TOPOGRAPH: Topology-Preserving Graph Reduction with Adaptive Structure for Persistent Homology

Zonghao Chen[1], **Yuncheng Jiang**[1], **Gang Li**[2*]

[1]School of Computer Science, South China Normal University, Guangzhou, China
[2]School of Information Technology, Deakin University, Burwood, Australia
zh.chen@m.scnu.edu.cn, ycjiang@scnu.edu.cn, gang.li@deakin.edu.au

Abstract

Topological Data Analysis (TDA) provides artificial intelligence (AI) systems with mathematically rigorous geometric descriptors through Persistent Homology (PH), capturing essential shape characteristics in high-dimensional data. Yet, PH's combinatorial complexity and sensitivity to outliers hinder its scalability and reliability, especially for Intrinsic PH (IPH) that relies on accurate geodesic distances. While state-of-the-art landmark-based subsampling methods, PH Landmarks, ameliorate computational costs and improve outlier robustness by selecting representative points based on local PH scores, it remains computationally intensive and at low sampling rates struggles to reconstruct the global topology. In this work, we introduce TOPOGRAPH, a simple yet powerful framework that preserves intrinsic topology. The resulting coarsened graph supports efficient IPH computations using Fermat distances. Experiments on both synthetic and real-world datasets show that TOPOGRAPH outperforms state-of-the-art sampling-based methods by achieving an order-of-magnitude speedup and substantially improved topological fidelity in persistence diagrams, demonstrating its ability for robust and scalable topological data analysis.

Code — https://github.com/tulip-lab/topograph

Introduction

Topological Data Analysis (TDA) has emerged as a powerful and principled framework for extracting geometric and topological features from complex, high-dimensional data (Xia and Wei 2014; Otter et al. 2017; Chazal and Michel 2021; Dey and Wang 2022). By producing features that do not rely on domain-specific heuristics and remain inherently interpretable. TDA represents a paradigm at the intersection of data science and artificial intelligence AI theory, providing a mathematically-grounded toolkit to address the dual challenges of robustness and explainability. Over the past decade, TDA has transcended its theoretical origins, establishing itself as a robust methodological framework with demonstrably profound utility across a multitude of practical and scientific domains, including protein–ligand binding affinity prediction (Liu et al. 2022), drug discovery (Liu et al. 2021), disease modeling (Wu et al. 2023), image-based tumor classification (Vandaele et al. 2023), materials design (Anand et al. 2022), materials property prediction (Jiang et al. 2021), artificial text detection (Kushnareva et al. 2021), and integrate into the generative model (Gupta, Samaras, and Chen 2024; Xu et al. 2025).

Central to these successes is Persistent Homology (PH) (Edelsbrunner, Letscher, and Zomorodian 2002), which constructs a filtered simplicial complex to capture topological features, such as connected components, loops, and voids, across multiple scales. These features are visualized via a Persistence Diagram (PD) or Barcode (Cohen-Steiner, Edelsbrunner, and Harer 2005), recording each feature's birth and death parameter. Among various simplicial-complex constructions, the Vietoris-Rips (VR) complex is most widely used because it can be built directly from a distance matrix: a k-simplex appears whenever all pairwise distances among its $k+1$ vertices fall below a threshold ε. Compared to deep neural methods, PH offers interpretable, geometry-grounded insights, making it particularly appealing for scientific domains.

However, computing PH is notoriously expensive. The time and space complexity scale rapidly with the number of points and the resulting simplices (Malott, Chen, and Wilsey 2022; Stolz 2023; Zhou, Dong, and Lin 2022). Early algorithms had a worst-case complexity of $O(|K|^3)$, where $|K|$ is the number of simplices, though later advances reduced this to $O(|K|^{2.37})$ (Otter et al. 2017). Unfortunately, $|K|$ still grows exponentially with the number of points. For example, constructing the VR complex for a 3D point cloud with 10000 points requires considering approximately $C^2_{10000} \approx 5.0\times10^7$ 1-simplices and $C^3_{10000} \approx 1.67\times10^{11}$ 2-simplices, as the radius parameter increases to infinity. This scalability bottleneck severely limits the applicability of PH to large-scale data.

Subsampling methods (Chazal et al. 2015; Cao and Monod 2022; Stolz 2023; Malott, Chen, and Wilsey 2022) address PH complexity by selecting landmarks for witness or Vietoris-Rips complexes (De Silva and Carlsson 2004), yet remain sensitive to outliers and disrupt intrinsic geometry (Stolz 2023). Although `PH landmarks` improves robustness via local persistence (Stolz 2023), its high overhead and topological distortions limit effectiveness for IPH (Fernández et al. 2023). As Figure 1 illustrates, these

*Corresponding author.

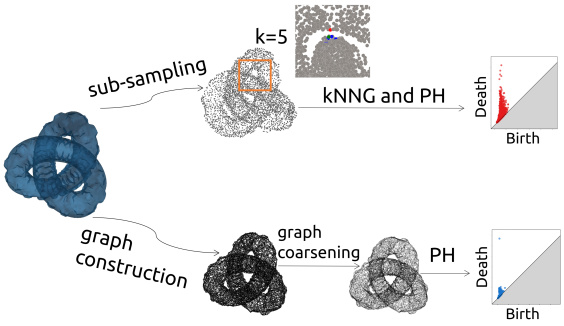

Figure 1: Persistence diagrams of a subsampled and a graph-coarsened torus knot using Fermat distance. In the subsampled case, a green point and its k-nearest neighbors are shown: blue points are both Euclidean- and geodesic-close, while red points are Euclidean neighbors but geodesically distant. These red points reveal that subsampling disrupts geodesic structure, distorting topological inference. In contrast, graph coarsening preserves geodesic relationships and yields more faithful persistence diagrams.

distortions cause kNNGs to create wrong bridges (Berry and Sauer 2019; Chen, van der Smagt, and Cseke 2022), yielding unreliable PH even with small k by failing to capture underlying topology.

