DMT-EV: An Explainable Deep Network for Dimension Reduction

Zelin Zang©, Shenghui Cheng©, Hanchen Xia, Liangyu Li, Yaoting Sun, Yongjie Xu©,
Lei Shang, Baigui Sun, and Stan Z. Li©, Fellow, IEEE

Abstract—Dimension reduction (DR) is commonly utilized to capture the intrinsic structure and transform high-dimensional data into low-dimensional space while retaining meaningful properties of the original data. It is used in various applications, such as image recognition, single-cell sequencing analysis, and biomarker discovery. However, contemporary parametric-free and parametric DR techniques suffer from several significant shortcomings, such as the inability to preserve global and local features and the poor generalisation performance. On the other hand, regarding explainability, it is crucial to comprehend the embedding process, especially the contribution of each part to the embedding process, while understanding how each feature affects the embedding results that identify critical components and help diagnose the embedding process. To address these problems, we have developed a deep neural network method called DMT-EV, which provides not only excellent performance in structural maintainability but also explainability to the DR therein. DMT-EV starts with data augmentation and a manifold-based loss function to improve embedding performance. The explanation is based on saliency maps and aims to examine the trained DMT-EV parameters and contributions of components during the embedding process. The proposed techniques are integrated with a visual interface to help the user to adjust DMT-EV to achieve better DR performance and explainability. The interactive visual interface makes it easier to illustrate the data features, compare different DR techniques, and investigate DR. An in-depth experimental comparison shows that DMT-EV consistently outperforms the state-of-the-art methods in both performance measures and explainability.

Index Terms—Dimension reduction, explainability of DR models, deep learning, parametric model

1 INTRODUCTION

Dimension reduction (DR) techniques aim to find a low-dimensional representation of a high-dimensional dataset that retains as much as possible its original structure. When used for visualization, the output is usually set to two or three dimensions. The results are commonly visualized with scatterplots, where nearby points model similar objects and distant points model dissimilar ones. DR techniques have a wide range of applications. For example, some DR techniques are used to analyze gene expression patterns and find new cell types [52] as well as eye movement patterns [23]. Some DR techniques are even used to analyze the network parameter distribution and diagnose the training state of neural networks [39].

Explaining DR models and their output has become a critical scientific concern. Although DR techniques are excellent for studying high-dimensional data, they are like a black box, as they are challenging to diagnose and understand [24] their outcome. Furthermore, researchers are interested in how DR algorithms make decisions. It is because figuring out which features (features) impact the DR output is crucial to discover new scientific knowledge. For example, when biology researchers use DR to process mRNA data and find a unique cell type, they want to see which mRNA are associated with it. As another example, medical researchers want to discover biomarkers highly correlated with disease progression in the DR of big data.

The explanation of DR technology is not a new topic. Some classical linear DR methods have natural explainability, such as Principal Component Analysis (PCA) [55]. PCA measures the contribution of each feature to the DR output through eigenvalues. The users can refine the dataset or explore the role of certain features with the help of the relative importance provided by eigenvalues. However, to get good results on complex high-dimensional data, there is a tendency to use parametric-free DR methods (e.g., t-SNE (tSNE) [20] and UMAP [33]). Since those methods optimize only the output results and not the generalizable model parameters, resulting in a catastrophic decrease in explainability.

We observe that the work on the explanation of DR methods is classified into two categories [4], [34]. (1) Color annotation. These methods demonstrate the relationship between features and DR output by adding a particular background color or annotation [28], [46]. Although these methods aid in exploring the data, they cannot reveal the implicit patterns and essential features. (2) Proxy explanation. These methods explain a DR approach by analyzing the parameters of the proxy model [5], [13], [15]. Although they are more scalable as they can work with DR, such methods are only a simple explanation of DR.

© 2022 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Moreover, the proxy models cannot accurately fit DR output, causing the explanation results to be ambiguous.

The above method can be seen as a post-hoc analysis of the DR method. With the development of neural network based explainability methods, a new scheme, named Neural network-based (NN-based) explanation, has the potential to enhance the explainability of DR methods. If an NN-based neural network DR model is trained, then mining the information embedded in the model by explanation tools (e.g., saliency map [45] and counterfactual explanations [53]) can effectively achieve the explanation of DR.

Unfortunately, despite the potential of NN-based DR methods, few high-performance NN-based DR methods have been proposed. There are two reasons for this. (R1) Parametric models are challenging to train. Based on previous theoretical studies [17] and experiments comparisons in this paper, it is more difficult to fit a parametric model than a parametric-free model because the optimized objects are parameters carrying generalized knowledge. Moreover, the current methods, like PUMAP [43] and Ivis [49], do not design a suitable loss function and network structure for parametric models. As a result, the performance of these parametric methods is lower than that of the original parameter-free methods. (R2) Explanation scheme is challenging to design. Most of the current explainable analyses rely on supervised signals to accomplish. Explainable analysis theories and schemes for DR methods have not been discovered yet.

To address the issues mentioned above, we propose Deep Manifold Transformation for Explainable Visualization (DMT-EV) in unsupervised, supervised, and semi-supervised DR tasks. Furthermore, based on DMT-EV, we propose an explanation scheme for explainable analysis based on the model parameter and DR output.

For (R1), richer prior knowledge is introduced for training a high-performance DR model. Inspired by unsupervised contrastive learning (UCL [7]), we design a new network architecture and a structure-preserving loss function surrounding data augmentation. The new design strengthens the assumption of local connectivity [26] - the localities of manifolds are semantically similar - leading to better DR performance. For (R2), we polish the network structure and design three explanation schemes to provide an in-depth exploration of the parametric DR model. A gate layer is proposed to evaluate the importance of each feature and filter out unimportant features. In addition, an interaction mechanism between DR input and DR output is established to explore the critical features of the local areas of DR output. The novel contributions in this paper are as follows:

- **A novel deep-network-based model for DR embedding.** The model introduces data augmentation as prior knowledge to train the parametric DR model.
- **Explanation schemes for DR model.** The explanation schemes of features are proposed based on the decoupled network parameters and the interaction mechanism between DR input and DR output.
- **Excellent performance.** Experiments demonstrate that our proposed method not only has stronger explainability but also can produce better visualization results. The experiments on multiple image and biological datasets indicate that the proposed DMT-EV better preserves the intrinsic structure of the original manifold.

## 2 RELATED WORK

In this section, we briefly review the concepts of DR methods from a visualization perspective and explainable AI techniques that are commonly used by researchers.

### 2.1 Dimension Reduction and Data Visualization

The dimensional reduction (DR) methods include parametric-free and parametric methods.

The parametric-free DR method directly optimizes the embedding output corresponding to the input data rather than learning a continuous mapping function from the input to the embedding result. These methods are widely acknowledged because of the proven system and low computation costs. Some approaches focus on preserving global information and are built around the principle of distance (or ‘dissimilarity’) preservation. For example, MDS [22] characterizes the data relationships in low-dimensional space according to pairwise distance. ISOMAP [51] improves MDS by describing the data relationships based on geodesic similarity. LLE [42] improves the MDS by introducing the linearity assumption for localized manifolds, allowing it to model linearly only locally. Laplacian Eigenmap (LapEig) [27] uses graph embedding methods to improve the representation process by introducing undirected graphs into the modeling process. In addition, some methods introduce the manifold assumption to produce a non-linear embedding by emphasizing the local manifold structure. The SNE method [19], for example, assumes that the distance between the data satisfies the normal distribution and transforms the pairwise distance into pairwise similarity by the normal kernel function. The tSNE [32] and LargeVis [50] improve SNE by substituting the t-distribution kernel function, resulting in a better global structure-preserving performance. UMAP [33] makes a significant reduction in computational cost by introducing negative samples. Such above methods are widely used in data analysis fields such as life sciences and physics.

