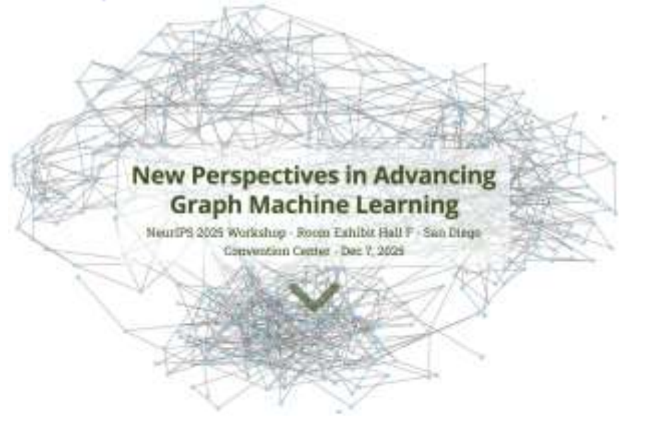
Predict Training Data Quality via Its Geometry in Metric Space

Yang Ba, Mohammad Sadeq Abolhasani, Rong Pan

School of Computing and Augmented Intelligence (SCAI), Arizona State University

NEURAL INFORMATION PROCESSING SYSTEMS

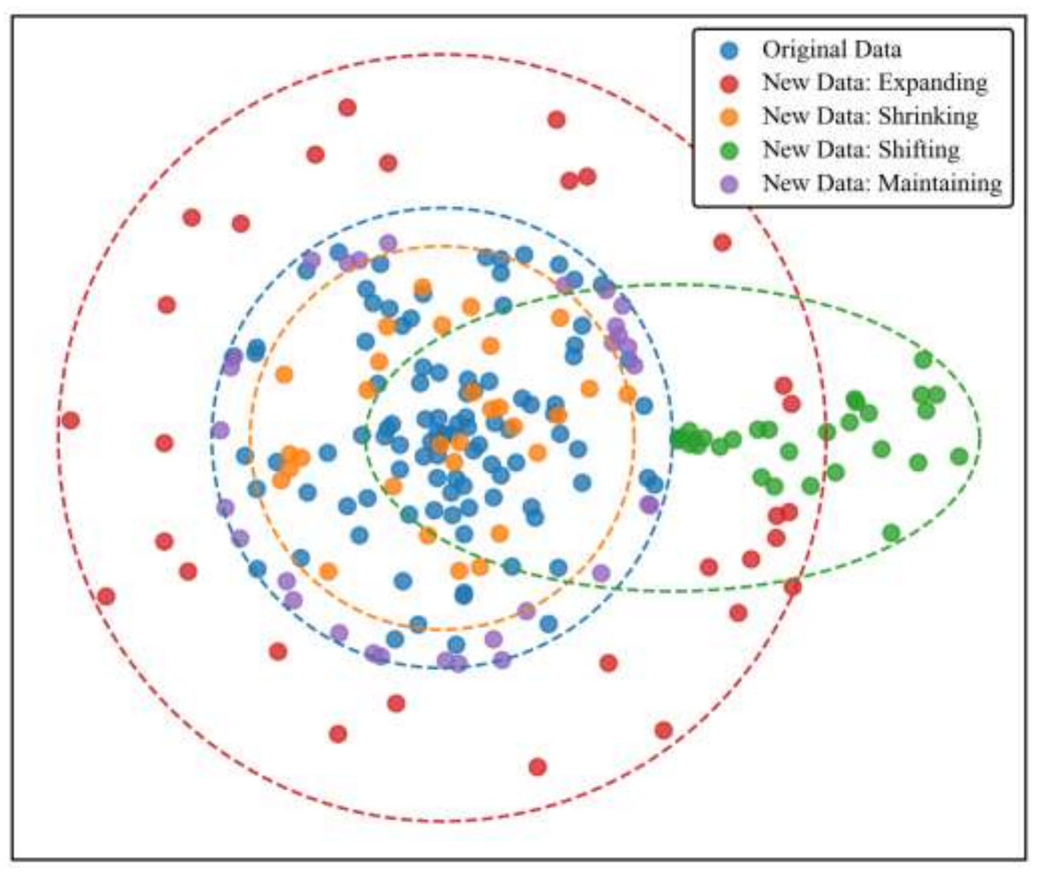

New Perspectives in Advancing Graph Machine Learning
NeurIPS 2025 Workshop · Room Exhibit Hall F · San Diego Convention Center · Dec 7, 2025

