 50$ and added 50, 100, or 200 outlier points. These were sampled uniformly from axis-aligned cubes around the data in ambient space. The size of the cube was set just large enough that it contained the data even with the strongest added Gaussian noise.

In low dimensionality, Euclidean distance suffered in the low Gaussian noise setting already when only 50 outliers were added. Adding 100 or 200 outliers severely lowered the detection score for Euclidean distance across the entire range of Gaussian noise strength. Fermat distances suffered from high random seed variability when adding outliers in low ambient dimension. Diffusion distance and effective resistance were much more outlier-resistant than the Euclidean distance and were only affected by 200 outliers. Even then they performed better than the Euclidean distance without any outliers. DTM excelled in this setting, being completely insensitive to outliers and achieving top score for all noise levels (Figure S16a – d).

The volume of the bounding box in $d = 50$ ambient dimensions is much larger and thus the same number of outlier points are distributed much more sparsely. In particular, it is much less likely that an outlier happens to fall into the middle of the circle. As a result, even Euclidean and Fermat distances were very outlier-robust in $d = 50$ ambient dimensions (Figure S16e – g). Similarly, DTM's performance did not change at all in the face of outliers. However, all three methods suffered strongly from the high-dimensional Gaussian noise.

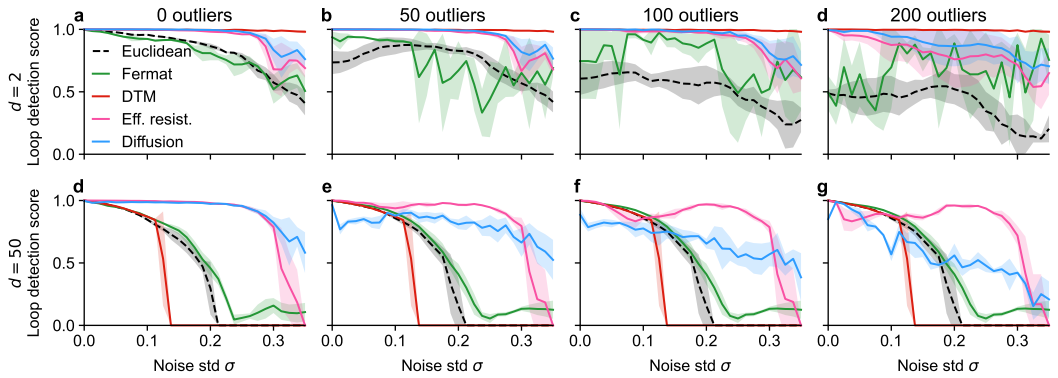

Figure S16: Loop detection performance of various methods on the noisy circle in the presence of outliers in low- and high-dimensional ambient space. Outliers were sampled uniformly from an axis-aligned cube around the data. **a – d.** In low ambient dimension ($d = 2$) adding outliers hurt the performance of the Euclidean and Fermat distances, but barely affected the performance of the spectral methods and not at all DTM's excellent performance. **e – g.** In high ambient dimensionality ($d = 50$) outliers did not further decrease the weak performance of non-spectral methods. Diffusion distance was somewhat outlier-sensitive, but could still detect the loop structure in the high Gaussian noise setting. Effective resistance performed best overall and was not very outlier-sensitive.

As diffusion distance and effective resistance in our implementation rely on the unweighted kNN graph, they were somewhat more susceptible to outliers. Performance of diffusion distance decreased steadily with the number of outliers in high ambient dimension. When outliers were present, it performed worse than the Euclidean distance in the low Gaussian noise setting, but much better in the high Gaussian noise setting, even for 200 outliers. Effective resistance performed best overall, deteriorating only slightly in the low Gaussian noise setting when outliers were added. Both spectral methods clearly outperformed other methods in the high-Gaussian noise regime even in the presence of numerous outliers.

To sum up, effective resistance (and to a lesser extent diffusion distance) can handle both outliers and high-dimensional Gaussian noise, while other methods can handle at most one type of noise.

O UMAP with higher embedding dimension

Following the original publication [53], UMAP is typically used to embed data into two dimensions. This is an obvious issue for datasets sampled from manifolds which are not embeddable into two dimensions, such as the sphere or the torus. Therefore, we also experimented with the less common approach of higher embedding dimensionality (3, 5, 10) for UMAP. Save for the successful detection of the sphere's void with at least three embedding dimensions, we observed few consistent effects of the embedding dimension either on the toy (Figure S17) or on the real data (Figure S19). Nevertheless, on the linked circles dataset, which is not embeddable into two dimensions, we saw a small improvement for UMAP when going beyond two embedding dimensions for the low noise setting.

Independent of the embedding dimension, UMAP struggled in the low-noise setting on the eyeglasses dataset for $k = 15$. Moreover, we observed very poor performance both for loop and void detection on the torus. We believe the reasons may be UMAP's tendency to over-fragment the manifold [22] and UMAP's use of a heavy-tailed kernel. We visualized three-dimensional UMAP embeddings of the noiseless eyeglasses, sphere, and torus for $k = 15$ in Figure S18. UMAP left a gap in the embedding of the eyeglasses dataset, so that the true loop only had the second highest persistence. Short-cutting at the bottleneck yielded the most persistent loop. For the torus, the surface of the embedding was very fragmented, preventing the detection of any void. While the main loop of the torus was detected well, the second most persistent detected loop was already in the noise cloud as the fragmented surface of the embedding allowed for many fairly persistent loops. In a similar way, the surface of the sphere's embedding got fragmented, leading to many loops. Our score does not penalize this, because the detection score for $m = 1$ loop is low as there is nearly no gap in persistence between the most and the second-most persistent loop. Higher levels of noise and also

https://doi.org/10.52202/079017-1328

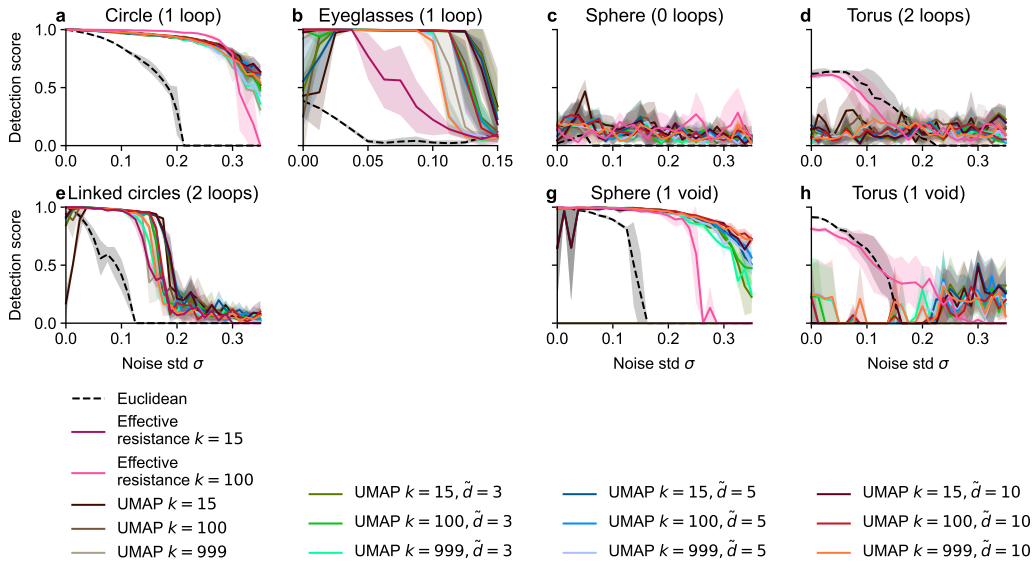

Figure S17: Hole detection scores in UMAP embeddings of the toy datasets in different embedding dimensions. Changing the embedding dimension only has a small effect.

