Predict Training Data Quality via Its Geometry in Metric Space

Yang Ba, **Mohammad Sadeq Abolhasani**, and **Rong Pan**
School of Computing and Augmented Intelligence, Arizona State University
{yangba, mabolhas, Rong.Pan}@asu.edu

Abstract

High-quality training data is the foundation of machine learning and artificial intelligence, shaping how models learn and perform. Although much is known about what types of data are effective for training, the impact of the data's geometric structure on model performance remains largely underexplored. We propose that both the richness of representation and the elimination of redundancy within training data critically influence learning outcomes. To investigate this, we employ persistent homology to extract topological features from data within a metric space, thereby offering a principled way to quantify diversity beyond entropy-based measures. Our findings highlight persistent homology as a powerful tool for analyzing and enhancing the training data that drives AI systems.

1 Introduction

Data serves as the cornerstone of the artificial intelligence (AI) revolution, with its quality directly shaping the performance and reliability of AI models. Specifically, the features, patterns, and information embedded in data determine how effectively models can learn [3, 17]. Yet not all data carries equal value. Coverage and diversity, in particular, play a pivotal role in model performance, influencing the generalization, fairness, and robustness of AI systems [27, 4, 16]. This raises an important question: which specific properties of data make it most valuable for model training, and how can we systematically construct high-quality datasets that embody them?

A recent study [2] has shown a strong link between training data diversity and model performance, demonstrating that greater diversity improves both in-distribution (ID) and out-of-distribution (OOD) generalization. The importance of diversity has long been recognized in machine learning, with data augmentation serving as a common strategy for introducing data variability [26, 28, 30]. By exposing models to a broader range of scenarios and feature variations, diverse training data reduces overfitting and enhances generalization to unseen cases. Building on this perspective, we characterize data quality in terms of its diversity and hypothesize that higher-quality, more diverse data can lead to better model performance. To quantify such diversity, metrics such as the Vendi Score [6] have been introduced. This entropy-based approach is inspired by the concept of "community diversity" in ecology and biology [5, 18]. However, excessive diversity can also be detrimental, potentially introducing distribution shifts that degrade performance. This raises several important questions: *Aside from class balance, if all data points are equally important, will more data always be beneficial? When augmenting a dataset, what types of data are more valuable to add?* These questions motivate our focus on the geometry of data and its role in shaping model performance. Considering a dataset embedded in a metric space (Figure 1) that we wish to augment, four augmentation strategies are possible – shrinking, expanding, maintaining, or shifting the scope of the data. The central challenge, then, is to determine which of these strategies most effectively enhances generalization.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: New Perspectives in Graph Machine Learning .

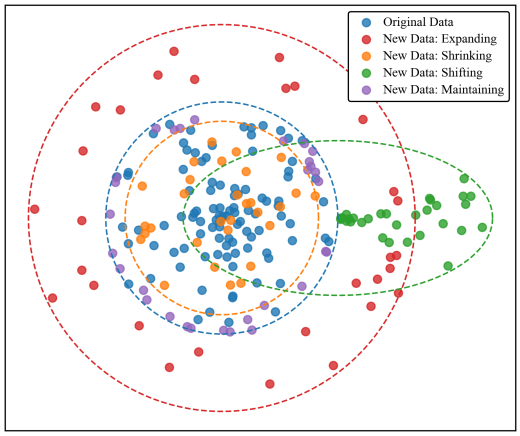

Figure 1: There are four scenarios for adding new data points to augment the current dataset (blue). However, without prior knowledge of the unseen data, clear guidance on which approach is most effective is lacking.

The biology-inspired, entropy-based diversity measures can capture how evenly data are distributed from a purely distributional perspective. Building on the fundamental connection between persistent homology (PH) and agglomerative hierarchical clustering [22], we propose a PH-based diversity measure that extends this idea to capture the topological features of data. First, we show that the PH-based measure satisfies the axiomatic definition of diversity. Next, we demonstrate how the geometry of training data influences model performance through a transfer learning framework for classification tasks. Finally, we provide practical guidance for training dataset augmentation and data point selection informed by these findings.

Our contributions are summarized as follows:

- We show that persistent homology, as a diversity measure, can capture richer structural information than conventional distribution- or entropy-based metrics.
- We develop multiple PH–based diversity measures that can quantify data quality and reveal their connection to model performance, highlighting the value of higher-order data geometric features (e.g., H_1) playing a key role in capturing meaningful structural patterns.

