SpaceMAP: Visualizing High-dimensional Data by Space Expansion

Xinrui Zu [1] **Qian Tao** [1]

Abstract

Dimensionality reduction (DR) of high-dimensional data is of theoretical and practical interest in machine learning. However, there exist intriguing, non-intuitive discrepancies between the geometry of high- and low-dimensional space. We look into such discrepancies and propose a novel visualization method called Space-based Manifold Approximation and Projection (SpaceMAP). Our method establishes an analytical transformation on distance metrics between spaces to address the "crowding problem" in DR. With the proposed equivalent extended distance (EED), we are able to match the capacity of high- and low-dimensional space in a principled manner. To handle complex data with different manifold properties, we propose hierarchical manifold approximation to model the similarity function in a data-specific manner. We evaluated SpaceMAP on a range of synthetic and real datasets with varying manifold properties, and demonstrated its excellent performance in comparison with classical and state-of-the-art DR methods. In particular, the concept of space expansion provides a generic framework for understanding nonlinear DR methods including the popular t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).

1. Introduction

Real-world data, including images, videos, genetic expressions, natural languages, etc., commonly have a high ambient dimension (i.e. number of pixels or measurements). The intrinsic dimension of these data, however, is typically much lower, which fact is recognized as a fundamental reason for modern machine learning to work well (Levina & Bickel, 2004; Pope et al., 2021; Wright & Ma, 2022). Effective and accurate dimensionality reduction (DR) to discover the intrinsic low-dimensional data structure is of significant value for machine learning research.

DR is also essential for data visualization, which provides valuable intuition for researchers from different disciplines dealing with high-dimensional data. A space with a dimensionality higher than 3, however, is already beyond our accustomed way of seeing data. Here we enumerate a number of challenges in visualizing high-dimensional data onto a drastically reduced dimension (2 or 3):

Firstly, our intuition of geometry in 2- or 3-dimensional space often does not generalize to high-dimensional space. Theories on high-dimensional geometry revealed a number of intriguing, non-intuitive phenomena (Giraud, 2021), one of which is "concentration on crust". Imagine a ball with a radius r in a d-dimensional Euclidean space. Consider a "crust" of the d-dimensional ball, which is between the surfaces of this ball and a slightly smaller concentric ball with radius $(1-\epsilon)r$, where ϵ is small (Fig.1 a.1). The ratio of the volume of between the "crust" $C_d(r)$ to the ball is $V_d(r)$ is $\frac{C_d(r)}{V_d(r)} = 1-(1-\epsilon)^d$. Take $\epsilon = 0.01$, it is intuitive to see that the ratio is tiny with small d, however this ratio grows fast to near 100% when d increases, as illustrated in Fig.1 a.2. The mass of a high-dimensional ball is therefore counter-intuitively concentrated on a crust. Such concentration underlies the "crowding problem" of DR (van der Maaten & Hinton, 2008): a faithful preservation of distances of high-dimensional space would lead to crowded data points in low-dimensional space (all pairwise distances become similar as shown in Fig.1 a.3). This suggests that distances need to be defined *differently* in high- and low-dimensional spaces to mitigate the crowding.

Secondly, many real-world high-dimensional datasets exhibit hierarchical structures, with sub-manifolds on large manifolds, governed by the underlying data generation process (Bengio, 2013; Abdolali & Rahmati, 2019). The state-of-the-art DR methods, such as Barnes-Hut t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten, 2014) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), commonly consider a restricted $k-$nearest neighborhood, hence part of the manifold. Isomap (Tenenbaum et al., 2000), on the other hand,

*Equal contribution [1]Department of Imaging Physics, Delft University of Technology. Correspondence to: Xinrui Zu <X.Zu-1@tudelft.nl>, Qian Tao <Q.Tao@tudelft.nl>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

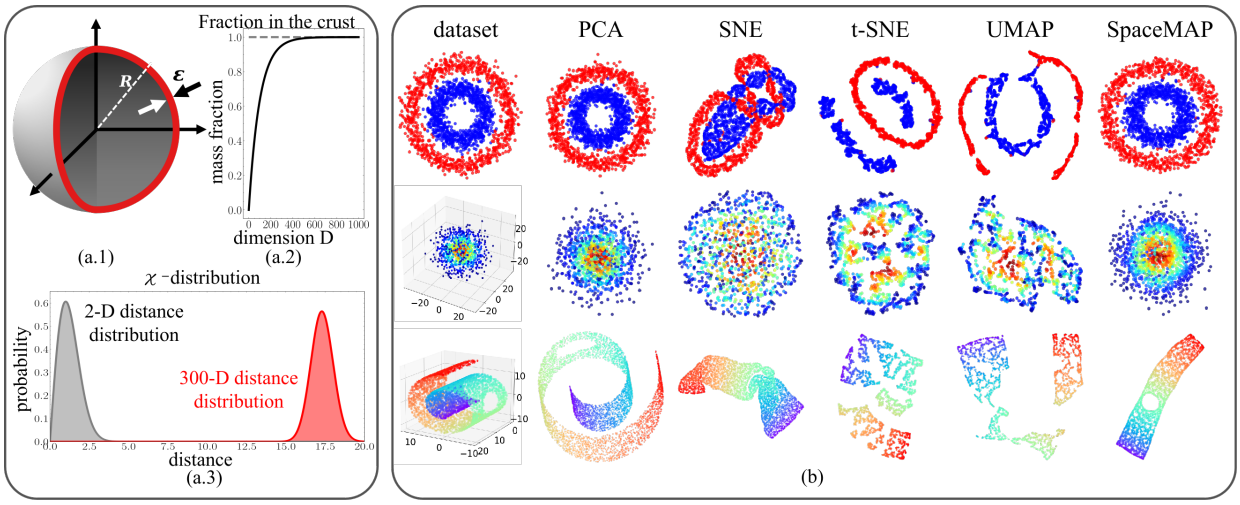

Figure 1. Left panel a: counter-intuitive high-dimensional geometry. (a.1) The "crust" of a ball in 3-dimensional spaces with Euclidean distance metric. (a.2) The fraction of volume between crust and ball, as a function of the dimensionality of the ball. As the dimensionality grows, the mass concentrates on the crust. (a.3) Distributions of the pairwise distances in a 2-dimensional (gray) and 300-dimensional (red) spaces. Right panel b: visualization results of three synthetic datasets (2D concentric rings, 3D Gaussian distributed point cloud, and Swiss roll with a hole), by PCA, SNE, t-SNE, UMAP, and the proposed SpaceMAP.

takes the entire manifold into account by inferring geodesic distances on a connected graph. Empirically, t-SNE and UMAP work well on data of disjoint manifolds such as MNIST and ImageNet, while Isomap works on data of continuous manifold such as Swiss roll dataset. It remains challenging, nevertheless, for one method to succeed in both disjoint and continuous data manifolds.

A third concern is related to the first observation, which shows that to mitigate the "crowding problem" one must define distances differently in high- and low-dimensional space. This essentially means a geometrical *distortion* between two spaces. By using a "heavier tail" similarity function in low-dimensional space, t-SNE and UMAP realize such distortion implicitly. However, the underlying distortion of distance metric has never been analytically expressed or validated. The implicity in turn deters us from integrating prior information on data when it is known. For example, for the Swiss roll dataset, we know *a priori* that the intrinsic dimensionality of data is 2. Even for real-world datasets such as MNIST or ImageNet, a meaningful estimation of the intrinsic dimensionality can often be made (Pope et al., 2021; Ansuini et al., 2019).

We strive to tackle the three challenges by the proposed Space-based Manifold Approximation and Projection (SpaceMAP) method. We made three contributions:

- *Space Expansion:* We explicitly define, and match the "capacity" of high- and low-dimensional space, by analytical transformation of distances, and show that this transformation is justified by previous theory of intrinsic dimensionality estimation;
- *Hierarchical Manifold Approximation:* We propose dataset-specific, hierarchical modeling of similarity, to accommodate both disjoint and continuous manifolds.
- *Prior Information for DR:* The hyper-parameters of SpaceMAP are explainable, and amendable to integrating data prior.

2. Related Work

2.1. Mini Review

We follow the categorization in van der Maaten et al. 2009 and divide DR method into full spectral and sparse spectral methods.

Full Spectral Methods: Full spectral methods are based on eigen-decomposition of a completed matrix that captures the covariances between dimensions or pairwise similarities among all data points. The most well-known method is the principal component analysis (PCA) (Pearson, 1901), which finds a low number of principle directions that account for the largest variations in data. PCA also has nonlinear versions called kernel PCA (Ross et al., 2007; Tipping & Bishop, 1999; Zou et al., 2006). Alternatively, the similarity matrix can be based on geodesic distance (Isomap, landmark Isomap) on a manifold (Tenenbaum et al., 2000; Chen et al., 2006). Maximum variance unfolding (MVU) (Weinberger & Saul, 2006) also belongs to this category of methods.