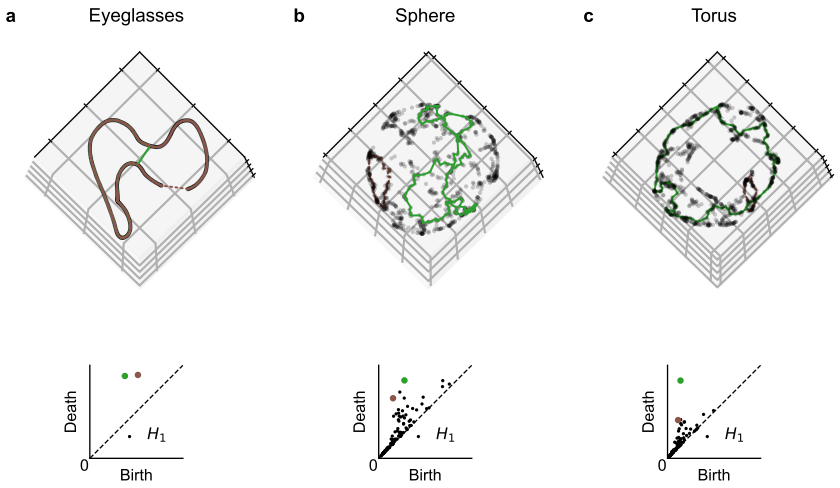

Figure S18: Exemplary UMAP embeddings with $k = 15$ of the noiseless eyeglasses dataset, the sphere, and the torus into three dimensions. We show the persistence diagrams for loops and highlight the two most persistent loops and superimposed them on the embedding. We see strong over-fragmentation of the surfaces that challenges the loop detection and in the case of the torus the void detection (persistence diagram for the voids not shown).

higher k could overcome UMAP's over-fragmentation tendency for the eyeglasses dataset. This may be due to high-dimensional noise better matching UMAP's heavy-tailed kernel. However, it did not resolve the over-fragmentation for the torus.

The runtime of UMAP scales linearly with the embedding dimension, but typical t-SNE implementations scale exponentially. In fact, the implementation we used here, openTSNE [63], does not implement embedding dimensions higher than two, which is why we explored higher-dimensional embeddings only for UMAP.

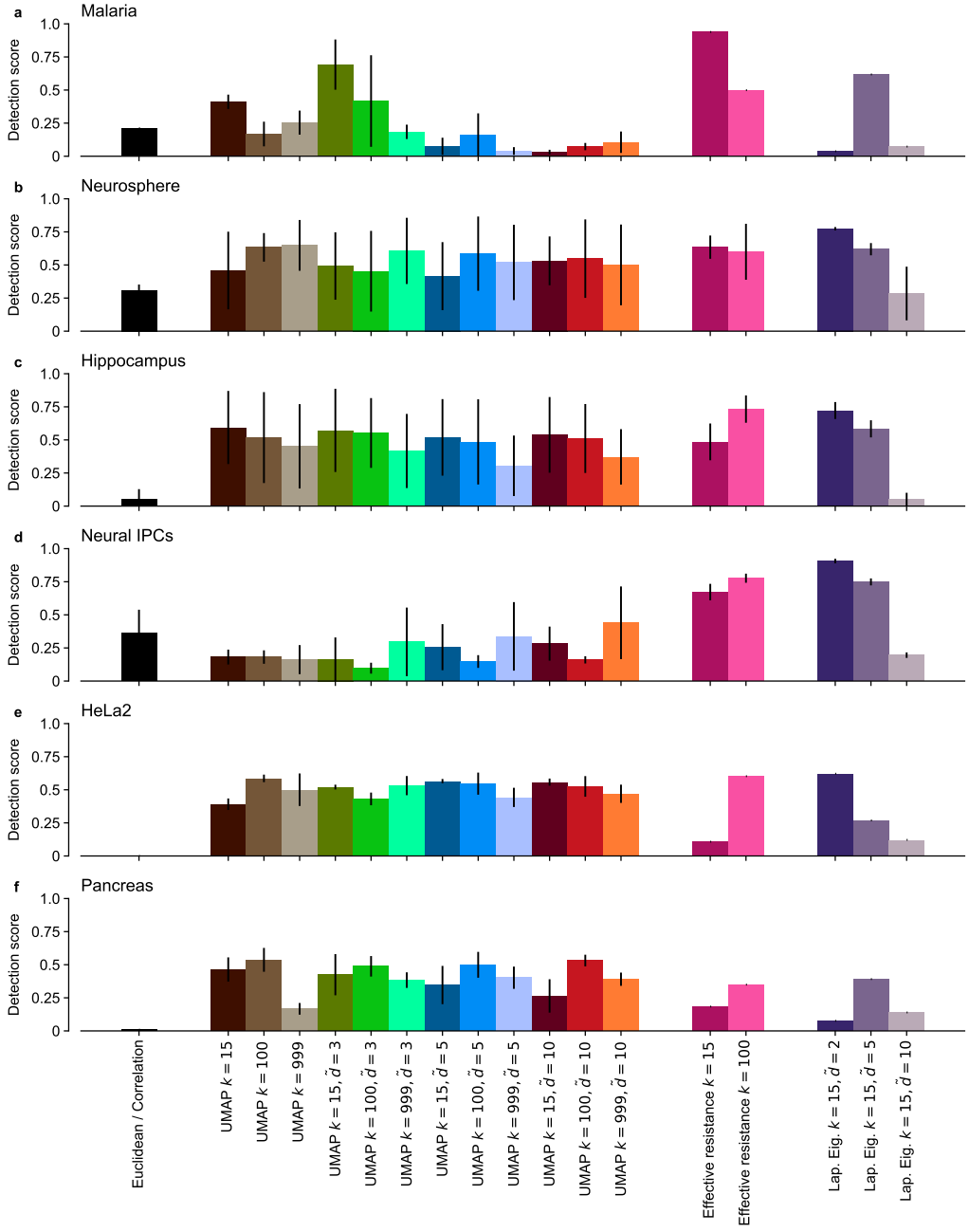

Figure S19: Detection scores for UMAP with different embedding dimensions on the single-cell datasets with some other methods for reference. Changing the embedding dimension did not have a consistent effect on the detection scores, while higher embedding dimension usually hurt for Laplacian Eigenmaps.

https://doi.org/10.52202/079017-1328

P Additional figures

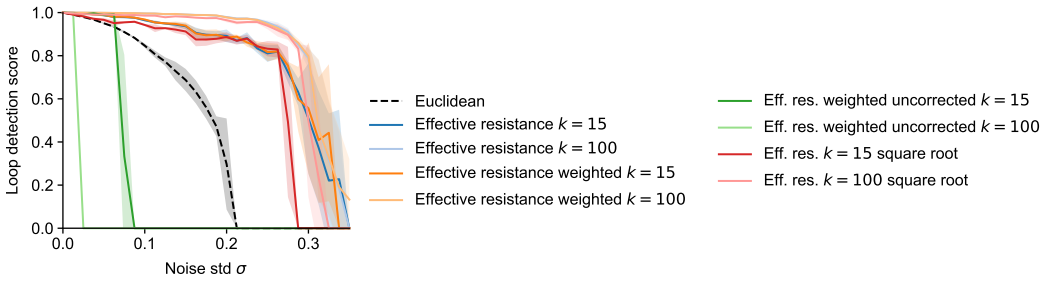

Figure S20: Loop detection score on noisy $S^1 \subset \mathbb{R}^{50}$ for various versions of effective resistance. There was little difference between using the weighted kNN graph, unweighted kNN graph, and using the square root of effective resistance based on the unweighted kNN graph. The latter got filtered out for high noise levels. Using $k = 100$ instead of $k = 15$ helped only marginally in this dataset. The uncorrected (naive) version of effective resistance collapsed already at very small noise levels.