Our study offers a deeper understanding of the role of data geometry in the generalizability of AI models, paving the way for its integration into workflows for data augmentation, data selection, and synthetic data generation.

2 Background

2.1 Persistent Homology in Metric Space

Persistent Homology (PH) [9, 10] is a central tool in Topological Data Analysis (TDA) for uncovering the underlying shape of data, which is typically represented as a point cloud. This technique constructs a sequence of geometric objects, called simplicial complexes, over continuously expanding scales to connect nearby data points. By applying an algebraic tool, called *homology*, PH tracks the birth and death of topological features such as connected components, loops, and voids as the scale grows. The result is a multi-scale summary, commonly visualized as a barcode or persistence diagram, which distinguishes significant, long-lived features from noise. The Stability Theorem ensures that these summaries are robust to small perturbations in the data, making PH a reliable tool for extracting meaningful data structure information. While PH has been widely applied in various domains [31, 14, 25], its potential as a direct measure of data diversity remains largely unexplored.

2.2 Diversity Measurement

Several metrics have been developed to quantify the diversity of a dataset. Among them, Vendi Score (VS) [6] is the most prominent one and is often used in various data augmentation tasks. This score, derived from a set of samples and their pairwise similarity functions, quantifies the similarities among the data in a dataset. Mathematically, VS is given by the exponential of the Shannon entropy, which is obtained from the eigenvalues of the scaled similarity matrix $X^\top X$:

$$VS = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right)$$

where λ_i are the eigenvalues of scaled $X^\top X$. Another work [20] introduces several magnitude-based diversity measures that quantify the effective size of a space across scales. The magnitude function, $Mag_X(t)$, captures data diversity from local to global scopes. Specifically, the authors proposed two metrics, MAGAREA and MAGDIFF, which provide robust measures of intrinsic diversity and enable meaningful comparisons between datasets, particularly for detecting mode collapse.

3 Methodology

3.1 PH-Based Diversity Measure

We define PH-based diversity measures by using persistent homology (PH) lifetimes derived from a Vietoris–Rips complex constructed on the pairwise distance matrix of the dataset. Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be the dataset and D denote its pairwise distance matrix. The choice of distance metric depends on the application, with Euclidean distance and cosine distance being common options. Intuitively, longer lifetimes correspond to more persistent topological structures, reflecting the importance of the underlying data geometry; thus, these PH lifetimes provide a principled way to quantify diversity beyond conventional distributional measures.

To start, we construct a Vietoris–Rips filtration $\{\mathrm{VR}_\epsilon(D)\}_{\epsilon \geq 0}$ [29], where simplices are included whenever the corresponding pairwise distances in D are less than or equal to ϵ. Persistent homology yields a set of intervals [13]:

$$\mathcal{B}_k = \{(b_i, d_i)\}_{i=1}^{m_k}, \quad k = 0, 1, 2, \ldots$$

where $\mathcal{B}_k$ denotes the set of persistence intervals in homological dimension k, and $m_k = |\mathcal{B}_k|$ is the total number of such intervals. Each interval corresponds to a k-dimensional topological feature (e.g., connected components when $k = 0$, loops when $k = 1$, voids when $k = 2$, etc.), with birth scale b_i and death scale d_i. The persistence length (or lifetime) of the $i - th$ feature in homological dimension k is then defined as

$$l_i = d_i - b_i.$$

In this work, we restrict our analysis to 0-dimensional features (H_0, connected components) and 1-dimensional features (H_1, loops). To summarize diversity, we consider persistence-based analogues of entropy and Hill numbers. Define the normalized persistence weights as

$$p_i = \frac{l_i}{L}, \quad \text{where} \quad L = \sum_{i=1}^{m_k} l_i.$$

The Rényi persistence entropy of order $q \geq 0$, $q \neq 1$ is then defined as

$$\mathrm{PE}_k^{(q)} = \frac{1}{1-q} \log\Big(\sum_{i=1}^{m_k} p_i^{\,q}\Big).$$

As $q \to 1$, this reduces to the Shannon persistence entropy [21]:

$$\mathrm{PE}_k^{(1)} = -\sum_{i=1}^{m_k} p_i \, \log p_i,$$

Finally, the corresponding PH-based Hill numbers (PEH) can be expressed as the exponential of the Rényi persistence entropy, analogous to Hill numbers in ecology :

$$\mathrm{PEH}_k^q(X) = \exp\big(\mathrm{PE}_k^{(q)}\big).$$

PEH quantifies the effective number of topologically significant features in the dataset. By varying q, one can emphasize either rare or dominant features.