Sparse Spectral Methods: Different from full spectral

methods, sparse spectral methods only focus on retaining local structure of data and solve a sparse eigen-problem. In general, preserving only local properties allows for better embedding of nonlinear manifolds (van der Maaten et al., 2009). Important work in this category include locally linear embedding (LLE), modified LLE, Hessian LLE (HLLE), Laplacian eigenmaps(LE), diffusion maps (Coifman & Lafon, 2006), local tangent space alignment (LTSA), locally linear coordination (LLC), etc (Roweis, 2000; Donoho & Grimes, 2003; Belkin & Niyogi, 2001; Weinberger & Saul, 2006; Zhang & Zha, 2004; Coifman & Lafon, 2006; Roweis et al., 2002).

Non-convex Extensions: Instead of solving eigen-decomposition problems, both category of methods extends to non-convex variants, which can be solved by gradient descent. This significantly enriches the possibilities of DR. Most of today's popular DR methods are non-convex sparse spectral methods, including the t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008) originated from stochastic neighbor embedding (SNE) (Hinton & Roweis, 2002), and its variants: Barnes-Hut t-SNE, fast Fourier-transform-accelerated interpolation-based t-SNE) (van der Maaten, 2013; Linderman et al., 2019). More methods are proposed in recent years to re-formulate the similarity and loss functions, including among others LargeVis (Tang et al., 2016), uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), TriMAP (Amid & Warmuth, 2019) and PaCMAP (Wang et al., 2021). PHATE (Moon et al., 2019) and multiscale PHATE (Kuchroo et al., 2022) were proposed based on diffusion probability (Coifman et al., 2005). We will expand specifically on UMAP in Section 2.2, as our proposed SpaceMAP largely inherits its framework.

Finally, we briefly note that there also exist parametric DR methods, which amortize the computation on parameters of neural networks (NN) instead of directly on embedding. This category includes the early parametric t-SNE method (van der Maaten, 2009), autoencoder (Hinton & Salakhutdinov, 2006), deep learning multidimensional projection (Espadoto et al., 2020), and deep recursive embedding (Zhou et al., 2022).

2.2. UMAP: A Contrastive Learning Perspective

UMAP is a newly established visualization method which has gained popularity in recent years (McInnes et al., 2018). In its original presentation UMAP is rooted in the theory of Riemannian geometry and algebraic topology. The mathematical savviness however casts certain degree of opaqueness, and has initiated many discussions and debates in the community (Damrich & Hamprecht, 2021a; Kobak & Linderman, 2021; Wang et al., 2021). The core question is: what has led to UMAP's unique success in visualization? Here we take a fresh look on UMAP from the lens of contrastive learning, a category of unsupervised learning methods that effectively learns representations without labels (Wu et al., 2018; Ye et al., 2019; He et al., 2020; Tian et al., 2020; Chen et al., 2020; Chen & He, 2021).

Like t-SNE, UMAP starts from computing similarities, denoted as w_{ij} representing the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the high-dimensional space, and v_{ij} between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ in the low-dimensional space. The original loss function of UMAP is defined as:

$$\mathcal{L}_{\text{umap}} = \sum_i \sum_{j \neq i} \left(w_{ij} \log \frac{w_{ij}}{v_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - v_{ij}} \right) \tag{1}$$

Given that w_{ij} is only dependent on $\boldsymbol{x}$, not on $\boldsymbol{y}$, the effective loss to derive gradient on $\boldsymbol{y}$ is:

$$\mathcal{L}'_{\text{umap}} = -\sum_i \sum_{j \neq i} \left(w_{ij} \log v_{ij} + (1 - w_{ij}) \log(1 - v_{ij}) \right) \tag{2}$$

Here the first term can be interpreted as attractive force, and the second term repulsive force, realized by negative sampling (e.g. sampling the non-neighbors ourside KNN) (McInnes et al., 2018).[1] Negative sampling is the essence of contrastive learning: if $\boldsymbol{x}_j$ belongs to the KNN of $\boldsymbol{x}_i$ (category *similar*), the first term asserts that their representations $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are also close; if $\boldsymbol{x}_j$ is outside of the KNN of $\boldsymbol{x}_i$ (category *dissimilar*), the second term asserts that their representations are distant. Here the representation is the 2D output $\boldsymbol{y}$, instead of intermediate tensors in neural networks for representation learning. This hidden contrastive mechanism likely underlies the success of UMAP to a large extent.

3. Method

3.1. Overview

We follow the generic framework of graph matching: first define the pairwise similarities P_{ij} among the high-dimensional data points $\boldsymbol{x}_i$, $i = 1, ..., N$, and Q_{ij} among the low-dimensional data points $\boldsymbol{y}_i$, then optimize a contrastive loss function $\mathcal{L}(P_{ij}, Q_{ij})$ with respect to $\boldsymbol{y}_i$'s. The framework resembles that of UMAP; the major novelty of this work, however, is a principled and explainable way to compute similarities that construct the two graphs.

[1] We note that by using negative sampling to compute the repulsive force, UMAP's true loss is different from the one in the original form, as closely examined in (Damrich & Hamprecht, 2021b).

3.2. SpaceMAP

Space Expansion: The mismatch of capacity (expressiveness of data) between high- and low-dimensional space is the fundamental difficulty of DR. Here we define "capacity" as a Hausdorff measure of volume in the D-dimensional space:

Definition 3.1 (Space Capacity)**.** *Let $R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$ be the distance between data point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the D-dimensional space. The space capacity $\mathcal{V}_D(R_{ij})$ from point i to point j is defined as the volume of a D-dimensional ball with a radius of R_{ij}.*

$l(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can be any valid distance metric such as Euclidean, Manhattan, cosine, etc. Throughout this paper we use Euclidean distance for simplicity. In Euclidean space, the capacity $\mathcal{V}_D(R_{ij})$ is simply the volume of a D-dimensional hyper-sphere $S_D(R_{ij})$:

$$\mathcal{V}_D(R_{ij}) = \frac{\pi^{D/2}}{\Gamma(D/2+1)} R_{ij}^D \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function.

To match capacity between two spaces of different dimensionality, we can change the definition of "distance". We hence introduce the concept of *equivalent extended distance (EED)*:

Definition 3.2 (Equivalent Extended Distance: EED)**.** *Let $R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$ be the distance between data point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the D-dimensional space. The equivalent extended distance (EED) $\tilde{\mathcal{R}}_{ij,D\to d}$ is defined as the equivalent distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in d-dimensional space such that the Space Capacity matches: $\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij})$*

Example 3.1 (2-dimensional EED of Euclidean distance)**.** *Consider a D-dimensional Euclidean space, the distance between two points i and j is $R_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. To match capacity, the 2-dimensional EED of R_{ij} is: $\tilde{\mathcal{R}}_{ij,D\to 2} = \sqrt{\frac{S_D(R_{ij})}{\pi}} = \alpha R_{ij}^{D/2}$ where $\alpha = \frac{\pi^{(D-2)/4}}{\sqrt{\Gamma(D/2+1)}}$.*

It is straightforward to prove that when embedding D-dimensional Euclidean space into d-dimensional space, EED can be expressed in a simple form:

$$\tilde{\mathcal{R}}_{D\to d} = \alpha R^{\frac{D}{d}} \tag{4}$$

where α is a constant determined by D and d. In the special case of $d = D$, the distance definition is unchanged.

Intrinsic Dimension Estimation: Most natural high-dimensional data lie on or near to a low-dimensional manifold $\mathcal{M}$, with its intrinsic dimension (ID) $d \ll D$, where D is the ambient dimension. Here we briefly describe two ID estimation methods (Levina & Bickel, 2004; MacKay & Ghahramani, 2005).

Levina & Bickel 2004 proposed the Maximum Likelihood Estimation (MLE) method to estimate the ID based on constant density assumption in a small neighborhood and the Poisson process to model the random sampling in this neighborhood. The MLE method provides a way to estimate ID at point x_i in its $k-$neighborhood. Let R be the distance metric (e.g. Euclidean) and $R_{ij} \in \mathbb{R}$ be the distance between point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ under this metric, the maximum likelihood estimation (MLE) of the ID around point $\boldsymbol{x}_i$, with the distance metric of R, is computed as:

$$\hat{d}_k(x_i; R) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{R_{ik}}{R_{ij}} \right)^{-1} \tag{5}$$

where the summation is over the k-nearest neighbors of point $\boldsymbol{x}_i$. We note that $\hat{d}_k(x_i; R)$ is point-specific, dependent on k and the distance metric R. ID therefore uniquely characterizes the sub-manifolds around $\boldsymbol{x}$.

MacKay & Ghahramani 2005 proposed a robust way to estimate global ID of the whole dataset:

$$\bar{d}_k = \left(\frac{1}{N} \sum_{i=1}^{N} \hat{d}_k(x_i)^{-1} \right)^{-1} \tag{6}$$

where N is the number of data points. We use Equation 5 to estimate the point-specific local IDs and Equation 6 to estimate the global ID.

