Unifying Global Topology Manifolds and Local Persistent Homology for Data Pruning

Arjun Roy ROY208@PURDUE.EDU
Prajna G. Malettira PMALETTI@PURDUE.EDU
Manish Nagaraj MNAGARA@PURDUE.EDU
Kaushik Roy KAUSHIK@PURDUE.EDU
Purdue University

Editors: List of editors' names

Abstract

Geometric coreset selection is often compromised by architectural variance in feature embeddings. We propose a solution grounded in topological invariance, which first standardizes the data's global manifold before a differentiable persistence-based optimizer distills local sample importance from each point's corrective displacement. The resulting framework yields coresets that are fundamentally robust to the geometric shifts between diverse pretrained models, enabling universal applicability.

1. Introduction

The immense scale of modern datasets and foundation models has made training and fine-tuning computationally prohibitive. Data pruning (or commonly called coreset selection) addresses this by creating a small, representative subset of the data that preserves the full dataset's essential learning characteristics. This enables faster training, efficient fine-tuning, and reduced storage costs with minimal degradation in model performance.

Coreset selection methods that require model training can be broadly categorized into two groups. *Optimization-based* methods select a subset whose loss (Killamsetty et al., 2021b; Mindermann et al., 2022) or gradient dynamics (Mirzasoleiman et al., 2019; Killamsetty et al., 2021a; Tan et al., 2023) closely match the full dataset, but often rely on computationally intensive second-order (Pooladzandi et al., 2022) or bilevel optimization (Borsos et al., 2020). Similarly, *score-based* methods rank samples using training dynamics (Toneva et al., 2019; Garg and Roy, 2023; Zheng et al., 2025) or uncertainty scores (Paul et al., 2021; He et al., 2023, 2024; Cho et al., 2025), but these metrics are dynamic throughout training and model-dependent. The critical limitation for both is their reliance on training-time information, making them incompatible with the vast ecosystem of pre-trained models where only final weights are available.

To overcome this constraint, geometric-based coreset selection methods operate on static feature embeddings extracted from a suitably trained model, which removes the need for costly training-based analysis. Approaches in this domain range from analyzing the penultimate layer feature embedding space (Xia et al., 2023), using optimal transport to measure distributional similarity (Xiao et al., 2024), or leveraging the geometric reconstruction error of samples (Yang et al., 2024). However, a significant limitation of existing geometric

methods is their reliance on metrics that are sensitive to the extrinsic geometry of the embedding space, a vulnerability we term "geometric brittleness." This leads to two primary shortcomings. First, these methods tend to prioritize samples from dense, highly represented regions of the data manifold, often at the expense of informative samples from the sparse tails of the distribution, an issue that is exacerbated at high pruning rates (Zheng et al., 2023). Second, their performance is not stable across different network architectures, as each model produces a unique geometric embedding. This instability is particularly apparent in graph-based methods (Maharana et al., 2024; Xie et al., 2025), where Euclidean distance-based metrics are used for hyperparameter tuning, making them highly sensitive to changes in the geometric embedding space and not easily transferable across architectures (see Appendix E for more details).

In this work, we introduce TopoCore, a novel framework that resolves the challenge of geometric brittleness by leveraging the principles of topology. The inherent invariance of topology to geometric deformation allows TopoCore to achieve exceptional stability. TopoCore achieves this through a unique combination of topological structures at two distinct scales:

1. **Global Structure via Manifold Projection**: We employ topology-aware manifold approximation (McInnes et al., 2018a; Wang et al., 2021) to project the high-dimensional feature space into a standardized, low-dimensional representation. This provides a stable global view that preserves the data's density distribution, ensuring comprehensive coverage.

2. **Local Structure via Differentiable Persistent Homology:** While global scores can group similarly important samples, they fail to distinguish which ones to prioritize within a localized region. Existing methods often resort to random sampling within these strata or use geometric heuristics like message-passing (Maharana et al., 2024). We propose a more principled approach using differentiable persistent homology (Cohen-Steiner et al., 2005; Birdal et al., 2021; Mukherjee et al., 2024) to assess a sample's importance relative to its immediate neighbors. Instead of using a static topological measure, we perform a topological optimization that maximizes the persistence of samples w.r.t. it's closest neighbors (Loiseaux et al., 2023). Through this optimization, samples in topologically ambiguous positions are repositioned to maximize the persistence of their local structures. The magnitude of this displacement quantifies a sample's local structural importance, enabling the selection of the most informative examples from groups of otherwise indistinguishable points.

By defining sample importance through the stable, intrinsic properties of topology, TopoCore moves beyond brittle geometric heuristics to deliver a truly architecture-agnostic coreset selection framework (see Appendix A for a visual representation of the pipeline).

2. Methodology

Our proposed method, TopoCore, constructs a coreset by analyzing the data's topological structure at two distinct scales. The first stage, *Global Manifold Representation*, addresses the challenge of architectural variance by projecting the original high-dimensional embeddings into a standardized low-dimensional space. This ensures a stable, global view of the

data's overall shape. The second stage, *Local Topological Scoring*, then analyzes the intricate local structure of this manifold. We use differentiable multi-parameter persistent homology to derive an importance score for each sample based on its contribution to the local topological complexity, effectively measuring its importance relative to its neighbors.

2.1. Global Structure: Dataset Representation with Topological Manifold Embedding

Given a well-trained deep model, denoted by $f(\cdot)$, we can express it as a composition of a feature extractor $h(\cdot)$ and a classifier $g(\cdot)$, such that $f(\cdot) = g(h(\cdot))$. Here, $h(\cdot)$ represents the network up to the *penultimate layer*, which maps an input data point $\mathbf{x}$ to a high-dimensional hidden representation $\mathbf{z} = h(\mathbf{x}) \in \mathbb{R}^D$. The full training dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ can thus be transformed into a high-dimensional feature set $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$. While this high-dimensional space Z contains rich semantic information, its extrinsic geometry is often complex and architecture-dependent. To obtain a stable and standardized representation, we project Z onto a low-dimensional manifold using topology-based manifold approximation and projection techniques (McInnes et al., 2018a; Wang et al., 2021). See Appendix D for a more detailed explanation on these techniques. Through a detailed investigation into different manifold approximation and projection techniques presented in Appendix F and Figure 5 we use the Uniform Manifold Approximation & Projection (McInnes et al., 2018a) algorithm as it creates more uniform feature embeddings across network architectures.

