

# DMT-HI: MOE-based Hyperbolic Interpretable Deep Manifold Transformation for Unsupervised Dimensionality Reduction

Zelin Zang\*, Member, IEEE, Yuhao Wang\*, Student Member, IEEE, Jinlin Wu, Student Member, IEEE, Hong Liu, Member, IEEE, Yue Shen, Member, IEEE, Stan.Z Li†, Fellow, IEEE and Zhen Lei†, Fellow, IEEE

**Abstract**—Dimensionality reduction (DR) plays a crucial role in various fields, including data engineering and visualization, by simplifying complex datasets while retaining essential information. However, the challenge of balancing DR accuracy and interpretability remains crucial, particularly for users dealing with high-dimensional data. Traditional DR methods often face a trade-off between precision and transparency, where optimizing for performance can lead to reduced interpretability, and vice versa. This limitation is especially prominent in real-world applications such as image, tabular, and text data analysis, where both accuracy and interpretability are critical. To address these challenges, this work introduces the MOE-based Hyperbolic Interpretable Deep Manifold Transformation (DMT-HI). The proposed approach combines hyperbolic embeddings, which effectively capture complex hierarchical structures, with Mixture of Experts (MOE) models, which dynamically allocate tasks based on input features. DMT-HI enhances DR accuracy by leveraging hyperbolic embeddings to represent the hierarchical nature of data, while also improving interpretability by explicitly linking input data, embedding outcomes, and key features through the MOE structure. Extensive experiments demonstrate that DMT-HI consistently achieves superior performance in both DR accuracy and model interpretability, making it a robust solution for complex data analysis. The code is available at [https://github.com/zangzelin/code\\_dmhi](https://github.com/zangzelin/code_dmhi).

**Index Terms**—Dimensionality Reduction, Mixture of Experts (MOE), Hyperbolic Embedding, Interpretability

## I. INTRODUCTION

Dimensionality reduction [1], [2], [3], [4] simplifies complex datasets while preserving their intrinsic structure [5], [6]. This is crucial for managing high-dimensional data, which presents challenges in computational complexity, storage, and visualisation [7], [8]. Reducing data dimensionality allows for more efficient analysis, pattern recognition, and interpretation [9]. However, balancing high performance [10] and interpretability [11], [12] remains challenging, as efficient processing [13] often conflicts with human interpretability [14], [15].

Zelin Zang is with Centre for Artificial Intelligence and Robotics (CAIR), HKISI-CAS and Westlake University. email: zangzelin@westlake.edu.cn

Yuhao Wang and Stan.Z Li are with Westlake University. Hong Liu is with School of Information and Electrical Engineering, Hangzhou City University, Hangzhou, 310015 China and Academy of Edge Intelligence Hangzhou City University, Hangzhou City University, 310015 China. Jinlin Wu is with CAIR, HKISI-CAS; State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), CASIA. Zhen Lei is with CAIR, HKISI-CAS; MAIS, CASIA; and School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS). Yue Shen is with Ant Group.

Manuscript received Oct 8, 2024

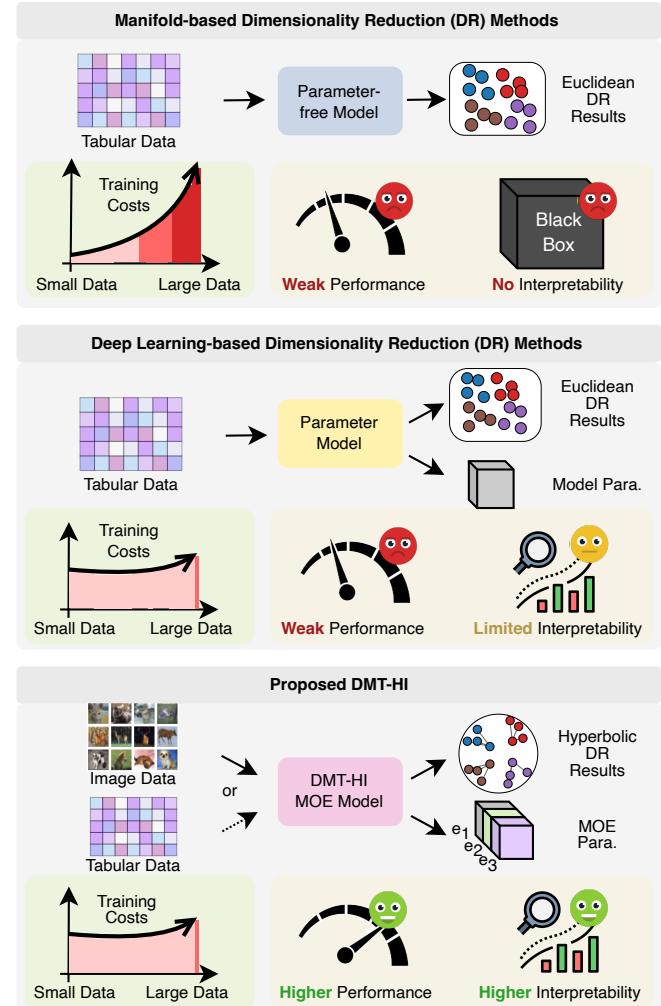


Fig. 1. **Overview of the proposed DMT-HI network.** The figure compares three dimensionality reduction methods, manifold-based, deep learning-based, and the proposed DMT-HI. DMT-HI leverages a Mixture of Experts (MOE) strategy to efficiently process both image [16] and tabular [17] data, offering better performance, lower training costs, and improved interpretability across different data sizes.

Dimensionality reduction methods fall into two categories, manifold-based parameter-free approaches [18], [19], [20] and deep learning-based methods [21]. Manifold-based methods like t-SNE [22] and UMAP [23] are known for their speed (on small dataset) and adaptability [7], projecting high-dimensional data into low-dimensional spaces through nonlinear mappings, revealing underlying structures. Deep learning

methods, such as parametric UMAP (PUMAP) [24], [25] and DMT-EV [26], handle complex data more effectively, especially high-dimensional data, leveraging neural networks to capture intricate patterns. DMT-EV, in particular, excels in both performance and explainability [27], pruning irrelevant features for clearer, more interpretable results. Deep learning methods stand out for their ability to scale and generalize well across large datasets, positioning them as central to current dimensionality reduction research.

In terms of method efficiency, performance, and interpretability, manifold-based parameter-free methods and deep learning-based methods each have distinct strengths and weaknesses, driven by their theoretical and design differences [7]. Parameter-free methods are more efficient for small datasets with lower time costs since they don't rely on parametric models and focus on optimizing the output [11]. However, their efficiency declines with increasing data size due to the rising complexity of neighborhood search and distance calculations. In contrast, deep learning methods handle large-scale data more efficiently due to model scalability and hardware acceleration, though their training on small datasets is costlier. In terms of performance, parameter-free methods excel at capturing local structures but struggle with complex global hierarchies due to their reliance on Euclidean space [28]. Deep learning methods, by contrast, can capture both local and global structures through multilayer transformations but require significant data and computational resources. Regarding interpretability, parameter-free methods rely on similarity metrics, making them harder to interpret and inconsistent globally. While deep learning methods can capture richer features, their "black-box" nature and complexity make their decision-making process harder to explain.

To address these challenges in global structure characterization and interpretability, As shown in Fig. 1, we propose the **MOE-based Hyperbolic Interpretable Deep Manifold Transformation (DMT-HI)**, which integrates hyperbolic mapping and a Mixture of Experts model (MOE) [29], [30]. Hyperbolic mapping uses negative curvature to better capture complex hierarchical structures and global dependencies, preserving global information in lower dimensions. The MOE strategy enhances performance and efficiency by dynamically assigning tasks to specialized expert networks that handle different input features, thereby avoiding bottlenecks from a single model. This model provides interpretability by allowing users to track expert decisions and understand the internal model workings. Additionally, MOE serves as a bridge between raw data, embedded data, and feature subsets, enabling clear interpretation of how features influence data representations at different levels. By combining hyperbolic mapping's structural advantages and MOE's efficient task allocation, DMT-HI aims to improve performance, efficiency, and interpretability, offering a comprehensive solution for reducing the dimensionality of complex data.

By integrating these innovations, our approach not only improves dimensionality reduction performance but also enhances interpretability, offering a comprehensive solution for handling complex data types and extracting insights from high-dimensional datasets. DMT-HI's performance in dimen-

sionality reduction is enhanced through the MOE strategy, which dynamically assigns tasks to the most suitable experts, improving both processing efficiency and overall model performance. Additionally, the redesigned manifold loss optimizes the training process, enabling the model to capture both local and global structures more effectively. Overall, the key contributions of this paper include,

- A hyperbolic embedding and deep manifold loss function, which improve the accuracy of dimensionality reduction by better capturing the global structure of data.
- The introduction of the MOE strategy, establishing a clear connection between input data, embedding results, and key features, thus enhancing model interpretability and stability.
- Comprehensive tests evaluating global and local performance, time efficiency, and other dimensions to validate the advantages of the proposed model.

## II. RELATED WORK

**Dimension Reduction & Visualization.** Dimensionality reduction (DR) methods can be classified into parametric-free and parametric approaches. *Parametric-free* methods, like MDS [31], ISOMAP [32], LLE [33], and t-SNE [34], optimize the output directly, focusing on preserving distances, particularly local manifold structures. However, these methods often lack generalizability and interpretability, as they do not capture underlying feature relationships and cannot handle new datasets well. Parametric methods, such as Topological Autoencoders (TAE) [35] and PUMAP [36], address these shortcomings by using neural networks for continuous mappings and input space constraints, improving both generalizability and interpretability. NeuroDAVIS [37] further enhance DR performance and visualization by focusing on both local and global structure retention. While progress has been made, many methods still struggle with effective loss functions and training approaches, especially in capturing complex relationships like those in biological data. Approaches integrating techniques like causal representation learning, as seen in Cells2Vec [38], represent a future direction for enhancing both interpretability and robustness in DR.

**Explainability & Interpretability of DR Methods.** Explainability focuses on revealing a model's internal workings, while interpretability ensures the model's outputs are easily understood by users [11], [12]. In DR, explainability methods like DMT-EV [26] utilize saliency maps and interactive visual tools to provide insights into how different components influence the embedding process. Other approaches, such as DimReader [39] and DataContextMap [40], enhance the interpretability of embeddings by visually linking features to their impact on reduced data spaces. Interpretability is particularly critical in domains like healthcare and finance, where understanding the relationship between input data and predictions is crucial for decision-making [41], [42]. Approaches such as DT-SNE [43], which combines t-SNE embeddings with decision tree rules, offer clear explanations of how data points relate in the embedding space, further advancing the interpretability of machine learning models in these high-stakes areas.

**Hyperbolic Embeddings.** Hyperbolic embeddings have gained prominence for their ability to efficiently represent hierarchical data, outperforming Euclidean spaces in capturing both local and global structures [44]. These embeddings have proven valuable in tasks such as zero-shot learning [45], graph embeddings, and hierarchical data visualization. In visualization, hyperbolic embeddings are used in tools like scDHMap [46] for single-cell RNA-seq data, where they reveal complex developmental trajectories and reduce noise. Methods like D-Mercator [47] extend hyperbolic spaces for network analysis, while techniques like Hyperbolic Informed Embedding (HIE) [48] improve task performance by maintaining hierarchical relationships. Despite their strengths, challenges with numerical stability remain [49], highlighting areas for further development in making hyperbolic embeddings more robust and scalable for broader applications. The Detailed related work is in Appendix.

### III. PROBLEM DEFINITION AND PRELIMINARIES

#### A. Data Description and Data Augmentation.

In this section, we provide a formal description of the data preprocessing, data augmentation [50], dimensionality reduction, and interpretability concepts used in the proposed method. These concepts are essential for understanding the subsequent sections of the paper. The input data  $\mathbf{X}$  can be image data, tabular data, or sequential data. Assume we have  $N$  samples, with each sample represented as  $\mathbf{x}_i$ :

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N\}. \quad (1)$$

Data augmentation is applied to enhance the robustness and diversity of the training samples. Formally, we define the augmented data  $\mathbf{X}^{aug}$  as:

$$\begin{aligned} \mathbf{X}^{aug} &= \{\mathbf{x}_1^{aug}, \mathbf{x}_2^{aug}, \dots, \mathbf{x}_i^{aug}, \dots, \mathbf{x}_N^{aug}\}, \\ \mathbf{x}_i^{aug} &= \tau(\mathbf{x}_i), \end{aligned} \quad (2)$$

where each  $\mathbf{x}_i^{aug}$  is obtained through specific augmentation techniques based on the data type.

**Tabular Data.** For tabular data, we normalize each feature to have zero mean and unit variance. This helps in stabilizing the training process and ensures that each feature contributes equally [51]. We further perform data augmentation through neighbor discovery and linear interpolation among neighbors. Specifically, for each data point  $\mathbf{x}_i$ ,

$$\begin{aligned} \tau^{\text{tab}}(\mathbf{x}_i) &= \lambda \mathbf{x}_i + (1 - \lambda) \mathbf{x}_j, \\ \mathbf{x}_j &\in \mathcal{N}^k(\mathbf{x}_i), \quad \lambda \sim \mathcal{U}(0, 1) \end{aligned} \quad (3)$$

where  $\lambda$  is a random interpolation factor sampled from a uniform distribution between 0 and 1, and  $\mathbf{x}_j$  is a randomly selected neighbor of  $\mathbf{x}_i$ ,  $\mathcal{N}^k(\mathbf{x}_i)$  is the function to discover the  $k$  nearest neighbors of  $\mathbf{x}_i$ ,  $k$  is a hyperparameter.

**Image Data.** For image data, we employ common data augmentation techniques, including random cropping, flipping, and rotation, to enhance the diversity and robustness of the training dataset [52], [53], [50]. Formally, for a given image  $\mathbf{x}_i$ , we apply a set of transformations denoted as  $\tau^{\text{img}}$ ,

$$\mathbf{x}_i^{aug} = \tau^{\text{img}}(\mathbf{x}_i), \quad (4)$$

where  $\tau^{\text{img}}$  includes color jittering [54], random cropping [55], applying Gaussian blur [56], Mixup [57] and other domain-specific augmentations [58]. The details of the augmentation techniques for image data are provided in the Appendix.

#### B. Interpretability in Dimensionality Reduction.

Let  $f$  to be a DR mapping function,  $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$ ,  $\mathbf{h}_i = f(\mathbf{x}_i)$ ,  $d \ll D$ . The parameterized methods like autoencoders learn mappings through optimization, allowing generalization, whereas non-parametric methods like t-SNE optimize for specific datasets.

**Feature Interpretability** ( $I_{\text{feat}}$ ) assesses the influence of individual features on DR outputs,

$$I_{\text{feat}} : (\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{h}_i\}_{i=1}^N, f) \rightarrow \mathcal{I}_{\text{feat}}, \quad (5)$$

where  $\mathcal{I}_{\text{feat}}$  measures the importance of features in determining low-dimensional representations.

**Group Interpretability** ( $I_{\text{group}}$ ) evaluates the collective importance of feature groups,

$$I_{\text{group}} : (\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{h}_i\}_{i=1}^N, f) \rightarrow \mathcal{I}_{\text{group}}, \quad (6)$$

where  $\mathcal{I}_{\text{group}}$  quantifies the impact of interacting feature subsets on the embeddings.

#### C. Gumbel-Softmax for Task Allocation.

Gumbel-Softmax [59] enables differentiable sampling from a categorical distribution, useful for task allocation in neural networks. Given logits  $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, \dots, \mathbf{L}_D)$  and temperature  $\tau$ , the  $D$  is the number of the dimensionality in raw data. The Gumbel-Softmax distribution introduces Gumbel noise  $\mathbf{g}_i$  and applies a softmax function,

$$\mathbf{y}_i = \frac{\exp((\mathbf{L}_i + \mathbf{g}_i)/\tau)}{\sum_{j=1}^D \exp((\mathbf{L}_j + \mathbf{g}_j)/\tau)} \quad (7)$$

where  $\mathbf{g}_i$  is drawn from a Gumbel(0,1) distribution, and  $\tau$  controls the smoothness, As  $\tau \rightarrow 0$ , the distribution approaches a categorical one, making hard selections. Higher  $\tau$  allows softer outputs for exploration. For task allocation in a Mixture of Experts (MOE) model, this technique assigns input  $\mathbf{x}_i$  to an expert by computing the logits  $\mathbf{L}_i$  and sampling task probabilities. The expert with the highest probability is selected,

$$\text{Expert Assignment} = \arg \max_i \mathbf{y}_i, \quad (8)$$

This allows for dynamic task allocation while preserving gradient-based optimization.

### IV. METHODS

To enhance both performance and interpretability in dimensionality reduction (DR), as shown in Fig. 2, we propose the Hyperbolic Interpretable Deep Manifold Transformation (DMT-HI), leveraging a Mixture of Experts (MOE) framework [29], [60]. DMT-HI addresses limitations in existing DR methods by capturing complex, non-Euclidean structures [61] while offering a more interpretable mapping process. It incorporates three key components, Multiple Gumbel Operator-Based Matchers, a MOE network, and a Hyperbolic Mapper. These components work together to improve DR quality while ensuring transparency in the data transformation process.