Our proposed EED fits in this framework of MLE ID estimation. By applying EED to the distance metric on a manifold of ID D, the following Proposition holds:

Proposition 3.1 (EED transforms ID provably)**.** *For any neighborhood size k, if the MLE of the intrinsic dimension around point x_i under the distance metric R is $\hat{d}_k(x_i; R) = D$, the MLE of the intrinsic dimension after applying EED to the distance metric is d: $\hat{d}_k(x_i; \tilde{\mathcal{R}}_{D\to d}) = d$.*

Proof. By replacing metric R with the EED-transformed metric $\tilde{\mathcal{R}}_{D\to d}$ (Equation 4) in Equation 5, we have:

$$\begin{aligned}
\hat{d}_k(x_i; \tilde{\mathcal{R}}_{D\to d}) &= \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{\alpha R_{ik}^{D/d}}{\alpha R_{ij}^{D/d}} \right)^{-1} \\
&= \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \frac{D}{d} \log \frac{R_{ik}}{R_{ij}} \right)^{-1} \\
&= \frac{d}{D} \hat{d}_k(x_i; R) = \frac{d}{D} D = d
\end{aligned} \tag{7}$$

□

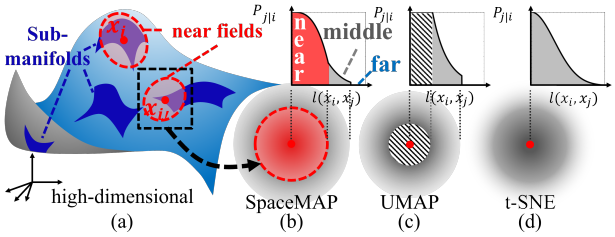

Figure 2. (a) Illustration of a hierarchical manifold: sub-manifolds (dark blue) on a global manifold (light blue). (b) The high-dimensional similarity functions P for SpaceMAP (b), UMAP (c), and t-SNE (d). SpaceMAP has subtle perception of the neighborhood in near and middle fields.

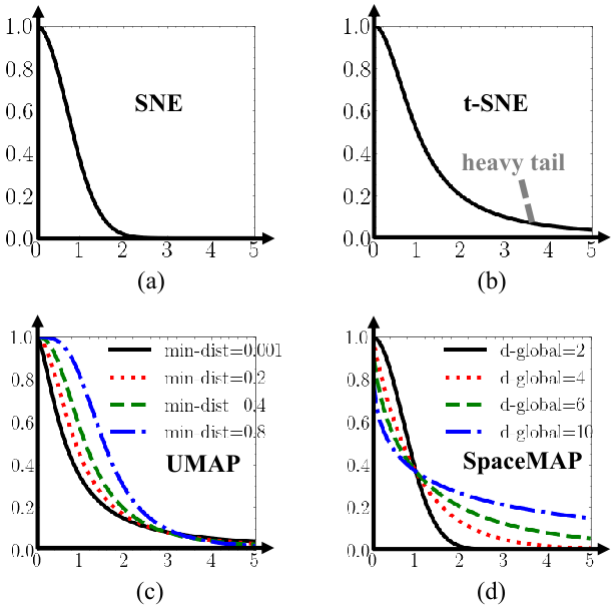

Figure 3. The low-dimensional similarity functions Q for (a) SNE (Gaussian), (b) t-SNE (t-distribution), (c) UMAP (inverse polynomial) and (d) SpaceMAP (exponential). SpaceMAP alters the tail based on dataset-specific ID.

The same holds true for the global ID in Equation 6. Therefore, when the targeted dimensionality d is set to 2 for visualization, the MLE of ID *under the EED-transformed distance metric* will change to be exactly 2. Essentially, the validity of Proposition 3.1 is due to the exponential form in EED transformation by Equation 4, while in the MLE Equation 5 and 6 the distances are taken logarithmic.

Hierarchical Manifold Approximation: Real-world data often exhibits a hierarchical structure, given the underlying data generation process (Bengio, 2013). The data manifold can be continuous, disjoint, or locally continuous and globally disjoint. It is informative to analyze both the near field (local neighborhood) and middle field (further range) of data to capture a meaningful range of inter-relationship. For SpaceMAP, we divide the space with respect to each data point i into the near field $\mathbb{S}_{i,\text{near}}$, middle field $\mathbb{S}_{i,\text{middle}}$, and far field $\mathbb{S}_{i,\text{far}}$. Given unknown scaling of data, a common approach is to set the neighborhoods by the number (or percentage in the whole dataset) of nearest neighbors. Two numbers are empirically set: $k_{\text{near}} = 20$, and $k_{\text{middle}} = 50$. Beyond the middle field, data points are usually so distant that their similarity is trivial.

We subsequently estimate two IDs: d_{local} is computed using Equation 5 for each data point using k_{near}, and d_{global} using Equation 6 for the entire dataset using $k_{\text{near}} + k_{\text{middle}}$. Pseudo code for calculating d_{local} and d_{global} is provided in Appendix A.2. The two IDs characterize the hierarchical manifold $\mathcal{M}$, and enable us to analytically compute EED. Different from t-SNE or UMAP where similarity is defined in an implicit manner, the similarities in SpaceMAP are computed from explicitly transformed distances, which provably reduces the intrinsic dimension (Proposition 3.1). Alternatively, both d_{local} and d_{global} can also be set manually with prior knowledge on data, to impose strong regularization on the DR results.

Similarity: Similarity is typically modeled as a nonlinear function of distance. Suppose the similarity function is defined as $P = S(\cdot)$ in the D-dimensional space, which is to be matched by $Q = S'(\cdot)$ in the d-dimensional space. With EED (Equation 4), matching P and Q leads to distortion of the similarity function:$S'(R) = S\left(\left(\frac{R}{\alpha}\right)^{\frac{d}{D}}\right)$, where R is the distance metric (here Euclidean) in the D-dimensional space. Given the exponential coefficient $\frac{d}{D} < 1$, such an operation amounts to deforming the high-dimensional similarity function towards a heavier-tail one (Fig. 3 d). This is exactly the same principle as t-SNE, where the use of t-distribution with heavier tail fundamentally underlies its success. In SNE, in contrast, both spaces are treated equally in terms of distance metric.

We take the Gaussian similarity function as in SNE, and define P_{ij} and Q_{ij} for SpaceMAP:

$$P_{j|i} = \begin{cases} \exp\left(-\frac{R_{ij}^{2d_{\text{local}}/d_{\text{global}}}}{\sigma_{i,\text{near}}}\right), & x_j \in \mathbb{S}_{\text{near}} \\ \exp\left(-\frac{(R_{ij}-\gamma_i)^2}{\sigma_{i,\text{middle}}}\right), & x_j \in \mathbb{S}_{\text{middle}} \\ 0, & x_j \in \mathbb{S}_{\text{far}} \end{cases}$$

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2} \tag{8}$$

$$Q_{ij} = \exp\left(-\left(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^{\frac{2}{d_{\text{global}}}}\right)^2\right) \tag{9}$$

where $P_{j|i}$ denotes the conditional probability $P(x_j|x_i)$, $R_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, d_{local} and d_{global} are the estimated local

and global ID, and γ_i is a parameter to ensure continuity from near field to middle field. Equation 8 symmetrizes the conditional probability as in t-SNE (van der Maaten, 2014). Q_{ij} takes a simple exponential form (note that the EED parameter α can be absorbed to $\boldsymbol{y}$ hence omitted), which is favorable for gradient computation during optimization. More details on how to set up $\sigma_{i,\text{near}}$, $\sigma_{i,\text{middle}}$, and γ_i are found in Appendix A.1.

The resulting hierarchical similarity function is visually simple, as illustrated in Fig. 2, where P of near, middle, and far fields are plotted for SpaceMAP (b). It can be seen that t-SNE makes no differentiation between the fields, with a single Gaussian, while UMAP better differentiates them, but with a uniform profile in the near field (1-nearest neighbor). SpaceMAP has a subtle perception of near and middle fields: the hierarchical modeling of manifold is reflected by the field-wise similarity, where exponential coefficients of R_{ij} differ depending on the dataset-specific local and global ID. In the middle field, the exponential remains 2, same as in SNE and t-SNE. In the near field, however, a scaling of $d_{\text{local}}/d_{\text{global}}$ is applied depending on the data-specific manifold property.

Fig. 3 shows the Q function for SNE, t-SNE, UMAP, and SpaceMAP. Heavier tails can be observed in the later three methods. For SpaceMAP, heavier tail naturally arises from EED.

