ROBUST GEODESIC LEARNING VIA HIERARCHICAL CONNECTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Geodesic rank is a fundamental signal for manifold learning, but it is difficult to recover when low-dimensional structure is embedded in a high-dimensional noisy ambient space. Irrelevant ambient coordinates can dominate nearest-neighbor relations, making graph-based methods unstable. We study robust geodesic learning as the problem of recovering both geodesic rank and topological structure under such noise. We propose AGD-SS, a dissimilarity measure built from random agglomerative hierarchical clustering forests. For each subsample, AGD-SS projects query points onto the sampled dendrogram and aggregates merge heights in the subtree rooted at their lowest common ancestor, using hierarchical connectivity rather than raw ambient neighborhoods. This construction connects geodesic learning with minimum spanning trees and 0-dimensional persistent homology, and supports a stability analysis of the estimator under perturbation. Experiments on synthetic manifolds, non-Euclidean settings, the Drosophila connectome, and single-cell datasets show that AGD-SS preserves local and global geodesic rank while recovering topological structure under ambient noise. These results suggest that hierarchical connectivity provides a simple inductive bias for robust representation learning that preserves both geometry and topology.

1 INTRODUCTION

Geodesic distance (Crane et al., 2020; Lee & Verleysen, 2007) is defined as the shortest path connecting two points on a Riemannian manifold. Correspondingly, geodesic learning (Madhyastha et al., 2020) is the task of inferring these intrinsic distances between all pairs of points within a given dataset. The manifold assumption (Bernstein et al., 2000) states that finite data points in ambient space $\mathbb{R}^{d_a}$ are assumed to lie on a manifold $\mathcal{M}^{d_i}$ of low intrinsic dimension $d_i \ll d_a$. In high-dimensionality case, distances will be dominated by the noise rather than by the latent manifold structure (Hein & Maier, 2006; Vandaele et al., 2022; Ting et al., 2024). Therefore, it is challenging to perform geodesic learning in a high-dimensional dataset with noise. The central challenge of Robust Geodesic Learning (RGL) is therefore to recover both geodesic rank (Madhyastha et al., 2020) and manifold topology (Fernández et al., 2023) in terms of Persistent Homology (Chazal & Michel, 2021; Hausmann et al., 1995) under this noisy circumstance.

Existing methods obtain geometric structure from different forms of local relation. Isomap (Bernstein et al., 2000) and UMAP (McInnes et al., 2018) construct neighborhood graphs from distances in the ambient space. Geodesic Forests improve robustness through sparse linear feature combinations and ensembles of data partitions (Madhyastha et al., 2020). Diffusion distance aggregates graph connectivity across random walk scales and provides a strong representation of geometry and topology (Coifman & Lafon, 2006; Damrich et al., 2024). These methods improve the reliability of the local relations used to infer intrinsic structure. Our starting point is that intrinsic structure is reflected not only by individual neighborhoods, but also by the way local connections collectively organize the data. A hierarchical representation captures this organization by recording how local connections form progressively larger structures. We therefore ask whether hierarchical connectivity can preserve geodesic rank and topology under ambient noise.

Persistent Homology (PH) (Chazal & Michel, 2021) provides a natural route to this representation. In zero-dimensional PH, the merging of connected components is represented by the edges of a Minimum Spanning Tree (MST). The same MST determines single linkage Agglomerative Hierarchical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Clustering (AHC) (Gower & Ross, 1969; Boutry et al., 2022). An AHC dendrogram (Carlsson et al., 2010; Murtagh & Contreras, 2012) therefore records connectivity across scales, linking the local attachment of observations with the larger components in which they are organized. This connection motivates hierarchical connectivity as an inductive bias for robust geodesic learning.

To turn hierarchical connectivity into a pairwise representation, we construct an ensemble of single linkage AHC trees from random subsamples. Each query point is assigned to its nearest sampled representative, and the lowest common ancestor of two representatives identifies the component that connects the pair. Rather than representing the pair by a single merge height, we aggregate the merge heights throughout the corresponding subtree. Averaging this statistic across the ensemble gives *AHC-based Geodesic Distance with Subtree Summation* (AGD-SS).

Subtree summation is the central design of AGD-SS. It describes a pair through the internal connectivity of the component that joins it, rather than through an individual local relation alone. By aggregating connectivity throughout the subtree, AGD-SS captures the multiscale organization encoded by the dendrogram.

Our contributions are summarized as follows:

- We formulate hierarchical connectivity as an inductive bias for robust geodesic learning, linking single linkage AHC, Minimum Spanning Trees, and zero-dimensional Persistent Homology to the joint preservation of geodesic rank and topology.
- We propose AGD-SS, a subtree-summation dissimilarity from random single-linkage AHC forests that aggregates merge heights in subtrees rooted at lowest common ancestor.
- We relate the expected subtree statistic to geodesic volume and establish geodesic rank preservation under this relation. We also derive a perturbation stability bound for the AGD-SS estimator.
- We show empirically that AGD-SS preserves local and global geodesic rank while recovering topological structure under ambient noise across synthetic manifolds, hyperbolic and SPD spaces, the Drosophila connectome, and single-cell datasets.

2 Problem Formulation

Given a set of data points sampled from an unknown Riemannian manifold, Robust Geodesic Learning (RGL) aims to infer a distance that approximately preserves the true geodesic rank while remaining robust to noise. We characterize an effective RGL method through four properties:

1. *Geodesic Rank Preservation*: The method must preserve both **local** and **global** geodesic rank w.r.t. the ground-truth geodesic distance $d_{\mathcal{M}}$.
2. *Manifold Topology Preservation*: The method must correctly identify the topology (Persistent Homology (Chazal & Michel, 2021; Hausmann et al., 1995)) of the dataset.
3. *Metricity*: The learned distance must have all the properties of a valid metric, including the triangle inequality, as geodesic distance is inherently a metric.
4. *Noise Robustness*: In geodesic rank and manifold topology preservation, the method must maintain its performance under noise contamination such as (i) additive noise (Damrich et al., 2024), e.g., Gaussian noise, in the original data dimensions, or (ii) additional noise dimensions that inflate the ambient space (Madhyastha et al., 2020).

2.1 Evaluation of RGL

We propose to use existing tasks: Geodesic Retrieval Madhyastha et al. (2020) for evaluating geodesic rank preservation and Loop Detection Damrich et al. (2024) for evaluating manifold topology preservation.

2.1.1 Geodesic Retrieval

In order to avoid subjective visualization or downstream performance with model specific biases, Madhyastha et al., 2020 proposed Geodesic Retrieval task where geodesic precision and recall

compare estimated neighbors to true latent manifold neighbors, providing a direct quantitative measure.

We adopt these evaluation metrics by first defining key notation. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ denote the input dataset, where each $x_i \in \mathbb{R}^{d_a}$ (ambient space) is sampled from a latent Riemannian manifold $\mathcal{M}^{d_i}$ ($d_i \ll d_a$). Let $d_{\mathcal{M}}(x_i, x_j)$ be the *true geodesic distance* between x_i and x_j on $\mathcal{M}^{d_i}$ (unknown in practice, approximated via ground-truth noise-free latent structure or directly computed on simulated manifolds).

For a query point x_q and query size k, Madhyastha et al., 2020 defined $\mathcal{N}_{\text{true}}(x_q, k)$, referred as manifold neighbors, are the k samples from $\mathcal{X}$ that are nearest to x_q based on the true geodesic distance; $\mathcal{N}_{\text{est}}(x_q, k)$ are the retrieved neighbors, i.e., the k samples that the geodesic learner reports are nearest.

Geodesic Precision quantifies the proportion of retrieved neighbors that are true manifold neighbors (i.e., the accuracy of retrieved neighbors): Geodesic Precision$(x_q, k) = \frac{|\mathcal{N}_{\text{true}}(x_q,k) \cap \mathcal{N}_{\text{est}}(x_q,k)|}{|\mathcal{N}_{\text{est}}(x_q,k)|}$.

Geodesic Recall quantifies the proportion of true manifold neighbors that are correctly retrieved by the retrieved neighbors: Geodesic Recall$(x_q, k) = \frac{|\mathcal{N}_{\text{true}}(x_q,k) \cap \mathcal{N}_{\text{est}}(x_q,k)|}{|\mathcal{N}_{\text{true}}(x_q,k)|}$.

Following Madhyastha et al., 2020, these measures work in two cases: a continuous geodesic in which there is a finite geodesic distance between all pairs of points and a discrete geodesic in which there are classes of points. In the latter case, for class-labeled datasets, we treat all points within the same class as manifold neighbors; under this setting, geodesic precision and recall are generally not identical.

2.1.2 Loop Detection

Beyond retrieval, we evaluate the capacity to preserve manifold topology in terms of Persistent Homology Chazal & Michel (2021); Hausmann et al. (1995), with a particular focus on topological features like loops. Loop detection Damrich et al. (2024) is a fundamental topological learning task that aims to identify intrinsic cyclic structures (such as 1D holes) in point clouds. It is an essential goal across diverse machine learning and scientific applications, such as the cyclical nature of the asexual intra-erythryocytic developmental cycle Howick et al. (2019). To quantify the ability of a method to recover the correct number of topological features (e.g., loops), we utilize the *Loop Detection Score* (LDS), proposed by Damrich et al., 2024. For a target manifold with l ground-truth loops, let $\{(b_i, d_i)\}_i$ be the Persistence Diagram (PD) [1] of 1-dimensional topological features, sorted by persistence $p_i = d_i - b_i$ in descending order ($p_1 \geq p_2 \geq \ldots$). The LDS for l features is defined as the relative gap between the l-th and $(l+1)$-th persistence: $s_l = (p_l - p_{l+1})/p_l$. Appendix A.2.1 shows an example for s_1 in Figure 10 .

3 Related Work

Isomap constructs a k-NN graph from ambient Euclidean distances and approximates geodesic distance by shortest path distance on this graph (Bernstein et al., 2000; Balasubramanian & Schwartz, 2002). UMAP similarly begins with an ambient neighborhood graph and learns a low-dimensional representation that preserves its local manifold structure (McInnes et al., 2018). However, Euclidean distances in the ambient space fail to approximate manifold structure when noise dimensions dominate, leading to distorted k-NN graphs and unsatisfactory geodesic estimates for Isomap and UMAP.

Diffusion distance replaces individual shortest paths with transition behavior on a graph (Coifman & Lafon, 2006). It has demonstrated stable loop detection under noise over a broad range of diffusion steps (Damrich et al., 2024), supporting Manifold Topology Preservation and Noise Robustness. Its preservation of geodesic rank, however, depends on the diffusion time m. Small m retains local distinctions but remains sensitive to noisy transitions, whereas large m captures broader structure but

[1] Persistent Homology (PH) studies the evolution of topological features (connected components, loops, voids, etc.) across a filtration of nested spaces. The Persistence Diagram (PD) is a multi-set of points $(b_i, d_i) \in \mathbb{R}^2$, where each point represents a topological feature born at scale b_i and destroyed at scale $d_i (\geq b_i)$. The lifetime (or persistence) of a feature is $d_i - b_i$, with longer lifetimes typically indicating more significant features. The details of PD are given in Appendix A.2.

Table 1: Summary of Euclidean distance, Diffusion distance, Geodesic Forest and AGD-SS. $\sim$ indicates the Diffusion distance's ability of preserving geodesic rank depends on its hyperparameter choice, i.e., the number of diffusion steps m.

Desired Properties		Euclidean	Diffusion	GF	AGD-SS
Geodesic Rank	Local	$\times$	$\sim$	✓	✓
	Global	$\times$		$\times$	✓
Manifold Topology		$\times$	✓	$\times$	✓
Metricity		✓	✓	✓	✓
Noise Robustness		$\times$	✓	✓	✓

can smooth local geometry, as shown in Figure 62. Thus, the same parameter controls the balance between local and global Geodesic Rank Preservation.

Geodesic Forests (GF) [2] reduce sensitivity to ambient noise using sparse linear feature combinations and an ensemble of unsupervised partitions (Madhyastha et al., 2020). Each tree uses sparse projections and the Fast BIC splitting criterion, and the forest distance averages whether two points fall into different leaves. This construction provides Noise Robustness without first constructing a neighborhood graph and is effective for local geodesic retrieval. Since shared leaf membership records whether two points fall into the same terminal partition rather than their ordering over long manifold distances, global geodesic rank and manifold topology are not its primary targets.

Table 1 summarizes these methods through the four RGL properties. Unlike Isomap and UMAP, AGD-SS does not rely on a single ambient neighborhood graph. It shares the ensemble strategy of GF but replaces terminal leaf membership with a subtree statistic, and unlike diffusion distance, it does not use diffusion time to balance local and global rank. AGD-SS thus jointly targets geodesic rank and topology under ambient noise while retaining metricity.

4 Achieving Robust Geodesic Learning via Agglomerative Hierarchical Clustering

We first connect agglomerative hierarchical clustering with geodesic learning through AHC based Geodesic Distance (AGD), which represents a pair by the merge heights along the paths joining it. Motivated by the equivalence between single linkage dendrograms, minimum spanning trees, and zero-dimensional persistent homology (Gower & Ross, 1969; Boutry et al., 2022; Wu et al., 2025), we then introduce AGD with Subtree Summation (AGD-SS).

4.1 AHC-based Geodesic Distance

Definition 4.1 (AHC-based Geodesic Distance)**.** Let $\mathcal{X} \subset \mathbb{R}^d$ be a dataset and $\psi \in \mathbb{N}$ a subsample size. Let $S \subset \mathcal{X}$ be a random subset of cardinality ψ, and $\mathcal{T}_S$ be the agglomerative hierarchical clustering dendrogram constructed on S with node heights (linkage values[3]) $h(\cdot)$. We define the nearest neighbor (NN) projection $\phi_S : \mathcal{X} \to S$ as $\phi_S(x) = \arg\min_{y \in S} \|x - y\|$. The AHC-based Geodesic Distance (AGD) between $x, y \in \mathcal{X}$, i.e., $d_{\mathrm{AGD}}(x, y)$, is defined as

$$d_{\mathrm{AGD}}(x, y) = \mathbb{E}_S \left[\sum_{w \in \mathcal{P}(\phi_S(x), \mathrm{LCA})} h(w) + \sum_{w \in \mathcal{P}(\phi_S(y), \mathrm{LCA})} h(w) \right], \tag{1}$$

where LCA denotes the lowest common ancestor of $\phi_S(x)$ and $\phi_S(y)$ in $\mathcal{T}_S$, and $\mathcal{P}(u, v)$ denotes the set of nodes on the path from u to its ancestor v. The height $h(\mathrm{LCA})$ is counted once on each path. The empirical d_{AGD} is estimated by averaging over t random subsets.

[2] We prove that GF has *Metricity* in Appendix B.3.

[3] In the main body of this work, we only consider single linkage function. Empirical comparisons working with other linkage functions are given by Figures 36-44 in Appendix E.7.

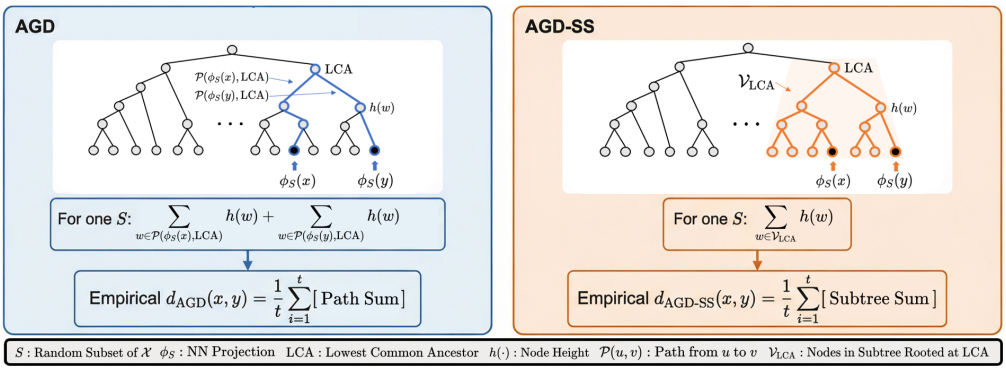


Figure 1: AGD (*Left*) and AGD-SS (*Right*) over one random sample subset S from dataset.

Relation to Geodesic Rank. The path statistic accumulates the merge scales encountered while connecting the two representatives, as illustrated in Figure 1 (*left*). It therefore provides a hierarchical ordering of pairs by their cumulative connection cost, which is used to recover local and global geodesic rank.

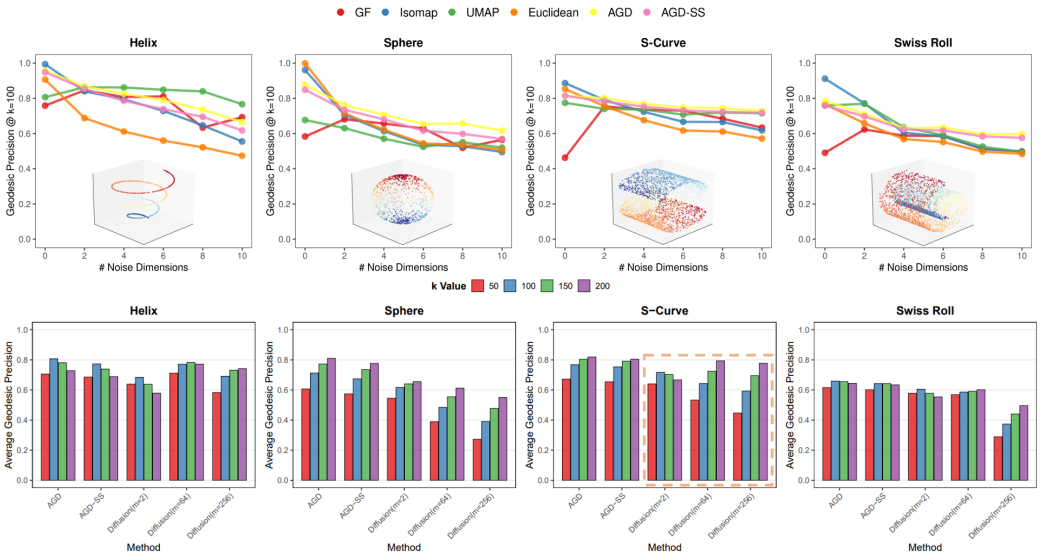


Figure 2: *Top*) Artificial datasets and geodesic precisions of different methods when $k = 100$. All metrics are used on normalized data. The Gaussian noise std of noise dimension is $\sigma = 20$. The results under different k values are given in Figure 25 in Appendix E.5. (*Bottom*) Comparison of average geodesic precision over different numbers of noise dimensions for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$). More detailed results of each number of noise dimensions are given in Figure 26 in Appendix E.5. For Helix and Sphere datasets, the ground-truth geodesic distances are directly computed in the latent manifold. For S-Curve and Swiss Roll datasets generated by the scikit-learn API Buitinck et al. (2013); Pedregosa et al. (2011), the ground-truth geodesic distances are approximated as the shortest path distances on the 10-NN graph on the data with no noise dimensions. Analysis on connectivity and non-convexity of manifold is given in Appendix E.4.

Validation on Artificial Datasets. Figure 2 shows that AGD retains geodesic neighborhoods as irrelevant ambient dimensions are added. Across retrieval sizes, the comparison with GF further distinguishes local retrieval from global geodesic rank: shared leaf membership is effective at small k, whereas the hierarchical path statistic remains informative as k increases.

Comparison with Diffusion Distance. Diffusion time m balances local detail and broader connectivity: small m retains local distinctions but can be sensitive to noise, while large m may smooth away signal Damrich et al. (2024); Moon et al. (2019). Figure 2 (*bottom*) compares $m = 2, 64, 256$ on a 10-NN graph, following the range considered by Damrich et al., 2024. Smaller retrieval sizes k assess local geodesic neighborhoods, whereas larger k probe ordering over broader regions of the manifold. In the highlighted region, $m = 2$ gives the highest average precision for local retrieval ($k = 50$), while $m = 64$ performs best at the more global scale ($k = 200$); increasing m to 256 reduces precision. Swiss Roll shows similar behavior. With fixed settings $\psi = 32$ and $t = 100$, AGD and AGD-SS generally achieve comparable or higher precision than Diffusion distance.

4.2 AGD WITH SUBTREE SUMMATION

Building on AGD, we propose AGD-SS, AHC-based Geodesic Distance with Subtree Summation, to incorporate the internal connectivity of the component joining a pair.

Definition 4.2 (AHC-based Geodesic Distance with Subtree Summation)**.** Following the same notation in AGD, for any two points $x, y \in \mathcal{X}$, let LCA denote the lowest common ancestor of $\phi_S(x)$ and $\phi_S(y)$ in $\mathcal{T}_S$, and let $\mathcal{V}_{\text{LCA}}$ denote the set of all nodes in the subtree rooted at LCA. The AHC Subtree Distance (AGD-SS) between $x, y \in \mathcal{X}$, i.e., $d_{\text{AGD}-\text{SS}}(x, y)$, is given by

$$d_{\text{AGD}-\text{SS}}(x, y) = \mathbb{E}_S \left[\sum_{w \in \mathcal{V}_{\text{LCA}}} h(w) \right], \tag{2}$$

where the sum aggregates the linkage values of all nodes in the subtree. The empirical distance, denoted by $\widehat{d}_{\text{AGD}-\text{SS}}$, is estimated as the average over t random subsets. For a tree built on one subset S, we denote the corresponding subtree sum by $d_{\text{AGD-SS}}(x, y \mid S)$.

A demonstration is given in Figure 1 (*right*). Whereas AGD records two selected paths, AGD-SS represents the pair through the full hierarchical component rooted at their LCA.

Relation to Persistent Homology. The nodes in $\mathcal{T}_S$ record the merging events of connected components in zero-dimensional persistent homology, and $h(w)$ is the scale of the corresponding merge (Gower & Ross, 1969; Boutry et al., 2022), as shown in Figure 3. Summing these heights over $\mathcal{V}_{\text{LCA}}$ therefore aggregates the persistence information involved in connecting x and y. Since the persistence diagram is stable under perturbations (Bottleneck Stability Theorem Chazal & Michel, 2021), the derived distance $d_{\text{AGD-SS}}$ inherits this robustness, making it resilient to noise, as shown in Theorem 4.5. Proof is given in Appendix B.2 and validation is given in Appendix E.2. Dendrogram visualizations are given in Appendix E.1.

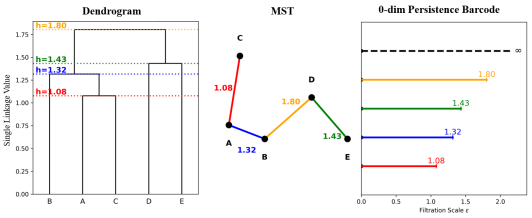


Figure 3: A single linkage dendrogram, its minimum spanning tree, and zero-dimensional persistent homology encode the same component merging events.

Theorem 4.3 (Metricity of Population AGD-SS)**.** *Assume that the sampling distribution separates every distinct pair with positive probability, so that* $\mathbb{P}_S[\phi_S(x) \neq \phi_S(y)] > 0$ *for all* $x \neq y$*. Then population AGD-SS is a valid metric.*

For a finite forest, empirical AGD-SS is generally a pseudometric because distinct queries may share the same representative in every sampled tree. It can be converted into a metric by $\widehat{d}_{\text{AGD-SS},\varepsilon}(x, y) = \widehat{d}_{\text{AGD-SS}}(x, y) + \varepsilon \mathbf{1}[x \neq y]$, which preserves every off diagonal ranking. Thus, metricity can be enforced without changing the geodesic retrieval order.

Definition 4.4 (δ perturbation)**.** For a finite dataset $\mathcal{X}$, a δ perturbation is a set $\mathcal{X}'$ in which each $x \in \mathcal{X}$ has a corresponding point $x' \in \mathcal{X}'$ satisfying $\|x - x'\| \leq \delta$.

Theorem 4.5 (Conditional Stability of AGD-SS under Bounded Perturbation)**.** *Let* $\mathcal{X}'$ *be a* δ *perturbation of* $\mathcal{X}$*. Fix* $x, y \in \mathcal{X}$ *and their perturbed counterparts* $x', y' \in \mathcal{X}'$*. Let* S *be a uniform subsample of size* ψ *and let* S' *be its paired perturbation. Assume that, with probability at least* $1 - \alpha$ *over* S*, the query leaves attach to corresponding nearest representatives in* S *and* S' *when*

constructing AHC from points $S \cup \{x, y\}/S' \cup \{x', y'\}$. *If* $d_{\text{AGD-SS}}(\cdot, \cdot \mid S) \leq B_\psi$, *then*

$$|d_{\text{AGD-SS}}(x, y) - d_{\text{AGD-SS}}(x', y')| \leq 2(\psi + 3)\delta + \alpha B_\psi. \tag{3}$$

For t *independent subsets, the empirical version satisfies the same bound plus* $2B_\psi\sqrt{\log(4/\rho)/(2t)}$ *with probability at least* $1 - \rho$.

Theorem 4.5 gives a conditional stability result for AGD-SS under bounded perturbation. Consequently, the normalized empirical version $|d_{\text{AGD-SS}}(x, y) - d_{\text{AGD-SS}}(x', y')|/\psi$ is upper bounded by 2δ as $\psi, t \to \infty$. Theorem 4.5 does not guarantee the noise robustness in Section 2, which is based on task performance. Compared with Euclidean distance who also has a similar 2δ bound, AGD-SS better reflects geodesic rank. We analyze rank preservation in Appendix B.4 with verified assumptions in Appendix B.4.3.

Validation on Artificial Datasets. The synthetic results in Figure 2 show that AGD-SS remains effective as the number of ambient noise dimensions increases and preserves geodesic rank across retrieval sizes. The close behavior of AGD and AGD-SS on these controlled manifolds establishes that subtree aggregation retains the geometric information of the path statistic; its contribution on heterogeneous biological structure is examined in the following experiment section.

Effect of Subtree Summation. On a fixed tree, AGD-SS is invariant to query assignment perturbation that preserves the LCA. Paired experiments across five manifolds and 45 noise conditions confirm this property and show greater average distance rank stability than AGD when assignments are computed from noisy rather than clean coordinates (Appendix E.12).

5 Experiments

5.1 Geodesic Retrieval on Drosophila Connectome Dataset

Besides the validation on artificial datasets, we conduct Geodesic Retrieval task on a bioinformatics dataset: Drosophila Connectome (Eichler et al., 2017), which is the same as that used in the Geodesic Forests (Madhyastha et al., 2020).