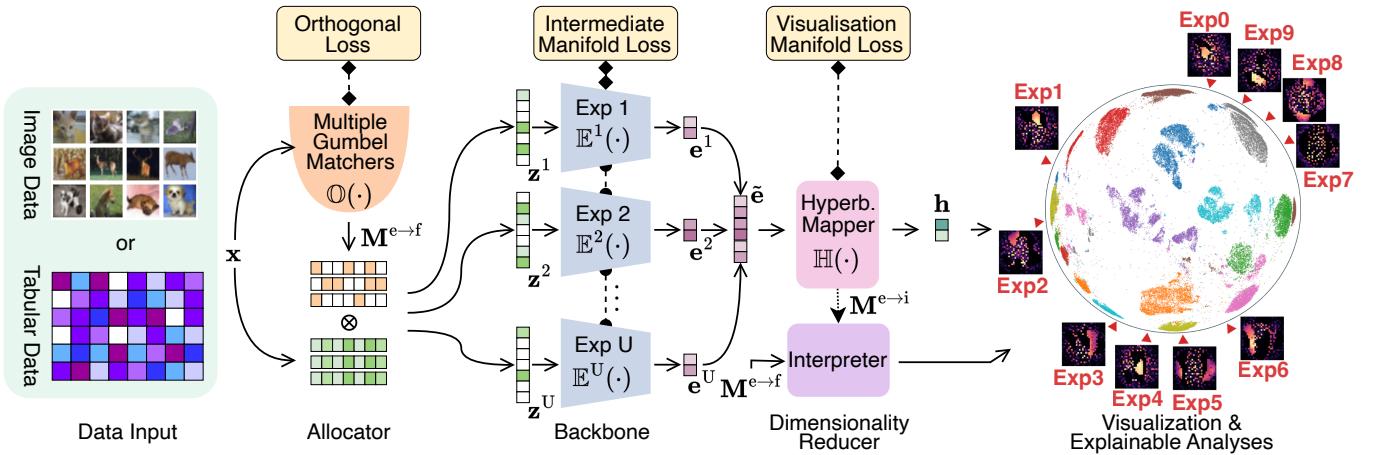


Fig. 2. **Overview of the proposed MOE-based Hyperbolic Interpretable Deep Manifold Transformation (DMT-HI) model.** The model processes input data (images or tabular) through four key components. First, the Allocator, using Multiple Gumbel Operator-Based Matchers, assigns different data segments to expert networks in the MOE backbone, ensuring diverse task allocation with orthogonal loss. The Backbone extracts features through expert networks while preserving manifold structure. The Hyperbolic Mapper then maps the data into a hyperbolic space to capture non-Euclidean relationships. Finally, the Interpreter provides enhanced interpretability and visualization, refining the output through a visualisation manifold loss. This approach ensures robust, interpretable dimensionality reduction across data types.

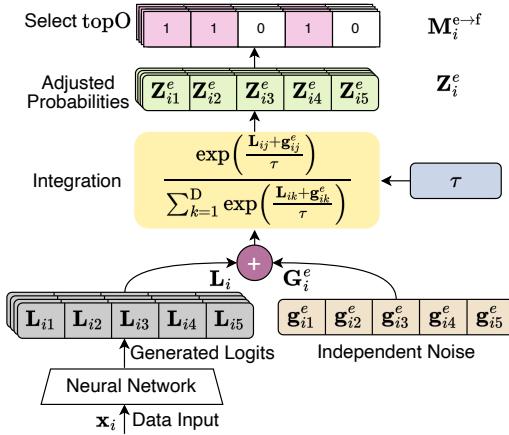


Fig. 3. **Overview of the Multiple Gumbel Matchers.** This figure illustrates a selection and integration mechanism for adjusting generated logits and independent noise. The topO selection identifies relevant elements, which are then integrated with independent noise through a softmax based normalization process to produce adjusted probabilities, enhancing the decision making or generation tasks.

### A. Multiple Gumbel Matchers

Effective task allocation is crucial in Mixture of Experts (MOE) models [29], as the way features are distributed to experts impacts both performance and interpretability. Random allocation can lead to suboptimal specialization. We address this by extending Gumbel-Softmax for selecting feature subsets, improving expert allocation efficiency.

As shown in Fig. 3, we propose the Multiple Gumbel Operator  $\textcircled{\text{O}}(\cdot)$ , which extends traditional Gumbel-Softmax to select multiple features per expert. A neural network (NN) first processes the input data  $\mathbf{x}_i$  to generate logits,  $\mathbf{L}_i = [\mathbf{L}_{i1}, \mathbf{L}_{i2}, \dots, \mathbf{L}_{iD}]$ , where  $D$  is the number of features. We then sample independent noise values  $\mathbf{g}_i^e = [g_{i1}^e, g_{i2}^e, \dots, g_{iD}^e]$  from the Gumbel distribution, which are added to the logits and scaled by the temperature parameter  $\tau$ . This yields adjusted probabilities,

$$Z_{ij}^e = \exp\left(\frac{\mathbf{L}_{ij} + \mathbf{g}_{ij}^e}{\tau}\right) / \sum_{k=1}^D \exp\left(\frac{\mathbf{L}_{ik} + \mathbf{g}_{ik}^e}{\tau}\right), \quad (9)$$

where  $Z_{ij}^e$  is probability of selecting feature  $j$  for expert  $e$ .

To control feature allocation, we introduce the hyperparameter  $O$ , which specifies how many features each expert receives. The top-O features with the highest probabilities are selected. Formally, for each expert  $e$ , the mask element  $\mathbf{m}_{ij}^e$  is defined as,

$$\begin{aligned} \mathbf{M}_i^{e-f} &= \{\mathbf{m}_i^1, \mathbf{m}_i^1, \dots, \mathbf{m}_i^e, \dots, \mathbf{m}_i^U\} \\ \mathbf{m}_i^e &= \{\mathbf{m}_{i1}^e, \mathbf{m}_{i2}^e, \dots, \mathbf{m}_{ij}^e, \dots, \mathbf{m}_{iD}^e\} \\ \mathbf{m}_{ij}^e &= \begin{cases} 1, & \text{if } j \in \text{Top-O}(Z_{i1}^e, Z_{i2}^e, \dots, Z_{iD}^e), \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (10)$$

where each  $\mathbf{m}_i^e$  is a binary vector indicating the features selected for expert  $e$ .

We apply the mask matrix using the Hadamard product between  $\mathbf{m}_i^e$  and the input data  $\mathbf{x}_i$ , generating the input for each expert,

$$\mathbf{x}_i^e = \mathbf{x}_i \odot \mathbf{m}_i^e. \quad (11)$$

This approach ensures each expert processes a relevant subset of features, improving performance. The explicit selection mechanism also enhances interpretability by revealing which features are assigned to which experts, clarifying their contributions to the model's output.

### B. Mixture of Experts Network

In our proposed model, the integration of outputs from multiple experts is key to capturing the intricate, multi-faceted nature of the input data. Each expert model  $\mathbb{E}^e(\cdot)$  operates on the allocated input data  $\mathbf{x}_i^e$  using the corresponding mask vector  $\mathbf{m}_i^e$  generated by the Multiple Gumbel Operator. This allocation ensures that each expert focuses on specific subsets of features, allowing for specialized processing that enhances the overall model's expressiveness and interpretability.

The ensemble of expert models is formally defined as,

$$\mathbb{E} = \{\mathbb{E}^1(\cdot), \mathbb{E}^2(\cdot), \dots, \mathbb{E}^U(\cdot)\}, \quad (12)$$

where each expert can be instantiated as either structurally identical or distinct backbone networks, such as Convolutional Neural Networks (CNNs) or Multi-Layer Perceptrons (MLPs).

In our experiments, we employ structurally identical backbone networks to maintain consistency and isolate the impact of feature allocation. Each expert processes its respective masked input to produce an output. To synthesize the contributions from all experts, we concatenate their outputs to obtain the final representation,

$$\tilde{\mathbf{e}}_i = \text{concat}(\mathbf{e}_i^1, \mathbf{e}_i^2, \dots, \mathbf{e}_i^e, \dots, \mathbf{e}_i^U), \mathbf{e}_i^e = \mathbb{E}^e(\mathbf{z}_i^e), \quad (13)$$

where  $\text{concat}(\cdot)$  denotes the concatenation operation. This fusion of expert outputs integrates the diverse representations learned by each expert, potentially capturing a richer set of features and improving the model's expressiveness.

Since each expert receives different subsets of the input data, there is sufficient diversity among them. This diversity guides the model to explore more stable representations, making the final output more robust. Additionally, this setup allows us to perform interpretability analysis based on the specific contributions of each expert, providing insights into how different parts of the input data influence model's performance.

### C. Hyperbolic Mapper

To transform the concatenated expert outputs  $\tilde{\mathbf{e}}_i$  into low-dimensional embeddings, we propose a Hyperbolic Multi-Layer Perceptron (HMLP) [62]. The HMLP leverages hyperbolic geometry to map high-dimensional expert representations into a lower-dimensional hyperbolic space, preserving the hierarchical structures often missed in Euclidean spaces.

The transformation of  $\tilde{\mathbf{e}}_i$  into a low-dimensional embedding  $\mathbf{h}_i \in \mathbb{R}^k$  (with  $k < d$ ) in hyperbolic space is defined as,

$$\text{HMLP}(\tilde{\mathbf{e}}_i) = \exp_{\mathbb{B}^n}(\mathbf{W} \cdot \log_{\mathbb{B}^n}(\tilde{\mathbf{e}}_i) + \mathbf{b}), \quad (14)$$

where  $\mathbf{W}$  is the weight matrix and  $\mathbf{b}$  the bias vector. The logarithmic  $\log_{\mathbb{B}^n}(\cdot)$  and exponential  $\exp_{\mathbb{B}^n}(\cdot)$  maps project data to and from the tangent space of the Poincaré ball  $\mathbb{B}^n$ , ensuring the model captures complex hierarchical relationships in a compact, hyperbolic form.

The HMLP comprises multiple layers, each performing a hyperbolic linear transformation followed by a hyperbolic activation function, such as the hyperbolic tangent ( $\tanh$  function). The utilization of hyperbolic activation functions enables the network to effectively model complex relationships and hierarchical structures present in the data, which are challenging to represent in Euclidean geometry. By embedding the data into a hyperbolic space, we aim to better capture these relationships and improve the performance of downstream tasks.

### D. Sub-manifold Matching Loss Function

To ensure that similarities in the high-dimensional expert output space are preserved in the lower-dimensional hyperbolic embeddings, we introduce a manifold loss function, denoted as  $\mathcal{L}_{\text{SMM}}$ . Let  $\mathbf{h}$  is a batch of hyperbolic embeddings of the original raw data, and  $\mathbf{h}^+$  is the augmented data, respectively, from the Hyperbolic MLP. The manifold loss is formulated as,

$$\begin{aligned} \mathcal{L}_{\text{SMM}}(\mathbf{h}, \mathbf{h}^+) = & \frac{1}{2} \left( \sum_i \log \sum_j \mathbf{S}_{ij}^{\mathbf{h}\mathbf{h}^+} + \sum_i \log \sum_j \mathbf{S}_{ij}^{\mathbf{h}^+\mathbf{h}} \right) \\ & - \gamma \cdot \sum_i \log \text{diag}(\mathbf{S}_{ii}^{\mathbf{h}\mathbf{h}^+}), \end{aligned} \quad (15)$$

where  $\gamma > 0$  is an exaggeration factor that emphasizes preserving local similarities in hyperbolic space. The similarity matrices  $\mathbf{S}_{ij}^{h\mathbf{h}^+}$  are computed using a t-distribution kernel, with pairwise distances calculated as,

$$\mathbf{S}_{ij} = (1 + \mathbf{D}_{ij}^2 / \nu)^{-\frac{\nu+1}{2}}, \quad (16)$$

where  $\nu$  the degrees of freedom of the t-distribution,  $\mathbf{D}_{ij}$  is the distance metric of  $\mathbf{h}_i$  and  $\mathbf{h}_j$ . These t-Student similarity matrices ensure that hyperbolic embeddings preserve local pairwise similarities, maintaining hierarchical and relational structures.

The loss in Eq. (15) is an unsupervised function that guides experts in dimensionality reduction by preserving structural information. Unlike parameter-free methods, our model learns mapping parameters that can be used to interpret the model. By balancing alignment and uniformity, this loss ensures that hyperbolic embeddings capture the relational structure of the original data, crucial for tasks that rely on preserving topology.

### E. Expert Exclusive Loss Function

To promote diversity among the representations of different experts and reduce redundancy, we introduce an orthogonal loss function, denoted as  $\mathcal{L}_{\text{Exc}}$ . This loss ensures that features learned by each expert are uncorrelated with those learned by others, while encouraging coherence within each expert's feature set.

Cosine similarity is computed both within the same expert and across different experts. For a batch of expert outputs  $\mathbf{e}$ , the expert orthogonal loss is defined as,

$$\mathcal{L}_{\text{Exc}}(\mathbf{e}) = \frac{1}{U^2 N_B} \sum_{i=1}^{N_B} \sum_{e_a \neq e_b} \{1 + \frac{\mathbf{e}_i^{e_a \top} \cdot \mathbf{e}_i^{e_b}}{\|\mathbf{e}_i^{e_a}\| \|\mathbf{e}_i^{e_b}\|}\}. \quad (17)$$

where  $N_B$  is the batch size, the  $\mathbf{e}_i^{e_a}$  and  $\mathbf{e}_i^{e_b}$  from expert  $e_a$  and  $e_b$  are two different experts' outputs. The expert orthogonal loss penalizes high similarity between different experts, promoting inter-expert diversity. The orthogonal loss minimizing this loss balances intra-expert cohesion with inter-expert diversity, enhancing the model's ability to capture diverse features and improving generalization.

The overall loss function is the sum of the sub-manifold matching loss  $\mathcal{L}_{\text{SMM}}$  and the expert orthogonal loss  $\mathcal{L}_{\text{Exc}}$ , ensuring both structural preservation and diverse expert,

$$\mathcal{L} = \mathcal{L}_{\text{SMM}}(\mathbf{h}, \mathbf{h}^+) + \mathcal{L}_{\text{SMM}}(\tilde{\mathbf{e}}, \tilde{\mathbf{e}}^+) + \lambda \mathcal{L}_{\text{Exc}}(\mathbf{e}), \quad (18)$$

where  $\lambda$  controls the trade-off between structural preservation and expert diversity.

### F. Interpretability through MOE – Connecting Raw Data, Representation, and Raw Features

The Mixture of Experts (MOE) technique improves model performance by dividing input data into segments, each processed by specialized expert models. A gating network assigns samples to experts probabilistically, and the final prediction is a weighted sum of expert outputs. This structure ensures efficient resource allocation and enhanced feature representation through expert specialization. In our MOE-based model, task

**Algorithm 1** The DMT-HI algorithm

**Input:** Data:  $\mathbf{X}$ , Learning rate:  $\alpha$ , Epochs:  $E$ , Batch size:  $B$ , Number of experts:  $U$ , Temperature:  $\tau$ , Hyperbolic parameters:  $\gamma, \nu$ , Loss weights:  $\lambda$ , **Output:** Low-dim embeddings:  $\mathbf{H}^l$ , Feature masks:  $\mathbf{M}$ , Expert:  $\mathbf{E}$ .

```

1: Let  $t = 0$ .
2: while  $i = 0; i < E; i++$  do
3:   while  $b = 0; b < [\mathbf{|X|}/B]; b++$  do
4:      $\mathbf{X}_b \leftarrow \text{Sampling}(\mathbf{X}, b); \mathbf{X}'_b \leftarrow \text{Augment}(\mathbf{X}_b)$ ; # Sample a batch of data and Data augmentation
5:      $\mathbf{M}_b \leftarrow \text{Multiple_Gumbel_Operator}(\mathbf{X}'_b, \tau, U)$ ; # Select features for each expert using Gumbel-Softmax, see (10)
6:     for each expert  $e = 1$  to  $U$  do
7:        $\mathbf{Z}_b^e \leftarrow \mathbf{X}_b \odot \mathbf{M}_b^e$ ; # Apply feature mask for each expert, see (11)
8:        $\mathbf{E}_b^e \leftarrow \text{Expert}(\mathbf{Z}_b^e)$ ; # Process features through each expert, see (13)
9:     end for
10:     $\mathbf{E}_b \leftarrow \text{Aggregate}(\{\mathbf{E}_b^1, \mathbf{E}_b^2, \dots, \mathbf{E}_b^U\})$ ; # Aggregate expert outputs, see (13)
11:     $\mathbf{H}_b^h \leftarrow \text{HyperbolicMLP}(\mathbf{E}_b); \mathbf{H}_b^l \leftarrow \text{HyperbolicMapper}(\mathbf{H}_b^h)$ ; # Map to hyperbolic space, see (14)
12:     $\mathbf{S}^h \leftarrow \text{SimilarityMatrix}(\mathbf{H}_b^h); \mathbf{S}^l \leftarrow \text{SimilarityMatrix}(\mathbf{H}_b^l)$ ; # Calculate similarity matrices
13:     $\mathcal{L}_{\text{SMM}} \leftarrow \mathcal{L}_{\text{SMM}}(\mathbf{H}_b^h, \mathbf{H}_b^l) + \mathcal{L}_{\text{SMM}}(\mathbf{E}_b, \mathbf{E}'_b)$ ; # Sub-manifold matching loss
14:     $\mathcal{L}_{\text{Exc}} \leftarrow \frac{1}{N_B} \sum_i \mathcal{L}_{\text{Exc}}(\mathbf{E}_b^1, \mathbf{E}_b^2, \dots, \mathbf{E}_b^U)$ ; # Orthogonal loss to ensure expert diversity
15:     $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{SMM}} + \lambda \cdot \mathcal{L}_{\text{Exc}}$ ; # Total loss combining manifold and orthogonal losses
16:     $\text{Optimize}(\mathcal{L}_{\text{total}}, \alpha)$ ; # Update model parameters
17:  end while
18: end while
19: Return  $\mathbf{H}^l, \mathbf{M}, \mathbf{E}$ ; # Return final low-dimensional embeddings, feature selection masks, and expert outputs

```

TABLE I

GLOBAL STRUCTURE PRESERVATION PERFORMANCE (SVM CLASSIFICATION ACCURACY) COMPARISON ON TEN DATASETS. **BOLD** DENOTES THE BEST RESULT AND UNDERLINE DENOTES 1% HIGHER THAN OTHERS (NO BOLDED). ‘-’ MEANS THE OFFICIAL IMPLEMENTATION IS NOT AVAILABLE.