Loss and Optimization: The loss function of SpaceMAP is similar to that of UMAP:

$$\mathcal{L} = \sum_i \sum_j \left[P_{ij} \log \frac{P_{ij}}{Q_{ij}} + (1 - P_{ij}) \log \frac{1 - P_{ij}}{1 - Q_{ij}} \right] \tag{10}$$

where the first term exerts attractive force on similar points, and the second term repulsive force on dissimilar points. We used stochastic gradient descent and negative sampling to optimize $\mathcal{L}$ as in UMAP. For motivation of using this loss we refer to Section 2.2 and the recent publication on UMAP's true loss function (Damrich & Hamprecht, 2021b).

4. Experiments and Results

4.1. Datasets and Evaluation Metrics

We benchmarked our SpaceMAP with other classical or state-of-the-art methods including PCA, Laplacian Eigenmaps, t-SNE, and UMAP. Experiments were performed on a wide range of datasets, including the standard MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), Swiss roll 1 (on the surface we dig a hole to test if visualization methods can preserve the local property on a continuous manifold), Swiss roll 2 (consisting of parallel lines to test the hierarchical manifold assumption), COIL-20 (Nene et al., 1996), RNA-seq (Tasic et al., 2018). In addition, we also tested our method on the extremely large GoogleNews Word2Vec 3 million dataset (Mikolov et al., 2013) and the prime-number-divisibility 1 million dataset(McInnes et al., 2018) (Results are presented in Appendix).

For quantitative evaluation, we computed the 20-fold cross-validated KNN classification accuracy, trustworthiness, continuity, Shepard goodness, and normalized stress to evaluate both local and global structure preservation (Espadoto et al., 2021; Nonato & Aupetit, 2019). KNN accuracy measures the local structure preservation along different neighborhood size k. Trustworthiness and continuity evaluate the local pattern of the embedding by calculating the true neighbor rate and missing neighbor rate. The Shepard goodness and the normalized stress are two measures of goodness in global structure preservation. Details on how to compute the metrics are found in (Espadoto et al., 2021). A newly proposed metric called DeMAP (denoised manifold affinity preservation) score (Moon et al., 2019) is also included and reported in Appendix A.9.

4.2. Results

Fig. 1 (b) shows the results of different methods on 3 synthetic datasets, on which the validity of visualization can be checked.[2] The first row illustrates the simplest situation where the ambient and intrinsic dimensionalities are both 2, the second row shows a dataset with ambient and intrinsic dimensionality 3, and the third row shows a Swill roll with a hole, with the ambient dimensionality of 3 and intrinsic dimensionality of 2. By correctly integrating prior information into the formulation of similarity, the SpaceMAP map achieves the best performance in all examples. We also note that SNE surpasses t-SNE and UMAP in unfolding the Swiss roll, as it correctly (albeit implicitly) assumes the intrinsic dimensionality to be 2.

We further present the visualization results on two different synthetic Swiss rolls in Fig. 4. The first Swiss roll consists of parallel curves with disconnected dots, with $d_{\text{local}} = 1$ and $d_{\text{global}} = 2$. t-SNE in Fig. 4 (a.2) maintains the local parallelism but fails to preserve the global continuity. UMAP (a.4) enforces local connectivity but ignores the parallelism between curves. The performance improves on larger neighborhood (a.5) but remains non-ideal. Given the local and global ID, SpaceMAP (a.6) correctly unfolds both local and global structure, surpassing t-SNE (a.3) and UMAP (a.5) with equivalent number of neighbors. For the Swiss roll with a hole, SpaceMAP (b.6) remains the only method that correctly unfolds local and global structures. Both t-SNE and UMAP tend to break around the hole.

[2]Despite a number of quantitative evaluation metrics as described in Section 4.1, it is typically difficult to know the ground-truth manifold for real-world data. Therefore synthetic datasets serve well to check the theoretical validity of the proposed EED.

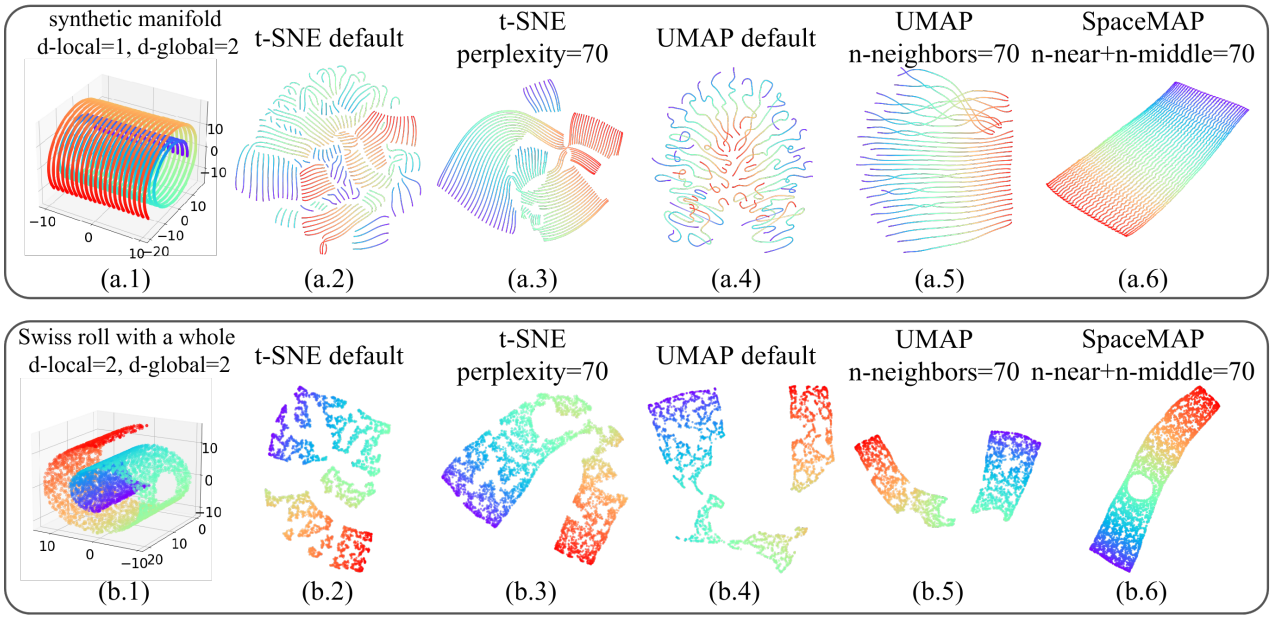

Figure 4. Visualization results of the synthetic datasets by t-SNE, UMAP and SpaceMAP. Upper panel: The synthetic Swill roll with $d_{\text{local}} = 1$ and $d_{\text{global}} = 2$. Lower panel: Swiss roll with a hole, with $d_{\text{local}} = 2$ and $d_{\text{global}} = 2$.

Table 1. Quantitative measure of visualization performance by SpaceMAP and other reference methods: M_t, M_c, M_s and M_σ indicate the trustworthiness, continuity, Shepard goodness and normalized stress, respectively. M_t and M_c are local metrics (left side of the table) while M_s and M_σ are global metrics (right side of the table).

Local Metrics (M_t and M_c)

Experiments		PCA	Laplacian	t-SNE	UMAP	SpaceMAP
MNIST	M_t	0.74	0.81	**0.98**	0.96	0.97
	M_c	0.94	0.93	0.97	0.97	**0.98**
FMNIST	M_t	0.91	0.89	0.98	0.98	**0.99**
	M_c	0.98	0.87	0.98	**0.99**	**0.99**
COIL-20	M_t	0.86	0.92	0.99	0.99	**1.00**
	M_c	0.93	0.79	0.99	0.99	**1.00**
RNA-Seq	M_t	0.89	0.85	**1.00**	0.99	**1.00**
	M_c	0.98	0.94	0.99	**1.00**	**1.00**

Global Metrics (M_s and M_σ)

Experiments		PCA	Laplacian	t-SNE	UMAP	SpaceMAP
MNIST	M_s	**0.50**	0.23	0.35	0.32	0.35
	$1 - M_\sigma$	0.42	0.33	0.54	0.53	**0.55**
FMNIST	M_s	**0.88**	0.41	0.64	0.58	0.67
	$1 - M_\sigma$	0.65	0.36	0.66	0.62	**0.67**
COIL-20	M_s	**0.89**	0.72	0.80	0.56	0.61
	$1 - M_\sigma$	0.55	0.19	**0.62**	0.58	0.60
RNA-Seq	M_s	**0.80**	0.11	0.61	0.22	0.63
	$1 - M_\sigma$	**0.60**	0.15	0.59	0.50	0.52