Therefore, the research gap lies in the challenge of selecting landmarks that can simultaneously balance (i) computational efficiency, (ii) robustness to outliers, and (iii) faithful preservation of the intrinsic topological structure of the data. To address these challenges, we propose a novel framework, `TOPOGRAPH` (TOPOlogy-preserving Graph Reduction with Adaptive structure for Persistent Homology), for effective and efficient computation of intrinsic persistent homology.

`TOPOGRAPH` integrates outlier removal, graph construction, and graph reduction into a unified pipeline. Specifically, it (i) removes potential outliers using the Distance-to-Measure (DTM) method (Chazal et al. 2018), (ii) constructs an adaptive k-nearest neighbor graph (`AkNNG`) that adjusts neighborhood size according to local density (Cai, Huang, and Yin 2022), (iii) and simplifies the resulting graph via efficient spectral graph coarsening (Deng et al. 2019). The output includes both landmark points and a reduced graph structure, enabling more reliable IPH computations. Experiments show that, compared with the state-of-the-art `PH landmarks`, `TOPOGRAPH` achieves faster runtimes while better preserving intrinsic topological structures.

Related Works

PH computations are notoriously expensive due to the combinatorial growth of simplicial complexes and the high cost of boundary matrix reduction (Malott, Chen, and Wilsey 2022; Stolz 2023). This challenge is further exacerbated in large or noisy datasets, motivating the development of efficient PH computation frameworks. Existing techniques can be broadly classified into three categories:

Matrix Reduction Optimization Methods This category of methods aims to accelerate the most computationally intensive step in persistent homology, boundary matrix reduction, through algorithmic innovations. Key approaches include the twist reduction algorithm (Chen and Kerber 2011), which reorders matrix columns to minimize overhead; cohomology dualities (De Silva, Morozov, and Vejdemo-Johansson 2011), which leverage topological duality to reduce matrix size; and implicit reduction strategies such as clearing and compression in `Ripser` (Chen and Kerber 2011; Bauer 2021), which skip redundant computations. While effective for moderate-sized data, these methods cannot mitigate the exponential simplex growth in large-scale settings, often requiring additional parallel or approximate solutions.

Parallel and Distributed Computation Methods This category of methods enhance the scalability of persistent homology by distributing the construction of filtered complexes and boundary matrix reduction across multiple processors or nodes. Notable frameworks include `Ripser++` (Zhang, Xiao, and Wang 2020), which exploits CUDA-enabled GPUs for fast and exact reductions; `DIPHA` (Bauer, Kerber, and Reininghaus 2014), an MPI-based system that partitions tasks across compute nodes for distributed evaluation; and the `LHF framework` (Malott 2020), which performs localized computations on subregions and merges results to approximate global features. While these methods offer substantial speedups for large or high-dimensional data, they often depend on specialized hardware and complex environments, limiting their accessibility in practice.

Subsampling and Approximation Methods This category of methods aim to reduce the size of filtered complexes either before or after construction to mitigate computational costs. Sparsification techniques such as edge collapse (Malgouyres and Francés 2008; Tancer 2016) simplify existing complexes by pruning simplices while retaining topological features. Landmark-based subsampling selects representative points in advance to curb simplex growth, with variants like random sampling (fast but unreliable for sparse regions (Stolz 2023)), max-min sampling (better coverage but sensitive to noise (De Silva and Carlsson 2004)), and PH Landmarks (Stolz 2023), which leverage local persistent homology for improved robustness and accuracy at higher computational cost. While effective in lowering complexity, these methods often involve trade-offs between efficiency, outlier robustness, and fidelity of topological representation.

TOPOGRAPH

To address the challenges of computational inefficiency, sensitivity to outliers, and the difficulty of preserving intrinsic topological structures in persistent homology computations, we propose `TOPOGRAPH`, a graph-based framework that integrates adaptive graph construction and spectral graph coarsening to significantly enhance the efficiency, robustness, and topological fidelity of persistent homology analysis.

Notation	Description
$\mathbf{X} \in R^{n \times d}$	Data
G	A graph
G_c	A reduced graph
$\mathcal{E}$	Set of graph edges
$\mathbf{A} \in R^{n \times n}$	Adjacency matrix
$\mathbf{S} \in R^{n \times n}$	Similarity matrix
$\mathbf{L} \in R^{n \times n}$	Graph Laplacian matrix
$\mathbf{P}$	Node assignment/mapping matrix
N	Number of nodes
$\mathcal{N}_k(\bullet)$	k-nearest-neighbor of $\bullet$

Table 1: Notations used in this paper

The notations used in this section are summarized in Table 1.

Overview

The architecture of TOPOGRAPH is illustrated in Figure 2, corresponding. addresses three major challenges in intrinsic persistent homology: outlier contamination, faithful neighborhood representation, and combinatorial explosion of simplicial complexes. It comprises three components: (1) DTM-based outlier removal (Chazal et al. 2018), which efficiently filters noisy points with statistical stability; (2) adaptive k-NN graph (AkNNG) construction, which captures local geometry via density-aware connectivity; and (3) spectral graph coarsening, which compresses the graph while preserving topological structure, enabling scalable filtration on the cleaned and simplified representation.