A significant recent development on the field of brain networks, known as connectomics (Vogelstein et al., 2019), was the estimation and public release of the complete larval Drosophila mushroom body connectome, i.e., the neural circuit responsible for learning and memory in the fruit fly (Eichler et al., 2017).

Comprising 200 nodes, each represents a distinct neuron within the mushroom body, interconnected by approximately 75000 edges. An edge is defined between any pair of neurons if they share at least one synapse. These connections link vertices across four identified cell classes: Kenyon cells, input neurons, output neurons, and projection neurons (Priebe et al., 2017), as shown in Figure 4 (*Left*). This dataset does not provide an explicit latent geodesic distance. Following existing GF benchmark (Madhyastha et al., 2020), neurons within the same class are treated as manifold neighbors. Therefore, precision/recall on the geodesic retrieval task on Drosophila Connectome dataset should be interpreted as a label-based proxy for geodesic rank preservation rather than a direct measurement of exact geodesic distances.

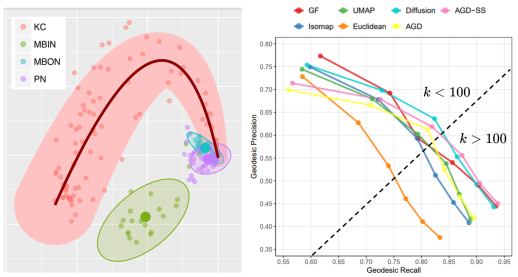


Figure 4: *Left*. Two of six dimensions of the Drosophila Connectome dataset. This is a direct reuse of Figure 10 in Priebe et al. (2017). *Right*. Geodesic Precisions/Recalls.

Figure 4 (*Right*) reports retrieval precision and recall for neighborhood sizes k from 50 to 150. For $k < 100$, GF and Diffusion distance show competitive retrieval performance. For $k > 100$, AGD-SS remains effective as the retrieved neighborhoods expand, with Diffusion distance showing similar performance. These results suggest that AGD-SS captures relationships consistent with biological cell classes at larger retrieval sizes, supporting the use of hierarchical connectivity to represent intrinsic neighborhood structure.

5.2 Loop Detection

5.2.1 Artificial Data

We provide a comparison with Diffusion distance Damrich et al. (2024), a baseline for loop detection task. While Diffusion distance is stable w.r.t. noise, it has distorted distances (the Diffusion distances between many points far from a reference point are all the same) and hence perturbs the manifold structure when the noise std σ is small, as shown in Figure 5. But this does not affect the Loop Detection Score because the global topology structure is still preserved (although distorted).

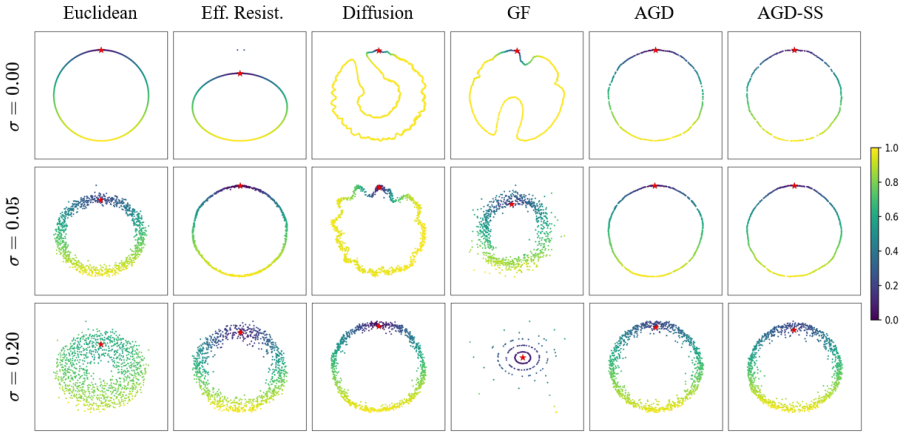


Figure 5: Multidimensional scaling Carroll & Arabie (1998) of Euclidean, Diffusion distances, GF, AGD and AGD-SS for a noisy circle in $\mathbb{R}^{50}$, under different noise stds (σ). Color indicates the distance to the red reference point. Results of higher ambient dimension ($\mathbb{R}^{100}, \mathbb{R}^{200}$) are provided in Appendix E.11.

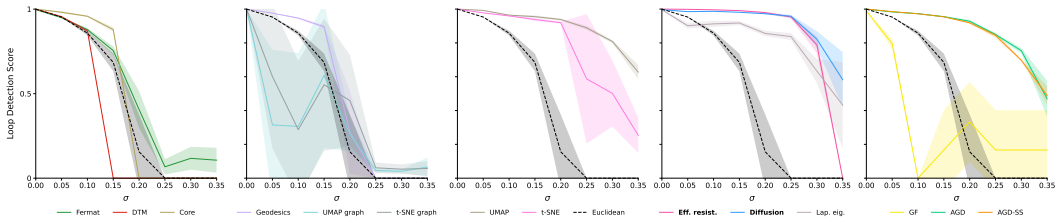


Figure 6: Loop Detection Score s_1 for persistent homology with various distances on a noisy circle in $\mathbb{R}^{50}$. Methods are grouped into panels for visual clarity. Euclidean distance is used in each subplot for reference. Loop Detection Score on noisy circle in higher ambient dimension ($\mathbb{R}^{100}, \mathbb{R}^{200}$) are provided in Figure 61.

In Figure 5, as noise std σ increases, we observe that Euclidean distance and GF fail to maintain the loop topology, collapsing the data into a cluster. Diffusion distance preserves loop topology but not global geodesic rank. In contrast, AGD and AGD-SS robustly preserve the global circular structure and the correct geodesic ordering (indicated by the smooth color gradient from the reference point), across all σs.

Figure 6 reports the corresponding Loop Detection Score s_1. The results support the ability of AGD-SS to identify the single loop under the tested noise levels, consistent with the circular structure observed in the MDS embeddings.

5.2.2 Single-Cell RNA-Sequencing Data

We extend our evaluation to real-world single-cell RNA-sequencing (scRNA-seq) data, which consists of expression vectors for thousands of genes across individual cells. These datasets are characteristically noisy. We select datasets known to exhibit circular structures, driven by the cyclic dynamics of the cell division cycle, to provide a biological reference for interpreting the detected loops.

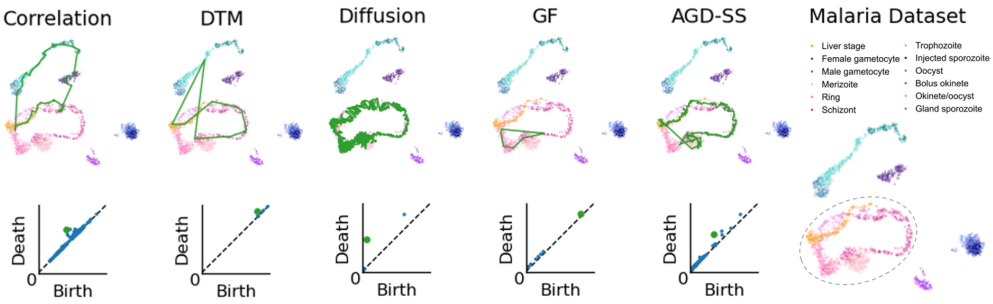


Figure 7: Representatives of the most persistent loop overlaid on UMAP embedding (rightmost) and persistence diagrams using five methods. Biology and the UMAP embedding dictate that there should be one major loop (marked in gray dashed line): Merozoites, rings, trophozoites and schizonts formed a circle capturing the cyclical nature of the asexual intra-erythryocytic developmental cycle.

Case Study on Malaria Dataset Howick et al. (2019). Figure 7 illustrates the performance of different methods. Similar to Diffusion distance, AGD-SS successfully recovers the intrinsic cyclic structure, i.e. a circle capturing the cyclical nature of the asexual intra-erythryocytic developmental cycle. While Correlation and DTM can detect persistent loops, these loops are not corresponding to the actual cyclical nature of the Malaria dataset. For GF, it only detects a very small loop.

Loop Detection on More Datasets. We also conduct Loop Detection task on three additional scRNA-seq datasets Neurosphere & Hippocampus from Zheng et al., 2022 and Neural IPCs from Braun et al., 2023. The results in Figure 8 extend the evaluation to additional biological settings and provide further evidence that AGD-SS can identify persistent loops in scRNA-seq data.

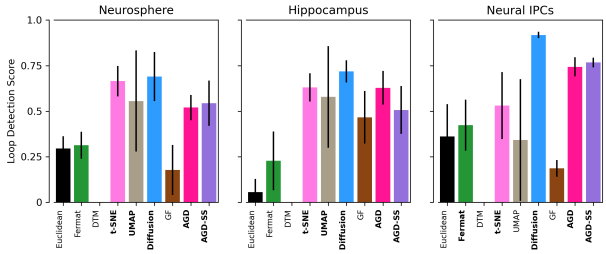


Figure 8: Loop Detection Scores (s_1) on additional datasets.

Topology & Geometry. The geodesic retrieval and loop detection results provide complementary evidence that AGD-SS captures both geodesic rank and cyclic structure under the tested noise conditions. The MDS visualizations in Figure 5 further illustrate these properties through the visible circular structure and smooth color variation around the circle. Together, these results support hierarchical connectivity as a useful signal for recovering intrinsic structure from noisy observations.

Additional Results. We provide hyperparameter sensitivity analysis in Appendix E.8 and E.9. Hyperparameter settings are given in Appendix D. Time cost is reported in Appendix E.10.

6 Conclusion

We study robust geodesic learning through hierarchical connectivity and propose AGD-SS, a dissimilarity measure that aggregates merge heights within LCA subtrees across random AHC forests. Its connection to minimum spanning trees and zero-dimensional persistent homology provides a structural interpretation of subtree summation.

Our analysis establishes geodesic rank preservation under a geodesic volume relation and characterizes perturbation stability under suitable structural conditions. Experiments on synthetic and biological datasets support the use of AGD-SS for capturing geodesic ordering and cyclic structure in noisy observations. These findings highlight hierarchical connectivity as a useful signal for learning intrinsic geometry and topology under ambient noise.

AI USE STATEMENT

In this work, we used generative AI tools for polishing, including enhancing clarity, coherence, and correcting minor grammatical issues. Additionally, we use AI tools to implement our proposed method. We have reviewed all AI-assisted work. we checked LLM-generated text. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We use synthetic and publicly available biological datasets. No new human participant studies or animal experiments were conducted. We have identified no specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Detailed experimental settings and complete proofs of the theoretical results are provided in the appendix. The source code is submitted as supplementary material.

REFERENCES

Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Citeseer, 2000.

Nicolas Boutry, Laurent Najman, and Thierry Géraud. Some equivalence relation between persistent homology and morphological dynamics. *Journal of Mathematical Imaging and Vision*, 64(7): 807–824, 2022.

Emelie Braun, Miri Danan-Gotthold, Lars E Borm, Ka Wai Lee, Elin Vinsland, Peter Lönnerberg, Lijuan Hu, Xiaofei Li, Xiaoling He, Žaneta Andrusivová, et al. Comprehensive cell atlas of the first-trimester developing human brain. *Science*, 382(6667), 2023.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Gunnar E Carlsson, Facundo Mémoli, et al. Characterization, stability and convergence of hierarchical clustering methods. *J. Mach. Learn. Res.*, 11(Apr):1425–1470, 2010.

J Douglas Carroll and Phipps Arabie. Multidimensional scaling. *Measurement, judgment and decision making*, pp. 179–250, 1998.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, 2021.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Keenan Crane, Marco Livesu, Enrico Puppo, and Yipeng Qin. A survey of algorithms for geodesic paths and distances. *arXiv preprint arXiv:2007.10430*, 2020.

Sebastian Damrich, Philipp Berens, and Dmitry Kobak. Persistent homology for high-dimensional data based on spectral methods. *Advances in Neural Information Processing Systems*, 37:41954–42014, 2024.

Katharina Eichler, Feng Li, Ashok Litwin-Kumar, Youngser Park, Ingrid Andrade, Casey M Schneider-Mizell, Timo Saumweber, Annina Huser, Claire Eschbach, Bertram Gerber, et al. The complete connectome of a learning and memory centre in an insect brain. *Nature*, 548(7666): 175–182, 2017.

Ximena Fernández, Eugenio Borghini, Gabriel Mindlin, and Pablo Groisman. Intrinsic persistent homology via density-based metric learning. *Journal of Machine Learning Research*, 24(75):1–42, 2023.

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

Jean-Claude Hausmann et al. On the vietoris-rips complexes and a cohomology theory for metric spaces. *Annals of Mathematics Studies*, 138:175–188, 1995.

Matthias Hein and Markus Maier. Manifold denoising. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/a0b83c02d720415dada82e08bc09e9f3-Paper.pdf.

Virginia Howick, Andrew Russell, Tallulah Andrews, Haynes Heaton, Adam Reid, Kedar Natarajan, Hellen Butungi, Tom Metcalf, Lisa Verzier, Julian Rayner, et al. The malaria cell atlas: Single parasite transcriptomes across the complete plasmodium life cycle. *Science*, 365(6455), 2019.

John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

Meghana Madhyastha, Gongkai Li, Veronika Strnadová-Neeley, James Browne, Joshua T. Vogelstein, Randal Burns, and Carey E. Priebe. Geodesic forests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 513–523, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403094. URL https://doi.org/10.1145/3394486.3403094.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Kevin R Moon, David Van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12): 1482–1492, 2019.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 2(1):86–97, 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Carey E Priebe, Youngser Park, Minh Tang, Avanti Athreya, Vince Lyzinski, Joshua T Vogelstein, Yichen Qin, Ben Cocanougher, Katharina Eichler, Marta Zlatic, et al. Semiparametric spectral modeling of the drosophila connectome. *arXiv preprint arXiv:1705.03297*, 2017.

Kai Ming Ting, Takashi Washio, Ye Zhu, Yang Xu, and Kaifeng Zhang. Is it possible to find the single nearest neighbor of a query in high dimensions? *Artificial Intelligence*, 336:104206, 2024.

Robin Vandaele, Bo Kang, Tijl De Bie, and Yvan Saeys. The curse revisited: When are distances informative for the ground truth in noisy high-dimensional data? In *International Conference on Artificial Intelligence and Statistics*, pp. 2158–2172. PMLR, 2022.

Joshua T Vogelstein, Eric W Bridgeford, Benjamin D Pedigo, Jaewon Chung, Keith Levin, Brett Mensh, and Carey E Priebe. Connectal coding: discovering the structures linking cognitive phenotypes to individual histories. *Current opinion in neurobiology*, 55:199–212, 2019.

Zitian Wu, Arkaprava Roy, and Leo L Duan. Graphical model-based inference on persistent homology. *arXiv preprint arXiv:2511.11996*, 2025.

Hang Zhang, Kaifeng Zhang, Kai Ming Ting, and Ye Zhu. Towards a persistence diagram that is robust to noise and varied densities. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 41952–41972, 2023.

Shijie C Zheng, Genevieve Stein-O'Brien, Jonathan J Augustin, Jared Slosberg, Giovanni A Carosso, Briana Winer, Gloria Shin, Hans T Bjornsson, Loyal A Goff, and Kasper D Hansen. Universal prediction of cell-cycle position using transfer learning. *Genome biology*, 23(1):41, 2022.

A Background

A.1 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) Murtagh & Contreras (2012) is a class of clustering algorithms that recursively merge data points or clusters of points based on a distance metric. The result is a hierarchy of clusters, often represented as a dendrogram. AHC is fundamentally different from divisive hierarchical clustering, which starts with all data points in a single cluster and recursively splits them.

Formally, AHC begins with each data point as an individual cluster. At each step, it merges the two closest clusters based on a linkage criterion until only one cluster remains or a stopping criterion is met. The linkage criterion determines the distance between clusters, guiding the merging process. Common linkage criteria include single linkage (minimum distance), complete linkage (maximum distance), and average linkage.

A.2 Persistent Homology

Let $g : \mathcal{X} \to \mathbb{R}^+$ denote a function on space $\mathcal{X}$, where $\mathbb{R}^+$ stands for positive real numbers, $\mathcal{X}$ is a bounded and open subset of Euclidean space. Function g is computed based on a finite point cloud X sampled from manifold $\mathcal{M} \subset \mathcal{X}$. This manifold assumption means that the finite point cloud X is sampled from an underlying k-dimensional manifold $\mathcal{M}$, where $\mathcal{M} \subset \mathbb{R}^d$ is a smooth, low-dimensional subspace embedded in a higher-dimensional Euclidean space $\mathbb{R}^d$. At scale $\eta \geq 0$, the sublevel set $\mathcal{X}_\eta^g = \{x \in \mathcal{X} \mid g(x) \leq \eta\}$ encodes the topological information in $\mathcal{X}$. For $\gamma \leq \eta$, we can have the nested sublevel sets $\mathcal{X}_\gamma^g \subseteq \mathcal{X}_\eta^g$. By increasing scale η from 0, we obtain a filtration $\mathcal{K}(X) = \{\mathcal{X}_\eta^g\}_{0 \leq \eta < \infty}$, a nested sequence of topological spaces.

A cycle is considered 'born' at $b \in \mathbb{R}$ when it first emerges in $\mathcal{X}_b^g$, and it 'dies' at $d \in \mathbb{R}$ when it ceases to exist in $\mathcal{X}_p^g$ for any $p > d$. 0-dimensional cycles are connected components; 1-dimensional cycles are rings or loops; 2-dimensional cycles are voids, etc. This homology dimension is denoted as q. Topological feature $r = (b, d)$, which corresponds to a cycle, is presented in the form of PD $\mathcal{D} = \{r_i = (b_i, d_i) \in \Omega | 1 \leq i \leq N(\mathcal{D})\}$, where $N(\mathcal{D})$ is the number of topological features. An example using Rips filtration (Hausmann et al., 1995) $g(\cdot) = 2 \min_{x \in X} \ell(\cdot, x)$, where ℓ is Euclidean distance, and corresponding PD are provided in Figures 9. A PD can be equivalently expressed as Persistence Barcode, as shown in Figure 3.

A.2.1 Loop Detection Score

We show an example of Loop Detection Score in Figure 10.

B Proofs & Additional Theoretical Analysis

B.1 Proof of Theorem 4.3

Theorem 4.3: Population $d_{\text{AGD-SS}}(\cdot, \cdot)$ is a valid distance metric.

Proof. First we prove the case for each random subset $S \subset \mathcal{X}$. Let $S \subset \mathcal{X}$ be the set of leaves in the dendrogram built by AHC with single linkage function. The distance between $x, y \in \mathcal{X}$ is defined as

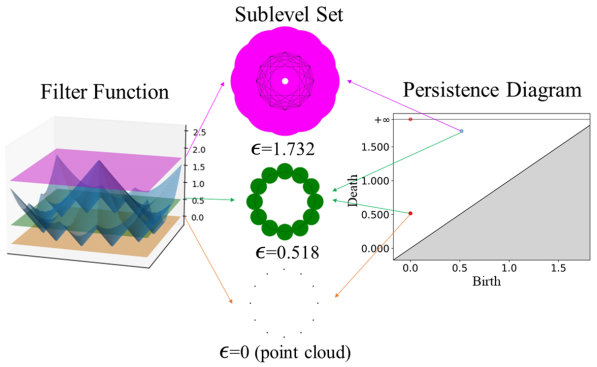


Figure 9: Example filter function, three sublevel sets and a PD. This is a direct reuse of Figure 1 in Zhang et al., 2023.

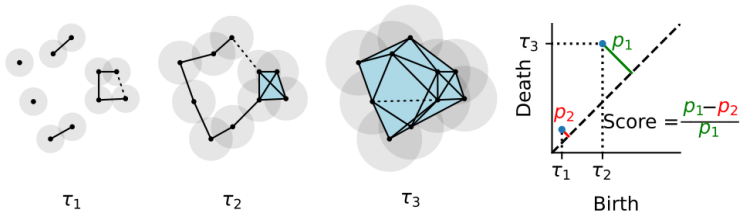


Figure 10: Illustration of applying Persistent Homology to a noisy circle and computing Loop Detection Score. This is directly plotted via this script of Damrich et al., 2024. p_i indicates persistence value. The Loop Detection Score for $l = 1$ captures the relative gap $(p_1 - p_2)/p_1$, quantifying the prominence of the main loop.

$d(x, y|S) = \sum_{w \in \mathcal{V}_{\text{LCA}(\phi_S(x), \phi_S(y))}} h(w)$, where $\text{LCA}(\phi_S(x), \phi_S(y))$ is the lowest common ancestor of $\phi_S(x)$ (x's nearest neighbor in S) and $\phi_S(y)$ (y's nearest neighbor in S) in $\mathcal{T}_S$, $\mathcal{V}_u$ is the set of nodes in the subtree rooted at u, and $h(w) \geq 0$ is the height (single linkage value) of node w.

- **Non-negativity:** Since the linkage height $h(w)$ is non-negative for all w, the sum over any subtree is non-negative. Thus, $d(x, y|S) \geq 0$ for all x, y.
- **Identity:** For $x = y$, $\text{LCA}(\phi_S(x), \phi_S(x)) = \phi_S(x)$ (the leaf itself). By the definition of AHC, leaves have height 0 (representing original points). Thus, $d(x, x|S) = h(x) = 0$. Conversely, if $x \neq y$, then $\text{LCA}(\phi_S(x), \phi_S(y))$ is an internal node corresponding to a merge event at some height $h > 0$ (assuming all points are distinct from each other) when $\phi_S(x) \neq \phi_S(y)$ (This will happen with a positive probability. The same technique is used in Zhang et al., 2023 when proving Λ-distance is a metric.). The sum includes $h(\text{LCA}(\phi_S(x), \phi_S(y)))$, $d(x, y|S) > 0$. Therefore, it holds that
$$d_{\text{AGD-SS}}(x, y) = 0 \iff x = y.$$
- **Symmetry:** Since $\text{LCA}(\phi_S(x), \phi_S(y)) = \text{LCA}(\phi_S(y), \phi_S(x))$, the associated subtree $\mathcal{V}_{\text{LCA}}$ is identical. Hence, it holds that $d(x, y|S) = d(y, x|S)$.
- **Triangle Inequality:** Let $u_{xy} = \text{LCA}(\phi_S(x), \phi_S(y))$, $u_{yz} = \text{LCA}(\phi_S(y), \phi_S(z))$, and $u_{xz} = \text{LCA}(\phi_S(x), \phi_S(z))$. Hierarchical clustering trees satisfy the ultrametric property on the heights (single linkage values) of LCAs:
$$h(u_{xz}) \leq \max(h(u_{xy}), h(u_{yz})).$$
Let $F(u) = \sum_{w \in \mathcal{V}_u} h(w)$. Note that if u is a descendant of v, i.e, v is an ancestor of u in the dendrogram $\mathcal{T}_S$, then $\mathcal{V}_u \subset \mathcal{V}_v$ and thus $F(u) < F(v)$. From the ultrametric property, the subtree rooted at u_{xz} is a subset of (or equal to) the subtree rooted at the higher of u_{xy} and u_{yz}. Therefore, $d(x, z|S) = F(u_{xz}) \leq \max(F(u_{xy}), F(u_{yz}))$, i.e., it holds that
$$d(x, z|S) \leq \max(d(x, y|S), d(y, z|S)) \leq d(x, y|S) + d(y, z|S).$$

Thus, $d(\cdot,\cdot|S)$ is a valid metric. Since $d_{\text{AGD-SS}}$ is the expectation of $d(\cdot,\cdot|S)$ over S, $d_{\text{AGD-SS}}$ is also a valid metric. □

B.2 Proof of Theorem 4.5

Theorem 4.5. Let $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^{d_a}$ and $X' = \{x'_1, \ldots, x'_N\} \subset \mathbb{R}^{d_a}$ be paired point clouds such that $\|x_i - x'_i\|_2 \le \delta$ for every i, i.e., δ-perturbation. Let $x, y \in X$ be two fixed query points, and let $x', y' \in X'$ be their perturbed counterparts. Let S be sampled uniformly from all ψ-subsets of X, and let $S' \subset X'$ be the paired perturbed subsample. Recall our notation

$$d_{\text{AGD-SS}}(x, y \mid S) = \sum_{w \in V_{\text{LCA}_S(\phi_S(x),\phi_S(y))}} h_S(w),$$

and define $d_{\text{AGD-SS}}(x', y' \mid S')$ analogously on $T_{S'}$. AGD-SS distance is

$$d_{\text{AGD-SS}}(x, y) = \mathbb{E}_S[d_{\text{AGD-SS}}(x, y \mid S)]\,.$$

Let $\widetilde{T}_S$ be the tree obtained from T_S by attaching x as a leaf to $\phi_S(x)$ at height $r_S(x) = \|x - \phi_S(x)\|_2$ and attaching y as a leaf to $\phi_S(y)$ at height $r_S(y) = \|y - \phi_S(y)\|_2$. Define $\widetilde{T}_{S'}$ analogously for x', y', S'. On these augmented trees, define

$$\widetilde{d}_{\text{AGD-SS}}(x, y \mid S) = \sum_{w \in V_{\text{LCA}_{\widetilde{T}_S}(x,y)}} h_{\widetilde{T}_S}(w),$$

and define $\widetilde{d}_{\text{AGD-SS}}(x', y' \mid S')$ analogously. Assume there is an event $\mathcal{E}_S$ with $\mathbb{P}_S(\mathcal{E}_S) \ge 1 - \alpha$. $\mathcal{E}_S$ means that after adding query leaves x, y and x', y' to S and S', the query leaves first attach to their nearest representatives in S/S'. (This implies that the merge height multisets inside $V_{\text{LCA}_{\widetilde{T}_S}(x,y)}$ and $V_{\text{LCA}_{\widetilde{T}_{S'}}(x',y')}$ admit a matching under which every matched height differs by at most 2δ.) We conduct validation on this event in Appendix B.2.1.