A significant drawback of parametric-free DR method is their poor explainability. Most of them are non-linear methods. The non-linear property implies that these methods are difficult to understand. The parameter-free property leads to the learned embedding knowledge not being stored in the model as parameters, making it impossible to generalize existing embeddings to newly collected datasets.

Some parameter variants of the above methods are proposed to improve generability and explainability. Topological autoencoders (TAE) [37] and Geometry Regularized AutoEncoders (GRAE) [11], for example, train autoencoders directly with local distance constraints in the input space. Ivis [49] propose a triplet loss function with distance as a constraint to train the neural network. Based on tSNE and UMAP, parametric tSNE (PtSNE) [31], and parametric UMAP (PUMAP) [43] introduce neural networks to learn a continues mapping from the input data to the output. Recently, Zang et al. [56] design a DR neural network that can process biological and image data based on the assumption of manifold flatness. While these methods have the potential to address generalization and explanation issues, these methods do not achieve satisfactory performance because they do not make practical improvements to the loss function and training form. The loss function of the...
parameter-free model is no longer suitable for the parametric approach, which inevitably causes a degradation of the optimization results. As a result, more applications, such as biological data visualization [6], [48], use non-parametric models rather than parametric models. We expect to design a naturally explainable DR model that achieves high performance and explainability.

2.2 Explanation of DR Methods

Non-linear DR techniques are commonly used for data exploration [38]. It includes finding better projections, communicating errors, and automatically detecting and visualizing features (features). Furthermore, some researchers have looked into the explainability of DR techniques to improve their understanding of the data and increase the reliability of DR embedding techniques.

Some methods add a particular background color or annotation to the reduced vector, demonstrating the features’ embedded relationship with the embedding [28]. Aupetit et al. [3] proposes that any measure linked with a projected datum or a pair of projected data be visualized by coloring the matching Voronoi cell in the projection space. Pagliosa et al. [9] presents a web-based visualization tool that enriches multidimensional projection layout with statistical measures derived from input data. The proposed method used statistical measures, such as variance, to highlight relevant attributes contributing to each region’s similarities. ProbingProjection [47] aims to reconstruct a continuous scalar field that is directly visualized using field visualization techniques. NoLiES [46] uses a set-based visualization method for binned attribute values, enabling users to observe the structure and perform interactive analysis quickly and allowing users to monitor the data structure.

Some methods use a simple (linear, in most cases) proxy model to fit the embedding output. The DR model’s behavior is explained by looking at the proxy model’s behavior and feature importance. VisExPreS [14], [15] produces a local approximation of the neighborhood structure to generate interpretable explanations on the preserved locality for a single instance. Chatzipimpapas et al. [5] present t-visSNE, an interactive tool for the visual exploration of tSNE projection. The t-visSNE enables analysts to inspect different aspects of their accuracy and meaning, such as the effects of hyper-parameters, distance and neighborhood preservation, densities and costs of specific communities, and the correlations between dimensions and visual patterns. Fujiwara et al. [13] propose a visual analysis method that effectively highlights the essential features of clusters in the DR results. To extract the important features, Fujiwara introduces ccPCA, which calculates the relative contribution of each feature to the comparison between one cluster and other clusters by building a local agent model.

Some methods follow the concept of a continuous field that tries to recover the non-linear axes or illustrate regions of maximal attribute values. The DataContextMap [8] enriches the embedding with additional data points that locate regions of high attribute values and augment the visualization with other attribute-based contours on the reconstructed scalar field. DimReader[12] augments the embedding with non-linear grid lines, and prolines display the non-linear axes.

2.3 Neural Network-Based Explanation

The studies of the explainability of neural networks have received wide attention since the success of deep neural networks for tasks such as computer vision and natural language processing. The explanation of machine learning models can be generally divided into an exploration of global aspects or local aspects [40].

On the global level, the scope of interpretation is on understanding how the model has produced the predictions [1]. On this level, we can examine which features affect the predictions most and what interactions there are between features [16]. On the local level, it is possible to explore and understand how and why the model produces predictions locally since these can be approximated with simple interpretable models [35], such as LIME [41] and SHAP [30]. So far, we have not found any neural network-based DR explainable methods.

3 DEEP NETWORK FOR EMBEDDING

In this section, we aim to design a high-performance and highly interpretable parametric DR model based on neural networks named DMT-EV. The design details of DMT-EV are described, including data augmentation, neural network architecture, and loss functions. It also explains the motivation and necessity of DMT-EV.

3.1 DR Embedding and Explainability Analysis

Given a training dataset \( X \), DR methods learn a continuous function \( F(\cdot) \) to map data \( X = \{x_1, \ldots, x_M\} \) (\( x \in \mathbb{R}^n \), \( n \) is the number of the features, \( M \) is the number of the training data) to a latent space \( Z \) (\( z \in \mathbb{R}^2 \) or \( z \in \mathbb{R}^3 \)). Furthermore, to better generalize the DR model and explain the DR results, the learned continuous function \( F(\cdot) \) is desired to be generalized to the newly collected dataset (test dataset) \( X^{test} \). The test dataset is independent and identically distributed (iid) with the training dataset.

The explanation of the DR method aims to discover the importance of features for DR associated with the embedding output \( Z \) and the embedding process \( F(\cdot) \). In this paper, the explanation output of the DR method includes the following aspects \( I = \{I^{F}, I^{F'}, I^{F''}\} \): (a) Global explanation \( (I^{F}) \), discovers the essential features of the global embedding process and evaluates the importance of these features. Simultaneously, irrelevant or noisy features are discarded by the DR methods. (b) Local explanation \( (I^{F'}) \), for any local area of the embedding output (selected by the user or through clustering methods), local explanation discovers the key features that lead the model to map the data into the local embedding area and evaluates the importance of those features. (c) Transformation explanation \( (I^{F''}) \), for interactions between local areas of the embedding output, the transformation explanation discovers the key features that lead the model to map data into interactions of local embedding areas and evaluates the importance of these features.

3.2 Data Augmentation in DR Embedding

Data augmentation (in Fig. 1a) is a well-known neural network (NN) training strategy for image classification and signal processing [44]. However, to the best of our knowledge, few works have been found to improve the DR
methods with data augmentation. It is first attributed to the difficulty of applying data augmentation directly in a parametric-free model. In addition, it is challenging to design the loss functions compatible with data augmentation in the parameter DR training process. Therefore, we introduce data augmentation as a substantial prior knowledge to improve the generalizability and explainability for solving the above problem.

Data augmentation improves the performance of the methods by precisely fitting the data distribution. First, data augmentation enhances the diversity of data, thereby overcoming overfitting. Second, data augmentation essentially reinforces the fundamental assumption of DR, i.e., local connectivity of neighborhoods. It learns refined data distribution by generating more intra-manifold data based on sampled points.

In addition, data augmentation can be guided by labels or other prior knowledge. We can use such data augmentation to be compatible with supervised and semi-supervised data settings without modifying the algorithm. A practical application is in Section 6.

