CLC _____ Number _____

UDC _____ Available for reference ☐Yes ☐No

SUSTech Southern University of Science and Technology

Undergraduate Thesis

Thesis Title: **A Comparative Analysis of Subsampling Strategies in Persistent Homology Computation**

Student Name: **Runfeng Yang**

Student ID: **12212840**

Department: **Department of Mathematics**

Program: **Mathematics**

Thesis Advisor: **Yifei Zhu**

COMMITMENT OF HONESTY

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.

2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the study of the paper have been indicated clearly in the paper.

3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.

4. If there is violation of any intellectual property right, I will take legal responsibility myself.

Signature: 杨润锋

Date:2026/04/29

A Comparative Analysis of Subsampling Strategies in Persistent Homology Computation

[ABSTRACT]:

We study the effect of landmark subsampling on the preservation of topological structure in persistent homology. To quantify fidelity, we compare persistence diagrams of full datasets and their sketches using Wasserstein distance, assisted by a geometric-feature-based metric.

We evaluate several subsampling methods on synthetic datasets with known topology and a natural image dataset exhibiting Klein bottle structure. Experiments examine performance under varying compression levels and noise, as well as the potential denoising effect of subsampling.

Our results reveal clear trade-offs between methods in terms of fidelity and stability, and provide practical guidance for selecting sketching strategies in topological data analysis.

[Keywords]: persistent homology; subsampling; Wasserstein distance; topological data analysis

[摘要]：本文研究采样对持续同调拓扑结构保持性的影响。为量化保真度，比较完整数据集与其草图数据集的持续图，并采用 Wasserstein 距离作为主要评价指标，辅助使用基于几何特性的一种指标。

我们在已知拓扑结构的合成数据集及具有克莱因瓶结构特征的自然图像数据集上评估多种子采样方法，系统考察不同压缩率与噪声强度下的表现，并分析子采样可能的“去噪”效应。

实验结果表明，各方法在保真度、稳定性与鲁棒性方面存在明显差异，并随采样比例而产生性能的变化，可为拓扑数据分析中的草图策略选择提供实践参考。

[关键词]：持续同调；抽样； Wasserstein 距离；拓扑数据分析

Contents

1. Introduction

Persistent homology is an extremely powerful tool to assess the topology of a structure from a dataset. Its computation is nonetheless very expensive in time and space, as in [1], in particular if we consider dense datasets. But it then suffices a concrete need to choose a representative subset of points, landmarks, before computing the persistence homology. Many landmark selection strategies have been proposed and formulated, such as random sampling [6], farthest-point sampling, clustering-based approaches [3], and density-based approaches. More recently, a method using local topology information has also been proposed. Before subsampling can significantly reduce the computational cost, it still remains an open question on how different methods affect the preservation of true topological structure.

Such methods of persistent homology can sometimes seem mysterious in this paper, we explain a principled method to study the faithfulness of subsampling methods. Instead of considering one data set or criterion, we consider a whole collection of synthetic data sets with known topological struc- ture, plus a natural image data set with the topology of a Klein bottle [5]. Using this, we can evaluate subsampling methods over a range of geometric and topological regimes. To study the effect of subsampling, we compare persistence diagrams of the full data set with those of the sample sketch using Wasserstein distance. We examine methods over a range of levels of compression and noise, and also consider if the methods might be able to have a denoising effect.

These results highlight clear tradeoffs between different subsampling strategies in terms of fidelity, stability and robustness. For example, we see that those subsampling strategies defined in terms of geometric coverage or geometric features tend to better preserve persistence for aggressive subsampling and noise. Such observations provide insight into practical application of landmark-based sketches for TDA.

2. Background

2.1 Simplicial Complexes

Given a dataset, a simplicial complex provides a combinatorial representation of the geometric and topological structure of a dataset. Let V be a finite set of vertices. A *simplicial complex* K on V is a collection of subsets $\sigma \subseteq V$ such that if $\sigma \in K$ and $\tau \subseteq \sigma$, then $\tau \in K$. Each $\sigma \in K$ is called a *simplex*.

A k-simplex is a simplex $\sigma = \{v_0, \ldots, v_k\}$ with $k+1$ vertices. Geometrically, a 0-simplex is a point, a 1-simplex is an edge, a 2-simplex is a filled triangle, and higher-dimensional simplices generalize this construction. The dimension of a simplicial complex K is defined as

$$\dim K = \max_{\sigma \in K}(|\sigma| - 1).$$

Given a point cloud $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, a simplicial complex can be built from points, by introducing simplices based on proximity relations between points. In this paper, we are only concerned with the Vietoris–Rips complex. For a scale parameter $\varepsilon > 0$, the Vietoris–Rips complex $\mathrm{VR}_\varepsilon(X)$ is defined as

$$\mathrm{VR}_\varepsilon(X) = \{\sigma \subseteq X \mid \|x_i - x_j\| \leq \varepsilon \text{ for all } x_i, x_j \in \sigma\}.$$

That is, a simplex is included whenever all its vertices are pairwise within distance ε.

Simple, the complex is useful because it only depends on pairwise distances, which is then easy to obtain from the raw data, with the advantage and disadvantage being the fast growth of computation and storage cost as data grow. This motivates the use of subsampling methods to reduce the size of the data before constructing the complex.

2.2 Persistent Homology

Persistent homology gives a multiscale description of the topology of a dataset by monitoring the persistence of homology features in a filtration. Given a finite point cloud $X \subset \mathbb{R}^d$, one constructs a filtered simplicial complex $\mathcal{K}\epsilon(X)\epsilon \geq 0$, such as the Vietoris–Rips complex,

where simplices are added as the scale parameter ϵ increases.

For each scale ϵ, the complex $\mathcal{K}\epsilon(X)$ has associated homology groups $H_k(\mathcal{K}\epsilon(X))$, whose rank counts the number of k-dimensional topological features, such as connected components for $k = 0$ and loops for $k = 1$. As ϵ increases, these features may appear and disappear, leading to intervals (b, d) that record their birth and death scales.

The collection of these intervals forms the persistence diagram, which encodes the lifetime of features of different dimensions over the scales. Features with long lifetime are often viewed as giving rise to meaningful underlying structure and short-lived features are often considered noise. In practice, persistent homology is computed using Vietoris Rips complexes.

In practice, persistent homology is commonly computed using Vietoris–Rips complexes. However, the number of simplices in $\mathrm{VR}_\epsilon(X)$ grows combinatorially with $|X|$, as a k-simplex corresponds to a $(k + 1)$-tuple of points with pairwise distances below ϵ. This fast growth brings computational and memory costs, so several subsampling or landmark methods are used to alleviate this.

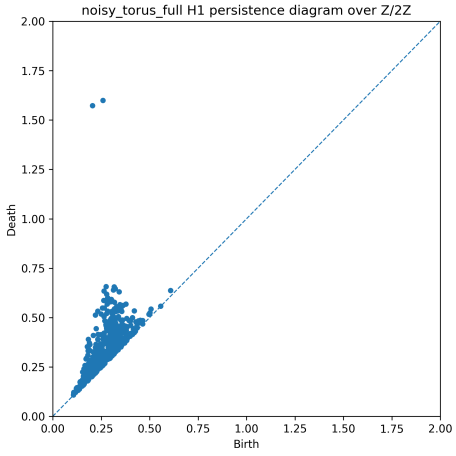

Figure 1. H_1 persistence diagram for a torus adjoined with Gaussian noise in coefficient field $\mathbb{Z}_2$. Two prominent points on top denote the 2 nontrivial homology classes.

2.3 Wasserstein Distance Between Persistence Diagrams

To compare the topology of a full dataset to that of a sketch, we need a metric on the persistence diagrams. In this work, we use the Wasserstein distance between persistence di-

agrams, which is a typical choice in TDA for measuring the discrepancy between the outputs of persistent homology [3, 2].

Let D_1 and D_2 be two persistence diagrams. A point $p = (b, d) \in D_i$ represents a homological feature born at scale b and dying at scale d. The persistence of this feature is

$$\text{pers}(p) = d - b.$$

Long-persisting points lie far from the diagonal

$$\Delta = \{(x, x) : x \in \mathbb{R}\},$$

while short-lived features lie close to Δ.

The q-Wasserstein distance between D_1 and D_2 is defined by

$$W_q(D_1, D_2) = \left(\inf_{\gamma} \sum_{p \in D_1} \|p - \gamma(p)\|_\infty^q \right)^{1/q},$$

where γ ranges over bijections between $D_1 \cup \Delta$ and $D_2 \cup \Delta$. The diagonal is included with infinite multiplicity, allowing features in one diagram to be matched either to features in the other diagram or to the diagonal. Matching a point to the diagonal corresponds to treating that feature as absent from the other diagram.

Wasserstein distance is well suited for this task of comparison of sketches with a full dataset if a sketch is a good approximation of the dominant features of the full dataset: the corresponding high-persistence points in both diagrams should be matched with small cost.As opposed to this,if a sketch loses some interesting feature,or vice versa,producing a false persistence of features at the wrong death point,there should be a larger cost for matching.Thus,Wasserstein distance measures both shifts in the birth–death scales and differences in the number of topological features of interest.

In our experiments, we compute persistence diagrams for the full dataset and for each subsampled sketch, then use 2-Wasserstein distance as the primary measure of topological difference. A smaller value indicates that the sketch more faithfully preserves the persistent

homology of the reference dataset. For repeated randomized subsampling methods, we also calculate the mean and standard deviation of the Wasserstein distance across runs.