DTM-based Outlier Removal

Persistent homology (PH) can be severely distorted by even a handful of extreme outliers, which introduce spurious homological features and obscure the true topology of the underlying manifold (Blumberg et al. 2014). To obtain a robust filtration, the distance-to-a-measure (DTM) (Chazal et al. 2018) was originally introduced as a noise-stable alternative to distance functions, specifically designed to tolerate outliers.

For a probability measure μ on R^d and mass parameter $m_0 \in (0, 1)$, the DTM at x is defined as

$$d_{\mu,m_0}(x) = \left(\frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(x)^2 \, dm \right)^{1/2}, \qquad (1)$$

where $\delta_{\mu,m}(x)$ is the smallest radius r such that $\mu(B(x, r)) \geq m$. In the discrete setting, a finite point cloud P of size N with uniform measure, this becomes

$$d_{P,k}(x) = \left(\frac{1}{k} \sum_{p \in \mathcal{N}_k(x)} \|x - p\|^2 \right)^{1/2}, \qquad k = m_0 N. \qquad (2)$$

Subsequent theoretical work further established that, under Huber's contamination model (Gu, Akoglu, and Rinaldo 2019), the observed data distribution P is assumed to be a mixture of a normal distribution P_0 and an anomalous (contaminating) distribution P_1:

$$P = (1 - \varepsilon)P_0 + \varepsilon P_1, \qquad 0 \leq \varepsilon < 1. \qquad (3)$$

In this setting, anomalous points drawn from P_1 provably attain larger DTM values than normal points drawn from P_0, enabling consistent separation even in the presence of adversarial outliers. These results provide theoretical support for using DTM as an outlier score.

In practice, we sort all DTM scores and automatically select a cutoff via the `kneed` algorithm (Satopaa et al. 2011), which detects the elbow of the score curve, removing all points above the knee, avoids manual threshold tuning. This DTM-based filtering runs in $O(N \log N)$ time using standard k-NN data structures.

Adaptive k-Nearest Neighbors Graph

In the second stage of TOPOGRAPH, a graph is constructed to reflect the intrinsic geometry of the cleaned data. Standard k-NN graphs assign uniform degree, often introducing shortcuts across low-density regions and distorting local connectivity, which degrades both spectral coarsening and subsequent PH computation.

We construct an `Adaptive k-Nearest Neighbors Graph` (AkNNG) (Cai, Huang, and Yin 2022), in which each point's degree is adjusted according to local density, up to a global maximum $k_{\max}$. Formally, given pairwise distances d_{ij}, the edge weights are obtained by solving the following expression:

$$\begin{aligned} &\min_S \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{m=0}^{\infty} d_{ij} \, s_{ij}^m, \\ &\text{s.t.} \quad 0 \leq s_{ij} \leq 1, \quad \sum_{j=1}^{n} s_{ij} = 1, \end{aligned} \qquad (4)$$

whose Karush-Kuhn-Tucker solution yields the closed-form weights:

$$s_{ij} = \begin{cases} 1 - \dfrac{(k_{\max} - 1 - \delta)\sqrt{d_{ij}}}{\sum_{\ell=1}^{k_{\max}} \sqrt{d_{i\ell}}}, & x_j \in \mathcal{N}_k(x_i), \\ 0, & \text{otherwise}, \end{cases} \qquad (5)$$

where $\delta (\leq 0)$ is a parameter controlling graph sparsity.

AkNNG's density-aware connectivity preserves meaningful local structures and avoids false inter-cluster links, improving topological fidelity in downstream PH. Moreover, its construction retains the same $O(n \log n)$ complexity as standard kNN (Cai, Huang, and Yin 2022), making it scalable for large point clouds while enhancing robustness to non-uniform sampling.

Spectral Graph Coarsening

Even after constructing the AkNNG, large datasets yield massive graphs that render Fermat distance and PH computation intractable. To address this, we apply a coarsening strategy that significantly reduces node count while preserving geodesic distances and local connectivity, ensuring faithful recovery of intrinsic topology.

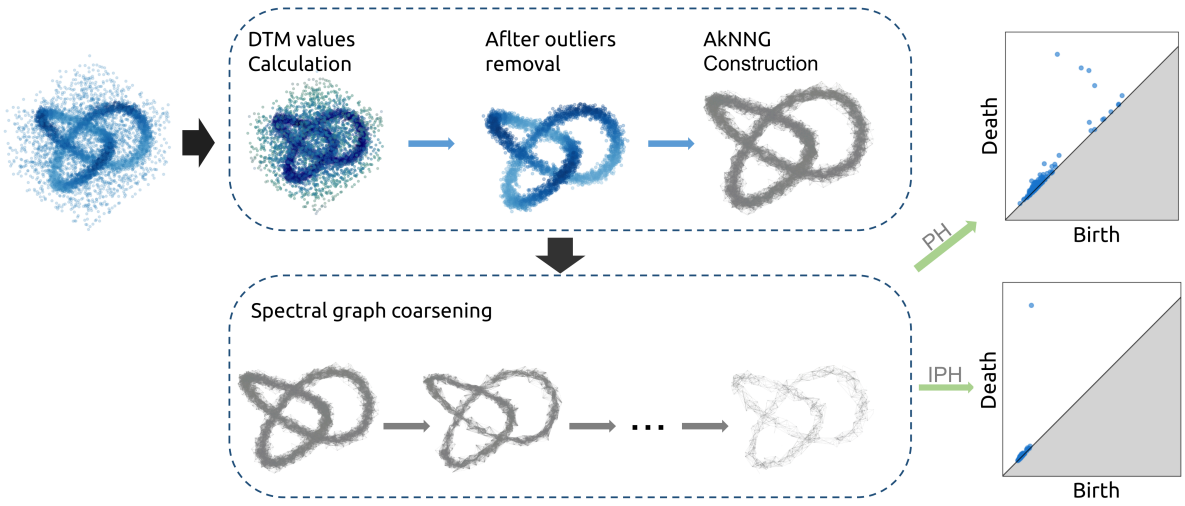

Figure 2: TOPOGRAPH pipeline.