2.2. Local Structure: Sample Neighbors with Persistence-Based Optimizer

The *Global Manifold Embedding* provides a low-dimensional representation that faithfully preserves the global structure of the data manifold. While this ensures a stable, high-level representation, a purely global perspective is insufficient for identifying the most informative samples, whose importance is often defined by the complex local interactions with their nearest neighbors. To capture this fine-grained structure, we leverage persistent homology not as a static descriptor, but as the foundation for a dynamic topological optimization process. The objective of this process is to iteratively adjust the position of each point within its class manifold to maximize topological persistence. This is performed independently for each class $c \in \{1, \ldots, C\}$ to analyze the specific intra-class structure. For each class, we begin with its point cloud from the global manifold embedding $Y_c = \{\mathbf{y}_i \mid \text{label}(\mathbf{y}_i) = c\}$ and construct a Vietoris-Rips filtration (Oudot, 2015) on Y_c due to it's computational scalability compared to Alpha and Čech complexes, as shown in Otter et al. (2017) and Mishra and Motta (2023).

Similar to work from Scoccola et al. (2024) we define a differentiable loss function, $\mathcal{L}_{\text{pers}}(Y_c)$, whose negative gradient, $-\nabla_{Y_c}\mathcal{L}_{\text{pers}}$, points in the direction that *maximally increases the total persistence of the features*. This loss is formulated using a multi-parameter filtration considering two parameters: (1) the class-manifold Vietoris-Rips filtration (VR_{Y_c}) and (2) the class-manifold Kernel Density Estimator ($\hat{f} = KDE_{Y_c}$). The persistence of this two-parameter filtration is summarized using the Hilbert decomposition signed measure, denoted $\mu^{Hil}_{H(VR_{Y_c},\hat{f})}$ (Botnan and Lesnick, 2022).

This descriptor represents the persistence diagram as a finite collection of positive point masses (representing feature births) and negative point masses (representing feature deaths)

in the parameter space of (distance, density). Our objective is to maximize the persistence, which is accomplished by maximizing the Optimal Transport (OT) distance between this signed measure and the zero measure, $\mathbf{0}$ (Carriere et al., 2021). The differentiable loss function for a given class c is therefore defined as:

$$\mathcal{L}_{\text{pers}}(Y_c) := \text{OT}(\mu^{Hil}_{H(VR_{Y_c},\hat{f})}, \mathbf{0}) \tag{1}$$

The optimization seeks a new point configuration Y'_c that minimizes this loss (see Appendix H exploring the number of optimization steps). This formulation ensures that the optimization *enhances topological stability while preserving the original density of the class manifold*, as the density is recomputed at each epoch and is an integral part of the loss calculation. We define the **Persistence Score** for each sample $\mathbf{y}_i$ belonging to class c as the magnitude of its total displacement during its class-specific optimization, where $\mathbf{y}_i$ is the initial position and $\mathbf{y'}_i$ is the final, optimized position:

$$\text{Score}_{\text{pers}}(\mathbf{y}_i) = \|\mathbf{y}_i - \mathbf{y'}_i\|_2, \quad \text{for } \mathbf{y}_i \in Y_c, \mathbf{y'}_i \in Y'_c \tag{2}$$

Interpreting this notion of local dataset structure. A high Persistence Score quantifies the degree of topological instability a sample introduces within its own class manifold. Crucially, our optimization process is designed to be density-preserving, it enhances local topological features without altering the overall density distribution of the class manifold. This is vital for coreset selection, as it ensures our search for structurally important samples does not distort the global representativeness of the data. The optimization process repositions these points to clarify the underlying intra-class structure and increase its persistence. Therefore, the magnitude of this corrective displacement serves as a direct, dynamic measure of a sample's contribution to the topological complexity of its class, derived from the collective interaction of every point in the manifold.

2.3. Comprehensive Coreset with Global and Local Dataset Structures

To create a comprehensive sample importance metric, we formulate a final score that synergizes the local, topological information from our Persistence Score with a global measure of data representativeness. This global component is a **Density Score**, derived from a Kernel Density Estimator (KDE) applied to the projected features within each class, Y_c. The final score for a sample $\mathbf{y}_j$ is a weighted combination of these two metrics:

$$\text{TopoScore}(\mathbf{y}_j) = \alpha \cdot \text{Persistence}(\mathbf{y}_j) + \beta \cdot \text{Density}(\mathbf{y}_j) \tag{3}$$

The hyperparameters $\alpha, \beta \in [0, 1]$ modulate the influence of local topological complexity (Persistence Score) versus global distributional rarity (Density Score). This allows our framework to construct a coreset that is not only rich in challenging, boundary-defining examples but also maintains a faithful representation of the full dataset's underlying distribution.

For brevity, our main Conclusion is presented in Appendix B. We present coreset accuracy for different dataset pruning rates for CIFAR-10 and CIFAR-100 and show that TopoCore outperforms previous geometric-based coreset selection techniques (see Appendix C). We also provide a comprehensive set of additional experiments and theoretical discussions in

the appendix, including validation of the architectural transferability of topological versus Euclidean metrics (Appendix E). Ablation study comparing different manifold projection techniques (Appendix F) as well as other supporting results and experiments.

Acknowledgments

This project was supported in part by the Purdue Center for Secure Microelectronics Ecosystem – CSME#210205 and the Center for the Co-Design of Cognitive Systems (CoCoSys), a DARPA-sponsored JUMP 2.0 center.

References

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in Neural Information Processing Systems*, 2020.

Magnus Bakke Botnan and Michael Lesnick. An introduction to multiparameter persistence. *ArXiv*, 2022.

Mathieu Carriere, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In *International conference on machine learning*. PMLR, 2021.

Yeseul Cho, Baekrok Shin, Changmin Kang, and Chulhee Yun. Lightweight dataset pruning without full training via example difficulty and prediction uncertainty. In *International Conference on Machine Learning*, 2025.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. In *Symposium on Computational Geometry*, 2005.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *ACM SIGACT Symposium on Theory of Computing*, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 2020.