2.4 Computational Complexity and Subsampling

A central problem in the computation of persistent homology is the vast increase in size of the simplicial complexes constructed over a point cloud. For a dataset X with $n = |X|$ points, the Vietoris–Rips complex $\mathrm{VR}_\epsilon(X)$ contains a k-simplex for every $(k+1)$-tuple of points whose pairwise distances are bounded by ϵ.

In the worst case, this leads to

$$\#\{k\text{-simplices}\} = \binom{n}{k+1},$$

and hence the total number of simplices grows exponentially with n. As a consequence, both the time and memory complexity of persistent homology computation scale combinatorially with the size of the input.

As a consequence, both the time and memory cost of the persistent homology computation are combinatorially dependent on the size of the input. In practice, this growth is already prohibitive for any form of practical use. Even for relatively small data sizes of, say, $n \approx 1e4$, direct computation can be extremely time consuming. There are several approaches to tackle this complexity. Algorithmic optimisations, such as those implemented in Ripser [1], exploit sparsity and minimise the number of simplices that need to be considered during computation, making use of additional information such as persistence lifetimes. Other approaches replace the Vietoris–Rips complex with alternative constructions, including sparsified filtrations [6]. Despite these advances, the computational cost remains large for large or dense data sets and becomes even larger when persistent homology must be computed repeatedly. In such situations, it is useful to reduce the size of the input data set. A common strategy is to replace the full data set X by a smaller subset $X' \subset X$ of size $k \ll n$, referred to as a set of landmarks. The complex $\mathrm{VR}(X')$ is substantially smaller, and its complexity depends on k rather than n. For example, reducing the data size from n to $2k$ changes the worst-case

number of simplices from $\binom{n}{d+1}$ to $\binom{2k}{d+1}$, yielding substantial computational savings.

But this reduction also raises a natural question: to what extent is the persistence computed on the landmarks stable with respect to the topological structure of the original dataset? To find out the balance between computational efficiency and structural fidelity is, in essence, the question we address in this work. Methods

3. Methods

3.1 Synthetic Datasets

To gain an understanding of subsampling methods in a simple controlled setting, we consider a collection of synthetic point clouds with known geometric and topological structure. These data include standard examples such as circles, spheres, and tori, as well as more complicated constructions including wedge spaces and quotient spaces (for example, $\mathbb{RP}^3$).

Each data set is generated by sampling points from a prescribed geometric model and, in some cases, embedding the resulting space into a Euclidean ambient space. We also consider noisy variants obtained by perturbing the sampled points with Gaussian noise of varying magnitude. Specifically, we introduce ambient Gaussian noise by applying perturbations of the form

$$x \mapsto x + \eta, \qquad \eta \sim \mathcal{N}(0, \varepsilon^2 I),$$

where the noise acts in the ambient space rather than intrinsically on the underlying manifold. These synthetic data sets provide a controlled environment in which the ground-truth topology is known, allowing us to systematically evaluate the extent to which different subsampling methods preserve topology under increasing levels of compression and noise.

3.2 Natural Image Pack Dataset

As a simpler example, we also work with a collection of high-contrast natural image patches of fixed dimen- sion coming from grayscale images and stored on disk as chunked arrays. This dataset was taken from the image patch model developed in [5], a canonical example of TDA with a non-trivial structure. For each patch, one computes a representation

for the patch as a point in a Euclidean feature space after normalizing the patches. As in standard preprocessing, the patches are mean centred and rescaled to remove global effects on the intensity. To capture the most interesting regions, we restrict to a subset of patches selected by local contrast: we keep the patches within the top 10% measured in k-nearest-neighbour distance (here k = 100). It was pointed out in [5] that a Klein-bottle structure can be seen with this choice.

Because of the size of the full dataset,computations are done by loading data from disk as required rather than loading the whole dataset into memory.These methods allow us to work with large collections of patches without prohibitive memory costs. To facilitate landmark selection, we first construct a small sample set X_{small} by randomly sampling a fixed number of patches from the full dataset. This subset is used only for selecting landmark indices, after which the corresponding landmark points are retrieved from disk using their global indices.

This dataset provides a well-studied and interpretable benchmark for evaluating topological data analysis pipelines, as it combines large scale with a known underlying topological signal. Let $X_{\text{small}} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ denote the dataset, and let $L \subset \{1, \ldots, n\}$ denote the index set of selected landmarks with $|L| = m \ll n$.

Random Sampling. Landmarks are selected uniformly at random without replacement:

$$L \sim \text{Unif}\big(\{S \subset \{1, \ldots, n\} : |S| = m\}\big).$$

Random subsampling, in expectation, preserves certain topo- logical aspects of the underlying space if suitable conditions hold[4]. More precisely, sufficiently dense random samples recover the persistent homology of the underlying space with high probability. This provides a natural baseline against which to compare, since random sampling does not aim to select any topological structure at all, but, nevertheless, can do so correctly on average, and hence provide a baseline against which more structured landmark selection strategies can be compared.

3.3 Subsampling Methods for Evaluation

Farthest-Point Sampling (FPS). Starting from an initial index $i_1 \in \{1, \ldots, n\}$ (chosen randomly or deterministically), we iteratively select

$$i_{k+1} = \arg\max_{j \in \{1,\ldots,n\}} \min_{i \in L_k} \|x_j - x_i\|,$$

where $L_k = \{i_1, \ldots, i_k\}$. The final landmark set is $L = L_m$. This procedure approximates a greedy max–min covering of the dataset.

k-Means Landmarking. We perform k-means clustering with $k = m$ clusters:

$$\min_{\{\mu_\ell\}_{\ell=1}^m} \sum_{i=1}^n \min_{1 \le \ell \le m} \|x_i - \mu_\ell\|^2.$$

Landmarks are chosen as either the cluster centroids $\{\mu_\ell\}$ or the nearest data points to each centroid:

$$L = \left\{ \arg\min_j \|x_j - \mu_\ell\| : \ell = 1, \ldots, m \right\}.$$

This favors representatives of high-density regions.

Density-Based Selection. Let $\rho_k(x_i)$ denote a density estimate based on the distance to the k-th nearest neighbor:

$$\rho_k(x_i) = \|x_i - x_{(k)}(i)\|,$$

where $x_{(k)}(i)$ is the k-th nearest neighbor of x_i. Points with smaller ρ_k correspond to higher local density. We select the m points with smallest ρ_k:

$$L = \arg\min_{S \subset \{1,\ldots,n\},\, |S|=m} \sum_{i \in S} \rho_k(x_i).$$

This concentrates landmarks in dense regions.

Hybrid Density–FPS. We first restrict to a high-density subset

$$X_{\text{dense}} = \{x_i : \rho_k(x_i) \leq \tau\},$$

for a chosen threshold τ (e.g. top p% densest points), and then apply FPS within this subset:

$$i_{k+1} = \arg\max_{j \in X_{\text{dense}}} \min_{i \in L_k} \|x_j - x_i\|.$$

This balances coverage with concentration in the data manifold.

Persistence-Based Landmark Selection. Apart from subsampling methods, we also consider persistence-based landmark selection method of Stolz [7]. This method was motivated by the Mayer-Vietoris sequence, which provides a principled relation between the global homology of a space and the homology of local substructures of that space.

Let X be a simplicial complex and let $\hat{y}$ be a vertex corresponding to a data point y. One can decompose X as

$$X = (X \setminus \hat{y}) \cup \text{St}(\hat{y}), \qquad \text{Lk}(\hat{y}) = (X \setminus \hat{y}) \cap \text{St}(\hat{y}),$$

where $\text{St}(\hat{y})$ denotes the closed star of $\hat{y}$ and $\text{Lk}(\hat{y})$ its link. The Mayer–Vietoris sequence then relates the homology groups of these spaces via a long exact sequence

$$\cdots \to H_n(\text{Lk} * \hat{y}) \to H_n(X \setminus \hat{y}) \oplus H_n(\text{St} * \hat{y}) \to H_n(X) \to H_{n-1}(\text{Lk}_{\hat{y}}) \to \cdots .$$

Since the closed star $\text{St} * \hat{y}$ is contractible, its homology vanishes in positive dimensions. Consequently, if both $H_n(\text{Lk} * \hat{y})$ and $H_{n-1}(\text{Lk}_{\hat{y}})$ are trivial, the sequence implies that

$$H_n(X \setminus \hat{y}) \cong H_n(X),$$

That is, removing the point y does not affect the global homology in dimension n .

In practice, we work with point clouds and persistent homology rather than exact homology groups. Following [7], we use the link $\mathrm{Lk}_{\hat{y}}$ by a δ-neighborhood

$$\Delta_y = \tilde{y} \in D \setminus y \mid d(\tilde{y}, y) \leq \delta,$$

and compute persistent homology on Δ_y using a Vietoris–Rips filtration. This yields a local persistence diagram $\mathrm{Dgm}_k(\Delta_y)$, which serves as a proxy for the topology of the δ-link of y.

To quantify the influence of y, we define the local persistence score

$$\mathrm{out}_{\mathrm{PH}}^{(k)}(y) = \max_{(b,d)\in \mathrm{Dgm}_k(\Delta_y)} (d - b),$$

This score measures the extent to which nontrivial local topology is present around y. Specifically, the maximum persistence of a finite feature in dimension k.