We apply spectral graph coarsening (Hashemi et al. 2024) to compress the AkNNG with adjacency matrix A into a smaller graph G_c with adjacency matrix A_c and $(N_c \ll N)$ super-nodes. Concretely, we define a surjective node mapping $(P \in \{0,1\}^{N \times N_c})$ and compute

$$A_c = PAP^\top \tag{6}$$

$$L_c = PLP^\top \tag{7}$$

where L and L_c are the Laplacians of G and G_c, respectively.

Specifically, We employ `GraphZoom` (Deng et al. 2019), which avoids costly eigen-decomposition by employing iterative low-pass filtering and clustering. First, it constructs the filter:

$$F = D^{-\frac{1}{2}}(D - L + \lambda I)D^{-\frac{1}{2}}, \tag{8}$$

where L is the graph Laplacian, D its degree matrix, and $\lambda = 0.1$ the regularization parameter.

Applying F^p times to random test vectors $\{v^{(i)}\}$ yields:

$$V = F^p[v^{(1)}, \ldots, v^{(7)}], \tag{9}$$

where each $v^{(i)}$ is orthogonalized against $\mathbf{1}$. Node affinities are computed as:

$$\begin{aligned} a_{i,j} &= \frac{|(\boldsymbol{V}_{i,:}, \boldsymbol{V}_{j,:})|^2}{\|V_{i,:}\|^2 \|\boldsymbol{V}_{j,:}\|^2}, \\ (\boldsymbol{V}_{i,:}, \boldsymbol{V}_{j,:}) &= \sum_{k=1}^{t} \left(\boldsymbol{x}_i^{(k)} \cdot \boldsymbol{x}_j^{(k)}\right) \end{aligned} \tag{10}$$

Finally, nodes with a_{ij} above a threshold are merged into super-nodes via the resulting projection matrix until the graph reaches the desired size.

Graph coarsening via GraphZoom offers key advantages: it preserves essential low-frequency Laplacian modes to maintain topological features (Deng et al. 2019); scales efficiently with runtime $O(|\mathcal{E}|)$, leveraging low-pass filtering on sparse graphs instead of costly eigen-decomposition; and enables fast intrinsic (Fermat) distance computation by converting it into weighted shortest-path problems on the coarsened graph.

After graph coarsening, the reduced graph provides a reliable and geometry-aware structure for computing IPH. Instead of using Euclidean distances, IPH replaces pairwise distances with the Fermat distance (Fernández et al. 2023):

$$d^p_{fermat}(x_i, x_j) = \inf_{\gamma} \sum_{i=0}^{r} |x_{i+1} - x_i|^p \tag{11}$$

where the *infimum* is taken over all paths $\gamma = (x_0, x_1, ..., x_{r+1})$ with $x_0 = x$, $x_{r+1} = x_j$. The parameter p adjusts the robustness to outliers, but may introduce manifold deformation, and computing the shortest paths on the kNN graph can improve efficiency (Fernández et al. 2023).

In practice, the choice of distance metric depends on the downstream task. It should be noted that our framework, while supporting Euclidean distance, also accommodates the Fermat distance, which respects the intrinsic geometry of the data.

Empirical Analysis

`TOPOGRAPH` is complementary to existing acceleration techniques such as sparsification methods, distributed methods, and GPU-based implementations. It produces a topology-preserving coarsened graph that can be directly fed into standard PH pipelines, providing additional speedups without modifying downstream tools.

This section presents a comprehensive evaluation of the proposed `TOPOGRAPH` framework by addressing three key research questions:

1. **Topological fidelity under outlier removal:** Does `TOPOGRAPH` effectively filter outliers while preserving the essential structural features required for accurate persistence diagrams?

Dataset	#Points	#Dimension	Topological Features
Three-rings	9,000	2	Three concentric circles (H_1=3)
Eyeglasses	6,000	2	One circle (H_1=1)
LinkedCircle	5,000	3	Two circles(H_1=2)
3D-Torus	20,000	3	Two circles (H_1=2)
Knot	478,704	3	$(3, 2)$-torus knot, one circle (H_1=1)
Neural IPCs	26,625	10	one circle (H_1=1)

Table 2: Summary of datasets used in our experiments.

2. **Efficiency:** Does it deliver competitive runtime performance in its key components of DTM-based outlier filtering, AKNNG construction, and spectral graph coarsening?

We conduct extensive experiments on four synthetic datasets and two real-world datasets, evaluating both quantitative and qualitative aspects of topological preservation and computational performance.

Experimental Settings

Benchmark Datasets We evaluate our method on four synthetic benchmarks with known topological structures (Fernández et al. 2023; Damrich, Berens, and Kobak 2024) and two real-world datasets: the `Knot` dataset (Cao and Monod 2022), which exhibits complex geometric structure, and the `Neural IPCs` dataset (Damrich, Berens, and Kobak 2024; Braun et al. 2023), which is characterized by substantial noise. All datasets are summarized in Table 2.

To assess robustness across low and high dimensional settings, each synthetic dataset is evaluated in two forms: an original 2D/3D version with injected Gaussian noise and outliers, and a 50-dimensional embedded version[1]. In the high-dimensional setting, Gaussian noise is intentionally removed to isolate the effect of dimensionality. This setup provides an initial assessment of the robustness of `TOPOGRAPH` and `PH landmarks` in high-dimensional regimes. In both cases, outliers constitute 10% or 30% of the total samples.