Isha Garg and Kaushik Roy. Samples with low loss curvature improve data efficiency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Isha Garg, Deepak Ravikumar, and Kaushik Roy. Memorization through the lens of curvature of loss function around samples. In *International Conference on Machine Learning*, 2024.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer, 2022.

Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yang He, Lingao Xiao, and Joey Tianyi Zhou. You only condense once: Two rules for pruning condensed datasets. In *Advances in Neural Information Processing Systems*, 2023.

Krishnateja Killamsetty, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2021b.

David Loiseaux and Hannah Schreiber. Multipers: Multiparameter Persistence for Machine Learning. *Journal of Open Source Software*, 9(103):6773, November 2024. ISSN 2475-9066. doi: 10.21105/joss.06773.

David Loiseaux, Mathieu Carrière, and Andrew Blumberg. A framework for fast and stable representations of multiparameter persistent homology decompositions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. $\mathbb{D}ˆ2$ pruning: Message passing for balancing diversity & difficulty in data pruning. In *International Conference on Learning Representations*, 2024.

Clément Maria, Pawel Dlotko, Vincent Rouvreau, and Marc Glisse. Rips complex. In *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.11.0 edition, 2025.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, 2018a.

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

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, 2022.

Baharan Mirzasoleiman, Jeff A. Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 2019.

Amish Mishra and Francis C Motta. Stability and machine learning applications of persistent homology using the delaunay-rips complex. *Frontiers in Applied Mathematics and Statistics*, 2023.

Soham Mukherjee, Shreyas N Samaga, Cheng Xin, Steve Oudot, and Tamal K Dey. D-gril: End-to-end topological learning with 2-parameter persistence. *ArXiv*, 2024.

Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 2017.

S.Y. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis*. Mathematical Surveys and Monographs. 2015. ISBN 9781470425456.

Mathilde Papillon, Sophia Sanborn, Johan Mathe, Louisa Cornelis, Abby Bertics, Domas Buracas, Hansen J Lillemark, Christian Shewmake, Fatih Dinc, Xavier Pennec, et al. Beyond euclid: an illustrated guide to modern machine learning with geometric, topological, and algebraic structures. *Machine Learning: Science and Technology*, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 2021.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901.

Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In *International Conference on Machine Learning*, 2022.

Luis Scoccola, Siddharth Setlur, David Loiseaux, Mathieu Carrière, and Steve Oudot. Differentiability and optimization of multiparameter persistent homology. In *International Conference on Machine Learning*, 2024.

Suryaka Suresh, Bishshoy Das, Vinayak Abrol, and S Dutta Roy. On characterizing the evolution of embedding space of neural networks using algebraic topology. *Pattern Recognition Letters*, 2024.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *ArXiv*, 2020.

Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. In *Advances in Neural Information Processing Systems*, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

Renata Turkes, Guido F Montufar, and Nina Otter. On the effectiveness of persistent homology. *Advances in Neural Information Processing Systems*, 2022.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 2021.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *International Conference on Learning Representations*, 2023.

Weiwei Xiao, Yongyong Chen, Qiben Shan, Yaowei Wang, and Jingyong Su. Feature distribution matching by optimal transport for effective and robust coreset selection. *AAAI Conference on Artificial Intelligence*, 2024.

Tianchi Xie, Jiangning Zhu, Guozu Ma, Minzhi Lin, Wei Chen, Weikai Yang, and Shixia Liu. Structural-entropy-based sample selection for efficient and effective learning. In *International Conference on Learning Representations*, 2025.

Shuo Yang, Zhe Cao, Sheng Guo, Ruiheng Zhang, Ping Luo, Shengping Zhang, and Liqiang Nie. Mind the boundary: Coreset selection via reconstructing the decision boundary. In *International Conference on Machine Learning*, 2024.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations*, 2023.

Haizhong Zheng, Elisa Tsai, Yifu Lu, Jiachen Sun, Brian R. Bartoldson, Bhavya Kailkhura, and Atul Prakash. ELFS: Label-free coreset selection with proxy training dynamics. In *International Conference on Learning Representations*, 2025.

Appendix A. Overview of TopoCore Pipeline

See Figure 1

Appendix B. Conclusion

In this work, we addressed the critical challenge of "geometric brittleness" in coreset selection, where methods fail to transfer effectively across different neural network architectures due to their sensitivity to extrinsic embedding geometry. We introduced TopoCore, a framework that resolves this issue by leveraging the principles of topology. By combining a global manifold projection for a stable overall representation with a novel local importance score derived from differentiable persistent homology, TopoCore captures the intrinsic, stable

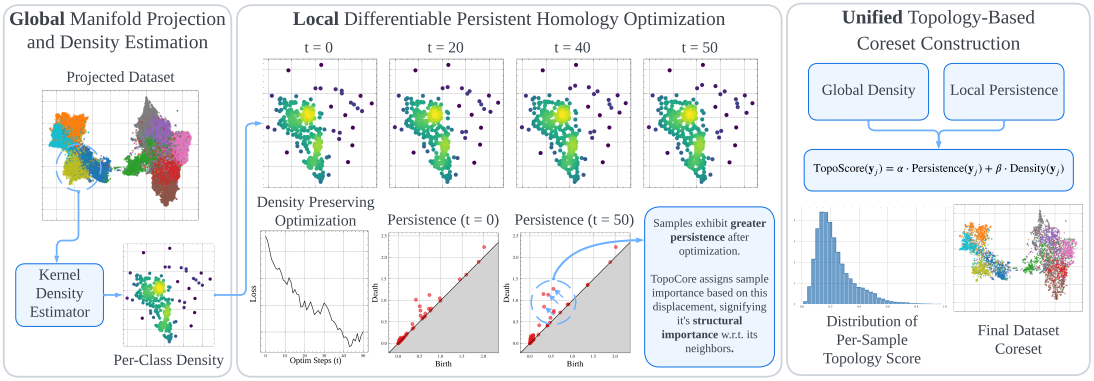


Figure 1: **An overview of TopoCore**. **(Left)** A topology-aware projection visualizes the *global* data manifold. **(Middle)** Within each class, a density-preserving persistent homology optimization derives a *local* persistence score per sample. The color map indicates high (yellow) to low (blue) density. **(Right)** The final coreset is constructed via stratified sampling on a unified score combining *global density* and *local persistence*. This not only prioritizes the most topologically informative samples but also faithfully represents the density distribution of the original dataset.

structure of the data rather than its transient geometric layout. Our central contribution is a coreset selection method that demonstrates exceptional robustness to perturbations in feature embedding across architectures. This allows for the direct and efficient application of TopoCore to a wide variety of pretrained models without the need for costly retraining or architecture-specific tuning. Ultimately, TopoCore highlights the promise of topological data analysis in developing more fundamental and universal principles for data-efficient deep learning.