From the perspective of Mayer–Vietoris, small values of $\mathrm{out}_{\mathrm{PH}}^{(k)}(y)$ indicate that the local topology around y is close to trivial, suggesting that removing y will have little effect on the global persistent homology. Large values, on the other hand, signal a deviation from this idealized regime, indicating that y lies in a region whose removal is more likely to alter the global topological structure.

In our experiments, we focus on the case $k = 1$ and define

$$\mathrm{out}_{\mathrm{PH}}(y) := \mathrm{out}_{\mathrm{PH}}^{(1)}(y),$$

corresponding to the most persistent local loop feature.

Landmarks are selected by ranking points according to this score. We primarily adopt the strategy of selecting points with large $\mathrm{out}_{\mathrm{PH}}(y)$, which emphasizes regions that are locally rich in topological structure and therefore most relevant for preserving global features under subsampling.

3.4 Persistent Homology Computation

Given a set of selected landmark points $X \subset \mathbb{R}^d$, we compute persistent homology using the Vietoris–Rips filtration. For each scale parameter $\epsilon \geq 0$, the Vietoris–Rips complex is defined as

$$\mathrm{VR}_\epsilon(X) = \{\sigma \subseteq X; |x - y| \leq \epsilon, \forall x, y \in \sigma\}.$$

As ϵ increases, these complexes form a filtration

$$\mathrm{VR}_{\epsilon_0}(X) \subseteq \mathrm{VR}_{\epsilon_1}(X) \subseteq \cdots,$$

which induces a corresponding sequence of homology groups

$$H_k(\mathrm{VR}_\epsilon(X); \mathbb{F}),$$

where $\mathbb{F}$ is a chosen coefficient field.

In this work, we restrict our attention to one-dimensional homology ($k = 1$), which captures loop-like structures in the data. For each homology class, persistent homology records a birth time b and death time d, yielding a multiset of intervals (b, d) known as the persistence diagram.

All computations are performed up to a fixed maximum filtration scale $E_{\max}$. This truncation controls the computational cost and ensures consistency across different subsampling methods.

To explore the effect of coefficient choice,persistence is computed over different fields $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. Differences in the resulting persistence diagrams reflect variations in the underlying homological structure detectable over different characteristics.

For each interval (b, d), we define its *lifetime* as

$$\ell = d - b.$$

In our experiments, persistence diagrams are summarized through the distribution of these

lifetimes, with particular emphasis on the most persistent H_1 features.

In addition to qualitative inspection, persistence diagrams can also be used as quantitative objects for comparison. Given two data sets X and Y, we compute their persistence diagrams and measure their discrepancy using the Wasserstein distance (see Section 2.3).

This provides a way to quantify how closely the persistent homology of a subsampled data set approximates that of the full data set: a smaller Wasserstein distance indicates better preservation of the underlying topology under subsampling.

4. Experiments

Next, we study the effect of landmark selection on the topological structure preservation in persistent homology for several data sets and experimental settings of interest. Our goal is to quantify how different subsampling strategies trade off computational efficiency with topological fidelity.

4.1 Experimental Setup

We consider both synthetic data sets with known topology and a natural image-patch data set exhibiting the structure of a Klein bottle. For each data set, we compute a reference persistence diagram and compare it with the persistence diagrams obtained from subsampled point clouds.

Given a data set $X \subset \mathbb{R}^d$, we select a subset of k landmark points using one of the subsampling methods described in Section 2.4. Persistent homology is then computed on these landmark sets using the Vietoris–Rips filtration, restricted to one-dimensional homology.

For synthetic data, to make the situation more concrete, Figure 2 shows an example of one of the benchmark point clouds used in the experiments. The torus dataset is sampled from the standard embedding of $S^1 \times S^1$ in $\mathbb{R}^3$. We use similar synthetic data sets for the experiments throughout, with different topology and ambient dimension.

For projective spaces, we use standard matrix-valued embeddings rather than arbitrary coordinate representatives. To sample $\mathbb{RP}^3$, we first sample unit vectors $x \in S^3 \subset \mathbb{R}^4$ and

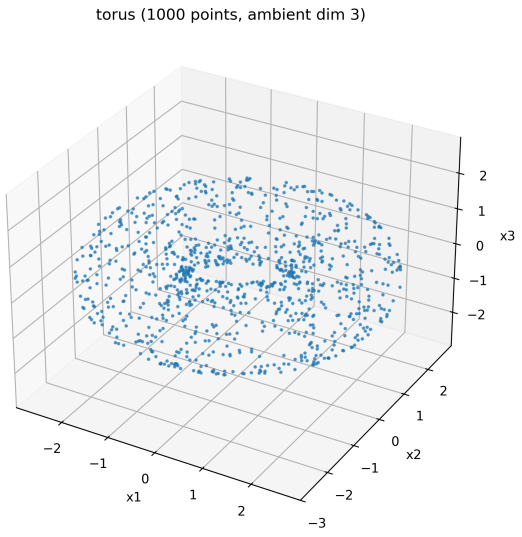

Figure 2. Example synthetic dataset: a point cloud sampled from a torus embedded in $\mathbb{R}^3$.

map the antipodal class $[x]$ to the centered rank-one projection matrix

$$\frac{1}{\sqrt{2}}\left(xx^T - \frac{1}{4}I\right).$$

This construction is invariant under $x \mapsto -x$ and embeds $\mathbb{RP}^3$ into the space of traceless symmetric 4×4 matrices, which we identify with $\mathbb{R}^{10}$ using Frobenius-norm-preserving coordinates.

The Wasserstein distance is the principal topological similarity measure used in this work, quantifying global differences between persistence diagrams in a geometrically meaningful way.

However, the Wasserstein distance can be sensitive to the presence of many low persistence features close to the diagonal, which can be noise or sampling effects. To complement this global measure, we also have feature based criteria for which we only focus on the most persistent topological features. Criteria of this type allow us to explore if the main features of the data are preserved under subsampling.

For stochastic subsampling methods, experiments are done for several independent runs and results are summarized with means and standard deviations. For deterministic or intensive computation methods, they are tested once for each setting of the parameters.

4.2 Evaluation Protocol

We study the behavior of subsampling methods under three experimental settings.

Compression vs. Fidelity. We vary the sketching ratio $r = k/|X|$ while keeping the underlying dataset fixed.For each value of r, we compute the persistence diagram of the subsampled data and measure its Wasserstein distance to the reference diagram of the full dataset. This experiment reflects how much the underlying data can be compressed while still preserving the topology of the data set.

Noise Robustness. We fix the sketching ratio and add noise to the data, perturb it in Gaussian manner with different noise magnitude. For each noise magnitude ϵ, we compare the persistence diagram of the subsampled noisy data with the persistence diagram of its full noisy version. This experiment eliminates the influence of subsampling in presence of noise.

Denoising Effect. Denoising Effect. To see whether or not subsampling can help to"denoise", we compare the persistence diagram of a subsampled noisy dataset with that of the original clean dataset. We also compare to the full, noisy dataset to see if subsampling is helping to better or worse reflect topology in the presence of noise.

4.3 Feature-Based Diagnostics and Separation Ratio

While the Wasserstein distance gives a global sense of the mismatch between persistence diagrams, it cannot distinguish between errors on origin for higher occurring topological features and errors coming from many low-persistence artifacts near the diagonal. For feature-based diagnostics that better capture preservation of meaningful structure, we introduce a feature-based diagnostic that focuses on the most persistent features.

Let $\ell_1 \geq \ell_2 \geq \cdots$ denote the ordered lifetimes of persistence intervals in a fixed homological dimension. We define the *k-th separation ratio* as

$$S_k = \frac{\ell_k}{\ell_{k+1} + \varepsilon},$$

where $\varepsilon > 0$ is a small constant added for numerical stability.

The separation ratio measures the gap between the k-th and $(k+1)$-th most persistent features. A large value of S_k indicates that the diagram consists of k dominant features that are clearly separated from lower-persistence noise, while smaller values suggest the presence of competing or spurious features.

In practice, the parameter k is taken with respect to the anticipated topology of the dataset. For example, for the torus in dimension H_1, we take $k = 2$, corresponding to its two independent loops. This gives a unified perspective of evaluating whether subsampling leads to preservation of the dominant topological feature for various datasets.

For more structured and dataset-dependent topology of datasets such as natural image patches, with a more specific topology (for example, the Klein bottle topology), we use other specialised diagnostics (like the Klein bottle criterion) that incorporate information across the coefficient field. They are complementary and informative tools that aid the interpretation of the persistence diagram along with separation ratio and Wasserstein distance.

4.4 Visualization and Reporting

Results are presented by plots of Wasserstein distance as a function of sketching ratio or noise level: one plot per subsampling method. For stochastic methods, the results are reported by average values and the variability between runs. Results are also presented by some persistence diagrams to report qualitatively the effect of subsampling on the features' structure. For the case of Klein bottle data set, we report also the values of the diagnostic criterion to report the presence or absence of a detectable topological structure.

5. Results

5.1 Overview

Here, we evaluate subsampling methods for topological data analysis along three complementary dimensions: compression fidelity, robustness to noise, and potential denoising effects. Our objective is to understand how effectively a reduced set of landmark points preserves the topology of the original data set, as measured through persistent homology. We

present experiments on a collection of synthetic data sets with known topology, including manifolds such as the torus and more structured spaces such as $\mathbb{RP}^3$, together with a natural image-patch data set exhibiting the structure of a Klein bottle. Taken together, these examples provide a range of geometric and topological complexity for studying subsampling behaviour.