Introduction and Background

Data Geometry Matters for Model Performance
High-quality training data determines how well AI models learn and generalize.
Standard diversity metrics—like **Vendi Score** —only measure distributional spread.
They **miss geometric structure**: how data clusters, separates, or forms topological patterns.
But not all "diverse" data helps; poorly structured diversity can harm generalization.

Research Question:
Which geometric properties of data make it truly useful for model training?

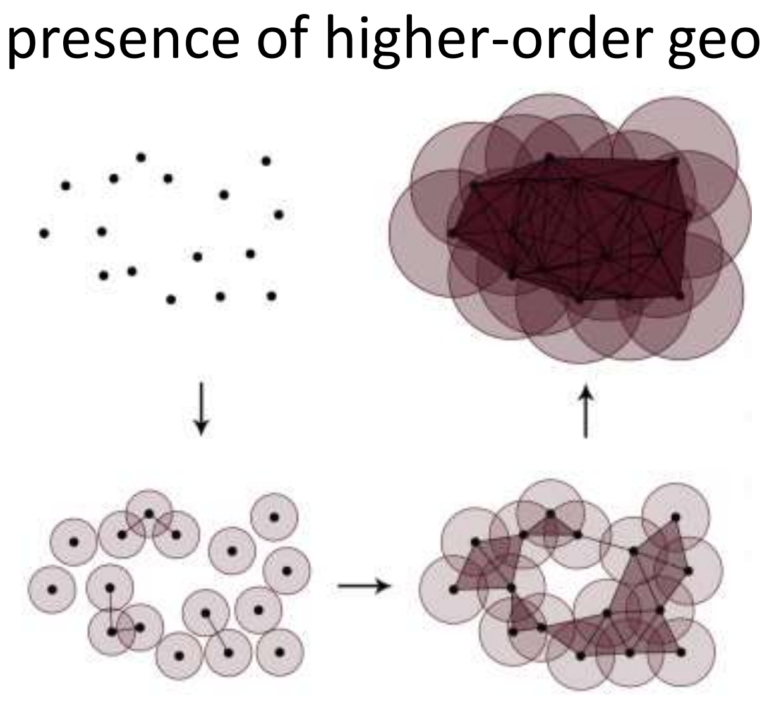
Four geometric augmentation scenarios that motivate our focus on data structure.



Our idea:
Use **Persistent Homology (PH)** to capture richer topological features (H_0 clusters, H_1 loops) that go beyond entropy-based measures.

We show:
PH-based diversity reveals structural patterns that Vendi Score cannot.
These measures (proved to satisfy standard diversity axioms) correlate with model performance.
Understanding data geometry guides better augmentation, selection, and synthetic data design.

Our Approach: Geometry via Persistent Homology
Goal: Capture dataset structure—not just distribution—by analyzing its geometry in a metric space.
Persistent Homology (PH) provides:
H_0 (clusters): how well-separated the data is
H_1 (loops): presence of higher-order geometric patterns



Lifetimes: how long these features persist as we expand the distance scale

We summarize these lifetimes into:

- **Persistence Entropy**
- **PH-based Hill Numbers (PEH)**
  These give a principled measure of **structural diversity** in the data.

PH captures topological structure that Vendi Score(VS) cannot:

- VS → measures entropy of similarity
- **PH → measures geometric structure (clusters + loops) across scales**

Our PH-based diversity measures **satisfy classic diversity axioms** (effective size, twin property, multi-scale, symmetry), with proofs included in the paper.

Methodology

PH-Based Diversity
Goal: Quantify dataset structural diversity using persistent homology lifetimes extracted from a Vietoris–Rips complex built on the pairwise distance matrix.