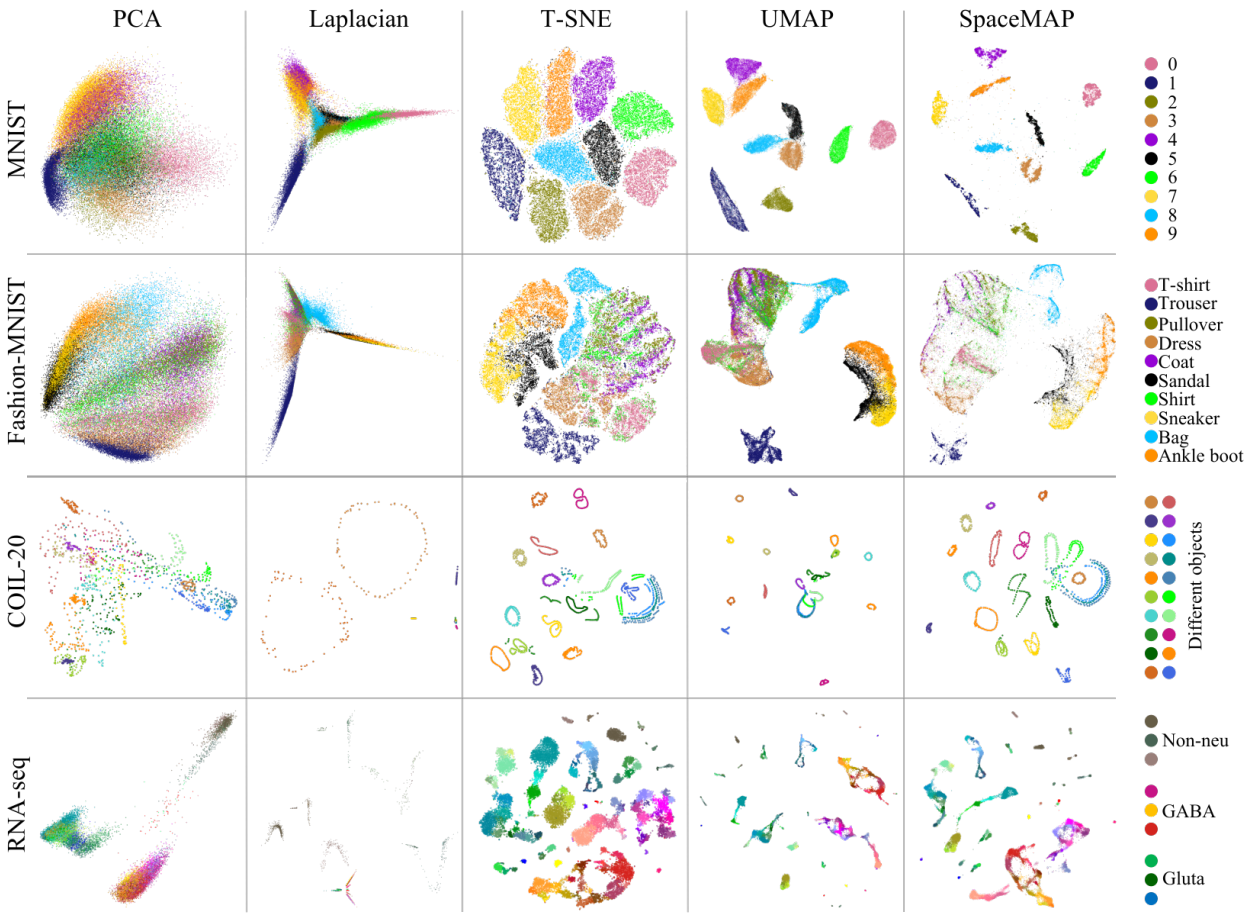

Figure 5. Visualization results by PCA, Laplacian eigenmaps, t-SNE, UMAP, and SpaceMAP, on the high-dimensional datasets MNIST, Fashion-MNIST, COIL-20, and RNA-seq.

Fig. 5 presents 2D visualization of the real-world datasets, comparing PCA, Laplacian Eigenmap, t-SNE, UMAP, and SpaceMAP. SpaceMAP shows competitive performance on data of disjoint manifolds. We also report the quantitative metrics in Table 1, showing superior local and global metrics by SpaceMAP. We also report the estimated global ID in Table 3 for different real-world datasets. In Appendix A.6 we present the visualization results of GoogleNews Word2Vec 3M and prime-number-divisibility 1M in Fig. 10 and 11, where we observe interesting and interpretable hierarchies in the 2 gigantic datasets. Runtime is reported in Table A.8.

4.3. Ablation Studies

We present the ablation study on the hierarchical manifold assumption. By setting $d_{\text{local}} = d_{\text{global}}$, we could remove the hierarchical manifold assumption, imposing the same similarity functions in near- and middle-fields. Fig. 6 shows the comparison between SpaceMAP visualization with and without hierarchy. A close examination shows that SpaceMAP with hierarchy reveals more subtle structures in data: the sub-clusters within the same class (e.g. different ways of writing a digit) are more pronounced. This subtle improvement is also reflected by the quantitative measure in Table 5, including the KL divergence of local similarity (M_{KL}) in Appendix A.5, which measures the preservation of sub-manifolds in data, and local normalization stress (M'_σ), a local version of normalized stress (Espadoto et al., 2021).

The hyper-parameters of k_{near} and k_{middle} were empirically set to 20 and 50. We show in Appendix Fig. 7 that the results are not particularly sensitive to the hyper-parameter choices, because what essentially matters is d_{local} and d_{global}, see Appendix A.2 for details. d_{local} is computed from k_{near}-NN, and the value of d_{local} can be stable over a range of k_{near} choices. d_{global} is a value over all data points, hence stable by nature.

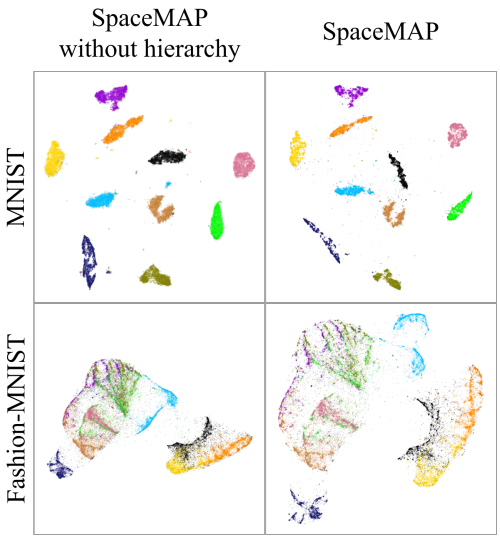

Figure 6. Visualization results of **the ablation study on hierarchical manifold assumption in SpaceMAP** with MNIST and Fashion-MNIST datasets. SpaceMAP with hierarchy reveal more details of the local pattern (i.e. sub-clusters within a class), as also quantitatively evaluated in Table 5.

Table 2. Quantitative results of **the ablation study on hierarchical manifold assumption in SpaceMAP** with two quantitative metrics: KL divergence of local similarity (M_{KL}) and local normalization stress (M'_{σ}). For both metrics, lower value is seen as an indicator of better local structure preservation quality.

Local structure preservation quality

Experiments		UMAP	SpaceMAP (no hierarchy)	SpaceMAP
MNIST	M_{KL}	1.0014	0.9961	**0.9915**
	$M_{\sigma'}$	0.9413	0.6714	**0.6576**
FMNIST	M_{KL}	1.3239	1.2915	**1.2594**
	$M_{\sigma'}$	0.9277	0.9403	**0.9218**

5. Discussion and Conclusions

We proposed a new visualization method called SpaceMAP, which visualizes data of any dimensionality on a 2-dimensional map. Different from previous DR methods, we analytically derived a transformation of distance between high- and low-dimensional spaces to match their capacity. We further show that the transformation provably reduces the intrinsic dimension of high-dimensional data, within the framework of maximum likelihood intrinsic dimensionality estimation.

We argue that all successful DR methods, including t-SNE and UMAP, make use of the rationale of *space expansion* to realize high-dimensional data visualization on a drastically reduced dimension. However, previous methods did such transformation in an implicit manner, with predefined similarity function (as long as they have a heavy tail, e.g. t-distribution for t-SNE). Despite their empirical success, it is difficult to examine the optimality or impose prior knowledge. For SpaceMAP, the capability to integrate useful data prior explains the success on all synthetic datasets evaluated in this work.

Furthermore, SpaceMAP models similarity in a hierarchical, dataset-specific manner, hence compatible to both continuous and disjoint manifolds. Data property should be an integral part of DR methods, and SpaceMAP presents a way to incorporate it into the construction of similarity functions.

Finally, we motivated the use of UMAP loss for SpaceMAP by showing its hidden connection to contrastive learning. Recently, contrastive learning has demonstrated substantial success in representation learning. A 2-dimensional visualization map can be regarded as representation of the original high-dimensional data; the contrastive mechanism likely explains the success of UMAP, as well as SpaceMAP.

In conclusion, we have proposed a novel visualization method called SpaceMAP, based on a principled way to transform distances between high- and low-dimensional spaces. It further models the hierarchical structure in a dataset-specific manner based on the local and global intrinsic dimensionalities of data. Our experiments demonstrated its excellent performance on a wide range of synthetic and real-world datasets, in comparison with other state-of-the-art DR methods.