Finally, assume the one-subsample AGD-SS values are uniformly bounded for evary S:

$$0 \le d_{\text{AGD-SS}}(x, y \mid S), d_{\text{AGD-SS}}(x', y' \mid S') \le B_\psi.$$

Then it holds that

$$|d_{\text{AGD-SS}}(x, y) - d_{\text{AGD-SS}}(x', y')| \le 2(\psi + 3)\delta + \alpha B_\psi.$$

Moreover, for t independent subsamples $S_1, \ldots, S_t$, with probability at least $1 - \rho$,

$$\left| \frac{1}{t} \sum_{\ell=1}^{t} d_{\text{AGD-SS}}(x, y \mid S_\ell) - \frac{1}{t} \sum_{\ell=1}^{t} d_{\text{AGD-SS}}(x', y' \mid S'_\ell) \right| \le 2(\psi + 3)\delta + \alpha B_\psi + 2B_\psi \sqrt{\frac{\log(4/\rho)}{2t}}.$$

Proof. Fix a subsample S and work on the event $\mathcal{E}_S$. Since $\widetilde{T}_S$ has at most $\psi + 2$ leaves, any LCA subtree in $\widetilde{T}_S$ contains at most $\psi + 1$ nonzero merge heights. By the event $\mathcal{E}_S$,

$$\left| \widetilde{d}_{\text{AGD-SS}}(x, y \mid S) - \widetilde{d}_{\text{AGD-SS}}(x', y' \mid S') \right| \le 2(\psi + 1)\delta.$$

This is the precise use of the 0-dimensional PH stability intuition: under a paired δ-perturbation of points, pairwise distances change by at most 2δ, so matched single-linkage death times change by at most 2δ. We apply this only to the matched merge heights inside the two LCA subtrees; we do not assume the entire dendrogram has a fixed combinatorial structure.

Let

$$r_S(x) = \|x - \phi_S(x)\|_2, \qquad r_{S'}(x') = \|x' - \phi_{S'}(x')\|_2.$$

The paired perturbation implies

$$|r_S(x) - r_{S'}(x')| \le 2\delta.$$

Indeed, if $a = \phi_S(x)$ and a' is its paired point in S', then

$$r_{S'}(x') \le \|x' - a'\|_2 \le \|x' - x\|_2 + \|x - a\|_2 + \|a - a'\|_2 \le r_S(x) + 2\delta,$$

and the reverse inequality follows by symmetry. The same argument gives $|r_S(y) - r_{S'}(y')| \le 2\delta$.

By construction of $\widetilde{T}_S$, the augmented subtree sum is the projected AGD-SS subtree sum plus the two query leaf attachment heights:

$$\widetilde{d}_{\text{AGD-SS}}(x, y \mid S) = d_{\text{AGD-SS}}(x, y \mid S) + r_S(x) + r_S(y),$$

and analogously for (x', y', S'). Hence, on event $\mathcal{E}_S$,

$$\begin{aligned}
&|d_{\text{AGD-SS}}(x, y \mid S) - d_{\text{AGD-SS}}(x', y' \mid S')| \\
&\quad \le \left|\widetilde{d}_{\text{AGD-SS}}(x, y \mid S) - \widetilde{d}_{\text{AGD-SS}}(x', y' \mid S')\right| + |r_S(x) - r_{S'}(x')| + |r_S(y) - r_{S'}(y')| \\
&\quad \le 2(\psi + 1)\delta + 2\delta + 2\delta = 2(\psi + 3)\delta.
\end{aligned}$$

On $\mathcal{E}_S^c$, the uniform boundedness assumption gives

$$|d_{\text{AGD-SS}}(x, y \mid S) - d_{\text{AGD-SS}}(x', y' \mid S')| \le B_\psi.$$

Taking expectation over the uniform random subsample S yields

$$\begin{aligned}
&|d_{\text{AGD-SS}}(x, y) - d_{\text{AGD-SS}}(x', y')| \\
&\quad \le \mathbb{E}_S\, |d_{\text{AGD-SS}}(x, y \mid S) - d_{\text{AGD-SS}}(x', y' \mid S')| \\
&\quad \le (1 - \alpha)2(\psi + 3)\delta + \alpha B_\psi \le 2(\psi + 3)\delta + \alpha B_\psi.
\end{aligned}$$

For the empirical estimator, Hoeffding's inequality gives

$$\Pr\left[\left|\frac{1}{t}\sum_{\ell=1}^{t} d_{\text{AGD-SS}}(x, y \mid S_\ell) - d_{\text{AGD-SS}}(x, y)\right| \ge u\right] \le 2\exp\left(-\frac{2tu^2}{B_\psi^2}\right).$$

The same bound holds for (x', y'). Set $u = B_\psi\sqrt{\log(4/\rho)/(2t)}$ and take a union bound over the two empirical means. Combining this empirical deviation with the population bound proves the result.

□

B.2.1 Validation of Event $\mathcal{E}_S$

We empirically test the event $\mathcal{E}_S$ used in Theorem 4.5 on Swiss roll, S-curve, and Sphere. For each dataset, we sample query pairs (x, y) and random subsamples S with $\psi = 32$. We then perturb every data point by a random displacement with norm at most δ, construct the paired perturbed subsample S', and compare the selected AGD-SS LCA subtree before and after perturbation. A trial is counted as satisfying the event $\mathcal{E}_S$ if the selected LCA-subtree merge-height multisets in $\widetilde{T}_S$ and $\widetilde{T}_{S'}$ have the same cardinality and can be matched with per-height error at most 2δ.

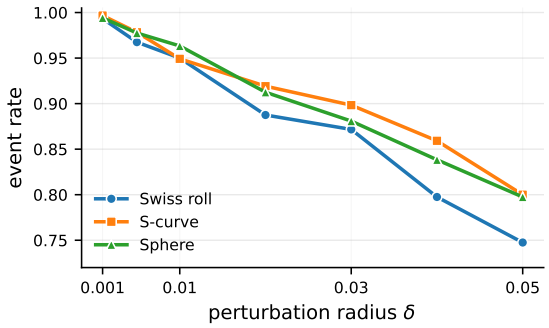


Figure 11: Empirical validation of the stable LCA-subtree event $\mathcal{E}_S$ in Theorem 4.5. The horizontal axis is the perturbation radius δ, and the vertical axis is the fraction of random query-pair/subsample trials for which the selected LCA-subtree merge-height multisets before and after perturbation can be matched with per-height error at most 2δ. Higher values indicate that the conditional stability event required by the theorem holds more frequently under bounded perturbations.

As shown in Figure 11, the stable LCA-subtree event holds with high probability for small perturbations on all three synthetic manifolds. The event rate is close to one for $\delta = 0.001$ and remains above 0.88 for all datasets at $\delta = 0.02$, which is the perturbation range most relevant to the small noise stability regime. As δ increases further, the event rate decreases, especially on Swiss roll. The overall validation gives a rough estimation $\alpha \approx 0.25$ on these datasets.

B.3 Proof of Geodesic Forest's Metricity

We prove that population GF is a metric using the same proof process in Zhang et al., 2023, i.e., Λ-distance. GF partitions the ambient space using an unsupervised decision tree. For a given point x in a specific partition (i.e., a leaf node), the distance to any other point is set to 0 if the point lies within the same leaf node, and 1 otherwise. Λ-distance is very similar to GF, except that the partition uses random Voronoi Diagram instead of the leaf nodes in unsupervised decision tree. We begin with some notation: the distance based on one unsupervised decision tree T is denoted as $d(\cdot, \cdot|T)$, the population distance based on Geodesic Forest is then $d_{\mathrm{GF}}(\cdot, \cdot) = \mathbb{E}_T[d(\cdot, \cdot \mid T)]$, with empirical estimator $\hat{d}_{\mathrm{GF}}(\cdot, \cdot) = \frac{1}{t}\sum_{i=1}^{t} d(\cdot, \cdot|T_i)$

Proof. We prove d_{GF} is a distance metric from four aspects:

- **Identity:** It is trivial that the distance from a point to itself is zero: $d(x, x|T) = 0$. So if $x = y$, $d_{\mathrm{GF}}(x, y) = 0$. If $x \neq y$, then there exists a random unsupervised decision tree T that has x, y are in different leaf nodes with positive probability, i.e., $d(x, y|T) = 1$. Therefore, it holds that
$$d_{\mathrm{GF}}(x, y) = 0 \iff x = y.$$
- **Non-negativity:** The distance between two distinct points x, y is always positive since $d(\cdot, \cdot|T_i)$ is either 0 or 1. Hence $d_{GF} > 0$.
- **Symmetry:** It is trivial that the distance from x to y is always the same as the distance from y to x: $d_{\mathrm{GF}}(x, y) = d_{\mathrm{GF}}(y, x)$.
- **Triangle Inequality**: $d_{\mathrm{GF}}(x, y)$ can be re-expressed as
$$d_{\mathrm{GF}}(x, y) = \frac{1}{2t}\sum_{k=1}^{t} d'(x, y|T_i),$$
where $d'(x, y|T_i) = \|\Phi(x|T_i) - \Phi(y|T_i)\|_2^2$. $\Phi(x|T) = [0, 0, ..., 1, .., 0, 0]$ is a vector in one-hot form, its length is the number of leaf nodes in T_i. Giving any fixed order of these leaf nodes, 1 lies in index of the leaf node which contains x.

  So it suffices to prove that triangle inequality holds for $d'(x, y|T)$. The positions of x, y, z with respect to the leaf nodes of T can be divided into 4 cases:

  a) x, y, z are in the same leaf node θ_i: $d'(x, y|T) = d'(x, z|T) = d'(y, z|T) = 0$. Hence $d'(x, z|T) \leq d'(x, y|T) + d'(y, z|T)$ holds.

  b) x, z are in the same leaf node θ_i; and y is in a different leaf node θ_j: $d'(x, z|T) = 0$, $d'(x, y|T) = d'(y, z|T) = 2$. Hence $d'(x, z|T) \leq d'(x, y|T) + d'(y, z|T)$ holds.

  c) x, z are in different leaf nodes θ_i and θ_j, respectively; and y is in either θ_i or θ_j: $d'(x, z|T) = 2$, $d'(x, y|T) = 0$ (or 2) and $d'(y, z|T) = 2$ (or 0). Hence $d'(x, z|T) \leq d'(x, y|T) + d'(y, z|T)$ holds.

  d) x, y, z are in three different leaf nodes θ_i, θ_j and θ_k, respectively: $d'(x, y|T) = d'(x, z|T) = d'(y, z|T) = 2$. Hence $d'(x, z|T) \leq d'(x, y|T) + d'(y, z|T)$ holds.

□

B.4 Analysis on the ability of AGD-SS to reflect geodesic rank

B.4.1 Reminder of the AGD-SS Notation

Let

$$X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^{d_a}$$

be the observed dataset. We assume that the clean signal lies on a latent d_i-dimensional Riemannian manifold $\mathcal{M}^{d_i}$ with geodesic distance $d_{\mathcal{M}}$, where $d_i \ll d_a$. This follows the notation of the main paper.

Let $\psi \in \mathbb{N}$ be the AGD-SS subsample size. For a random subset $S \subset X$ with $|S| = \psi$, let T_S be the single-linkage AHC dendrogram on S, and let $h(w) \geq 0$ be the linkage height of a dendrogram node w. Define the nearest-neighbor projection

$$\phi_S : X \to S, \qquad \phi_S(x) = \arg\min_{y \in S} \|x - y\|.$$

For $x, y \in X$, let

$$u_S(x, y) = \mathrm{LCA}_{T_S}(\phi_S(x), \phi_S(y))$$

be the lowest common ancestor of the projected points in T_S, and let $V_{u_S(x,y)}$ be the set of nodes in the subtree rooted at $u_S(x, y)$.

Definition B.1 (One-subsample and population AGD-SS)**.** Following the original AGD-SS definition, define

$$d(x, y \mid S) = \sum_{w \in V_{u_S(x,y)}} h(w).$$

The population AGD-SS distance is

$$d_{\text{AGD-SS}}(x, y) = \mathbb{E}_S[d(x, y \mid S)].$$

Given t independent random subsets $S_1, \ldots, S_t$, the empirical AGD-SS distance is

$$\widehat{d}_{\text{AGD-SS}}(x, y) = \frac{1}{t} \sum_{\ell=1}^{t} d(x, y \mid S_\ell).$$

B.4.2 Persistence Interpretation

Lemma B.2 (AGD-SS subtree sum as local zero-dimensional persistence (Wu et al., 2025))**.** *For a fixed subsample S, the single-linkage merge heights inside the subtree rooted at $u_S(x, y)$ are the finite death times of zero-dimensional persistent homology computed on the leaves of that subtree. Therefore,*

$$d(x, y \mid S) = \sum_{w \in V_{u_S(x,y)}} h(w)$$

is the sum of zero-dimensional persistence inside that LCA subtree, up to the deterministic tie-breaking convention used by single linkage.

B.4.3 Global Assumptions on a Queried Radius Range

Fix a query point $x_q \in X$ and a finite candidate set $\mathcal{C}_q \subset X$. For $z \in \mathcal{C}_q$, define

$$r_z = d_{\mathcal{M}}(x_q, z).$$

Let $I_q \subset [0, \infty)$ be the radius range on which rank preservation is claimed. The assumptions below are stated in a global form on I_q. They therefore cover both local and global geodesic retrieval by changing only the choice of I_q:

$$\text{local case: } I_q = [0, r_0] \text{ with } r_0 < \mathrm{inj}(\mathcal{M}),$$

and

$$\text{global case: } I_q = [0, R_q] \text{ or any larger queried radius range,}$$

where $R_q = \sup_{x \in \mathcal{M}} d_{\mathcal{M}}(x_q, x)$. Here $\mathrm{inj}(\mathcal{M})$ denotes the injectivity radius of $\mathcal{M}$, i.e., a radius below which geodesic balls are locally well-behaved and avoid cut-locus effects.

Assumption B.3 (Global volume rank transform on I_q)**.** There is a deterministic function

$$V_q : I_q \to \mathbb{R}_{\geq 0}$$

that is strictly increasing on the radii appearing in the queried comparisons. In the geometric case,

$$V_q(r) = \mathrm{Vol}(B_{\mathcal{M}}(x_q, r)).$$

For local radii below the injectivity radius,

$$V_q(r) = \omega_{d_i} r^{d_i}(1 + O(r^2)).$$

For global radii, $V_q(r)$ is the full geodesic ball volume function; the increase requirement excludes radius pairs whose geodesic balls have the same volume. An illustration is given in Figure 12.

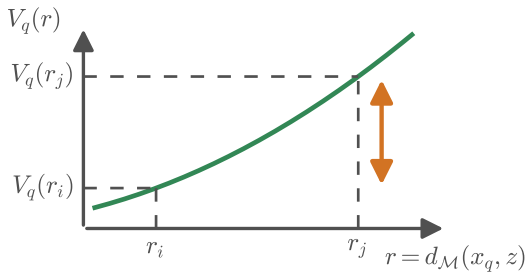


Figure 12: Illustration of Assumption B.3. The volume rank transform $V_q(r)$ is strictly increasing on the queried radius range I_q, so a geodesic margin $r_i < r_j$ induces a positive transformed margin $V_q(r_j) - V_q(r_i) > 0$.

Assumption B.4 (Global persistence-volume approximation on I_q)**.** There exist constants $a_N > 0$, $c_{d_i} > 0$, and an error $\xi_N \geq 0$ such that, for every $z \in \mathcal{C}_q$ with $r_z \in I_q$,

$$\left| a_N^{-1} d_{\text{AGD-SS}}(x_q, z) - c_{d_i} V_q(r_z) \right| \leq \xi_N.$$

Validation of Assumptions. The results in Figure 13 show that AGD-SS behaves consistently with the margin-dependent rank preservation theory on S-curve and Sphere, while Swiss roll is a harder global case. The empirical volume transform is monotone with respect to the ground-truth geodesic radius on all three datasets, supporting its role as a rank transform in Assumption B.3. AGD-SS also shows a clear monotone association with this volume transform, with Spearman correlations of 0.82 on S-curve and 0.95 on sphere for the queries, and lower but still positive association on Swiss roll. This supports Assumption B.4. More importantly, the pairwise rank violation rate decreases as the geodesic margin increases, indicating that AGD-SS is most reliable when candidate points are separated by a sufficient intrinsic distance gap. This matches the theorem's margin condition: AGD-SS does not need to resolve arbitrarily small geodesic gaps, but it should preserve order once the transformed margin dominates approximation and sampling errors. The weaker Swiss-roll behavior suggests that global rank preservation should be stated as conditional on the queried radius range and margin size, rather than as an unconditional guarantee.

B.4.4 Rank Preservation Theorem

Theorem B.5 (Unified persistence-based geodesic rank preservation)**.** *Suppose Assumptions B.3 and B.4 hold. Let $i, j \in \mathcal{C}_q$ satisfy*

$$r_i = d_{\mathcal{M}}(x_q, i) < r_j = d_{\mathcal{M}}(x_q, j), \qquad r_i, r_j \in I_q.$$

If the transformed volume margin satisfies

$$c_{d_i}\left[V_q(r_j) - V_q(r_i)\right] > 2\xi_N,$$

then the population AGD-SS ranking is correct:

$$d_{\text{AGD-SS}}(x_q, i) < d_{\text{AGD-SS}}(x_q, j).$$

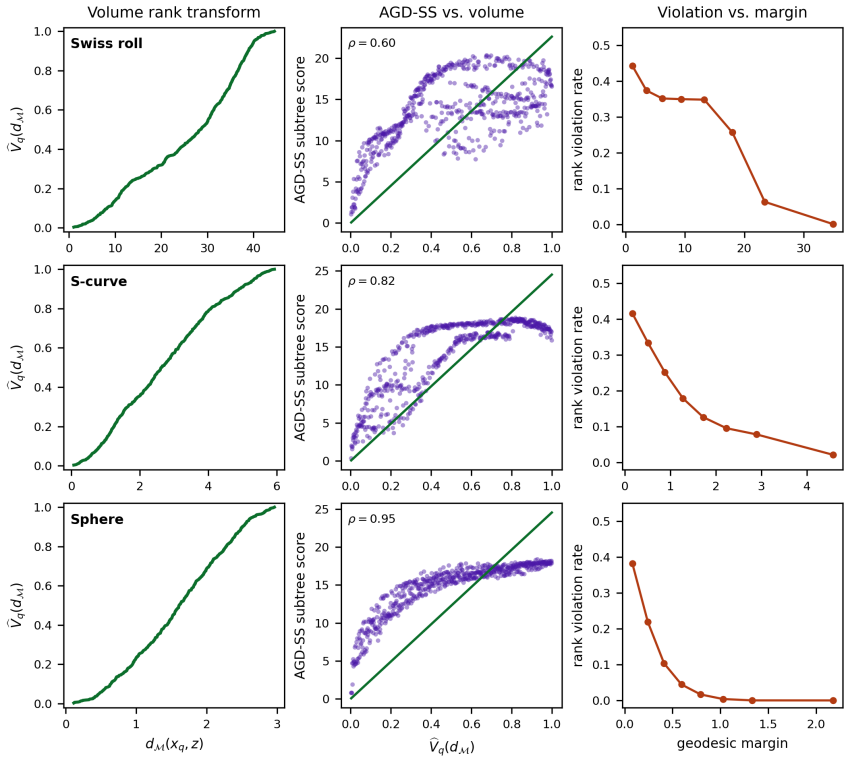


Figure 13: Empirical validation of the assumptions on Swiss roll, S-curve, and Sphere using $N = 600$, $\psi = 32$, $t = 100$, and 12 query points. Each row corresponds to one dataset. Left: the empirical volume transform $\widehat{V}_q(d_{\mathcal{M}}(x_q, z))$ is monotone in the ground-truth geodesic radius. Middle: the horizontal axis is $\widehat{V}_q(d_{\mathcal{M}}(x_q, z))$, and the vertical axis is the AGD-SS subtree score, i.e., the averaged single-linkage/MST LCA-subtree sum used by $d_{\text{AGD-SS}}(x_q, z)$ before any monotone rank interpretation; the reported ρ is the Spearman correlation between this score and the empirical volume transform for the displayed query. Right: the horizontal axis is the geodesic margin $d_{\mathcal{M}}(x_q, z_j) - d_{\mathcal{M}}(x_q, z_i)$ over pairs satisfying $d_{\mathcal{M}}(x_q, z_i) < d_{\mathcal{M}}(x_q, z_j)$, and the vertical axis is the fraction of such pairs for which AGD-SS reverses the order.

Proof. By Assumption B.4,

$$a_N^{-1} d_{\text{AGD-SS}}(x_q, i) \le c_{d_i} V_q(r_i) + \xi_N$$

and

$$a_N^{-1} d_{\text{AGD-SS}}(x_q, j) \ge c_{d_i} V_q(r_j) - \xi_N.$$

The margin condition gives

$$c_{d_i} V_q(r_i) + \xi_N < c_{d_i} V_q(r_j) - \xi_N.$$

Therefore,

$$a_N^{-1} d_{\text{AGD-SS}}(x_q, i) < a_N^{-1} d_{\text{AGD-SS}}(x_q, j).$$

Since $a_N > 0$, multiplying by a_N preserves the inequality. □

Corollary B.6 (Empirical AGD-SS rank preservation)**.** *Suppose Assumptions B.3 and B.4 hold. Let $i, j \in \mathcal{C}_q$ satisfy*

$$r_i = d_{\mathcal{M}}(x_q, i) < r_j = d_{\mathcal{M}}(x_q, j), \qquad r_i, r_j \in I_q.$$

Suppose also that, for $z \in \{i, j\}$ and every subsample S,

$$0 \le d(x_q, z \mid S) \le B_N.$$

Then, over the independent subsamples $S_1, \dots, S_t$ used in $\widehat{d}_{\text{AGD-SS}}$, with probability at least $1 - \delta$,

$$\widehat{d}_{\text{AGD-SS}}(x_q, i) < \widehat{d}_{\text{AGD-SS}}(x_q, j)$$

whenever

$$c_{d_i}\big[V_q(r_j) - V_q(r_i)\big] > 2\xi_N + 2a_N^{-1}B_N\sqrt{\frac{\log(4/\delta)}{2t}}.$$

Proof. For a fixed $z \in \{i, j\}$, define

$$Y_\ell(z) = d(x_q, z \mid S_\ell), \qquad \ell = 1, \dots, t.$$

The subsets $S_1, \dots, S_t$ are independent, hence $Y_1(z), \dots, Y_t(z)$ are independent random variables. By the definition of the population and empirical AGD-SS scores,

$$\mathbb{E}\, Y_\ell(z) = d_{\text{AGD-SS}}(x_q, z), \qquad \widehat{d}_{\text{AGD-SS}}(x_q, z) = \frac{1}{t}\sum_{\ell=1}^{t} Y_\ell(z).$$

The assumed boundedness gives $0 \le Y_\ell(z) \le B_N$. Therefore Hoeffding's inequality gives, for every $u > 0$,

$$\Pr\Big(\Big|\widehat{d}_{\text{AGD-SS}}(x_q, z) - d_{\text{AGD-SS}}(x_q, z)\Big| \ge u\Big) \le 2\exp\left(-\frac{2tu^2}{B_N^2}\right).$$

Choose

$$u_\delta = B_N\sqrt{\frac{\log(4/\delta)}{2t}}.$$

Then $2\exp(-2tu_\delta^2/B_N^2) = \delta/2$. Applying the bound to $z = i$ and $z = j$ and taking a union bound shows that the event

$$\mathcal{E} = \Big\{\Big|\widehat{d}_{\text{AGD-SS}}(x_q, z) - d_{\text{AGD-SS}}(x_q, z)\Big| \le u_\delta \text{ for both } z = i, j\Big\}$$

has probability at least $1 - \delta$.

On $\mathcal{E}$,

$$a_N^{-1}\widehat{d}_{\text{AGD-SS}}(x_q, i) \le a_N^{-1}d_{\text{AGD-SS}}(x_q, i) + a_N^{-1}u_\delta$$

and

$$a_N^{-1}\widehat{d}_{\text{AGD-SS}}(x_q, j) \ge a_N^{-1}d_{\text{AGD-SS}}(x_q, j) - a_N^{-1}u_\delta.$$

By Assumption B.4,

$$a_N^{-1}d_{\text{AGD-SS}}(x_q, i) \le c_{d_i}V_q(r_i) + \xi_N$$

and

$$a_N^{-1}d_{\text{AGD-SS}}(x_q, j) \ge c_{d_i}V_q(r_j) - \xi_N.$$

Combining these inequalities on $\mathcal{E}$ gives

$$a_N^{-1}\widehat{d}_{\text{AGD-SS}}(x_q, i) \le c_{d_i}V_q(r_i) + \xi_N + a_N^{-1}u_\delta$$

and

$$a_N^{-1}\widehat{d}_{\text{AGD-SS}}(x_q, j) \ge c_{d_i}V_q(r_j) - \xi_N - a_N^{-1}u_\delta.$$

The stated margin condition is exactly

$$c_{d_i}V_q(r_i) + \xi_N + a_N^{-1}u_\delta < c_{d_i}V_q(r_j) - \xi_N - a_N^{-1}u_\delta.$$

Hence, on $\mathcal{E}$,

$$a_N^{-1}\widehat{d}_{\text{AGD-SS}}(x_q, i) < a_N^{-1}\widehat{d}_{\text{AGD-SS}}(x_q, j).$$

Since $a_N > 0$, multiplying both sides by a_N yields

$$\widehat{d}_{\text{AGD-SS}}(x_q, i) < \widehat{d}_{\text{AGD-SS}}(x_q, j).$$

Because $\Pr(\mathcal{E}) \ge 1 - \delta$, the claim follows. □

Examples on the relation between geodesic margin and geodesic rank preservation are given in the rightmost column in Figure 13.

Corollary B.7 (Top-k retrieval)**.** *Let $\mathcal{N}_{\mathrm{true}}(x_q, k)$ be the k candidates in $\mathcal{C}_q$ with smallest true geodesic distances to x_q, where $1 \le k < |\mathcal{C}_q|$. Define the transformed top-k margin*

$$\Gamma_k(x_q) = \min_{\substack{i \in \mathcal{N}_{\mathrm{true}}(x_q,k) \\ j \in \mathcal{C}_q \setminus \mathcal{N}_{\mathrm{true}}(x_q,k)}} c_{d_i}\big[V_q(r_j) - V_q(r_i)\big].$$

Assume that $r_i, r_j \in I_q$ for every boundary pair $i \in \mathcal{N}_{\mathrm{true}}(x_q, k)$ and $j \in \mathcal{C}_q \setminus \mathcal{N}_{\mathrm{true}}(x_q, k)$. If

$$\Gamma_k(x_q) > 2\xi_N,$$

then population AGD-SS recovers the true geodesic top-k set. If the boundedness condition in Corollary B.6 also holds, then empirical AGD-SS recovers the true geodesic top-k set with probability at least $1 - \delta$ whenever

$$\Gamma_k(x_q) > 2\xi_N + 2a_N^{-1} B_N \sqrt{\frac{\log(4M_k/\delta)}{2t}}, \qquad M_k = k\big(|\mathcal{C}_q| - k\big).$$

Proof. Let

$$\mathcal{P}_k = \{(i, j) : i \in \mathcal{N}_{\mathrm{true}}(x_q, k),\ j \in \mathcal{C}_q \setminus \mathcal{N}_{\mathrm{true}}(x_q, k)\}$$

be the set of boundary comparisons, so $|\mathcal{P}_k| = M_k$. For every $(i, j) \in \mathcal{P}_k$, the definition of $\Gamma_k(x_q)$ gives

$$c_{d_i}\big[V_q(r_j) - V_q(r_i)\big] \ge \Gamma_k(x_q).$$

If $\Gamma_k(x_q) > 2\xi_N$, then the margin condition in Theorem B.5 holds for every boundary pair. Hence

$$d_{\text{AGD-SS}}(x_q, i) < d_{\text{AGD-SS}}(x_q, j) \qquad \text{for all } (i, j) \in \mathcal{P}_k.$$

This strict separation means that every member of $\mathcal{N}_{\mathrm{true}}(x_q, k)$ has smaller population AGD-SS score than every candidate outside $\mathcal{N}_{\mathrm{true}}(x_q, k)$. Therefore the set of the k smallest population AGD-SS scores is exactly $\mathcal{N}_{\mathrm{true}}(x_q, k)$; possible ties within the true top-k set or within its complement do not affect the recovered set.