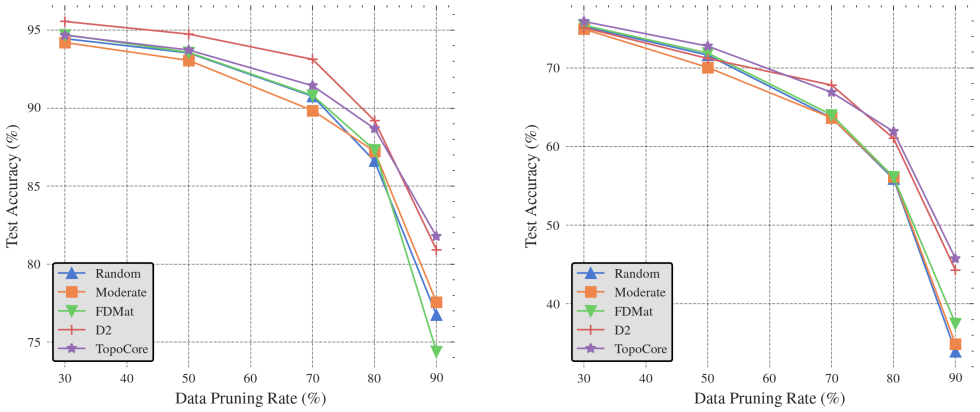
Appendix C. Coreset Performance on CIFAR-10 and CIFAR-100

C.1. Experimental Setup

TopoCore utilizes several tools and frameworks. Manifold projection is performed using `UMAP` (McInnes et al., 2018b), `multipers` (Loiseaux and Schreiber, 2024) facilitates differential persistent homology which uses the `Gudhi` C++ library (Maria et al., 2025) as a backend, and `DeepCore` (Guo et al., 2022) is used to standardize coreset selection and training across different methods.

Inspired by the findings in Zheng et al. (2023), we incorporate a crucial filtering step, during sample selection, to handle potentially mislabeled data. Since noisy or mislabeled examples can often receive high importance scores but ultimately degrade model accuracy (Swayamdipta et al., 2020), we preemptively remove all training samples that are misclassified by the base model. This ensures that our subsequent topological scoring and selection process operates on a cleaner data distribution, allowing us to focus on samples that are genuinely "hard" rather than simply erroneous.

We compare **TopoCore** with several geometry-based coreset selection methods: A) **Random** selection. B) **Moderate** (Xia et al., 2023) uses samples near the median distance to a class prototype. C) **FDMat** (Xiao et al., 2024) matches data distribution between dataset and coreset using optimal transport. D) **D2** (Maharana et al., 2024) uses a message-passing graph network. All reported coreset accuracies and standard deviations are computed over five independent training runs, with each run using a different random seed. Please see Figure 2 and Table 1.



(*a*) Results on CIFAR-10 (*b*) Results on CIFAR-100

Figure 2: Comparison of the test accuracy (averaged over five random seeds) across various geometric-based coreset selection methods on (a) CIFAR-10 and (b) CIFAR-100

Appendix D. Further Explanation on Topology-based Manifold Projection

This class of methods begins by constructing a topological representation of the high-dimensional data in the form of a *fuzzy simplicial set*. A simplicial set is a collection of simplices (0-simplices are points, 1-simplices are edges, 2-simplices are triangles, etc.) that captures the shape of the data. The "fuzzy" aspect assigns a membership strength to each simplex, representing the belief that it exists in the true underlying manifold. This is achieved by examining the local neighborhood of each point $\mathbf{z}_i$ and assigning a membership strength p_{ij} to the 1-simplex (edge) connecting it to its neighbor $\mathbf{z}_j$, based on their distance and normalized by the local density.

The algorithm then seeks to learn a low-dimensional embedding $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where $\mathbf{y}_i \in \mathbb{R}^d$ and $d \ll D$, whose own fuzzy simplicial set is as similar as possible to the one derived from the high-dimensional data. A corresponding set of membership strengths, $Q = \{q_{ij}\}$, is defined for the 1-simplices in the low-dimensional space, typically using a heavy-tailed kernel to allow for effective separation of dissimilar points.

The final low-dimensional representation Y is found by optimizing the positions of the points $\{\mathbf{y}_i\}$ to minimize the divergence between the high-dimensional and low-dimensional

Table 1: Accuracy and standard deviation across various geometric-based coreset selection methods on CIFAR-10 and CIFAR-100.

	CIFAR-10 (ResNet-18)				
Pruning Rate (→)	30%	50%	70%	80%	90%
Random	94.5±0.1	93.5±0.1	90.8±0.2	86.6±0.3	76.7±0.9
Moderate (Xia et al., 2023)	94.2±0.1	93.1±0.1	89.9±0.2	87.2±0.2	76.9±1.0
FDMat (Xiao et al., 2024)	94.7±0.1	93.6±0.2	90.8±0.2	87.3±0.4	74.4±0.7
D2 (Maharana et al., 2024)	**95.6±0.1**	**94.8±0.1**	**93.1±0.1**	**89.2±0.2**	80.9±1.5
TopoCore	94.7±0.2	93.7±0.2	91.6±0.1	88.7±0.4	**82.1±0.3**
	CIFAR-100 (ResNet-18)				
Pruning Rate (→)	30%	50%	70%	80%	90%
Random	75.3±0.2	71.6±0.1	63.7±0.5	55.9±1.0	34.0±1.1
Moderate (Xia et al., 2023)	74.9±0.3	70.1±0.3	63.7±0.2	56.1±0.5	34.9±2.1
FDMat (Xiao et al., 2024)	75.4±0.2	71.9±0.3	64.0±0.6	56.1±1.5	37.5±1.6
D2 (Maharana et al., 2024)	75.1±0.5	71.2±0.2	**67.8±0.9**	61.1±1.4	44.3±2.6
TopoCore	**75.9±0.4**	**72.8±0.3**	66.9±0.5	**61.9±0.6**	**45.8±0.7**

fuzzy simplicial sets. The objective function, often a form of cross-entropy, can be expressed as:

$$\mathcal{L}(Y) = \sum_{(i,j)} \text{AttractiveForce}(p_{ij}, q_{ij}) + \sum_{(i,j)} \text{RepulsiveForce}(p_{ij}, q_{ij}) \tag{4}$$