	SVM Classification Accuracy - training set							SVM Classification Accuracy - testing set						
	tSNE (2014)	UMAP (2018)	IVIS (2019)	PaCMAP (2021)	PUMAP (2023)	DMT-EV (2024)	DMT-HI (Ours)	tSNE (2014)	UMAP (2018)	IVIS (2019)	PUMAP (2023)	DMT-EV (2024)	DMT-HI (Ours)	
20News	34.4	28.4	25.2	28.9	33.8	34.6	<b>43.3</b>	33.5	28.2	25.5	32.5	32.3	<b>41.1</b>	
MNIST	95.2	96.4	76.9	95.9	96.7	97.1	<b>98.0</b>	95.1	95.6	77.0	95.2	96.2	<b>97.4</b>	
K-MNIST	69.9	73.7	58.6	73.2	73.1	73.3	<u>74.3</u>	48.7	47.7	37.4	38.9	70.9	<u>74.1</u>	
E-MNIST	65.2	66.6	29.6	64.9	64.1	68.6	<b>71.8</b>	63.4	63.6	28.9	60.8	67.8	<b>69.9</b>	
Cifar10	22.3	21.8	21.1	22.3	-	22.2	<u>77.5</u>	22.9	23.4	22.0	-	23.0	<b>74.9</b>	
Cifar100	4.8	5.3	4.8	4.9	-	5.2	<u>39.1</u>	3.8	4.4	4.3	-	4.6	<u>38.9</u>	
GAST	65.3	57.7	64.4	79.9	63.7	<b>82.7</b>	80.3	61.4	49.5	63.5	61.7	75.0	<b>78.4</b>	
SAM	69.1	69.7	64.7	71.5	69.2	<b>71.8</b>	69.5	69.0	68.0	63.2	69.0	<b>72.4</b>	71.3	
HCL	68.7	41.6	53.4	78.5	46.3	78.3	<u>84.5</u>	63.7	38.6	47.4	42.7	72.6	<b>76.7</b>	
MCA	46.0	37.8	71.2	76.2	-	78.1	<u>80.7</u>	44.7	38.2	69.7	-	<b>77.4</b>	75.5	
AVE RANK	54.1	49.9	47.0	59.6	-	61.2	<b>71.9</b>	50.6	45.7	43.9	-	59.2	<b>69.8</b>	
	4	5	6	3	-	2	1	3	4	5	-	2	1	

allocation masks  $M_i$  are assigned using the Multiple Gumbel Operator, targeting specific features for expert processing and strengthening the link between data and features, thereby improving interpretability.

To quantify the impact of each expert’s output, we perturb the output of the  $e$ -th expert for the  $i$ -th sample,  $\mathbf{z}_i^e + \delta$ , and calculate the change in model predictions,

$$\Delta^\delta \mathbf{e}_i^e = \mathbb{E}^e(\mathbf{z}_i^e + \delta) - \mathbb{E}^e(\mathbf{z}_i^e), \quad (19)$$

where  $\mathbb{E}^e(\cdot)$  is expert output and  $\delta$  is a perturbation.

We define an importance metric  $\mathbf{I}_i$  for each sample  $\mathbf{x}_i$ , capturing the sensitivity of features processed by each expert,

$$\mathbf{I}_{i,e,j} = |\Delta^\delta \mathbf{e}_{ij}^e / \delta|. \quad (20)$$

This importance tensor  $\mathbf{I}_i$  measures how experts contribute to features, aiding in understanding the model’s decision process.

To bridge the experts, input features, and data points, we define two connection matrices, one linking experts to features ( $\mathbf{M}^{e,f}$ ) and another linking experts to data points ( $\mathbf{M}^{e,i}$ ). These are computed by averaging the importance tensor across relevant dimensions. The expert-to-data connection matrix is,

$$\mathbf{M}_{e,i}^{e \rightarrow i} = \frac{1}{D} \sum_{j=1}^D \mathbf{I}_{i,e,j}, \quad (21)$$

where  $D$  is the number of features.

By visualizing these matrices, we can project expert representations onto the boundary of hyperbolic space, identifying key regions where experts are most active. This approach clarifies how experts specialize in processing different data subsets, enhancing both transparency and interpretability within the

TABLE II  
THE TRUSTWORTHINESS STRUCTURE-PRESERVING PERFORMANCE COMPARISON ON TEN DATASETS. **BOLD** DENOTES THE BEST RESULT,  
UNDERLINED DENOTES THE PROPOSED METHODS IS 1% BETTER THAN THE BASELINE METHODS.

Data Type	Dataset	tSNE (2014)	UMAP (2018)	IVIS (2019)	PaCMAP (2021)	PUMAP (2023)	DMT-EV (2024)	DMT-HI (Ours)
Text Data	NG20	74.5	74.0	66.7	73.7	73.6	75.6	<b>76.1</b>
Image Data	MNIST	92.2	93.0	86.7	92.5	93.0	<b>93.3</b>	93.0
	K-MNIST	87.5	88.7	84.0	88.4	<b>88.9</b>	88.1	86.7
	E-MNIST	86.7	88.4	78.8	87.8	88.8	<b>89.1</b>	87.2
	Cifar10	85.0	83.2	73.2	-	83.2	87.2	<b>89.2</b>
	Cifar100	86.2	85.3	71.4	-	85.7	88.9	<b>90.1</b>
Biological Data	GAST	59.7	57.6	58.4	58.1	61.8	61.5	<b>62.4</b>
	SAM	95.1	95.3	93.4	95.4	95.3	95.7	<b>95.8</b>
	HCL	73.1	65.4	70.8	66.8	74.3	74.4	<b>74.9</b>
	MCA	78.8	72.5	76.8	-	<b>85.9</b>	84.5	84.9
Statistics	AVERAGE	81.9	80.3	76.0	-	83.1	83.8	<b>84.0</b>
	RANK	4	5	6	-	3	2	1

TABLE III  
LOCAL PRESERVATION PERFORMANCE (KNN ACCURACY) COMPARISON ON TEN DATASETS. **BOLD** DENOTES THE BEST RESULT AND UNDERLINED DENOTES 1% HIGHER THAN OTHERS (NO BOLDED). ‘-’ MEANS THE OFFICIAL IMPLEMENTATION IS NOT AVAILABLE.

	KNN accuracy - training set							KNN accuracy - testing set						
	tSNE (2014)	UMAP (2018)	IVIS (2019)	PaCMAP (2021)	PUMAP (2023)	DMT-EV (2024)	DMT-HI (Ours)	tSNE (2014)	UMAP (2018)	IVIS (2019)	PUMAP (2023)	DMT-EV (2024)	DMT-HI (Ours)	
20News	54.5	49.3	21.4	44.6	46.4	55.9	<b>59.3</b>	45.0	41.3	20.6	36.5	<b>48.1</b>	<b>48.1</b>	
MNIST	95.2	96.3	78.5	96.0	96.5	96.8	<b>97.9</b>	94.8	95.3	78.4	95.0	95.8	<b>97.3</b>	
K-MNIST	93.7	95.5	75.9	94.3	94.6	95.4	<b>96.4</b>	84.2	85.2	59.1	78.7	92.2	<b>94.4</b>	
E-MNIST	70.9	70.3	30.8	68.1	66.2	72.0	<b>75.6</b>	68.1	67.3	30.6	63.5	71.3	<b>72.9</b>	
Cifar10	26.0	21.4	18.5	21.0	-	25.2	<b>75.3</b>	24.7	21.1	18.7	-	23.2	<b>74.3</b>	
Cifar100	9.3	6.2	3.5	5.8	-	7.6	<b>42.1</b>	6.7	5.8	2.8	-	6.2	<b>40.3</b>	
GAST	78.1	65.5	69.9	91.4	68.5	87.6	<b>93.2</b>	71.8	61.9	68.4	64.1	78.5	<b>85.1</b>	
SAM	75.7	74.4	72.6	74.3	75.3	<b>76.4</b>	75.6	74.6	74.5	71.1	74.6	<b>76.1</b>	75.6	
HCL	73.6	46.4	52.7	85.0	52.0	80.2	<b>86.9</b>	68.2	43.2	50.0	46.6	74.3	<b>79.8</b>	
MCA	67.0	47.2	72.9	92.6	-	87.4	<b>94.1</b>	59.9	46.1	71.8	-	83.4	<b>90.9</b>	
AVE	64.4	57.3	49.7	67.3	-	68.5	<b>79.6</b>	59.8	54.2	47.2	-	64.9	<b>75.9</b>	
RANK	4	5	6	3	-	2	1	3	4	5	-	2	1	

MOE system. The Aulgrithm. 1 is the pseudocode for the proposed MOE-based Hyperbolic.

## V. EXPERIMENTS

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed DMT-HI for High dimensional data visualizatoin. In Sec V-A, we provide the details of the datasets and the baselines method. In Sec V-B, we provide the details of our implantation. In Sec V-C, we compare our method with other baselines in ten datasets. Qualitative analysis are conducted in Sec V-D, which directly demonstrate the effect of the method.

### A. Datasets and Baselines

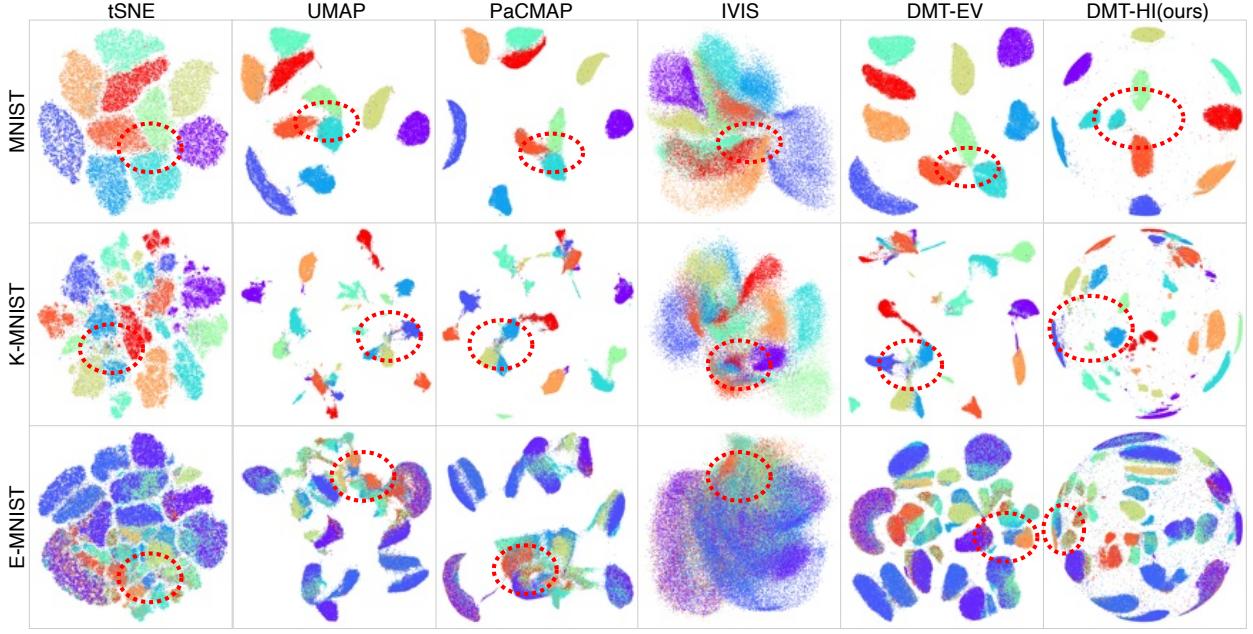
**Datasets.** Our comparative experiments include ten datasets (20News, MNIST, E-MNIST, K-MNIST, Cifar10, Cifar100, MCA, GAST, HCL, and SAM). The detailed descriptions of the datasets are provided in Table A1.

**Baseline Methods.** The methods compared in this study include tSNE [34], [63], UMAP [23], PUMAP [36], [25], Ivis [64], PaCMAP [65], HNNE [66] and DMT-EV [26] (The result of HNNE is in the appendix). Each method maps the input data to a two-dimensional latent space to meet consistent visualization requirements.

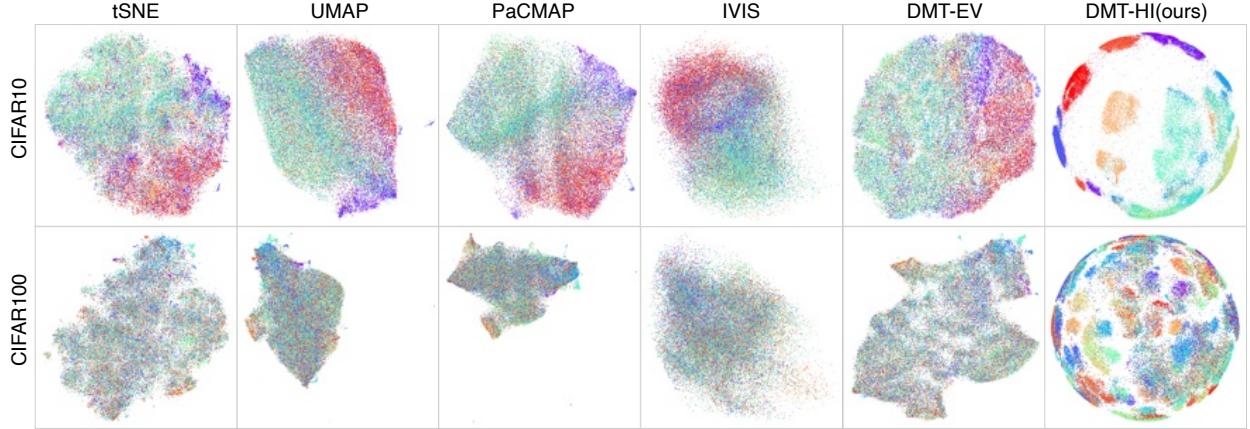
**Evaluation Metrics.** We evaluate the performance of the methods using three metrics: SVM classification accuracy, KNN classification accuracy and trustworthiness (TRUST) score [35]. The SVM classification accuracy is calculated using a linear SVM classifier trained on the latent space representations of the training set and tested on the test set. The Trust score is calculated using the Trust metric, which measures the preservation of the global structure of the data in the latent space. All models are trained on the training set (80%), validated on the validation set (10%), and evaluated on the test set (10%), with generalization performance evaluated using test metrics. The mean and variance of the results over 10 trials are provided in the Appendix.

### B. Implantation Details

We follow the basic implementation details as described in [26] to set up our experiments. In our proposed model, the Multiple Gumbel Matchers are implemented as a 2-layer Multilayer Perceptron (MLP) network, where the hidden layer contains 100 neurons. The O in Eq. (10) is set to  $\lceil 0.9 \times D \rceil$ , where D is the number of the features of the raw data. This structure allows the matchers to effectively learn task allocations for different feature subsets. In the Mixture of Experts (MOE) Network, each expert is designed as a 4-layer



**Fig. 4. Comparison of visualizations generated on MNIST, K-MNIST, and E-MNIST dataset.** The red circles highlight the same areas in each method's visualization to illustrate differences in clustering and separation of data points. These visualizations show how various methods capture local and global structures, with DMT-HI offering enhanced separation and clustering compared to other methods, as shown in the circled regions.



**Fig. 5. Comparison of visualizations generated on CIFAR-10 and CIFAR-100 dataset.** While all methods show some capability in preserving data structure, our proposed method (DMT-HI) demonstrates a clear advantage by producing more distinct and well-separated clusters, especially on the more challenging CIFAR-100 dataset. These results highlight the effectiveness of DMT-HI in handling high-dimensional and complex image data, providing superior cluster separation and structural preservation compared to other methods.