For the empirical statement, apply the proof of Corollary B.6 to each fixed boundary pair, but allocate failure probability δ/M_k to that pair. Equivalently, Hoeffding's inequality and a union bound over the $2M_k$ candidate deviations imply that, with probability at least $1 - \delta$,

$$\left|\widehat{d}_{\text{AGD-SS}}(x_q, z) - d_{\text{AGD-SS}}(x_q, z)\right| \le B_N \sqrt{\frac{\log(4M_k/\delta)}{2t}}$$

for every candidate z that appears in a boundary comparison. On this event, the empirical version of the margin argument in Corollary B.6 holds simultaneously for all $(i, j) \in \mathcal{P}_k$ when the empirical margin condition is satisfied. Thus every true top-k candidate has smaller empirical AGD-SS distance than every non-top-k candidate, and the empirical top-k set equals $\mathcal{N}_{\mathrm{true}}(x_q, k)$. □

C Limitation and Future Work

While AGD-SS demonstrates robust performance in preserving both geometry (geodesic rank) and topology (persistent homology), several avenues for improvement and exploration remain.

C.1 Limitation

Computational Complexity. The reliance on Agglomerative Hierarchical Clustering (AHC) implies a computational complexity that can be prohibitive for very large datasets. Although our subsampling strategy and NN projection mitigate this by operating on smaller subsets, scaling to massive datasets may require more efficient approximations or scalable hierarchical clustering algorithms.

Hyperparameter Sensitivity. The performance of AGD-SS can be influenced by hyperparameters such as the subsample size and the number of trees in the ensemble. While we found robust default settings, a more systematic study of optimal hyperparameter selection based on dataset properties would be beneficial.

C.2 Future Work

Extension to Higher Order Topology. Currently, AGD-SS leverages 0-dimensional persistent homology (connected components) via the MST-PH equivalence. Extending this framework to capture topological features with higher homology dimension (e.g., voids) is a promising direction.

Theoretical Bounds. While we established stability under perturbation, deriving tighter bounds on the error of the geodesic rank itself with respect to the manifold curvature and sampling density would strengthen the theoretical foundation.

More PH Summary. We just use a plain PH summary: sum of persistence, in the theoretical and experimental analysis. We could use more PH summary as the attribution of two points, and this could enable the derived distance to be useful in tasks other than Geodesic Retrieval and Loop Detection.

D Experiment Settings

All methods used in the experiments were run on a machine with i5-12500 CPU, 16 GB RAM, Intel Iris Xe Graphics GPU and Ubuntu 20.04 system.

For Geodesic Retrieval task, in Figure 2: the default hyperparameter setting of AGD and AGD-SS is $\psi = 32, t = 100$. For GF, Isomap, UMAP, we use the same setting as that in the Geodesic Forest paper Madhyastha et al. (2020). For Diffusion distance, we built symmetric 10-NN graph on the dataset and choose diffusion steps $m = 2, 64, 256$. In Figure 4, for GF, UMAP and Isomap, we use the same setting as that in the Geodesic Forest paper Madhyastha et al. (2020). For Diffusion distance, given the dataset size (around 200), we still build symmetric 10-NN graph with diffusion step $m = 2, 4, 8, 16, 32, 64$ and choose the m with highest average geodesic precision over different ks. For AGD and AGD-SS, we choose $t = 200$ and $\psi = [16, 32, 64, 128]$. We choose the ψ with highest average geodesic precision over different ks.

For Loop Detection task, in Figure 5, for Diffusion distance, we built symmetric 15-NN graph with diffusion steps $m = 8$, which is the same as that in Damrich et al. (2024). For GF, we use the default setting given in the R package of GF (trees=100, LinearCombo=True, splitCrit=bicfast, sparsity=1/20 and min_parent=100). For AGD and AGD-SS, we choose $\psi = 4, t = 100$. The same goes for Figure 6, where the hyperparameter setting of the rest methods are the same as those in Damrich et al., 2024. For Figure 7 and 8, for AGD and AGD-SS, we choose $t = 100$ and $\psi = [4, 8, 16, 32, 64, 128]$. For GF, we use the default setting given in the R package of GF (trees=100, LinearCombo=True, splitCrit=bicfast, sparsity=1/20 and min_parent=100). The rest methods use the same setting as that in Damrich et al., 2024.

Code is in the Supplementary Material.

E Supplementary Experiments

E.1 Examples of AHC dendrograms

We sample $N = 300$ points from each manifold (Helix, S-Curve, Sphere, Swiss Roll) and build a single AHC tree via random subsampling with subsample size $\psi = 32$ and single linkage. Each leaf represents one of the 32 randomly subsampled data points. Leaf nodes are coloured by their *geodesic position*: arc length for Helix; first intrinsic coordinate for S-Curve and Swiss Roll; azimuthal angle for Sphere, normalised to $[0, 1]$ and mapped to a viridis colour scale.

The results are shown in 14. For Helix and S-Curve, dendrograms show that geodesically nearby points are merged at low heights, while geodesically distant points are only joined at higher levels. Sphere and Swiss Roll show broader colour mixing at higher merge levels, but local geodesic clusters are still clearly visible within subtrees.

E.2 Stability under noise

For Euclidean d, the triangle inequality $|d(x, y) - d(x', y')| \leq 2\delta$ gives when points move by at most δ. However, this is not the robustness notion used in our paper. In Section 2, robustness is defined

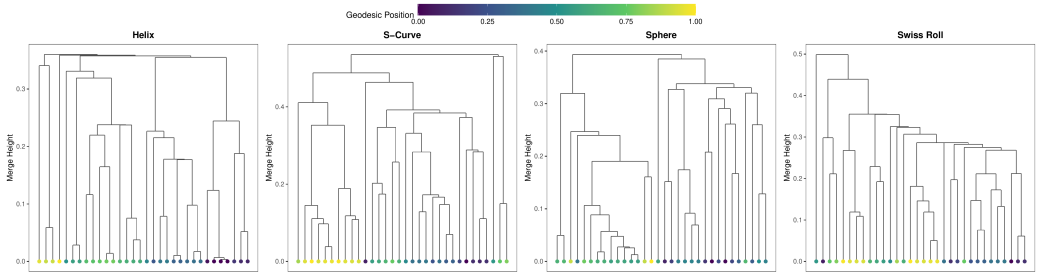


Figure 14: **Single-trial AHC dendrograms for four artificial datasets.**

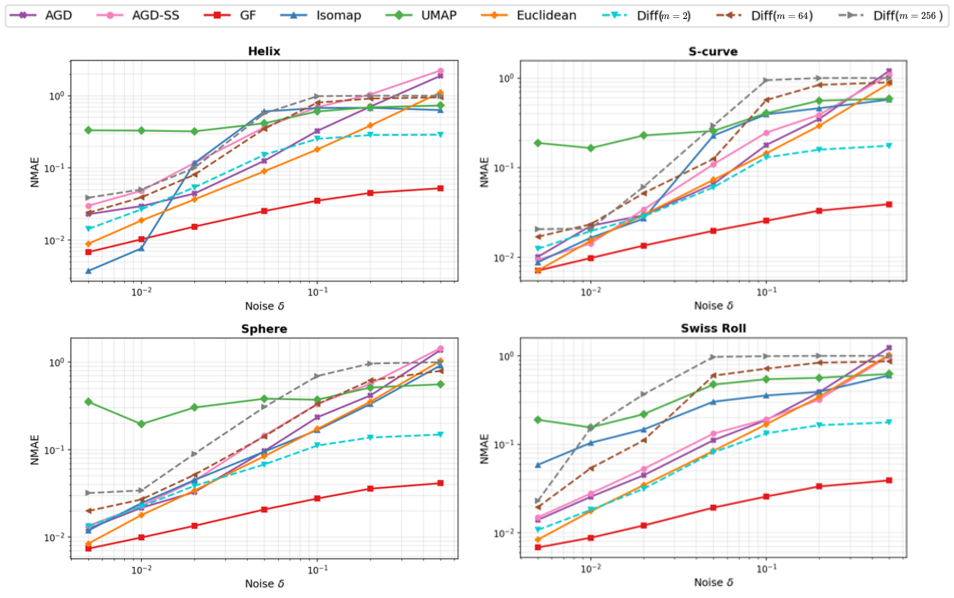


Figure 15: Normalized Mean Absolute Error (NMAE) of pairwise distance estimates under additive Gaussian noise on the four synthetic manifolds from the paper (Helix, Sphere, S-curve, Swiss Roll; $N = 500$ points each). For each noise level δ in $\{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$, independent Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \delta^2 I)$ is added to every point, all pairwise distances are re-estimated, and NMAE is computed relative to the clean distance matrix. Both axes are on a log scale.

at the task level: an effective RGL method should preserve geodesic rank and manifold topology, and also maintain its performance under noise contamination (Properties 1, 2, and 4). Under this definition, Euclidean can be Lipschitz-stable yet still non-robust for RGL: its pairwise distances change little under perturbation, but on curved manifolds they are not good geodesic estimators, so the induced topology and retrieval geodesic neighbors remain wrong.

The non-trivial problem is therefore not whether a fixed metric changes smoothly, but whether a data-dependent learned metric that is geodesically accurate before perturbation remains accurate after perturbation. Our stability theorem 4.5 provides exactly this missing guarantee, establishing that the geodesically accurate values are preserved under perturbation.

To validate this, we perturb the four paper datasets with Gaussian noise and measure the Normalized Mean Absolute Error (NMAE) between perturbed and clean distance matrices:

$$\mathrm{NMAE} = \frac{\sum_{i<j} |\hat{d}_\delta(i,j) - \hat{d}_0(i,j)|}{\sum_{i<j} \hat{d}_0(i,j)},$$

in Figure 15. Euclidean is triangle-inequality stable yet not robustly geodesic-accurate (Figure 2). At $\delta = 0.005$, Euclidean NMAE is small, consistent with triangle-inequality stability. The same goes for Isomap, GF and Diffusion. This illustrates that triangle-inequality stability of a metric and geodesic-accuracy stability are orthogonal properties. AGD and AGD-SS are more geodesically accurate than Euclidean (Figure 2) and also stable to perturbation (Figure 15). They have comparable NMAE to Euclidean.

E.3 Geodesic retrieval on hyperbolic space and SPD manifold

Here we conduct geodesic retrieval task on hyperbolic space and SPD manifold.

E.3.1 Scalability of intrinsic dimension on hyperbolic space

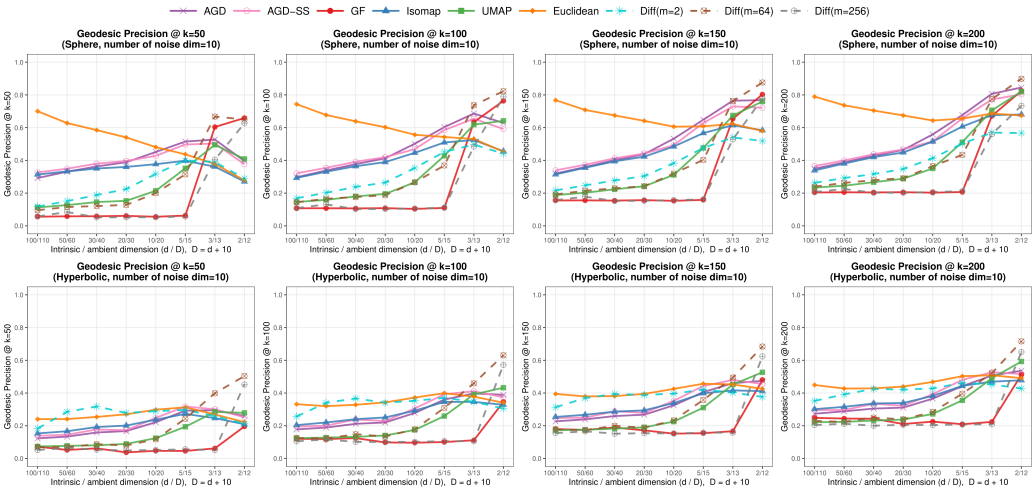


Figure 16: **Performance under fixed ambient noise dimension $c = 10$ as the intrinsic-to-ambient ratio d/D decreases, with $D = d + 10$. Top row:** Sphere S^{d-1}. **Bottom row:** Hyperbolic space $\mathbb{H}^d$. **Columns:** geodesic precision@k for $k \in \{50, 100, 150, 200\}$. This companion experiment reduces the amount of irrelevant ambient noise from 40 coordinates to 10, while still varying the intrinsic dimension and plotting performance against the ratio d/D. The intrinsic dimensions are $d \in \{2, 3, 5, 10, 20, 30, 50, 100\}$, so the corresponding ambient dimensions are $D \in \{12, 13, 15, 20, 30, 40, 60, 110\}$. All panels use $N = 1000$ points and noise constant 20, and compare the same nine methods used elsewhere in the response: AGD, AGD-SS, GF, Isomap, UMAP, Euclidean, and diffusion distances with $m \in \{2, 64, 256\}$. The horizontal axis is ordered by the displayed ratio d/D, so moving across the axis corresponds to increasing the fraction of informative coordinates in the ambient representation.

We design three experiments to isolate the effect of signal dilution in the ambient space while showing scalability of intrinsic dimension. We fix the number of noise dimensions and vary intrinsic dimension d, so that the ambient dimension becomes $D = d + c$ with $c \in \{10, 40, 100\}$. This makes the ratio d/D the primary control variable and allows us to study when ambient Euclidean distance is already adequate and when tree-based geodesic approximations become beneficial.

Number of noise dimensions (c) were chosen to span a low-noise regime ($c = 10$), an intermediate regime ($c = 40$), and a strongly noise-dominated regime ($c = 100$). Together, they provide a simple progression showing how performance changes as varies:

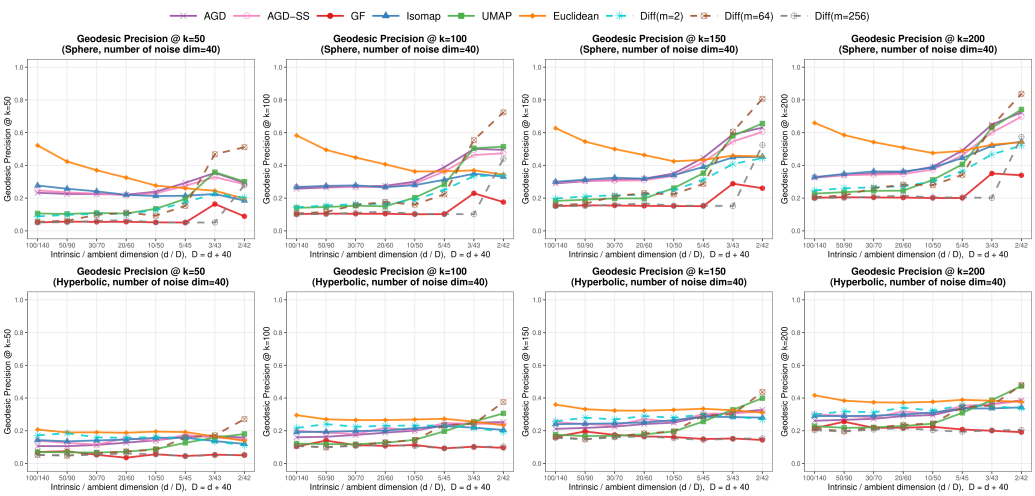


Figure 17: **Performance under fixed ambient noise dimension $c = 40$.**

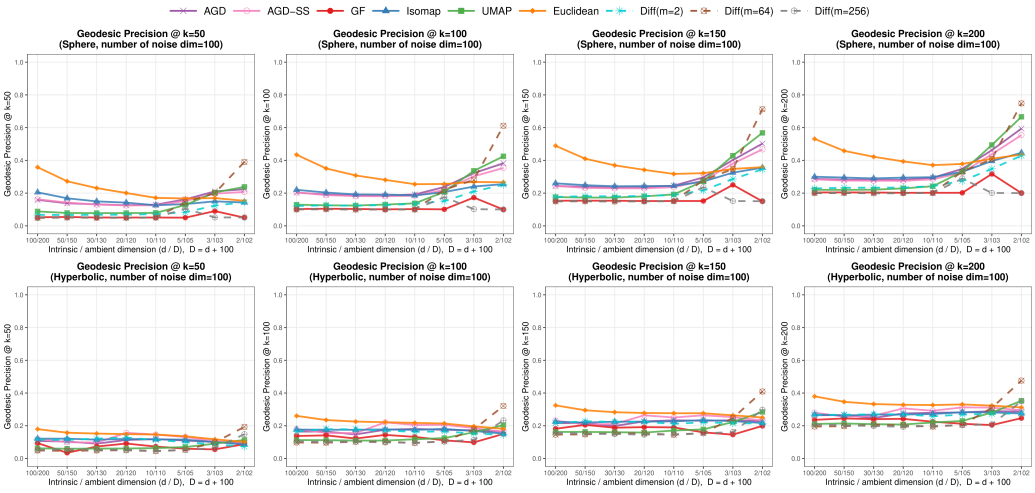


Figure 18: **Performance under fixed ambient noise dimension $c = 100$.**

1. With only $c = 10$ noise coordinates (Figure 16), Euclidean geometry is much more informative (large d/D), so the crossover toward Euclidean occurs earlier than in Figure 16. On the Sphere, AGD helps only at the smallest ratio, and the gain is already modest. In the hyperbolic panels, Euclidean is strongest for most dimensions because the bounded Poincaré ball samples preserve much of the local ordering under chord distance. Thus, when ambient noise is low, the extra estimation step in AGD and AGD-SS brings little additional benefit.
2. For $c = 40$ (Figure 17), AGD and AGD-SS help mainly in the lowest-ratio spherical regime ($d/D = 2/42, 3/43$), after which Euclidean becomes dominant. This crossover is expected

because, on the Sphere (top row), geodesic distance is monotone in Euc distance, so once the informative coordinates occupy a larger share of the ambient space, Euclidean distance is already a strong proxy. In the hyperbolic case (bottom row), Euclidean remains competitive over most of the range because the data are sampled in a bounded Poincaré ball rather than near the boundary (where geodesic distance from boundary point$=\infty$), so hyperbolic distance acts largely like a smooth reweighting of Euc distance. **Overall, tree-based approximations help most when the signal is a very small fraction of the ambient representation.**

3. When $c = 100$ (Figure 18), Euclidean distance deteriorates sharply at the smallest ratios, especially on the Sphere, where AGD and several geometry-aware baselines clearly outperform it. **This confirms that tree-based approximations are most useful in the extreme low-signal regime.** As d/D increases (right→left), Euclidean recovers and again becomes competitive. In the hyperbolic panels the gains from AGD-SS are still modest, because even under heavier noise the bounded Poincaré samples often preserve a local ordering close to Euclidean chord ordering except at the smallest ratios.

Together, the three figures show that the value of tree-based geodesic approximation grows with a smaller intrinsic-to-ambient ration (d/D).

E.3.2 Comparison on SPD manifold

Data generation. We fix a base point $A_0 = 2I_n$ and draw a random orthonormal basis $\{B_1, B_2\}$ of a 2-dimensional subspace of $\mathrm{Sym}(n)$ (each B_j is a symmetrised Gaussian matrix normalised to unit Frobenius norm). For each of the $N = 1000$ points, intrinsic coordinates $c \sim \mathcal{N}(0, 0.8^2 I_2)$ are sampled and the tangent vector $V = c_1 B_1 + c_2 B_2$ is mapped to the manifold via the Riemannian exponential map $\mathrm{Exp}_{A_0}(V) = A_0^{1/2} \exp\big(A_0^{-1/2} V A_0^{-1/2}\big) A_0^{1/2}$, yielding a 2-dimensional sub-manifold of $\mathcal{S}^n_{++}$.

Noise model. multiplicative congruence perturbations $P \mapsto LPL^\top$ with $L = I + \sigma E$, where E has i.i.d. standard Gaussian entries and $\sigma \in \{0, 0.05, 0.1, 0.2, 0.5, 1.0\}$. This preserves the SPD structure automatically and produces heteroscedastic Frobenius perturbations (larger matrices are perturbed more). In the baselines, F-norm is analogous to Euclidean in previous experiments. **For GF**, becasue GF needs feature selection in vector, each $n \times n$ SPD matrix is represented by its upper-triangular entries (off-diagonal scaled by $\sqrt{2}$), giving an $n(n+1)/2$-dimensional feature vector.

The ground-truth pairwise distances are computed using the affine-invariant Riemannian metric: $d_{\mathrm{AI}}(A, B) = \big\|\log(A^{-1/2} B A^{-1/2})\big\|_F = \sqrt{\sum_{i=1}^n \ln^2 \lambda_i(A^{-1}B)}$, where $\lambda_1, \ldots, \lambda_n$ are the generalized eigenvalues of $A^{-1}B$. Methods compared: AGD, AGD-SS (subsample variant, $\psi = 64$), Geodesic Forests (GF, 300 trees), Isomap, UMAP, Frobenius distance, and diffusion distances (k-NN with $k = 10$) with $m \in \{2, 64, 256\}$. All results are averaged over 100 trials with random seed 42.

We give the results of different methods on the SPD manifold $\mathcal{S}^n_{++}$ with $n \in \{3, 5, 10, 20, 50, 100\}$ in Figure 19. Under large noise level σ, performance degrades sharply for all methods once $\sigma \geq 0.1$, reflecting the multiplicative nature of the perturbation $P \mapsto LPL^\top$, which distorts both eigenvalues and eigenvectors simultaneously. At low noise ($\sigma = 0.05$), most methods retain reasonable precision. As the matrix dimension n increases, the noise effect intensifies: for $n = 100$, even $\sigma = 0.05$ causes a substantial drop in precision across all methods, because high-dimensional congruence transformations perturb an exponentially larger number of matrix entries. Among the geodesic-based approaches, AGD and AGD-SS maintain slightly higher precision than Isomap and diffusion distances at moderate noise levels ($\sigma \in [0.05, 0.2]$), particularly for smaller k values where local neighborhood fidelity matters most. Diffusion distances with small m (e.g., $m = 2$) decay fastest under noise, while larger m values ($m = 64, 256$) exhibit more graceful degradation, consistent with the smoothing effect of longer diffusion times. At $k = 800$ (rightmost column), all methods converge to similar precision since the neighborhood encompasses 80% of the dataset, reducing the discriminative power of the metric.

In summary, across three manifold types with analytically known geodesics, AGD-SS shows a consistent pattern: it provides the most benefit when ambient noise dominates the signal.

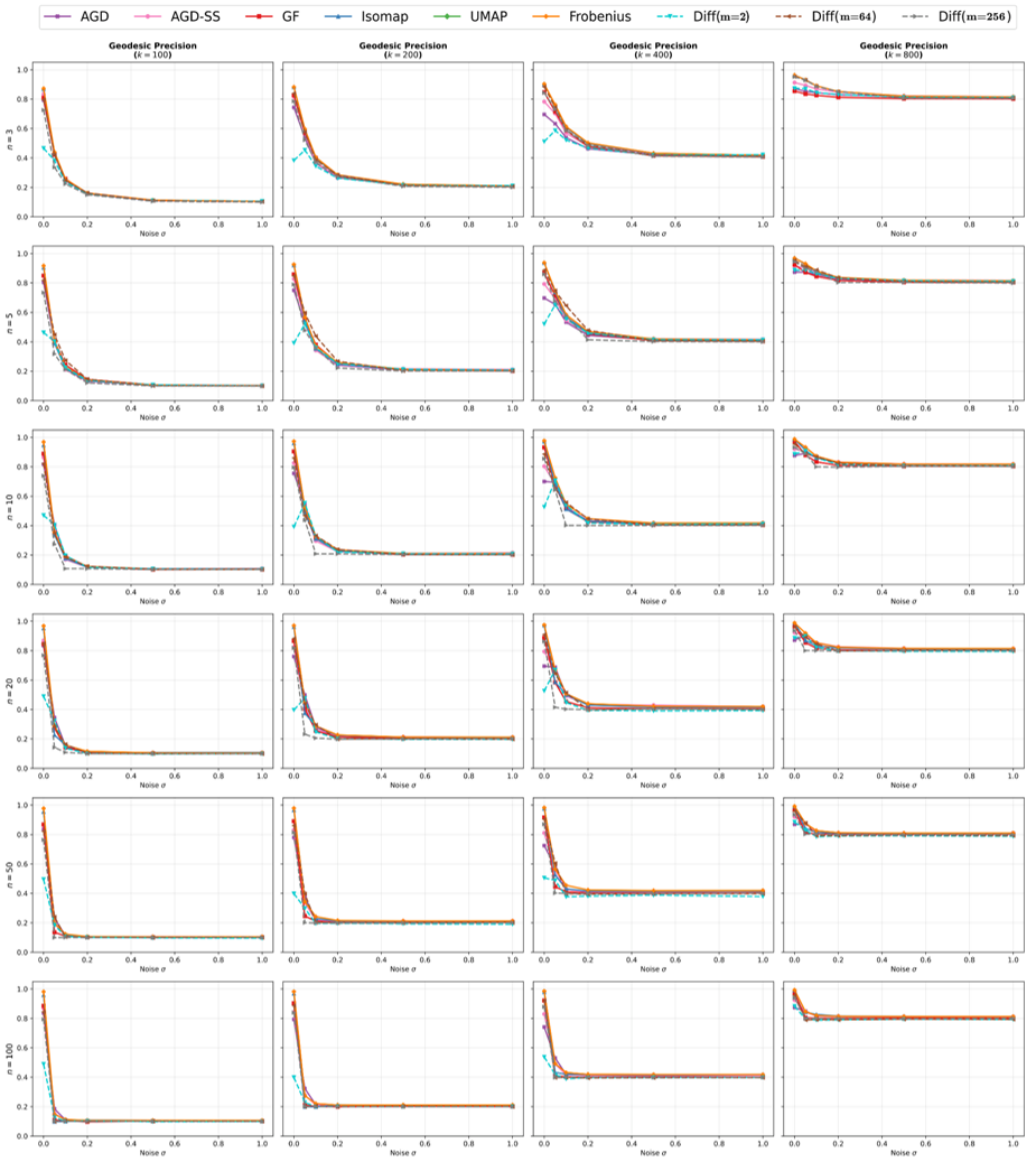


Figure 19: **Geodesic precision under congruence noise on the SPD manifold $\mathcal{S}^n_{++}$. Rows:** matrix sizes $n \in \{3, 5, 10, 20, 50, 100\}$ (corresponding feature dimensions $6, 15, 55, 210, 1275, 5050$). **Columns:** geodesic precision@k for $k \in \{100, 200, 400, 800\}$.