This optimization effectively arranges the points in the low-dimensional space such that the topological structure (clusters, voids, and connectivity) of the original high-dimensional manifold is preserved. The resulting representation Y is a standardized embedding robust for downstream tasks like coreset selection.

Appendix E. On the Transferability of Topological vs. Euclidean Features

We provide a formal argument for the superior transferability of topological features derived from persistent homology over traditional Euclidean metrics across different neural network architectures (Papillon et al., 2025). We demonstrate that the stability guarantees inherent to persistent homology ensure that its output is robust to the geometric variations common between different network embeddings. Conversely, we show that Euclidean-based metrics, such as the distance to a class prototype, are inherently sensitive to these variations.

E.1. Preliminaries and Notation

Let X be the input data space and $Y = \{1, \ldots, K\}$ be the set of K class labels. A neural network architecture is a function $f : X \to \mathbb{R}^n$ that maps input data to an n-dimensional embedding space. Let f_A and f_B denote two distinct network architectures (e.g., ResNet18

and ViT-L-16). The outputs of these networks for the entire dataset X are the point clouds $X_A = f_A(X)$ and $X_B = f_B(X)$ in their respective embedding spaces. We equip these embedding spaces with the standard Euclidean metric, d_E.

Definition 1 (Vietoris-Rips Filtration) *For a point cloud $P \subset \mathbb{R}^n$ and a scale parameter $r \geq 0$, the **Vietoris-Rips complex** $VR(P, r)$ is the simplicial complex whose vertices are the points in P and whose simplices are all finite subsets of P with a diameter of at most $2r$. A filtration is the nested sequence of complexes $\{VR(P, r)\}_{r \geq 0}$.*

Definition 2 (Persistence Diagram) *Applying the homology functor $H_k(\cdot)$ (for a fixed dimension k, e.g., $k = 0$ for connected components) to a filtration yields a set of birth-death pairs (b, d) representing the scales at which topological features appear and disappear. This multiset of pairs is the **persistence diagram**, denoted $Dgm(P)$. The **persistence** of a feature (b, d) is defined as $d - b$.*

Definition 3 (Bottleneck Distance) *The similarity between two persistence diagrams Dgm_1 and Dgm_2 is measured by the **bottleneck distance** $d_B(Dgm_1, Dgm_2)$, defined as the infimum over all bijections $\eta : Dgm_1 \to Dgm_2$ of the supremum of distances between matched points:*

$$d_B(Dgm_1, Dgm_2) = \inf_{\eta} \sup_{p \in Dgm_1} \|p - \eta(p)\|_\infty$$

Points may also be matched to the diagonal. The p-Wasserstein distance W_p is a related metric.

Definition 4 (Gromov-Hausdorff Distance) *The distance between two metric spaces (M_1, d_1) and (M_2, d_2) is measured by the **Gromov-Hausdorff distance** $d_{GH}(M_1, M_2)$, which is the infimum of distances over all possible isometric embeddings into a common metric space. It quantifies the "metric dissimilarity" of two spaces.*

E.2. Instability of Euclidean Distances to Prototypes

We now formalize the lack of such stability for Euclidean distances.

Definition 5 (Class Prototype and Distance Distribution) *For an embedding $f(X)$ and a class $k \in Y$, the class prototype (centroid) is $c_k = \frac{1}{|X_k|} \sum_{x \in X_k} f(x)$, where X_k are the samples of class k. The set of distances to the prototype is $S_k(f) = \{d_E(f(x), c_k) \mid label(x) = k\}$. Let $P(S_k(f))$ be the probability distribution of these distances.*

Proposition 6 (Sensitivity to Scaling) *Let f_A be a network embedding. Consider a new embedding f_B defined by a simple isotropic scaling transformation, $f_B(x) = \alpha f_A(x)$ for some scalar $\alpha > 0, \alpha \neq 1$. Then the distribution of distances to the prototype is scaled accordingly: $P(S_k(f_B)) = \alpha P(S_k(f_A))$.*

Proof The new class prototype c'_k under the embedding f_B is:

$$c'_k = \frac{1}{|X_k|} \sum_{x \in X_k} f_B(x) = \frac{1}{|X_k|} \sum_{x \in X_k} \alpha f_A(x) = \alpha \left(\frac{1}{|X_k|} \sum_{x \in X_k} f_A(x) \right) = \alpha c_k$$

The distance for any sample x of class k to the new prototype is:

$$\begin{aligned} d_E(f_B(x), c'_k) &= d_E(\alpha f_A(x), \alpha c_k) \\ &= \|\alpha f_A(x) - \alpha c_k\|_2 \\ &= |\alpha| \cdot \|f_A(x) - c_k\|_2 = \alpha \cdot d_E(f_A(x), c_k) \end{aligned}$$

Thus, every distance value in the set $S_k(f_A)$ is multiplied by α to obtain the set $S_k(f_B)$. The probability distribution of these distances is therefore a scaled version of the original. ■

E.3. Stability Guarantees for Persistent Homology

The transferability of persistence-based features is a direct consequence of the fundamental stability theorem of topological data analysis (Cohen-Steiner et al., 2005).