For each dataset, we compute a reference persistence diagram and compare it with diagrams resulting from subsampled point sets with different sketching methods. Our main quantitative measure is the Wasserstein distance between persistence diagrams, which measures the difference in both the location and persistence of the topological features. The experiments are organized into three main studies.

The first one studies the trade-off between compression and fidelity by varying the sketching ratio and measuring the resulting deviation from the reference persistence diagram.

In the second one we study robustness to noise by introducing some controlled noise and measure how well the subsampling method preserves the topology in presence of increasing noise.

The third one studies a possible denoising effect by comparing the persistence diagrams of noisy subsamples with the diagrams of the original, clean data, thus allowing us to investigate whether subsampling can increase topological signal quality.

Across all experiments, stochastic methods are tested for many independent runs and results are presented in terms of average and standard deviation value. Deterministic or computationally expensive methods are evaluated once for each value of the parameter setting. The following subsections provide more details of results for each experimental setting.

5.2 Fidelity vs Compression

We evaluate how subsampling affects topological fidelity as the sketching ratio $r = k/|X|$ varies. For each dataset, we compare persistence diagrams using the Wasserstein distance as a quantitative measure of fidelity. In parallel, we analyze the separation ratio S_k, which captures the prominence of the dominant topological features relative to lower-

persistence noise.

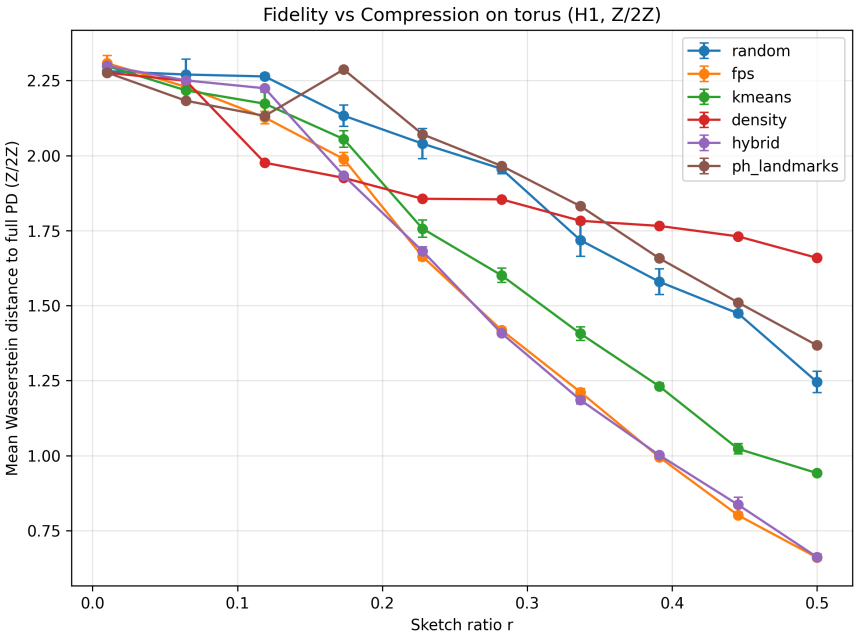

Figure 3. Fidelity vs. compression on the torus dataset. Mean Wasserstein distance between subsampled and full persistence diagrams (H_1, $\mathbb{Z}/2\mathbb{Z}$) as a function of sketch ratio r.

Torus. Figure 3 and Figure 4 show the behavior of the Wasserstein distance and separation ratio for the torus dataset. As the sketching ratio increases, the Wasserstein distance reduces more or less steadily for all methods, indicative of a better approximation of the full persistence diagram.

The separation ratio gives us further insights into structure. Even at rather low levels of compression, the ratio S_2 remains relatively large, indicating that the two dominant H_1 features of the torus remain well separated from lower-persistence features. This shows that a signature of the topology is retained even if the Wasserstein distance is still fairly large for some sketching methods.

Among the subsampling methods considered, FPS and the hybrid approach consistently produce lower Wasserstein distances and higher separation ratios across a broad range of compression levels. These methods achieve tighter geometric coverage of the data set, which appears to play an important role in preserving the two-loop structure of the torus, reflected in the persistence of the $H_1 \cong \mathbb{Z}^2$ homology class.

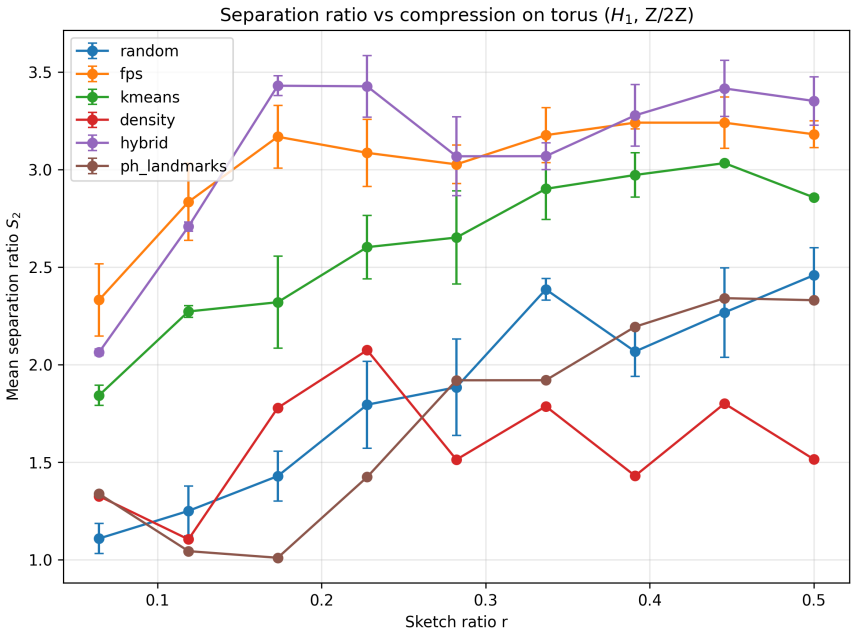

Figure 4. Separation ratio vs. compression on the torus dataset. The separation ratio S_2 measures the prominence of the second most persistent feature relative to lower-persistence noise (H_1, $\mathbb{Z}/2\mathbb{Z}$).

Real Projective Space $\mathbb{RP}^3$. Figure 5 to Figure 7 presents the corresponding results for $\mathbb{RP}^3$. The Wasserstein distance shows a non-monotonic behaviour, initially increasing for small sketching ratio, decreasing as the number of landmarks increases. This is due to the difficulty of recovering the topology of $\mathbb{RP}^3$ with very sparse samples.

The separation ratio again reveals a clearer trend. As the sketching ratio increases, the ratio S_1again grows steadily, indicating that the dominant topological feature separates more clearly from the noise. Compared with the torus, this separation develops less gradually, suggesting that $\mathbb{RP}^3$ requires a higher sampling density to stabilize its topological signature.

Consistent with the results of the torus, for both FPS and the hybrid method, the best overall performance is obtained, maintaining low Wasserstein distance and high separation ratios. The method K-means gives a reasonable approximation, but it lags behind in feature separation. The density-sbased method performs noticeably worse, with persistently low separation ratios, indicating that it fails to capture the global topology effectively.

Summary. We observed in both datasets that geometric subsampling strategies favouring uniform coverage, such as FPS and hybrid strategies, provide the best trade-off between

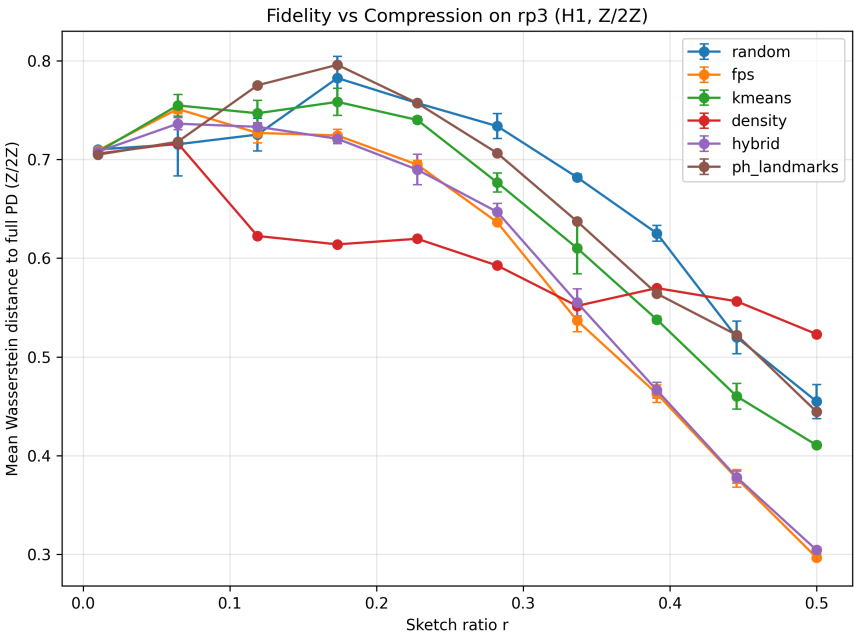

Figure 5. Fidelity vs. compression on the $\mathbb{RP}^3$ dataset. Mean Wasserstein distance between subsampled and full persistence diagrams (H_1, $\mathbb{Z}/2\mathbb{Z}$) as a function of sketch ratio r.