Procedure:
1. Build the pairwise distance matrix D.$\{\mathrm{VR}_\epsilon(D)\}_{\epsilon\geq 0}$
2. Construct a Vietoris–Rips filtration:
3. Compute persistence intervals

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

4. Compute lifetimes $l_i = d_i - b_i$

We focus on:

- H_0 (connected components) — cluster structure
- H_1 (loops) — higher-order geometric patterns

PH Entropy & Hill Numbers
Normalized persistence weights:

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

Rényi persistence entropy:

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

PH-based Hill numbers (PEH):

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

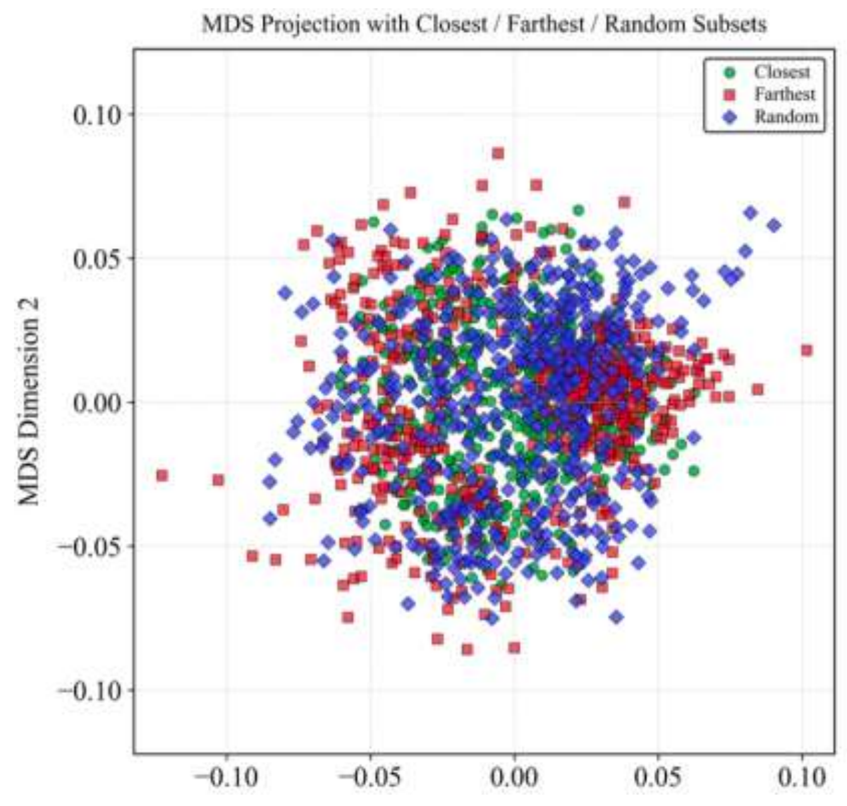
Why PH Diversity Works:

- Captures stability of geometric features across scales.
- Longer lifetimes = more significant topological structure.
- Provides a principled structural metric beyond distribution-based measures.

Experiment Setup

To study how **data geometry** affects training, we construct three balanced subsets based on each point's **maximum distance** to all others:

- **Closest:** points with smallest max-distance (core, redundant)
- **Farthest:** points with largest max-distance (peripheral, sparse)
- **Random:** uniform sampling across the dataset



A demonstration of three representative subsets construction for "Medical" dataset

We then analyze their PH summaries (H_0 & H_1 lifetimes) and compare how well models trained on each subset generalize.