References

Abdolali, M. and Rahmati, M. Neither global nor local: A hierarchical robust subspace clustering for image data. *CoRR*, abs/1905.07220, 2019. URL http://arxiv.org/abs/1905.07220.

Amid, E. and Warmuth, M. K. Trimap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *CoRR*, abs/1905.12784, 2019. URL http://arxiv.org/abs/1905.12784.

Belkin, M. and Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Nips*, volume 14, pp. 585–591, 2001.

Bengio, Y. Deep learning of representations: Looking forward, 2013.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual rep-

resentations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

Chen, Y., Crawford, M., and Ghosh, J. Improved nonlinear manifold learning for land cover classification via intelligent landmark selection. In *2006 IEEE International Symposium on Geoscience and Remote Sensing*. IEEE, jul 2006. doi: 10.1109/igarss.2006.144.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, jul 2006. doi: 10.1016/j.acha.2006.04.006.

Coifman, R. R., S. Lafon, A. B., Lee, Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 21:7426–7431, 2005.

Damrich, S. and Hamprecht, F. A. UMAP does not reproduce high-dimensional similarities due to negative sampling. *CoRR*, abs/2103.14608, 2021a. URL https://arxiv.org/abs/2103.14608.

Damrich, S. and Hamprecht, F. A. On UMAP's true loss function. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2021b.

Donoho, D. L. and Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, apr 2003. doi: 10.1073/pnas.1031596100.

Espadoto, M., Hirata, N. S. T., and Telea, A. C. Deep learning multidimensional projections. *Information Visualization*, 19(3):247–269, may 2020. doi: 10.1177/1473871620909485.

Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S. T., and Telea, A. C. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, mar 2021. doi: 10.1109/tvcg.2019.2944182.

Giraud, C. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Hinton, G. and Roweis, S. T. Stochastic neighbor embedding. In *NIPS*, volume 15, pp. 833–840. Citeseer, 2002.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, jul 2006. doi: 10.1126/science.1127647.

Kobak, D. and Linderman, G. C. UMAP does not preserve global structure any better than t-sne when using the same initialization. *Nature Biotechnology*, abs/210339: 156–157, 2021.

Kuchroo, M., Huang, J., Wong, P., Grenier, J.-C., Shung, D., Tong, A., Lucas, C., Klein, J., Burkhardt, D. B., Gigante, S., Godavarthi, A., Rieck, B., Israelow, B., Simonov, M., Mao, T., Oh, J. E., Silva, J., Takahashi, T., Odio, C. D., Casanovas-Massana, A., Fournier, J., Team, Y. I., Farhadian, S., Cruz, C. S. D., Ko, A. I., Hirn, M. J., Wilson, F. P., Hussin, J. G., Wolf, G., Iwasaki, A., and Krishnaswamy, S. Multiscale PHATE identifies multimodal signatures of COVID-19. *Nat Biotechnol*, 40:681–691, 2022.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Levina, E. and Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pp. 777–784, Cambridge, MA, USA, 2004. MIT Press.

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245, feb 2019. doi: 10.1038/s41592-018-0308-4.

MacKay, D. J. and Ghahramani, Z. Comments on 'maximum likelihood estimation of intrinsic dimension' by e. levina and p. bickel (2004). 2005. URL http://www.inference.org.uk/mackay/dimension/.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. February 2018.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn,

M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol*, 37: 1482–1492, 2019.

Nene, S. A., Nayar, S. K., Murase, H., et al. Columbia object image library (coil-100). 1996.

Nonato, L. G. and Aupetit, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, aug 2019. doi: 10.1109/tvcg.2018.2846735.

Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *CoRR*, abs/2104.08894, 2021. URL https://arxiv.org/abs/2104.08894.

Ross, D. A., Lim, J., Lin, R.-S., and Yang, M.-H. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, aug 2007. doi: 10.1007/s11263-007-0075-7.

Roweis, S. et al. Automatic alignment of hidden representations. In *Sixteenth Annual Conference on Neural Information Processing Systems, Vancouver, Canada*, volume 15, pp. 841–848, 2002.

Roweis, S. T. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, dec 2000. doi: 10.1126/science.290.5500.2323.

Tang, J., Liu, J., Zhang, M., and Mei, Q. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, apr 2016. doi: 10.1145/2872427.2883041.

Tasic, B., Yao, Z., Graybuck, L. T., Smith, K. A., Nguyen, T. N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M. N., Viswanathan, S., Penn, O., Bakken, T., Menon, V., Miller, J., Fong, O., Hirokawa, K. E., Lathia, K., Rimorin, C., Tieu, M., Larsen, R., Casper, T., Barkan, E., Kroll, M., Parry, S., Shapovalova, N. V., Hirschstein, D., Pendergraft, J., Sullivan, H. A., Kim, T. K., Szafer, A., Dee, N., Groblewski, P., Wickersham, I., Cetin, A., Harris, J. A., Levi, B. P., Sunkin, S. M., Madisen, L., Daigle, T. L., Looger, L., Bernard, A., Phillips, J., Lein, E., Hawrylycz, M., Svoboda, K., Jones, A. R., Koch, C., and Zeng, H. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, oct 2018. doi: 10.1038/s41586-018-0654-5.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, dec 2000. doi: 10.1126/science.290.5500.2319.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, aug 1999. doi: 10.1111/1467-9868.00196.

van der Maaten, L. Learning a parametric embedding by preserving local structure. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 384–391, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/maaten09a.html.

van der Maaten, L. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*, 2013.

van der Maaten, L. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15 (93):3221–3245, 2014. URL http://jmlr.org/papers/v15/vandermaaten14a.html.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

van der Maaten, L., Postma, E., and Van den Herik, J. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10:66–71, 2009.

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. 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*, 22(201):1–73, 2021.

Weinberger, K. Q. and Saul, L. K. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, volume 6, pp. 1683–1686, 2006.

Wright, J. and Ma, Y. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2022.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhang, Z. and Zha, H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, jan 2004. doi: 10.1137/s1064827502419154.

Zhou, Z., Zu, X., Wang, Y., Lelieveldt, B. P. F., and Tao, Q. Deep recursive embedding for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 28(2):1237–1248, 2022. doi: 10.1109/TVCG.2021.3122388.

Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, jun 2006. doi: 10.1198/106186006x113430.

A. Appendix

A.1. Parameters in SpaceMAP Similarity Functions

We provide more details on the computation of SpaceMAP similarity in Equation 8 and 9. In Equation 8, the border point x_{k_1} (the k_1-nearest points to x_i) at the intersection of near field and middle field satisfies $P_{k_1|i} = \eta$, where $\eta \in (0, 1)$ is a hyper-parameter representing the conditional probability differentiating near- and middle-fields. Therefore $\sigma_{i,\text{near}}$ can be computed as:

$$\sigma_{i,\text{near}} = -\frac{R_{ik_1}^{2d_{\text{local}}/d_{\text{global}}}}{\ln \eta} \tag{11}$$

The computation of $\sigma_{i,\text{middle}}$ in Equation 8 follows similar rationales in UMAP (McInnes et al., 2018) to ensure that middle-field neighbors take up certain probability mass: $\sum_{j=k_1+1}^{k_1+k_2} P_{j|i}(\sigma_{i,\text{middle}}) = \eta \log(k_2)$, where k_1 is the number of nearest neighbors in the near field, and k_2 is the number of nearest neighbors in the middle field. γ_i in Equation 8 can be accordingly computed to satisfy continuity of P_{ij}:

$$\gamma_i = R_{ik_1} - \sqrt{-\sigma_{i,\text{middle}} \ln \eta} \tag{12}$$

A.2. MLE Pseudo Code

Algorithm 1 describes the maximum likelihood estimation (MLE) of the intrinsic dimensions. The inputs of Algorithm 1 are the distance function of the nearest neighbors of each point knn_dist, the number of nearest neighbors in the near field and the middle field of each data k_{near}, k_{middle}. The outputs of the algorithm are the MLE of the intrinsic dimensions of the local sub-manifolds and the global manifold, on which the data distributes.

Algorithm 1 Maximum likelihood estimation (MLE) of the intrinsic dimension

function MLEINTRINSICDIMENSION($knn\_dist, n_{\text{near}}, n_{\text{middle}}$)
  $k_1 \leftarrow k_{\text{near}}$
  $k_2 \leftarrow k_{\text{near}} + k_{\text{middle}}$
  for $i \leftarrow 1, ...,$N **do**
    $R_{i,k_1} \leftarrow knn\_dist[i, k_1]$
    $R_{i,k_2} \leftarrow knn\_dist[i, k_2]$
    $\hat{d}_{\text{local}}[i] \leftarrow (\frac{1}{k_1-1} \sum_{j=1}^{k_1-1} \log \frac{R_{i,k_1}}{R_{i,j}})^{-1}$
    $\hat{d}_{\text{middle}}[i] \leftarrow (\frac{1}{k_2-1} \sum_{j=1}^{k_2-1} \log \frac{R_{i,k_2}}{R_{i,j}})^{-1}$
  end for
  $\hat{d}_{\text{global}} \leftarrow \left(\frac{1}{N} \sum_{i=1}^{N} \hat{d}_{\text{middle}}[i]^{-1}\right)^{-1}$
  return $\hat{d}_{\text{local}}, \hat{d}_{\text{global}}$
end function

A.3. SpaceMAP hyper-parameters

SpaceMAP has a small set of hyper-parameters, including the number of nearest neighbor in the near field k_{near}, number of nearest neighbor in the middle field k_{middle}, and the similarity $\eta \in (0, 1)$ at the border of near- and middle-field. We present the visualization results for SpaceMAP at different number of these three hyper-parameters Fig. 7 and Fig. 8. We observe that the final visualization results are not particular sensitive to the choice of parameters.