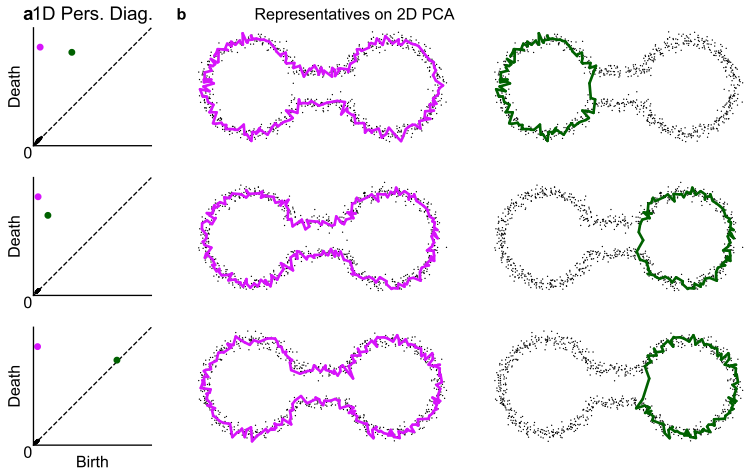

Figure S21: Illustration for the random seed variability of effective resistance with $k = 15$ on the noisy eyeglasses dataset in $\mathbb{R}^{50}$ with $\sigma = 0.075$. This refers to Figure 7b. **a.** One-dimensional persistence diagrams for three random seeds. **b.** Representatives of the most two most persistent features superimposed on a 2D PCA of the dataset. These always corresponded to the full shape and one of the two circle segments. For the first two random seeds, some points are distorted in such a way that they form a bridge in the 2D PCA, while in the third there is not such bridge and the second most persistent feature is much less persistent. Note that this is just a 2D PCA, in particular, much of the noise in 50D is not visible. A similar explanation applies for the diffusion distance in Figure 7b.

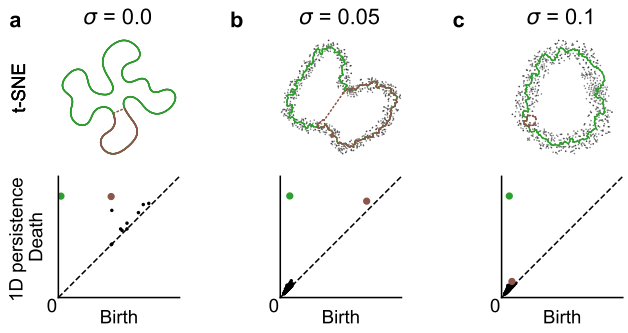

Figure S22: t-SNE embeddings with perplexity $\rho = 8$ and 1D persistence diagrams of the embedding for a circle in ambient $\mathbb{R}^{50}$ with Gaussian noise of low standard deviation σ. Perplexity $\rho = 8$ is rather small, such that each embedding point only feels attraction to very few other points. In the noiseless setting this very sparse attraction is only among immediate neighbors along the circle. This makes the embedding to have spurious curves. For higher noise, the sparse attraction pattern is less regular and less local such that the spurious curves disappear. The more spurious curves the embedding has, the more high persistent features, given by bottlenecks in the curvy embedding, exist. This explains the dip for the t-SNE $\rho = 8$ curve in Figure S23g.

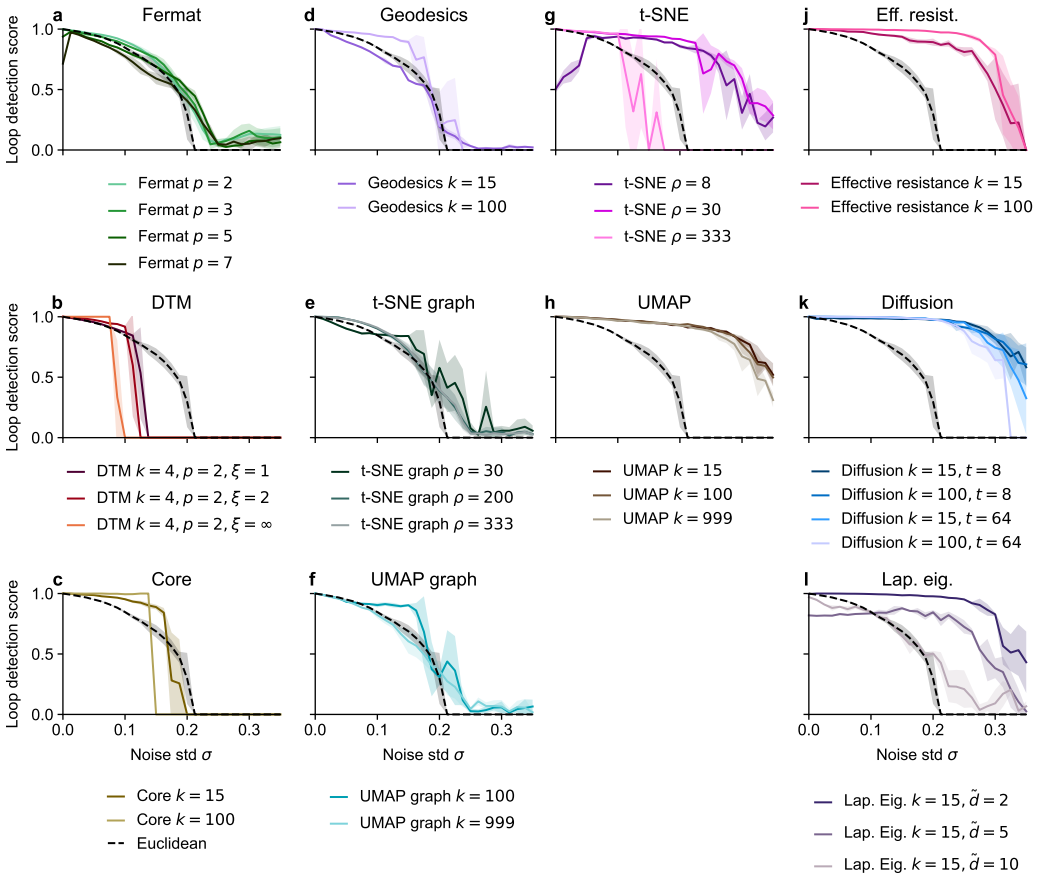

Figure S23: Loop detection score for persistent homology with various distances on a noisy circle in ambient $\mathbb{R}^{50}$. Extension of Figure 6. Spectral and embedding methods performed best. The reason for the dip for the low-perplexity t-SNE embedding is depicted in Figure S22.