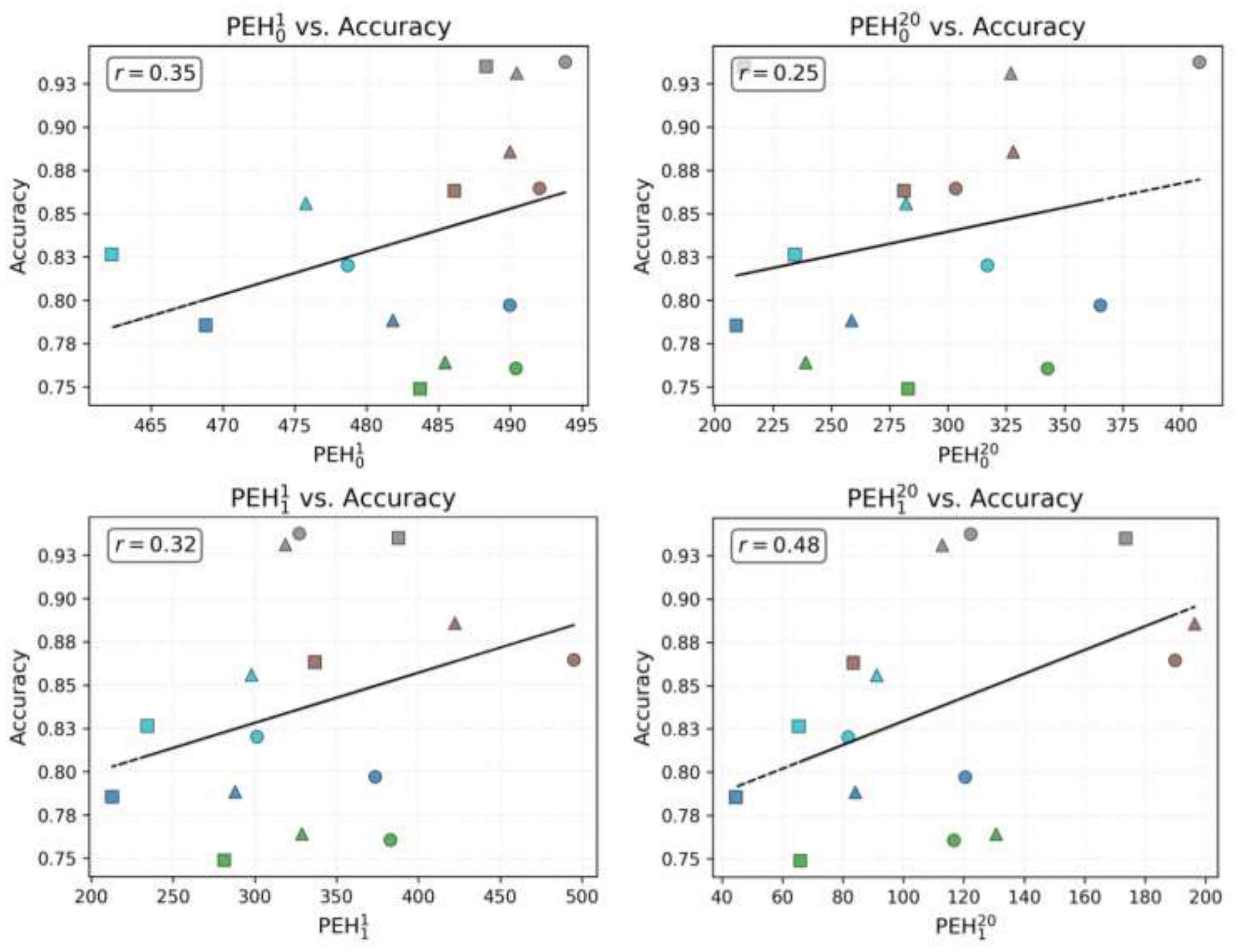
Experiment Setup

- Task: **fine-tune BERT** for text classification
- Datasets: TC, SUBJ, SentEval, Arxiv-10, Medical
- Each subset contains **500 samples (250 per class)**
- Training: **8 epochs**, learning rate 1×10^{-6} ,dropout 10% - 3 runs per subset for stability