Unlike entropy-based measures, which quantify distributional richness in terms of eigenvalue spectra, PH-based diversity quantifies the stability of the corresponding topological structure across scales. Longer persistence lengths correspond to more significant and robust features, thereby providing a natural foundation for defining diversity measures that reflect thegeometric and topological variability in the data.

3.2 Axiomatic Definition of Diversity

Entropy-based measures, such as magnitude function and Vendi Score, are strongly connected to the notion of diversity and are widely used in ecological studies due to their computational efficiency and adherence to core diversity axioms [19, 20]. The key principles are:

- **Effective size:** In a dataset of fixed size, diversity increases when data points are well-separated and decreases as they cluster, reaching a maximum when all points are distinct and a minimum when all are identical.
- **Twin property:** Adding a duplicate observation leaves diversity unchanged.
- **Multi-scale:** Diversity is evaluated across multiple scales of similarity, capturing both local and global structure in the data manifold.
- **Symmetry:** Diversity is invariant to the order of data points, exhibiting permutation invariance.

The proposed PH-based diversity measure satisfies these principles. For *Effective size*, when all data points overlap, only one connected component (H_0) merges and no loops (H_1) form, yielding short lifetimes and near-zero diversity. Conversely, multiple persistent clusters (H_0) or robust loops (H_1) produce long, varied lifetimes. Summary statistics such as total persistence, mean, and variance of H_0 and H_1, or the PH-based Hill number capture higher diversity, reflecting the geometric spread of the data. For *twin property*, adding a duplicate point has no effect, as a duplicate x_n of $x_i \in X$ has zero distance to its twin and identical distances to all other points. In the Vietoris–Rips filtration, it immediately merges with x_i and does not create any new feature with non-zero persistence. Consequently, all non-zero persistence intervals, their weights, and the final diversity measure remain unchanged. Different homological dimensions capture geometric features across multiple scales through the filtration parameter ϵ, while the Hill-number order q adjusts the emphasis between local and global structures. The computation relies solely on the pairwise distance matrix, ensuring invariance to the ordering of samples. Formal proofs are provided in Appendix A.

3.3 Relationship between Structural Diversity and Model Performance

To investigate the relationship between structural diversity in metric space and model performance, we constructed three balanced subsets (the closest, the farthest, and random) based on pairwise distance matrices. For each sample, the maximum distance from a data point to all other points was computed, and it is used to rank points. The closest subset was drawn from the lower half of this ranking (core samples), while the farthest subset was drawn from the upper half (peripheral samples). In each case, equal numbers of class 0 and class 1 samples were randomly selected to ensure a balanced dataset. As a baseline, the random subset was created by sampling equal numbers of class 0 and class 1 directly from all data points, regardless of their distance rankings. Figure 2 illustrates an example of how three subsets are distributed in 2D space using multidimensional scaling (MDS) [7]. By analyzing the persistent homology summaries of these representative subsets and comparing the performance (accuracy) of models trained upon them, we aim to uncover systematic connections between the geometric structure of data and model generalization.

4 Experiment & Analysis

4.1 Experiment Setup

We evaluate our hypothesis on text classification tasks across multiple domains using a transfer learning approach. Specifically, we fine-tune BERT$_{base}$ models [8] by adding a dropout layer and a

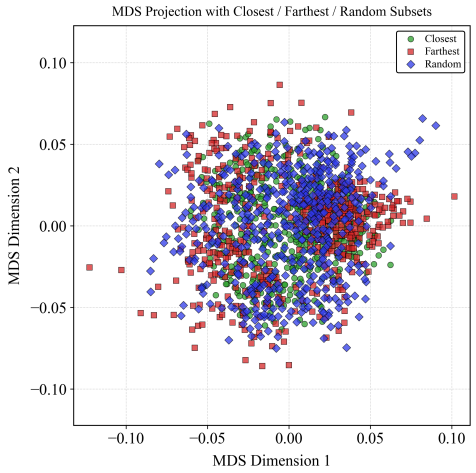

Figure 2: A demonstration of three representative subsets construction for "Medical" dataset

softmax classifier on top of the pre-trained architecture. Each model is trained for 8 epochs with a learning rate of 1e-6 and a dropout rate of 10%.