Proposition 7 (Invariance and Stability of Persistent Homology)

1. ***Isometry Invariance**: Let $P \subset \mathbb{R}^n$ be a point cloud and $g : \mathbb{R}^n \to \mathbb{R}^n$ be a Euclidean isometry (translation, rotation, reflection). Then, the persistence diagram is unchanged: $Dgm(P) = Dgm(g(P))$.*
2. ***Stability**: Let X_A and X_B be two point clouds in $\mathbb{R}^n$. The bottleneck distance between their respective persistence diagrams is bounded by the Gromov-Hausdorff distance between their metric spaces:*

$$d_B(Dgm(X_A), Dgm(X_B)) \leq d_{GH}((X_A, d_E), (X_B, d_E))$$

Proof (1) An isometry g preserves all pairwise Euclidean distances. Since the Vietoris-Rips filtration is constructed based solely on these distances, the filtration $\{VR(P, r)\}_{r\geq 0}$ is identical to $\{VR(g(P), r)\}_{r\geq 0}$. Applying the homology functor to identical filtrations yields identical persistence diagrams. (2) The proof is a cornerstone result in TDA. It formalizes the intuition that if two spaces are metrically similar (a small d_{GH}), their topological features as captured by persistence homology must also be similar (a small d_B). ■

E.4. Empirical Validation

To validate the above proof related to the stability of persistent homology across perturbations in the embedding space we look at (a) the euclidean distance of samples to their class prototype (Figure 3), (b) the density of samples after manifold projection (Figure 4(*a*)) and (c) the persistence score of samples (Figure 4(*b*)). We see that that as we move from euclidean-based metric (a) to global topology (b) to local topology (c) we see an increased uniformity in the metrics, showing increased stability to embedding space perturbations with topological information. As similarly shown in Turkes et al. (2022), the properties of topology render methods like TopoCore exceptionally stable across issues that arise in deep learning such as limited training data, noisy and out-of-distribution data. This insight is similar to findings in Suresh et al. (2024) which investigated the topological complexity of

the embedding space of different network architectures based on Betti numbers and Coleman et al. (2020) which showed that the feature representation of smaller proxy models can be used directly to determine the sample importance of larger more expensive models.

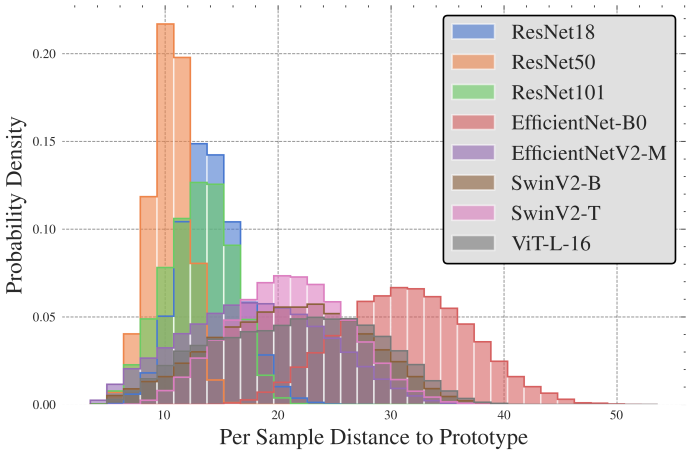
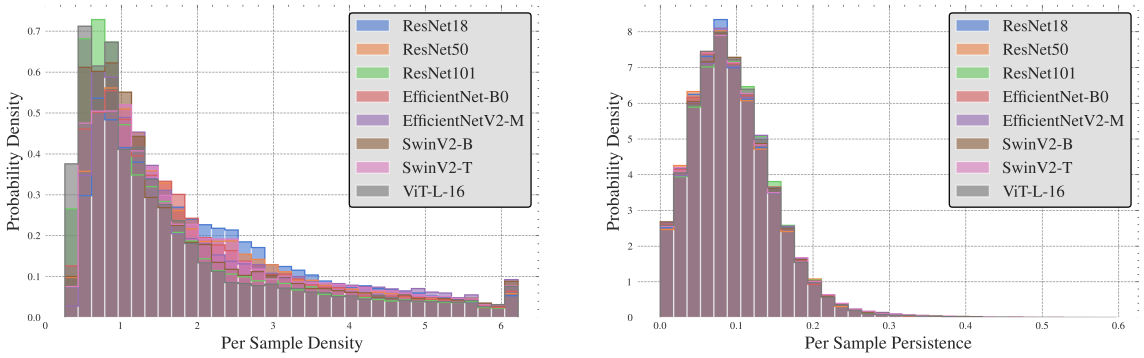


Figure 3: Euclidean distance of individual samples to their class prototype (barycenter) across a wide range of network architectures for CIFAR-100. We see that these distances, from the feature embedding space, are not uniform across architectures.