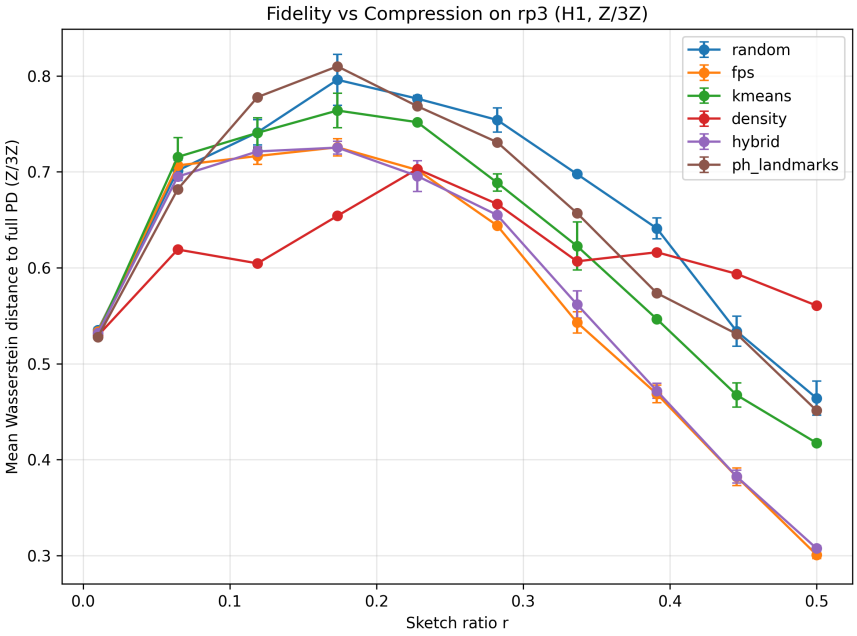

Figure 6. Fidelity vs. compression on the $\mathbb{RP}^3$ dataset. Mean Wasserstein distance between subsampled and full persistence diagrams (H_1, $\mathbb{Z}/3\mathbb{Z}$) as a function of sketch ratio r.

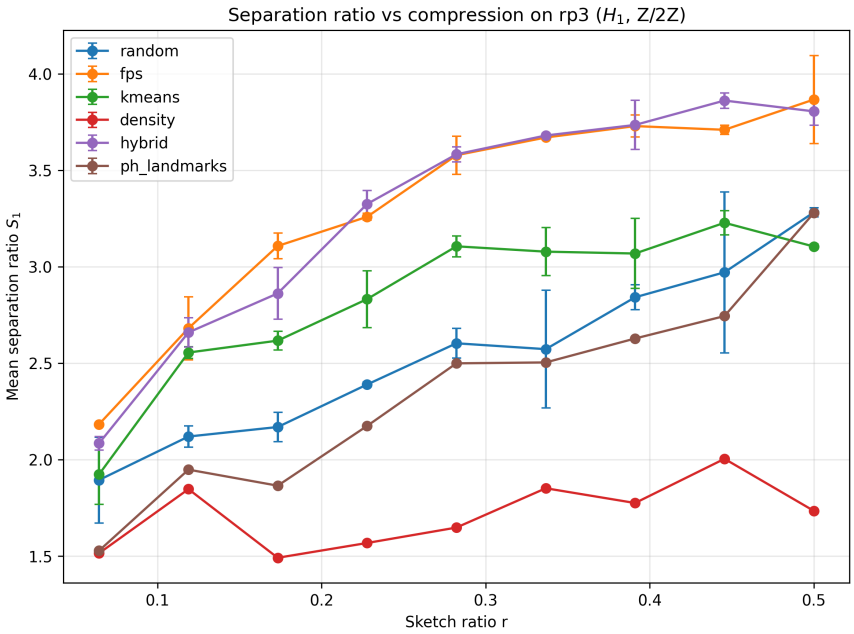

Figure 7. Separation ratio vs. compression on the $\mathbb{RP}^3$ dataset. The separation ratio S_1 captures the prominence of the dominant H_1 feature relative to noise (H_1, $\mathbb{Z}/2\mathbb{Z}$).

compression and geometric fidelity of the topology. The separation ratio is stable over a larger range of compression rates than the Wasserstein distance, indicating that dominant topological features can still be preserved despite the overall persistence diagram being very different from the reference.

5.3 Noise Robustness

Next, we look at how subsampling methods behave as the levels of ambient noise increase. For a fixed sketching ratio $r = 0.1$, we perturb the input point cloud with noise of magnitude ϵ and measure both the Wasserstein distance to the noisy full persistence diagram and the separation ratio S_2.

Torus. Figure 8 and Figure 9 indicate the effect of noise on both fidelity and feature separation. As noise level increases, all approaches observe a degradation of the separation ratio, meaning it is more difficult to identify the two main H1 features of the torus from noise. But they behave very differently under various sampling strategies. Geometric methods such as FPS and the hybrid approach achieve the highest separation ratios at low noise levels, indi-

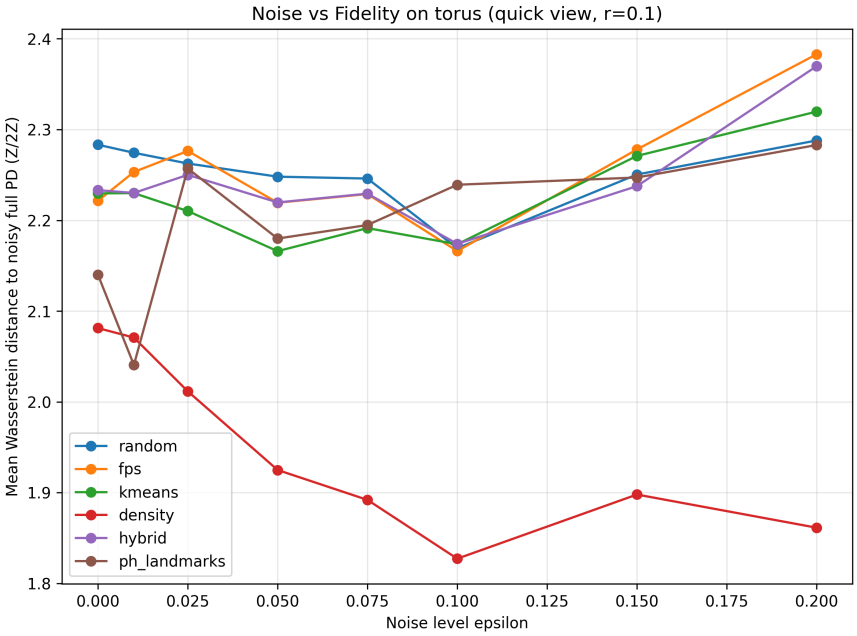

Figure 8. Noise vs. fidelity on the torus dataset at fixed sketch ratio $r = 0.1$. Mean Wasserstein distance between subsampled and full persistence diagrams (H_1, $\mathbb{Z}/2\mathbb{Z}$) as a function of noise level ϵ.

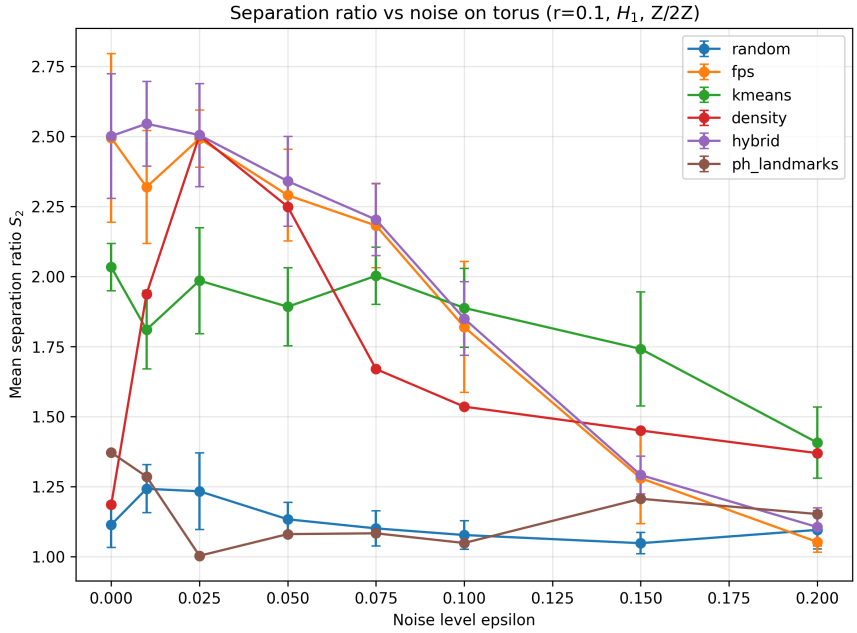

Figure 9. Separation ratio vs. noise on the torus dataset at fixed sketch ratio $r = 0.1$. The separation ratio S_2 measures the prominence of the second most persistent feature relative to lower-persistence noise (H_1, $\mathbb{Z}/2\mathbb{Z}$).

cating strong preservation of the torus structure in near-clean settings. As noise increases, their separation ratios decrease steadily, eventually approaching those of simpler methods.

In contrast, the density-based method has a different behaviour. Although it has a low initial separation ratio, it remains very stable as noise increases and it also does not deteriorate as sharply as does the geometric methods. It is therefore easy to see that density-based sampling is less sensitive to noise in terms of feature separation. The difference is even clearer in fidelity. The density-based method always produces the lowest Wasserstein distance for all values of noise, meaning that it follows the noisy persistence diagram the closest for any method.