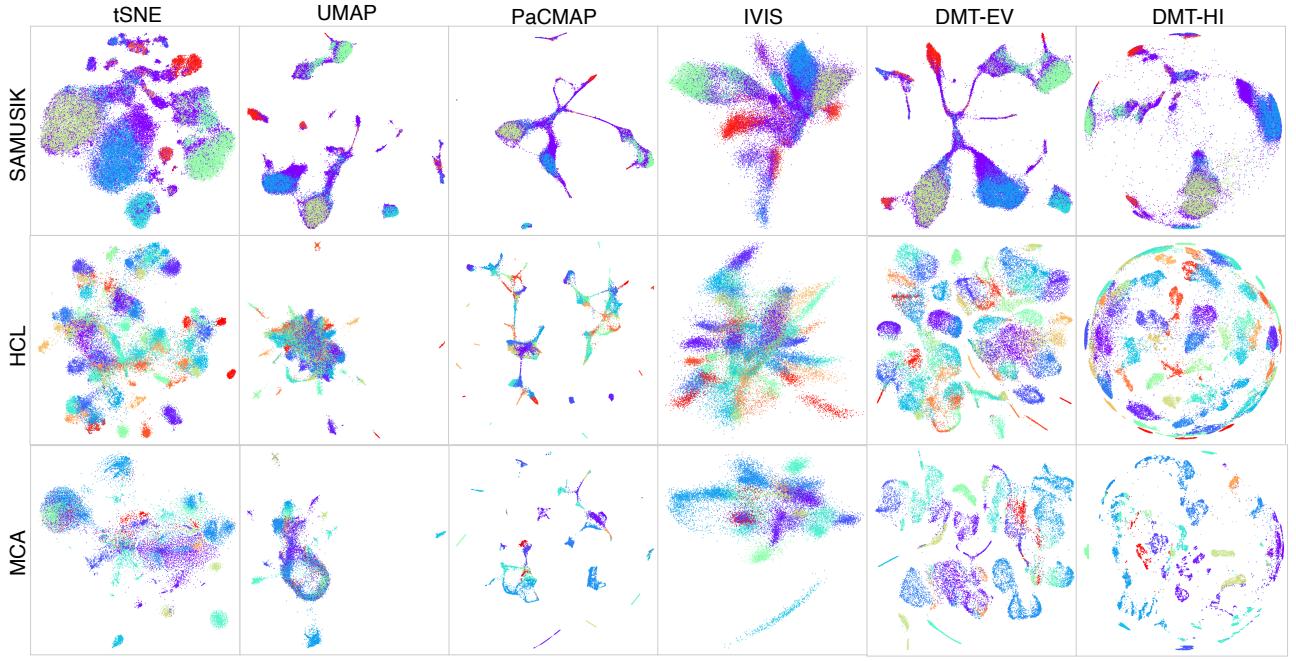
MLP, with each hidden layer comprising 512 neurons. The MOE network includes a total of 10 expert models, ensuring that each expert can specialize in processing distinct subsets of the input data, providing the necessary capacity and flexibility for complex data representations. The  $\nu$  latent space parameter  $\nu_{\text{lat}}$ , the embedding space parameter  $\nu_{\text{emb}}$ , the exaggeration parameter  $\gamma$ , the batch size (**batch\_size**), and the number of neighbors in augmentation (**K**) are provided in different datasets. The detailed optimal parameters for each dataset are summarized in Table A2.

### C. Global Structure Preservation

To evaluate the accuracy and effectiveness of DMT-HI, we use two key metrics: classification accuracy and the Trust

metric. These are standard for assessing the preservation of global and local structures in dimensionality reduction (DR).

**Results on linear SVM metric & global performance.** Classification accuracy is assessed using a linear Support Vector Machine (SVM) on the reduced data to evaluate the model's ability to preserve inter-cluster relationships from the original high-dimensional space. By using a simple linear classifier, we ensure the focus remains on structural preservation without the added complexity of a more sophisticated classifier. As shown in Table I, DMT-HI consistently achieves superior classification accuracy across various datasets, particularly excelling on complex image datasets like Cifar10 and Cifar100. For instance, on Cifar10, DMT-HI outperforms all other methods with an impressive training accuracy of



**Fig. 6. Comparison of visualizations generated on SAMUSIK, HCL, and MCA dataset.** Our method (DMT-HI) shows a clear advantage in preserving both local and global structures, achieving better separation and more distinct clusters across all datasets. In contrast to other methods, DMT-HI provides more interpretable and biologically relevant clusters, particularly on the highly complex HCL and MCA datasets. This demonstrates the superiority of DMT-HI in handling diverse and complex biological data, offering better representation of the underlying biological structures.

77.5% and a testing accuracy of 74.9%, compared to the much lower accuracy of traditional methods like t-SNE (22.3%) and UMAP (21.8%). Similarly, on the Cifar100 dataset, DMT-HI reaches a remarkable training accuracy of 39.1% and a testing accuracy of 38.9%, significantly higher than baseline methods like DMT-EV (5.2%) and UMAP (5.3%). The model also demonstrates strong performance on biological datasets, such as MCA and HCL. On MCA, DMT-HI achieves a training accuracy of 80.7%, outperforming other methods like UMAP (37.8%) and PaCMAP (76.2%). Similarly, on HCL, DMT-HI stands out with a training accuracy of 84.5% and a testing accuracy of 76.7%, while methods such as t-SNE and IVIS lag behind significantly. Moreover, DMT-HI exhibits robust performance across all datasets, maintaining the highest average classification accuracy (71.9% for training, 69.8% for testing), as compared to other methods like DMT-EV (61.2%) and PaCMAP (59.6%). This strong performance is reflected in DMT-HI’s overall ranking, where it consistently ranks first across both training and testing sets.

**Results on trustworthiness metric & global performance.** we evaluate the trustworthiness (TRUST) metric [35] to compare the local structure preservation capabilities of DMT-HI against other dimensionality reduction methods. Traditional methods like t-SNE and UMAP, which directly optimize the trustworthiness metric, have a natural advantage in preserving local neighborhood relationships, particularly when processing data in smaller batches. As shown in Table II, these methods perform slightly better in certain datasets, such as K-MNIST and MCA, where they achieve higher local structure preservation. However, DMT-HI demonstrates consistently competitive performance across a wide range of datasets, maintaining a

strong balance between global and local structure preservation. Despite not being specifically optimized for the trustworthiness metric, DMT-HI performs well on complex datasets such as SAM and Cifar100, surpassing traditional methods in average performance. This balance is reflected in DMT-HI’s ability to achieve the best or nearly best trustworthiness scores across most datasets (see Table II). Overall, while traditional methods may show slight advantages in optimizing for the trustworthiness metric, DMT-HI’s robust performance across all datasets, particularly in terms of average rankings, highlights its versatility and suitability for a wide range of high-dimensional data applications.

**More advantages on larger datasets.** DMT-HI consistently outperforms baseline methods, especially on large-scale datasets like MNIST and E-MNIST, where its higher classification accuracy and Trust scores demonstrate its robust performance. This is due to DMT-HI’s neural network model, which is well-suited for high-complexity data, and its hyperbolic embedding, which captures complex nonlinear features. On complex datasets like SAM and HCL, DMT-HI excels, preserving intricate structures that traditional methods often struggle with.

#### D. Local Structure Preservation

To evaluate local structure preservation in dimensionality reduction (DR), we used KNN clustering accuracy to assess how well neighborhood relationships were maintained in the reduced space. This metric is key for evaluating clustering and classification tasks as it ensures that close points in the original high-dimensional space remain close after reduction.

**Results on KNN metric & local performance.** As shown in Table III, DMT-HI consistently outperformed methods like

t-SNE, UMAP, PaCMAP, and Ivis, achieving an average KNN accuracy of 79.6% on the training set and 75.9% on the test set. Its superior performance on datasets like Cifar10 and Cifar100 demonstrates its ability to retain intricate local structures, where simpler methods often struggle. On biological datasets such as MCA and SAM, DMT-HI's hyperbolic embeddings excelled at capturing hierarchical and nonlinear structures, achieving 94.1% KNN accuracy on MCA, far surpassing DMT-EV. This adaptability across domains highlights DMT-HI's versatility.

**Strong generalization on complex image data**, maintaining 74.3% accuracy on the Cifar10 test set. In contrast, traditional methods like t-SNE displayed more variability, underscoring their limitations on complex datasets. For instance, on Cifar100, DMT-HI achieved 42.1% accuracy, while t-SNE managed only 9.3%, illustrating DMT-HI's ability to capture complex hierarchical structures.

#### E. Visualization Results

In this section, we provide a detailed analysis of the visualization results across various dimensionality reduction methods on multiple datasets, including both training and test sets. These results demonstrate the superior performance of our proposed DMT-HI method in representing complex data structures across different domains, such as image and biological datasets. A more extensive set of results, including additional datasets and methods, can be found in the Appendix, where comprehensive visualizations for both training and test sets are presented across multiple approaches. The detailed visualization results are shown in Appendix (Fig. A1, Fig. A2, Fig. A3, and Fig. A4).

**Clear boundaries and less manifolds overlapping.** In terms of performance metrics, we have previously demonstrated that DMT-HI achieves significant improvements over traditional dimensionality reduction techniques such as t-SNE, UMAP, PaCMAP, and IVIS. Fig. 4 further support these findings by showing that DMT-HI not only preserves the global data structure in both training and test sets but also offers superior separation between classes, especially in more complex datasets such as K-MNIST and E-MNIST. The red-circled regions highlight specific areas where DMT-HI excels in clustering, providing more granularity and reduced class overlap, which is crucial for tasks requiring high-resolution data representation.

**Clear advantage on complex image datasets (CIFAR-10 and CIFAR-100).** As shown in Fig. 5, while all methods exhibit some level of consistency between training and test visualizations, DMT-HI excels in generating well-separated clusters with minimal overlap, even in the highly complex CIFAR-100 dataset, where subclasses exhibit significant similarity. The deep neural architecture of DMT-HI allows it to capture and represent rich semantic information, resulting in enhanced class distinction and a more interpretable low-dimensional space. In more complex image datasets, such as CIFAR-100, DMT-HI shows substantial performance gains compared to other methods, particularly in terms of semantic representation. The architecture of DMT-HI enables it to

TABLE IV  
ABLATION STUDY: SVC TRAINING PERFORMANCE ON DIFFERENT VALUES OF K. THE BEST RESULTS ARE IN BOLD.

$\nu$	MNIST	E-MNIST	HCL	20News	AVERAGE
0.001	96.7	61.0	75.4	<b>42.0</b>	68.8
0.005	97.3	62.9	79.0	40.0	69.8
0.01	97.3	63.3	80.7	41.3	70.7
0.05	<b>98.0</b>	66.2	80.6	41.1	71.4
0.1	97.9	65.4	82.4	41.1	71.7
0.5	97.7	69.2	<b>84.0</b>	40.6	<b>72.9</b>
1	97.3	<b>69.3</b>	79.9	40.0	71.6

TABLE V  
ABLATION STUDY: SVC TRAINING PERFORMANCE ON DIFFERENT VALUES OF K. THE BEST RESULTS ARE IN BOLD.

K	MNIST	E-MNIST	HCL	20News	AVERAGE
2	97.8	67.4	82.6	41.2	72.2
3	97.9	68.0	83.6	40.8	72.5
4	97.9	67.0	83.7	40.8	72.3
5	97.9	<b>68.8</b>	83.4	<b>42.8</b>	<b>73.2</b>
6	<b>98.1</b>	67.7	<b>83.8</b>	36.1	71.4
7	98.0	68.2	82.7	39.8	72.1
8	97.9	68.7	82.9	37.6	71.7
9	98.0	68.0	82.0	38.5	71.6

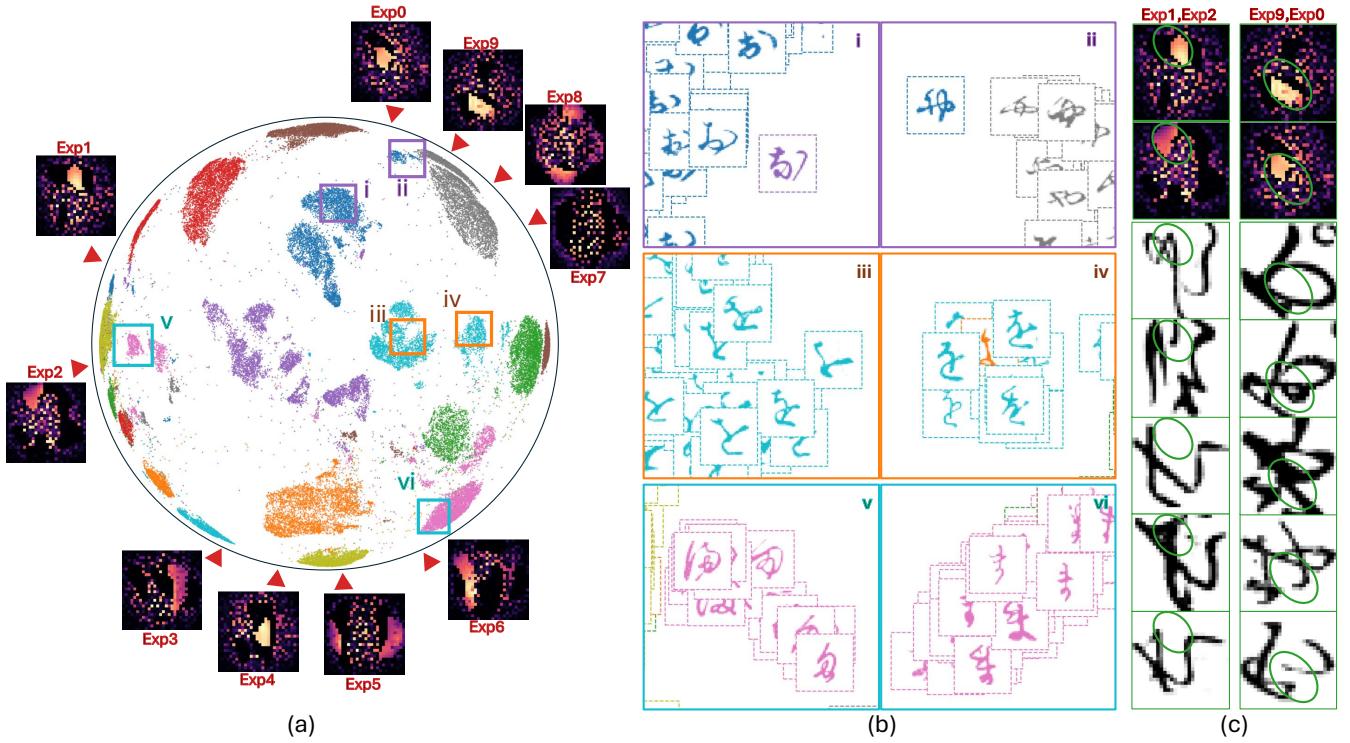
better capture the nonlinear relationships between subclasses, resulting in superior cluster separation in both training and test sets. This is evident in Fig. 5, where DMT-HI provides more distinct class boundaries and a clearer representation of the inherent hierarchical structure of the data, demonstrating its superiority.

**Better hierarchical relationships on biological datasets.** On biological datasets (SAMUSIK, HCL, and MCA), as shown in Fig. 6, DMT-HI demonstrates its ability to effectively capture and visualize complex hierarchical relationships. Compared to other methods, DMT-HI provides a more coherent representation of the biological data structure, preserving both local and global relationships. This is particularly important in biological data analysis, where understanding the hierarchical nature of the data is crucial. DMT-HI excels in this regard by providing a more interpretable and biologically relevant low-dimensional representation, as evidenced by its performance across the SAMUSIK, HCL, and MCA datasets.

**Stability in training and testing datasets.** To ensure the comprehensiveness of our analysis, we have extended our comparison to include a wider range of datasets and methods, considering both the training and test phases. These additional visualizations, which can be found in the Appendix, further reinforce the advantages of DMT-HI across various datasets and scenarios. By offering more detailed comparisons across both image and biological datasets, we demonstrate that DMT-HI consistently outperforms traditional methods in terms of cluster separation, hierarchical representation, and overall interpretability.

#### F. Time Consumption

In this experiment, we evaluate the time consumption of different dimensionality reduction methods on various datasets,



**Fig. 7. Visualization of the DMT-HI model’s performance on the K-MNIST dataset.** (a) The hyperbolic embedding of K-MNIST shows distinct clusters representing different Kanji characters, with clear boundaries, highlighting the model’s ability to handle complex structures. (b) Zoomed-in views of selected regions (i-vi) reveal the model’s capacity to separate characters with the same label but different structures, while regions v and vi show how mislabeled points are correctly embedded closer to their true categories. Sub-clusters indicate the model’s precision in identifying variations within a single category. (c) Expert representations and associated images showcase the model’s interpretability and specialization in character recognition.

TABLE VI

TIME CONSUMPTION PERFORMANCE COMPARISON ON SIX DATASETS,  
**BOLD** DENOTES THE BEST RESULT. (HH:MM:SS)

	tSNE	UMAP	PaCMAP	DMT-EV	DMT-HI
Mnist	00:14:02	00:01:09	<b>00:00:52</b>	00:01:09	00:01:19
K-Mnist	00:15:12	00:01:12	<b>00:00:56</b>	00:01:13	00:01:02
E-Mnist	00:34:22	00:21:02	00:18:56	00:17:13	<b>00:06:02</b>
Gask	00:03:06	<b>00:01:54</b>	00:02:20	<b>00:01:54</b>	00:01:57
MCA	00:09:28	00:08:51	<b>00:06:11</b>	00:06:20	00:07:03
HCL	00:13:08	00:12:28	00:10:48	00:10:34	<b>00:05:34</b>

TABLE VII

ABALATION STUDY: SVC TRAINING PERFORMANCE ON DIFFERENT VALUES OF NUMBER OF MOE. THE BEST RESULTS ARE IN **BOLD**.