E.4 Effect of connectivity and non-convexity in manifold

Here we demonstrate the effect of the connectivity and non-convexity in manifold to the geodesic retrieval task. We show that AGD-SS can handle different levels of connectivity and non-convexity.

E.4.1 Varying connectivity levels

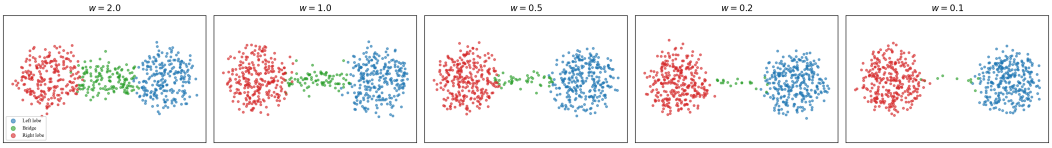


Figure 20: Dumbbell manifold overview at five bottleneck widths.

We designed a dumbbell manifold experiment: two 2D disks (radius $R = 2$) connected by a bridge of width w and length $L = 4$, embedded in $\mathbb{R}^{50}$ via random rotation with Gaussian noise ($\sigma = 0.3$) (Figure 20). We sweep $w \in \{2.0, 1.0, 0.5, 0.2, 0.1\}$ from strong to weak connectivity ($N = 600$).

Data generation. A 2D dumbbell manifold consists of two disks of radius $R = 2$ centered at $(\pm R + L/2,\ 0)$ connected by a rectangular bridge of width w and length $L = 4$. We sample $N = 600$ points distributed proportionally to area (two lobes + bridge, with a minimum of 10 bridge points). The 2D manifold is then embedded into $\mathbb{R}^{50}$ via a random orthogonal rotation $Q \in O(50)$ followed by additive Gaussian noise $\sigma = 0.3$ in all 50 dimensions.

Visualization. PCA projection from $\mathbb{R}^{50}$ back to 2D. Blue = left lobe, green = bridge, red = right lobe. Bottleneck widths sweep $w \in \{2.0, 1.0, 0.5, 0.2, 0.1\}$: at $w = 2.0$ the bridge is as wide as the lobe radius and points mix smoothly; at $w = 0.1$ the bridge is 20× narrower, leaving only around 10 bridge points and producing a near-disconnection. The visualization is given in Figure 20.

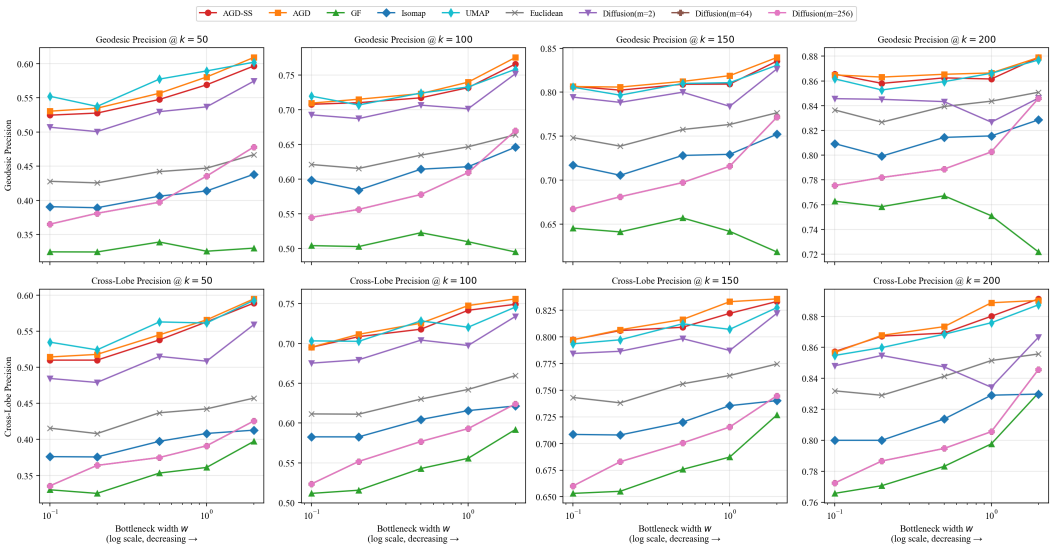


Figure 21: Geodesic Precision (top row) and Cross-Lobe Precision (bottom row) vs. bottleneck width.

Nine methods are evaluated on the dumbbell manifold ($N = 600$, $\mathbb{R}^{50}$, $\sigma = 0.3$) at widths $w \in \{2.0, 1.0, 0.5, 0.2, 0.1\}$. Columns correspond to $k \in \{50, 100, 150, 200\}$. Methods: AGD-SS/AGD ($\psi = 32$, $t = 100$); Diffusion ($m \in \{2, 64, 256\}$, 10-NN graph); Isomap ($k = 9$); Euclidean ($\mathbb{R}^{50}$); GF (200 trees).

Metrics. For a given distance estimate $\hat{d}$ and a query point x_i, let $\hat{\mathcal{N}}_k(x_i)$ denote the k nearest neighbours of x_i under $\hat{d}$, and let $\mathcal{N}_k^*(x_i)$ denote the k nearest neighbours under the **true** geodesic

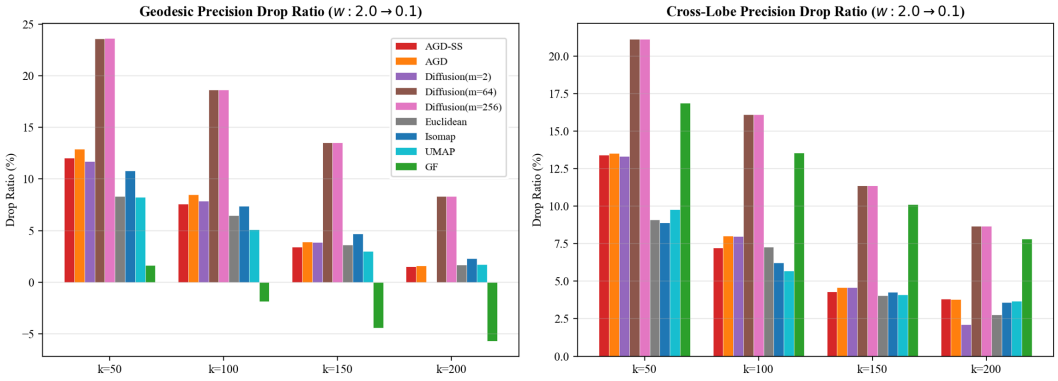


Figure 22: Performance drop ratio (%) from $w = 2.0$ to $w = 0.1$.** Drop ratio in GP (left) and CL (right) when moving from the widest bottleneck ($w = 2.0$, strong connectivity) to the narrowest ($w = 0.1$, near-disconnection), for each method at each k. Computation: Drop Ratio $= \frac{\text{metric}(w=2.0)-\text{metric}(w=0.1)}{\text{metric}(w=2.0)} \times 100\%$; larger bars indicate greater vulnerability to bottleneck narrowing.

distance. **Geodesic Precision (GP) @** $k = \frac{1}{N}\sum_{i=1}^{N} \frac{|\hat{\mathcal{N}}_k(x_i) \cap \mathcal{N}_k^*(x_i)|}{k}$, i.e. the average overlap between estimated and true geodesic k-NN, over all N points. **Cross-Lobe Precision (CL) @** k: let $\mathcal{L}$ denote the set of points in the left lobe. Then CL@$k = \frac{1}{|\mathcal{L}|}\sum_{x_i \in \mathcal{L}} \frac{|\hat{\mathcal{N}}_k(x_i) \cap \mathcal{N}_k^*(x_i)|}{k}$. CL isolates performance on the "bottleneck side" of the manifold and is more sensitive to connectivity-related errors than GP.

Figure 21 shows that AGD-SS/AGD consistently rank 1st–2nd across all connectivity levels. Figure 22 shows that all methods degrade as w decreases, but AGD-SS degrades slowly while maintaining high precision. Figure 22 left shows that the two Diffusion variants with $m = 64$ and $m = 256$ suffer the largest drop ratios for Geodesic precision, especially at $k = 50$ and $k = 100$, confirming their vulnerability to bottlenecks. AGD-SS/AGD show moderate drop ratios that decrease with k . Cross-lobe precision (Figure 22 right) shows Diffusion($m = 64, 256$) and GF collapse fastest under weak connectivity, while AGD-SS remains relatively stable.

Why AGD-SS handles bottlenecks? Single-linkage must route through the narrow neck, producing a high merge height. The subtree summation accumulates these bottleneck merges, encoding the geodesic detour. The ensemble average over random subsamples smooths out stochastic variation in bridge sampling.

E.4.2 Varying degrees of non-convexity

We designed a bent-strip experiment (Figure 23): a 2D strip of fixed arc length and width is bent into a cylindrical arc of angle θ, then embedded in $\mathbb{R}^{50}$ with noise. We sweep $\theta = \frac{\pi}{6}$ to 2π. At $\theta = \frac{pi}{6}$ the strip is nearly flat (geodesic≈ Euclidean); at $\theta = 2\pi$, it forms a full cylinder (maximally non-convex). The manifold area is held constant so that only non-convexity varies.

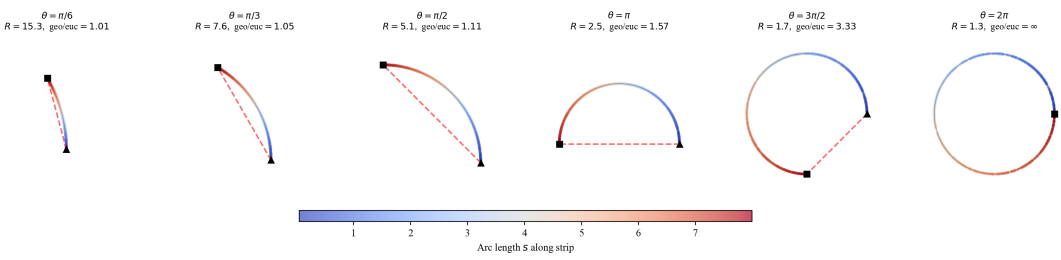


Figure 23: Bent-strip manifold at six bending angles (top-down view).

Data generation. A 2D rectangular strip of *fixed* arc length $\ell = 8$ and width $W = 3$ is bent into a cylindrical arc of angle θ and radius $R = \ell/\theta$. Keeping the arc length constant ensures that the manifold area and point density remain identical across all panels; only the curvature (and hence the degree of non-convexity) changes. We sample $N = 600$ points uniformly in intrinsic coordinates (s, z) with $s \in [0, \ell]$ (arc-length position) and $z \in [0, W]$ (height along the cylinder axis). The 2D surface is embedded in $\mathbb{R}^{50}$: the three non-trivial coordinates $(R\cos(s/R),\ R\sin(s/R),\ z)$ are zero-padded to 50 dimensions, then multiplied by a random orthogonal matrix and perturbed with i.i.d. Gaussian noise $\sigma = 0.3$.

Visualization. Each panel projects the 3D cylinder coordinates onto the (x, y) plane. Color encodes the arc-length coordinate s (blue $\rightarrow$ red).Black markers denote arc start (▲) and end (■); the red dashed line is the Euclidean chord. **Non-convexity ratio** = (geodesic arc length) / (Euclidean chord) between endpoints: it grows monotonically from 1.01 ($\theta = \pi/6$, nearly flat) through 1.57 ($\theta = \pi$, U-shape) to ∞ ($\theta = 2\pi$, full cylinder where endpoints coincide and the chord collapses to zero). Bending angles sweep $\theta \in \{\pi/6,\ \pi/3,\ \pi/2,\ \pi,\ 3\pi/2,\ 2\pi\}$.

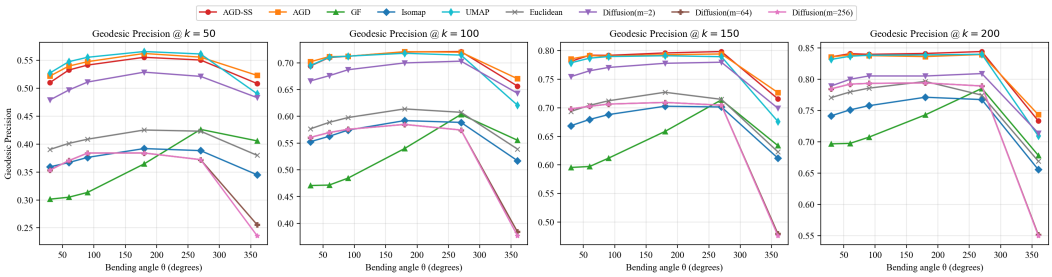


Figure 24: Geodesic Precision vs. bending angle θ (non-convexity) experiment on the bent-strip manifold..

The results are shown in Figure 24. AGD-SS/AGD and UMAP maintain stable precision across all θs. At $\theta = 2\pi$, AGD-SS achieves $GP@50 = 0.508$ while Diffusion(m=64) collapses to 0.256 and Diffusion(m=256) to 0.236. Euclidean also degrades at high non-convexity ($GP@50 = 0.380$). In addition, the dumbbell experiment also tests non-convexity: as w shrinks, the shortest path between lobes must detour through an increasingly narrow neck, making the manifold strongly non-convex. Furthermore, several datasets in our main paper are inherently non-convex (e.g., the Swiss roll), where AGD-SS already demonstrates superior geodesic preservation.

E.5 Geodesic precisions of AGD and AGD-SS (using single linkage) under different ks

We show the comparison of different methods in Geodesic Retrieval task under different k values in Figure 25 ($k = 50, 100, 150, 200$). GF achieves the best performance across all datasets when $k = 50$, but its geodesic precision degrades rapidly as k increases. In contrast, AGD consistently attains the highest precision across most datasets for larger k values. This result underscores that GF only preserves local geodesic structure, while AGD effectively captures both local and global geodesic distances.

In Figure 26, we compare AGD and AGD-SS with Diffusion distance with different number of diffusion steps ($m = 2, 64, 256$). AGD and AGD-SS usually give similar results. Compared with Diffusion distance, both of them give comparable or higher geodesic precision.

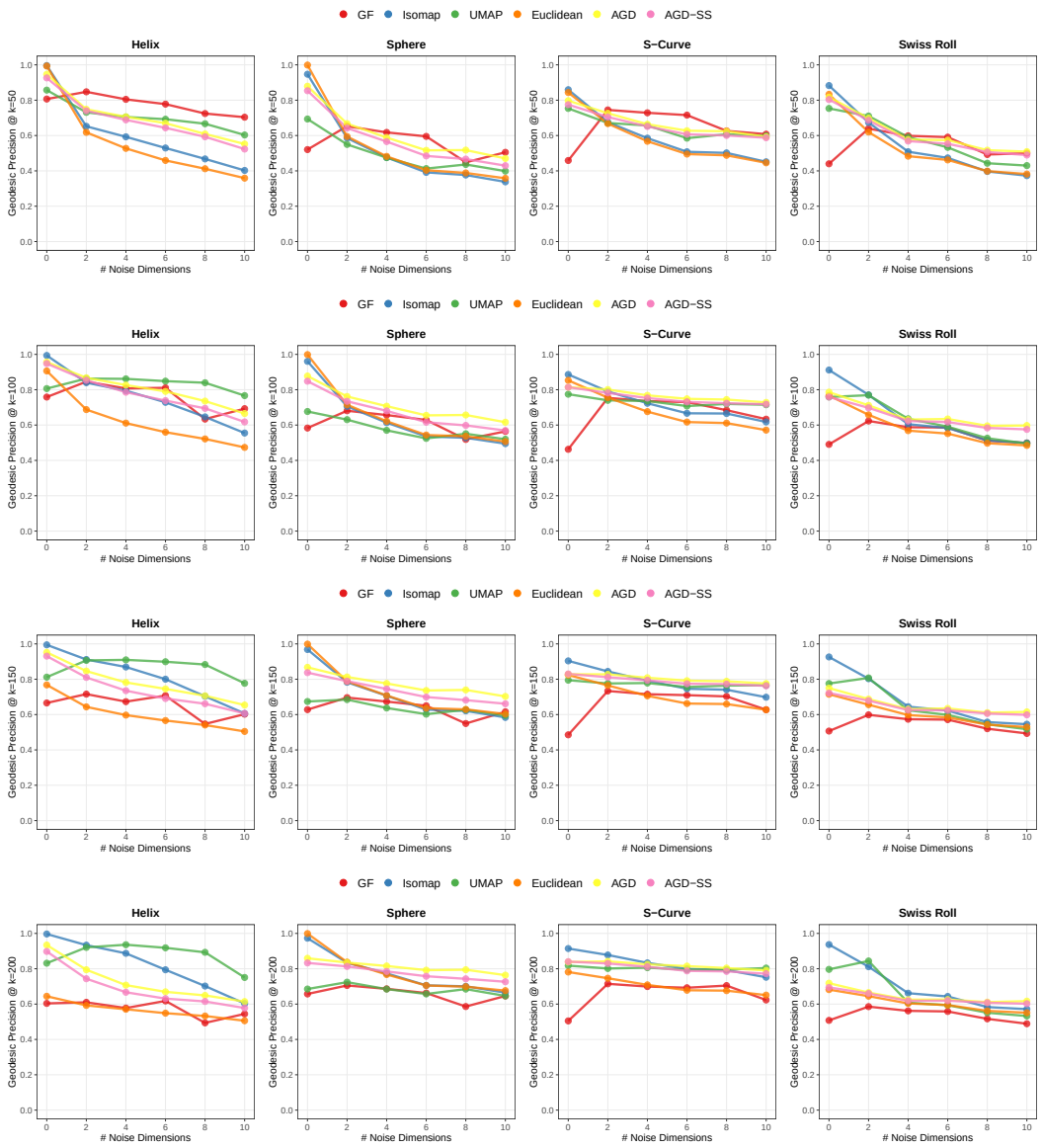


Figure 25: Geodesic Precisions of different metrics on artificial datasets under different k values ($k = 50, 100, 150, 200$, the same setting as that in Madhyastha et al., 2020). All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

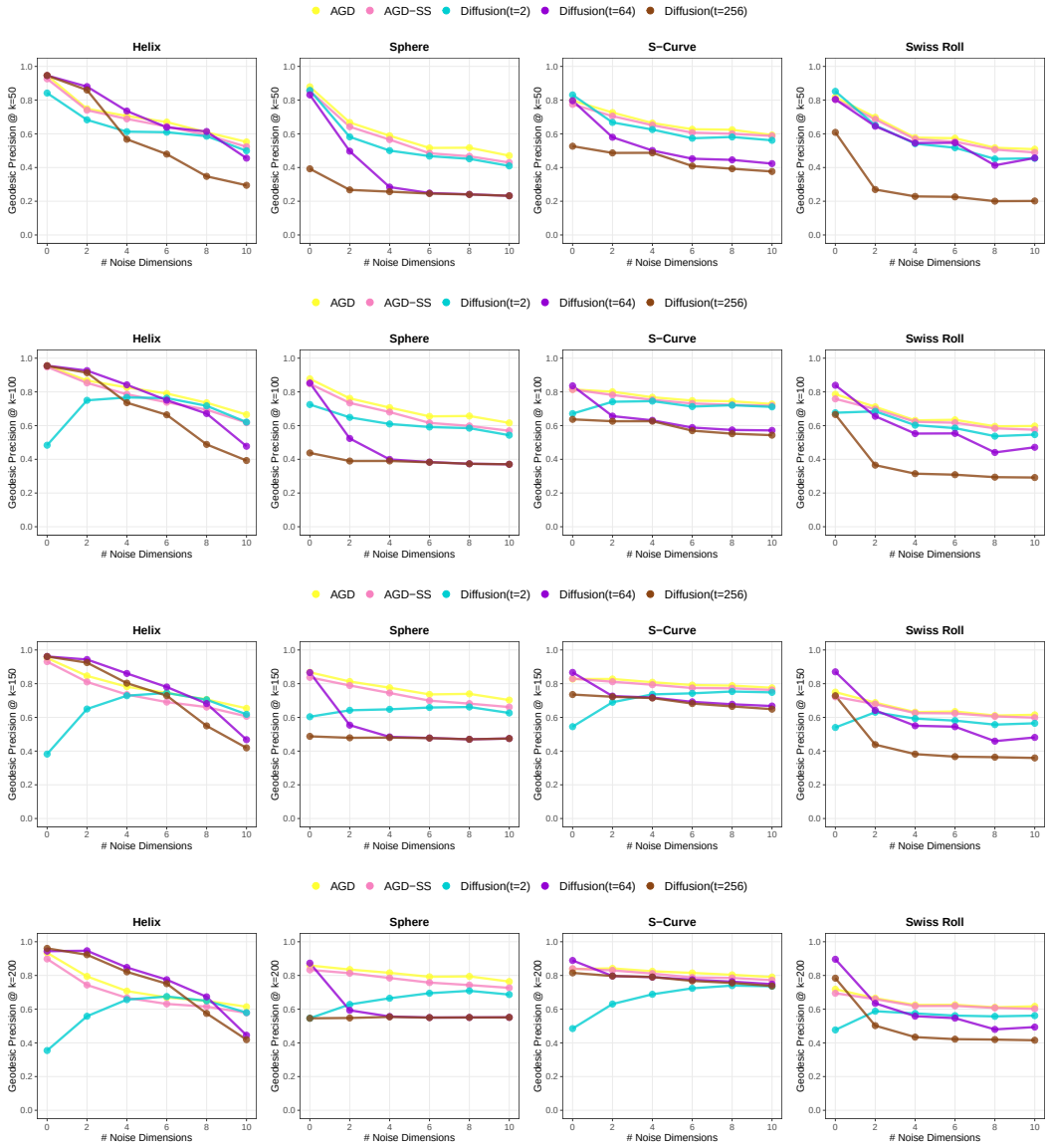


Figure 26: Geodesic Precisions of AGD, AGD-SS and Diffusion distance on artificial datasets under different k values ($k = 50, 100, 150, 200$, the same setting as that in Madhyastha et al., 2020). All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

E.6 GEODESIC PRECISIONS OF AGD AND AGD-SS (USING SINGLE LINKAGE) UNDER DIFFERENT NOISE STDS σ IN HIGH-DIMENSIONAL CASE

We show the comparison of different methods under different noise stds in the noise dimensions in Figure 27-29. Given that we use all the methods on normalized datasets, the effect of noise std σ is not obvious. Nearly all methods give the same result under different noise stds. AGD and AGD-SS remain the best methods across varying noise stds on most datasets.

In Figure 30 to 32, we show the average geodesic precision across different numbers of noise dimensions of Diffusion distance, AGD and AGD-SS. In Figure 33-35, we show the corresponding detailed results. For a given k value, compared with the best result achievable by Diffusion distance (based on 10-NN graph) under any specific m, the default setting of AGD and AGD-SS ($\psi = 32, t = 100$) generally yields comparable or higher geodesic precisions.

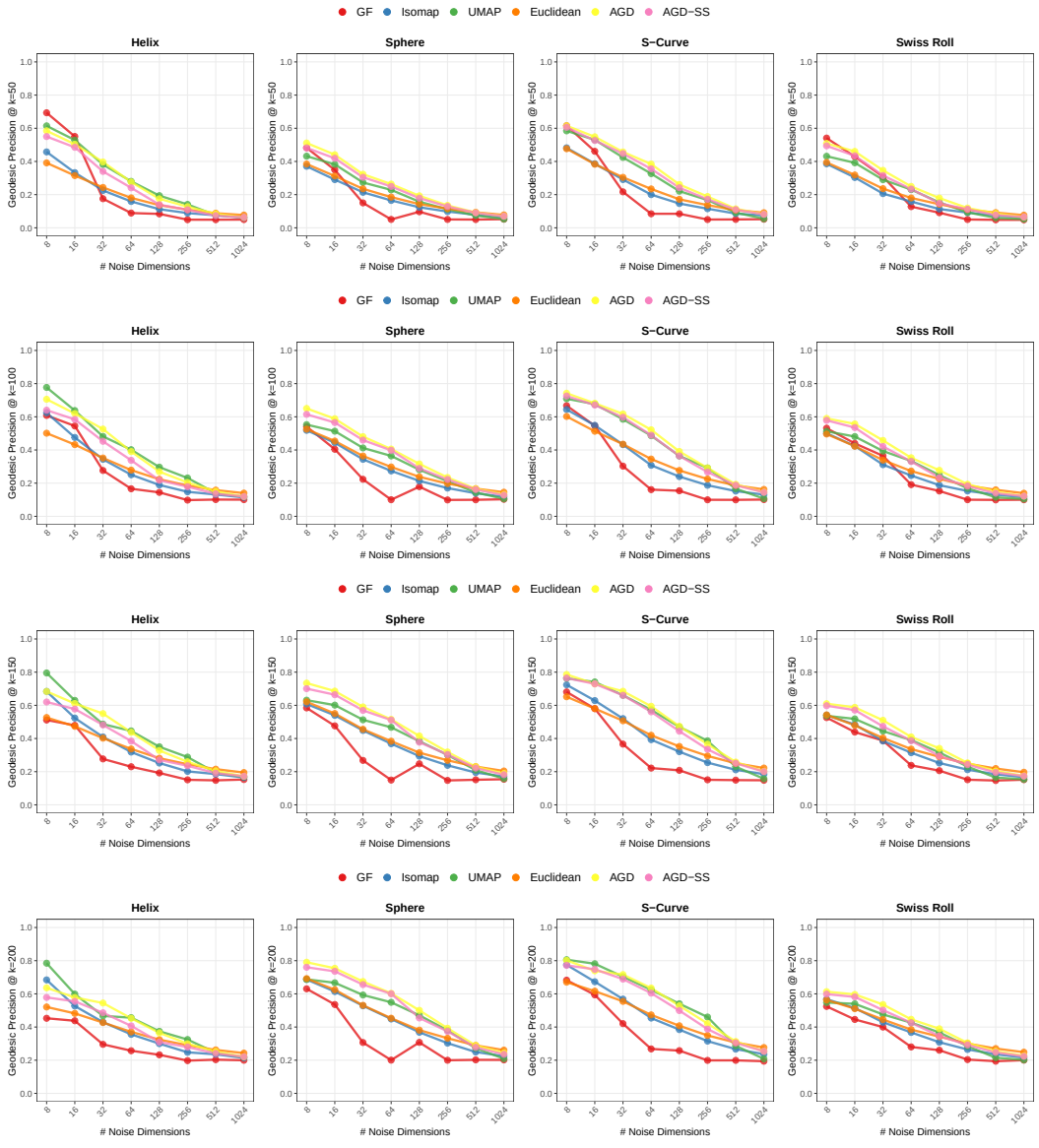


Figure 27: Geodesic Precisions of different metrics on artificial datasets. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

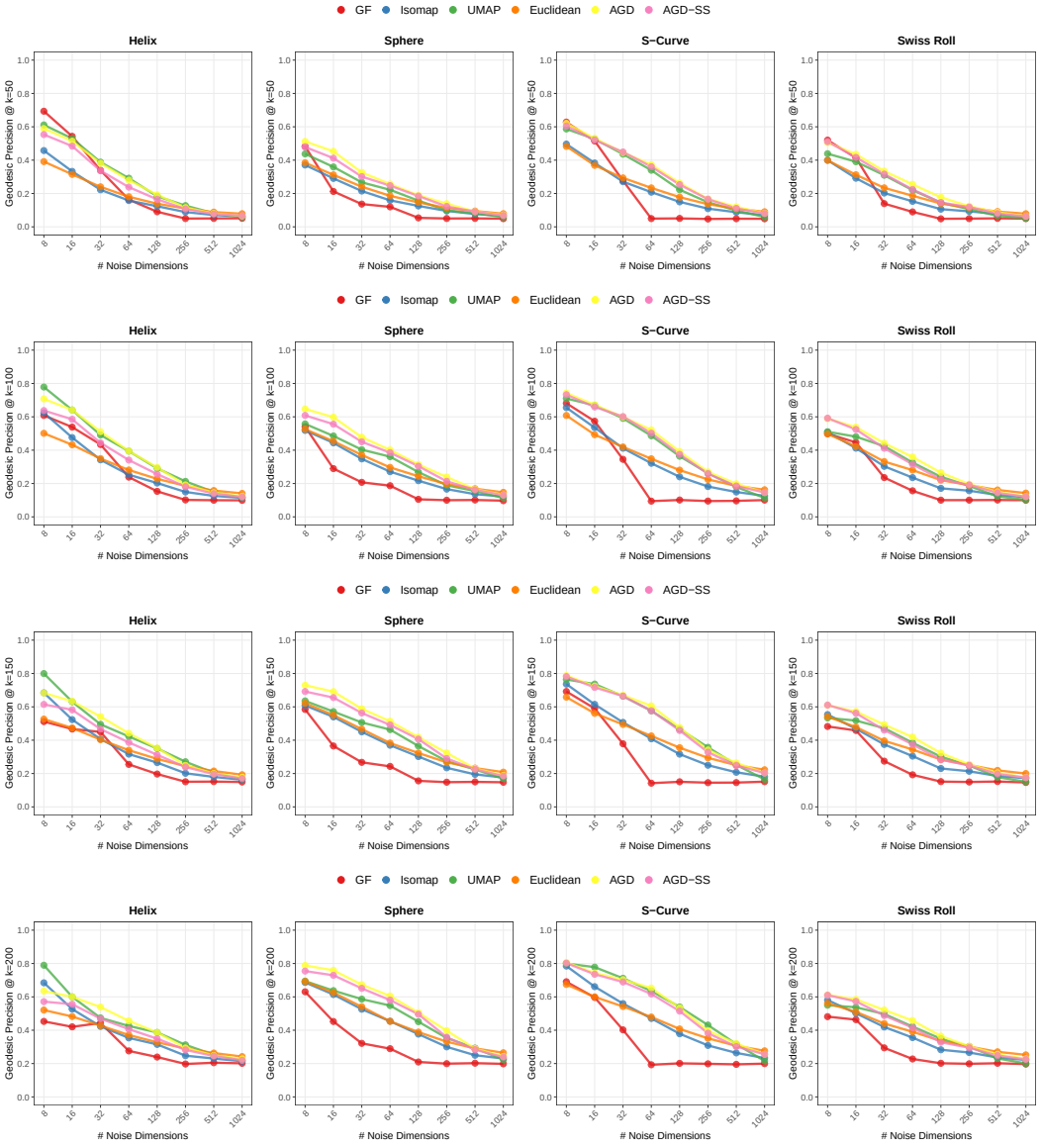


Figure 28: Geodesic Precisions of different metrics on artificial datasets. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 40$.

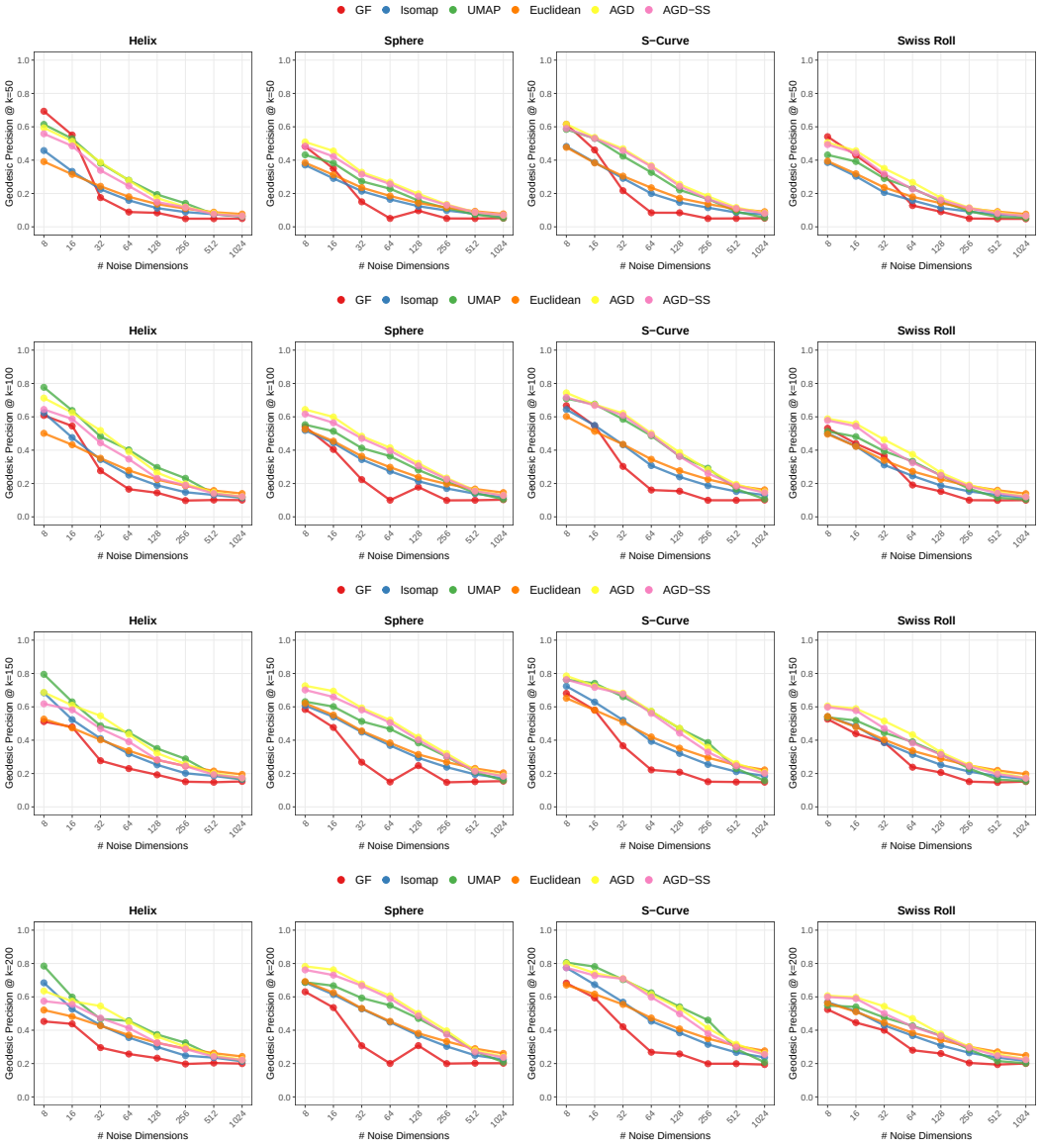


Figure 29: Geodesic Precisions of different metrics on artificial datasets. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 80$.

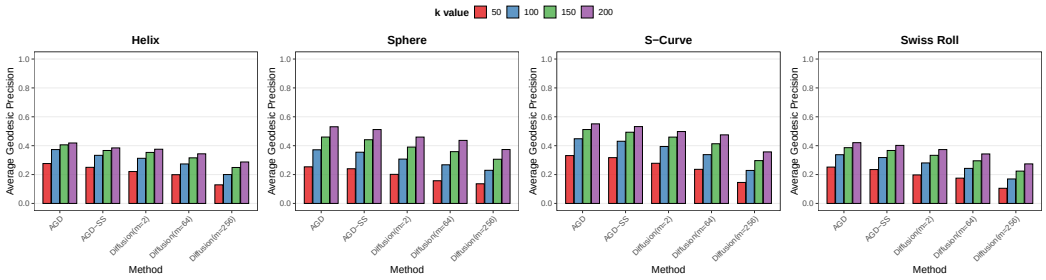


Figure 30: Comparison of average geodesic precision over different numbers of noise dimensions (8,16,32,64,128,256) for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$) when noise std $\sigma = 20$.

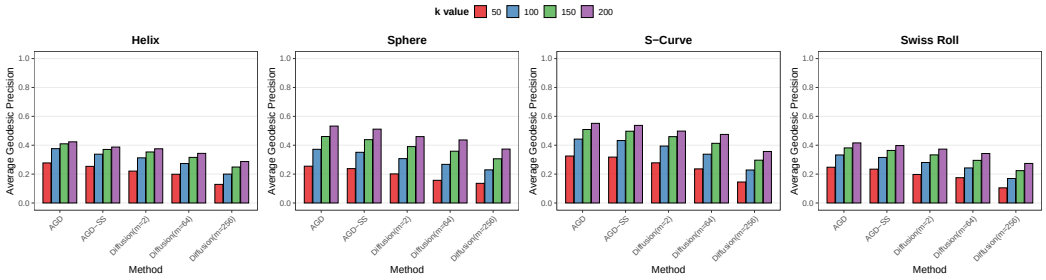


Figure 31: Comparison of average geodesic precision over different numbers of noise dimensions (8,16,32,64,128,256) for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$) when noise std $\sigma = 40$..

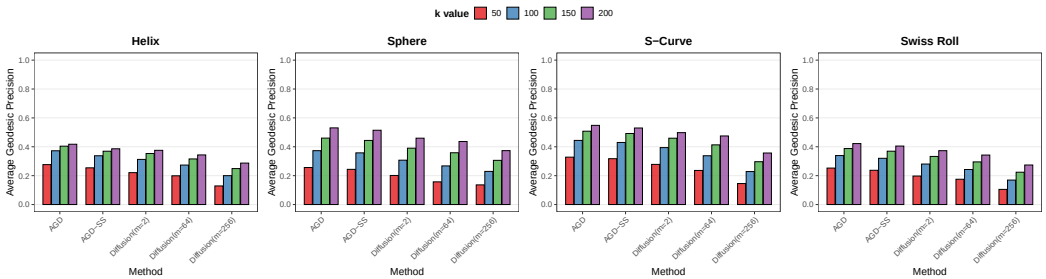


Figure 32: Comparison of average geodesic precision over different numbers of noise dimensions (8,16,32,64,128,256) for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$) when noise std $\sigma = 80$.

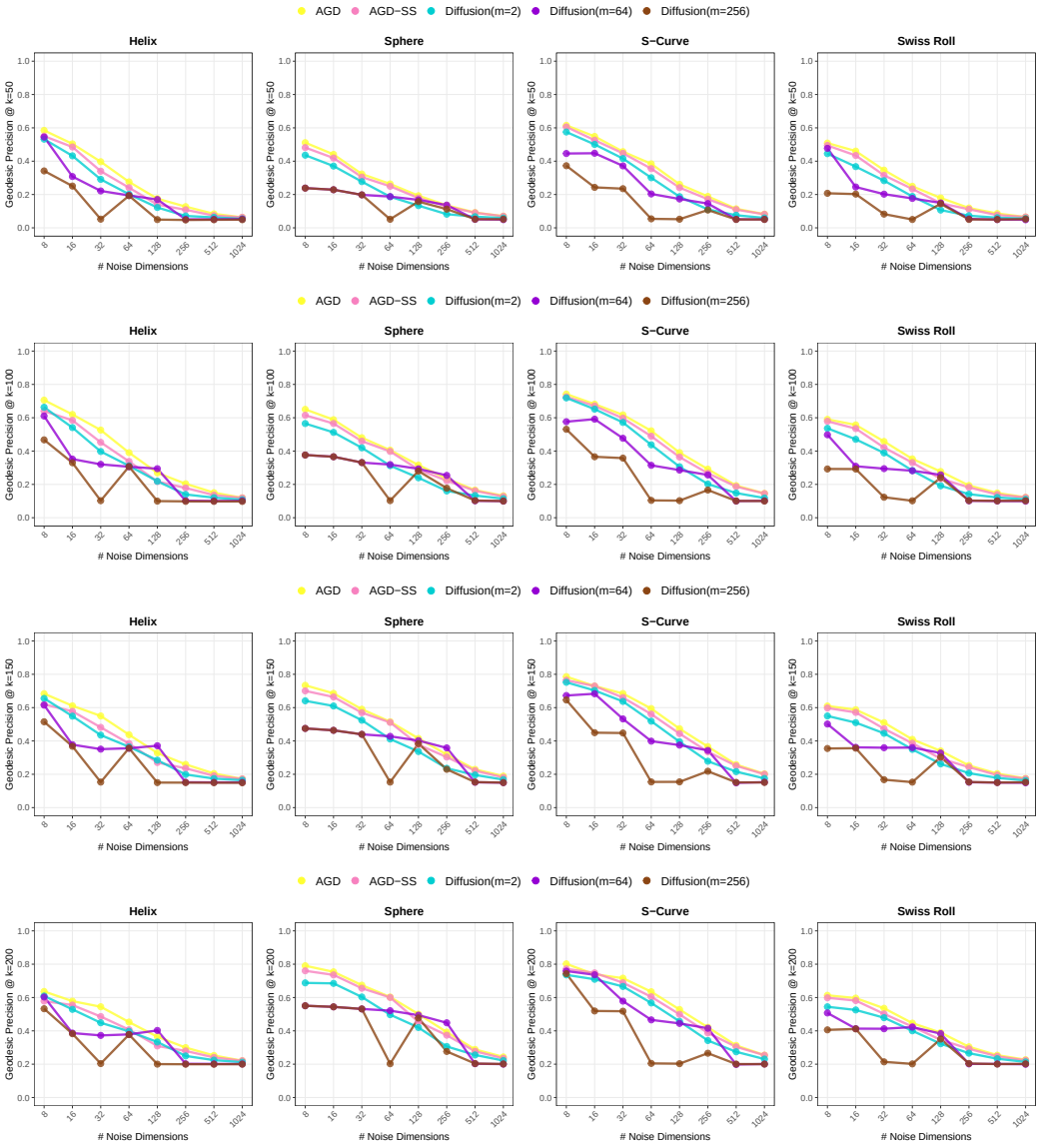


Figure 33: Comparison of geodesic precision over different numbers of noise dimensions (8,16,32,64,128,256) for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$) when noise std $\sigma = 20$.

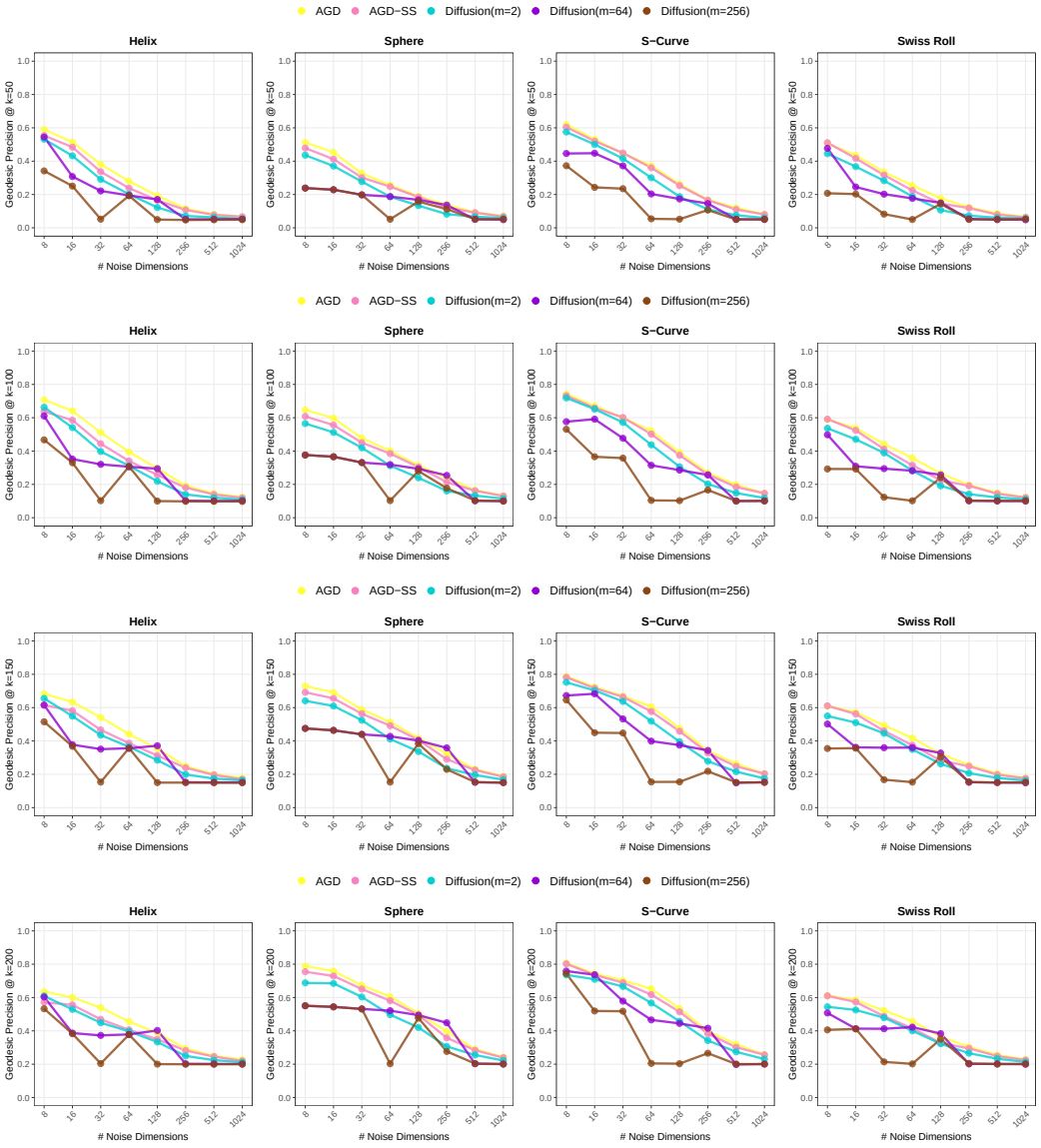


Figure 34: Comparison of geodesic precision over different numbers of noise dimensions (8,16,32,64,128,256) for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$) when noise std $\sigma = 40$.

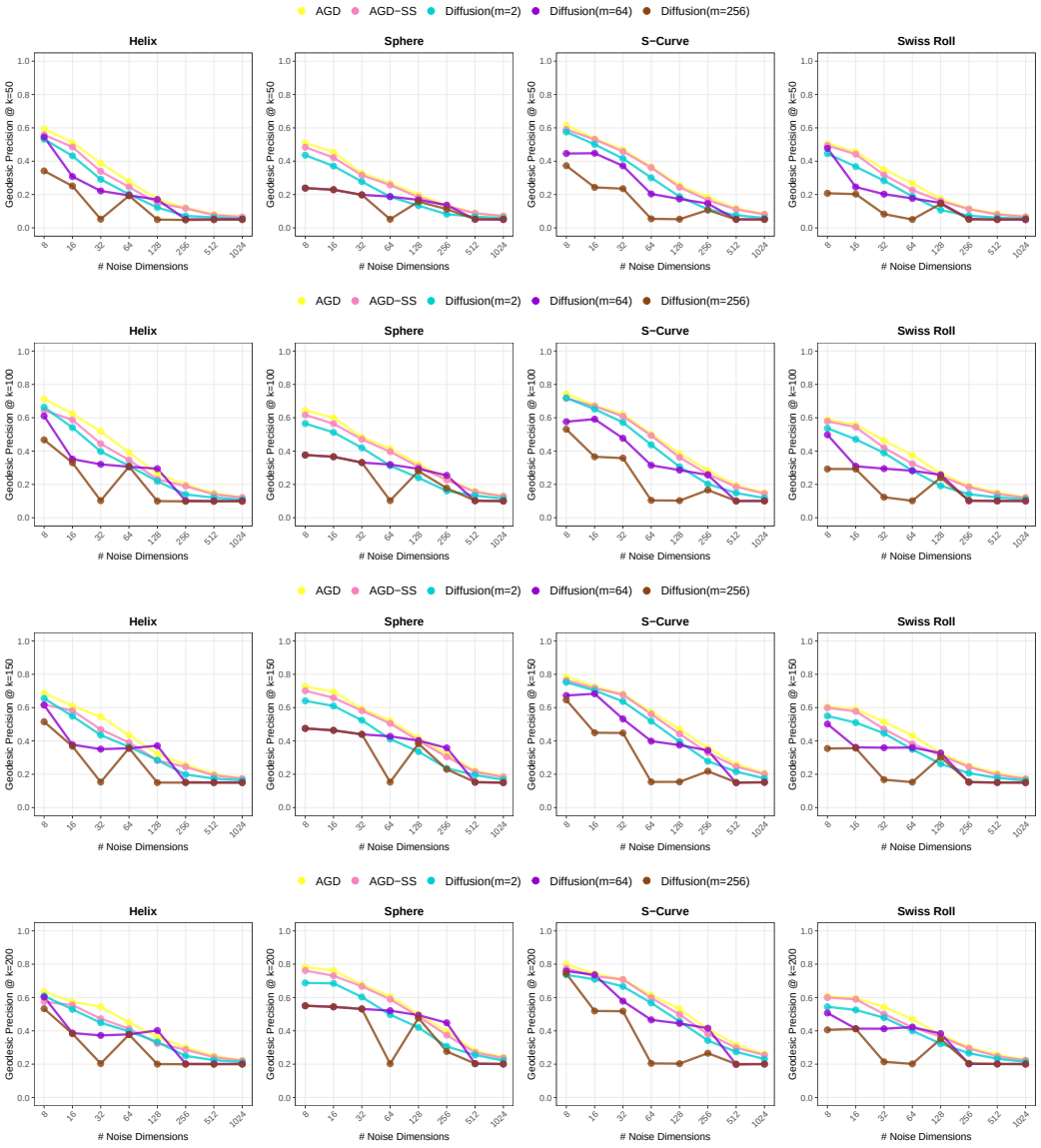


Figure 35: Comparison of geodesic precision over different numbers of noise dimensions (8,16,32,64,128,256) for AGD, AGD-SS and Diffusion distance with different diffusion steps ($m = 2, 64, 256$) when noise std $\sigma = 80$.

E.7 On the linkage function choice

Here we consider three different linkage function choices in AGD and AGD-SS: single linkage, average linkage and complete linkage. In additon, we consider using the same random projection technique as that in the Geodesic Forest paper Madhyastha et al. (2020) in AGD and AGD-SS as a preprocessing step.

E.7.1 Artificial datasets

We consider different linkage choices under the setting of noise std $\sigma = 20$ and varying k (50,100,150,200) in Geodesic Retrieval task in Figure 36-43. We observe that whether using sparse random projection, different linkage choices give similar geodesic precisions on the artificial datasets.

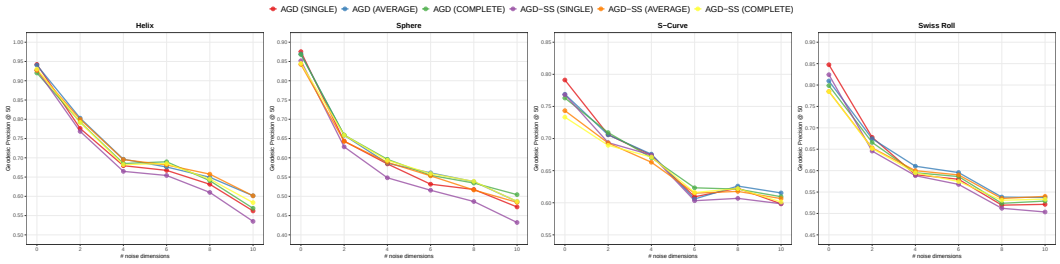


Figure 36: Geodesic Precisions of different metrics on artificial datasets when $k = 50$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

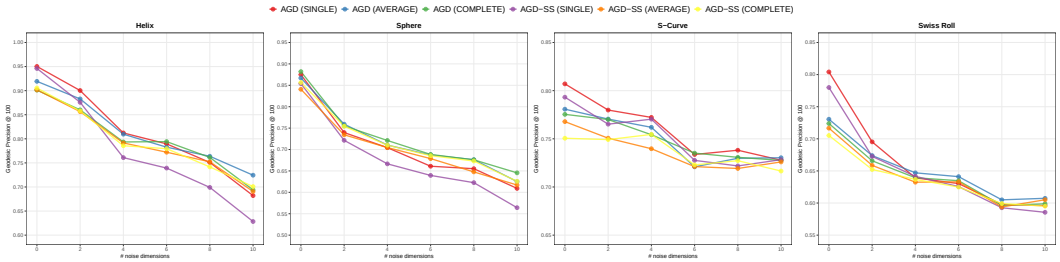


Figure 37: Geodesic Precisions of different metrics on artificial datasets when $k = 100$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

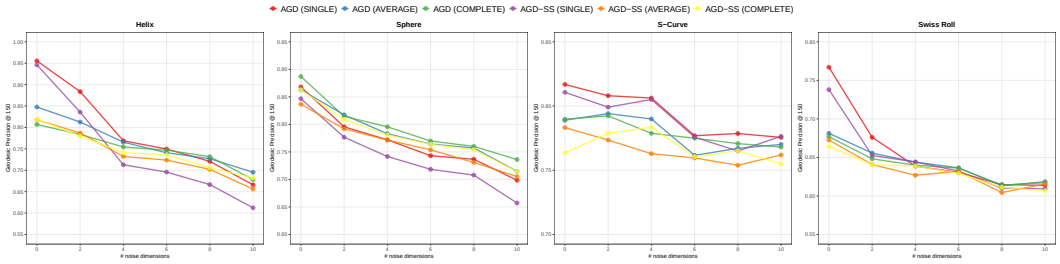


Figure 38: Geodesic Precisions of different metrics on artificial datasets when $k = 150$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

E.7.2 Drosophila Connectome dataset

We report the geodesic precision and recall curve of AGD and AGD-SS using different linkage functions on Drosophila Connectome dataset in Figure 44 (with random projection) and 45 (without

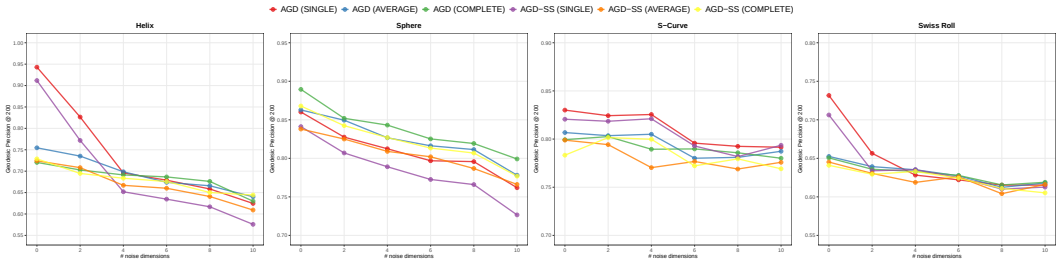


Figure 39: Geodesic Precisions of different metrics on artificial datasets when $k = 200$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$.

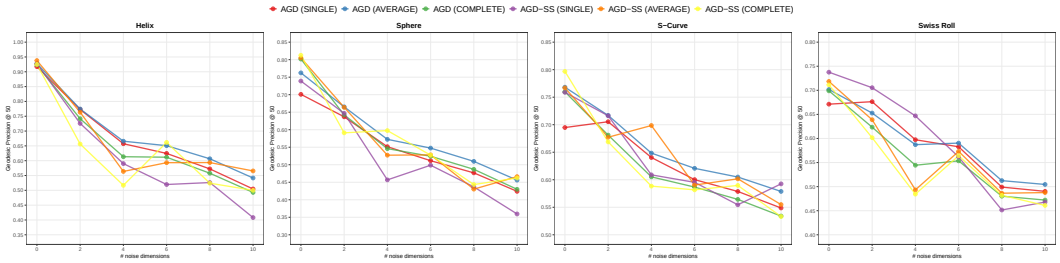


Figure 40: Geodesic Precisions of different metrics on artificial datasets when $k = 50$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$. Both AGD and AGD-SS use random projection.

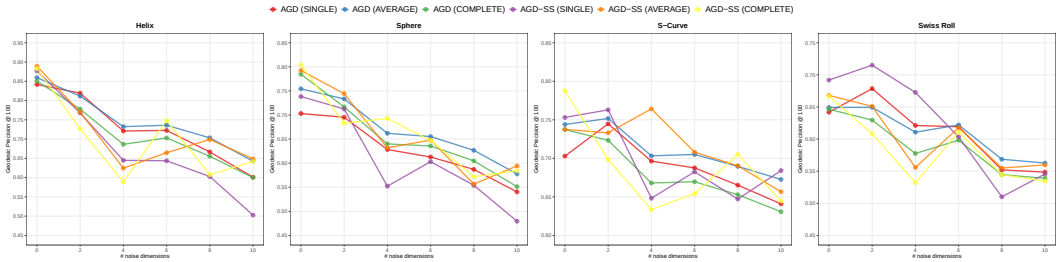


Figure 41: Geodesic Precisions of different metrics on artificial datasets when $k = 100$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$. Both AGD and AGD-SS use random projection.

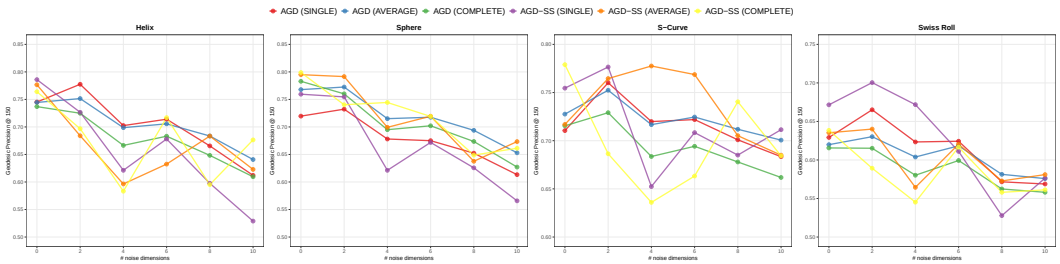


Figure 42: Geodesic Precisions of different metrics on artificial datasets when $k = 150$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$. Both AGD and AGD-SS use random projection.

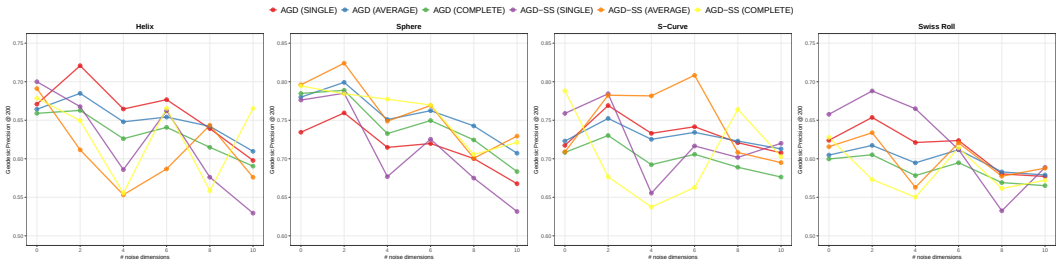


Figure 43: Geodesic Precisions of different metrics on artificial datasets when $k = 200$. All metrics are used on normalized data. The noise level of noise dimension is $\sigma = 20$. Both AGD and AGD-SS use random projection.

random projection). We observe that single linkage always gives the best performance over different k values (from 50 to 150).

We analyze that by establishing the equivalence between the AHC dendrogram using single linkage and 0-dimensional Persistent Homology, we ground our method in the stability theorems of TDA. This connection explains why AGD-SS shares the robustness properties of PH, providing a rigorous justification for its performance in noisy regimes where traditional geometric methods fail. Furthermore, while we utilize single linkage for its theoretical connection to Persistent Homology, empirical comparisons working with other linkage functions (Figure 44 and 45) confirm that single linkage is indeed the optimal choice for AGD-SS, yielding superior geodesic preservation.

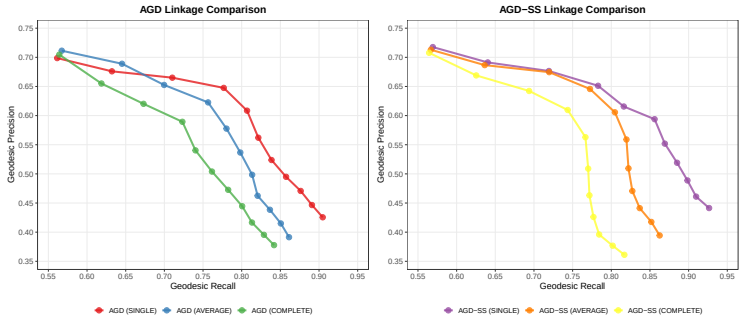


Figure 44: Geodesic Precisions and Recalls of AGD and AGD-SS (with different linkage functions) without random projection preprocessing step on Drosophila Connectome dataset.

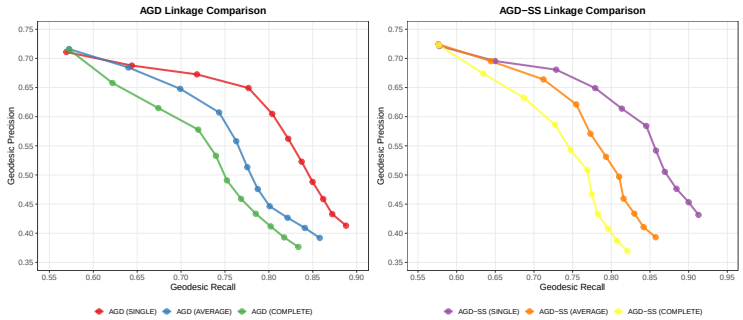


Figure 45: Geodesic Precisions and Recalls of AGD and AGD-SS (with different linkage functions) with random projection preprocessing step on Drosophila Connectome dataset.

E.8 HYPERPARAMETER SENSITIVITY

E.8.1 GEODESIC RETRIEVAL

We show the Geodesic Precision of AGD and AGD-SS under different ψ values on the artificial datasets in Figures 46-49. The geodesic precision is rather stable w.r.t. the ψ value.

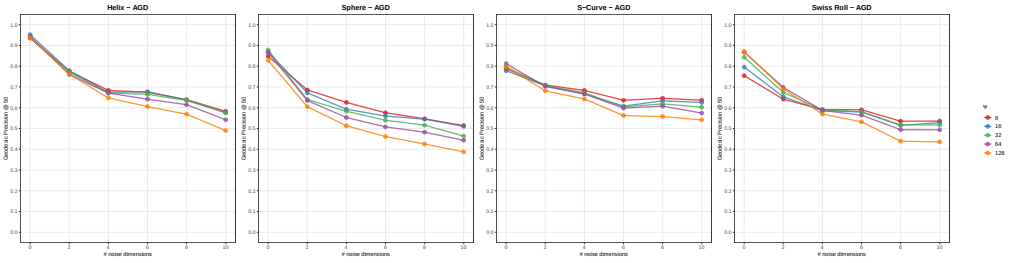


Figure 46: Geodesic Precisions (k=50) of AGD under different ψ values on artificial datasets.

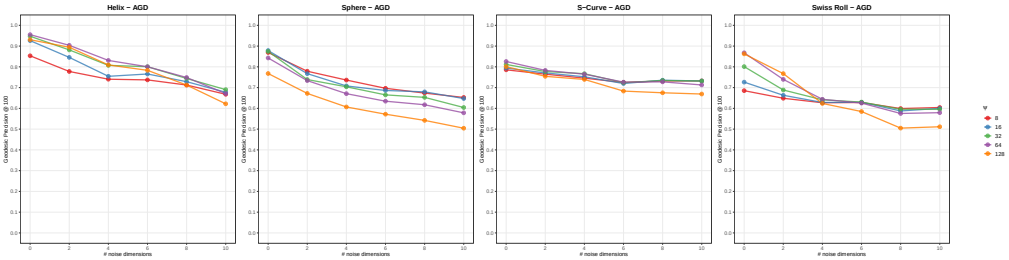


Figure 47: Geodesic Precisions (k=100) of AGD under different ψ values on artificial datasets.

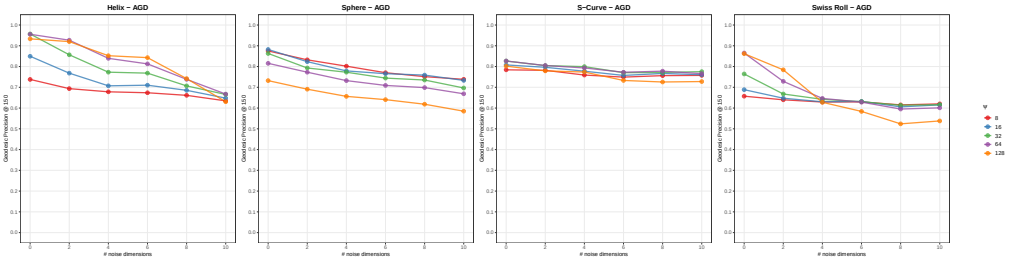


Figure 48: Geodesic Precisions (k=150) of AGD under different ψ values on artificial datasets.

We show the results on Drosophila Connectome dataset using different ψ values in Figure 54, given that this datasets is rather small (around 200 points), a larger ψ often gives better results. But in general AGD and AGD-SS is rather stable w.r.t. the hyperparameter choice, in comparison with Diffusion distance. In Figure 55, we show the Precision-Recall Curve of Diffusion distance with different diffusion steps ($m =, 2, 4, 8, 16, 32, 64$). A larger m (≥ 16) usually gives better results. The results are very sensitive to m when m is rather small.

E.8.2 LOOP DETECTION

We show the Loop Detection Score of AGD and AGD-SS using different ψ values on three scRNA-seq datasets in Figure 56. We observe that AGD-SS is less sensitive to ψ values than AGD.

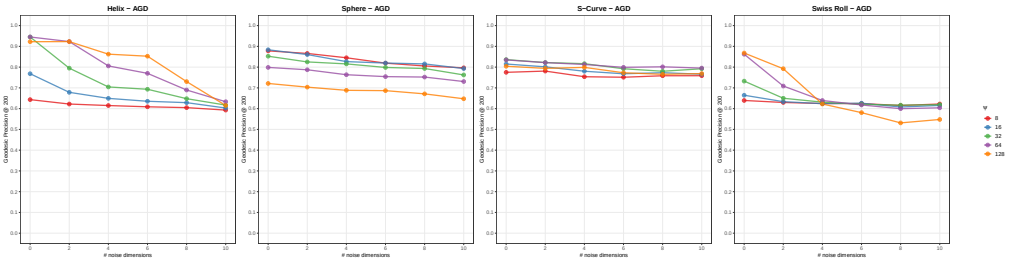


Figure 49: Geodesic Precisions (k=200) of AGD under different ψ values on artificial datasets.

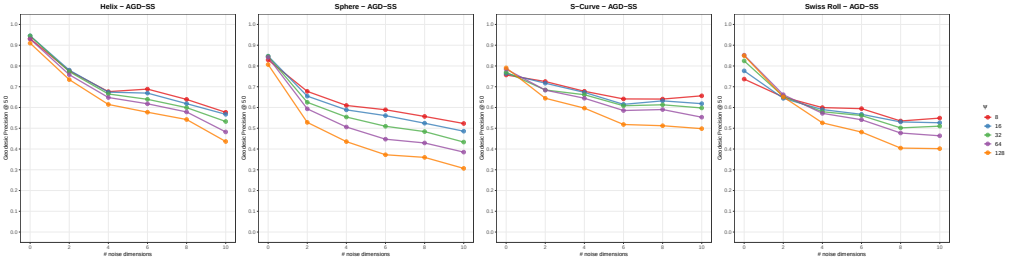


Figure 50: Geodesic Precisions (k=50) of AGD-SS under different ψ values on artificial datasets.

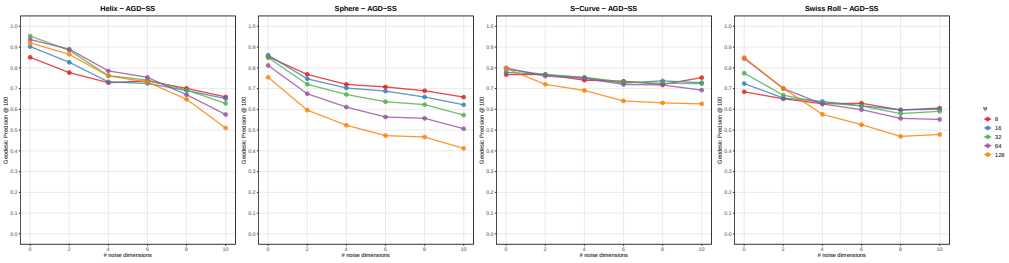


Figure 51: Geodesic Precisions (k=100) of AGD-SS under different ψ values on artificial datasets.

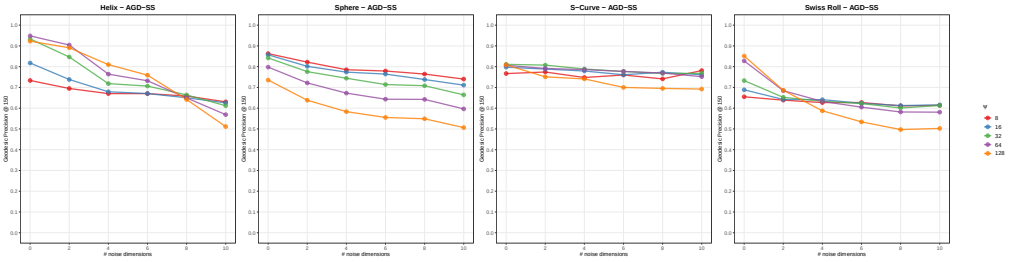


Figure 52: Geodesic Precisions (k=150) of AGD-SS under different ψ values on artificial datasets.

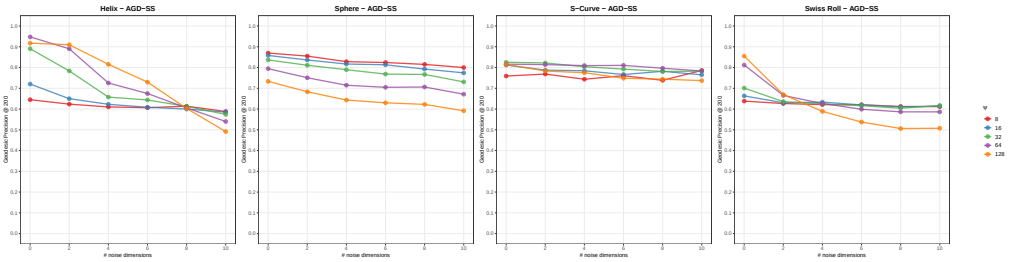


Figure 53: Geodesic Precisions (k=200) of AGD-SS under different ψ values on artificial datasets.

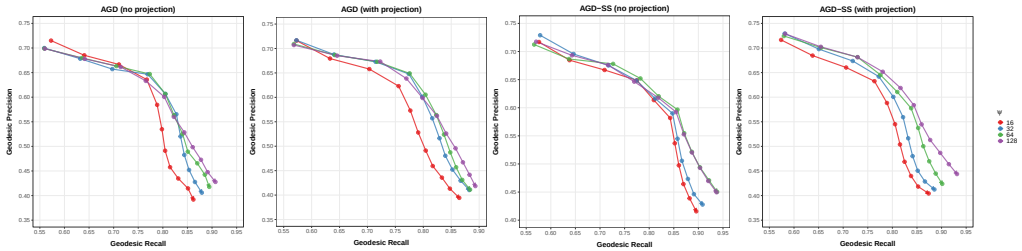


Figure 54: Geodesic Precisions and Recalls of AGD and AGD-SS under different ψ values on Drosophila Connectome dataset.

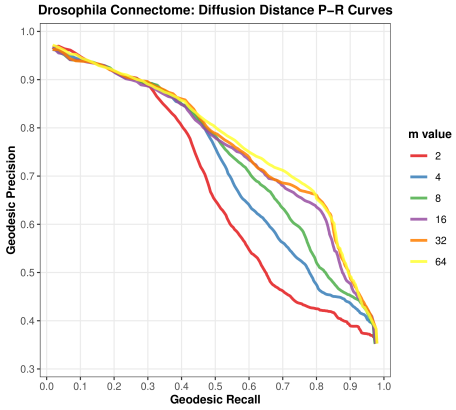


Figure 55: Geodesic Precisions and Recalls of Diffusion distance under different m values on Drosophila Connectome dataset.

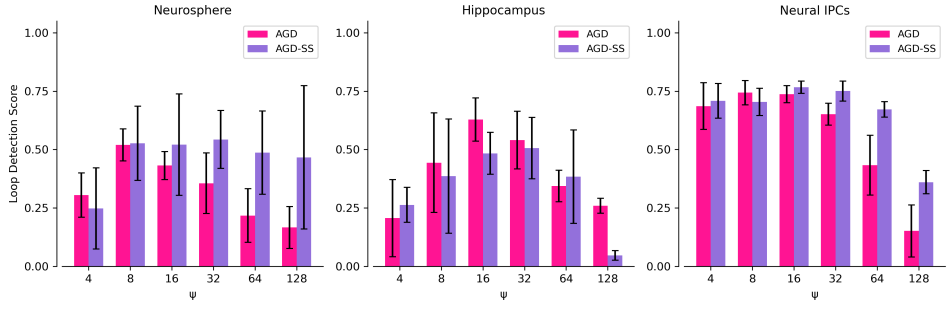


Figure 56: Loop Detection Score s_1 of AGD and AGD-SS using single linkage and different ψ values.

E.9 Effect of hyperparameter t (the number of trees) on concentration

Here we discuss the concentration for AGD-SS. Each subsample distance $d_{\text{AGD-SS}}(x, y|S) \in [0, B]$ where $B = \psi \cdot h_{\max}$ ($h_{\max}$ = maximum merge height in the single-linkage dendrogram). By Hoeffding's inequality, we have $\mathbb{P}(|\frac{1}{t}\sum_{i=1}^{t} d_{\text{AGD-SS}}(x, y|S_i) - \mathbb{E}[d_{\text{AGD-SS}}(x, y|S)]| \geq \epsilon] \leq 2\exp(-\frac{2t\epsilon^2}{B^2})$.

So $t = O(B^2/\epsilon^2) = O(\psi^2 h_{\max}^2)$ subsamples suffice for ϵ-accurate estimation with high probability. With our defaults $\psi = 32, t = 100$, the exponential tail ensures the estimator is well-concentrated. The estimator converges to $\mathbb{E}[d_{\text{AGD-SS}}(x, y|S)$ at rate $O(1/\sqrt{t})$.

We report an empirical sensitivity study in Figure 57–59, reporting variance across runs. The results confirm that $t = 100$ yields stable performance.

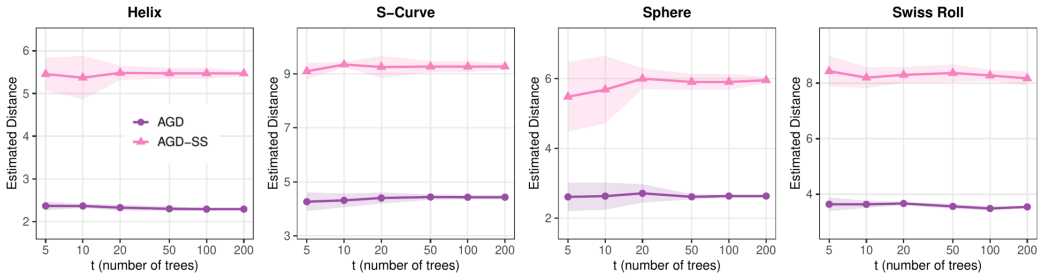


Figure 57: Pointwise concentration of AGD and AGD-SS for a single pair of points (5 independent runs). One pair (i, j) was randomly selected and held fixed across all runs.

E.10 Computation Time

AGD-SS's total cost is $O(t\psi N^2)$ with ψ, t as fixed constants, giving $O(N^2)$, the same asymptotic complexity as Diffusion Distance (which can directly computed with standard matrix operation). Our actual time cost is larger than Diffusion Distance due to t rounds, even when each round is parallel. With parallel execution, the wall-clock overhead is modest.

We report the computation time cost in Figure 60. Diffusion distance is the fastest as it only needs to power the $D^{-1}A$, where A is the adjacency matrix (with self loop) of symmetric k-NN graph and D is the inverse of the degree matrix. And the pairwise distance can be computed directly in the matrix form. Compared with GF, AGD-SS achieves competitive or superior runtime across all tested configurations, with particularly strong performance on larger datasets where the ensemble strategy provides substantial savings. AGD-SS scales sub-quadratically with n by operating on small random subsamples of size ψ. The advantages of AGD-SS over Diffusion distance come with the price of more computation time.

E.11 Multidimensional scaling of a noisy circle

We show the MDS results of a noisy circle in different ambient dimensions ($\mathbb{R}^{50}$, $\mathbb{R}^{100}$, $\mathbb{R}^{200}$) in Table 2-4, with corresponding Loop Detection Score in Figure 61. AGD-SS, AGD and Diffusion distance perform the best among all the methods, giving similar Loop Detection Scores across different noise stds and ambient dimensions.

Diffusion distance's preservation of geodesic rank hinges critically on the hyperparameter m (the number of diffusion steps), which balances the encoding of local versus global structure. Small m retains local manifold details but is sensitive to noise, while large m captures global structure but may smooth out signal Moon et al. (2019). We validate this on a loop embedded in $\mathbb{R}^{50}$ in Figure 62: Diffusion distance with small m ($m < 64$) fails to preserve global geodesic rank: Diffusion distances between many points far from a reference point are all the same, as shown in the first row of Figure 62. Although a larger m (128 or 256) improves global capture, noisy settings lead to the collapse of distances within a rather large local neighborhoods of the reference point. Consequently, local geodesic rank is poorly preserved, as indicated by the marked regions in Figure 62.

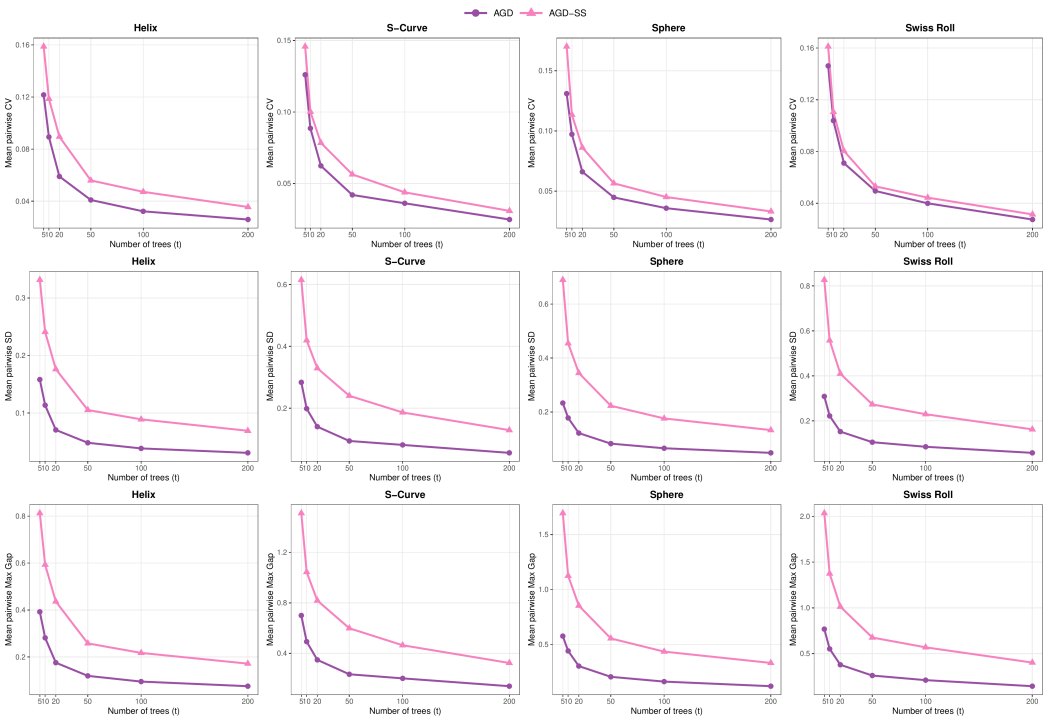


Figure 58: **Pairwise concentration across all $\binom{N}{2}$ point pairs** ($N = 300$, $\psi = 64$, single linkage, $R = 5$ independent runs). Let $\hat{d}_{ij}^{(r)}$ denote the estimated distance between points i and j in the r-th independent run ($r = 1, \ldots, R$). Define the per-pair mean $\mu_{ij} = \frac{1}{R}\sum_{r=1}^{R} \hat{d}_{ij}^{(r)}$, standard deviation $\sigma_{ij} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{d}_{ij}^{(r)} - \mu_{ij})^2}$, coefficient of variation $\mathrm{CV}_{ij} = \sigma_{ij} \,/\, \mu_{ij}$, and max gap $\mathrm{MG}_{ij} = \max_r \hat{d}_{ij}^{(r)} - \min_r \hat{d}_{ij}^{(r)}$. All statistics below are averaged over the $\binom{N}{2}$ pairs. **Top row (Mean Pairwise CV).** $\overline{\mathrm{CV}}(t) = \left(\binom{N}{2}\right)^{-1} \sum_{i<j} \mathrm{CV}_{ij}$, the mean coefficient of variation as a function of the number of trees t. **Middle row (Mean Pairwise SD).** $\bar{\sigma}(t) = \left(\binom{N}{2}\right)^{-1} \sum_{i<j} \sigma_{ij}$, the mean raw standard deviation (unnormalized) across all pairs. **Bottom row (Mean Pairwise Max Gap).** $\overline{\mathrm{MG}}(t) = \left(\binom{N}{2}\right)^{-1} \sum_{i<j} \mathrm{MG}_{ij}$, the mean of the per-pair range (maximum minus minimum) over R runs, measuring the worst-case single-run deviation for a typical pair.

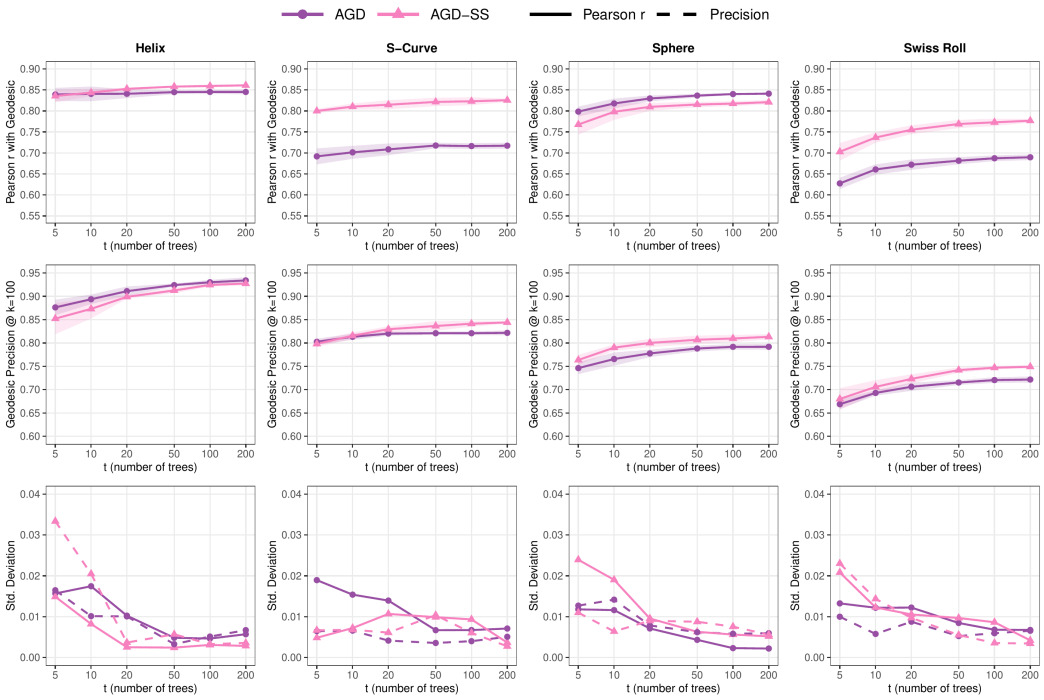


Figure 59: **Concentration of AGD and AGD-SS as a function of t (number of trees)** (5 independent runs). **Top row.** Pearson r with the true geodesic distance matrix. **Middle row.** Geodesic Precision at $k = 100$. **Bottom row.** Standard deviation across the 5 runs. Shaded bands in the top two rows denote standard deviation.

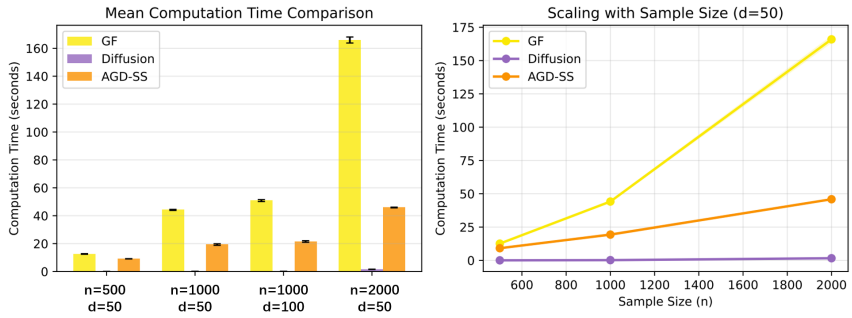


Figure 60: Mean computation time (±std) for GF, Diffusion, and AGD-SS across varying dataset sizes (*Left*) and scaling with sample size n at $d = 50$ (*Right*). Error bars are computed over 3 trials. In each trial, we sample n points from the same noisy circle in Figure 5.

Table 2: Multidimensional scaling of Euclidean, effective resistance, diffusion distances, AGD and AGD-SS for a noisy circle in $\mathbb{R}^{50}$ under different noise levels (σ). Color indicates the distance to the red reference point.

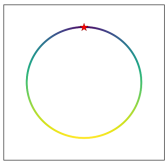
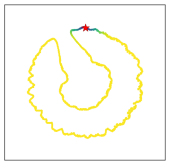
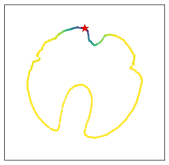
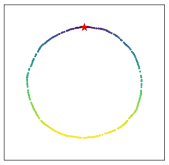
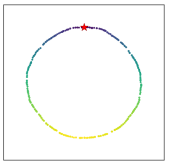
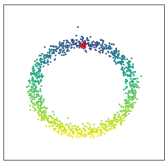
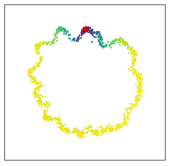
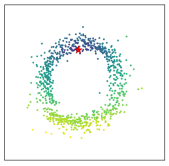
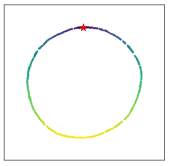
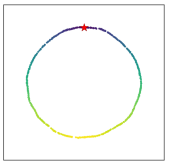
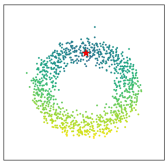
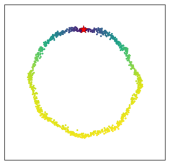
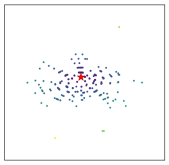
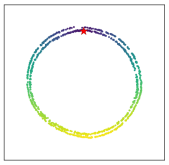
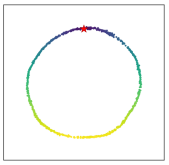
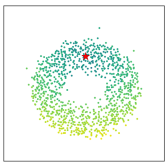
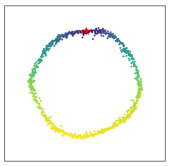
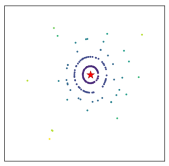
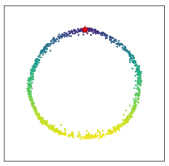
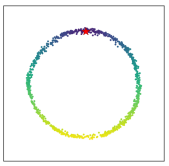
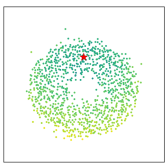
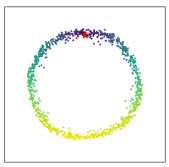
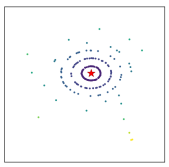
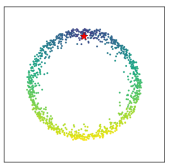
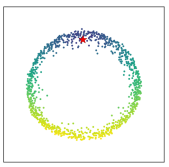
	Euclidean	**Diffusion**	**GF**	**AGD**	**AGD-SS**
$\sigma = 0.00$					
$\sigma = 0.05$					
$\sigma = 0.10$					
$\sigma = 0.15$					
$\sigma = 0.20$					

Table 3: Multidimensional scaling of Euclidean, effective resistance, diffusion distances, AGD and AGD-SS for a noisy circle in $\mathbb{R}^{100}$ under different noise levels (σ). Color indicates the distance to the red reference point.

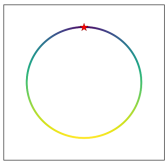
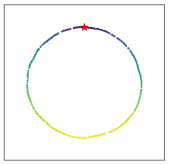
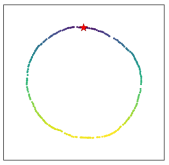
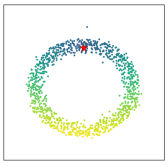
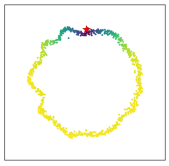
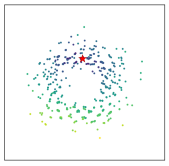
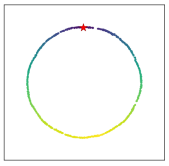
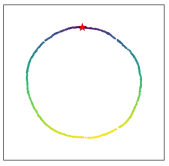
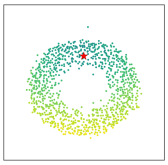
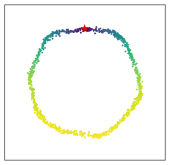
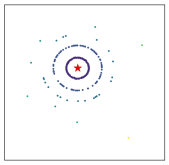
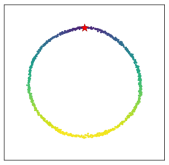
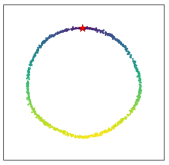
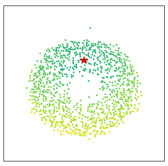
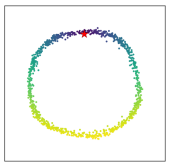
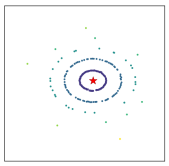
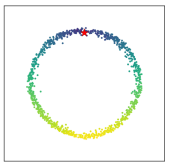
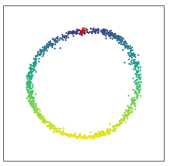
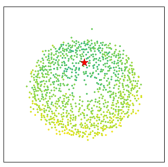
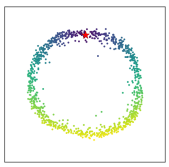
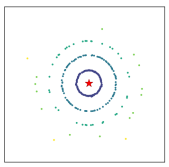
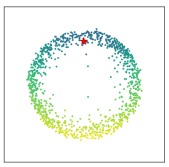
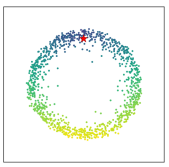
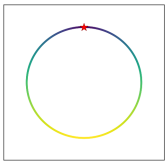
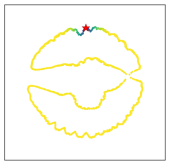
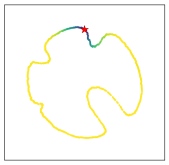
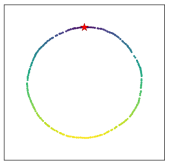
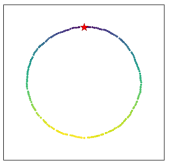
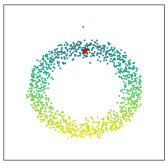
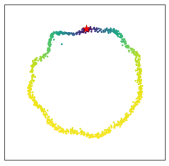
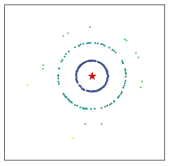
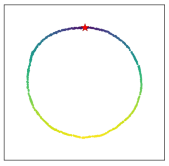
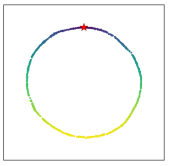
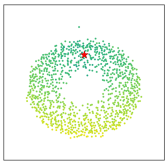
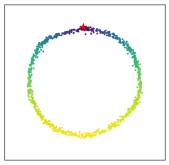
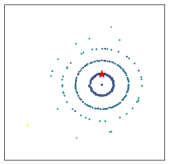
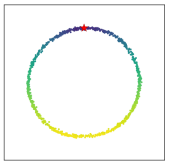
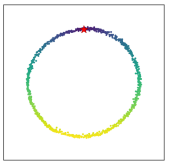
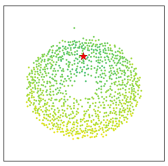
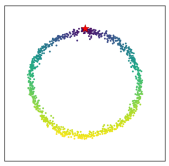
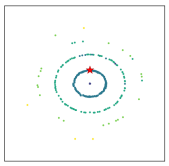
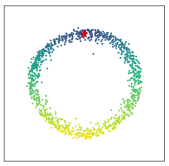
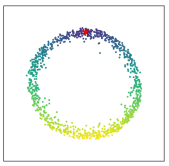
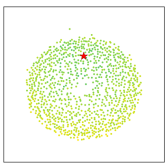
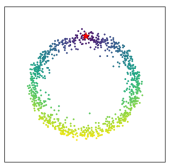
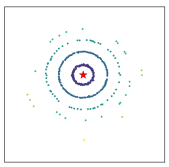
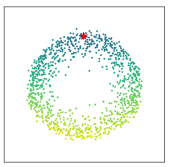
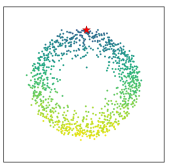
	Euclidean	**Diffusion**	**GF**	**AGD**	**AGD-SS**
$\sigma = 0.00$					
$\sigma = 0.05$					
$\sigma = 0.10$					
$\sigma = 0.15$					
$\sigma = 0.20$					

Table 4: Multidimensional scaling of Euclidean, effective resistance, diffusion distances, AGD and AGD-SS for a noisy circle in $\mathbb{R}^{200}$ under different noise levels (σ). Color indicates the distance to the red reference point.

	Euclidean	**Diffusion**	**GF**	**AGD**	**AGD-SS**
$\sigma = 0.00$					
$\sigma = 0.05$					
$\sigma = 0.10$					
$\sigma = 0.15$					
$\sigma = 0.20$					

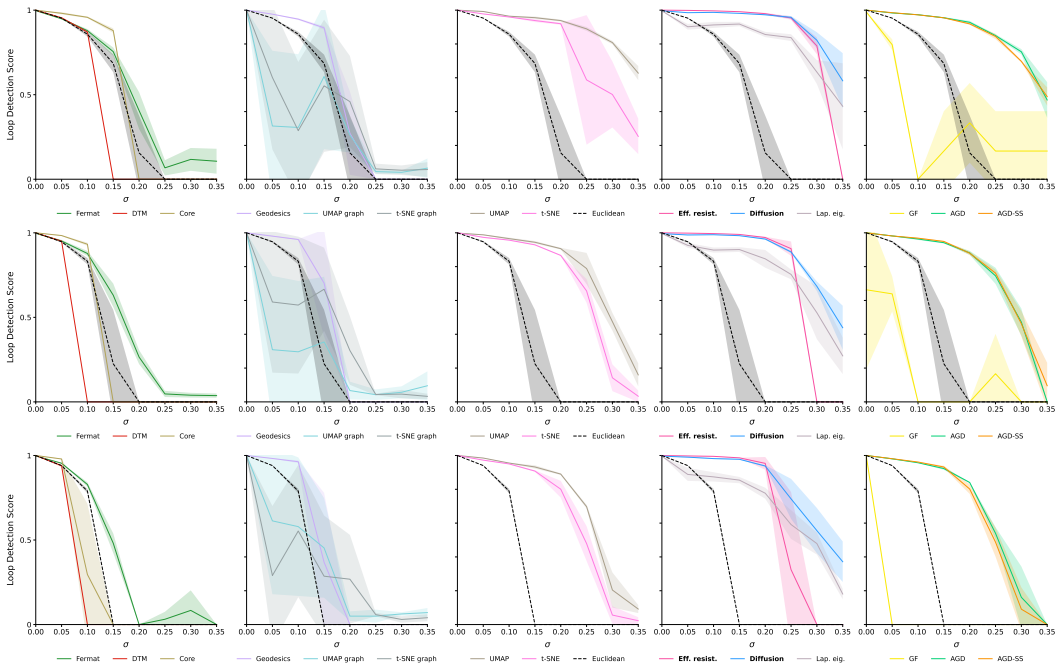


Figure 61: Loop Detection Score for persistent homology with various distances on a noisy circle in $\mathbb{R}^{50}$ (*Top*), $\mathbb{R}^{100}$ (*Middle*) and $\mathbb{R}^{200}$ (*Bottom*). Methods are grouped into panels for visual clarity.

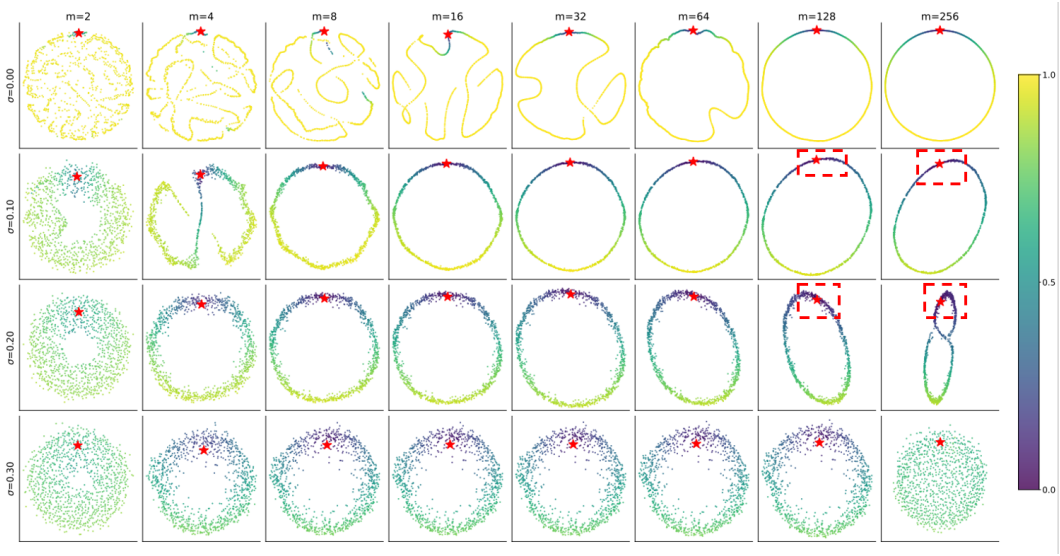


Figure 62: Multidimensional scaling of Diffusion distance with different ms, i.e., the number of diffusion steps, for a noisy circle in $\mathbb{R}^{50}$ under different noise levels (σ). Color indicates the distance to the red reference point. **The marked areas are the regions where the points have almost the same Diffusion distance to the red reference point.**

E.12 Effect of Representative Assignment on Subtree Summation

We examine how changes in nearest representative assignments affect AGD and AGD-SS. Both methods use the same data, sampled representatives, single linkage trees, and assignments. Only the aggregation rule differs, allowing us to isolate the effect of subtree summation from the shared hierarchy and ensemble construction.

Invariance on a fixed tree. Let T be a sampled tree whose structure and node heights are fixed. Suppose the representatives of a query pair change from (a, b) to (a', b'), while

$$\mathrm{LCA}_T(a, b) = \mathrm{LCA}_T(a', b') = u. \tag{4}$$

Let $F_T(u) = \sum_{w \in \mathcal{V}_u} h_T(w)$, both assignments give the same AGD-SS value $F_T(u)$. AGD can change because it depends on the individual paths from the representatives to their LCA[4].

Data and noise. We use Helix, S-Curve, Sphere, Swiss Roll, and Bent Strip, with 200 points per dataset. Each clean point cloud is centered and scaled to unit root mean square distance from its centroid. We append $D \in \{10, 50, 200\}$ independent Gaussian coordinates with standard deviation $\sigma \in \{0.05, 0.20, 0.35\}$, giving 45 manifold and noise conditions. Each condition uses ten paired seeds, $\psi = 32$ representatives sampled without replacement per tree, and $t = 100$ trees, for a total of 45,000 trees. The clean cloud for each manifold and seed is shared across its nine noise settings.

For the supplementary geodesic retrieval measurement, Helix uses its parameter separation, which is proportional to arc length, and Sphere uses angular distance. S-Curve and Swiss Roll use shortest paths on the symmetrized clean 10-NN graph. Bent Strip uses Euclidean distance in its unrolled strip coordinates as the reference ordering.

Paired assignments. For each representative subset, we construct one single-linkage tree from its noisy ambient coordinates. We then compute two assignments to the same representative indices: a reference assignment using Euclidean distances in the clean coordinates and a noisy assignment using Euclidean distances in the noisy ambient coordinates. The reference uses clean observed coordinates, not geodesic distances. We hold the tree and its heights fixed between these assignments. For each method, averaging the resulting statistics over the same 100 trees gives reference and noisy distance matrices. Thus, the experiment isolates sensitivity to assignment changes while retaining a hierarchy constructed from noisy data.

Measurements. The assignment mismatch rate is the fraction of queries whose representative changes, pooled over trees. Among pairs for which at least one assignment changes, we report the fraction retaining the same LCA. Let $\mathcal{C}_\ell$ contain these affected pairs with an unchanged LCA in tree ℓ. For each method, the relative distance change is

$$R = \frac{\sum_{\ell=1}^{t} \sum_{(i,j) \in \mathcal{C}_\ell} |d_\ell^{\mathrm{noise}}(i, j) - d_\ell^{\mathrm{ref}}(i, j)|}{\sum_{\ell=1}^{t} \sum_{(i,j) \in \mathcal{C}_\ell} |d_\ell^{\mathrm{ref}}(i, j)|}. \tag{5}$$

We set $R = 0$ if the denominator is zero. For the full ensemble matrices, we compute Spearman correlation over all unordered distinct query pairs. Neighborhood agreement at $k = 50$ is

$$A_{50} = \frac{1}{n} \sum_{i=1}^{n} \frac{|N_{50}^{\mathrm{ref}}(i) \cap N_{50}^{\mathrm{noise}}(i)|}{50}. \tag{6}$$

Finally, geodesic precision at 50 compares noisy assignment neighborhoods with the clean reference ordering described above. Self neighbors are excluded, and distance ties use the same deterministic, label independent pair hash for both methods. Rank correlation and neighborhood agreement measure stability relative to each method's reference matrix; geodesic precision measures agreement with the external reference ordering.

[4]This property concerns changes in assignments on a fixed tree; the bounded perturbation analysis considers a different setting in which the data used to construct the tree are perturbed.

Table 5: Paired assignment experiment, averaged across 45 conditions and ten seeds. Confidence intervals (CI) are for the paired difference, AGD-SS minus AGD. Lower relative distance change indicates greater stability; higher values are preferred for the other measurements.

Measurement	AGD	AGD-SS	Difference [95% CI]
Relative change, unchanged LCA	0.2325	0.0000	-0.2325 $[-0.2334, -0.2316]$
Distance rank correlation	0.7878	0.8149	$+0.0271$ $[0.0233, 0.0311]$
Neighborhood agreement at 50	0.7691	0.7786	$+0.0096$ $[0.0084, 0.0108]$
Geodesic precision at 50	0.7113	0.6966	-0.0147 $[-0.0160, -0.0134]$

Aggregation and uncertainty. For each metric, we average all 45 conditions within each seed, then compute the paired difference AGD-SS minus AGD. The 95% confidence intervals use the 2.5th and 97.5th percentiles of 10,000 bootstrap resamples of the ten paired seed averages. Conditions and trees are therefore not treated as independent replicates. For individual conditions, the same procedure uses their ten paired seed results. These intervals characterize variation across the tested seeds.

Across condition means, assignment mismatch rates range from 0.058 to 0.784. Among affected pairs, the fraction retaining the same LCA ranges from 14.5% to 90.0%, with a mean of 45.2%. Thus, representative changes that preserve the LCA occur frequently in the tested settings. AGD-SS has zero distance change for these pairs in every run, consistent with the fixed tree identity, while AGD has mean relative change 0.2325.

Table 5 reports the full ensemble measurements over all pairs, including those whose LCA changes. AGD-SS has mean distance rank correlation 0.8149 and neighborhood agreement 0.7786. The paired rank correlation difference is 0.0271, with a 95% interval of $[0.0233, 0.0311]$. Among the 45 individual conditions, 26 intervals lie above zero, five lie below zero, and 14 include zero.

The measured stability concerns preservation of each method's reference ordering. Mean geodesic precision is 0.6966 for AGD-SS and 0.7113 for AGD, showing that stability and agreement with the geodesic reference are distinct properties. **The experiment supports a specific role for subtree summation: it prevents assignment changes from affecting the statistic when the LCA is retained and improves rank stability on average in the tested settings.**