Geometric methods do well in low-noise regime, but their accuracy deteriorates with increasing noise level. Taken together, this indicates a trade-off between geometric coverage and robustness to noise. Methods such as FPS and hybrid sampling emphasise global structure and work best in the low-noise regime and density-based sampling behaves better in the noisy setting and has higher fidelity to the perturbed topology.

$\mathbb{RP}^3$. RP^3 is very different from a torus in that there is only a single dominant H_1 feature over $\mathbb{Z}/2\mathbb{Z}$, as is visible in Fig. 10 and Fig. 11. This setting is a useful contrast, since topological fidelity is primarily driven by preserving only one important cycle, rather than by preserving a collection of similarly sized features. In the compression experiment, all methods show a similar qualitative behaviour for Wasserstein distance: fidelity improves as sketching ratio increases.

In the compression experiments, all subsampling methods exhibit a broadly similar qualitative trend in Wasserstein distance: topological fidelity improves as the compression ratio decreases. The separation ratio S_1, however, reveals clearer distinctions between methods. Geometric approaches such as FPS and the hybrid method consistently achieve higher separation ratios, indicating better preservation of the underlying topological structure under aggressive compression.

The results on different sketching ratios all indicate that they preserve the prominent

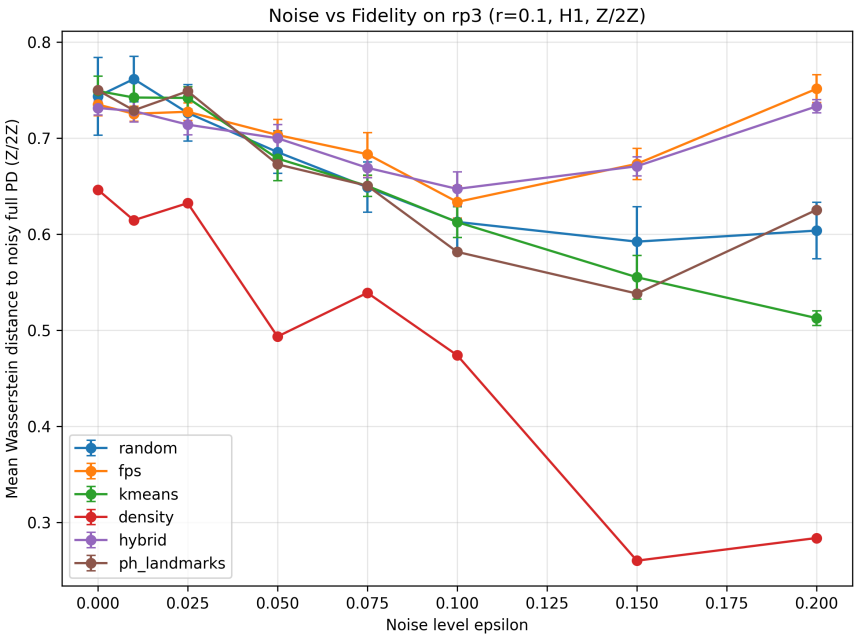

Figure 10. Noise vs. fidelity on the $\mathbb{RP}^3$ dataset at fixed sketch ratio $r = 0.1$. Mean Wasserstein distance between subsampled and full persistence diagrams (H_1, $\mathbb{Z}/2\mathbb{Z}$) as a function of noise level ϵ.

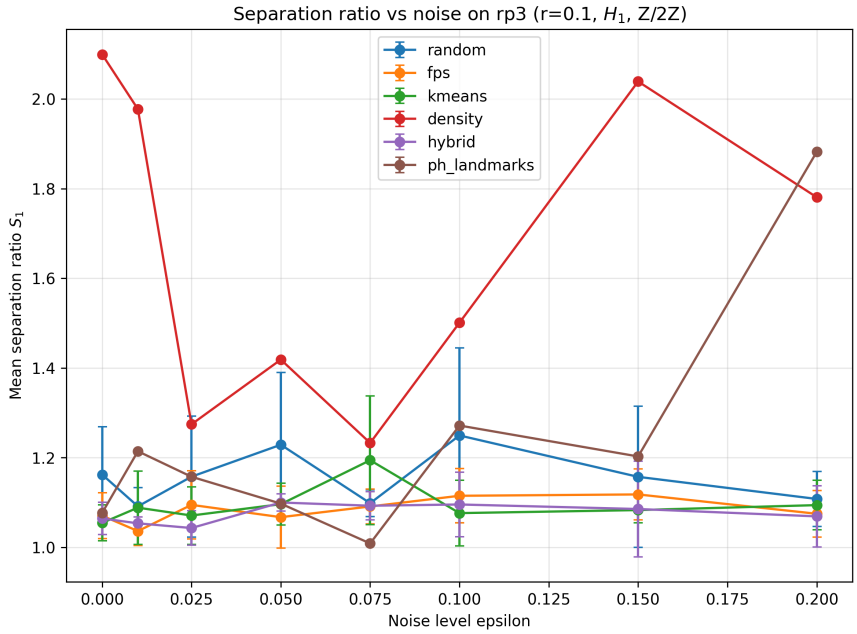

Figure 11. Separation ratio vs. noise on the $\mathbb{RP}^3$ dataset at fixed sketch ratio $r = 0.1$. The separation ratio S_1 captures the prominence of the dominant H_1 feature relative to lower-persistence noise (H_1, $\mathbb{Z}/2\mathbb{Z}$).

feature and suppress spurious intervals successfully with sufficient sampling ratio. Random and k-means sampling have more moderate separation, while density-based sampling have more different behaviours, reflecting its dependence on the underlying sampling distribution.

In the noise robustness experiment, the observed trends are different from the torus experiment. Although the Wasserstein distance tends to increase with the noise level for most methods, density-based sampling still keeps a fairly low Wasserstein distance even for higher noise levels, which means that it stays close to the corresponding noisy reference diagram. At the same time, its separation ratio keeps moderate level, meaning that the most significant topological feature is still preserved.

Another interesting phenomenon emerges when examining the separation ratio as a function of the noise level. For several methods, including density-based sampling, the separation ratio does not decrease monotonically with increasing noise and may even improve at intermediate noise levels.

This behaviour can be explained by the asymmetric effect of noise on the persistence diagram: secondary intervals are substantially more sensitive to perturbations and collapse earlier than the dominant interval. Consequently, the ratio l_1/l_2 may increase even as the overall topological signal deteriorates.

This highlights a distinction between data sets containing multiple competing topological features and those characterised by a single dominant feature. On $\mathbb{RP}^3$, methods that emphasise the dominant structure, such as FPS and hybrid sampling, remain particularly stable with respect to separation, whereas density-based sampling performs especially well in noisy settings by further suppressing non-dominant structures.

Interpretation. This behaviour reflects the combination of sampling strategy, noise and the structure of the persistence diagram.

Geometric methods like FPS and hybrid sampling do force spatial coverage, hence, they enforce spatial coverage in order to preserve the global topological structure if the noise level is low.

On the other hand, with higher noise levels, such coverage coverage can also introduce perturbed points that affect the persistence diagram and give a less faithful representation of the data and less separation of features.

Density based sampling adapts to the empirical distribution of the data. Concentrating sample points in high-density regions weakens the effect of scattered noisy points and gives persistence diagrams that are more faithful to the noisy reference, for instance, where the topology is dominant by one feature as in $\mathbb{RP}^3$ and where with this noise suppression technique the separation between intervals can be improved.

The behaviour of the separation ratio in the presence of noise illustrates this imbalance more. In particular, noise can degrade lower persistence features more rapidly than dominant ones, which can lead to an increasing separation ratio even as the overall topology signal deteriorates.

Summary. The choice of subsampling strategy depends strongly both on the noise regime and the nature of the underlying topology. Geometric methods are well suited for preserving many prominent features in low-noise regimes, while density methods are more robust to noise in general and in particular for data for which the topology is dominated by a few prominent features.

5.4 Denoising Effect

High-noise regime. High-noise regime. Across both datasets, increasing the noise monotonically damages the topo- logical signal of the noisy reference diagrams, as Figure 12 and Figure 13 show. In the torus, the separation ratio is reduced from about 2.89 in the clean setting to 1.16 at $\epsilon = 0.2$. A similar trend is observed for $\mathbb{RP}^3$, where the separation ratio drops more sharply from 3.98 to 1.03 over the same noise range . In both cases, the dominant H_1 features become increasingly indistinguishable from noise-induced features as the separation ratio approaches 1.