MOE	MNIST	E-MNIST	HCL	20News	AVERAGE
1	97.9	66.6	<b>85.2</b>	39.9	72.4
2	<b>98.0</b>	65.9	83.1	40.6	71.9
4	<b>98.0</b>	68.8	84.4	40.6	73.0
6	97.9	67.5	82.2	<b>44.2</b>	73.0
8	<b>98.0</b>	<b>69.0</b>	84.0	39.7	72.7
10	97.9	68.8	83.4	42.8	<b>73.2</b>
12	97.9	65.9	84.4	43.8	73.0
16	<b>98.0</b>	68.9	84.0	42.0	<b>73.2</b>

including image datasets (MNIST, K-MNIST, E-MNIST) and biological datasets (Gask, MCA, HCL). These datasets vary in size and complexity, allowing a thorough comparison of performance across tasks.

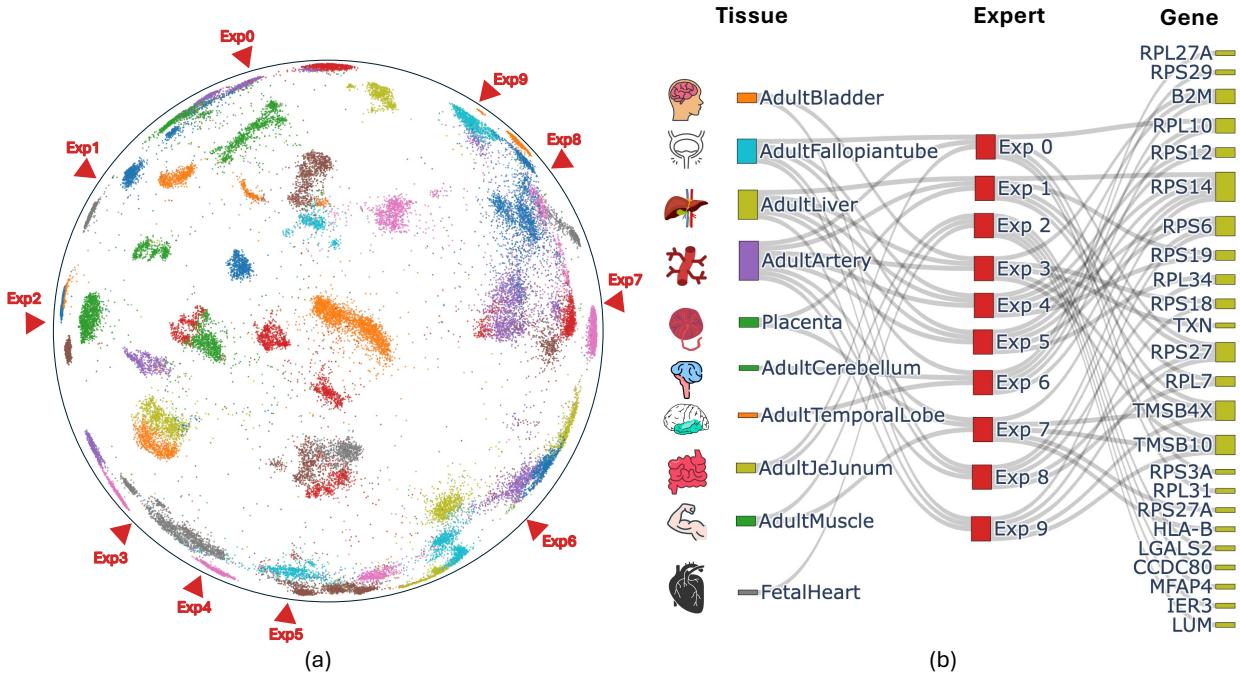
We selected popular non-parametric methods (tSNE,

UMAP, PaCMAP) and two deep learning methods (such as DMT-HI), adjusting parameters according to official guidelines. The experiments were conducted on a GPU-accelerated platform, though non-parametric methods typically do not fully utilize GPU, affecting their time performance.

Table VI shows the time consumption for each method. Non-parametric methods like PaCMAP perform well on smaller datasets such as MNIST and K-MNIST, completing tasks quickly. However, on larger datasets like E-MNIST and HCL, deep learning models (DMT-EV and DMT-HI) demonstrate a significant time advantage, benefiting from GPU parallelization. While non-parametric methods are efficient for smaller datasets, their time consumption rises sharply with data size and complexity.

## VI. ABLATION STUDY

In this section, we present ablation studies to analyze the impact of key parameters on the performance and stability of the DMT-HI model. We focus on three key parameters, the number of neighbors  $K$  in Eq. (III-A), the number of experts in the Mixture of Experts (MOE) network, and the hyperbolic parameter  $\nu$ . The results are summarized in Tables V, VII, and IV. We explore the influence of the number of neighbors  $K$ , which governs local structure preservation in the hyperbolic embedding. As shown in Table V,  $K = 5$  provides a good balance between capturing local and global structures. Smaller  $K$  values capture insufficient local information, while larger



**Fig. 8. Visualization of the DMT-HI model’s performance on the HCL (Human Cell Landscape) dataset.** (a) The hyperbolic embedding shows distinct clusters for different tissue types, demonstrating the model’s ability to differentiate and reveal relationships among tissues. The well-separated clusters indicate DMT-HI’s effectiveness in handling complex biological data, with triangle markers (Exp0 to Exp9) representing experts specializing in specific regions. (b) A bipartite network connects tissues, experts, and genes, illustrating the model’s interpretability by mapping tissue-specific experts to their associated genes, offering insights into gene-tissue interactions and expert specializations.

$K$  values introduce noise. Thus,  $K = 5$  is set as the default value, ensuring stable performance across datasets.

We assess the effect of the number of experts in the MOE network, which improves the model’s ability to handle diverse feature subsets. Table VII shows that increasing the number of experts enhances performance by capturing complex patterns. However, beyond 10 or 16 experts, diminishing returns and overfitting risks emerge. Therefore, 10 or 16 experts strike an optimal balance between performance and computational cost. We analyze the hyperbolic parameter  $\nu$ , which affects the curvature of the hyperbolic space and the model’s ability to capture hierarchical structures. Table IV indicates that  $\nu = 0.5$  balances capturing hierarchical relationships and maintaining stability. Lower  $\nu$  values capture stronger hierarchies, while higher values flatten the space, reducing its effectiveness. Thus,  $\nu = 0.5$  is chosen as the default value for stability across diverse datasets.

## VII. CASE STUDY & INTERPRETABILITY

### A. Performance & Interpretability on Image Datasets

To further validate the DMT-HI model, we conducted a case study on the K-MNIST dataset, which contains complex handwritten Kanji characters. Compared to MNIST, K-MNIST presents more intricate structures, making it a challenging benchmark for evaluating a model’s ability to handle high-dimensional data. We employed standard configurations to ensure fairness and reproducibility, using hyperbolic embedding for data analysis.

**Performance advantages on K-MNIST dataset & the ability to discover subclusters.** Despite the complexity of the dataset, the hyperbolic representations are well-organized, forming distinct character clusters in hyperbolic space (Fig. 7(a)). The sharp boundaries between clusters confirm the model’s ability to preserve data structure even after high-dimensional mapping. This highlights DMT-HI’s effectiveness in managing complex data. The model also excels in distinguishing characters with the same label but different structures. As seen in Fig. 7(b) (regions i, ii, iii, and iv), DMT-HI accurately separates characters that vary in strokes and shapes, reflecting its sensitivity to fine-grained features. This precise feature differentiation supports tasks like character recognition, laying a foundation for classification and pattern recognition. DMT-HI shows robustness in handling mislabeled data. Fig. 7(b) (regions v and vi) illustrates how the model embeds mislabeled points closer to their true categories, reducing the impact of label noise. This suggests that DMT-HI can identify and adjust anomalous points, making it useful for data cleaning in practical applications. The model can further subdivide clusters into meaningful sub-clusters (Fig. 7(b) regions v and vi), showcasing its ability to capture internal hierarchy and diversity in the data. This is crucial for analyzing datasets with complex structures, such as single-cell sequencing, enhancing DMT-HI’s applicability across domains.

**Interpretability of the key pixel patterns.** The expert analysis in Fig. 7(c) demonstrates the model’s interpretability. Each expert focuses on different character attributes, provid-

ing insight into how DMT-HI decomposes and reconstructs character features. This division of expertise enhances interpretability and shows how the model adapts to recognize different features optimally.

### B. Interpretability on Biological Datasets

To validate the effectiveness and interpretability of the DMT-HI model, we conducted a case study using the HCL (Human Cell Landscape) dataset [67], which contains a wide range of tissue types, offering a complex test for high-dimensional biological data analysis. This study focuses on the model's ability to generate clear representations and reveal meaningful connections between tissue clusters and key gene markers [68], [69].

**DMT-HI bridge between tissue and key genes.** The results highlight several strengths of the DMT-HI model. DMT-HI effectively generates distinct and well-separated clusters, as shown in Fig. 8(a), where different tissues are clearly represented in hyperbolic space with sharp boundaries. This clear separation supports tissue classification and provides deeper insights into the intrinsic structure of the data. DMT-HI excels in interpretability by linking tissue clusters to key genes. Fig. 8(b) illustrates how the model connects tissue types (left column) to specific experts (middle column, Exp0 to Exp9) and their associated genes (right column). This network shows how experts specialize in identifying tissue-specific characteristics and genes, providing a clear view of the model's decision-making process.

**Potential of Unsupervised target discovery .** DMT-HI demonstrates the ability to perform unsupervised biomarker selection. By automatically linking tissue clusters with important genes, the model identifies critical biological markers without needing labeled data. Fig. 8(b) shows how each expert aligns with tissue-specific genes, reflecting the model's capacity to uncover key features, which is crucial in biological research for discovering biomarkers.

## VIII. CONCLUSION

In this paper, we introduced the MOE-based Hyperbolic Interpretable Deep Manifold Transformation (DMT-HI), a novel approach to dimensionality reduction that addresses the limitations of traditional methods by combining hyperbolic embeddings with a Mixture of Experts (MOE) framework. DMT-HI enhances both performance and interpretability, preserving complex data structures while offering dynamic task allocation to efficiently process diverse datasets. Our experiments on image and biological datasets demonstrated DMT-HI's superior ability to retain local and global features, outperforming baseline methods. The model also provides clearer insights into data representation through expert-driven decision tracking and advanced visualization tools.

## ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (No.2022ZD0115100), the National Natural Science Foundation of China Project (No. U21A20427), and Project (No. WU2022A009) from the Center of Synthetic

Biology and Integrated Bioengineering of Westlake University. This work was supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2024C01140). This work was supported by Key Research and Development Program of Hangzhou (2023SZD0073). We thank the Westlake University HPC Center for providing computational resources. This work was supported by InnoHK program. This work was supported by Ant Group through CAAI-Ant Research Fund.

## REFERENCES

- [1] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, “Tensor canonical correlation analysis for multi-view dimension reduction,” *IEEE transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, 2015.
- [2] W. Wei, Q. Yue, K. Feng, J. Cui, and J. Liang, “Unsupervised dimensionality reduction based on fusing multiple clustering results,” *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [3] X.-R. Yu, D.-B. Wang, and M.-L. Zhang, “Dimensionality reduction for partial label learning: A unified and adaptive approach,” *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [4] D. Bera, R. Pratap, and B. D. Verma, “Dimensionality reduction for categorical data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3658–3671, 2021.
- [5] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [6] A. Gisbrecht and B. Hammer, “Data visualization by nonlinear dimensionality reduction,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 2, pp. 51–73, 2015.
- [7] S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Information Fusion*, vol. 59, pp. 44–58, 2020.
- [8] P. Ray, S. S. Reddy, and T. Banerjee, “Various dimension reduction techniques for high dimensional data analysis: a review,” *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3473–3515, 2021.
- [9] H. Hojjati and N. Armanfard, “Dasvdd: Deep autoencoding support vector data descriptor for anomaly detection,” *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [10] M. F. Kabir, T. Chen, and S. A. Ludwig, “A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction,” *Healthcare Analytics*, vol. 3, p. 100125, 2023.
- [11] Y. Zhang, P. Tiño, A. Leonardis, and K. Tang, “A survey on neural network interpretability,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.
- [12] F. Imrie, R. Davis, and M. van der Schaar, “Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare,” *Nature Machine Intelligence*, vol. 5, no. 8, pp. 824–829, 2023.
- [13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [14] J. P. Bharadiya, “A tutorial on principal component analysis for dimensionality reduction in machine learning,” *International Journal of Innovative Science and Research Technology*, vol. 8, no. 5, pp. 2028–2032, 2023.
- [15] S. Piaggesi, M. Khosla, A. Panisson, and A. Anand, “Dine: Dimensional interpretability of node embeddings,” *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [16] Z. Zang, S. Li, D. Wu, G. Wang, K. Wang, L. Shang, B. Sun, H. Li, and S. Z. Li, “Dlme: Deep local-flatness manifold embedding,” in *European Conference on Computer Vision*. Springer, 2022, pp. 576–592.
- [17] Z. Zang, Y. Xu, L. Lu, Y. Geng, S. Yang, and S. Z. Li, “Udrn: unified dimensional reduction neural network for feature selection and feature projection,” *Neural Networks*, vol. 161, pp. 626–637, 2023.
- [18] L. Hajderanj, D. Chen, S. Dudley, G. Gillopte, and B. Sivy, “Novel parameter-free and parametric same degree distribution-based dimensionality reduction algorithms for trustworthy data structure preserving,” *Information Sciences*, vol. 661, p. 120030, 2024.
- [19] J. You, S. W. Jeong, and C. Donnat, “Gnumap: A parameter-free approach to unsupervised dimensionality reduction via graph neural networks,” *arXiv preprint arXiv:2407.21236*, 2024.
- [20] J. Yi, H. Duan, J. Wang, Z. Yang, and F. Nie, “Structure preserved fast dimensionality reduction,” *Applied Soft Computing*, p. 111817, 2024.
- [21] F. Luo, L. Zhang, B. Du, and L. Zhang, “Dimensionality reduction with enhanced hybrid-graph discriminant learning for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5336–5353, 2020.