In unsupervised contexts (same settings as tSNE [32], UMAP [33], etc), data augmentation generates new data \( x' = T(x) \) by:

\[
x' = T(x) = \{ \tau(x^1), \ldots, \tau(x^j), \ldots, \tau(x^n) \}
\]

\[
\tau(x^j) = (1 - r_u) \cdot x^j + r_u \cdot \bar{x}^j, \quad \bar{x} \sim \mathcal{N}(x),
\]

where the augmented \( x' \) is the output of linear interpolation of input data \( x \) and \( \bar{x} \), and \( \bar{x} \) is sampled from the neighborhood \( \mathcal{N}(x) \) of \( x \). When no additional prior knowledge is available, the neighborhood is determined by the k-NN method \( \mathcal{N}^k(x) \). The combination parameter \( r_u \sim U(0, p_u) \) is sampled from the uniform distribution \( U(0, p_u) \), and \( p_u \) is the hyperparameter.

In supervised contexts, we only need to change the form of neighborhood selection:

\[
x' = T(x) = \{ \tau(x^1), \ldots, \tau(x^j), \ldots, \tau(x^n) \} \quad \tau(x^j) = (1 - r_u) \cdot x^j + r_u \cdot \bar{x}^j, \quad \bar{x} \sim \mathcal{N}^k(x) \cap Y(x)
\]

where \( Y(x) \) is the set of nodes with the same label as \( x \). In the above scenario, regardless of whether labels are collected for each data point, we can update the neighborhood relationship, thus incorporating the prior semantics of the category labels into the model training.

### 3.3 Deep Network Architecture

We adapt the neural network structure to achieve high performance and to improve the explainability of the DR model. The proposed neural network structure (Fig. 1a) consists of: lasso network and embedding network \( (m_u(\cdot)) \).

The lasso network includes gate layer \( g_W(\cdot) \), projection layers \( f_o(\cdot) \) is designed to handle noisy and irrelevant features and discover the globally important features for DR embedding. The gate layer \( g_W(\cdot) \) evaluates the importance of each feature with a gate operation and prevents unimportant features from forwarding propagation. The projection layer \( f_o(\cdot) \) maps the data with selected features \( \bar{x} \) into a latent space.

\[
x = g_W(x) = \{ x^j | x^j = x^j \cdot w^j \cdot 1_{w^j > \epsilon} \} \quad y = f_o(\bar{x})
\]

where \( x^j \in x \) is \( j \)-th feature of data \( x \). The \( w^j \in W \) is the parameter representing the importance of \( j \)-th feature, and \( W \) is the weight parameter vector of the gate layer. The gate operation \( 1_{w^j > \epsilon} \) ensures that less important features (less than the hyperparameter threshold \( \epsilon \)) do not leak to the following network layer. The latent space embedding vectors \( y \in Y \) is the output of projection layers \( f_o(\cdot) \). \( f_o(\cdot) \) is the backbone network and we use multilayer perceptron (MLP) for all datasets.
The lasso network is designed for pre-processing and building latent space data structures. The lasso network allows us to avoid the dilemma of vanilla DR methods (such as tSNE and UMAP) - constructing the neighborhood parameter $\sigma$ for each data point, the $\sigma$ is used to scale the pairwise distances to compute the similarity better, check Eq. (1) in [32] for detail. It also avoids the damage to the local structure caused by the inappropriate PCA for ultra-high-dimensional data [2].

Next, embedding vectors $y$ is mapped to lower space for visualization.

$$z = m_\theta(y)$$

where $z \in Z$ is the DR results in low-dimensional latent space, and $m_\theta$ is a simple reduced-dimensional mapping network.

### 3.4 Loss Functions

The network architecture and data augmentation are introduced, and the loss function used in concert with them is presented next.

The traditional DR loss function oriented to the parameter-free model cannot be adapted to the training of parametric models. The first reason comes from the pre-calculation of input data pairwise similarity. The vanilla parameter-free DR method (e.g., tSNE or UMAP) directly calculates the similarity of the input data $X$ and then uses the Kullback-Leibler divergence loss [32] or fuzzy set cross-entropy loss [33] to minimize the difference between the input space and the latent space. These methods use pre-calculated input data pairwise similarity to escape the duplication of calculations. A crucial pre-calculation step is using a binary search to compute the kernel function’s bandwidth $\sigma^2$. However, such pre-calculation is not compatible with data augmentation because the $\sigma^2$ of data $X$ can not be pre-calculated, and the time consumed by realtime-$\sigma$-calculation is unacceptable. The second reason comes from the noisy features of the input data. Considering that the input data contains redundant or noisy features, we cannot directly use the similarity of the input data as the fitting target. Instead, we should incorporate a priori information about the data augmentation to design the loss function.

Given a pire of augmented data $x_\alpha$ and $x_\beta$, a novel structure-preserving loss function $L_{sp}$ compatible with data augmentation is as follows:

$$L_{sp} = \frac{1}{(B - 1)^2} \sum_{i,j=1}^{B} \bar{S}^z_{ij} \log S^z_{ij} + (1 - \bar{S}^y_{ij}) \log (1 - S^y_{ij})$$

where $B$ is batch size. Unlike the fuzzy set cross-entropy loss function, the proposed loss function discards the similarity calculation of the input data and calculates the data similarity in the latent space, named augmentation invariant similarity $\bar{S}^y_{ij}$.

The data similarity in low dimensional latent space $S^z_{ij}$ and the augmentation invariant similarity $\bar{S}^y_{ij}$ are calculated from the distance metric with t-distribution kernel function:

$$\bar{S}^y_{ij} = \kappa(f_\theta \circ g_W \circ T^{-1}(x_i), f_\theta \circ g_W \circ T^{-1}(x_j))$$

$$S^z_{ij} = \kappa(z_i, z_j, v_j)$$

where $\circ$ is the superposition of the operator, $f_\theta \circ g_W \circ T^{-1}(x'_i) = g_W(f_\theta(T^{-1}(x'_i)))$. The $T^{-1}(\cdot)$ is the inverse function of $T(\cdot)$, $x'_i = T(x_i) \rightarrow x = T^{-1}(x')$. So $S^z_{ij}$ assumes the semantics of data $x$ is the same as the data augmentation $T(x)$, thus guiding the network map $x'$ to $T(x')$ to a similar space in high-dimensional latent space. The t-distribution kernel function $\kappa(\cdot, \cdot, v)$ maps the relationship of the input data to the similarity, and the degree of freedom $v$ is related to the number of the dimension.

Finally, the complete loss of the model contains structure-preserving loss $L_{sp}$ and L1 regularization loss $L_T$:

$$L = L_{sp} + \lambda L_T, \quad L_T = \|W\|_1,$$

where regularization loss $L_T$ is constraining the parameter $W$ to be small, and $\lambda$ is hyperparameter. To select a specific feature number, we use the idea of annealing to schedule $\lambda$. We give $\lambda$ a small initial value, then slowly increase $\lambda$ until the number of features satisfies the requirements.

### 4 PERFORMANCE EVALUATION

In this section, we evaluate the proposed DMT-EV by numerical metrics and try to demonstrate the advantages of DMT-EV in three aspects: performance, stability, and time consumption. Several questions of interest are answered in this section to discuss the advantages of DMT-EV.

(Q1) The global structure-preserving performance is the basis of global explanation. Can DMT-EV have better global structure-preserving performance?

(Q2) The local structure-preserving performance is the basis of local and transformation explanation. Can DMT-EV have better local structure-preserving performance?

(Q3) Is the output of DMT-EV’s DR embedding clear and easy to understand, and what are the differences between ours as compared with tSNE, UMAP, and other DR methods?