https://doi.org/10.52202/079017-1328

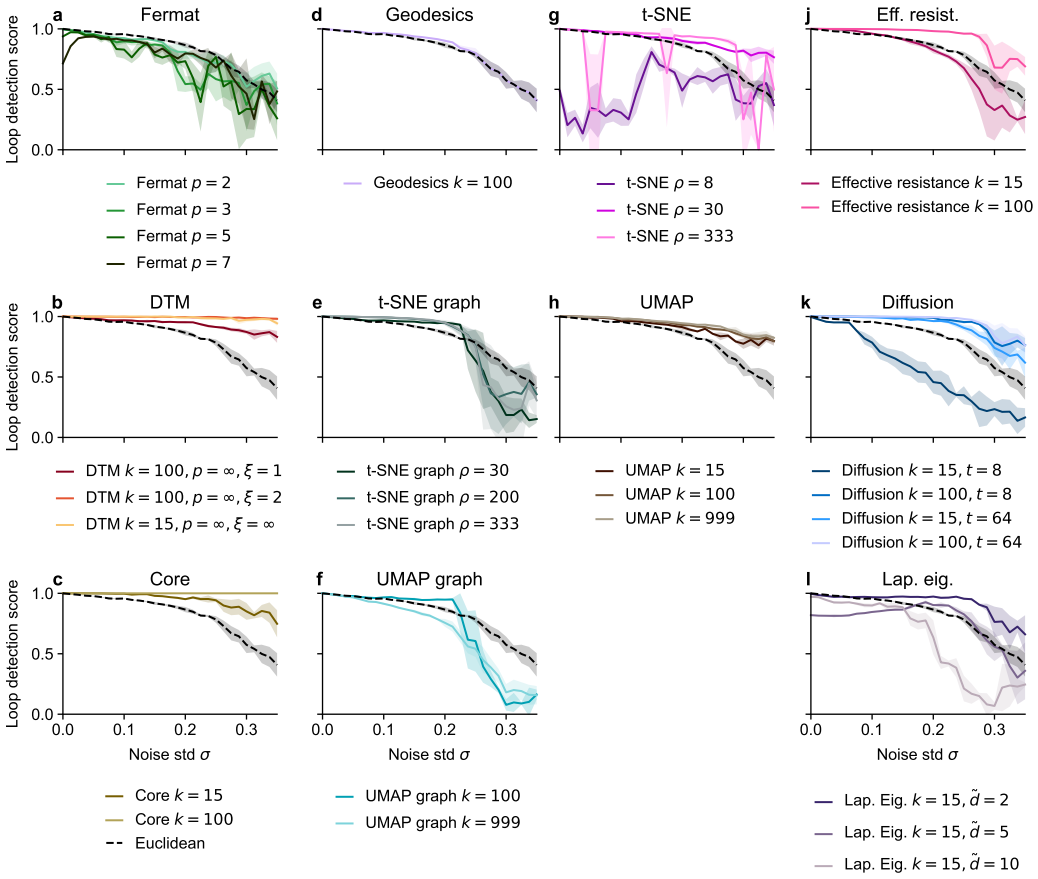

Figure S24: Loop detection score for persistent homology with various distances on a noisy circle in ambient $\mathbb{R}^2$. Our code for finding the geodesics for $k = 15$ did not terminate. Nearly all methods performed near perfectly for most noise levels. Note the striking difference to the 50D setting in Figure S23.

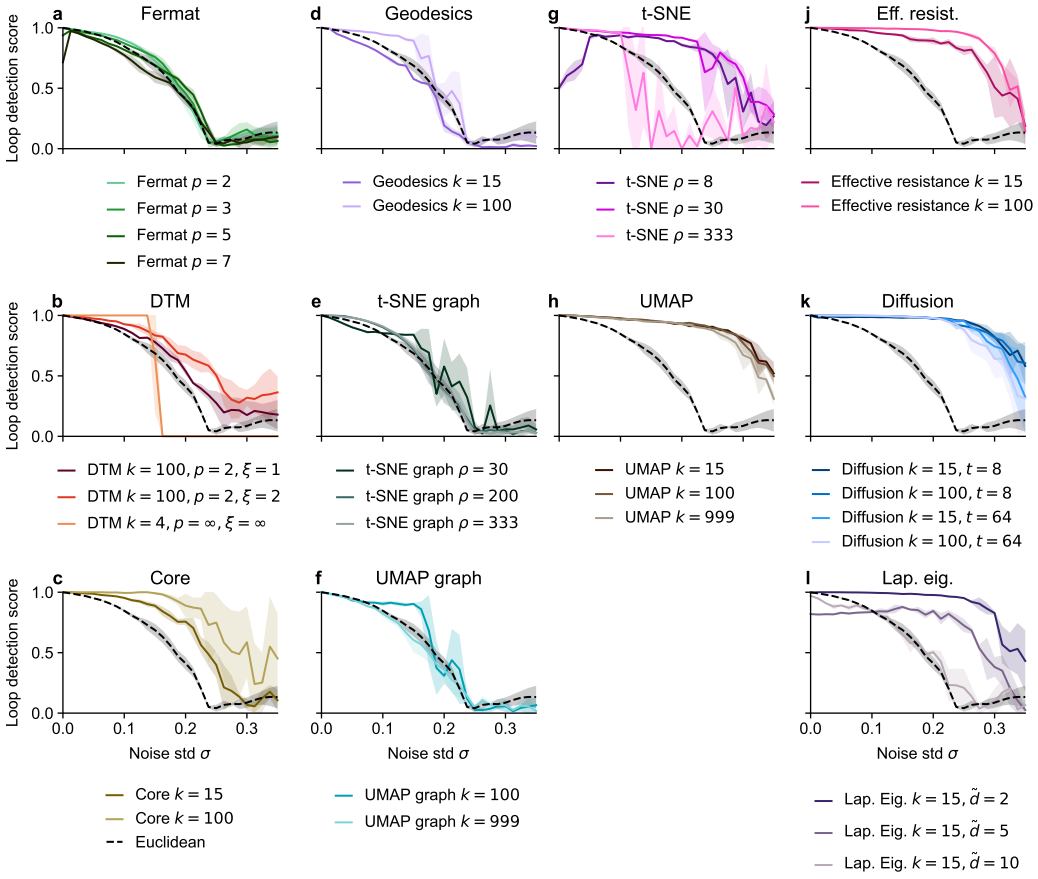

Figure S25: Loop detection score for persistent homology with various distances on a noisy circle in ambient $\mathbb{R}^{50}$. No thresholding was used for this figure, in contrast to Figure S23. Without thresholding, DTM had better performance, but not much beyond the level of Euclidean distance. Several issues such as high random seed variability for Core $k = 100$, t-SNE $\rho = 333$ and artifactually increasing performance for several methods at very high noise levels can be visible here; this is why we used the thresholding procedure in the main text.