- [22] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [23] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.
- [24] T. Sainburg, L. McInnes, and T. Q. Gentner, "Parametric umap embeddings for representation and semisupervised learning," *Neural Computation*, vol. 33, no. 11, pp. 2881–2907, 2021.
- [25] B. Xu and G. Zhang, "Robust parametric umap for the analysis of single-cell data," *bioRxiv*, pp. 2023–11, 2023.
- [26] Z. Zang, S. Cheng, H. Xia, L. Li, Y. Sun, Y. Xu, L. Shang, B. Sun, and S. Z. Li, "Dmt-ev: An explainable deep network for dimension reduction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 3, pp. 1710–1727, 2024.
- [27] R. Marcinkevičius and J. E. Vogt, "Interpretable and explainable machine learning: methods-centric overview with concrete examples," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 3, p. e1493, 2023.
- [28] B. D. Drnovsek and F. Forstneric, "Hyperbolic domains in real euclidean spaces," *arXiv preprint arXiv:2109.06943*, 2021.
- [29] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon *et al.*, "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.
- [30] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts," *arXiv preprint arXiv:2407.06204*, 2024.
- [31] J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.
- [32] J. B. Tenenbaum, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
- [33] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000, publisher: American Association for the Advancement of Science.
- [34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [35] M. Moor, M. Horn, B. Rieck, and K. Borgwardt, "Topological autoencoders," in *International conference on machine learning*. PMLR, 2020, pp. 7045–7054.
- [36] T. Sainburg, L. McInnes, and T. Q. Gentner, "Parametric UMAP embeddings for representation and semi-supervised learning," *arXiv:2009.12981 [cs, q-bio, stat]*, Apr. 2021, arXiv: 2009.12981.
- [37] C. Maitra, D. B. Seal, and R. K. De, "NeuroDAVIS: A neural network model for data visualization," *Neurocomputing*, vol. 573, p. 127182, Mar. 2024.
- [38] D. Rajwade, A. Ahmadi, and B. P. Ingalls, "Cells2Vec: Bridging the gap between experiments and simulations using causal representation learning," Oct. 2023.
- [39] R. Faust, D. Glickenstein, and C. Scheidegger, "Dimreader: Axis lines that explain non-linear projections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 481–490, 2019.
- [40] S. Cheng and K. Mueller, "The data context map: Fusing data and attributes into a unified display," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 121–130, 2016.
- [41] G. Stiglic, P. Kocabek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1379, 2020.
- [42] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," 2021, pp. 782–791.
- [43] A. Bibal, V. Delchevalerie, and B. Frénay, "DT-SNE: t-SNE discrete visualizations as decision tree structures," *Neurocomputing*, vol. 529, pp. 101–112, Apr. 2023.
- [44] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," *Advances in neural information processing systems*, vol. 30, pp. 6338–6347, 2017.
- [45] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, "Hyperbolic image embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6418–6428.
- [46] T. Tian, C. Zhong, X. Lin, Z. Wei, and H. Hakonarson, "Complex hierarchical structures in single-cell genomics data unveiled by deep hyperbolic manifold learning," *Genome Research*, vol. 33, no. 2, pp. 232–246, Feb. 2023.
- [47] R. Jankowski, A. Allard, M. Boguna, and M. A. Serrano, "The DMercator method for the multidimensional hyperbolic embedding of real networks," *Nature Communications*, vol. 14, no. 1, p. 7585, Nov. 2023.
- [48] M. Yang, M. Zhou, R. Ying, Y. Chen, and I. King, "Hyperbolic Representation Learning: Revisiting and Advancing," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 39639–39659, iSSN: 2640-3498.
- [49] G. Mishne, Z. Wan, Y. Wang, and S. Yang, "The Numerical Stability of Hyperbolic Representation Learning," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 24925–24949, iSSN: 2640-3498.
- [50] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [51] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13001–13008.
- [52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
- [54] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6688–6697.
- [55] B. Cheng, W. Wu, D. Tao, S. Mei, T. Mao, and J. Cheng, "Random cropping ensemble neural network for image classification in a robotic arm grasping system," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6795–6806, 2020.
- [56] J. Flusser, S. Farokhi, C. Höschl, T. Suk, B. Zitova, and M. Pedone, "Recognition of images degraded by gaussian blur," *IEEE transactions on Image Processing*, vol. 25, no. 2, pp. 790–806, 2015.
- [57] Z. Liu, S. Li, Z. C. Di Wu, L. Wu, J. Guo, and S. Z. Li, "Automix: Unveiling the power of mixup," 2021.
- [58] Z. Zang, H. Luo, K. Wang, P. Zhang, F. Wang, S. Z. Li, and Y. You, "Diffaug: Enhance unsupervised contrastive learning with domain-knowledge-free diffusion-based data augmentation," in *Forty-first International Conference on Machine Learning*.
- [59] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.
- [60] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.
- [61] A. Saa, E. Miranda, and F. Rouxinol, "Higher-dimensional euclidean and non-euclidean structures in planar circuit quantum electrodynamics," *arXiv preprint arXiv:2108.08854*, 2021.
- [62] S. Buchholz and G. Sommer, "A hyperbolic multilayer perceptron," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 2. IEEE, 2000, pp. 129–133.
- [63] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014, publisher: JMLR.org.
- [64] B. Szubert, J. E. Cole, C. Monaco, and I. Drozdov, "Structure-preserving visualisation of high dimensional single-cell datasets," *Scientific Reports*, vol. 9, no. 1, p. 8914, Jun. 2019.
- [65] Y. Wang, H. Huang, C. Rudin, and Y. 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*, vol. 22, no. 201, pp. 1–73, 2021.
- [66] M. S. Sarfraz, M. Koulakis, C. Seibold, and R. Stiefelhagen, "Hierarchical Nearest Neighbor Graph Embedding for Efficient Dimensionality Reduction," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 336–345.
- [67] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge *et al.*, "Construction of a human cell landscape at single-cell level," *Nature*, vol. 581, no. 7808, pp. 303–309, 2020.
- [68] Z. Zang, Y. Xu, C. Duan, J. Wu, S. Z. Li, and Z. Lei, "A review of artificial intelligence based biological-tree construction: Priorities, methods, applications and trends," *arXiv preprint arXiv:2410.04815*, 2024.
- [69] Z. Liu, J. Li, S. Li, Z. Zang, C. Tan, Y. Huang, Y. Bai, and S. Z. Li, "Genbench: A benchmarking suite for systematic evaluation of genomic foundation models," *arXiv preprint arXiv:2406.01627*, 2024.

## APPENDIX A THE DETAILS OF DATASETS

**Datasets.** Our comparative experiments include ten datasets (20News, MNIST, E-MNIST, K-MNIST, Cifar10, Cifar100, MCA, GAST, HCL, and SAM). The detailed descriptions of the datasets are provided in Table A1.

TABLE A1  
DATASETS INFORMATION OF DATASETS USED IN THIS PAPER.

	Dataset	Training	Validation	Testing	Dimension
Text Data	20News	15,076	1,884	1,884	100
Image Data	MNIST	48,000	6,000	6,000	$28 \times 28 \times 1$
	K-MNIST	48,000	6,000	6,000	$28 \times 28 \times 1$
	E-MNIST	651,404	81,410	81,410	$28 \times 28 \times 1$
	Cifar10	48,000	6,000	6,000	$32 \times 32 \times 3$
	Cifar100	48,000	6,000	6,000	$32 \times 32 \times 3$
Biological Data	Gast	10,638	1,064	1,064	1,457
	MCA	24,000	3,000	3,000	34,947
	SAMUSIK	69,491	8,686	8,686	38
	HCL	224,000	28,000	28,000	27,341

## APPENDIX B THE DETAILS OF RELATED WORKS

## APPENDIX C THE DETAILS OF BEST PARAMETERS FOR DMT-HI

Specifically, all methods utilize a grid search strategy to determine the optimal parameters from 20 candidate configurations. For t-SNE, we set the perplexity to values from {20, 25, 30, 35} and early exaggeration to {8, 10, 12, 14, 16}. For UMAP, the number of neighbors (`n_neighbors`) is selected from {10, 15, 20, 25}, and the minimum distance (`min_dist`) from {0.01, 0.05, 0.08, 0.1, 0.15}. Similarly, for PaCMAP, `n_neighbors` is chosen from {10, 15, 20, 25} and `min_dist` from {0.01, 0.05, 0.08, 0.1, 0.15}. For Ivis, the number of neighbors (`k`) is selected from {130, 140, 150, 160} and the number of trees (`ntrees`) from {40, 45, 50, 55, 60}. For DMT-EV, the hyperparameter  $\nu^Z$  is chosen from {0.001, 0.005, 0.01, 0.1} and the number of neighbors (`knn`) from {3, 5, 8, 10, 15}. Finally, for our proposed DMT-HI, we set  $\nu$  to values from {0.02, 0.04, 0.08, 0.16, 0.32} and  $\gamma$  from {1.5, 2, 3, 4}.

To optimize the performance of DMT-HI across various datasets, we conducted a hyperparameter search to identify the optimal values for key parameters, including the hyperbolic latent space parameter  $\nu$ , the exaggeration parameter  $\gamma$ , the batch size, and the number of neighbors used in the augmentation process (denoted as  $K$ ). These parameters significantly influence the model's ability to capture and preserve both global and local data structures during dimensionality reduction. Table A2 provides the optimal parameter settings for a range of datasets, including image datasets like Coil20, MNIST, and K-MNIST, as well as biological and text datasets such as GAST, SAM, and NG20. These carefully tuned parameters ensure that DMT-HI delivers consistent and robust performance across diverse types of data.

TABLE A2

OPTIMAL PARAMETERS FOR EACH DATASET. THE HYPERBOLIC LATENT SPACE PARAMETER  $\nu_{\text{LAT}}$ , THE EMBEDDING SPACE PARAMETER  $\nu_{\text{EMB}}$ , THE EXAGGERATION PARAMETER  $\gamma$ , THE BATCH SIZE (`BATCH_SIZE`), AND THE NUMBER OF NEIGHBORS IN AUGMENTATION ( $K$ ) ARE PROVIDED.

Dataset	$\nu$	$\gamma$	<code>batch_size</code>	$K$
Coil20	0.02	3	500	3
Coil100	0.04	4	500	3
MNIST	0.08	1.5	1000	5
K-MNIST	0.08	4	1000	5
E-MNIST	0.32	3	1000	5
ACT	0.02	1.5	1000	5
GAST	0.02	3	1000	5
SAM	0.04	1.5	1000	5
HCL	0.32	4	1000	5
MCA	0.08	2	300	3
NG20	0.08	3	1000	5

## APPENDIX D THE DETAILS OF RESULTS ON 10 INDIVIDUAL EXPERIMENTS

In this section, we present a detailed comparison of the global structure preservation performance across various datasets. The performance is evaluated using SVM classification accuracy on both the training and testing sets. The results are summarized in the following tables, where **bold** values indicate that the proposed method (DMT-HI) exceeds all baseline methods. A dash ('-') indicates that the official implementation is not available.

Tables A3 and A4 present the results of the global structure preservation performance, measured using SVM classification accuracy on thirteen datasets, for both the training and testing sets. The results provide a detailed comparison of the proposed DMT-HI model against several baseline dimensionality reduction methods, including tSNE, UMAP, PUMAP, Ivis, PaCMAP, HNNE, and DMT-EV. Each method's performance is evaluated across a diverse set of datasets, including image datasets like MNIST and CIFAR, as well as biological datasets like ACT and MCA. The mean accuracy, along with standard deviation, is reported for each dataset, highlighting the robustness and consistency of the methods.

For each dataset, the highest accuracy value is highlighted in **bold**, indicating the superior performance of the corresponding method. Notably, DMT-HI consistently outperforms the other methods across the majority of datasets, particularly for complex datasets such as CIFAR-10, CIFAR-100, and HCL, demonstrating its effectiveness in preserving the global structure during dimensionality reduction. The last row of each table shows the average accuracy across all datasets, further emphasizing DMT-HI's overall superiority. Specifically, in the training set (Table A3), DMT-HI achieves an average accuracy of 73.4%, which is significantly higher than the next best method, PaCMAP, at 62.9%. Similar trends are observed in the testing set (Table A4), where DMT-HI also leads with an average accuracy of 68.1%, surpassing the other methods by a notable margin.

These results clearly indicate that DMT-HI provides better global structure preservation, which is crucial for maintaining inter-cluster relationships and ensuring interpretability of the reduced-dimensional space. The high accuracy across both

training and testing sets demonstrates the model's generalization capabilities, making it a reliable choice for dimensionality reduction tasks involving complex data distributions.

Tables A5 and A6 present the local structure preservation performance on thirteen datasets, measured using KNN classification accuracy. Unlike global structure preservation, local structure preservation focuses on how well the dimensionality reduction method retains the structure within neighborhoods after the data is mapped to a lower-dimensional space. These tables provide a detailed comparison of the performance across different methods for both the training and testing sets, with the best value for each dataset highlighted in **bold**.

In the training set (Table A5), DMT-HI consistently achieves the best performance on most datasets, particularly on complex datasets such as CIFAR-10, CIFAR-100, and HCL, where it significantly outperforms other methods. DMT-HI demonstrates a strong ability to capture local structures, excelling especially in image datasets. Additionally, it shows outstanding performance in biological datasets such as MCA and gast. This indicates that DMT-HI not only excels at preserving global structure but also retains fine-grained local structures effectively. The average KNN accuracy on the training set for DMT-HI reaches 82.8%, which is significantly higher than other methods like PaCMAP (71.5%) and HNNE (70.8%).

In the testing set (Table A6), DMT-HI also outperforms other methods, especially in complex datasets like CIFAR-10, CIFAR-100, MNIST, and KMNST, showing the best local structure preservation performance. Its ability to capture local patterns within neighborhoods is particularly remarkable in these datasets, further validating its strong generalization capabilities across various tasks. The average KNN accuracy on the testing set is 77.0%, again outperforming other methods, indicating that DMT-HI is highly stable in retaining local structures on unseen data.

When compared to the global structure preservation results (Tables A3 and A4), DMT-HI shows excellent performance in both local and global structure preservation. This indicates that DMT-HI not only maintains the overall shape and inter-cluster relationships of the data but also effectively captures the fine-grained patterns within neighborhoods. This dual advantage in both global and local structure preservation makes DMT-HI a powerful tool for dimensionality reduction tasks, particularly in complex datasets where accurate retention of local relationships is crucial.

## APPENDIX E

### THE DETAILS OF DATA AUGMENTATION ON IMAGE DATASETS

#### A. Data Augmentation of the Compared Methods

a) *BYOL augmentation*.: The BYOL augmentation method is a hand-designed method. It is composed of four parts: random cropping, left-right flip, color jittering, and color dropping. The details of each part are as follows:

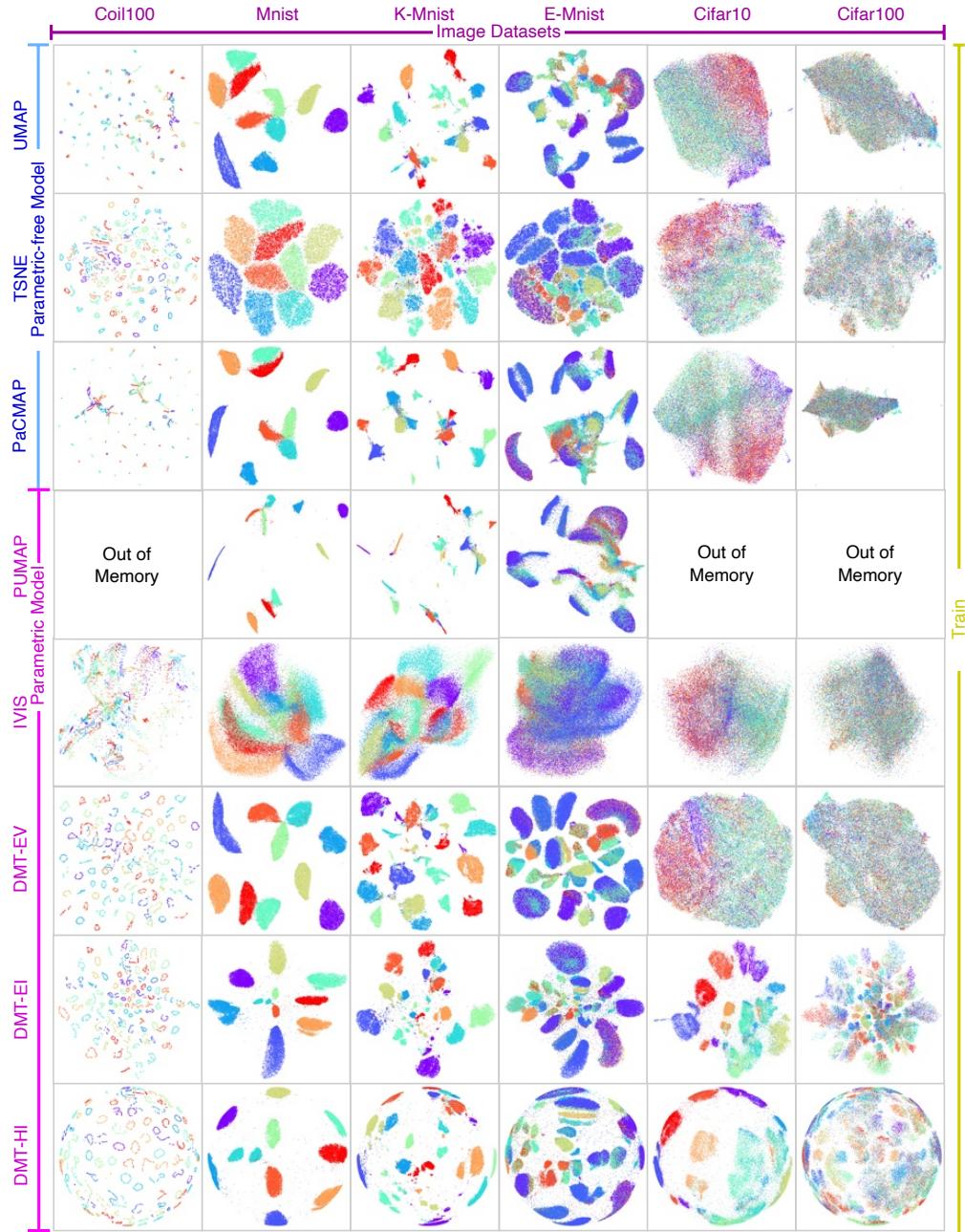
- Random cropping: A random patch of the image is selected, with an area uniformly sampled between 8% and 100% of that of the original image, and an aspect ratio logarithmically sampled between 3/4 and 4/3. This

patch is then resized to the target size of  $224 \times 224$  using bicubic interpolation.