A.4. Quantitative Evaluation of KNN Accuracy

We compare the 20-fold cross-validated KNN classification accuracy as a function of neighborhood size k on different datasets, namely, MNIST, COIL-20, RNA-seq. In general, SpaceMAP outperforms UMAP in different neighborhood sizes, while t-SNE is competitive to SpaceMAP when the neighborhood size is small. See the results in Fig. 9.

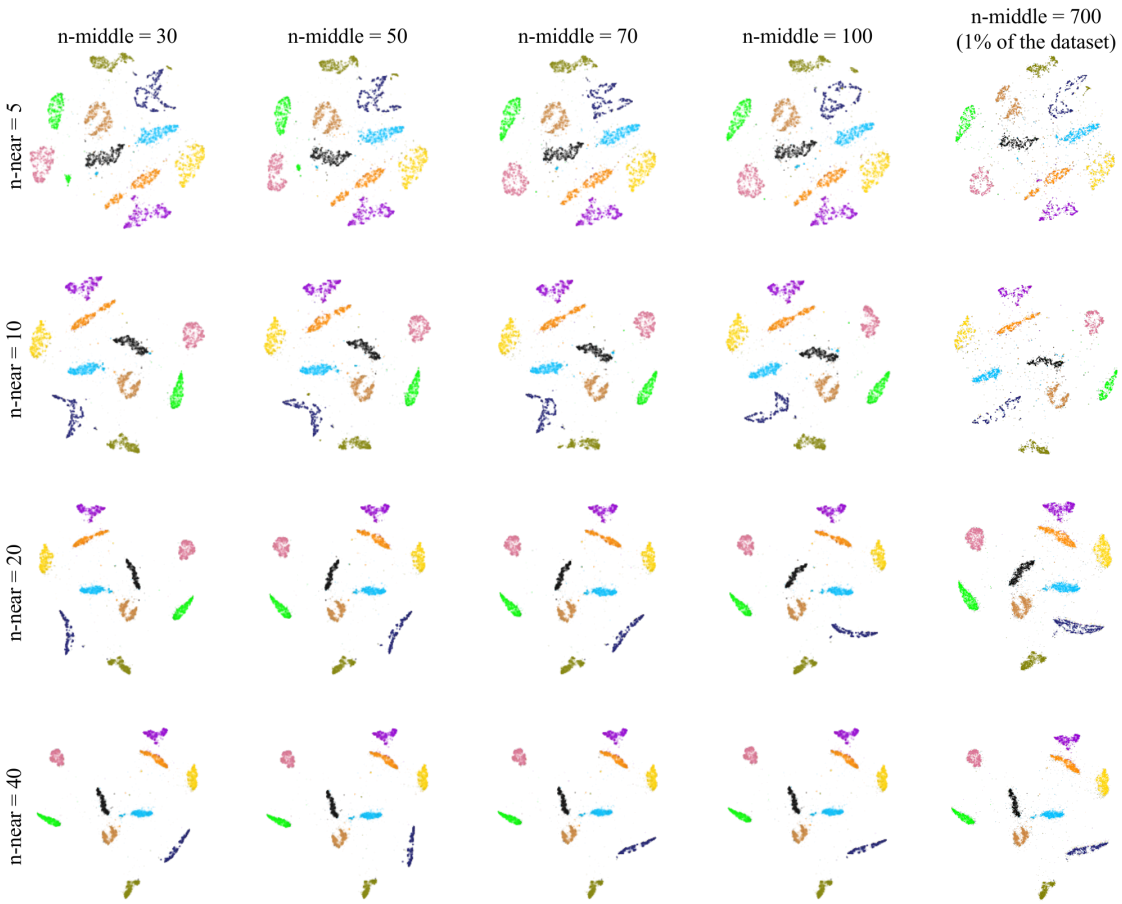

Figure 7. **Hyper-parameter selection in SpaceMAP**. The influence of n_{near} and n_{middle}, the two main hyper-parameters of SpaceMAP.

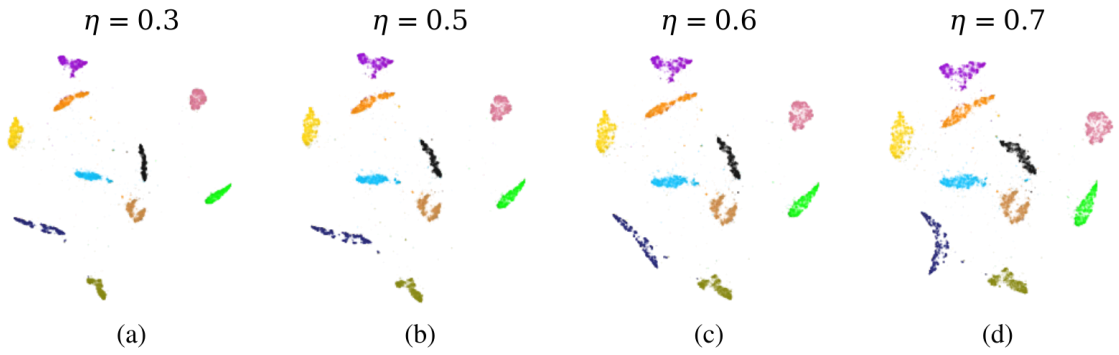

Figure 8. **Hyper-parameter selection in SpaceMAP**. The influence of η, similarity at the intersection of near- and middle field.

A.5. KL Divergence of Local Similarity

We introduce a new metric, called KL divergence of local similarity, to evaluate the preservation of local sub-manifolds in data. It computes the KL divergence between the KNN pairwise similarities (modeled as posterior probability) in the

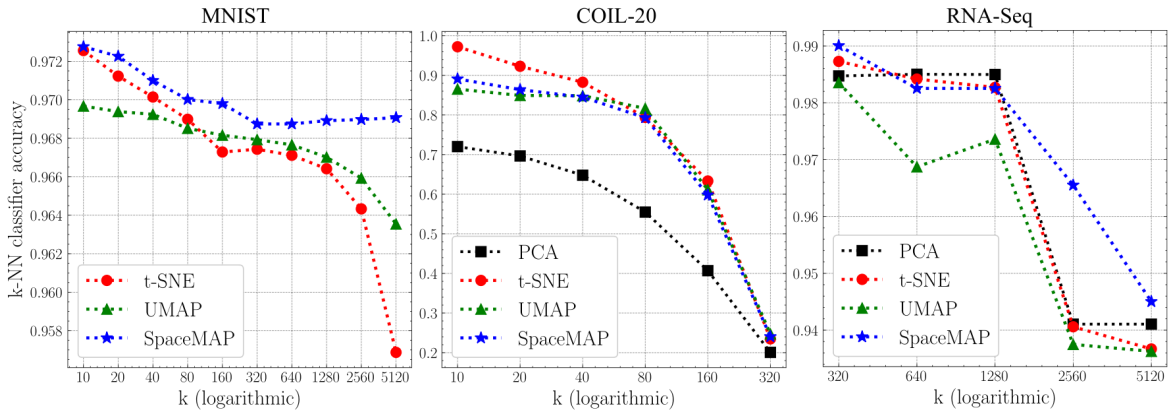

Figure 9. 20-fold cross-validated KNN classification accuracy as a function of neighborhood size k on different datasets. From left to right: MNIST, COIL-20, and RNA-seq.

high-dimensional space and the embedded low-dimensional space:

$$M_{\mathrm{KL}} = \mathrm{KL}(R\|r) = \sum_i \sum_j R_{ij} \log \frac{R_{ij}}{r_{ij}} \tag{13}$$

$$\tag{14}$$

where $R_{ij} = \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sum_{k \neq l} \|\boldsymbol{x}_k - \boldsymbol{x}_l\|}$ and $r_{ij} = \frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|}{\sum_{k \neq l} \|\boldsymbol{y}_k - \boldsymbol{y}_l\|}$ are the pairwise similarities in high-dimensional space and their embedding in low-dimensional space. This metric characterizes the local data structure preservation, better than other local metrics such as the KNN accuracy, trustworthiness, and continuity. We selected all the KNN neighbors in SpaceMAP (default setting is $k = n_{near} + n_{middle} = 20 + 50 = 70$) and set UMAP with $n_{neighbors} = 70$ as the baseline in Table 5. We note that this metric is defined independent of our loss function, hence unbiased.