The evaluation spans several datasets – the Complaints dataset (TC) [24], the SUBJectivity dataset (SUBJ) [23], SentEval (SE) [15], Arxiv-10 [12], and Medical [11]. For each dataset, we constructed three subsets—closest, farthest, and random—and ran experiments three times per subset. To ensure consistent experimental conditions, each training set contains 500 samples (250 per class) across all datasets and subsets. Our experiments are limited to text classification with BERT fine-tuning; extending the evaluation to other modalities and larger-scale settings is left for future work.

4.2 Experiment Result

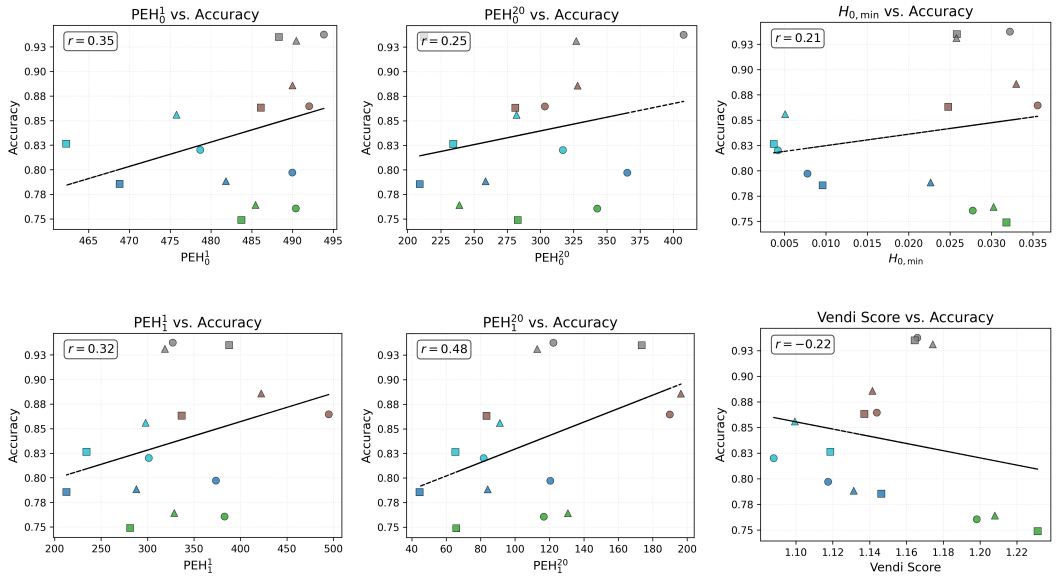

Figure 3: Model Accuracy vs. PH-based Diversity Measures

Can the data geometry predict its quality? We define high-quality data as data that enables models trained on it to achieve high accuracy. Figure 3 illustrates that PH-based diversity measures are positively correlated with model accuracy; i.e., greater diversity in both H_0 (connected components) and H_1 (loops) features tends to improve model performance. By contrast, the Vendi Score exhibits

a negative trend, with higher values associated with lower accuracy. This contrast highlights that geometry-aware indicators, such as PH-based measures, can serve as reliable predictors of data quality, whereas distributional, entropy-based measures, like Vendi Score, fail to do so (the last plot in Figure 3). We also observe a negative relationship between minimum H_0 values and the standard deviation of model accuracy. This suggests that greater geometric diversity not only boosts accuracy but also stabilizes model training, an effect detailed in the Appendix (Figure 4).

Which data geometry characteristics are most desirable for an effective model training? Table 1 shows that the topological features of different subsets strongly influence model performance. The *closest* subset has extremely high H_0-based Hill numbers, indicating fragmented clusters with many clusters of similar persistence, and high H_1-based Hill numbers with short lifetimes, reflecting noisy and unstable loops. It yields slightly better average accuracy than the farthest subset, but suffers from the largest variance, highlighting instability caused by redundancy and noise. The *farthest* subset shows both low H_0- and H_1-based Hill numbers, resulting in sparse, fragile clusters and minimal loop diversity. The scattered points provide poor coverage of the data manifold, producing the lowest accuracy, though with slightly better stability than the *closest* subset. The *random* subset strikes a balance. Moderate H_0-based Hill numbers correspond to well-separated, non-redundant clusters, while moderate H_1-based Hill numbers reflect stable loops. This balanced geometric profile produces the best accuracy and lowest standard deviation, representing the most desirable structural diversity for training.