Baselines We compare `TOPOGRAPH` with `PH landmarks`[2], a state-of-the-art outlier-robust subsampling method for persistent homology that has demonstrated clear advantages over `random sampling` and `MaxMin` strategies, particularly in outlier-contaminated settings (Stolz 2023).

`PH landmarks` has two variants: `PH landmarks-I`, which is more effective under unstructured noise or low-to-moderate signal density, and `PH landmarks-II`, which performs better when signal density is high or when one-dimensional features dominate. Since our datasets mainly involve unstructured noise, we adopt `PH landmarks-I` as the baseline.

Evaluation Metrics We evaluate performance using three metrics that capture complementary aspects of topological data reduction.

Residual Outliers counts the number of true outliers remaining after reduction, reflecting the effectiveness of outlier removal and its impact on preserving meaningful topology. **Runtime Performance** measures the total computation time of the pipeline, indicating efficiency and scalability. **Bottleneck Distance** (W_∞) quantifies the maximum discrepancy between persistence diagrams of the reduced and full datasets, and is a standard metric for assessing topological fidelity (Efrat, Itai, and Katz 2001).

Formally, the bottleneck distance between two persistence diagrams is defined as a variant of the Wasserstein distance with the ℓ_∞ norm:

$$\mathcal{W}_\infty\big(\mathrm{PD}(X), \mathrm{PD}(Y)\big) = \inf_{\phi} \max_{j} \|q_j - \phi(q_j)\|_\infty , \quad (12)$$

where $q_j = (b_j, d_j) \in \mathrm{PD}(X) \cup \mathrm{PD}(Y)$ and $\phi : \mathrm{PD}(X) \to \mathrm{PD}(Y)$ denotes a bijection between diagram features.

For synthetic datasets, the ground-truth persistence diagram is obtained by constructing a geodesic (shortest-path) filtration on the clean, outlier-free data.

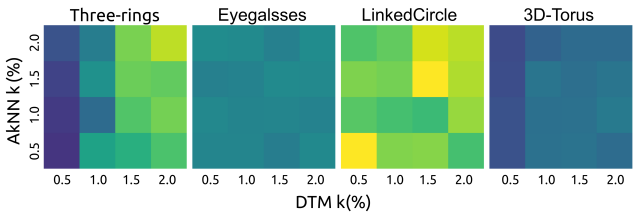

Figure 3: Hyperparameter sensitivity of `TOPOGRAPH` on the four synthetic datasets. Each heatmap reports the bottleneck distance between the persistence diagrams from the full data and from `TOPOGRAPH`, as a function of the DTM neighborhood size k and the AkNNG maximum degree $k_{\max}$, both varied independently over $0.5\%, 1.0\%, 1.5\%, 2.0\%$.

Parameter Settings `TOPOGRAPH` sets the DTM neighborhood size (k) and the maximum AkNNG degree ($k_{\max}$) as small fractions of the dataset size. For synthetic benchmarks, we use $k = k_{\max} = 0.5\%$, which yields sufficiently dense local neighborhoods for reliable topology recovery. For the corresponding 50D embeddings, we increase this setting to 2% to compensate for sparsity in high-dimensional space. For the large-scale Knot dataset, a smaller value of 0.1% is adopted to maintain computational efficiency. For Neural IPCs, we consider two settings, 0.5% and 2%, reflecting its moderate sample size and high-dimensional noise.

To assess robustness, we further conduct a sensitivity study with $k = k_{\max} \in \{0.5\,\%,\ 1\,\%,\ 1.5\,\%$ and $2.0\,\%\}$.

For `PH landmarks`, small neighborhood radii are recommended (Stolz 2023). Accordingly, we evaluate $\delta \in \{0.005,\ 0.01,\ 0.1,\ 0.2\}$ to cover both highly local and moderately large neighborhoods, as the optimal scale is not known a priori.

Experimental Environment Due to its large scale, the ground-truth persistence diagram for the `Knot` dataset was

[1]https://github.com/berenslab/eff-ph/blob/main/notebooks/fig_all_methods_on_toy.ipynb

[2]https://github.com/stolzbernadette/Outlier-robust-subsampling-techniques-for-persistent-homology/blob/main/getPHLandmarks.py