(Q4) Whether the time consumption of DMT-EV is unbearable?

(Q5) Whether the data augmentation hyper-parameter of DMT-EV is stable?

### 4.1 Datasets, Compared Methods, and Experiments Setup

**Datasets.** Our comparison experiments consist of six image datasets (Coil20$^1$, Coil100$^2$, Mnist$^3$, E-Mnist$^4$, K-Mnist$^5$, and IMAGENET$^6$) and five biological datasets (Activity$^7$, Gast10k$^8$, SAMUSIK10$^9$, and HCL$^{11}$). Following [54], for the

1. https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
2. https://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php
3. https://www.tensorflow.org/datasets/catalog/mnist
4. https://www.tensorflow.org/datasets/catalog/emnist
5. https://www.tensorflow.org/datasets/catalog/kmnist
6. https://image-net.org/
7. https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones
8. https://figshare.com/articles/dataset/MCA_DGE_Data/5435866
9. http://biogps.org/dataset/tag/gastric%20cancer
10. https://github.com/abhiobioinfo/CyAnno
11. http://biv.jzu.edu.cn/HCL/contact.html
The compared methods include four non-parametric methods (tSNE [20], UMAP [33], PHATE [12], PaCAMP [54]) and two parametric methods (Ivis [13], and parametric UMAP) [15, 43]).

All methods map the input data to a 2-dimensional latent space for evaluation to be consistent with the visualization requirements. All models are fitted on the training set (80%) and tested on the test set (20%). The metrics of the test set are introduced to evaluate the model's generalization performance. The output of the 'transform' function in the official code is used to test the generalization performance of the baseline.

Grid Search. The grid search strategy identifies the most suitable parameters from 20 candidate parameters. For tSNE, perplexity = [20, 25, 30, 35] and early exaggeration = [8, 10, 12, 14, 16]. For UMAP, n_neighbors = [10, 15, 20, 25] and min_dist = [0.01, 0.05, 0.08, 0.1, 0.15]. For PacMap, n_neighbors = [10, 15, 20, 25] and min_dist = [0.01, 0.05, 0.08, 0.1, 0.15]. For Ivis, k = [130, 140, 150, 160] and ntrees = [40, 45, 50, 55, 60]. For PHATE, k = [3, 5, 8, 10] and decay = [20, 30, 40, 50, 60]. For DMT-EV, $v^Z = [1e-3, 5e-3, 1e-2, 1e-1, 1e0]$ and $\lambda = [3, 5, 8, 10, 15]$. For DMT-EV: disables the gate layer; all features can be fairly compared with other DR methods through the gate layer. DMT-EV (entifier): contains the gate layer; the number of features passing through the gate layer is set to half of all features. The evaluation results are presented in Tables 1 and 2.

After analyzing the data in Tables 1 and 2, we obtain three conclusions. (1) The proposed DMT-EV outperforms the baseline methods in both classification and clustering. The advantage of clustering performance is even more pronounced, with an average performance that is more than 8% higher. It indicates that DMT-EV makes it easier to discover communities and facilitates a deeper interpretation of the embedded localities. (2) The results of DMT-EV show that the performance of DMT-EV does not degrade even if only half of the features are used. This indicates that the lasso layer of DMT-EV is valid and has redundant features in all datasets. (3) The proposed DMT-EV performs better on data with a large sample size or many features (features). This is due to the powerful fitting generalization ability of the neural network model and stable loss function avoids model collapse.
Global Structure Preservation Performance (K-Means Clustering Accuracy) Comparison on Nine Datasets

| Dataset   | iTSNE | UMAP | PUMAP | Ivis | PHATE | PaCMAP | DMT-EV | DMT-EV (▲) | PUMAP | Ivis | PHATE | PaCMAP | DMT-EV | DMT-EV (▲) |
|-----------|-------|------|-------|------|-------|--------|--------|-------------|-------|------|-------|--------|--------|-------------|
| Coi20     | 75.1  | 79.4 | 69.3  | 57.1 | 70.9  | 77.7   | 89.6   | 89.0        | 72.2  | 61.1 | 73.6  | 89.2   | 89.6   |            |
| Coi100    | 73.9  | 74.0 | —     | 50.8 | 55.6  | 82.3   | 92.1   | 92.5        | —     | 53.5 | 59.0  | 91.3   | 91.3   |            |
| Mnist     | 87.3  | 81.1 | 77.9  | 57.3 | 74.8  | 81.5   | 97.7   | 97.7        | 78.7  | 59.5 | 76.1  | 97.5   | 97.5   |            |
| KMNist    | 66.6  | 68.4 | 71.8  | 64.3 | 64.0  | 64.8   | 73.5   | 73.6        | 71.3  | 64.9 | 65.6  | 73.7   | 74.1   |            |
| EMnist    | 35.8  | 38.2 | 32.1  | 20.3 | 36.4  | 38.7   | 44.9   | 46.1        | 34.8  | 20.8 | 37.1  | 45.3   | 45.4   |            |

Activity 83.4 71.4 85.1 56.3 66.8 70.9 86.3 86.4 83.6 73.9 68.3 85.5 84.2
MCA 49.8 55.7 54.2 29.2 44.9 55.7 58.4 58.1 53.8 31.3 48.0 57.7 57.8
Gast 58.2 67.4 60.3 36.0 61.9 71.2 68.8 65.7 59.9 36.5 59.5 69.3 66.3
SAMUSIK 43.2 52.3 40.3 31.9 62.5 61.8 66.5 53.6 31.3 34.3 29.6 57.9 43.6
HCL 65.4 70.6 67.8 19.8 45.6 73.7 78.9 79.0 67.3 19.7 47.6 77.9 77.9

Average 66.1 67.7 57.9 45.4 59.9 68.4 73.1 73.0 58.5 48.8 62.0 72.9 72.9

Bold denotes that DMT-EV exceeds all baseline methods and Underlined denotes 2% higher than others (no bolded). The ▲ means the model only uses the half number of features (A₂ = n/2).

4.3 Local Structure Preservation Performance (Q2)

Local structure preservation performance is another crucial metric for evaluating DR methods. It measures the degree of damage to adjacent structures during DR embedding. Less local corruption indicates that the DR model preserves the local structure of input data much better. For explainability analysis, excellent local preservation performance is suitable for discovering and evaluating the locally essential features. We use an order-based metric (mean relative rank error, RRE) to evaluate how well the local structure of embedding data matches the input data. The RRE measures the average changes in neighbor ranking between two spaces:

$$RRE = \frac{(MR_{k}^{(i,j)} + MR_{k}^{(j,i)})}{2}$$

where $k = 10$ in this paper, and

$$MR_{k}^{(i,j)} = T_{RRE} \sum_{i=1}^{M} \sum_{j \in N_{i,k}^{(i)}} \frac{|r_{ij}^{(i)} - r_{ij}^{(j)}|}{r_{ij}^{(i)}}$$

$$MR_{k}^{(j,i)} = T_{RRE} \sum_{i=1}^{M} \sum_{j \in N_{i,k}^{(j)}} \frac{|r_{ij}^{(j)} - r_{ij}^{(i)}|}{r_{ij}^{(j)}}$$

where $T_{RRE}$ is the normalizing term. The $r_{ij}^{(i)}$ is the rank of $x_{j}$ in the k-NN of $x_{i}$. $M$ is the size of the dataset. $N_{i,k}^{(i)}$ is the set of indices to the k-NN of $x_{i}$, $k = 10$ for all the datasets.