https://doi.org/10.52202/079017-1328

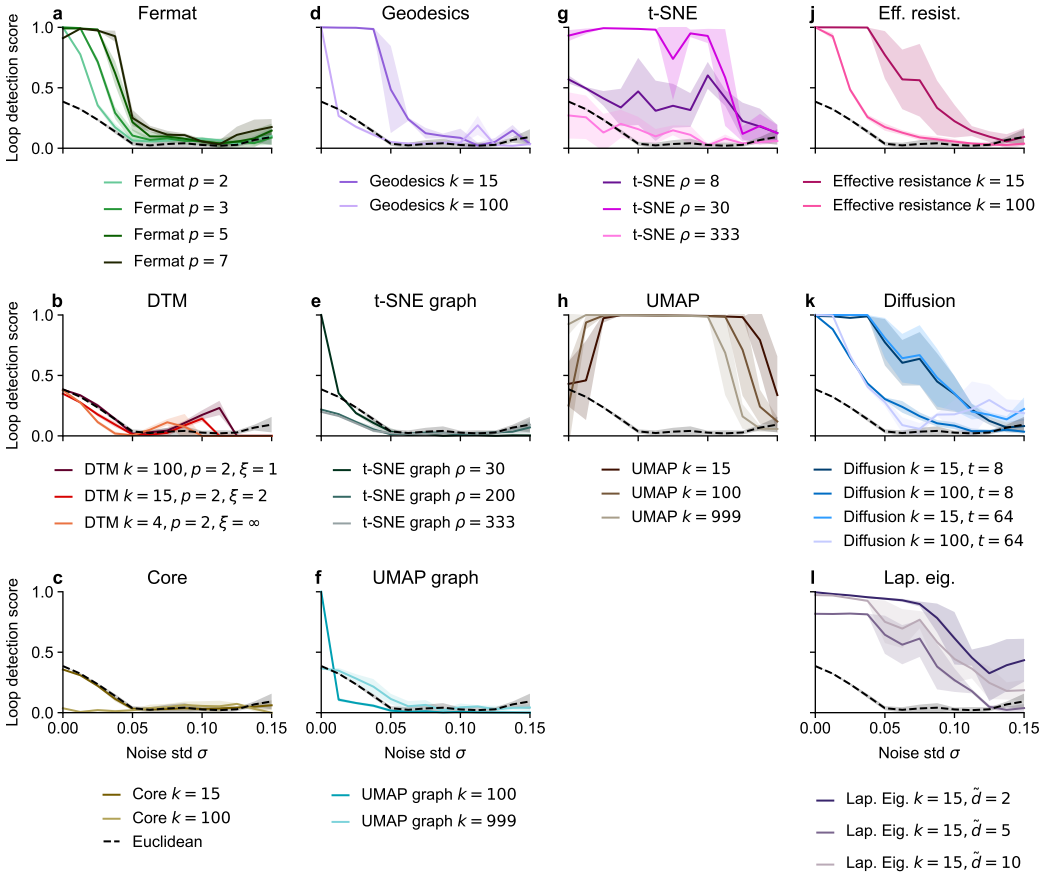

Figure S26: Loop detection score for persistent homology with various distances on the noisy eyeglasses dataset in ambient $\mathbb{R}^{50}$. Only Fermat distance, geodesics, t-SNE and UMAP, and spectral methods outperformed the Euclidean distance, but UMAP struggled in the low noise setting. The reason for the high random seed variability for effective resistance with $k = 15$ is depicted in Figure S21.

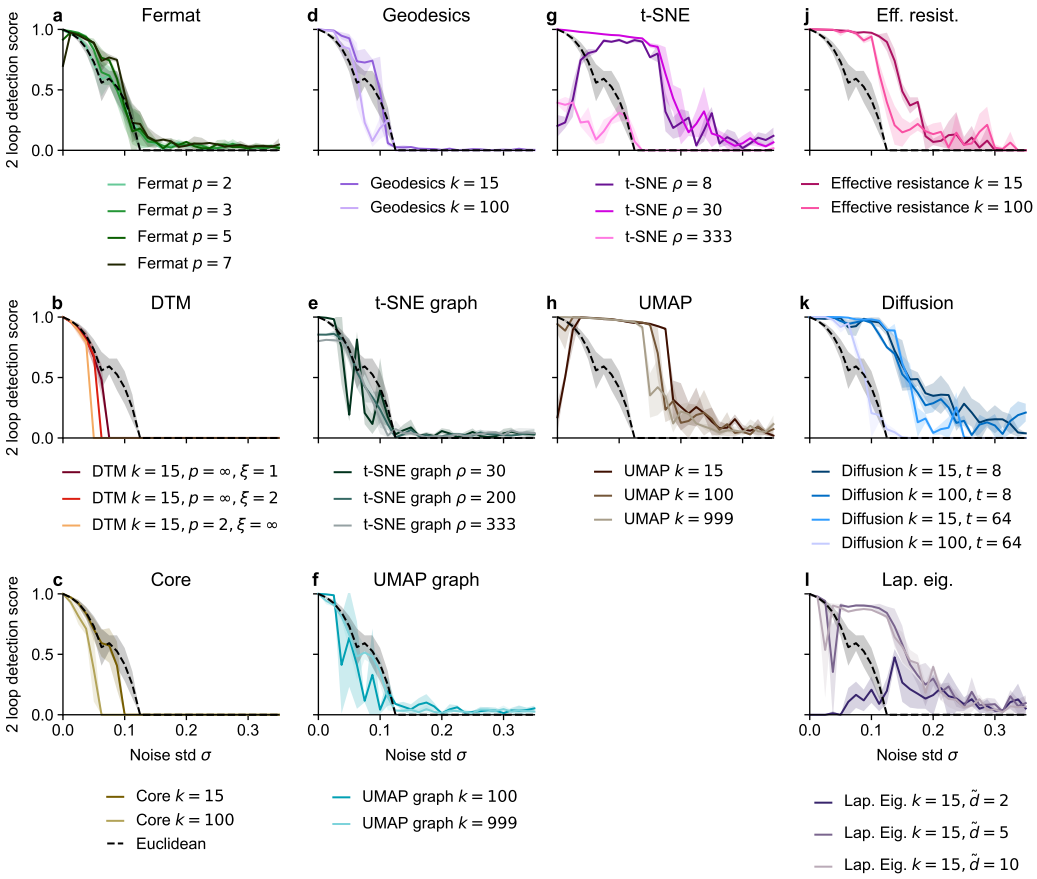

Figure S27: 2-loop detection score for persistent homology with various distances on two interlinked circles in ambient $\mathbb{R}^{50}$. Spectral and embedding methods performed best, but the latter sometimes had issues in the low noise setting.