Dataset	Out.	Method	Reduction 40%						Reduction 10%					
			Original (2D/3D)			Embedded (50D)			Original (2D/3D)			Embedded (50D)		
			Res. Out. ↓	Time ↓	B. Dist. ↓	Res. Out. ↓	Time ↓	B. Dist. ↓	Res. Out. ↓	Time ↓	B. Dist. ↓	Res. Out. ↓	Time ↓	B. Dist. ↓
Three-rings	0.1	PH-landmarks	138	20	**0.113**	383	42	0.347	30	23	0.332	97	40	0.374
		TOPOGRAPH	**116**	**2.5**	0.179	**0**	**5.7**	**0.134**	**2**	**2.4**	**0.117**	**0**	**5.3**	**0.110**
	0.3	PH-landmarks	619	36	0.323	835	83	0.275	181	38	0.249	370	83	0.345
		TOPOGRAPH	**428**	**3.7**	**0.154**	**0**	**7.4**	**0.151**	**37**	**2.6**	**0.060**	**0**	**6.1**	**0.113**
Eyeglasses	0.1	PH-landmarks	99	26	0.356	**0**	19	0.162	43	28	0.305	**0**	19	0.360
		TOPOGRAPH	**48**	0.97	**0.090**	**0**	**5.4**	**0.046**	**0**	**0.95**	**0.088**	**0**	5.4	**0.048**
	0.3	PH-landmarks	481	53	0.356	**0**	41	0.357	106	40	0.360	**0**	43	0.360
		TOPOGRAPH	**372**	1.2	**0.099**	**0**	**4.6**	**0.129**	**13**	**1.2**	**0.089**	**0**	**5.7**	**0.051**
LinkedCircle	0.1	PH-landmarks	20	12	**0.187**	**0**	21	0.374	9	11	0.191	**0**	19	0.374
		TOPOGRAPH	**12**	**1.4**	0.211	**0**	**5.4**	0.374	**0**	**1.3**	**0.158**	**0**	**5.5**	**0.189**
	0.3	PH-landmarks	231	18	0.374	**0**	40	0.374	84	18	0.375	**0**	43	0.360
		TOPOGRAPH	**99**	**1.7**	**0.149**	**0**	**5.6**	0.374	**3**	**1.6**	**0.222**	**0**	**4.6**	**0.242**
3D-Torus	0.1	PH-landmarks	268	118	**0.048**	**0**	314	**0.044**	119	77	0.064	**0**	314	**0.080**
		TOPOGRAPH	**232**	**6.8**	0.064	**0**	**26**	0.094	**22**	**6.6**	**0.057**	**0**	**24**	0.086
	0.3	PH-landmarks	1643	228	0.142	**0**	614	**0.024**	466	213	0.110	**0**	616	**0.085**
		TOPOGRAPH	**1298**	**12.6**	**0.068**	**0**	**38**	0.094	**178**	**8.4**	**0.050**	**0**	**32**	0.087

Table 3: Performance comparison of TOPOGRAPH and PH-landmarks under different reduction rates on synthetic datasets. Lower is better; best results are highlighted in bold.

computed on a workstation with dual Intel Xeon E5–2680 v4 CPUs (28 cores) and 256 GB RAM. All other experiments were conducted on a laptop with an AMD Ryzen 7 5800U CPU (8 cores) and 16 GB RAM, demonstrating the general applicability and efficiency of TOPOGRAPH. Reported runtimes are averaged over ten runs to reduce variance.

Results on Synthetic Datasets

Comparison with PH landmarks Table 3 shows the performance comparison between TOPOGRAPH and PH-landmarks under different reduction rates and outlier levels. TOPOGRAPH demonstrates markedly stronger robustness across all synthetic datasets. It consistently removes a larger portion of outliers and produces PDs that remain closer to the ground-truth topology under both moderate and heavy contamination levels. In contrast, PH-landmarks exhibits greater sensitivity to dataset characteristics, particularly on shapes with non-uniform sampling density such as *Three-rings*, where radius-based local PH computations are more susceptible to density fluctuations. As a consequence, PH-landmarks tends to leave more residual outliers and yields less stable persistence diagrams compared to TOPOGRAPH.

In the 50D embedded setting, where geometric perturbations are absent and only outliers are introduced, the two methods exhibit distinct behavior. PH-landmarks often succeeds in identifying most, or even all, outliers. However, this aggressive removal strategy could discard structurally informative points, which can distort the subsequent Fermat-distance construction and yield persistence diagrams that deviate from the ground-truth topology. By contrast, TOPOGRAPH, benefiting from its graph-structured coarsening pipeline, preserves more intrinsic connectivity, resulting in persistence diagrams that remain closer to those computed from the full dataset.

Across all datasets, dimensions, and reduction rates, TOPOGRAPH consistently achieves significantly lower runtime. It is typically an order of magnitude faster in low-dimensional settings, and the speed advantage becomes even larger in 50D, where PH-landmarks incurs heavy computational costs due to repeated local PH computations. This demonstrates that the proposed coarsening strategy provides a practical acceleration mechanism for downstream PH pipelines.

Overall, the results show that TOPOGRAPH delivers robust outlier handling, high-fidelity topology preservation, and substantial computational savings across both low- and high-dimensional settings.

Parameter Sensitivity of Bottleneck Distance Figure 3 illustrates the sensitivity analysis over the DTM mass percentage (horizontal axis) and the AKNNG neighborhood upper bound (vertical axis). Across all four datasets, TOPOGRAPH maintains low bottleneck distances over a broad region of the parameter space, indicating that its performance is generally stable with respect to hyperparameter tuning.

Among the datasets, the *Three-rings* case exhibits relatively larger variations in bottleneck distance across parameter settings. This behavior is attributed to its heterogeneous sampling, where substantial local density differences make the DTM scale and neighborhood size more influen-

Computation Stage	Random Sampling	TOPOGRAPH
Data Reduction	1 s	484 s
Fermat Distance Calculation	587 s	42 s
PH Computation (GUDHI, simplicial complex sparse=0.3)	7000 s	3 s
Total Time	7588 s	**529 s**

Table 4: Computation time comparison between random sampling and TOPOGRAPH

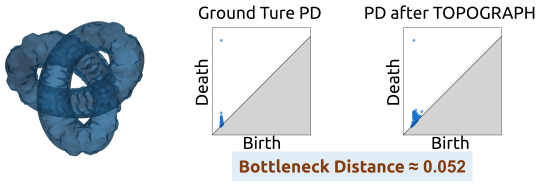

Figure 4: Result on Knot. Left: Original Knot structure. Middle: Ground-truth persistence diagram computed from 30,000 samples with Fermat distance ($k = 5$). Right: Persistence diagram generated by TOPOGRAPH after 99% data reduction. The bottleneck distance between middle and right diagrams is approximately 0.052.

tial. These observations highlight the need for mechanisms that more effectively adapt to local density variations.