A.6. Visualizing 3 Million Data

Like UMAP, SpaceMAP has the ability to embed millions of data with high dimensionality. Figure **??** shows the embedding result of the GoogleNews Word2Vec 3 million dataset by UMAP and SpaceMAP. The dataset contains 3 million pre-trained word vectors with 300 dimensions, widely used in the natural language processing (NLP) community. In the SpaceMAP result, several clear semantic clusters emerge. Zooming in the cluster with geographic words, we observed a hierarchical structure based on geographic meaning like the continental location. Some clusters even demonstrated a profound understanding on the cultural relationship among countries. For example, the South American country names are in the same sub-cluster as Spain and Portugal, and the North African country names are in the same sub-cluster as other Islamic countries. Figure 11 shows the embedding result of integers from 0 to 1,000,000 represented as binary vectors indicating their prime divisibility, inspired by McInnes et al. (2018). The result has a subtle fractal structure.

A.7. Intrinsic Dimensions of Real-world Datasets

Here we provide the intrinsic dimensions of the datasets (MNIST, Fashion-MNIST, COIL-20 and RNA-Seq) as calculated by Eq. 6 in Table 3. Further information is provided in Pope et al. 2021.

Table 3. The global intrinsic dimensions of the datasets (MNIST, Fashion-MNIST, COIL-20 and RNA-Seq) calculated by Eq. 6.

	MNIST	**Fashion-MNIST**	**COIL-20**	**RNA-Seq**
ID	8.8	9.3	3.3	16.7

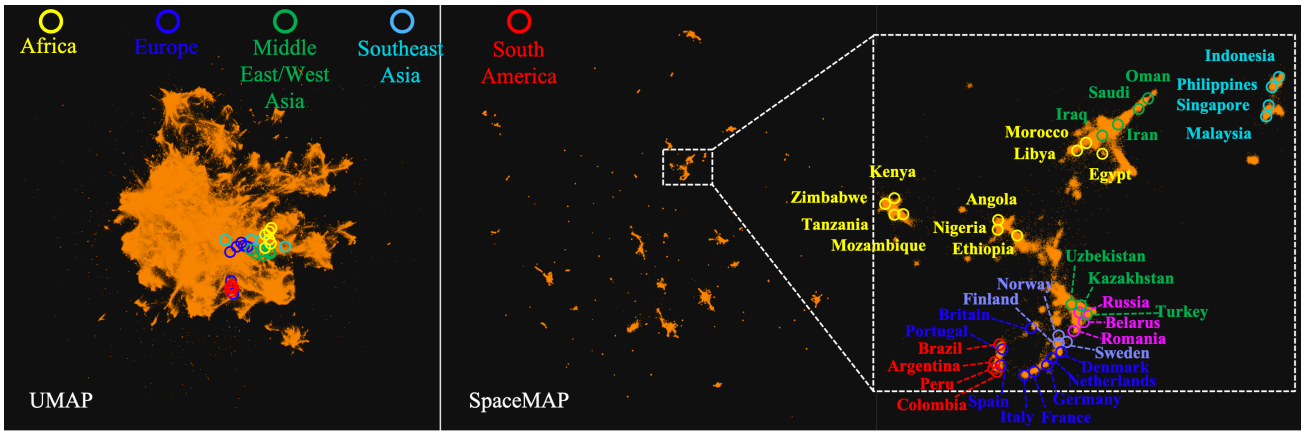

Figure 10. **The Word Map: visualization of the GoogleNews Word2Vec 3 Million Dataset** by UMAP (left) and SpaceMAP (middle). The SpaceMAP result is zoomed in for better visualization of the word semantics (right). The same data points are also marked in UMAP. We observed a better hierarchical embedding by SpaceMAP compared with UMAP. Remarkable, SpaceMAP demonstrated a profound understanding on the cultural relationship among countries.

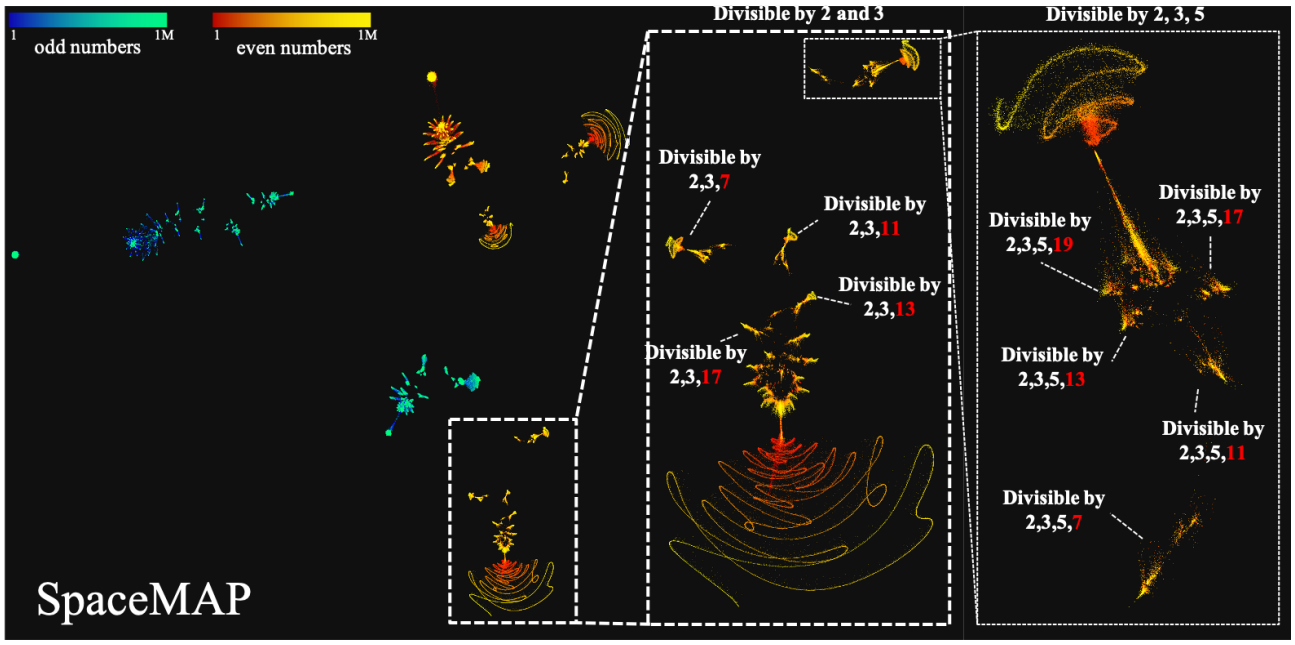

Figure 11. **The Prime Number Divisibility:** SpaceMAP embeds integers from 0 to 1,000,000, as represented by the sparse binary vectors of prime number divisibility. Details are found in the original UMAP paper (McInnes et al., 2018). Sub-structures of data can be identified, which corresponds to divisibility of multiple prime numbers (zoomed-in windows). The map has a beautiful fractal structure.

A.8. Runtime and Implementation

The runtime of the visualization methods (t-SNE, UMAP and SpaceMAP) are provided in Table 5. We implemented all the DR methods on a Ubuntu 20.04 LTS workstation platform with AMD 3900x 4.2GHz 12-core CPU, 64GB DDR4 RAM and NVidia RTX 3090 24GB GPU. For t-SNE, we use Scikit-learn Barnes-Hut t-SNE implementation with PCA initialization and the default perplexity is 30. For UMAP, we use the original *umap-learn* implementation with spectral embedding initialization, the default number of neighbors for each point is n-neighbors=15.

Table 4. Runtime of the methods (t-SNE, UMAP and SpaceMAP)

dataset	t-SNE	UMAP	SpaceMAP
MNIST	210s	34s	85s
Fashion-MNIST	202s	43s	88s
COIL-20	3s	3s	2s
RNA-Seq	44s	12s	39s

A.9. Manifold Preservation

We additionally report the DeMAP (denoised manifold affinity preservation) score proposed in (Moon et al., 2019) to evaluate the manifold preservation quality by different DR methods.

Table 5. The DeMAP score of different DR methods (PCA, t-SNE, UMAP and SpaceMAP). Higher DeMAP score indicates better manifold preservation. The synthetic Swiss roll dataset is shown in Fig. 4 a.1

dataset	PCA	t-SNE	UMAP	SpaceMAP
RNA-Seq	**0.8415**	0.6012	0.6315	0.7250
COIL-20	0.5080	0.6066	0.4086	**0.8022**
Swiss roll with a hole	0.4153	0.5717	0.8513	**0.9111**
synthetic Swiss roll	0.3929	0.4360	0.2969	**0.9579**