https://doi.org/10.52202/079017-1328

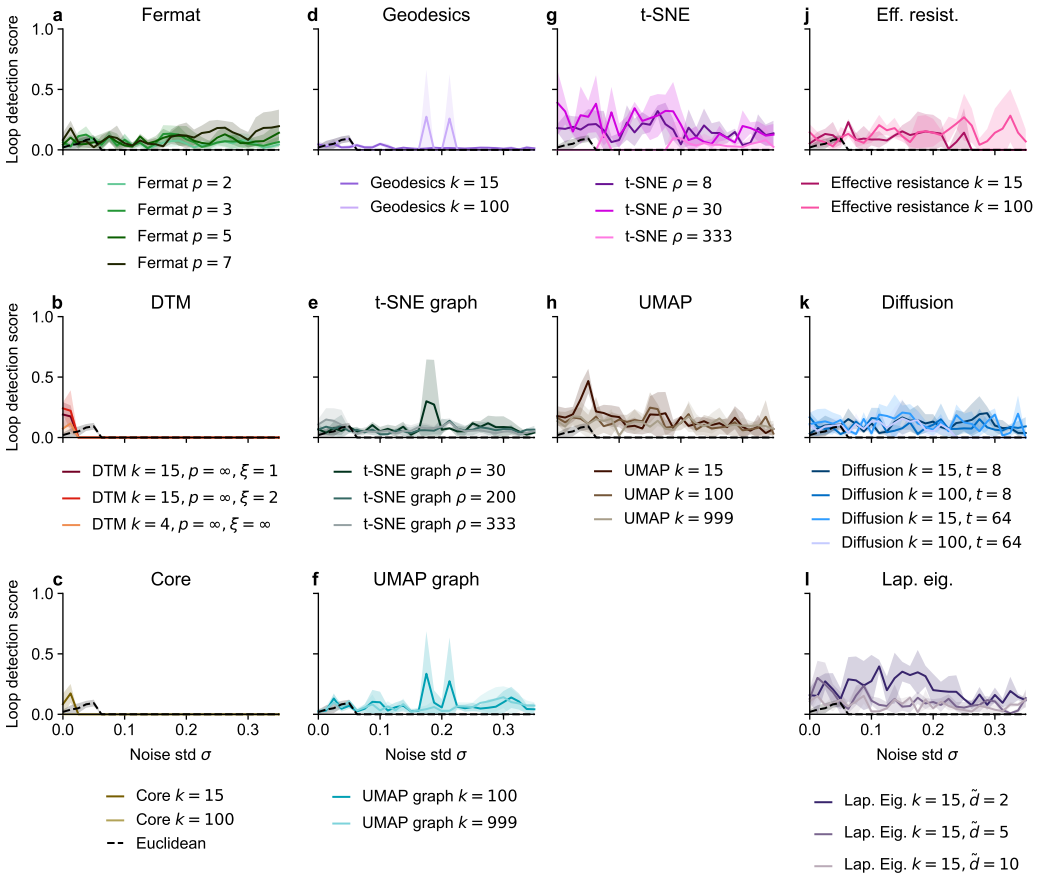

Figure S28: Loop detection score for persistent homology with various distances on a noisy sphere in ambient $\mathbb{R}^{50}$. Most methods passed this negative control.

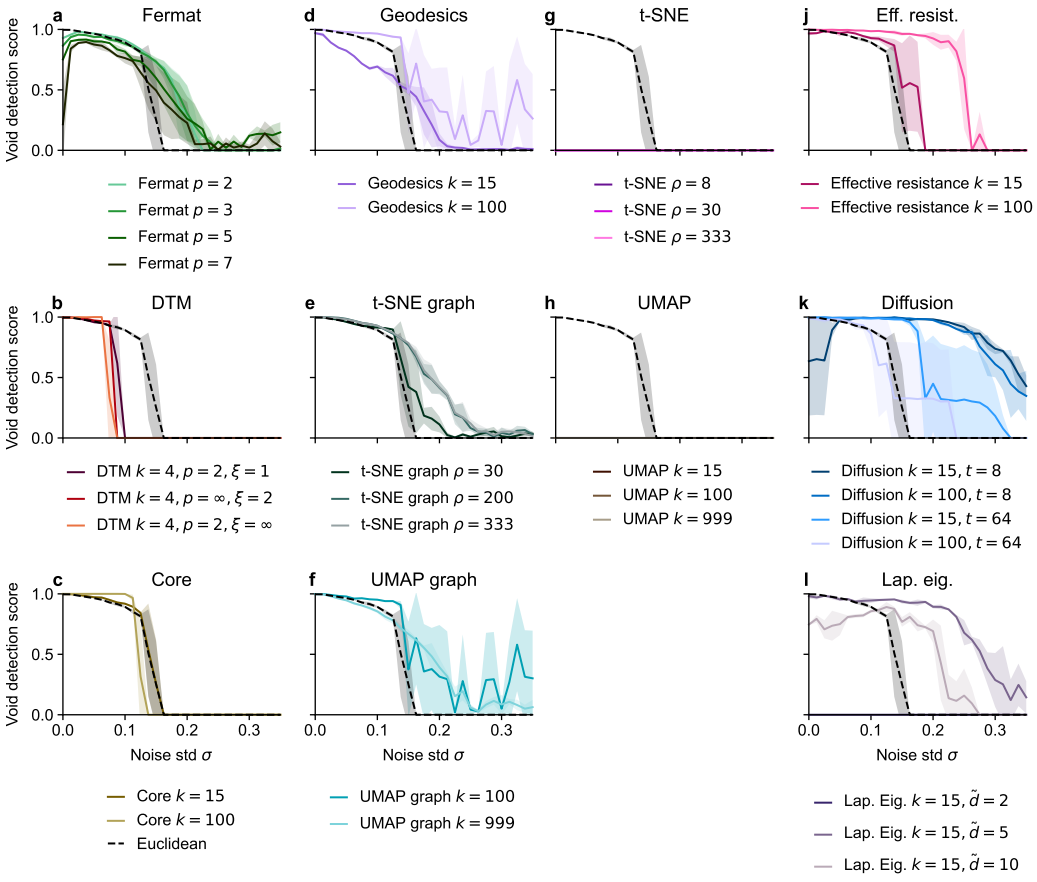

Figure S29: Void detection score for persistent homology with various distances on a noisy sphere in ambient $\mathbb{R}^{50}$. Methods relying on 2D embeddings did not find the loop for any noise level. Spectral methods performed best.

https://doi.org/10.52202/079017-1328

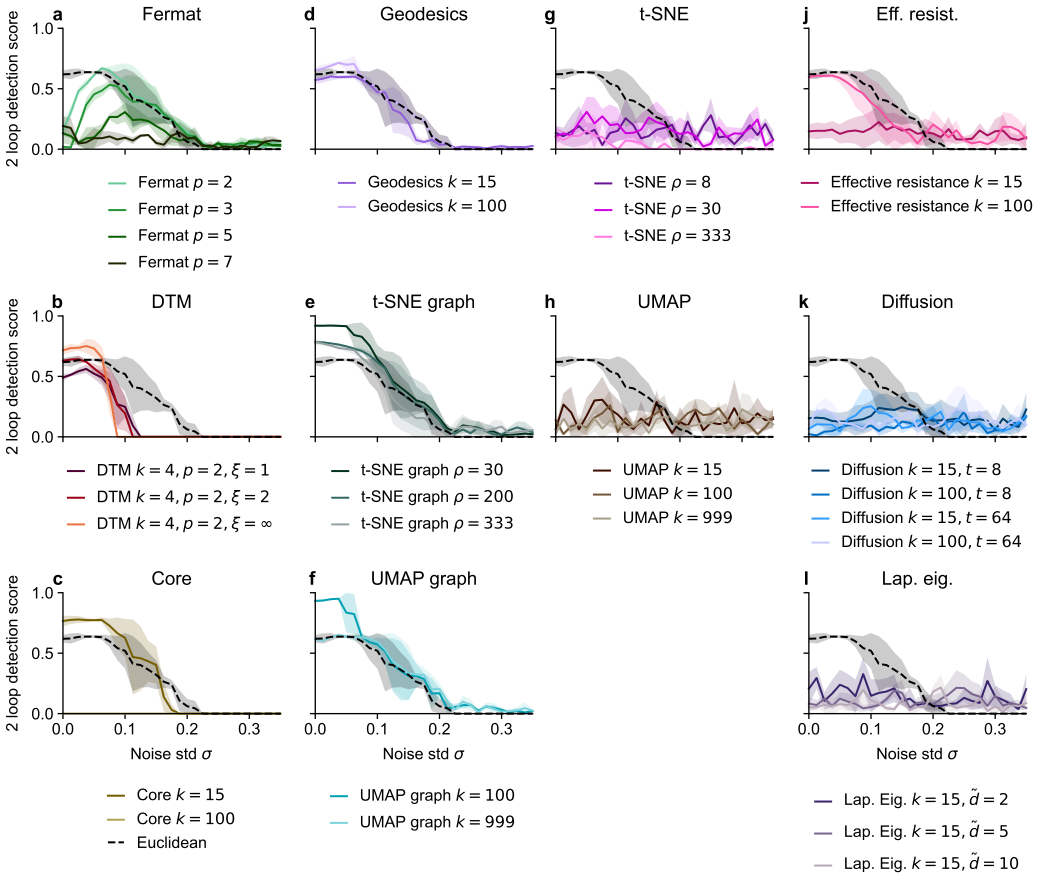

Figure S30: 2-loop detection score for persistent homology with various distances on a noisy torus in ambient $\mathbb{R}^{50}$. All methods struggled here, and only DTM, core, t-SNE graph and UMAP graph improved noticeably over the Euclidean distance. On a denser sampled torus effective resistance and diffusion distance outperformed other methods (Figure S31). Using fewer diffusion steps improved the performance on the torus (Figure S32).