Subsampling methods reflect this degradation rather than correcting it. For all methods, the denoising score Δ increases toward zero as ϵ grows, indicating that sketches become

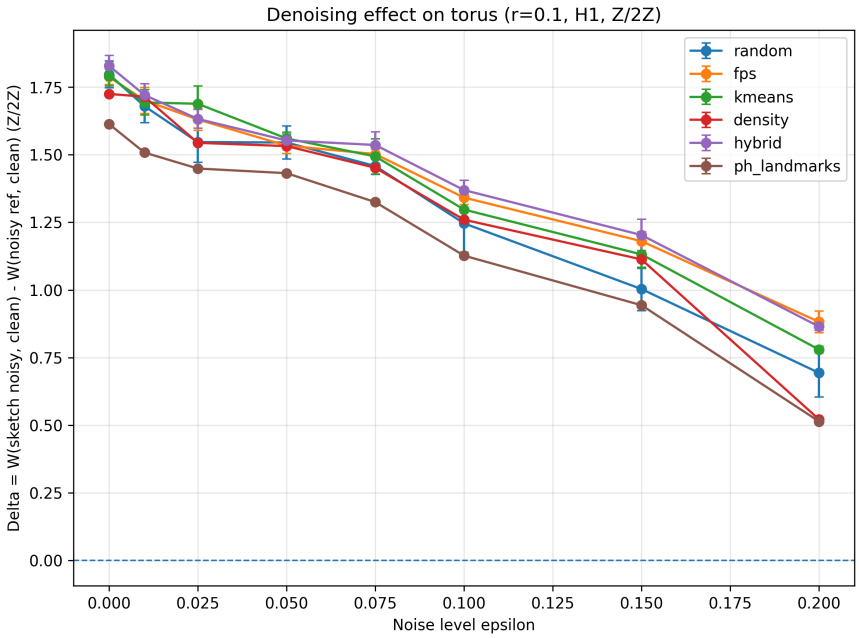

Figure 12. Denoising effect on the torus dataset at fixed sketch ratio $r = 0.1$. The quantity $\Delta = W(\text{sketch noisy, clean}) - W(\text{noisy reference, clean})$ measures whether subsampling improves or degrades fidelity relative to the noisy baseline (H_1, $\mathbb{Z}/2\mathbb{Z}$). Negative values indicate improved agreement with the clean topology.

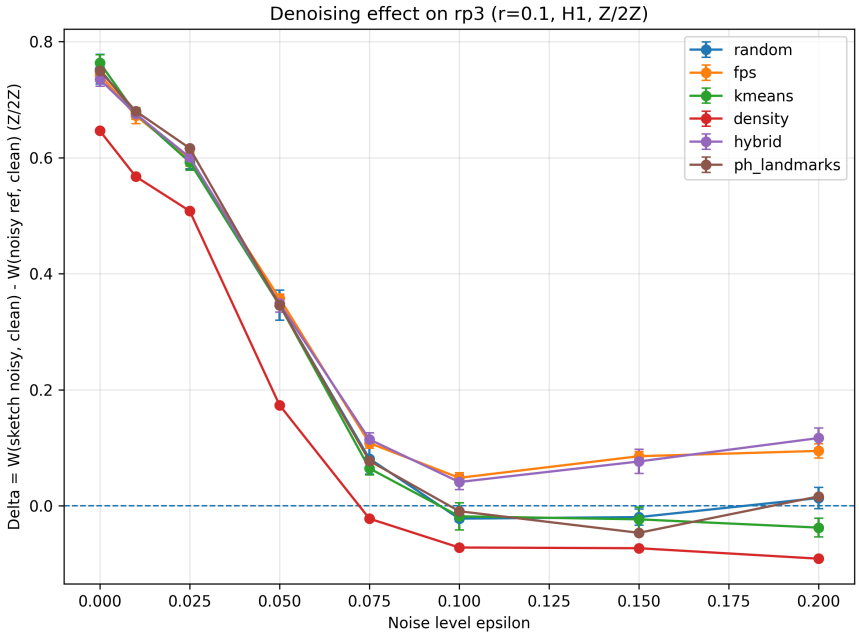

Figure 13. Denoising effect on the $\mathbb{RP}^3$ dataset at fixed sketch ratio $r = 0.1$. The quantity $\Delta = W(\text{sketch noisy, clean}) - W(\text{noisy reference, clean})$ measures whether subsampling improves or degrades fidelity relative to the noisy baseline (H_1, $\mathbb{Z}/2\mathbb{Z}$). Negative values indicate improved agreement with the clean topology.

comparable to the noisy reference rather than closer to the clean dataset. For instance, in the torus experiments with random sampling, the mean Δ increases from approximately -1.73 at $\epsilon = 0$ to -0.06 at $\epsilon = 0.2$, while in $\mathbb{RP}^3$ it transitions from approximately -2.06 to positive values at high noise levels, indicating slight divergence from the noisy reference.

5.5 Klein Bottle Criterion

How can we detect the presence of Klein bottle structure in our image patch dataset? We present here a simple feature-based diagnostic based on the behaviour H_1 persistence.

The Klein bottle has distinct homological signatures over different coefficient fields. In particular, it exhibits two prominent one-dimensional features over $\mathbb{Z}_2$, while only one persists over $\mathbb{Z}_3$. This asymmetry motivates the following criterion.

Definition (Klein bottle criterion). Let S_2 and S_1 denote the first and second separation ratios of the persistence diagram in H_1 computed respectively over $\mathbb{Z}_2$ and $\mathbb{Z}_3$, respectively. We define the Klein bottle criterion as the pair

$$(A, B) = (S_2, S_1).$$

A dataset is considered consistent with Klein bottle structure if both A and B are sufficiently large, suggesting presence of reasonably well-separated dominant features in both coefficient fields, with different multiplicities for the coefficient fields.

Empirical results. Figure 14 shows the values for (S_2, S_1) obtained for the different subsampling methods. Most of the methods return values for (S_2, S_1) equal to $(1, 1)$, indicating no real dominant features found for either coefficient field. This indicates that the Klein bottle structure is not recovered well by any of the sampling methods. Of all the methods, the density-based sampling method has quite different behaviour: it gives significantly higher values for S_1, but with $S_2 \approx 1$. This signals the presence of one dominant feature over $\mathbb{Z}_3$ and not the expected two dominant features over $\mathbb{Z}_2$. The PH-landmark method exhibits a weaker version of this behaviour; it has some separation over $\mathbb{Z}_3$ but only weak structure

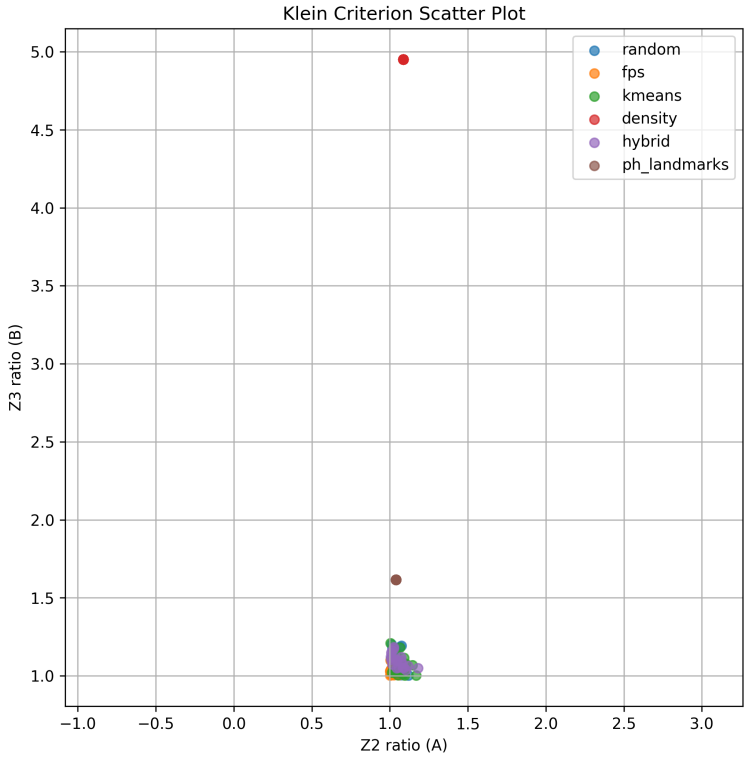

Figure 14. Klein criterion scatter plot for the image patch dataset. Each point represents a subsampled persistence diagram, plotted using its separation ratios (A, B), where A denotes the $\mathbb{Z}/2\mathbb{Z}$ separation ratio and B denotes the $\mathbb{Z}/3\mathbb{Z}$ separation ratio. Large values in both coordinates indicate stronger evidence of Klein bottle–like topological structure.

over $\mathbb{Z}_2$.

Interpretation. These results seem to suggest that recovering Klein bottle structure from this data is orders of magnitude harder than in the synthetic settings. There is no evidence of a strong second dominant feature over $\mathbb{Z}_2$. The signal corresponding to the full H_1 structure is either weak or shadowed by noise.

At the same time, the partial success of density-based sampling in showing a dominant feature over $\mathbb{Z}_3$ indicates that certain sampling strategies highlight lower complexity parts of the topology. No sampling method consistently recovers the full Klein bottle criterion.

Summary. The Klein bottle criterion provides a useful qualitative diagnostic for detecting coefficient-dependent topological structure. In this dataset, the criterion reveals only partial evidence of the expected topology, highlighting the difficulty of recovering non-orientable structure from large, noisy real-world data using subsampling alone.

5.6 Cross-Dataset Comparison

Across all datasets, several recurring themes emerge regarding the behavior of subsampling methods as the compression rate varies and noise level increases.

The first is that, across datasets, the relative behaviour of the subsampling methods is fairly stable in the low-noise regime. Geometric methods, like farthest point sampling and hybrid methods, all tend to better preserve the dominant topological features, as illustrated by the smaller Wasserstein distances and higher separation ratios. This shows that enforcing spatial coverage is beneficial when the underlying topology is well resolved.

The second theme in the effect of noise emerges similarly across all datasets. As noise level increases, the separation between signal topology and noise seen in the reference diagrams decreases and so the performance of all subsampling methods deteriorates. In high-noise regime, the differences between methods disappear and the different methods behave similarly with regard to fidelity.