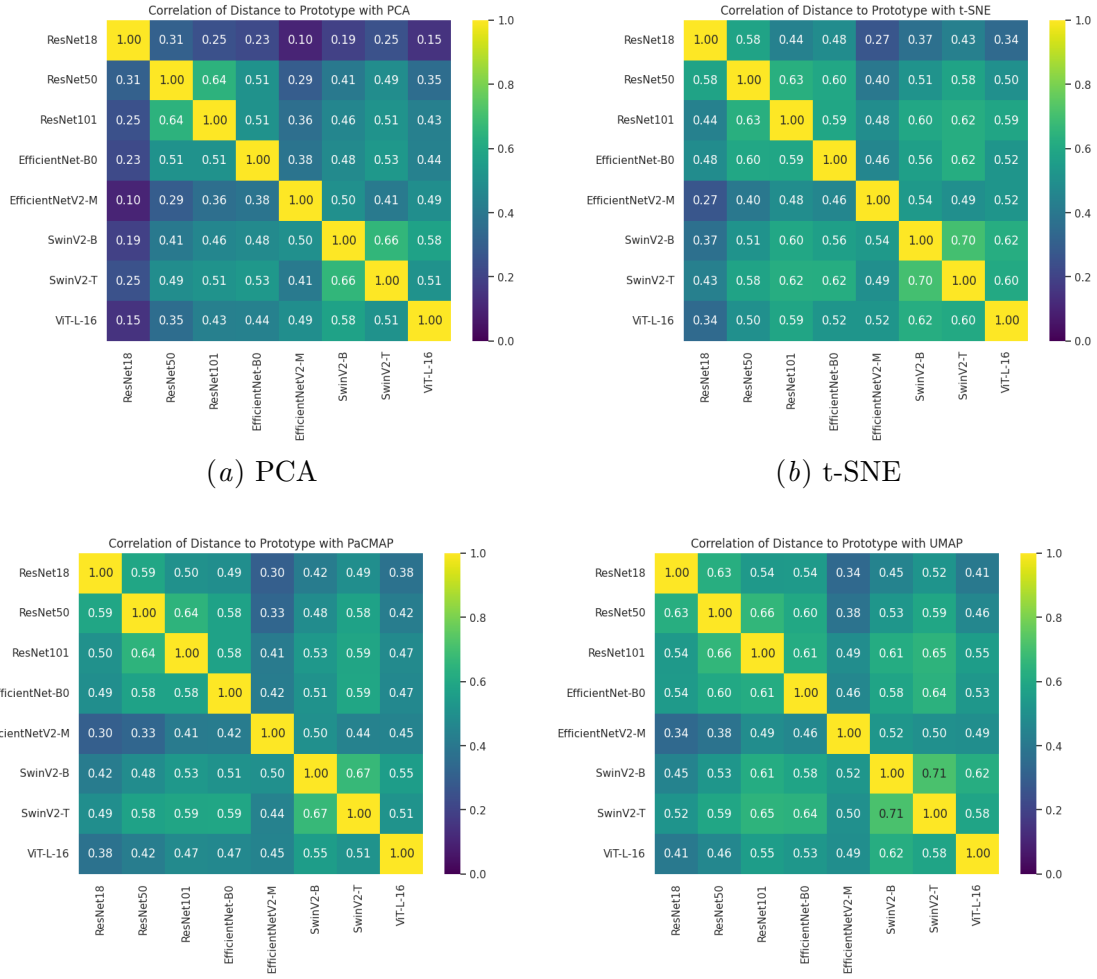
(*a*) Density (*b*) Persistence

Figure 4: (a) Kernel density estimation of individual samples within their class across a wide range of network architectures for CIFAR-100. We see that applying a "global" topology standardization via manifold projection starts improving stability across architectures. (b) Persistence score of individual samples within their class across a wide range of network architectures for CIFAR-100. By further adding the "local" topological structure, we see further improvement in stability where probability distributions across architectures almost fully match.

Appendix F. Manifold Projection as a Topological Standardization

Applying a UMAP projection as a preprocessing step is critical for achieving metric stability across diverse neural network architectures. While high-dimensional embeddings may vary in their extrinsic geometry, they share a common intrinsic topology. UMAP leverages this shared structure to construct a new, low-dimensional manifold that is not only topologically faithful but also geometrically standardized. A key consequence of this process is that the resulting low-dimensional embeddings are *density-preserving* across architectures. This standardization ensures that the global density score, a core component of our sample importance calculation, is a stable and reliable metric regardless of the source network.

To further elaborate the standardization of topology-based manifold approximation and projection across perturbations in the embedding space we look at correlation (Figure 5) of per-sample distance to prototypes across different manifold projection and feature reduction techniques (a) PCA (Pearson, 1901), (b) t-SNE (van der Maaten and Hinton, 2008) (c) PaCMAP (Wang et al., 2021) and (d) UMAP (McInnes et al., 2018a). We see that the topology-based methods, UMAP and PaCMAP, demonstrate significantly higher correlation and thus better transferability across architectures compared to linear PCA or the more locally-focused t-SNE. Notably, UMAP exhibits slightly superior transferability over PaCMAP, reinforcing its selection for our framework. This high correlation between smaller models (e.g., ResNet-18) and larger models is particularly valuable, as it validates the use of computationally inexpensive networks to generate manifold embeddings that remain effective for data selection on much larger models.



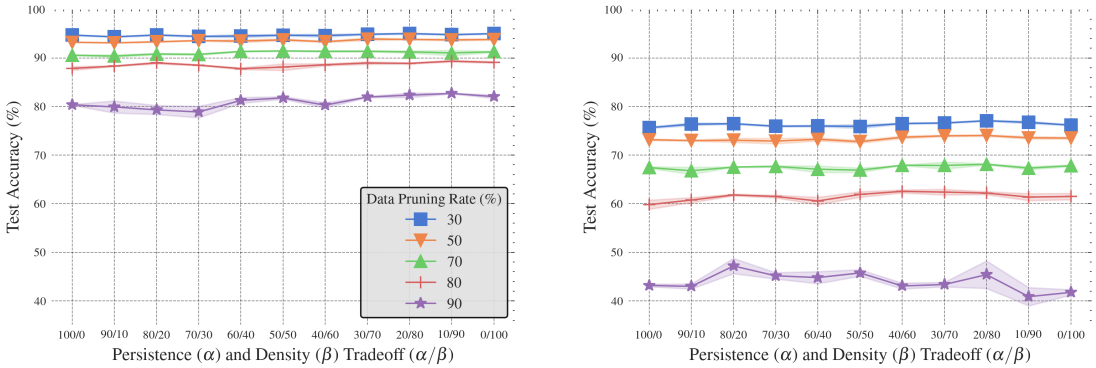
(*a*) PCA (*b*) t-SNE

(*c*) PaCMAP (*d*) UMAP

Figure 5: Correlation of per-sample distance to prototype across different architectures when applying different linear and non-linear manifold projection techniques.

Appendix G. Ablation on "Global" Density and "Local" Persistence

We investigate the impact of the hyperparameters α and β from Equation (3), which balance the influence of global density and local persistence (see Figure 7). Our analysis reveals that while the coreset quality is generally stable across a range of (α, β) values, a synergistic combination of both metrics consistently yields the best performance. Although using either density or persistence alone provides a reasonable baseline, combining them is particularly crucial at high pruning rates (e.g., 90%), where a balanced score improves accuracy by up to 5.4% over using either metric in isolation. This demonstrates that both global and local topology are vital for optimal selection and justifies our use of a fixed and balanced configuration set at (50/50) across all experiments, minimizing the need for extensive hyperparameter tuning.