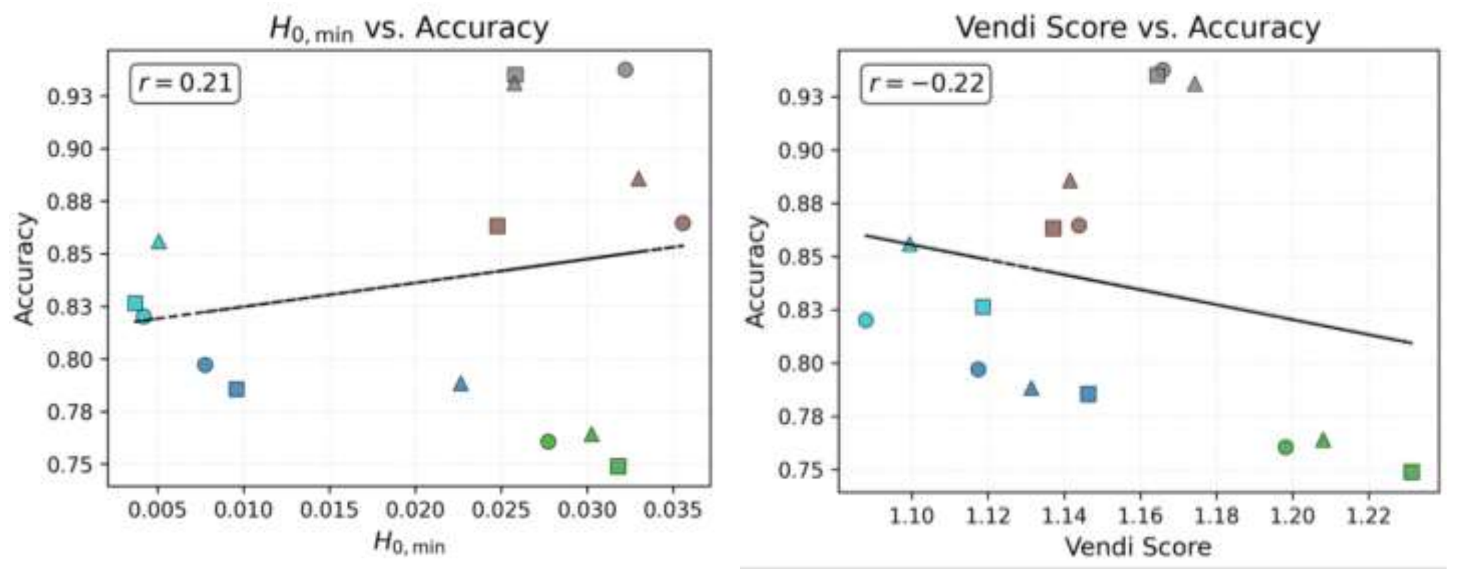
Experiment Result

Question: Can structural diversity predict which data subsets produce higher accuracy?

Yes — **PH-based diversity (H_0 + H_1) is positively correlated with accuracy.**



Vendi Score shows the opposite trend (higher value → lower accuracy).



- A High-quality datasets should exibit:
  - **well-separated clusters (H_0)**
  - **some stable loops (H_1)**
- Avoid extremes:
  Redundancy → instability,
  Sparsity → poor generalization

Subset	Accuracy (avg ± 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

- The **random** sampling often approximates ideal structure — the best balance of H_0 and H_1 structure, resulting in the **highest accuracy and lowest variance**.
- **Our experiments highlight that: More data ≠ better** — what matters is **geometric structural diversity**. With good structural diversity, **6–19%** of the full dataset can achieve **91–98.6%** of full-data accuracy

Conclusion

- Training data **geometry**, captured through persistent homology, is linked to model performance.
- Entropy-based metrics alone (e.g., Vendi Score) **cannot** predict data quality reliably.
- PH-based diversity distinguishes **meaningful structure** from **noise or redundancy**.
- The best-performing datasets show:
  - **well-balanced, well-separated clusters (H_0)**
  - **stable geometric loops (H_1)**
- Structural diversity leads to **higher accuracy** and **more stable training**.
- These insights support more principled dataset construction, augmentation, and synthetic data generation.

Future direction:
Leverage PH features directly to guide robust training and improve generalization without relying on ever-larger datasets.

yangba@asu.edu
mabolhas@asu.edu
rong.pan@asu.edu

Arizona State University
School of Computing and Augmented Intelligence (SCAI)