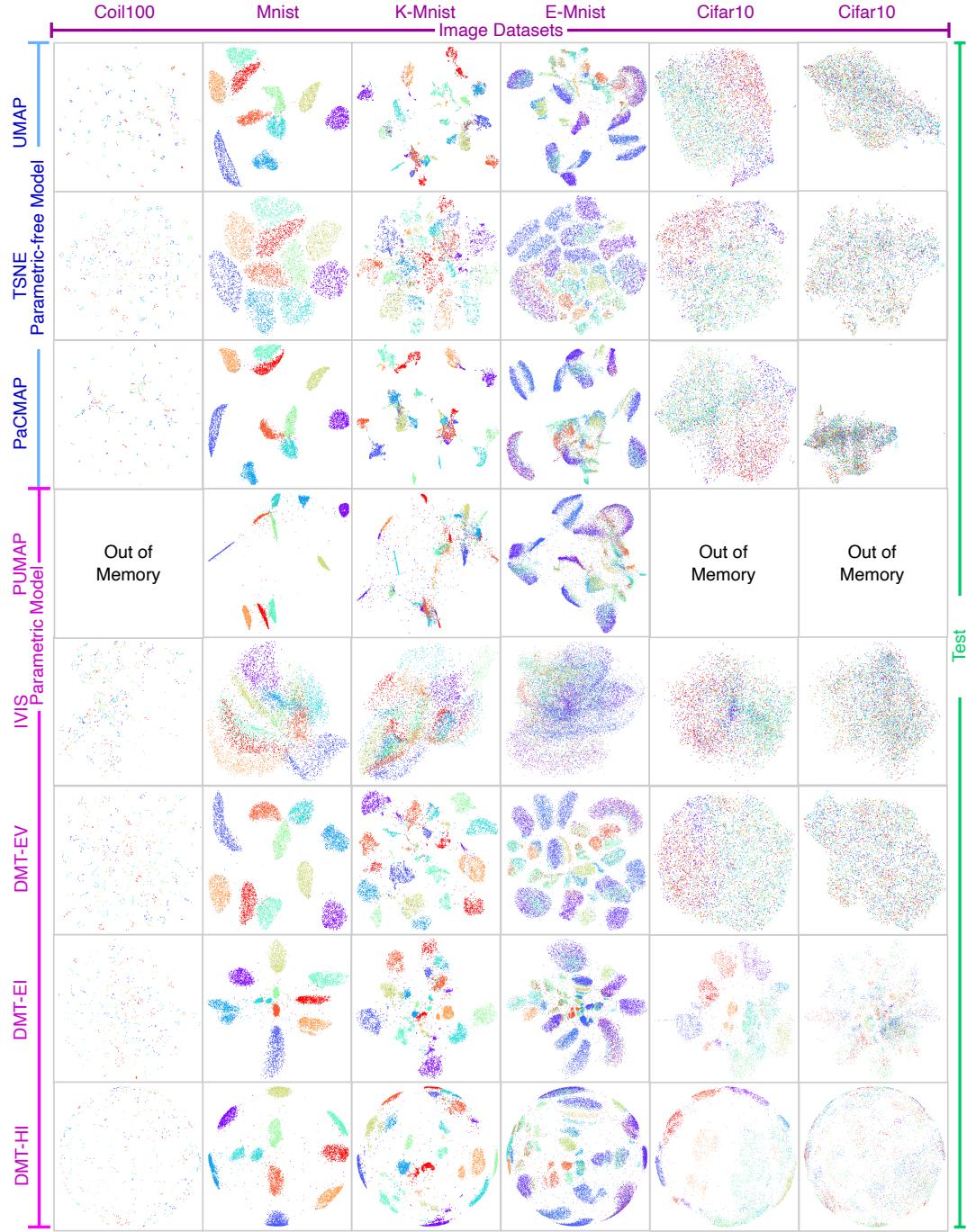
- Optional left-right flip.
- Color jittering: The brightness, contrast, saturation, and hue of the image are shifted by a uniformly random offset applied to all the pixels of the same image. The order in which these shifts are performed is randomly selected for each patch.
- Color dropping: An optional conversion to grayscale. When applied, the output intensity for a pixel  $(r, g, b)$  corresponds to its luma component, computed as  $0.2989r + 0.5870g + 0.1140b$ .

## APPENDIX F

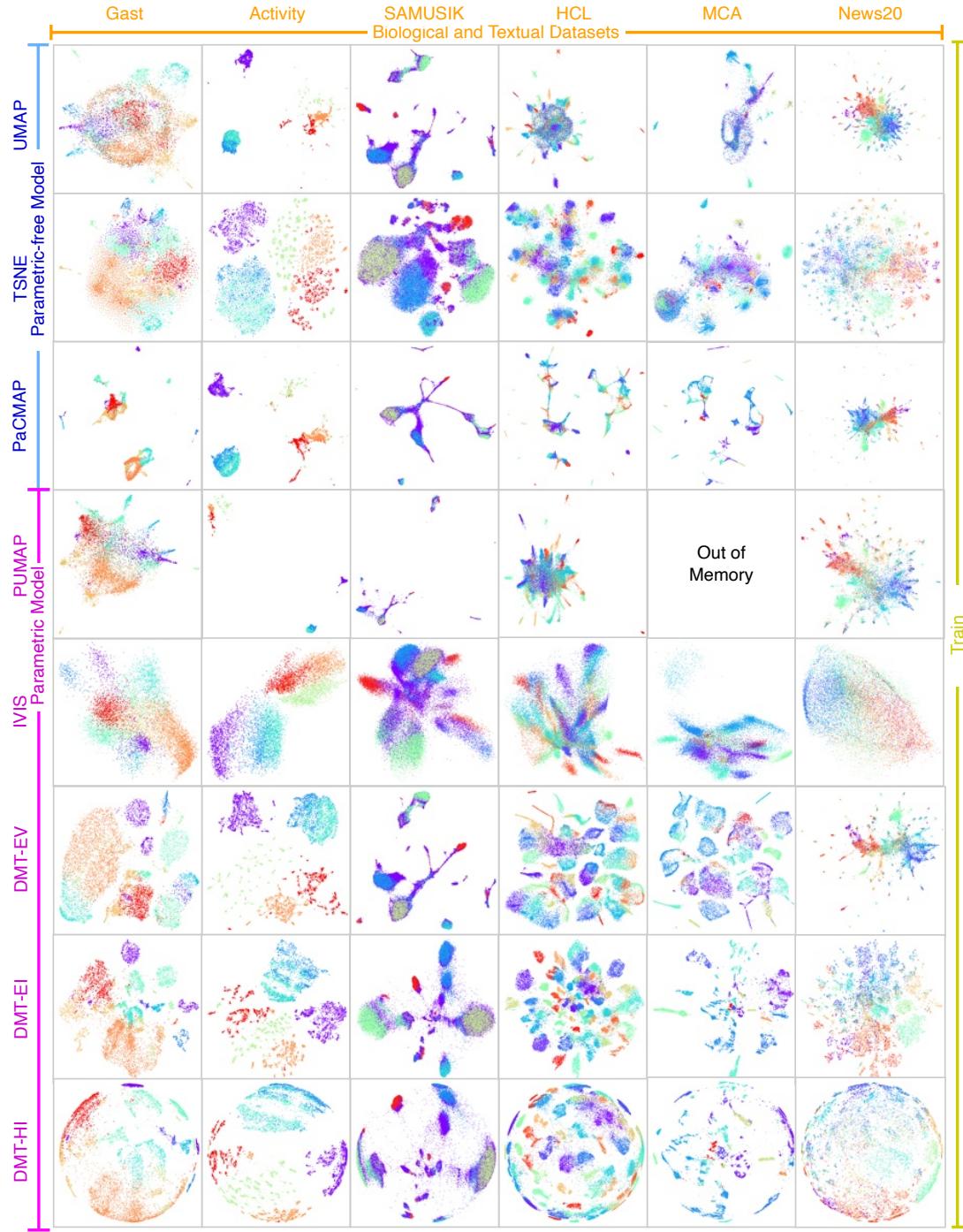
### THE DETAILS OF THE VISUALIZATION RESULTS



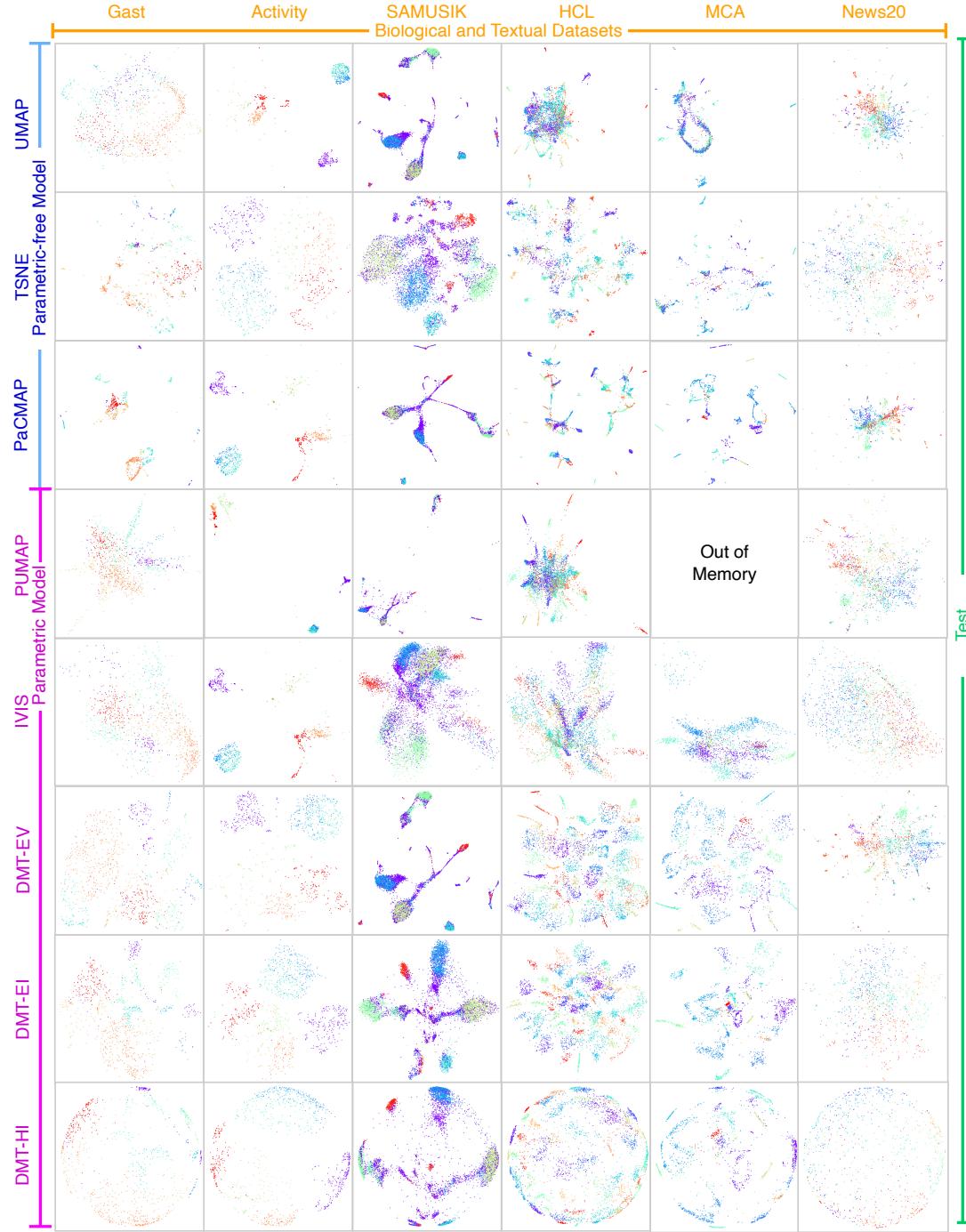
**Fig. A1. Detailed visualization results for dimensionality reduction methods applied to six image datasets: Coil100, MNIST, K-MNIST, E-MNIST, Cifar10, and Cifar100.** The comparison includes parametric-free models (UMAP, t-SNE, PaCMAP) and parametric models (PUMAP, IVIS, DMT-EV, DMT-EI, DMT-HI). While parametric-free models often encounter memory issues on large datasets such as Cifar10 and Cifar100, parametric models like DMT-HI demonstrate robust performance. DMT-HI, in particular, provides well-organized and clearly separated embeddings, even for complex datasets. These results highlight DMT-HI's capacity to handle both small and large datasets while preserving the global and local structures of the data. This visualization provides a more in-depth comparison across different techniques, showcasing the effectiveness of DMT-HI.



**Fig. A2. Detailed visualization results for dimensionality reduction methods applied to six image datasets: Coil100, MNIST, K-MNIST, E-MNIST, Cifar10, and Cifar100 (Test Set).** The comparison includes parametric-free models (UMAP, t-SNE, PaCMAP) and parametric models (PUMAP, IVIS, DMT-EV, DMT-EI, DMT-HI). For large datasets like Cifar10 and Cifar100, some parametric-free models encountered memory issues (marked as "Out of Memory"). DMT-HI consistently outperformed the other models, producing well-separated clusters and clear representations, demonstrating its robustness in handling both small and large datasets while maintaining local and global structure. These test set results confirm the consistent performance of DMT-HI across different dataset sizes and complexities.



**Fig. A3. Detailed visualization results for dimensionality reduction methods applied to six biological and textual datasets: GAST, Activity, SAMUSIK, HCL, MCA, and News20 (Train Set).** This comparison showcases parametric-free models (UMAP, t-SNE, PaCMAP) and parametric models (PUMAP, IVIS, DMT-EV, DMT-EI, DMT-HI). Similar to the previous visualization, certain parametric-free models encountered memory issues for larger datasets (marked as "Out of Memory"). DMT-HI consistently outperformed baseline models, particularly on complex biological datasets such as SAMUSIK and HCL, preserving local and global structures effectively. These results further demonstrate DMT-HI's robustness and adaptability across diverse dataset types, including both biological and textual data.



**Fig. A4. Detailed visualization results for dimensionality reduction methods applied to six biological and textual datasets: GAST, Activity, SAMUSIK, HCL, MCA, and News20 (Test Set).** This figure compares the performance of parametric-free models (UMAP, t-SNE, PaCMAP) and parametric models (PUMAP, IVIS, DMT-EV, DMT-EI, DMT-HI) in preserving structure during dimensionality reduction on test data. Similar to the train set, certain models faced memory constraints (marked as "Out of Memory"). DMT-HI shows clear advantages in preserving both local and global structures, particularly excelling on challenging datasets like SAMUSIK and HCL, demonstrating its strong generalization capabilities across both biological and textual test datasets.

TABLE A3

DETAILED GLOBAL STRUCTURE PRESERVATION PERFORMANCE (SVM CLASSIFICATION ACCURACY) COMPARISON ON THIRTEEN DATASETS - TRAINING SET. **BOLD** DENOTES THE BEST VALUE IN EACH ROW. THE LAST ROW SHOWS THE AVERAGE ACCURACY FOR EACH METHOD.

	tSNE	UMAP	PUMAP	Ivis	PaCMAP	HNNE	DMT-EV	DMT-HI
Coil20	73.4 ( $\pm 1.7$ )	79.8 ( $\pm 2.6$ )	81.3 ( $\pm 2.0$ )	59.6 ( $\pm 2.5$ )	81.4 ( $\pm 0.6$ )	64.2 ( $\pm 0.0$ )	84.0 ( $\pm 1.7$ )	<b>87.4</b> ( $\pm 2.5$ )
Coil100	76.8 ( $\pm 0.5$ )	71.6 ( $\pm 3.5$ )	-	42.8 ( $\pm 1.8$ )	75.5 ( $\pm 1.5$ )	40.4 ( $\pm 0.0$ )	86.3 ( $\pm 1.2$ )	<b>91.8</b> ( $\pm 1.2$ )
MNIST	95.4 ( $\pm 0.1$ )	96.5 ( $\pm 0.1$ )	96.7 ( $\pm 0.2$ )	67.9 ( $\pm 2.7$ )	95.8 ( $\pm 0.2$ )	72.2 ( $\pm 5.3$ )	96.9 ( $\pm 0.2$ )	<b>97.9</b> ( $\pm 0.2$ )
K-MNIST	68.3 ( $\pm 1.9$ )	70.6 ( $\pm 2.0$ )	70.4 ( $\pm 3.8$ )	54.5 ( $\pm 2.1$ )	<b>71.4</b> ( $\pm 2.5$ )	38.7 ( $\pm 4.8$ )	65.8 ( $\pm 4.4$ )	64.1 ( $\pm 5.0$ )
E-MNIST	64.3 ( $\pm 0.9$ )	65.4 ( $\pm 1.0$ )	62.2 ( $\pm 0.8$ )	28.1 ( $\pm 0.7$ )	63.8 ( $\pm 1.4$ )	35.0 ( $\pm 2.9$ )	66.4 ( $\pm 1.0$ )	<b>69.3</b> ( $\pm 0.7$ )
Cifar10	22.3 ( $\pm 0.4$ )	22.0 ( $\pm 0.5$ )	-	21.0 ( $\pm 0.7$ )	22.5 ( $\pm 0.5$ )	17.4 ( $\pm 0.9$ )	20.9 ( $\pm 1.0$ )	<b>77.3</b> ( $\pm 0.9$ )
Cifar100	4.3 ( $\pm 0.2$ )	4.4 ( $\pm 0.2$ )	-	4.7 ( $\pm 0.2$ )	4.6 ( $\pm 0.2$ )	3.1 ( $\pm 0.3$ )	4.3 ( $\pm 0.3$ )	<b>38.7</b> ( $\pm 0.7$ )
ACT	86.9 ( $\pm 0.3$ )	77.1 ( $\pm 4.0$ )	<b>88.3</b> ( $\pm 0.6$ )	81.9 ( $\pm 0.6$ )	83.6 ( $\pm 1.4$ )	76.0 ( $\pm 0.0$ )	85.1 ( $\pm 4.2$ )	86.9 ( $\pm 1.9$ )
GAST	62.5 ( $\pm 1.7$ )	40.8 ( $\pm 2.9$ )	62.2 ( $\pm 0.8$ )	62.2 ( $\pm 1.6$ )	75.3 ( $\pm 3.6$ )	54.9 ( $\pm 0.0$ )	75.8 ( $\pm 3.9$ )	<b>77.4</b> ( $\pm 4.7$ )
SAM	68.0 ( $\pm 0.5$ )	66.0 ( $\pm 3.5$ )	67.8 ( $\pm 1.7$ )	59.3 ( $\pm 2.9$ )	<b>69.3</b> ( $\pm 1.0$ )	61.2 ( $\pm 2.6$ )	67.7 ( $\pm 3.3$ )	64.8 ( $\pm 2.7$ )
HCL	64.9 ( $\pm 1.0$ )	32.6 ( $\pm 2.1$ )	45.3 ( $\pm 0.9$ )	51.7 ( $\pm 1.0$ )	77.9 ( $\pm 0.8$ )	34.1 ( $\pm 1.4$ )	75.9 ( $\pm 1.4$ )	<b>84.8</b> ( $\pm 1.1$ )
MCA	42.4 ( $\pm 1.6$ )	31.8 ( $\pm 2.0$ )	-	68.7 ( $\pm 1.0$ )	71.2 ( $\pm 4.1$ )	45.6 ( $\pm 1.2$ )	63.2 ( $\pm 4.9$ )	<b>76.7</b> ( $\pm 6.0$ )
NG20	32.5 ( $\pm 1.3$ )	25.9 ( $\pm 2.7$ )	31.5 ( $\pm 1.2$ )	21.7 ( $\pm 2.2$ )	25.9 ( $\pm 1.6$ )	15.1 ( $\pm 2.0$ )	29.3 ( $\pm 3.0$ )	<b>39.5</b> ( $\pm 1.4$ )
<b>Average</b>	58.6	52.7	-	48.0	62.9	42.9	63.2	<b>73.4</b>