(a) CIFAR-10 (b) CIFAR-100

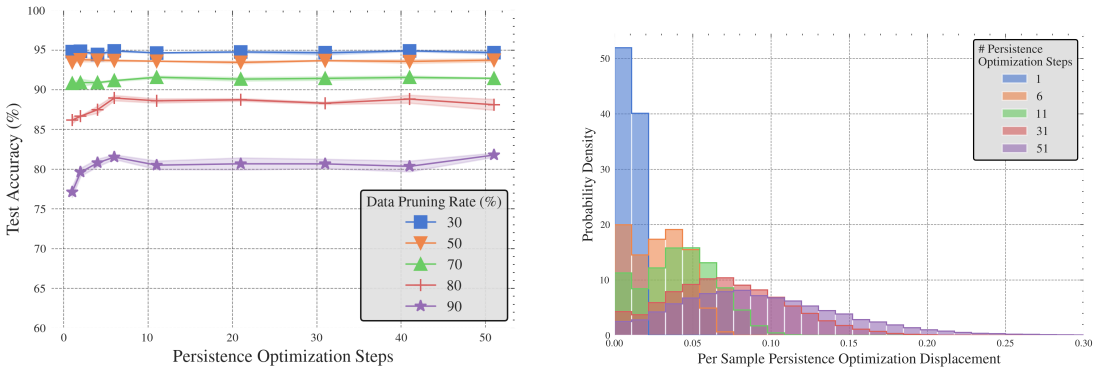
Figure 6: Topology hyperparameters across all ranges of data pruning rates for both (a) CIFAR-10 and (b) CIFAR-100.

Appendix H. Ablation on Persistence Optimization Steps

We investigate the impact of the number of optimization steps for "local" persistent homology (see Figure 7). The number of required persistence optimization steps is inversely correlated with the final coreset size. When selecting a large coreset (e.g., at a 30% pruning rate), the selection process is robust, and even a few optimization steps (1-2) suffice to identify a high-quality subset. However, at high pruning rates (e.g., 90%), the task of distinguishing the most crucial samples becomes more sensitive, necessitating a greater number of optimization steps ($\geq$6) to allow the point positions to converge and accurately reveal the most structurally important examples.

Appendix I. Relating Topology and Memorization

As a fun experiment, we examine the link between the intra-class density of our topology-based manifold projection and the established notion of sample memorization (Feldman, 2020; Feldman and Zhang, 2020), measured via the input curvature score (Garg et al., 2024). We see that high-density samples, which are prototypical examples near a class's barycenter, consistently exhibit low input curvature, characteristic of un-memorized, typical examples (see Figure 8). While low-density samples, which represent atypical data, show high input curvature, a key indicator of highly memorized samples (see Figure 9).



(*a*) Accuracy on CIFAR-10

(*b*) Distribution of Per-Sample Displacement

Figure 7: Smaller coresets have a lower margin for error, as the importance of each selected sample is magnified. Consequently, more optimization steps are needed to precisely distinguish the most critical samples. In contrast, larger coresets are more forgiving, requiring fewer steps to achieve a high-quality result.

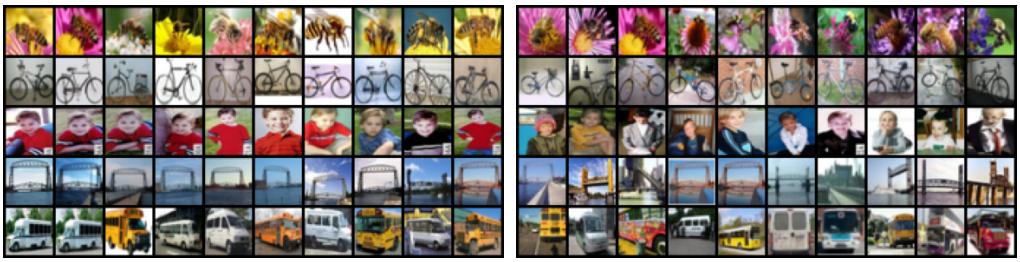


Figure 8: **Prototypical Samples:** Top-10 lowest curvature samples (left) vs. highest density samples (right) of the same class, for five CIFAR-100 classes.

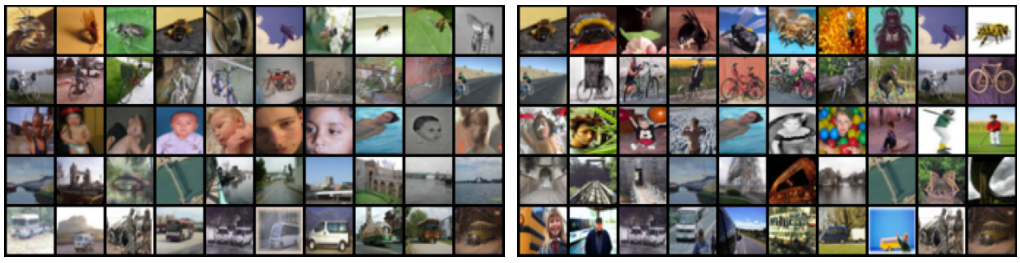


Figure 9: **Atypical Samples:** Top-10 highest curvature samples (left) vs. lowest density samples (right) of the same class, for five CIFAR-100 classes.

I.1. Training and Topological Hyperparameters

Table 2: Training and topological hyperparameters.

Section	Hyperparameter	CIFAR-10 CIFAR-100
Training (`DeepCore`)	Epochs	200
	Batch Size	256
	Optimizer	SGD
	Momentum	0.9
	Learning Rate	1e-1
	Weight Decay	5e-4
	Scheduler	CosineAnnealing
Global Manifold Projection (`UMAP`)	Number Neighbors	15
	Minimum Distance	0.1
	Metric	Cosine
	Dimensions	2
Kernel Density Estimation (`sklearn`)	Bandwidth	0.4
Local Persistent Homology (`multipers`)	Theta (Density Bandwidth)	0.4
	Function/Kernel	Gaussian
	Complex	Weak-Delaunay
	Homology Degree	1
	Optimization Steps	6
Topology Score	Global Density (α)	0.5
	Local Persistence (β)	0.5

I.2. Copyrights

Datasets: CIFAR-10 (unknown), CIFAR-100 (unknown). Libraries: Multipers (MIT License, Copyright (c) 2023 David Loiseaux), DeepCore (MIT License, Copyright (c) 2023 Zhao, Bo), UMAP (BSD 3-Clause License, Copyright (c) 2017, Leland McInnes)