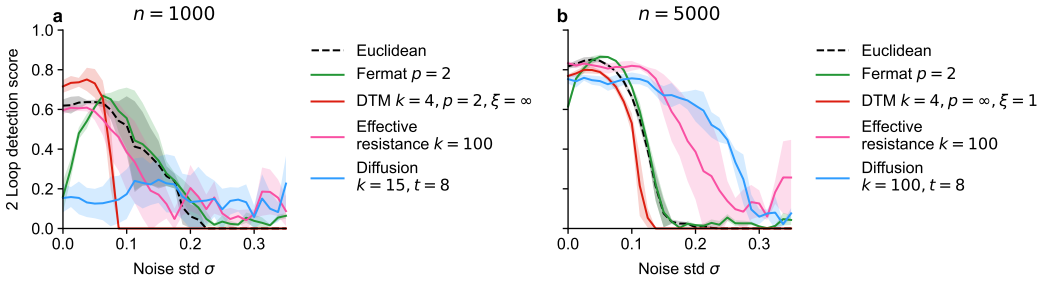

Figure S31: 2-loop detection score for persistent homology with various distances on a noisy torus with different sample size n. For more points, all methods performed better as the shape of the torus gets sampled more densely. The difference in performance is particularly striking for the spectral methods which outperformed the others for $n = 5\,000$ points, but did not for $n = 1\,000$.

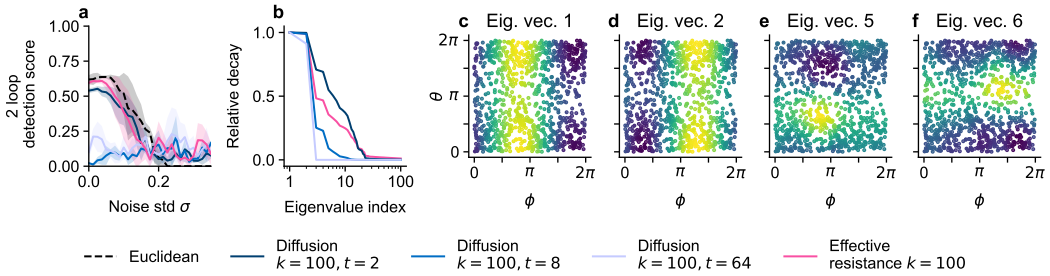

Figure S32: Diffusion distances failed on the torus with $n = 1000$ because its eigenvalue decay suppressed the relevant eigenvectors. **a.** 2-loop detection score for the torus in $d = 50$ ambient dimensions. Diffusion distances with $t = 2$ diffusion steps were on par with effective resistance and Euclidean distance. **b.** Decay of eigenvalues in various spectral distances on the noiseless torus. Diffusion distances with $t = 8, 64$ only had contribution below 0.1 for the fifth and sixth eigenvectors, while effective resistance and diffusion distance with $t = 2$ had substantial contributions from the first $\sim$10 eigenvectors. **c – f.** Eigenvectors of the symmetric graph Laplacian of a symmetric 100-nearest-neighbor graph of the noiseless torus. Coordinates are the angles of each point along (ϕ) and around (θ) the tube of the torus. The loop along the tube is encoded in the first two eigenvectors, the loop around the tube in the fifth and sixth eigenvectors.

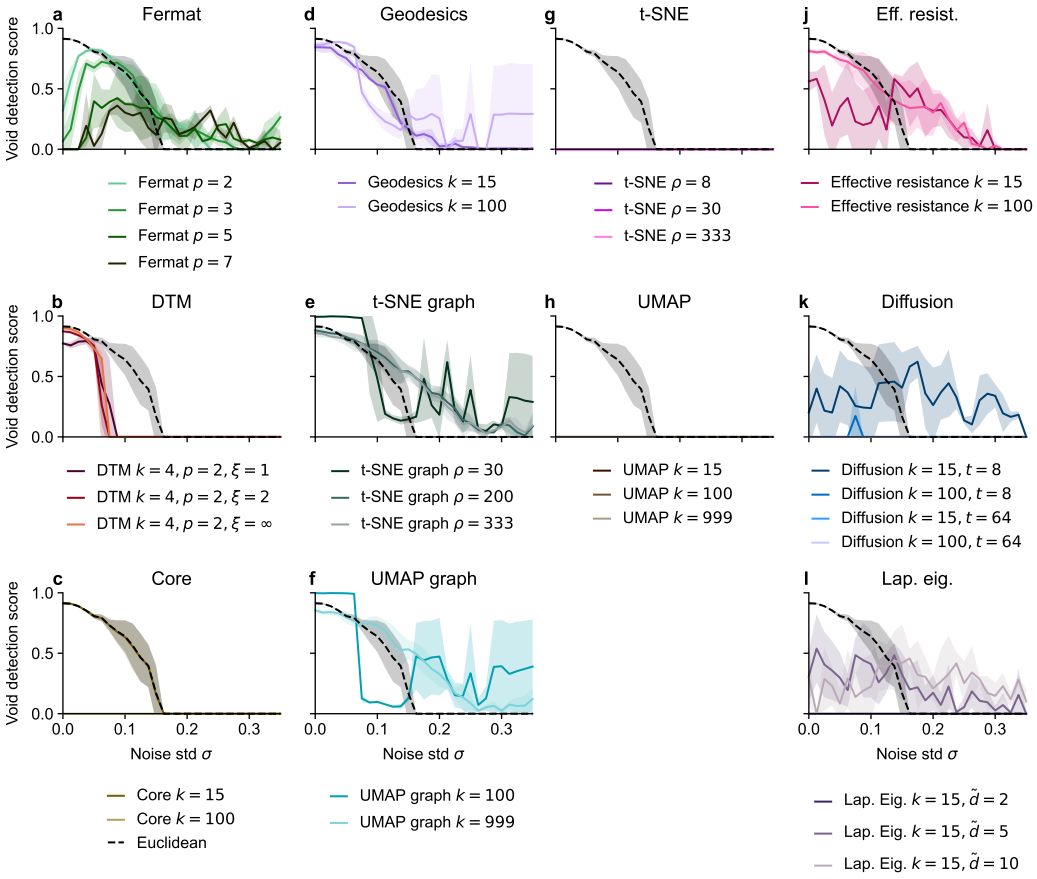

Figure S33: Void detection score for persistent homology with various distances on a noisy torus in ambient $\mathbb{R}^{50}$. Methods relying on 2D embeddings did not find the void for any noise level. Only t-SNE graph and UMAP graph could reliably improve above the Euclidean distance and only for low noise levels. However, they had unstable behavior for higher noise levels, resulting in high uncertainties. We suspect that a higher sampling density would benefit effective resistance and diffusion distance (as we saw for loop detection in Figure S31), but the computational complexity of persistent homology makes such experiments difficult.

https://doi.org/10.52202/079017-1328

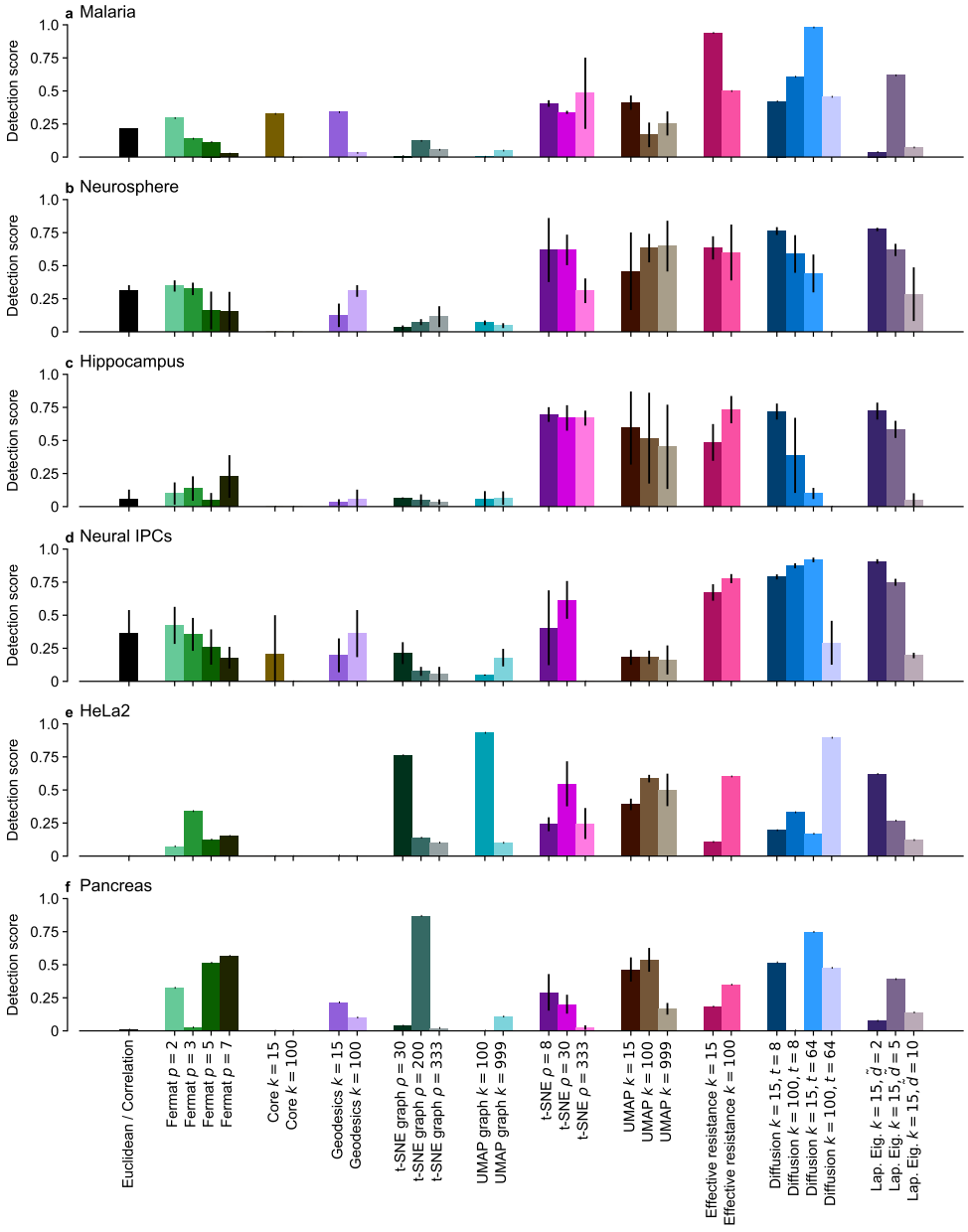

Figure S34: Detection scores for all hyperparameter settings for all six single-cell datasets. We omitted DTM as no setting passed the thresholding on any dataset. The black bar refers to correlation distance on the Malaria dataset and to Euclidean distance on the others. Extension of Figure 10. t-SNE graph and UMAP graph could perform very well, but were very hyperparameter-dependent. Their embedding variants often performed well, but collapsed on some datasets. The spectral methods behaved similarly, but on average performed better.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our contributions clearly in both the abstract and the introduction and explicitly list them at the end of the introduction. The failure modes of traditional persistent homology are described in Section 4 with formal statements and proofs in Appendix B. We state our novel closed-form expression for effective resistance in Section 6 (proof in Appendix F) and use it to relate effective resistance to other spectral methods. Finally, Section 7 contains our experimental validation of the quality of spectral methods over alternative distances as input to persistent homology.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention limitations of our work such as the ambiguity of inspecting representative cycles of detected topological features, persistent homology failing to distinguish non-isometric point clouds, the high computational complexity of any persistent homology computation, and sampling issues in high-dimensional spaces in the Limitations section 8. We also clearly state the need for selecting hyperparameters for our recommended methods in Section 7 and acknowledge that spectral methods need more samples for a good performance on the torus in Section 7.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

https://doi.org/10.52202/079017-1328

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We illustrate our claims about the failure of persistent homology with an example in Section 4 and provide full formal statements and proofs in Appendix B. The formal claims in Section 6 are rigorously shown in Appendices F and G and appropriately cross-referenced.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe every aspect of our experiments in detail, ensuring reproducibility. Datasets, distance measures, and the performance score are described in Section 7. Additionally, we give more details on the used distances and datasets in Appendices I and J. Further technical aspects of our experiments are described in the Appendix "Implementation details" M. Moreover, our code is publicly available at `https://github.com/berenslab/eff-ph/tree/neurips2024`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit our entire codebase with this submission and will make it open access at the camera ready stage. The codebase contains a README that explains how to setup the environment and which scripts to run to reproduce our experiments. Moreover, we detail which notebooks need to be run to reproduce each figure.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe our implementation in Sections 3 and 7. Moreover, we describe the explored hyperparameters in Appendix I and detail which hyperparameter values were chosen in Appendix K.

   Guidelines:

https://doi.org/10.52202/079017-1328

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For all experiments, we report the mean performance across three random seeds and the standard deviation. We specify this just before Section 7.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix M details the processor, memory, and GPU of the machine on which we ran our experiments. Moreover, we report the run times of individual experiments and an estimate for the total run time in this appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We read and conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have a Broader impact section in Appendix A, which also discusses potential negative societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any generative models or scraped datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

https://doi.org/10.52202/079017-1328

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The persistent homology computation relies on the `ripser` package, which we cite in Section 3 and we give the commit version we used in Appendix M. We cite the original authors of all the datasets we use both in Section 7.2 and Appendix J, where we also state the licenses and include the URLs from which we obtained the data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We obtained the Neural IPC data directly from Braun et al. [8] who allowed us to share it with this work (MIT license). We documented this asset in Appendix J.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1328