Third, the rate at which the topological structure of the dataset deteriorates depends

on the inherent characteristics of the dataset. A dataset with multiple prominent features, for example, the torus, has distinguishable structure available over a wider range of noise levels, while a dataset with fewer dominant features, such as RP^3, has a faster collapse of the separation ratios. This clearly shows the role played by feature multiplicity in robustness to noise.

Overall, these observations show that while the choice of subsampling method does matter, and influences performance under good circumstances, the ultimate dominant role governing the fidelity of topology is in the quality of the input data.

5.7 Qualitative Observations

Finally, in addition to quantitative information, persistence diagrams give interesting qualitative in-sights into the impact of subsampling and noise. In a low-noise situation, the diagrams of subsampled samples are rather similar to the diagrams of the full datasets.

The main H_1 features are represented as clusters of well-separated points far from the diagonal, with relatively few noise features. The geometric subsampling methods produce diagrams with a cleaner separation of signal from noise, reflecting their better capabilities for preserving global structure. As the noise level grows, we see a growing number of points of low persistence near the diagonal and also see that the main features are moving towards the diagonal. This leads to a gradual loss of contrast of signal with noise, which is quantitatively expressed by the decreasing separation ratios. At high noise levels, the diagrams become dominated by short-lived features and the topological signal is almost impossible to interpret.

Subsampling interacts with this effect in a relatively predictable way. Rather than removing noise, subsampling simplifies the diagram by decreasing the number of points without degrading the overall structure of the noisy distribution.

In some cases, this produces simpler looking diagrams (with better visual results), although the underlying topological signal is not necessarily recovered. These qualitative observations are consistent with the quantitative results and provide further support to the conclusion that subsampling preserves an observed structure but does not recover topology once

it has been substantially corrupted by noise.

6. Discussion

6.1 Summary of Findings

Experiments in our work show several consistent features in the behav- iour of subsampling methods for topological data analysis.

First, geometric subsampling methods, such as farthest-point sampling and hybrid methods, work well consistently in clean or low-noise settings. They preserve dominant topological features better than the other methods considered here, which is observed by both the Wasserstein distance and the separation ratio. Their geographic nature lets them capture the global structure when the true underlying geometry is sufficiently resolved.

Second, subsampling does not work as a denoising method. Across all the datasets, the obtained sketches are faithful representations of the distribution of the observed data and do not result in recovering the underlying ‘clean’ topology. In high-noise regimes, all methods do exhibit similar behaviour, indicating that subsampling preserves degraded structure rather than recovering it.

Third,the real image patch dataset is much more challenging than the synthetic datasets.Above intrinsic noise level is large,while the size of the dataset makes it costly to carry out computations. However,only very aggressive subsampling is computationally feasible and puts the analysis in a similar regime as low sketching ratios in our compression experiments. In this setting, recovered persistence diagrams do have small separation but only part of the topological structure.

Fourth,noise has an order of magnitude effect on the quality of the recovered topology.As the level of noise increases,separation ratios decrease and dominant features become more and more blurred to noise peaks.Beyond some threshold level, recovery of the topology is no longer possible,regardless of the method of subsampling used.

Fifth,the quality of subsampling methods is limited by both costs and quality of data.Parameter choices (especially, size of a sketch, scale of filtration) matter and the resulting performance

depends on the cost/risk resolution and computational tractability. Finally, the behaviour of methods such as persistence-homology-informed landmark selection depends greatly on the structure of the dataset. Methods such as these may be effective in sets with relatively uniform noise or features, but their advantage may be lower for highly structured or unevenly sampled datasets, such as the image patch dataset that we considered here. Taken together, all of these findings argue that subsampling is best thought of as a method for efficiently approximating observed topology, with its success depending largely on the strength and clarity of the underlying signal.

Finally, the behaviour of methods such as persistence-homology-informed landmark selection depends very strongly on the structure of the data set.

These methods are fruitful for settings of relatively uniform noise or even well-sampled features, but break down in settings of highly structured or very unequally sampled data, such as the image patch data that we consider here.

Taking these observations together, we can only recommend subsampling as a means of obtaining an efficient approximation of the observed topology, with the degree of success depending on the strength and quality of the underlying signal.

6.2 Limitations

Some limitations of this work are also worth pointing out.

First, the synthetic data sets that we consider here are not very big: a typical data sample size is of order 10^3 to 2×10^3 points.

The sizes are small enough to recover simple topology; the scale is already constraining in higher dimensions, where a much larger set of samples is often needed to reasonably represent the data.

Because of this, the experiments do not well represent the behaviour of subsampling methods for settings in which large data sets are available.

Second, the use of full persistence computations brings a high computational cost involved. In particular, computing Vietoris–Rips complexes for a big data set quickly becomes

unrealistic, so that we have to analyse data sets of moderate size.

This point is especially relevant in the case of the real image patch data set, in which aggressive subsampling has to be performed in order for the computations to be doable, which limits the amount of topology that can be recovered.

Third, the evaluation of the stochastic subsampling methods also relies mostly on summary statistics, such as the mean and variance of independent runs.

These quantities do give a flavour of the outcome, but they do not completely describe the difference between the different realizations. In particular, methods such as random sampling can display very heterogeneous behaviour, which is not reflected by the averages alone.

Finally, several aspects of the set-up of a given experiment are heuristic choices, such as the size of the sketch or parameters of a filtration, and feature-based diagnostics such as the separation ratio. Although these choices are motivated by practical reasons, they can influence quantitative results and limit the generality of conclusions.

In general, these limitations demonstrate the usual trade-off between computational efficiency and resolution of the topology, and motivate further developments of the method that can be applied to larger and more complex data sets.

6.3 Future Work

There are also several directions for future work coming naturally from the results of our work.

One obvious direction is a combination of subsampling with explicit denoising methods. Since subsampling alone will not recover topology in high-noise settings, combining it with preprocessing denoising steps such as smoothing or manifold denoising might improve the recovery of topology in real-world data sets.

Another direction would be to investigate richer and more informative evaluation metrics. Though we find the Wasserstein distance or the separation ratios useful summaries, they do not fully describe the structure of persistence diagrams.

Metrics that more directly separate signal from noise, or adapted to particular classes of

topological features, may enable more meaningful comparisons between methods.

Scalability is also an important challenge. Extending the analysis to larger datasets will likely need other constructions such as witness complexes, sparse filtrations, multi-scale methods etc., in order to decrease the computation cost due to full Vietoris-Rips complexes.

In addition, a more detailed study of stochastic subsampling methods would be valuable. Rather than relying only on aggregated statistics, future work could examine the distribution of outcomes across runs, as well as the stability of the resulting topological features.

Finally, a more detailed study of stochastic subsampling methods would be interesting. Rather than collecting only aggregated statistics, future work could consider the distribution of the outcomes of runs, as well as the stability of the resulting topological features. Finally, further work with real datasets is needed. The Klein bottle image dataset described here demonstrates the difficulty of recovering weak or noisy topological signals. Studying other datasets, along with improved preprocessing and feature extraction, could result in a clearer understanding of the practical applicability of subsampling methods in TDA.

7. Conclusion

In this work, we investigated the effect of subsampling on the preservation of topological structure in persistent homology. Using a synthetic data set for which the topological structure is known and a real-world data set of image patches with Klein bottle structure, we study how different landmark choice approaches trade computational cost for topological fidelity. The experiments show that geometric subsampling methods, such as farthest-point sampling and mixed methods perform similarly in clean and low noise regimes. In such regimes, they preserve prominent topological features and exhibit a clear separation between signal and noise. As the noise level increases, however, the differences between the methods become less important, and all methods tend towards a similar performance One central observation is that subsampling does not amount to denoising.

Although it can give accurate approximations of the observed data distribution, it does not recover topological structure once the structure has been hidden by noise. This issue is

particularly evident in both the denoising experiments and the real image dataset, in which intrinsic noise and computational restrictions limit the fidelity of the recovered topology.

The Klein bottle image dataset also illustrates the difficulty of real world topological analysis. Unlike in the synthetic dataset, in which the topology is well defined and recoverable, the real data only exhibits partial and unstable signatures of the topology. The proposed Klein bottle criterion illustrates that the expected coefficient-dependent topology is not consistently recovered even for carefully tuned subsampling strategies.

Future work could combine subsampling with explicit denoising or regularization methods, develop more informative diagnostics for persistence diagrams and topological features and to extend the method to bigger and more complicated data sets. Such directions will bridge together different abstract theory models from topology to very concrete real world applications of very high-dimensional noisy data.

References

[1] U. Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 2017.

[2] P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015.

[3] Y. Cao, P. Leung, and A. Monod. K-means clustering for persistent homology. *Advances in Data Analysis and Classification*, 19:95–119, 2025.

[4] Y. Cao and A. Monod. Approximating persistent homology for large datasets. *arXiv preprint arXiv:2204.09155*, 2022. arXiv: 2204.09155 [math.AT]. URL: https://arxiv.org/abs/2204.09155.

[5] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 2008.

[6] D. R. Sheehy. Linear-size approximations to the vietoris–rips filtration. *Discrete & Computational Geometry*, 49(4):778–796, 2013.

[7] B. J. Stolz. Outlier-robust subsampling techniques for persistent homology. *Journal of Machine Learning Research*, 24:1–35, 2023.