4.4 Time Consumption Comparison (Q3)

A stereotype is that parametric methods based on neural networks require expensive computational resources to complete training and are, therefore, unsuitable for visualization tasks. However, with the development of computing hardware and neural network theory, training neural networks have become fast and inexpensive. Sometimes, the efficiency of processing large datasets exceeds that of parametric-free methods. We select several representative works have become fast and inexpensive. Sometimes, the efficiency of processing large datasets exceeds that of parametric-free methods. We select several representative...
datasets to evaluate the training time. All tests are run in a container with one GPU and 12 cores, and we do not design additional acceleration schemes for algorithms that cannot be accelerated using GPU. The time consumption of different datasets is shown in Table 5.

After analyzing the data in Table 5, we arrived at two conclusions. (1) For a small dataset, parametric-free methods take less computing time. But for middle or large dataset, neural-network-based DMT-EV show comparable or higher efficiency. (2) The time complexity of DMT-EV network training consists of two main features: the initialization and model training features. For the initialization part, DMT-EV is the same as UMAP and requires the construction of a neighbor graph for each data node with a computational complexity of $O(M^{1.14})$ [21] for each node’s $k$-NN estimate. In the model training phase, the computational complexity of DMT-EV training with small batches is $O(BM)$, where $B$ is the batch size, which is much lower than $O(M^2)$, allowing DMT-EV training on massive datasets.

### 4.5 Visualization Performance Comparison (Q4)

The visualization section illustrates the reasons for the metric performance advantage by showing a 2-dimensional visualization. The visualization comparison on COIL-100, Mnist and HCL dataset are shown in Figs. 2, 3 and 4.

![Fig. 2. Visualization Comparison with baseline DR methods on COIL100 dataset. The black circles indicate the regions where the structure is broken, and DMT-EV has better embedding output.](image)

![Fig. 3. Visualization Comparison with baseline DR methods on Mnist dataset. The black circles indicate the regions where the structure is broken, and DMT-EV has better embedding output.](image)

![Fig. 4. Visualization Comparison with baseline DR methods on HCL dataset. The black circles indicate the regions where the structure is broken, and DMT-EV has better embedding output.](image)

The COIL-100 image dataset contains 100 toroidal manifolds. As shown in Table 1, the parameter-free methods outperform the parameter-based methods. Among the parameter-free methods, tSNE is good at local structure-preserving, so less manifold overlap exists. However, tSNE performs poorly on global structure-preserving as we cannot visually form judgments about inter-cluster relations according to the embedding (as shown in the black circle in Fig. 2). UMAP and PaCMAP, on the other hand, tend to preserve the global structure. UMAP and PaCMAP embed the manifold into local latent space and successfully preserve the inter-manifold relationship. Still, some manifolds are overlapped (as shown by the black circle in Fig. 2). We observe that the proposed DMT-EV outperforms all parametric models and has the potential to be the best in terms of both local and global performance. Furthermore, unlike PUMAP, DMT-EV avoids more manifold overlapping; DMT-EV has more pronounced cluster boundaries; DMT-EV has better generalization performance on the test set.
The Mnist image dataset is an extensively used dataset of handwritten digits (Fig. 3). Similar conclusions to COIL100 are available in the Mnist data. In addition, we focus on the overlapping phenomenon of the orange cluster (Digits 4) and the yellow cluster (Digits 9). The embedding output of UMAP and PaCMAP do not have distinguished cluster boundaries for orange and yellow clusters, affecting clustering and classification performance. However, the embedding of tSNE indicates that these two clusters can be distinguished, ruling out the possibility of unclear original data. The above evidence illustrates that DMT-EV can better distinguish different clusters and has good global performance. The same conclusions are obtained from the analysis of overlapping areas of the pink cluster (Digits 8), blue cluster (Digits 3), and purple cluster (Digits 5).

The detailed experimental results are shown in Figs. 5 and 6. The performance of biological data is shown in Fig. 4. DMT-EV also outperforms the comparative methods in the biological datasets and produces minimal overlap. For a more comprehensive comparison, the comparison of reduced vectors on five image datasets and five biological datasets are shown in Figs. 6 and 5.

5.1 Global Explanation
In global explanation, the indispensable features are identified, thus removing noisy features and improving the system’s stability. DMT-EV discovers the important features for the whole DR embedding without additional analysis, thanks to the proposed lasso network for DMT-EV. Specifically, the confrontation of structure-preserving loss $L_{SP}$ with regularization loss $L_{R}$ in the lasso layer enables DMT-EV to retain essential features and remove redundant features. The regularization loss $L_{R}$ guide $W$ reduction until the corresponding feature is discarded (unable to pass through the gate layer). The structure-preserving loss $L_{SP}$ resists the discarding of crucial features. The result of balancing these two losses is that the essential features are retained while redundant features are excluded from the forward propagation of the network. The lasso network decouples the feature importance parameter from other network parameters, allowing a transformation of $W$ to yield the global importance of the network $I_f$.

$$I_f = \{I^f_1, I^f_j = W_j/\max(W)\}, \quad (8)$$

where $f \in \{1, 2, \ldots, n\}$, $W$ is defined in Eq. (3). $I^f_j$ is the global importance of the feature $f$. The lasso network prevents the forward propagation of trivial features, and this can help solve the information leakage problem.

Fig. 7 demonstrates a global explanation of the Mnist dataset. DMT-EV discovers 196 essential features and 588 unimportant features in the Mnist dataset. In Fig. 7a, DMT-EV discards trivial features during training, $E$ is the epoch, and $F$ is the number of essential global features. The leftmost image is the initialized state of the network, where all features are retained. The rightmost image is the ended state, where only the selected features are shown. The gray background means that pixels (features) are discarded. Unimportant features are constantly discarded during the training process. We observe that different numbers can be easily distinguished based on 196 discovered features, suggesting that the global explanation does retain essential features for identifying numbers. Fig. 7b indicates that the discovery of globally important features is smooth, and we only need to set a suitable termination condition $A_j$ to obtain the required number of features. Fig. 7c shows the importance of the discovered features for in-depth exploration.

5.2 Local Explanation
In local explanation, the features critical to the particular local area of DR output are identified to help the users understand the behavior of the DR model. The local area is defined by the user or clustering algorithm, implying that these features are required to preserve the data skeleton of the local area.

Differing from the global explanation, the local explanation focuses on the difference in features between the data in a particular local area of DR output and the data in the other areas of DR output. In other words, we analyze why the DR model maps the data to a particular local area. Essentially, we assume the DR model maps the data into a different local area because these data have a distinct pattern dissimilar from the neighboring cluster. For example, some unique mRNAs specifically indicate a specific cell type. The data of these cell types are mapped to a particular local area by the DR model. The local explanation aims to
Fig. 5. Visualization comparison with baseline DR methods in the unsupervised case. The dataset includes five image datasets.
Fig. 6. Visualization comparison with baseline DR methods in the unsupervised case. The dataset includes five biological datasets.
discover the above unique mRNAs by analyzing the DR model and DR output.

An essential precondition of local explanation is the definition of each local area. In addition to defining the local area through human-machine interaction, we can also use a clustering algorithm. This paper uses the K-means clustering algorithm, a simple linear cluster, to distinguish the local area. Table 2 shows that the clustering performance of DMT-EV is significantly better than the baseline methods. It makes the cluster-based local analysis more appropriate. As evidence, the clustering performance comparison on the Mnist dataset is shown in Fig. 8, whereas the red hexagonal star is the cluster center C k. We observe that the cluster obtained by DMT-EV are consistent with the category labels.