Table 1: Relationship Between PH-based Diversity Measures and Average Accuracy by Category

Subset	Accuracy (avg $\pm$ std)	H_0 Measure	H_1 Measure	Vendi Score
Closest	0.836 ± 0.021	PEH_0^1:489 PEH_0^{20}:347 H_0 min: 0.0215	PEH_1^1:376 PEH_1^{20}:126 H_1 mean: 0.0025	1.143
Farthest	0.832 ± 0.014	PEH_0^1:478 PEH_0^{20}:244 H_0 min: 0.0191	PEH_1^1:291 PEH_1^{20}:86 H_1 mean: 0.0029	1.160
Random	0.845 ± 0.013	PEH_0^1:485 PEH_0^{20}:287 H_0 min: 0.0234	PEH_1^1:331 PEH_1^{20}:123 H_1 mean: 0.0028	1.151

Takeways. A high-quality dataset should exhibit well-separated clusters (H_0) and contain some stable loops (H_1), while avoiding the extremes of redundancy (too many overlapping data points) or sparsity (scattered, fragile structures). Random sampling often achieves this balance, but more deliberate strategies for data augmentation can explicitly target high H_0 minimum values (ensuring separation) combined with moderate H_1 mean lifetimes (capturing stable geometric features). We further observe that a relatively smaller training dataset (6%–19% of the original dataset) can achieve 91%–98.6% of the accuracy reported using the full dataset during fine-tuning [1]. This finding underscores the critical role of data selection in model performance and highlights that more data is not always better – what matters is the right structural diversity.

5 Conclusion

In this work, we demonstrated that the training data geometry, captured through persistent homology, is closely linked to model performance. Traditional entropy-based diversity metrics alone prove insufficient for predicting training data quality, whereas PH-based diversity measures offer clear advantages by effectively quantifying the structural richness of a dataset. This persistent homology–based method provides insight by separating noise-induced fragmentation from meaningful structural richness in the data. Our analysis of different data subsets shows that well-balanced clusters (H_0) combined with stable loops (H_1) yield the best accuracy and lowest variability, showing the importance of structural diversity. These insights pave the way for more principled strategies in constructing training datasets, data augmentation, and synthetic data generation. Future research could explore how topological features can be directly leveraged to guide robust model training, improving generalization while reducing dependence on large-scale data.

References

[1] Yang Ba, Michelle V Mancenido, and Rong Pan. Fill in the gaps: Model calibration and generalization with synthetic data. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17211–17225, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.955. URL https://aclanthology.org/2024.emnlp-main.955/.

[2] Yang Ba, Michelle V Mancenido, and Rong Pan. How does data diversity shape the weight landscape of neural networks? *arXiv preprint arXiv:2410.14602*, 2024.

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[4] Line H Clemmensen and Rune D Kjærsgaard. Data representativity for machine learning and ai systems. *arXiv preprint arXiv:2203.04706*, 2022.

[5] Aisling J Daly, Jan M Baetens, and Bernard De Baets. Ecological diversity: measuring the unmeasurable. *Mathematics*, 6(7):119, 2018.

[6] Dan Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*, 2023.

[7] Mark L Davison and Stephen G Sireci. Multidimensional scaling. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 323–352. Elsevier, 2000.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[9] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & computational geometry*, 28(4):511–533, 2002.

[10] Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282, 2008.

[11] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318, 2022.

[12] Ashkan Farhangi, Ning Sui, Nan Hua, Haiyan Bai, Arthur Huang, and Zhishan Guo. Protoformer: Embedding prototypes for transformers. In *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part I*, pages 447–458, 2022.

[13] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[14] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26):7035–7040, 2016.

[15] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[16] Beomjun Kim, Jaehwan Kim, Kangyeon Kim, Sunwoo Kim, and Heejin Ahn. A computation-efficient method of measuring dataset quality based on the coverage of the dataset. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4744–4752. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/kim25f.html.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[18] Tom Leinster. *Entropy and diversity: the axiomatic approach*. Cambridge university press, 2021.

[19] Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.

[20] Katharina Limbeck, Rayna Andreeva, Rik Sarkar, and Bastian Rieck. Metric space magnitude for evaluating the diversity of latent representations. *Advances in Neural Information Processing Systems*, 37:123911–123953, 2024.