Results on Real-word Datasets

Results on `Knot` Directly computing PH on the full `Knot` dataset is infeasible. To obtain a reliable reference, we compute a ground-truth diagram from a random subsample 30,000 points using Fermat distance with k=5. This sample size is sufficient to preserve the intrinsic topological structure, whereas smaller subsamples distort the topology (see Figure 1). The computation takes about two hours.

In contrast, TOPOGRAPH performs principled reduction through AkNNG construction and spectral graph coarsening, reducing the entire dataset to only 1 % of its original size before computing IPH. The resulting diagram (Figure 4) achieves a bottleneck distance of **0.052** from the ground truth, demonstrating strong topological fidelity under aggressive reduction.

Moreover, the full pipeline finishes in under 10 minutes (Table 4), yielding a 12 times speedup over the two-hour reference computation.

Results on `Neural IPCs` Figure 5 presents a visualization of the `Neural IPCs` dataset (Damrich, Berens, and Kobak 2024), along with the corresponding one-dimensional persistence diagrams obtained after applying TOPOGRAPH and `PH landmarks` at two topological neighborhood radius.

The PD produced by TOPOGRAPH reveals a single, dominant H_1 feature, highlighted by the red box in Figure 5, which accurately captures the known cell-cycle-dependent loop in the dataset. Importantly, this H_1 structure is recovered consistently across both $k = k_{\max} = 0.5\%$ and 2%, demonstrating the method's stability with

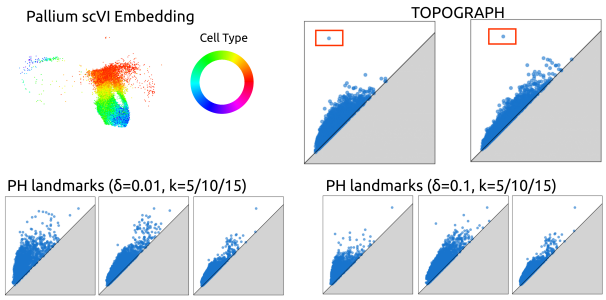

Figure 5: Result on Neural IPCs. Top-left: 2D projection of the Neural IPCs 10-dimensional embedding. Top-right: PD after reduction by TOPOGRAPH with $k = k_{max} = 0.5\%$ and $k = k_{max} = 2\%$. Bottom: PD after PH landmarks with $\delta = 0.01$ and $\delta = 0.1$. TOPOGRAPH cleanly captures the dominant one-dimensional loop (red box).

respect to neighborhood size. Notably, the diagram contains only a few short-lived off-diagonal points, underscoring TOPOGRAPH's robustness to noise and its effectiveness in preserving essential topological features.

In contrast, `PH landmarks` at $\delta = 0.01$ yields a persistence diagram characterized by a proliferation of short-lived bars, suggesting a heightened sensitivity to noise and the presence of spurious topological features. Even at the larger radius $\delta = 0.1$, the method continues to exhibit multiple mid-persistence features that do not correspond to the underlying topology. These observations indicate that `PH landmarks` struggles to isolate meaningful topological structures in the presence of noise, particularly in complex, high-dimensional biological data.

Conclusions

TDA offers interpretable shape analysis for high-dimensional data, enhancing AI with topology-aware representations. However, adoption faces three challenges: computational complexity, outlier sensitivity, and structure preservation during subsampling.

Our contributions are:

1. We introduce TOPOGRAPH, a unified and efficient TDA framework that integrates DTM-based outlier removal, adaptive kNN graph construction, and spectral graph coarsening. The framework enables robust, IPH-compatible data reduction while jointly addressing outlier sensitivity, topological fidelity, and computational efficiency in a hardware-friendly manner.
2. We conduct extensive experiments on synthetic and real-world datasets. The results show that TOPOGRAPH consistently achieves effective outlier removal, accurate topological preservation, and fast computation across diverse data regimes.

Future work includes extending TOPOGRAPH to multi-parameter persistence and to even larger and more complex datasets, where its coarsening strategy may further benefit downstream topological analysis and learning tasks.

Acknowledgments

The works described in this paper are supported by The National Natural Science Foundation of China under Grant No. 62477015; Key Research and Development Program of Guangdong of China under Grant No. 2023B0303010004; The Innovation Team Project for Universities in Guangdong Province in China under Grant No. 2023KCXTD011.

References

Anand, D. V.; Xu, Q.; Wee, J.; Xia, K.; and Sum, T. C. 2022. Topological feature engineering for machine learning based halide perovskite materials design. *npj Computational Materials*, 8(1): 203.

Bauer, U. 2021. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3): 391–423.

Bauer, U.; Kerber, M.; and Reininghaus, J. 2014. Distributed computation of persistent homology. In *2014 proceedings of the sixteenth workshop on algorithm engineering and experiments (ALENEX)*, 31–38. SIAM.

Berry, T.; and Sauer, T. 2019. Consistent manifold representation for topological data analysis. *Foundations of Data Science*.

Blumberg, A. J.; Gal, I.; Mandell, M. A.; and Pancia, M. 2014. Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. *Foundations of Computational Mathematics*, 14: 745–789.

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

Cai, Y.; Huang, J. Z.; and Yin, J. 2022. A new method to build the adaptive k-nearest neighbors similarity graph matrix for spectral clustering. *Neurocomputing*, 493: 191–203.

Cao, Y.; and Monod, A. 2022. Approximating persistent homology for large datasets. *arXiv preprint arXiv:2204.09155*.