We expect to obtain explainability while maintaining excellent performance. Inspired by the saliency map [45], we measure the local importance of features by analyzing the effect of changes in input features on the clustering similarity P x in the embedding space.

$$I'_{c,f} = \{I'_{c,1}, \ldots, I'_{c,j}, \ldots, I'_{c,n}\}$$

$${I'_{c,f}} = \frac{1}{M} \sum_{m=1}^{M} \left[ \frac{P_{c}(x^+ f) - P_{c}(x^- f)}{\tau(x') - x^f} \right]$$

where $I'_{c,f}$ is the local explanation results of feature f for cluster c. The $P_{c}(x^+ f)$ is a item of $P(x^+ f)$ for cluster c. To measure the impact of changes in the features of input data on the embedding output, we define

$$P_x = P(x) = \text{Softmax}\left( \{S_{x^f}\}_{f=C1, \ldots, CK} \right)$$

where $C_1, \ldots, C_K$ are the cluster centers, $K$ is the number of clusters.

The $x^+ f$ is the positive augmentation of feature f, and $x^- f$ is the negative augmentation of feature f:

$$x^+ f = \{r(x^1), \ldots, r(x^{f-1}), x', r(x^{f+1}), \ldots, r(x^n)\}$$

$$x^- f = \{r(x'), \ldots, r(x^{f-1}), r(x^f), r(x^{f+1}), \ldots, r(x^n)\}$$

where augmentation operator $r(\cdot)$ is defined in Eq. (1).

Next, we illustrate the local explanation with examples from the Mnist data (in Fig. 9). We analyzed the locally important features of digit 0 (brown color, C0) and digits 3 (gray color, C3). We first display the local area to be explained with a red circle and randomly select three images in the local area for display (yellow border). Then, we display one image from each potential adversarial area (grey border). Finally, we show the results of the local explanation (red border) for readers to compare. We have marked the critical features of the local explanation with a purple ellipse. The features of the image in the target local area have lower values, while features of the image in the surrounding adversarial images have higher values. At the same time, these features are the brightest part of the local explanation, indicating that these features are essential to distinguish the local area. It proves the validity of local explanation.

5.3 Transformation Explanation

In transformation explanation, the features critical to the interaction of the local areas of DR output are identified to help the users understand the behavior of the DR model. Local explanation finds features that identify whether they are embedded in a particular locality, while transformation explanation finds features that transform from one
locality to another specific locality. It is analogous to the marker discovery of cellular transformation processes.

The transformation explainability from cluster $c_1$ to cluster $c_2$ is defined as:

$$I_{t_{c_1,c_2}} = \sum_{f} I_{t_{c_1,c_2};f}$$

where $I_{t_{c_1,c_2};f}$ is the transformation importance of features $f$. It is calculated from the similarity distance of cluster $c_1$ and $c_2$.

Next, we illustrate the transformation explanation analysis with examples from the Mnist dataset. The transformation explanation example of $c_3 \rightarrow c_8$, $c_4 \rightarrow c_9$, $c_3 \rightarrow c_3$ on the Mnsit dataset is shown in Fig. 10. First, we display the images randomly selected from clusters (green border and orange border) and calculate the different pixels of the images (yellow border), followed by comparing the disparity pixels with the generated transformation explanation.

6 INTERFACE OF DMT-EV

This section introduces the interface of DMT-EV. Then, to facilitate in-depth analysis and exploration, we design an DMT-EV-interface to display the embedding and interpret the results (in Fig. 11). DMT-EV-interface consists of datasets & parameters control panel, main view, and explanations view.

The datasets & parameters control panel enables users to select datasets and basic parameters. Users can analyze and validate the built-in data in this panel or upload their dataset and download the analysis results. In addition, users can experiment with DMT-EV and explainable analysis parameters for best results by adjusting the settings in the panel. The main view displays the reduced vector, clustering results, and transformation routes. Users can control the displayed content, color, and scatter size through the drop-down box. The DMT-EV interface presents the embedding output of different pixels with the generated transformation explanation. The highlighted region in the transformation explanation corresponds precisely to the disparity part. It proves the validity of the local explanation. At the same time, we compare the differences between local and transformation explanations. The local explanation is a combination of all possible transformation explanations. For example, the sum of the transformation explanations $I_{t_{c_3,c_5}}$ and transformation explanations $I_{t_{c_3,c_8}}$ is much closer to local explanations $I_{t_{c_3}}$. 

![Fig. 9. Local explanation of Mnist dataset. We first display the local area to be explained with a red circle and randomly select images in the local area for display (yellow border). Then, we display one image from each potential adversarial cluster (green border). Finally, we show the results of the local explanation (red border) for the readers to compare.](image1)

![Fig. 10. Transformation explanation of Mnist dataset. First, we display the images randomly selected from clusters (green border and orange border) and calculate the different pixels of the images (yellow border), followed by comparing the disparity pixels with the generated transformation explanation.](image2)

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
methods through the ‘plotly’\textsuperscript{16} interactive interface. The ‘plotly’ interface can display data for particular class tags individually to find outliers and analyze the advantages of different methods. The explanations analysis view provides users with the distribution of the selected dataset and an in-depth analysis of the DR results. Users can choose a local area in the main view and execute the explanation analysis program online to obtain the corresponding results. Three different explanation analyses are available by interacting in the interface. Furthermore, we can verify the validity of explainable analysis by observing the input data’s heatmap.

With DMT-EV-interface, users can focus on exploring the knowledge embedded in the data without needing a deep understanding of deep learning and an explanation of deep learning methods. We can better understand the algorithm with the interface rather than looking at static images. First, we explore the suitability of the DR method for outlier processing by looking at data with different labels individually. Second, we cross-reference each node of embedding through an interactive interface. Next, based on the interactive interface, we intuitively understand the impact of DMT-EV’s hyperparameters on the analysis results. Finally, users are free to choose the local area, not limited to the fixed area provided by K-means, and the analysis results are more targeted. All results in the case study (from Figs. 13, 14, and 15) are the results of DMT-EV-interface analysis.

7 Qualitative Evaluation

7.1 Case Study
This section discusses the effectiveness of DMT-EV and explanation analysis through a case study on biological datasets.

We discuss the validity of this paper by comparing the empirical rules with the explanation results of the Thyroid Nodule Proteome (TNP) dataset and assess the positive impact of DMT-EV explanations on biological knowledge discovery.

The TNP dataset includes 579 data samples containing 3707 protein enrichment features. Characters represent the names of the proteins (e.g., ‘RPS27’). TNP data is a challenging problem for traditional DR methods because of the large number of noise features and irrelevant features in the data, which are closer to real-world data. Fortunately, the collector of the data provides partial labeling information. For such a problem, DMT-EV can introduce prior knowledge by data augmentation designed by Eq. (1) to complete the semantic embedding. Also, the lasso layer of DMT-EV eliminates useless features in the training process, thus improving the stability and concentrating on essential features. DMT-EV is compatible with unsupervised, supervised, and semi-supervised setups.