[21] Emanuela Merelli, Matteo Rucco, Peter Sloot, and Luca Tesei. Topological characterization of complex systems: Using persistent entropy. *Entropy*, 17(10):6872–6892, 2015.

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

[23] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035.

[24] Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. Automatically identifying complaints in social media. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1495. URL https://aclanthology.org/P19-1495.

[25] Chi Seng Pun, Si Xian Lee, and Kelin Xia. Persistent-homology-based machine learning: a survey and a comparative study. *Artificial Intelligence Review*, 55(7):5169–5213, 2022.

[26] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29935–29948. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fb4c48608ce8825b558ccf07169a3421-Paper.pdf.

[27] Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*, pages 9040–9051. PMLR, 2021.

[28] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[29] Leopold Vietoris. Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen. *Mathematische Annalen*, 97(1):454–472, 1927.

[30] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[31] Qi Zhao and Yusu Wang. Learning metrics for persistence-based summaries and applications for graph classification. *Advances in neural information processing systems*, 32, 2019.

A Technical Appendices and Supplementary Material

A.1 Proof of Diversity Axioms for PH-Based Measures

A diversity measure derived from Persistent Homology (PH) is defined as a summary statistic of the persistence lifetimes generated from a dataset's Vietoris-Rips filtration. We prove that such a measure satisfies the key principles of effective size, the twin property, multi-scale analysis, and symmetry. $\mathrm{Div}(X)$ denotes as PH-based Diversity Measure in general.

- **Effective Size**: In a fixed-cardinality dataset, diversity increases when the data points spread out, and decreases when points concentrate or fully coincide, reaching a minimum when all points are identical. For a dataset $X = \{x_1, \ldots, x_n\}$ where $x_i = x_j$ for all i, j, $\mathrm{Div}(X)$ attains its theoretical minimum value; for a dataset X_s containing distinct points, $\mathrm{Div}(X_s) \geq \mathrm{Div}(X)$.

  Proof. Minimum Diversity (Collapsed Data). If $x_1 = x_2 = \cdots = x_n$, then D is the zero matrix. In the Vietoris–Rips filtration, all points form a single connected component at $\epsilon = 0$. No loops (H_1 features) appear. Persistence intervals thus have zero or infinite length, degenerating to a trivial diagram. Any summary statistic (e.g., total persistence, entropy, Hill number) computed from this single lifetime yields its minimum possible value, correctly reflecting a minimal diversity or an effective size of one.

  Higher Diversity (Separated Data). Conversely, if the dataset X_s consists of well-separated points, as ϵ increases, components merge, creating multiple H_0 features with non-zero lifetimes. Moreover, geometric arrangements can generate robust higher-dimensional features, such as loops (H_1), that persist across a wide range of scales. The resulting persistence diagram is richer and has a varied distribution of lifetimes. Summary statistics applied to this richer distribution yield a higher value, reflecting the greater effective size and topological complexity of the data.

  Thus, the PH-based measure maps collapsed, redundant data to low diversity and structurally rich, separated data to high diversity.

  □

- **Twin Property**: Duplicating a data point does not change the measured diversity. Let X be a dataset and let $x_i \in X$. For the set $X' = X \cup \{x_n\}$ where $x_n = x_i$, the diversity is unchanged:
$$\mathrm{Div}(X') = \mathrm{Div}(X).$$

  Proof. By definition, the distance between the twin points is $d(x_i, x_n) = 0$. For any other point $x_j \in X$, the distance from the duplicate is identical to the distance from the original: $d(x_n, x_j) = d(x_i, x_j)$.

  In the Vietoris–Rips filtration, x_n forms a 0-distance edge with x_i and never contributes a feature with nonzero persistence (its lifetime is $\ell = 0$, because the connected components corresponding to x_i and x_n are born at $\epsilon = 0$ and merge immediately, generating a persistence interval of $(0, 0)$). All other interpoint distances are unchanged, so the persistence diagram of nonzero intervals is invariant. Hence, $\mathrm{Div}(X') = \mathrm{Div}(X)$.

  □

- **Multi-Scale**: The measure accounts for geometric/topological features across the full range of distance scales, capturing both local and global structure. The diversity measure $\mathrm{Div}(X)$ is a function of the entire multiset of persistence lifetimes that integrates information from all scales in the filtration.

  Proof. Persistent homology tracks connected components, loops, and higher-dimensional holes across all filtration scales $\epsilon \in [0, \max d(x_i, x_j)]$. Long persistence intervals correspond to large-scale, global topological features, while short intervals capture local or

noise-induced structures. By tuning the filtration bounds—using a smaller $\epsilon_{\max}$ to focus on fine-scale neighborhoods and local diversity, or a larger $\epsilon_{\min}$ to filter out small-scale noise and emphasize global structure—one can selectively highlight different geometric aspects of the data. The resulting diversity measure $\mathrm{Div}(X)$ then summarizes the complete distribution of persistence lifetimes, thereby integrating information across both local and global geometric scales.

Moreover, parameter choices (e.g. Hill number order q) can adjust sensitivity to rare vs. dominant features. For example, $q > 1$ gives more weight to long-lived (global) features and $q < 1$ gives more weight to short-lived (local) features. Thus, $\mathrm{Div}(X)$ is multi-scale by construction.

□

- **Symmetry**: The diversity measure is invariant under permuting (re-ordering) the data points. Let $X = (x_1, \ldots, x_n)$ be an ordered sequence of points and let π be any permutation of $\{1, \ldots, n\}$. For the permuted sequence $X_\pi = (x_{\pi(1)}, \ldots, x_{\pi(n)})$, we have

$$\mathrm{Div}(X_\pi) = \mathrm{Div}(X).$$

Proof. The PH pipeline begins with the pairwise distance matrix D, where $D_{ij} = d(x_i, x_j)$. Let X_π be the reordered dataset. The distance matrix D_π for the permuted data has entries $(D_\pi)_{ij} = d(x_{\pi(i)}, x_{\pi(j)})$. Importantly, the set of all unique pairwise distances

$$\{d(x_i, x_j)\}_{1 \le i < j \le n}$$

is unchanged for both X and X_π. The construction of the Vietoris–Rips filtration depends only on these distances. Hence, the persistence diagrams and lifetimes $\{l_i\}$ are identical. Therefore, any diversity measure computed from these lifetimes is invariant under permutation of the data. □

A.2 Correlation Between PH-based Diversity Measure and the Standard Deviation of Model Accuracy

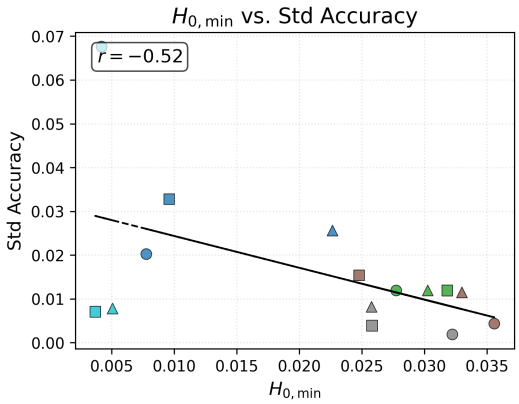

Figure 4: Minimum H_0 values vs. the standard deviation of model accuracy. A negative correlation between the minimum H_0 values and the standard deviation of model accuracy indicates that greater separation between clusters improves the stability of model training.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract summarizes our main contributions and experiment results. In the introduction section, we sum up our inspiration, methodology, and experiments, along with the contributions of our work.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We include the limitations of our work in the Appendix to discuss potential improvements of our analysis.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We provide theoretical proof and support to our methodology in Section 3: Methodology. We also provide a Preliminary section to help readers better understand the theory part. The result of proofs can be seen in Appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We describe the experiment setup at the beginning of Section 4. Hyperparameters for model training can be found in Appendix A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide code in the supplement material with the instructions.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We describe the experiment setup at the beginning of Section 4. Hyperparameters for model training can be found in Appendix A.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: In our comparison analysis, we use ANOVA and pairwise t-test; we report their results on top of the figures in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We specify the Pytorch version and library, along with the computing environments we used to run experiments, like GPU types, in Appendix A.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We reviewed and followed the reviewed the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We include a section in the Appendix to describe the impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: Our models and datasets are used in this paper both are widely-used and open-sourced models and benchmarks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We stated all existing assets used in the theory derivations and experiments. We cited the original paper and datasets employed in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: Our paper does not release new assets. The dataset and models we used have existed. We provide a theory and robust analysis.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [NA]

    Justification: Our paper does not contain anything related to crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [NA]

    Justification: Our paper does not contain anything related to crowdsourcing or research with human subjects

    Guidelines:

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

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is used only for writing, editing, or formatting purposes in our paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.