Chazal, F.; Fasy, B.; Lecci, F.; Bertr; Michel; Aless; ro Rinaldo; and Wasserman, L. 2018. Robust Topological Inference: Distance To a Measure and Kernel Distance. *Journal of Machine Learning Research*, 18(159): 1–40.

Chazal, F.; Fasy, B.; Lecci, F.; Michel, B.; Rinaldo, A.; and Wasserman, L. 2015. Subsampling methods for persistent homology. In *International Conference on Machine Learning*, 2143–2151. PMLR.

Chazal, F.; and Michel, B. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4: 108.

Chen, C.; and Kerber, M. 2011. Persistent homology computation with a twist. In *Proceedings 27th European workshop on computational geometry*, volume 11, 197–200.

Chen, N.; van der Smagt, P.; and Cseke, B. 2022. Local distance preserving auto-encoders using continuous k-nearest neighbours graphs. *arXiv preprint arXiv:2206.05909*.

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

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

De Silva, V.; and Carlsson, G. E. 2004. Topological estimation using witness complexes. In *PBG*, 157–166.

De Silva, V.; Morozov, D.; and Vejdemo-Johansson, M. 2011. Dualities in persistent (co) homology. *Inverse Problems*, 27(12): 124003.

Deng, C.; Zhao, Z.; Wang, Y.; Zhang, Z.; and Feng, Z. 2019. Graphzoom: A multi-level spectral approach for accurate and scalable graph embedding. *arXiv preprint arXiv:1910.02370*.

Dey, T. K.; and Wang, Y. 2022. *Computational topology for data analysis*. Cambridge University Press.

Edelsbrunner; Letscher; and Zomorodian. 2002. Topological Persistence and Simplification. *Discrete Comput. Geom.*, 28(4): 511–533.

Efrat, A.; Itai, A.; and Katz, M. J. 2001. Geometry helps in bottleneck matching and related problems. *Algorithmica*, 31: 1–28.

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

Gu, X.; Akoglu, L.; and Rinaldo, A. 2019. Statistical analysis of nearest neighbor methods for anomaly detection. *Advances in Neural Information Processing Systems*, 32.

Gupta, S.; Samaras, D.; and Chen, C. 2024. Topodiffusionnet: A topology-aware diffusion model. *arXiv preprint arXiv:2410.16646*.

Hashemi, M.; Gong, S.; Ni, J.; Fan, W.; Prakash, B. A.; and Jin, W. 2024. A Comprehensive Survey on Graph Reduction: Sparsification, Coarsening, and Condensation. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Jiang, Y.; Chen, D.; Chen, X.; Li, T.; Wei, G.-W.; and Pan, F. 2021. Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *npj computational materials*, 7(1): 28.

Kushnareva, L.; Cherniavskii, D.; Mikhailov, V.; Artemova, E.; Barannikov, S.; Bernstein, A.; Piontkovskaya, I.; Piontkovski, D.; and Burnaev, E. 2021. Artificial Text Detection via Examining the Topology of Attention Maps. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 635–649. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Liu, X.; Feng, H.; Wu, J.; and Xia, K. 2022. Dowker complex based machine learning (DCML) models for protein-ligand binding affinity prediction. *PLOS Computational Biology*, 18(4): 1–17.

Liu, X.; Wang, X.; Wu, J.; and Xia, K. 2021. Hypergraph-based persistent cohomology (HPC) for molecular representations in drug design. *Briefings in Bioinformatics*, 22(5): bbaa411.

Malgouyres, R.; and Francés, A. R. 2008. Determining whether a simplicial 3-complex collapses to a 1-complex is NP-complete. In *Discrete Geometry for Computer Imagery: 14th IAPR International Conference, DGCI 2008, Lyon, France, April 16-18, 2008. Proceedings 14*, 177–188. Springer.

Malott, N. O. 2020. *Partitioned persistent homology*. Master's thesis, University of Cincinnati.

Malott, N. O.; Chen, S.; and Wilsey, P. A. 2022. A survey on the high-performance computation of persistent homology. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4466–4484.

Otter, N.; Porter, M. A.; Tillmann, U.; Grindrod, P.; and Harrington, H. A. 2017. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6: 1–38.

Satopaa, V.; Albrecht, J.; Irwin, D.; and Raghavan, B. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, 166–171. IEEE.

Stolz, B. J. 2023. Outlier-robust subsampling techniques for persistent homology. *Journal of Machine Learning Research*, 24.

Tancer, M. 2016. Recognition of collapsible complexes is NP-complete. *Discrete & Computational Geometry*, 55: 21–38.

Vandaele, R.; Mukherjee, P.; Selby, H. M.; Shah, R. P.; and Gevaert, O. 2023. Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction. *Patterns*, 4(1).

Wu, S.; Liu, X.; Dong, A.; Gragnoli, C.; Griffin, C.; Wu, J.; Yau, S.-T.; and Wu, R. 2023. The metabolomic physics of complex diseases. *Proceedings of the National Academy of Sciences*, 120(42): e2308496120.

Xia, K.; and Wei, G.-W. 2014. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30(8): 814–844.

Xu, M.; Gupta, S.; Hu, X.; Li, C.; Abousamra, S.; Samaras, D.; Prasanna, P.; and Chen, C. 2025. TopoCellGen: Generating Histopathology Cell Topology with a Diffusion Model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, S.; Xiao, M.; and Wang, H. 2020. GPU-Accelerated Computation of Vietoris-Rips Persistence Barcodes. In *36th International Symposium on Computational Geometry (SoCG 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Zhou, C.; Dong, Z.; and Lin, H. 2022. Learning persistent homology of 3D point clouds. *Computers & Graphics*, 102: 269–279.