![Fig. 11. DMT-EV interface is a visual analytics tool that helps users understand reduced vector and analysis explanation output. DMT-EV interface consists of datasets & parameters control panel, the main view, and explanations analysis view.](image)

![Fig. 12. (a) DR embedding of the TNP dataset. (b) K-means clustering the TNP dataset. C0: malignant samples, C1: tumor samples (intermediate stage of malignant and cancer), and C2: cancer samples.](image)
The reduced vector on the TNP dataset is shown in Fig. 12a, with the blue node representing benign samples and the orange node representing malignant samples. The K-means clustering results are shown in Fig. 12b. All data are clustered into three clusters: malignant samples, tumor samples (intermediate stage of malignant and cancer), and cancer samples. We observe a smooth transition from benign to malignant in the embedding results of DMT-EV, indicating that the model finds practical features to distinguish between the two semantics. Next, we perform an explainable analysis based on the model and embedding results to discover practical knowledge.

Global Explanations. DMT-EV identifies 21 features important for semantic embedding from 3707 features based on the parameters of the lasso layer. The detailed importance of these features is shown in Fig. 13a.

To verify the biological significance of the discovered features, we enrich 21 features (proteins) with Ingenuity Pathway Analysis software (version 7075971). The 21 discovered biomarkers are associated with two protein networks, one associated with cancer and developmental disorder (includes LAMB2, VIM, LGALS3BP, DDB1, RPS27t; RPDX1, PSMD2, APOA1, MATNA, PRDX5, SUCGL2, AK1, FUCAI) and the other with immunological disease (includes CTBP1, SNX5, GTF2I, ANXA3, CLINT1, AP1B1, H1-10, DCXR). We merge the two networks into one by deleting all indirect effects and matching discovered features and domain knowledge indicates that DMT-EV can find critical protein markers associated with the target disease.

Local Explanations. Next, we analyze the important features for embedding the local areas by local explanations. We start with cluster analysis using K-means and then analyze the significant features of each cluster based on data augmentation. As shown in Fig. 14, different clusters favor different features meaning that different cancer cycles correspond to marker proteins. In addition, some critical features, such as ‘RPS27t’, are given high importance by all three clusters. It may indicate that ‘RPS27t’ means the entire process of thyroid cancer development.

To illustrate the effectiveness of the discovered features in local explanations, we design experiments based on newly collected test data. Similar to one-versus-rest classification, we use the features found in the local explanation of DMT-EV to identify the corresponding clusters. The results are shown in Table 7. We find that the top-ten importance features (DMT-EV (I\(_i\)-I\(_c\)-I\(_s\)-I\(_s\))) have the highest average accuracy in classification for the corresponding clusters. The sub-optimal performance is the global optimum of 21 features found by DMT-EV, since using all features as input would incorporate too much noise to affect the overall performance. This experiment demonstrates the validity of the local explanation, as it can identify features that are important for discriminating the local and assign high importance.

Transformation Explanations. Finally, we explore the essential features of transferring from one local area to another. The local explanation is confined to a local area but cannot explore the transfer process with direction. The pathway from C0 (benign samples) to C2 (cancer samples) are shown in Fig. 15a. The importance of the features I\(_t\)-I\(_c\)-I\(_s\) discovered by the transformation explanation is shown in Fig. 15b. When comparing the two images, we find that the essential features correspond to a long step in the semantic space.

To illustrate the effectiveness of the discovered features in transformation explanations, we design experiments based on newly collected test data. Similar to one-versus-one classification, we use the features found by transformation explanations to identify the corresponding clusters (in Table 8). The top-ten importance features (DMT-EV (I\(_t\)-I\(_c\)-I\(_s\))) have the highest average accuracy in classifying the corresponding clusters. The sub-optimal performance is the global optimum of 21 features found by DMT-EV, since
using all features as input would incorporate too much noise to affect the overall performance.

7.2 Expert Interview

We conducted expert interviews to verify further the effectiveness of recommended design choices and the correctness of generated rules.

Tasks. In our expert interviews, we designed three tasks to evaluate the quality of DMT-EV-interface. Task (a). The experts are presented with 11 datasets and the corresponding visualizations of DMT-EV and five baseline methods (tSNE, UMAP, PaCAMP, and Ivis). They were asked to score each visualization to determine whether it maintains the local structure, holds the global structure, has explicit embedding, forms clusters, and has outliers. Each question was scored on a scale from 0 to 1. The sum of the five questions was scored on a scale from 0 to 1.

Table 1: Expert Evaluation of the Reasonableness of the Explanations Generated by DMT-EV

| Experts in CS | Experts in PG | Experts in BI | Average |
|---------------|---------------|---------------|---------|
| Global Exp.   | 4.4(±0.6)     | 4.9(±0.3)     | 4.1(±0.4) |
| Local Exp.    | 4.7(±0.3)     | 4.3(±0.5)     | 4.4(±0.3) |
| Transformation Exp. | 4.0(±0.2) | 4.7(±0.1) | 4.2(±0.2) |

Using the Friedman Nemenyi statistical test for multiple comparisons of the experts’ scores, the average score of DMT-EV (4.8) was statistically different from other visualization methods (p = 0.028 < 0.05), implying that DMT-EV is significantly better than other baseline methods.

Fig. 16. Results of Nemenyi Test. The numerical axis marks the average rank performance of different models (rank = 1 means best). The bolded horizontal line under the numerical axis marks the models that cannot be judged as having statistical differences. The critical difference (CD) of Friedman statistic = 10.82 represents statistically significant differences in the evaluation results of several visualization methods. The p-value = 0.028(< 0.05), implying that DMT-EV is significantly better than other baseline methods.

Participants and Procedure. We invited 30 researchers (10 females, age_mean = 29.2, age_std = 4.93) major in computer science (10/30), proteomics&genomics (10/30), and bioinformatics (10/30). The length of an expert interview was about 1 hour. Before starting the interview, we collected the experts’ consent for collecting their feedback. Each expert interview started with a 5-min brief introduction to our entire project. After that, experts were asked to finish the three tasks. For Task (a) and Task (b), to ensure that experts provide effective feedback for each question, they were only allowed to submit their answers on each dataset after 10 seconds. After finishing all three tasks, experts were asked to provide general comments on our approach, including the advantages and disadvantages.

Feedback on the DMT-EV. The rules with the highest average scores are shown in Tables 9 and 10.

Overall, our DMT-EV is appreciated by experts. An expert comments that ‘some features are inspiring’, and another says that ‘the visualization results are clear and easy to understand’. We observed that although experts in different fields do not fully agree on the visualization tools. But DMT-EV received positive comments from all experts. Meanwhile, the affirmation of the explanation by DMT-EV by experts in biology and bioinformatics experts indicates that DMT-EV can indeed select essential features. To further demonstrate the advancement of DMT-EV in task a from a statistical perspective. We evaluate DMT-EV and several baseline methods using the Friedman Nemenyi Statistical Test[10] (in Fig. 16). The critical difference (CD) of Friedman statistic = 10.82 represents statistically significant differences in the evaluation results of several visualization
methods. The p_value = 0.028( < 0.05) implies that DMT-EV is significantly better than other baseline methods.

8 Conclusion

We propose DMT-EV, a deep learning-based parametric approach to achieve dimension reduction embedding and three different levels of explanation. It can effectively capture visual knowledge from data structures and semantics and provide users with meaningful rules to help them understand the recommendation results. Furthermore, the data augmentation and the proposed loss function improve the performance and stability of DMT-EV. As a result, the DMT-EV outperforms the SOTA techniques in local and global performance, although DMT-EV is considered the more difficult parametric model to optimize. We designed three levels of descriptive analysis to explain the parameters and knowledge learned by the model and explore the components (features) important for dimensional reduction embedding. Then, based on DMT-EV, we proposed DMT-EV-interface, an interactive visualization tool for facilitating in-depth analysis and exploration. DMT-EV-interface presents the embedding output and analysis results of DMT-EV in an interactive interface. It also deepens the user’s understanding through various graphs and charts.