TABLE A4

DETAILED GLOBAL STRUCTURE PRESERVATION PERFORMANCE (SVM CLASSIFICATION ACCURACY) COMPARISON ON THIRTEEN DATASETS - TESTING SET. **BOLD** DENOTES THE BEST VALUE IN EACH ROW. THE LAST ROW SHOWS THE AVERAGE ACCURACY FOR EACH METHOD.

	tSNE	UMAP	PUMAP	Ivis	PaCMAP	HNNE	DMT-EV	DMT-HI
Coil20	61.2 ( $\pm 1.3$ )	71.1 ( $\pm 1.8$ )	65.9 ( $\pm 1.5$ )	47.8 ( $\pm 2.9$ )	72.5 ( $\pm 1.4$ )	52.8 ( $\pm 0.0$ )	71.8 ( $\pm 3.1$ )	<b>76.2</b> ( $\pm 2.1$ )
Coil100	41.8 ( $\pm 0.8$ )	40.1 ( $\pm 3.4$ )	-	25.0 ( $\pm 2.0$ )	46.7 ( $\pm 2.0$ )	24.3 ( $\pm 0.0$ )	<b>51.3</b> ( $\pm 1.1$ )	45.7 ( $\pm 1.8$ )
MNIST	94.8 ( $\pm 0.2$ )	94.4 ( $\pm 0.2$ )	95.1 ( $\pm 0.1$ )	67.9 ( $\pm 2.9$ )	95.5 ( $\pm 0.2$ )	71.5 ( $\pm 5.9$ )	95.9 ( $\pm 0.2$ )	<b>97.3</b> ( $\pm 0.1$ )
K-MNIST	56.3 ( $\pm 2.2$ )	58.0 ( $\pm 2.8$ )	56.8 ( $\pm 3.0$ )	46.2 ( $\pm 1.4$ )	64.1 ( $\pm 2.2$ )	40.1 ( $\pm 3.7$ )	63.8 ( $\pm 3.6$ )	<b>64.5</b> ( $\pm 4.3$ )
E-MNIST	63.6 ( $\pm 0.6$ )	61.2 ( $\pm 0.4$ )	61.0 ( $\pm 0.5$ )	27.9 ( $\pm 0.5$ )	63.0 ( $\pm 0.5$ )	34.3 ( $\pm 2.5$ )	65.9 ( $\pm 1.8$ )	<b>68.9</b> ( $\pm 0.7$ )
Cifar10	22.9 ( $\pm 0.2$ )	22.7 ( $\pm 0.3$ )	-	20.6 ( $\pm 0.4$ )	23.7 ( $\pm 0.2$ )	18.1 ( $\pm 0.8$ )	21.4 ( $\pm 0.7$ )	<b>75.1</b> ( $\pm 0.7$ )
Cifar100	4.7 ( $\pm 0.2$ )	4.9 ( $\pm 0.2$ )	-	4.3 ( $\pm 0.2$ )	5.3 ( $\pm 0.2$ )	3.2 ( $\pm 0.5$ )	4.1 ( $\pm 0.4$ )	<b>38.7</b> ( $\pm 0.7$ )
ACT	86.4 ( $\pm 0.4$ )	74.8 ( $\pm 4.0$ )	81.5 ( $\pm 1.8$ )	81.9 ( $\pm 0.9$ )	82.6 ( $\pm 1.3$ )	76.3 ( $\pm 0.0$ )	84.3 ( $\pm 4.2$ )	<b>86.8</b> ( $\pm 2.3$ )
GAST	63.9 ( $\pm 1.7$ )	44.0 ( $\pm 4.3$ )	62.1 ( $\pm 1.2$ )	67.4 ( $\pm 1.5$ )	72.6 ( $\pm 3.8$ )	47.2 ( $\pm 0.0$ )	71.8 ( $\pm 3.6$ )	<b>77.2</b> ( $\pm 5.4$ )
SAM	69.1 ( $\pm 0.1$ )	67.0 ( $\pm 3.1$ )	68.9 ( $\pm 1.7$ )	59.5 ( $\pm 2.5$ )	<b>70.6</b> ( $\pm 0.1$ )	62.1 ( $\pm 2.4$ )	68.7 ( $\pm 3.1$ )	66.0 ( $\pm 3.0$ )
HCL	59.5 ( $\pm 1.1$ )	28.5 ( $\pm 1.8$ )	44.1 ( $\pm 0.6$ )	49.1 ( $\pm 1.6$ )	70.3 ( $\pm 0.7$ )	34.4 ( $\pm 0.0$ )	67.6 ( $\pm 1.5$ )	<b>77.3</b> ( $\pm 1.5$ )
MCA	41.6 ( $\pm 1.0$ )	32.4 ( $\pm 2.2$ )	-	67.3 ( $\pm 0.7$ )	71.7 ( $\pm 3.5$ )	42.8 ( $\pm 0.0$ )	62.9 ( $\pm 5.3$ )	<b>76.5</b> ( $\pm 3.6$ )
NG20	30.6 ( $\pm 0.8$ )	25.3 ( $\pm 2.0$ )	30.3 ( $\pm 1.1$ )	22.1 ( $\pm 2.3$ )	26.2 ( $\pm 1.7$ )	13.7 ( $\pm 1.6$ )	27.0 ( $\pm 2.6$ )	<b>37.4</b> ( $\pm 1.3$ )
<b>Average</b>	53.6	48.0	-	45.2	58.8	40.1	58.2	<b>68.1</b>

TABLE A5

DETAILED LOCAL STRUCTURE PRESERVATION PERFORMANCE (KNN CLASSIFICATION ACCURACY) COMPARISON ON THIRTEEN DATASETS - TRAINING SET. **BOLD** DENOTES THE BEST VALUE IN EACH ROW. THE LAST ROW SHOWS THE AVERAGE ACCURACY FOR EACH METHOD.

	tSNE	UMAP	PUMAP	Ivis	PaCMAP	HNNE	DMT-EV	DMT-HI
Coil20	92.9 ( $\pm 0.3$ )	87.2 ( $\pm 0.3$ )	86.2 ( $\pm 0.3$ )	70.1 ( $\pm 1.1$ )	86.1 ( $\pm 0.5$ )	90.0 ( $\pm 0.0$ )	90.0 ( $\pm 0.5$ )	<b>97.3</b> ( $\pm 0.5$ )
Coil100	94.4 ( $\pm 0.2$ )	90.6 ( $\pm 0.1$ )	-	62.6 ( $\pm 1.2$ )	89.9 ( $\pm 0.4$ )	89.0 ( $\pm 0.0$ )	93.9 ( $\pm 0.3$ )	<b>97.8</b> ( $\pm 0.3$ )
MNIST	95.2 ( $\pm 0.1$ )	96.3 ( $\pm 0.1$ )	96.5 ( $\pm 0.2$ )	71.1 ( $\pm 2.7$ )	96.0 ( $\pm 0.1$ )	94.9 ( $\pm 0.2$ )	96.8 ( $\pm 0.2$ )	<b>97.7</b> ( $\pm 0.2$ )
K-MNIST	93.4 ( $\pm 0.2$ )	95.3 ( $\pm 0.2$ )	93.8 ( $\pm 0.3$ )	73.7 ( $\pm 0.6$ )	94.2 ( $\pm 0.2$ )	88.7 ( $\pm 0.3$ )	95.2 ( $\pm 0.2$ )	<b>96.0</b> ( $\pm 0.1$ )
E-MNIST	69.5 ( $\pm 0.5$ )	68.6 ( $\pm 0.6$ )	64.1 ( $\pm 0.5$ )	29.1 ( $\pm 0.8$ )	66.5 ( $\pm 0.6$ )	65.7 ( $\pm 0.5$ )	71.9 ( $\pm 0.6$ )	<b>73.8</b> ( $\pm 0.4$ )
Cifar10	26.2 ( $\pm 0.4$ )	20.5 ( $\pm 0.3$ )	-	18.3 ( $\pm 0.2$ )	20.7 ( $\pm 0.3$ )	28.2 ( $\pm 0.4$ )	24.1 ( $\pm 0.4$ )	<b>74.5</b> ( $\pm 0.3$ )
Cifar100	8.6 ( $\pm 0.2$ )	5.1 ( $\pm 0.2$ )	-	3.3 ( $\pm 0.2$ )	5.3 ( $\pm 0.3$ )	9.3 ( $\pm 0.2$ )	6.3 ( $\pm 0.2$ )	<b>39.9</b> ( $\pm 0.2$ )
ACT	<b>94.0</b> ( $\pm 0.1$ )	91.3 ( $\pm 0.1$ )	91.0 ( $\pm 0.2$ )	80.4 ( $\pm 0.7$ )	88.8 ( $\pm 0.2$ )	92.4 ( $\pm 0.0$ )	91.6 ( $\pm 0.2$ )	92.1 ( $\pm 0.2$ )
GAST	77.0 ( $\pm 0.8$ )	57.0 ( $\pm 0.9$ )	67.8 ( $\pm 0.4$ )	68.2 ( $\pm 1.0$ )	90.0 ( $\pm 0.5$ )	87.8 ( $\pm 0.0$ )	86.3 ( $\pm 0.8$ )	<b>93.0</b> ( $\pm 0.2$ )
SAM	<b>75.8</b> ( $\pm 0.4$ )	73.9 ( $\pm 0.3$ )	74.7 ( $\pm 0.6$ )	70.9 ( $\pm 1.1$ )	74.4 ( $\pm 0.4$ )	72.8 ( $\pm 0.3$ )	74.4 ( $\pm 0.4$ )	74.8 ( $\pm 0.8$ )
HCL	72.7 ( $\pm 0.5$ )	39.2 ( $\pm 0.9$ )	49.7 ( $\pm 0.5$ )	51.5 ( $\pm 1.6$ )	84.5 ( $\pm 0.5$ )	72.9 ( $\pm 0.5$ )	79.7 ( $\pm 0.8$ )	<b>86.9</b> ( $\pm 0.3$ )
MCA	66.3 ( $\pm 1.4$ )	38.1 ( $\pm 0.7$ )	-	70.8 ( $\pm 0.9$ )	92.3 ( $\pm 0.3$ )	69.6 ( $\pm 0.2$ )	86.6 ( $\pm 0.4$ )	<b>93.8</b> ( $\pm 0.2$ )
NG20	54.0 ( $\pm 0.2$ )	45.7 ( $\pm 0.9$ )	44.8 ( $\pm 0.3$ )	18.2 ( $\pm 1.6$ )	40.9 ( $\pm 0.4$ )	<b>59.6</b> ( $\pm 0.4$ )	47.6 ( $\pm 0.9$ )	58.7 ( $\pm 0.6$ )
<b>Average</b>	70.8	62.2	-	52.9	71.5	70.8	72.6	<b>82.8</b>

TABLE A6

DETAILED LOCAL STRUCTURE PRESERVATION PERFORMANCE (KNN CLASSIFICATION ACCURACY) COMPARISON ON THIRTEEN DATASETS - TESTING SET. **BOLD** DENOTES THE BEST VALUE IN EACH ROW. THE LAST ROW SHOWS THE AVERAGE ACCURACY FOR EACH METHOD.

	tSNE	UMAP	PUMAP	Ivis	PaCMAP	HNNE	DMT-EV	DMT-HI
Coil20	59.0 ( $\pm 0.8$ )	62.2 ( $\pm 2.5$ )	59.9 ( $\pm 2.7$ )	52.8 ( $\pm 1.9$ )	62.6 ( $\pm 1.5$ )	54.8 ( $\pm 0.0$ )	62.4 ( $\pm 1.5$ )	<b>66.2</b> ( $\pm 2.3$ )
Coil100	63.7 ( $\pm 0.5$ )	59.7 ( $\pm 1.6$ )	-	43.2 ( $\pm 1.6$ )	62.5 ( $\pm 0.7$ )	58.3 ( $\pm 0.0$ )	<b>69.8</b> ( $\pm 0.8$ )	62.6 ( $\pm 1.2$ )
MNIST	94.4 ( $\pm 0.2$ )	94.0 ( $\pm 0.2$ )	94.7 ( $\pm 0.2$ )	70.4 ( $\pm 2.9$ )	95.6 ( $\pm 0.1$ )	94.1 ( $\pm 0.2$ )	95.7 ( $\pm 0.2$ )	<b>97.1</b> ( $\pm 0.1$ )
K-MNIST	88.7 ( $\pm 0.3$ )	88.3 ( $\pm 0.1$ )	84.2 ( $\pm 0.4$ )	65.6 ( $\pm 0.8$ )	90.4 ( $\pm 0.2$ )	86.5 ( $\pm 0.3$ )	91.7 ( $\pm 0.2$ )	<b>93.9</b> ( $\pm 0.2$ )
E-MNIST	67.9 ( $\pm 0.3$ )	65.3 ( $\pm 0.4$ )	62.3 ( $\pm 0.4$ )	29.2 ( $\pm 0.6$ )	65.8 ( $\pm 0.4$ )	65.0 ( $\pm 0.4$ )	70.9 ( $\pm 0.6$ )	<b>72.5</b> ( $\pm 0.3$ )
Cifar10	24.1 ( $\pm 0.4$ )	20.2 ( $\pm 0.4$ )	-	18.6 ( $\pm 0.4$ )	20.5 ( $\pm 0.3$ )	26.8 ( $\pm 0.6$ )	22.8 ( $\pm 0.6$ )	<b>74.2</b> ( $\pm 0.4$ )
Cifar100	7.1 ( $\pm 0.2$ )	5.6 ( $\pm 0.2$ )	-	3.0 ( $\pm 0.2$ )	5.5 ( $\pm 0.3$ )	7.9 ( $\pm 0.3$ )	5.4 ( $\pm 0.2$ )	<b>39.7</b> ( $\pm 0.3$ )
ACT	<b>90.6</b> ( $\pm 0.4$ )	89.6 ( $\pm 0.5$ )	89.8 ( $\pm 0.3$ )	80.8 ( $\pm 0.9$ )	88.7 ( $\pm 0.2$ )	86.5 ( $\pm 0.0$ )	89.7 ( $\pm 0.6$ )	90.3 ( $\pm 0.4$ )
GAST	72.8 ( $\pm 1.1$ )	55.0 ( $\pm 1.7$ )	69.0 ( $\pm 0.7$ )	69.6 ( $\pm 1.0$ )	<b>87.0</b> ( $\pm 0.6$ )	87.4 ( $\pm 0.0$ )	78.8 ( $\pm 1.6$ )	86.9 ( $\pm 1.0$ )
SAM	<b>75.6</b> ( $\pm 0.1$ )	74.9 ( $\pm 0.2$ )	75.7 ( $\pm 0.3$ )	71.2 ( $\pm 1.2$ )	75.1 ( $\pm 0.2$ )	73.2 ( $\pm 0.3$ )	74.6 ( $\pm 0.2$ )	75.5 ( $\pm 0.5$ )
HCL	67.9 ( $\pm 0.3$ )	33.5 ( $\pm 0.8$ )	45.3 ( $\pm 0.5$ )	49.2 ( $\pm 1.8$ )	<b>83.0</b> ( $\pm 0.5$ )	71.3 ( $\pm 0.0$ )	72.1 ( $\pm 0.5$ )	79.5 ( $\pm 0.5$ )
MCA	59.9 ( $\pm 1.6$ )	39.4 ( $\pm 1.2$ )	-	70.1 ( $\pm 1.2$ )	91.6 ( $\pm 0.3$ )	70.0 ( $\pm 0.0$ )	84.1 ( $\pm 0.8$ )	<b>90.0</b> ( $\pm 0.6$ )
NG20	42.1 ( $\pm 0.5$ )	40.0 ( $\pm 1.0$ )	36.6 ( $\pm 0.7$ )	17.7 ( $\pm 2.1$ )	37.7 ( $\pm 0.8$ )	<b>45.9</b> ( $\pm 0.4$ )	43.3 ( $\pm 0.8$ )	46.0 ( $\pm 0.9$ )
Average	65.4	59.6	-	51.8	71.2	69.4	71.5	<b>77.0</b>