Acknowledgments

The authors thank Linyan Lu for helping to improve the writing of the article. Zelin Zang and Shenghui Cheng contribute equally to the article.

References

[1] P. Adler et al., “Auditing black-box models for indirect influence,” KDD, Inf. Syst., vol. 54, pp. 95–122, 2018.
[2] T. S. Andrews, V. Y. Kiselev, D. J. McCarthy, and M. Hemberg, “Tutorial: Guidelines for the computational analysis of single-cell RNA sequencing data,” Nature Protoc., vol. 16, pp. 1–9, 2021.
[3] M. Aupetit, “Visualizing distortions and recovering topology in continuous projection techniques,” Neurocomputing, vol. 70, pp. 3043–3050, 2007.
[4] A. C. Belkin, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, “Automated optimized parameters for Tdistributed stochastic neighbor embedding improve visualization and analysis of large datasets,” Nature Commut., vol. 10, pp. 5415–5415, 2019.
[5] A. Chajetimpampas, R. M. Martins, and A. Kerren, “t-viSNNE: Interactive assessment and interpretation of t-SNE projections,” IEEE Trans. Vis. Comput. Graph., vol. 26, no. 8, pp. 2696–2714, Aug. 2020.
[6] C. Chen, J. Hou, J. J. Tanner, and J. Cheng, “Bioinformatics methods for mass spectrometry-based proteomics data analysis,” Int. J. Mol. Sci., vol. 21, 2020, Art. no. 2873.
[7] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple and powerful parameter-free moment matching loss,” Proc. ACM Symp. Eye Tracking Res. Appl., 2020, Art. no. 12.
[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Dec. 2015, arXiv:1512.03385.
[9] G. E. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2003, pp. 857–864.
[10] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, 2006.
[11] D. Kobak and G. C. Linderman, “UMAP does not preserve global structure any better than t-SNE when using the same initialization,” 2019, doi:10.1101/2019.12.19.877522.
[12] J. B. Kruska, “Nonmetric multidimensional scaling: A numerical method,” Psychometrika, vol. 29, no. 2, pp. 115–129, 1964.
[13] A. Kumar, P. Howlader, R. Garcia, D. Weiskopf, and K. Mueller, “Challenges in interpreting parameter regression t-SNE data,” in Proc. ACM Symp. Eye Tracking Res. Appl., 2020, Art. no. 12.
[14] D. Lähnemann et al., “Eleven grand challenges in single-cell data science,” Genome Biol., vol. 21, pp. 31–35, 2020.
[15] J. A. Lee and M. Verleysen, “Quality assessment of dimensionality reduction: Rank-based criteria,” Neurocomputing, vol. 72, pp. 1431–1443, 2009.
[16] M. Lin and H. Zha, “Riemannian manifold learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 5, pp. 796–809, May 2008.
[17] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” Neural Comput., vol. 15, no. 6, pp. 1373–1396, 2003.
[18] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, “Visualizing high-dimensional data: Advancing in the past decade,” IEEE Trans. Vis. Comput. Graph., vol. 23, no. 3, pp. 1249–1268, Mar. 2017.
[19] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, arXiv:1711.05101.
[20] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2017, pp. 4768–4777.
[21] N. V. D. Maaten, “Learning a parametric embedding by preserving local structure,” in Proc. Int. Conf. Artif. Intell. Statist., 2009, pp. 384–391.
[22] N. V. D. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. Nov., pp. 2579–2605, 2008.
[23] L. McNes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” Sep. 2018, arXiv:1802.03426.
[24] L. McNes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” Feb. 2018, arXiv:1802.03426.
[25] C. Molnar, “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book
[26] K. R. Moon et al., “Visualizing structure and transitions in high-dimensional biological data,” Nature Biotechnol., vol. 37, pp. 1482–1492, Aug. 2019.
[27] M. Moor, M. Horn, B. Rieck, and K. Borgwardt, “Topological autoencoders,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7045–7054.
[28] L. G. Nonato and M. Aupetit, “Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment,” IEEE Trans. Vis. Comput. Graph., vol. 25, no. 8, pp. 2650–2673, Aug. 2019.
[29] L. G. Nonato, T. Horiguchi, C. van Gemert, B. Pf. Leieventd, E. Eisemann, and A. Vilanova, “DeepEyes: Progressive visual analytics for designing deep neural networks,” IEEE Trans. Vis. Comput. Graph, vol. 24, no. 1, pp. 98–108, Jan. 2018.
[30] K. Puolamaki, A. Jokullond, and J. Mäkelä, “SISEMAP: Explainable dimensionality reduction,” 2022, arXiv:2201.04455.
[31] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” 2016, arXiv:1602.04338.
Zelín Zang received the MEng degree from the Zhejiang University of Technology. He is currently working toward the PhD degree with Westlake University, supervised by Stan Z. Li. His research interests include manifold learning, dimension reduction, and applications on biological and image data.

Shenghui Cheng received the PhD degree in computer science from Stony Brook University. He is a Westlake fellow with Westlake University. He held research positions with the Institute of Medical Research, University of Leipzig, Germany, Brookhaven National Lab, and Harvard Medical School.

Hanchen Xia received the BEng degree from the Zhejiang University of Technology. He is currently working toward the master degree with Shanghai Jiaotong University. His work in this article was completed during his visit to Westlake University.

Yaoting Sun received the PhD degree in proteomics from the Laboratory of Proteomic Big Data, Westlake University. She learned clinical pharmacy (2010-2015) and pharmacology (2015-2018) with the College of Pharmacy, Dalian Medical University. She is currently a postdoc with Westlake University. She was awarded National Scholarship and Outstanding Graduates in Hangzhou and Dalian.

Yongjie Xu received the BS and MS degrees from the Beijing University of Chemical Technology (BUCT), Beijing, China, in 2017 and 2020, respectively. He is currently working toward the PhD degree with Zhejiang University (ZJU) and Westlake University, Hangzhou, China, supervised by Chair Professor Stan Z. Li. His research interests include transfer learning, manifold learning, and its applications in biological and image data. He has published several papers in journals or conferences such as AAAI etc.

Lei Shang received the MEng degree from Zhejiang University, in 2016. He joined Alibaba Group as a computer vision algorithm researcher in 2018. His research interests include deep metric learning, semi-supervised learning, and computer vision applications.

Baigui Sun received the MEng degree from Zhejiang University, Hangzhou, China, in 2014. He joined Alibaba Group as a computer vision algorithm researcher in 2014. His research interests include face detection/recognition, deep metric learning, domain adaption, zero/few-shot learning, semi-supervised learning, and computer vision applications. He has published more than ten articles in international journals and conferences.

Stan Z. Li (Fellow, IEEE) received the BEng degree from Hunan University, Changsha, China, in 1982, the MEng degree from the National University of Defense Technology, Changsha, in 1985, and the PhD degree from the University of Surrey, Guildford, U.K., in 1991. He joined Westlake University, Hangzhou, China, as a chair professor of artificial intelligence in February 2019. He has published more than 400 articles in international journals and conferences, with more than 50,000 Google Scholar citations.

Lianguy Li currently working toward the BE degree with Shanghai Maritime University.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdli.